The evolution of mRNA splicing in mammals

by

Jason Jay Merkin

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author . . . .
Department of Biology
March 12, 2014

Signature redacted

Certified by . . . .
Christopher B. Burge
Professor
Thesis Supervisor

Signature redacted

Accepted by . . . .
Amy Keating
co-Chair, Biology Graduate Committee

The evolution of mRNA splicing in mammals

by

Jason Jay Merkin

Submitted to the Department of Biology
on March 12, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, I describe investigations into the evolution of splicing in mammals. I first investigate a small class of alternative splicing events, tandem splice sites, and show how they are used to introduce and remove coding sequence in a species-specific manner. I then describe the generation and analysis of a large RNA-seq dataset from 9 matched tissues in 5 species, with the aim to investigate the evolution of splicing in mammals. I first investigate the evolution of exons that predate the most ancient divergence of species studied, finding that their splicing is frequently poorly conserved. For a subset of these exons, I identify unique regulatory properties and provide evidence linking alternative splicing to phosphorylation potential of proteins. I then consider sources of novel exons, in these species. I use these and other published data to identify one way in which splicing of novel exons impacts the biology of the cell. I also present evidence implicating genomic indels in exon creation and splicing variation.

Thesis Supervisor: Christopher B. Burge
Title: Professor

for Darya and Lilah,

for always giving me a smile.

and for Cal and Yaffa,

for always giving.

Acknowledgments

I would like to open by thanking my wife, Darya. After starting a family with a graduate student and dealing with issues that arose with such aplomb, I must begin by thanking her.

I would like to thank my advisor, Chris. He has been a wonderful mentor, tolerating my idiosyncracies, teaching me to become a rigorous computational biologist, and providing support and encouragement regardless of the direction my life or research took.

I would like to thank my wonderful daughter, Lilah. She continues to inspire me each and every minute I spend with her, and gave me an extra impetus to get myself in gear and graduate.

I would like to thank my parents: Sara and Ted; and Joel and Karen. They have always pushed me to achieve (sometimes to my chagrin) and supported me even when I didn't want them to do so.

I would like to thank my grandmother, Yaffa. She has given my everything she could, simply because she could.

I would like to thank my grandfather, Cal. He gave me everything he could, and then some that he couldn't. I blame him for getting a PhD.

I would like to thank my high school biology teacher, Rich Tempsick. It was in his class that I ran my first gels (and learned what agarose was). I reflect back on it as where I began my PhD training and blame him as well.

I would also like to thank my friends. There are too many to list, but they were always there for a coffee or a beer when I needed to complain about everyone I just thanked.

Contents

List of Figures

Chapter 1

Introduction

mRNA splicing was discovered more than three decades ago in viruses (8, 20) and later found to be ubiquitous in eukaryotes (12, 93, 58, 84). Despite its near universal presence in eukaryotic organisms (89, 98, 74) and involvement in many areas of biology and disease (22, 31), the conservation of splicing patterns remained unclear (16, 76, 113). DNA microarrays have been applied successfully to investigate the evolution of gene expression patterns (17, 117, 16, 91), but are not as suitable to investigating splicing due in part to technical reasons (108). The application of high-throughput sequencing to splicing studies has led to a revolution in the field and enabled global investigations into the evolution of splicing and alternative splicing in mammals that I will describe in this thesis.

1.1 mRNA splicing

Genes in virtually all eukaryotes are found in discontinuous segments in the genome. The process of removing the intervening sequences ("introns") from the sequences encoding the final gene product ("exons") following transcription by RNA polymerase II is known as pre-mRNA splicing (9, 88, 70, 19, 72). After the completion of splicing, capping, cleavage and polyadenylation, and in some cases other RNA modifications, the processed mRNA is exported from the nucleus to the cytoplasm (72). Splicing consists of a two-step reaction catalyzed by the spliceosome, a large ribonucleoprotein machine consisting of 5 core RNA and protein complexes (short nuclear ribonucleoprotein complexes, or snRNPs) and a variety of accessory proteins present in stoichiometric, substoichiometric, and superstoichiometric amounts to the core machinery (82, 119, 9). At least 200 individual gene products are thought to generally be involved in the splicing of a single intron (82, 119, 99). Components of the spliceosome recognize a sequence at the 5' splice site at the 5' end of the intron, the 3' splice site and poly-pyrimidine tract at the 3' end of the intron, and the branch point sequence that is usually located near the 3' splice site (9). Interactions between spliceosome snRNPs and other factors generally bound near alternative exons can regulate splicing of an exon in a process known as alternative splicing (105).

1.1.1 Mechanism of mRNA splicing

Splicing consists of two trans-esterification reactions and proceeds through a concerted series of macromolecular ATP-dependent rearrangements (Fig. 1-1) (9, 88, 70, 19, 72). First, U1 snRNP binds to the 5' splice site. The 5' end of U1 snRNA pairs with a stretch of sequence that includes the last few nucleotides within the exon and the first few nucleotides of the intron. There is some degeneracy at most of the bases, but the first two bases of the intron are 5'-GU-3' in more than 99% of introns, which is a perfect reverse complement to that region in the U1 snRNA. A mutation at either position severely disrupts and often ablates splicing, frequently leading to loss of function alleles (21).

Following binding of U1 snRNP, two proteins that associate with U2 snRNP bind at the 3' splice site: U2AF35 recognizes and binds to the 3' splice site located at the end of the intron and the first few nucleotides of the exon while U2AF65 recognizes the poly-pyrimidine tract. U1 and the U2AF complex interact (mediated by the binding of other proteins) either across the exon or across the intron. In humans, where introns

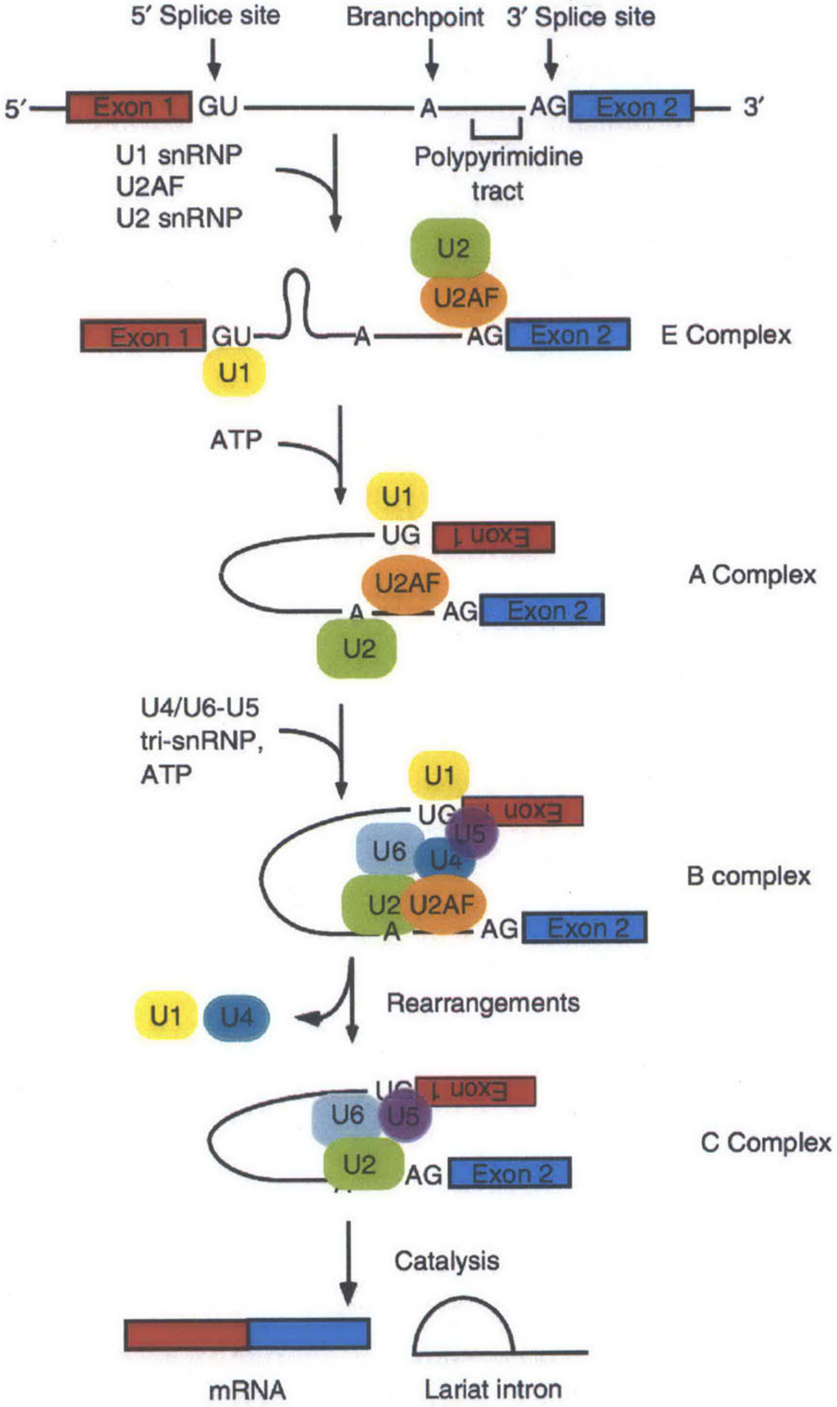


Figure 1-1: Splicing consists of two reactions and is mediated by the spliceosomal snRNPs. Adapted from (78).

are generally much longer than exons, the U1-U2AF interactions usually occur across the body of each exon in a process known as "exon definition," while in organisms with small introns, such as many yeasts, the U1-U2AF interactions occur across the intron in a process known as "intron definition"(7). Organisms such as fly with a wider variety of intron lengths are thought to use both exon definition and intron definition (92).

Once U1 and U2AF bridging interactions occur, U2 binds to the branch point. The branch point in spliceosomal introns is thought to almost always be an "A." In humans and mammals, the branch point has a very weak, degenerate consensus sequence while in yeast such as Saccharomyces cerevisiae, the consensus is much stronger, with as much information content in the branch point as in the 5' or 3' splice sites (66). U5/U4-U6 tri-snRNP then binds, with U5 and U6 binding at the 5' splice site, displacing U1, and U6 interacting with U2. U6 replaces U5 at the 5' splice site, and U6/U2 catalyzes the first splicing reaction, a nucleophilic attack on the 5' splice site by the 2'OH of the branch point A. This frees the 3'OH of the 5' splice site and produces a 2'-5' linkage between the 5' end of the intron and the branch point nucleotide. The closed loop with the unique 2'-5' linkage is known as a lariat. The 3'OH of the 5' splice site then attacks the 3' splice site in the second reaction of splicing, effecting the joining of the two exons together and the release of the lariat RNA species.

1.1.2 Alternative splicing

Transcripts derived from a single locus can be joined in various combinations to produce multiple unique transcripts in a process known as alternative splicing (101, 77). Entire exons can be included or skipped (cassette or skipped exons), or a choice can be made between multiple splice sites that are generally relatively close in location (alternative 3' or 5' splice sites). Other less common classes of alternative splicing include mutually exclusive exons (where exactly one of a set of exons is included in each mRNA), exon cassettes (distinct from cassette exons, adjacent exons that are included or excluded as a group or set), and retained introns (where an entire intron is included in the final transcript). The splicing level, generally measured as the proportion of gene's transcripts that contain that splice isoform, is often apparently regulated and can vary by the tissue or cell line of origin as well as in response to changing conditions or extracellular signaling. Alternative splicing is generally regulated by the binding of *trans*-factors to *cis*-elements located within the exon and the surrounding intron (105), and the final decision to include or exclude an exon can be thought as the sum of the positive and negative interactions between splicing regulators and the spliceosomal snRNPs.

1.1.3 Timing of mRNA splicing

mRNA splicing often occurs co-transcriptionally (79, 51, 94, 54). After RNA polymerase II transcribes an exon or even part of an exon, the splicing machinery begins recognizing the core splicing elements described previously, as well as other elements located within and around the exon. Cross-exon or cross-intron interactions occur and lead to the completion of splicing for most exons often before the completion of gene transcription. Alternative exons, which tend to have suboptimal splice sites, are more frequently spliced later or post-transcriptionally than constitutive exons (79).

The transcription rate has been hypothesized to alter exon recognition and splicing regulation (26, 41). This hypothesis is supported by observations wherein mutations that slow polymerase elongation are associated with higher levels of exon inclusion. If the decision to include or exclude an exon is considered an extreme case of alternative splice site selection between the 3' splice sites of the cassette exon and the subsequent exon, then the slower elongation rate would yield a longer interval during which only the 3' splice site of the cassette exon has been transcribed. During this period where the cassette exon's 3' splice site is the only 3' splice site available, splicing factors located in the nucleoplasm or associated with RNA polymerase can bind to the pre-mRNA, begin the process of exon recognition, and possibly commit to splicing the cassette exon to the upstream exon before the downstream exon is even transcribed.

1.1.4 Role of *trans*-factors in splicing regulation

Splicing is frequently regulated in *trans*, where a factor, generally a protein binds to an RNA motif and promotes or inhibits splicing (Fig. 1-2). A single exon can be regulated by the presence/absence of binding of many factors. The complexity is further increased by the fact that the location of binding can determine the effect a protein has on splicing, for instance binding in the exon may promote splicing while binding nearby in the intron inhibits it. Some proteins, such as those in the SR family (114, 32) and the hnRNP class (90, 5, 39, 40) families, are often broadly expressed, while others are more tissue-specific in their expression (and thus regulation), such as the RBFox (96, 33, 34), Nova (14, 43), and Muscleblind (80, 100) families of proteins. Splicing regulators will generally have multiple domains, some that govern RNA binding, such as RNA recognition motif (RRM) or hnRNPK homology (KH) domains, and domains that mediate interactions with other proteins, splicing factors, or the spliceosome. The RNA binding activity is separable from the effect upon splicing, such that the domains can be swapped or artificially tethered to pre-mRNA with predictable effects upon the chimeric protein's activity (36). Furthermore, by using an RNA binding domain whose target sequence can be modified in conjunction with one or more regulatory domains, splicing factors targeting new or custom motifs can be produced (102).

SR proteins comprise a class of splicing factors that generally bind ESE motifs and promote exon inclusion (114, 32, 105). They contain one or more RNA recognition motif (RRM) domains generally located near the N-terminus and an RS domain, consisting of many RS amino acid repeats near the C-terminus. The RS domain is typically heavily phosphorylated by SR protein kinases (SRPKs) (35) and other ki-

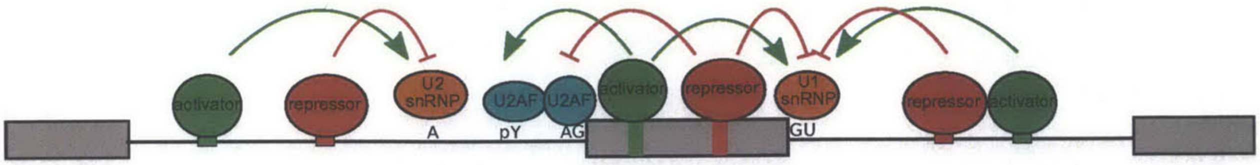


Figure 1-2:
Splicing regulation by binding of *trans*-factors to *cis*-elements.
Splicing factors recognize short sequence motifs located in the target exon and surrounding introns to regulate the exon's splicing. Enhancing elements located in the exon (exonic splicing enhancers, or ESEs) or intron (intronic splicing enhancers, or ISEs) are often bound by SR proteins, a family of splicing factors that tend to promote exon inclusion. Silencing elements located in the exon (exonic splicing silencers, or ESSs) or intron (intronic splicing silencers, or ISSs) are often bound by hnRNPs, a large class of generally repressive splicing factors. After binding, the splicing factors can promote or inhibit the binding of snRNPs, or block progression through the splicing reactions.

nases, a modification often required for their regulatory activity (36, 110, 56, 118). They are thought to often be a general splicing factor and have been found to affect the splicing of both alternative and constitutive exons (67). Recent work has found many diverse roles for SR proteins, including genome stability through creation of RNA:DNA hybrid structures (63) and promoter-proximal pause release through interactions with the 7SK RNA complex (44). At least one SR protein has been found to be a proto-oncogene and have transformative potential (49).

The hnRNPs family comprise another common class of splicing factors. They generally act to suppress splicing upon binding (90, 39, 40). One such protein, the polypyrimidine tract binding protein (PTB), can bind to some polypyrimidine tract and interfere with U2AF binding, thus acting to repress splicing (86).

1.1.5 Role of *cis*-elements in splicing regulation

Splicing regulatory motifs can have different effects depending upon the factors that targets them (105). Motifs are generally classified into 4 categories based on their location and effect upon splicing: intronic splicing enhancers (ISEs) (103), intronic splicing silencers (ISSs) (104), exonic splicing enhancers (ESEs) (30), and exonic splicing silencers (ESSs) (Fig. 1-2) (106). These sets of motifs are frequently related: a motif that promotes exon inclusion when located in an exon often promotes skipping of the same exon when located nearby in an adjacent intron and vice versa (ISEs tend to overlap with ESSs, ISSs tend to overlap with ESEs, etc)(104).

Global approaches to identify the relevant sequences that govern splicing regulation have fallen into a few general areas. Some groups have used libraries of splicing reporter constructs, where each member of the library is largely the same and varies in a small, well-defined region, generally in an alternative exon or surrounding intron (106, 103, 104). These libraries are screened for constructs that have higher or lower levels of splicing. This approach has been successful in identifying sets of k-mers, where k ranges anywhere from 5 to 10, that can promote or inhibit splicing when located within or near alternative exons, but it generally cannot be used to predict differences in inclusion across conditions. Other groups have attempted to learn a "splicing code" by training a machine learning algorithm on splicing levels across different conditions and allowing it to identify the relevant and informative features (4, 111). This approach has recently been successful at predicting relative differences in splicing levels between broadly defined conditions (such as "muscle" grouping together "skeletal muscle" and "cardiac muscle"), but these algorithms are only able to make predictions in a conditions upon which they have been trained. A third approach is to analyze variation in splicing levels across individuals and try to associate particular genetic variants with higher or lower levels of splicing (splicing quantitative trait loci, or sQTLs)(81, 59). This approach has identified thousands of genetic variants correlated with splicing levels, but proving causality is often difficult. Targeted *in-vivo* approaches, such as CLIP-Seq/HiTS-CLIP (64) and its variants (37, 107, 55) or *in-vitro* approaches such as HT-SELEX (46) or Bind-N-Seq (57), give information on the *in vivo* direct targets of splicing factors, but alone cannot be used to infer direction of regulation (109, 112, 115). No one approach has been shown to be optimal, and thus this is an area of active, ongoing work.

1.1.6 Effects of alternative splicing

Regulated alternative splicing has been implicated in many biological processes, including mRNA or protein localization (25, 27, 83, 100), cellular homeostasis (60), differentiation (38), apoptosis (10), and immunity (69). It can impact these processes a number of different ways. One is by altering the localization of the resulting protein (Fig. 1-3). It is vital that the antibody-producing b-cells must be activated by the same antigen that its antibodies recognize or else it will produce antibodies targeting a different antigen, possibly leading to an auto-immune response and failing to target the intended antigen. This precise correspondence between the target recognized by the cell and that bound by the antibodies produced is mediated by alternative splicing (25, 27, 83). The cell surface receptor that leads to the b-cell's activation and the antibody that the b-cell would ultimately produce are coded for by the same sets of genes. The transcripts coding for the receptor contain a different 3' end (and thus C-terminus at the protein level) than that of the antibody-coding transcripts. The difference between the transcripts is mediated by alternative mRNA processing and leads to the receptor protein being membrane-bound while the protein lacking the C-terminal trans-membrane domain is secreted from the cell.

Alternative splicing is also a highly conserved mechanism for maintaining homeostatic levels of splicing factors (Fig. 1-4) (60). Genes coding for proteins in the SR family of splicing factors frequently contain exons that encode a premature termination codon (PTC). These exons, termed poison cassette exons, can then be utilized in the regulation of the protein's expression through the induction of nonsense mediated mRNA decay (NMD) (3), a quality control and regulatory process that degrades transcripts containing a PTC or overly long 3' UTR. Thus, a splicing factor can autoregulate its own expression through regulation of these poison cassette exons, leading to an equilibrium between the protein level of the splicing factor and the inclusion level of the poison cassette exon. If the protein level becomes high enough such that it promotes the inclusion of a poison exon, then the transcripts produced will not lead to significant protein accumulation until overall splicing factor concentrations drop

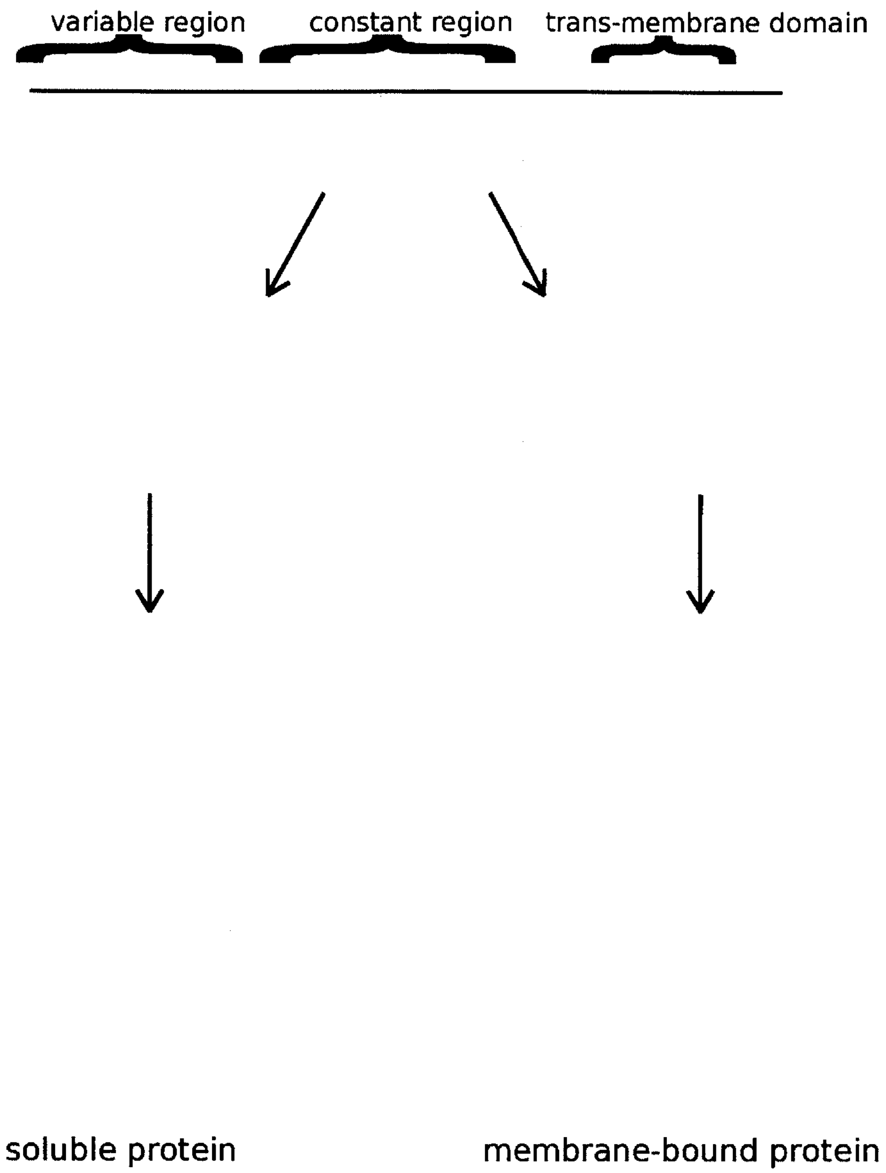


Figure 1-3: Alternative mRNA processing of IgM transcripts leads to protein isoforms with the same binding affinities but different localizations.
The pre-mRNA of the IgM gene is processed into two isoforms containing a common set of exons (in blue and grey) but differing in their 3' end (shaded in green). These green exons code for a trans-membrane domain. Signals in the mRNA lead to the localization of the nascent peptide to the golgi, where it is processed and assembled. Upon completion, vesicles containing the protein are exocytosed. If the green exons are included in the protein, then when the vesicle membrane fuses with the cell's plasma membrane, the protein will remain in the membrane and it will act as a receptor specific to the antigen it recognizes. However, if the green exons are not included, then the soluble antibody will be secreted (25, 27, 83).

below a level such that inclusion of the poison exon is reduced.

Another role for alternative splicing that is becoming increasingly well defined is in animal development. Here, for instance, muscleblind proteins, which are generally more tissue-specific splicing factors, are required for proper splicing of hundreds of cassette exons in differentiated cells (100, 38). Disrupting their expression in differentiated cells leads to reversion of exon inclusion of many targets to levels that more closely resemble those observed in embryonic stem cells. Similarly, their overexpression in embryonic stem cells has the opposite effect, leading to patterns of splicing that resemble differentiated tissues. This splicing switch is further driven by another family of splicing factors, the CUG-BP and ETR-3-like factors (CELF) family, which often opposes the effects of muscleblinds and promotes a more embryonic splicing pattern. This antagonistic regulatory relationship between CELF (higher expression in embryonic cells/tissues) and muscleblind (higher in differentiated cells/tissues) proteins effects coordinated changes in splicing during differentiation (48). Furthermore, in line with a reversion to an embryonic splicing phenotype upon their knockdown, disrupting muscleblind proteins leads to enhanced reprogramming of differentiated cells to induced pluripotent stem cells (38).

While it has been possible to interrogate such splicing changes in various biological systems for a decade or more through the use of splicing microarrays and (more recently) RNA-seq, a global and general effect (if one exists) for these splicing changes has remained elusive. A few recent studies employing RNA-seq, perhaps aided by the improved power to detect such changes compared to earlier technologies, have identified a number of global changes in the proteome brought about by alternative splicing. Recently, it was shown that brain-regulated isoforms tend to have different protein-protein interaction profiles relative to other isoforms of the same proteins, suggesting a role for alternative splicing regulation in rewiring protein-protein interactions between tissues (29). Other studies have suggested links between phosphorylation and alternativc splicing. Darnell and colleagues found that exons regulated by Nova, a neuronal splicing factor, are enriched for phosphorylation sites (115), a finding later

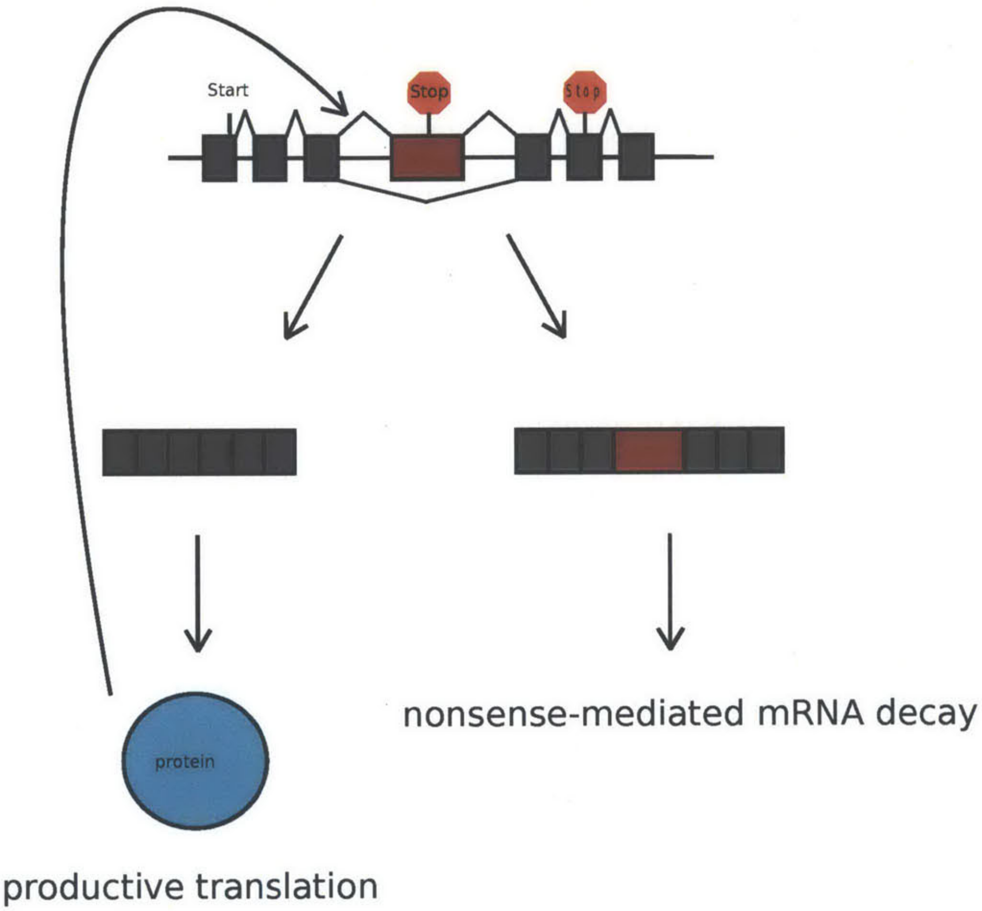


Figure 1-4: Splicing factors can regulate gene expression through regulation of exons that lead to transcript degradation.
The purple exon in the diagram contains a premature stop codon induces nonsense-mediated decay (NMD) when it is included. When the protein levels of the splicing factor are low, the exon is not included, leading to a stable final transcript product that yields protein when translated. When the protein level reaches a sufficiently high level, the splicing factor binds to its own pre-mRNA transcript and promotes the inclusion of the NMD-inducing exon, destabilizing the transcript. This exon is then included in the gene's transcripts until protein levels decrease such that the levels are not sufficient to promote the exon's inclusion, leading to an equilibrium between the NMD-inducing exon and the splicing factor's protein level (60).

generalized to exons differentially spliced between brain and liver (15). In this thesis, I present work showing that this phenomenon extends beyond brain, and that this connection to phosphorylation is a conserved feature of alternatively spliced exons.

1.2 Gene expression evolution

A goal of evolutionary biology is to understand the changes driving the differences between species. Decades ago, it was appreciated that there were relatively few coding sequence changes between closely related species such as human and chimp (53). It was therefore hypothesized that evolution of gene regulation, rather than protein sequence, would play a dominant role in driving evolutionary changes (13). The evolution of gene expression levels has been investigated globally and found to largely be under stabilizing selection, with many other changes evolving neutrally (11). Investigations into the evolution of splicing and alternative splicing, ubiquitous in mammalian gene expression, have been complicated by their more involved nature and hampered by technological limitations (42). Recent technological advancements, however, have made global and less biased investigations into splicing evolution possible (108).

1.2.1 Evolution of overall levels of gene expression

The first global studies examining the evolution of gene expression employed DNA microarrays to compare gene expression levels between species (17, 117, 16, 91). These studies sought to differentiate between different models for gene expression evolution (Fig. 1-5) and have yielded a number of important insights into the evolution of gene expression levels. They generally look for deviations from a null model and may require the specification of an alternative model and parameters. Neutral evolution is often used as this null model and predicts that most observed changes are effectively neutral and caused by random genetic drift (52). The observation of less variation between species than expected is suggestive of stabilizing selection, which acts to maintain particular levels of a trait (here, expression) (18) and may be indicative of negative selective pressure . Similarly, observing a trait with low variation between most species and a large change in a specific lineage or species suggests either positive/directional selection or a lineage-specific relaxation of stabilizing selection (24).

Microarray studies (and later RNA-seq work) comparing the expression profiles in different species and tissues have found that most genes evolve under purifying selection (17, 117, 16, 91, 11). When comparing the divergence rates of different tissues, defined as a measure of how rapidly expression patterns change within that tissue, expression in neural tissues (such as the brain and brain stem) has generally been found to have the slowest rate of change (17, 11). One interpretation for this observation is pressure to maintain core gene expression relating to, for instance, synaptic function, in these tissues. The low rate of expression divergence in neural tissues stands in stark contrast with a reproductive tissue like testes, which has been found to have the highest rate of divergence (11).

1.2.2 Evolution of alternative splicing and isoform levels

The evolution of splicing can be queried at different levels (42), but in general has been more difficult to investigate than gene expression. At the most basic level is the binary question—is an exon present in a genome? This can be made more precise by querying the specific splice sites. Beyond the position of the exon, the splicing status (e.g. relating to the skipping of the exon) can be interrogated. Finally, the most involved (and most difficult) questions to ask pertain to the actual pattern of splicing across tissues or conditions, similar to investigations into the patterns of gene expression across species described above.

A few recent studies have approached the first question posed above and looked at patterns of presence or absence at the exon level across species; primarily using two

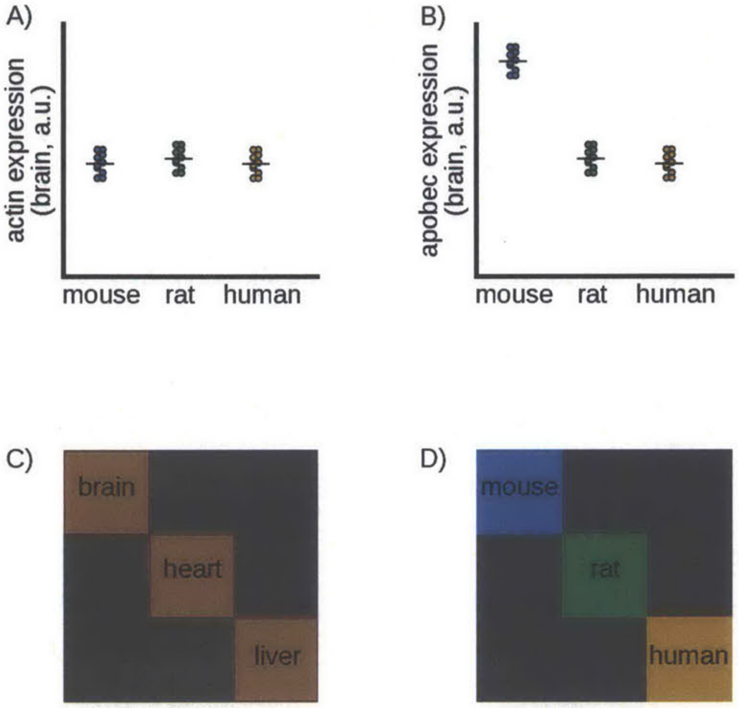


Figure 1-5: Different evolutionary pressures lead to unique predictions regarding gene expression comparisons between species

(a) There is little variance in gene expression between or within species. This pattern is suggestive of stabilizing selection, where there would be selection against changes in the trait (here: gene expression) across evolution.

(b) There is little variance in gene expression between or within most species while the levels are very different in one species. This pattern is suggestive of stabilizing selection in the majority of species and possibly positive selection in the species or lineage with different expression levels.

(c) When analyzed globally via a clustering analysis, one way stabilizing selection will present itself is in samples clustering by the tissue of origin.

(d) When analyzed globally via a similar clustering analysis, frequent positive selection may present itself as samples clustering by the species of origin. Note that this is not the only interpretation, and further analyses are necessary to conclude that positive selection has acted upon the expression of the genes considered.

approaches. One approach is to take the set of exons annotated in a query species (for instance, human), and look at the the aligned regions in a second species. An exon can be inferred to be spliced in that species if the "GT" and "AG" are conserved in that species. A caveat to this approach is that changes in exon boundaries between species will lead to exons being detected as lost or missing in some species, leading to patterns that are difficult to interpret, or biased towards younger ages. Nevertheless, careful application of this approach by Lee and colleagues found that novel exons are typically less included in EST databases (2). Further, they found that exon age tended to correlate with overall levels of inclusion in EST databases. In Chapter 2 of this thesis, I take this approach in a new direction by focusing not on cases where the splice sites have been only maintained or lost, but instead studying cases where the exon boundaries have moved. In investigating the evolution of alternative tandem 3' splice sites, I found that 3' splice sites are "hotspots" for changes in coding sequence between species.

An alternative to this approach is to use libraries of expressed sequence tags (ESTs), consisting of shotgun sequencing of cloned cDNA fragments, or full length cDNA sequences to compare exons between species (1). This approach would be less sensitive to splice site turnover, but being undetected in EST libraries does not mean that the exon does not exist in that species, since EST libraries are often low coverage and likely miss lowly included (and even some moderately or highly included) exons. This approach was used to compare exons and their splicing between mouse and human. Blencowe and colleagues found that only 10% of exons were alternatively spliced in both mouse and human, while nearly half of the alternative exons considered are unique to one species (76). Earlier work from Modrek and Lee found that exons lowly included in EST databases are frequently specific to mouse, rat, or human (71). Further, Zhang and Chasin, employing an EST approach rather than the genomic approach used by Lee, found that species-specific exons identified by comparing ESTs between species are lowly included in EST databases and frequently repeat-associated (116). These results suggest the frequent occurrence of species-specific exons at low inclusion levels and a general low level of splicing conservation, though this observation can be complicated by tissue-specific splicing impacting detection. In Chapter 4 of this thesis, I consider causes and effects of exon creation utilizing whole genome alignments in conjunction with RNA-seq data.

While gene expression estimates from microarrays have the potential to be robust to changes in splice site usage provided that care is taken in probe sequence choice, estimates of alternative splicing could be very biased by changes in the specific exon boundaries probed in microarrays. Splicing microarray analyses are further complicated by smaller informative regions than for analysis of gene expression, cross-hybridization between related regions, and other issues common to microarrays. They can nevertheless be used for inferring splicing levels of exons, even using the ability to cross-hybridize related sequences to apply a microarray from one species to a closely related species. This approach was employed to probe the expression and splicing profiles in RNA extracted from heart and brain of human, chimpanzee, and mouse (16). The authors found that the similarity in splicing between species was comparable to that of gene expression, which is generally well conserved and under stabilizing selection. In chapter 3 of this thesis, I use RNA-seq to ask similar questions, which avoids some of the problems inherent in microarrays and (importantly) allows for the discovery of species-specific splicing patterns.

1.3 Relevant technology

The advent of high-throughput sequencing of DNA fragments has led to a revolution in computational approaches to biological problems (75). The cost per nucleotide in recent years has been decreasing at rates that far surpass an exponential rate of improvement (47). In addition to sequencing fragments of genomic DNA, alternate sample preparation protocols can be used to generate data representing many other sources of information (75). One such protocol, ChIP-seq, uses crosslinking followed by immunoprecipitation of a target protein to enrich for and sequence regions of

DNA that are bound to the protein (45). Similarly, RNA can be reverse-transcribed into cDNA and then sequenced (RNA-seq, Fig. 1-6), yielding information about gene expression and alternative splicing (105, 73). Many other protocols have been developed, providing information on transcription kinetics (23), nucleosome positioning (97), chromosomal conformation (65), and many others. I will focus on Illumina sequencing technology here as it is the technology I used, though other approaches exist (28, 85, 87).

1.3.1 High-throughput sequencing of DNA fragments by Illumina sequencing

High-throughput sequencing of DNA fragments is now a widely used technique. The DNA of interest is ligated to adapter sequences that serve two purposes. First, they are used for PCR amplification of the library after ligation. Following amplification, the adapter is then used for generating clusters on a glass flowcell. The flowcell has short oligos covalently linked that are complementary to the adapter sequence. The DNA fragment hybridizes via the adapter, after which DNA polymerase extends the flowcell's oligo to the fragment's end. The other adapter then hybridizes to complementary oligo (again covalently attached to the flowcell) and is similarly elongated. This process, termed "bridge amplification" is repeated many times to generate clusters of identical oligos covalently attached to the flowcell. These clusters are then sequenced using DNA polymerase incorporation of reversibly terminated nucleotides with base-specific fluorophores. Only a single nucleotide can be incorporated at a time because of the unique chemistry of the nucleotides used, so imaging after each fluorophore-conjugated nucleotide is individually excited allows for the deciphering of the oligo's sequence. Sequencing starting from both oligos (individually) can be used to generate a pair of reads for a given starting fragment (6).

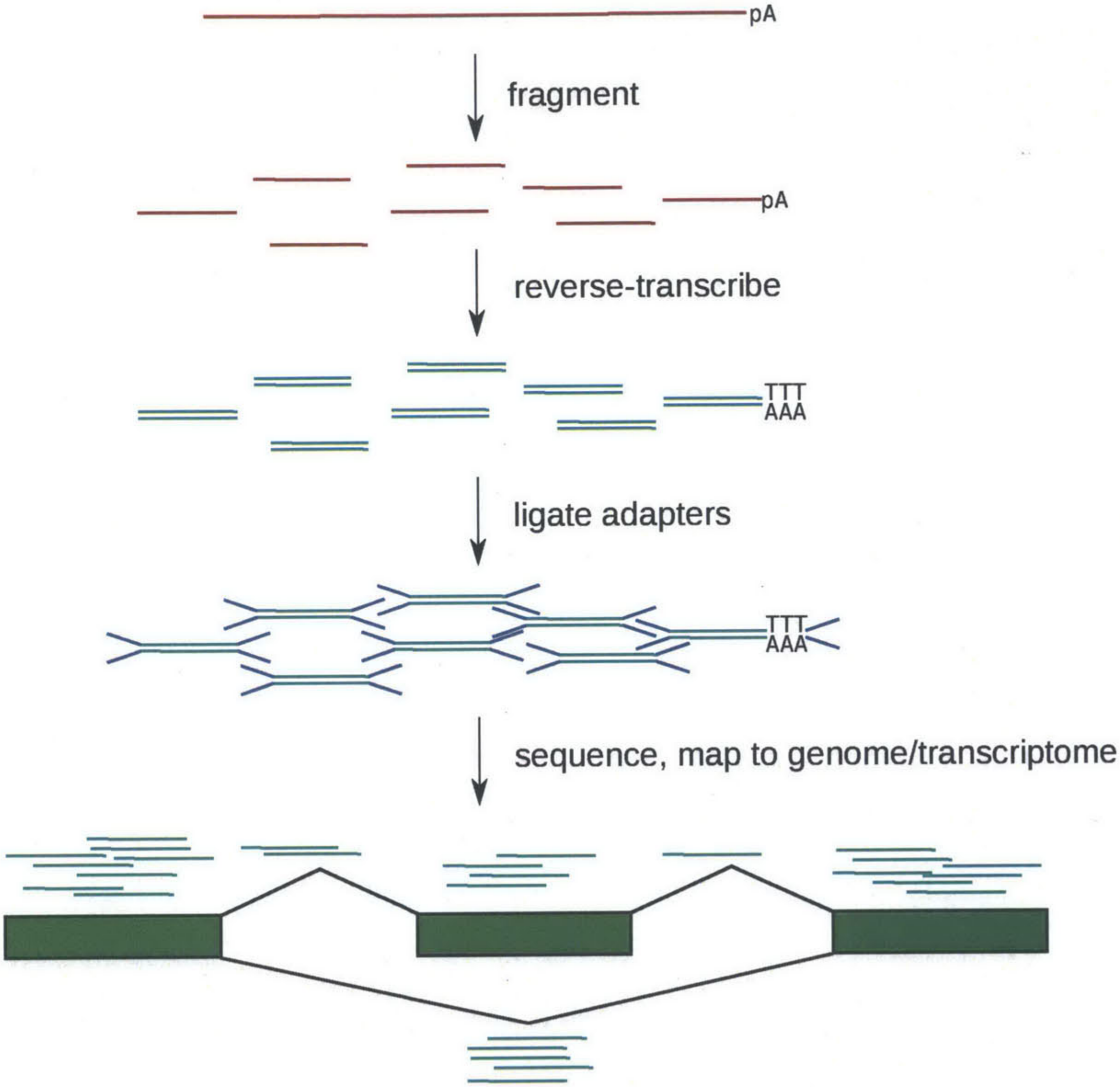


Figure 1-6: Application of high-throughput sequencing to probe mRNA transcript levels.

Total RNA is poly-A selected to isolate processed mRNA. The RNA is reverse-transcribed using an oligo dT primer or fragmented and reverse-transcribed using random hexamer priming. The cDNA fragments are ligated to adapters and sequenced using a next-generation sequencing technology (such as Illumina, Abi SOLiD, Ion Torrent, or others). The sequences are aligned to the genome or transcriptome and used to infer gene expression levels, splicing levels, or other information.

1.3.2 Applications of high-throughput sequencing to RNA

Reverse-transcribed RNA (cDNA) can be used in place of DNA in the above sequencing protocol. This can be used to measure global gene expression (often measured in FPKM: Fragments per Kilobase of interval Per Million mapped reads (73) or TPM: Transcripts Per Million (61)) by starting with poly-A-selected or ribosome-subtracted RNA in a manner similar to microarrays, though with a number of advantages over the preceding technology. These expression estimates can be scaled with the use of spike-in oligos to control for different efficiencies in library preparation and input amounts and convert relative expression values to absolute amounts (68). It can also be used to interrogate alternative splicing. Perhaps the simplest measure of alternative splicing compares reads that overlap splice junctions that support inclusion of an exon to those that support its exclusion (105). However, a number of more rigorous approaches have been developed, using more sophisticated statistical models to incorporate additional information or control for sequencing biases, leading to improved splicing estimates (95, 50, 62). The presence of splice junction reads can be used to construct a splice graph (similar to what was done with ESTs) to construct transcript models inferred from the data (95, 62) and discover novel isoforms not found in annotations.

Bibliography

[1] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, Jun 1991.

[2] Alexander V. Alekseyenko, Namshin Kim, and Christopher J. Lee. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, 13(5):661–670, May 2007.

[3] Kristian E. Baker and Roy Parker. Nonsense-mediated mrna decay: terminating erroneous gene expression. *Curr Opin Cell Biol*, 16(3):293–299, Jun 2004.

[4] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[5] Nuno L. Barbosa-Morais, Maria Carmo-Fonseca, and Samuel Aparcio. Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. *Genome Res*, 16(1):66–77, Jan 2006.

[6] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser,

Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung-Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W. Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008.

[7] S. M. Berget. Exon recognition in vertebrate splicing. *J Biol Chem*, 270(6):2411–2414, Feb 1995.

[8] A. J. Berk and P. A. Sharp. Sizing and mapping of early adenovirus mrnas by gel electrophoresis of s1 endonuclease-digested hybrids. *Cell*, 12(3):721–732, Nov 1977.

[9] Douglas L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336, 2003.

[10] L. H. Boise, M. Gonzlez-Garca, C. E. Postema, L. Ding, T. Lindsten, L. A. Turka, X. Mao, G. Nuez, and C. B. Thompson. bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. *Cell*, 74(4):597–608, Aug 1993.

[11] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gbor Csrdi, Patrick Harrigan, Manuela Weier, Anglica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grtzner, Sven Bergmann, Rasmus Nielsen, Svante Pbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348, Oct 2011.

[12] R. Breathnach, J. L. Mandel, and P. Chambon. Ovalbumin gene is split in chicken dna. *Nature*, 270(5635):314–319, Nov 1977.

[13] R. J. Britten and E. H. Davidson. Gene regulation for higher cells: a theory. *Science*, 165(3891):349–357, Jul 1969.

[14] R. J. Buckanovich, J. B. Posner, and R. B. Darnell. Nova, the paraneoplastic ri antigen, is homologous to an rna-binding protein and is specifically expressed in the developing motor system. *Neuron*, 11(4):657–672, Oct 1993.

[15] Marija Buljan, Guilhem Chalancon, Sebastian Eustermann, Gunter P. Wagner, Monika Fuxreiter, Alex Bateman, and M Madan Babu. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*, 46(6):871–883, Jun 2012.

[16] John A. Calarco, Yi Xing, Mario Cceres, Joseph P. Calarco, Xinshu Xiao, Qun Pan, Christopher Lee, Todd M. Preuss, and Benjamin J. Blencowe. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev*, 21(22):2963–2975, Nov 2007.

[17] Esther T. Chan, Gerald T. Quon, Gordon Chua, Tomas Babak, Miles Trochesset, Ralph A. Zirngibl, Jane Aubin, Michael J H. Ratcliffe, Andrew Wilde, Michael Brudno, Quaid D. Morris, and Timothy R. Hughes. Conservation of core gene expression in vertebrate tissues. *J Biol*, 8(3):33, 2009.

[18] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, Aug 1993.

[19] Mo Chen and James L. Manley. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, 10(11):741–754, Nov 2009.

[20] L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8, Sep 1977.

[21] 1000 Genomes Project Consortium, Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, Nov 2012.

[22] Thomas A. Cooper, Lili Wan, and Gideon Dreyfuss. Rna and disease. *Cell*, 136(4):777–793, Feb 2009.

[23] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, Dec 2008.

[24] Charles Darwin. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life.* John Murray, London, 1859.

[25] M. M. Davis, K. Calame, P. W. Early, D. L. Livant, R. Joho, I. L. Weissman, and L. Hood. An immunoglobulin heavy-chain gene is formed by at least two recombinational events. *Nature*, 283(5749):733–739, Feb 1980.

[26] Manuel de la Mata, Claudio R. Alonso, Sebastin Kadener, Juan P. Fededa, Matas Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R. Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Mol Cell*, 12(2):525–532, Aug 2003.

[27] P. Early, J. Rogers, M. Davis, K. Calame, M. Bond, R. Wall, and L. Hood. Two mrnas can be produced from a single immunoglobulin mu gene by alternative rna processing pathways. *Cell*, 20(2):313–319, Jun 1980.

[28] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[29] Jonathan D. Ellis, Miriam Barrios-Rodiles, Recep Colak, Manuel Irimia, Taehyung Kim, John A. Calarco, Xinchen Wang, Qun Pan, Dave O'Hanlon, Philip M. Kim, Jeffrey L. Wrana, and Benjamin J. Blencowe. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*, 46(6):884–892, Jun 2012.

[30] William G. Fairbrother, Ru-Fang Yeh, Phillip A. Sharp, and Christopher B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, Aug 2002.

[31] Nuno Andr Faustino and Thomas A. Cooper. Pre-mrna splicing and human disease. *Genes Dev*, 17(4):419–437, Feb 2003.

[32] X. D. Fu. The superfamily of arginine/serine-rich splicing factors. *RNA*, 1(7):663–680, Sep 1995.

[33] Thomas L. Gallagher, Joshua A. Arribere, Paul A. Geurts, Cameron R T. Exner, Kent L. McDonald, Kariena K. Dill, Henry L. Marr, Shaunak S. Adkar, Aaron T. Garnett, Sharon L. Amacher, and John G. Conboy. Rbfox-regulated alternative splicing is critical for zebrafish cardiac and skeletal muscle functions. *Dev Biol*, 359(2):251–261, Nov 2011.

[34] Lauren T. Gehman, Peter Stoilov, Jamie Maguire, Andrey Damianov, Chia-Ho Lin, Lily Shiue, Manuel Ares, Jr, Istvan Mody, and Douglas L. Black. The splicing regulator rbfox1 (a2bp1) controls neuronal excitation in the mammalian brain. *Nat Genet*, 43(7):706–711, Jul 2011.

[35] Thomas Giannakouros, Eleni Nikolakaki, Ilias Mylonis, and Eleni Georgatsou. Serine-arginine protein kinases: a small protein kinase family with a large cellular presence. *FEBS J*, 278(4):570–586, Feb 2011.

[36] B. R. Graveley and T. Maniatis. Arginine/serine-rich domains of sr proteins can function as activators of pre-mrna splicing. *Mol Cell*, 1(5):765–771, Apr 1998.

[37] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Jr, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–141, Apr 2010.

[38] Hong Han, Manuel Irimia, P Joel Ross, Hoon-Ki Sung, Babak Alipanahi, Laurent David, Azadeh Golipour, Mathieu Gabut, Iacovos P. Michael, Emil N. Nachman, Eric Wang, Dan Trcka, Tadeo Thompson, Dave O'Hanlon, Valentina Slobodeniuc, Nuno L. Barbosa-Morais, Christopher B. Burge, Jason Moffat, Brendan J. Frey, Andras Nagy, James Ellis, Jeffrey L. Wrana, and Benjamin J. Blencowe. Mbnl proteins repress es-cell-specific alternative splicing and reprogramming. *Nature*, 498(7453):241–245, Jun 2013.

[39] Siew Ping Han, Yue Hang Tang, and Ross Smith. Functional diversity of the hnrnps: past, present and perspectives. *Biochem J*, 430(3):379–392, Sep 2010.

[40] Stephanie C. Huelga, Anthony Q. Vu, Justin D. Arnold, Tiffany Y. Liang, Patrick P. Liu, Bernice Y. Yan, John Paul Donohue, Lily Shiue, Shawn Hoon, Sydney Brenner, Manuel Ares, Jr, and Gene W. Yeo. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnrnp proteins. *Cell Rep*, 1(2):167–178, Feb 2012.

[41] Joanna Y. Ip, Dominic Schmidt, Qun Pan, Arun K. Ramani, Andrew G. Fraser, Duncan T. Odom, and Benjamin J. Blencowe. Global impact of rna polymerase ii elongation inhibition on alternative splicing regulation. *Genome Res*, 21(3):390–401, Mar 2011.

[42] Manuel Irimia, Jakob L. Rukov, Scott W. Roy, Jeppe Vinther, and Jordi Garcia-Fernandez. Quantitative regulation of alternative splicing in evolution and development. *Bioessays*, 31(1):40–50, Jan 2009.

[43] Nejc Jelen, Jernej Ule, Marko Zivin, and Robert B. Darnell. Evolution of nova-dependent splicing regulation in the brain. *PLoS Genet*, 3(10):1838–1847, Oct 2007.

[44] Xiong Ji, Yu Zhou, Shatakshi Pandit, Jie Huang, Hairi Li, Charles Y. Lin, Rui Xiao, Christopher B. Burge, and Xiang-Dong Fu. Sr proteins collaborate with 7sk and promoter-associated nascent rna to release paused polymerase. *Cell*, 153(4):855–868, May 2013.

[45] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, Jun 2007.

[46] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. Sillanp, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R. Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome Res*, 20(6):861–873, Jun 2010.

[47] Wetterstrand KA. Dna sequencing costs: Data from the nhgri genome sequencing program (gsp) available at: www.genome.gov/sequencingcosts. accessed 3/1/2014.

[48] Auinash Kalsotra, Xinshu Xiao, Amanda J. Ward, John C. Castle, Jason M. Johnson, Christopher B. Burge, and Thomas A. Cooper. A postnatal switch of celf and mbnl proteins reprograms alternative splicing in the developing heart. *Proc Natl Acad Sci U S A*, 105(51):20333–20338, Dec 2008.

[49] Rotem Karni, Elisa de Stanchina, Scott W. Lowe, Rahul Sinha, David Mu, and Adrian R. Krainer. The gene encoding the splicing factor sf2/asf is a proto-oncogene. *Nat Struct Mol Biol*, 14(3):185–193, Mar 2007.

[50] Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12):1009–1015, Dec 2010.

[51] Yevgenia L. Khodor, Joseph Rodriguez, Katharine C. Abruzzi, Chih-Hang Anthony Tang, Michael T Marr, 2nd, and Michael Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mrna splicing in drosophila. *Genes Dev*, 25(23):2502–2512, Dec 2011.

[52] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, Feb 1968.

[53] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, Apr 1975.

[54] Alberto R. Kornblihtt, Ignacio E. Schor, Mariano All, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J. Muoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol*, 14(3):153–165, Mar 2013.

[55] Julian Knig, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. iclip reveals the function of hnrnp particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–915, Jul 2010.

[56] M. C. Lai, R. I. Lin, and W. Y. Tarn. Transportin-sr2 mediates nuclear import of phosphorylated sr proteins. *Proc Natl Acad Sci U S A*, 98(18):10154–10159, Aug 2001.

[57] Robertson A. D. Jangi M. McGeary S. Sharp P. A. Burge C. B. Lambert, N. J. Rna bind-n-seq: quantitative assessment of the sequence and structural binding specificity of rna binding proteins. *Molecular Cell*, 2014.

[58] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium . Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[59] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedlnder, Peter A C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle

Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, Geuvadis Consortium , Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Hsler, Ann-Christine Syvnen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guig, Ivo G. Gut, Xavier Estivill, Emmanouil T. Dermitzakis, and Geuvadis Consortium . Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[60] Liana F. Lareau, Maki Inada, Richard E. Green, Jordan C. Wengrod, and Steven E. Brenner. Unproductive splicing of sr genes associated with highly conserved and ultraconserved dna elements. *Nature*, 446(7138):926–929, Apr 2007.

[61] Bo Li and Colin N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, 2011.

[62] Heng Li. Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, 27(5):718–719, Mar 2011.

[63] Xialu Li and James L. Manley. Inactivation of the sr protein splicing factor asf/sf2 results in genomic instability. *Cell*, 122(3):365–378, Aug 2005.

[64] Donny D. Licatalosi, Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, Anthony C. Schweitzer, John E. Blume, Xuning Wang, Jennifer C. Darnell, and Robert B. Darnell. Hits-clip yields genome-wide insights into brain alternative rna processing. *Nature*, 456(7221):464–469, Nov 2008.

[65] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, Oct 2009.

[66] L. P. Lim and C. B. Burge. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A*, 98(20):11193–11198, Sep 2001.

[67] Shengrong Lin and Xiang-Dong Fu. Sr proteins and related factors in alternative splicing. *Adv Exp Med Biol*, 623:107–122, 2007.

[68] Jakob Lovn, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee, and Richard A. Young. Revisiting global gene expression analysis. *Cell*, 151(3):476–482, Oct 2012.

[69] R. Maki, W. Roeder, A. Traunecker, C. Sidman, M. Wabl, W. Raschke, and S. Tonegawa. The role of dna rearrangement and alternative rna processing in the expression of immunoglobulin delta genes. *Cell*, 24(2):353–365, May 1981.

[70] Arianne J. Matlin, Francis Clark, and Christopher W J. Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–398, May 2005.

[71] Barmak Modrek and Christopher J. Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, 34(2):177–180, Jun 2003.

[72] Melissa J. Moore and Nick J. Proudfoot. Pre-mrna processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, Feb 2009.

[73] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628, Jul 2008.

[74] Julie E J. Nixon, Amy Wang, Hilary G. Morrison, Andrew G. McArthur, Mitchell L. Sogin, Brendan J. Loftus, and John Samuelson. A spliceosomal intron in giardia lamblia. *Proc Natl Acad Sci U S A*, 99(6):3701–3705, Mar 2002.

[75] Lior Pachter. *seq. http://liorpachter.wordpress.com/seq/. Regularly updated list of sequencing protocols.

[76] Qun Pan, Malina A. Bakowski, Quaid Morris, Wen Zhang, Brendan J. Frey, Timothy R. Hughes, and Benjamin J. Blencowe. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet*, 21(2):73–77, Feb 2005.

[77] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, Dec 2008.

[78] Amy Pandya-Jones. Pre-mrna splicing during transcription in the mammalian system. *Wiley Interdiscip Rev RNA*, 2(5):700–717, 2011.

[79] Amy Pandya-Jones and Douglas L. Black. Co-transcriptional splicing of constitutive and alternative exons. *RNA*, 15(10):1896–1908, Oct 2009.

[80] Maya Pascual, Marta Vicente, Lidon Monferrer, and Ruben Artero. The muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing. *Differentiation*, 74(2-3):65–80, Mar 2006.

[81] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, Apr 2010.

[82] Juri Rappsilber, Ursula Ryder, Angus I. Lamond, and Matthias Mann. Large-scale proteomic analysis of the human spliceosome. *Genome Res*, 12(8):1231–1245, Aug 2002.

[83] J. Rogers, P. Early, C. Carter, K. Calame, M. Bond, L. Hood, and R. Wall. Two mrnas with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell*, 20(2):303–312, Jun 1980.

[84] Igor B. Rogozin, Yuri I. Wolf, Alexander V. Sorokin, Boris G. Mirkin, and Eugene V. Koonin. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol*, 13(17):1512–1517, Sep 2003.

[85] M. Ronaghi, M. Uhln, and P. Nyrn. A sequencing method based on real-time pyrophosphate. *Science*, 281(5375):363, 365, Jul 1998.

[86] Shalini Sharma, Arnold M. Falick, and Douglas L. Black. Polypyrimidine tract binding protein blocks the 5' splice site-dependent assembly of u2af and the prespliceosomal e complex. *Mol Cell*, 19(4):485–496, Aug 2005.

[87] Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, Sep 2005.

[88] Chanseok Shin and James L. Manley. Cell signalling and the control of pre-mrna splicing. *Nat Rev Mol Cell Biol*, 5(9):727–738, Sep 2004.

[89] Alastair G B. Simpson, Erin K. MacQuarrie, and Andrew J. Roger. Eukaryotic evolution: early origin of canonical introns. *Nature*, 419(6904):270, Sep 2002.

[90] O. P. Singh. Functional diversity of hnrnp proteins. *Indian J Biochem Biophys*, 38(3):129–134, Jun 2001.

[91] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct 2003.

[92] M. Talerico and S. M. Berget. Intron definition in splicing of small drosophila introns. *Mol Cell Biol*, 14(5):3434–3445, May 1994.

[93] S. M. Tilghman, D. C. Tiemeier, J. G. Seidman, B. M. Peterlin, M. Sullivan, J. V. Maizel, and P. Leder. Intervening sequence of dna identified in the structural portion of a mouse beta-globin gene. *Proc Natl Acad Sci U S A*, 75(2):725–729, Feb 1978.

[94] Hagen Tilgner, David G. Knowles, Rory Johnson, Carrie A. Davis, Sudipto Chakrabortty, Sarah Djebali, Joo Curado, Michael Snyder, Thomas R. Gingeras, and Roderic Guig. Deep sequencing of subcellular rna fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. *Genome Res*, 22(9):1616–1625, Sep 2012.

[95] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[96] Jason G. Underwood, Paul L. Boutz, Joseph D. Dougherty, Peter Stoilov, and Douglas L. Black. Homologues of the caenorhabditis elegans fox-1 protein are neuronal splicing regulators in mammals. *Mol Cell Biol*, 25(22):10005–10016, Nov 2005.

[97] Anton Valouev, Steven M. Johnson, Scott D. Boyd, Cheryl L. Smith, Andrew Z. Fire, and Arend Sidow. Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520, Jun 2011.

[98] Stepnka Vancov, Weihong Yan, Jane M. Carlton, and Patricia J. Johnson. Spliceosomal introns in the deep-branching eukaryote trichomonas vaginalis. *Proc Natl Acad Sci U S A*, 102(12):4430–4435, Mar 2005.

[99] Markus C. Wahl, Cindy L. Will, and Reinhard Lhrmann. The spliceosome: design principles of a dynamic rnp machine. *Cell*, 136(4):701–718, Feb 2009.

[100] Eric T. Wang, Neal A L. Cody, Sonali Jog, Michela Biancolella, Thomas T. Wang, Daniel J. Treacy, Shujun Luo, Gary P. Schroth, David E. Housman, Sita Reddy, Eric Lcuyer, and Christopher B. Burge. Transcriptome-wide regulation of pre-mrna splicing and mrna localization by muscleblind proteins. *Cell*, 150(4):710–724, Aug 2012.

[101] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[102] Yang Wang, Cheom-Gil Cheong, Traci M Tanaka Hall, and Zefeng Wang. Engineering splicing factors with designed specificities. *Nat Methods*, 6(11):825–830, Nov 2009.

[103] Yang Wang, Meng Ma, Xinshu Xiao, and Zefeng Wang. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol*, 19(10):1044–1052, Oct 2012.

[104] Yang Wang, Xinshu Xiao, Jianming Zhang, Rajarshi Choudhury, Alex Robertson, Kai Li, Meng Ma, Christopher B. Burge, and Zefeng Wang. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol*, 20(1):36–45, Jan 2013.

[105] Zefeng Wang and Christopher B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008.

[106] Zefeng Wang, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–845, Dec 2004.

[107] Zhen Wang, Melis Kayikci, Michael Briese, Kathi Zarnack, Nicholas M. Luscombe, Gregor Rot, Bla Zupan, Toma Curk, and Jernej Ule. iclip predicts the dual splicing effects of tia-rna interactions. *PLoS Biol*, 8(10):e1000530, 2010.

[108] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[109] Joshua T. Witten and Jernej Ule. Understanding splicing regulation through rna splicing maps. *Trends Genet*, 27(3):89–97, Mar 2011.

[110] S. H. Xiao and J. L. Manley. Phosphorylation of the asf/sf2 rs domain affects both protein-protein and protein-rna interactions and is necessary for splicing. *Genes Dev*, 11(3):334–344, Feb 1997.

[111] Hui Yuan Xiong, Yoseph Barash, and Brendan J. Frey. Bayesian prediction of tissue-regulated splicing using rna sequence and cellular context. *Bioinformatics*, 27(18):2554–2562, Sep 2011.

[112] Gene W. Yeo, Nicole G. Coufal, Tiffany Y. Liang, Grace E. Peng, Xiang-Dong Fu, and Fred H. Gage. An rna code for the fox2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–137, Feb 2009.

[113] Gene W. Yeo, Eric Van Nostrand, Dirk Holste, Tomaso Poggio, and Christopher B. Burge. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A*, 102(8):2850–2855, Feb 2005.

[114] A. M. Zahler, W. S. Lane, J. A. Stolk, and M. B. Roth. Sr proteins: a conserved family of pre-mrna splicing factors. *Genes Dev*, 6(5):837–847, May 1992.

[115] Chaolin Zhang, Maria A. Frias, Aldo Mele, Matteo Ruggiu, Taesun Eom, Christina B. Marney, Huidong Wang, Donny D. Licatalosi, John J. Fak, and Robert B. Darnell. Integrative modeling defines the nova splicing-regulatory network and its combinatorial controls. *Science*, 329(5990):439–443, Jul 2010.

[116] Xiang H-F. Zhang and Lawrence A. Chasin. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci U S A*, 103(36):13427–13432, Sep 2006.

[117] Xiangqun Zheng-Bradley, Johan Rung, Helen Parkinson, and Alvis Brazma. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol*, 11(12):R124, 2010.

[118] Xiang-Yang Zhong, Jian-Hua Ding, Joseph A. Adams, Gourisankar Ghosh, and Xiang-Dong Fu. Regulation of sr protein phosphorylation and alternative splicing by modulating kinetic interactions of srpk1 with molecular chaperones. *Genes Dev*, 23(4):482–495, Feb 2009.

[119] Zhaolan Zhou, Lawrence J. Licklider, Steven P. Gygi, and Robin Reed. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–185, Sep 2002.

Chapter 2

Widespread Regulated Alternative Splicing of Single Codons Accelerates Proteome Evolution

As published in PLoS Biology:

Bradley RK, Merkin J, Lambert NJ, Burge CB (2012) Alternative Splicing of RNA Triplets Is Often Regulated and Accelerates Proteome Evolution. PLoS Biol 10(1): e1001229. doi:10.1371/journal.pbio.1001229

My contribution:

Text sections "NAGNAGs accelerate protein evolution at exon-exon boundaries" and "NAGNAG-accelerated protein evolution is highly biased", relevant methods, Figure 5, and Figures S9-S11.

2.1 Abstract

Thousands of human genes contain introns ending in NAGNAG (N any nucleotide), where both NAGs can function as 3' splice sites, yielding isoforms that differ by inclusion/exclusion of three bases. However, few models exist for how such splicing might be regulated, and some studies have concluded that NAGNAG splicing is purely stochastic and nonfunctional. Here, we used deep RNA-Seq data from sixteen human and eight mouse tissues to analyze the regulation and evolution of NAGNAG splicing. Using both biological and technical replicates to estimate false discovery rates, we estimate that at least 25% of alternatively spliced NAGNAGs undergo tissue-specific regulation in mammals, and alternative splicing of strongly tissue-specific NAGNAGs was ten times as likely to be conserved between species as was splicing of non-tissue-specific events, implying selective maintenance. Preferential use of the distal NAG was associated with distinct sequence features, including a more distal location of the branch point and presence of a pyrimidine immediately before the first NAG, and alteration of these features in a splicing reporter shifted splicing away from the distal site. Strikingly, alignments of orthologous exons revealed a ~15-fold increase in the frequency of 3 base pair gaps at 3' splice sites relative to nearby exon positions in both mammals and in *Drosophila*. Alternative splicing of NAGNAGs in human was associated with dramatically increased frequency of exon length changes at orthologous exon boundaries in rodents, and a model involving point mutations that create, destroy or alter NAGNAGs can explain both the increased frequency and biased codon composition of gained/lost sequence observed at the beginnings of exons. This study shows that NAGNAG alternative splicing generates widespread differences between the proteomes of mammalian tissues, and suggests that the evolutionary trajectories of mammalian proteins are strongly biased by the locations and phases of the introns that interrupt coding sequences.

2.2 Introduction

The split structure of eukaryotic genes impacts gene expression and evolution in diverse ways. Most directly, the presence of introns enables multiple distinct mRNA and protein products to be produced from the same gene locus through alternative splicing, which is often regulated between tissues or developmental stages (1, 2). Alternative inclusion or exclusion of exons — "exon skipping" — can generate protein isoforms with distinct subcellular localization, enzymatic activity or allosteric regulation, and differing, even opposing, biological function (3-5). Splicing is often regulated by enhancer or silencer motifs in the pre-mRNA that are bound by splicing regulatory proteins that interact with each other or with the core splicing machinery to promote or inhibit splicing at nearby splice sites (6). Such enhancer and silencer motifs are common throughout constitutive as well as alternative exons and their flanking introns (7-9). In turn, the presence of splicing regulatory motifs in exons, and their higher frequency near splice junctions, impacts protein evolution. For example, the frequencies of single nucleotide polymorphisms (SNPs) and amino acid substitutions are both reduced near exon-exon junctions relative to the centers of exons as a result of selection on exonic splicing enhancer motifs (10, 11). Thus, a genes exon-intron structure and its evolution are intimately linked.

Alternative 3' and 5' splice site use, in which longer or shorter versions of an exon are included in the mRNA, are among the most common types of alternative splicing in mammals (1) and can generate protein isoforms with subtly or dramatically differing function. For example, production of the pro-apoptotic Bcl-xS or the anti-apoptotic Bcl-xL protein isoforms is controlled through regulated alternative splice site usage (12). Binding of splicing regulatory factors between the alternative splice sites or immediately adjacent to one site or the other can shift splicing toward the (intron-) proximal or distal splice site (6, 13, 14), providing a means to confer cell type-specific regulation. The distance between the alternative splice sites can vary over a wide range, from hundreds of bases to as few as 3 bases in the case of NAGNAG alternative 3' splice sites.

NAGNAG alternative splicing (Fig. 2-1a) has been observed in vertebrates, insects, and plants, and is known to be very common. Bioinformatic analyses of expressed sequence tag (EST) databases have identified thousands of examples (15-18). However, most of the mechanisms known to regulate other alternative 3' splice site pairs, particularly those that involve binding of regulatory factors between the sites, or much closer to one site than the other, cannot apply to NAGNAGs because of the extreme proximity of the two sites. Thus, regulation of NAGNAGs is more difficult to envisage. Furthermore, analyses of select genes using PCR and capillary electrophoresis approaches reached differing conclusions about NAGNAG tissue specificity (15, 19, 20), and several authors have argued that NAGNAG splicing is purely stochastic, is not evolutionarily conserved, and is not physiologically relevant (21, 22). However, analyses of NAGNAG splicing at a genome-wide scale have been hampered by the impracticality of distinguishing such similar isoforms by microarray hybridization and the insufficient depth of EST databases for assessment of tissue specificity.

In order to assess the abundance and potential regulation of NAGNAG splicing events genome-wide, we analyzed polyA-selected RNA-Seq data generated using the Illumina HiSeq platform from sixteen human tissues at depths of ~8 Gbp per tissue, similarly deep RNA-Seq data that we generated from eight mouse tissues, and data generated by the modENCODE consortium across a developmental time course in *Drosophila*. NAGNAG isoforms can be uniquely distinguished by short reads that overlap the splice junction, and the quantity of data available from each tissue in human and mouse typically represented at least 80-fold mean coverage of the transcriptome, a depth sufficient to detect potential tissue-specific differences in many cases. Sequence features were identified which can shift splicing toward the proximal or distal NAG, providing clues to regulation. We also analyzed the impact of NAGNAGs on exon evolution, obtaining evidence that NAGNAGs dramatically accelerate addition and deletion of sequence at the beginnings of exons.

2.3 Results and Discussion

2.3.1 Many human NAGNAGs are regulated across tissues

Our initial analyses used the Illumina Body Map 2.0 dataset of polyA-selected RNA-Seq data from sixteen human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells) sequenced at depths of ~80 million paired-end 2 x 50 bp reads per tissue. This sequencing depth generates ~8 Gbp of data, representing >80-fold coverage of the human protein-coding transcriptome. Enumerating all possible NAGNAG splicing events, we mapped both ends of each read against NAGNAG splice junctions (Fig. 2-1a). Isoform ratios were estimated across all tissues as "percent spliced in" (PSI or ψ) values (Fig. 2-1b), representing the fraction of mRNAs that use the intron-proximal splice site, thereby including the second NAG in the mRNA. The reliability of such RNA-Seq-based estimates of isoform abundance has been established previously (23).

Using a conservative approach that has comparable power to detect each of the major types of alternative splicing events, we estimated that NAGNAGs comprise slightly more than twenty percent of reading frame-preserving alternative splicing events in coding regions, making NAGNAGs the most common form of protein producing alternative splicing after exon skipping (Fig. 2-1c). In all, more than two thousand NAGNAG events were detected in protein-coding regions of human genes where both isoforms were expressed at 5% in at least one tissue. Strikingly, 73% of these NAGNAGs showed evidence of tissue-specific regulation ($p < 0.01$ by multinomial test). Furthermore, approximately 42% were "strongly regulated", with changes in ψ of at least 25% between tissues. For example, a NAGNAG in the gene encoding FUBP1, a transcriptional regulator of MYC, undergoes dramatically different splicing between kidney and lymph node (Fig. 2-1b). Here, we report absolute rather than relative differences in splicing levels, e.g., a change from 10% to 35% between tissues

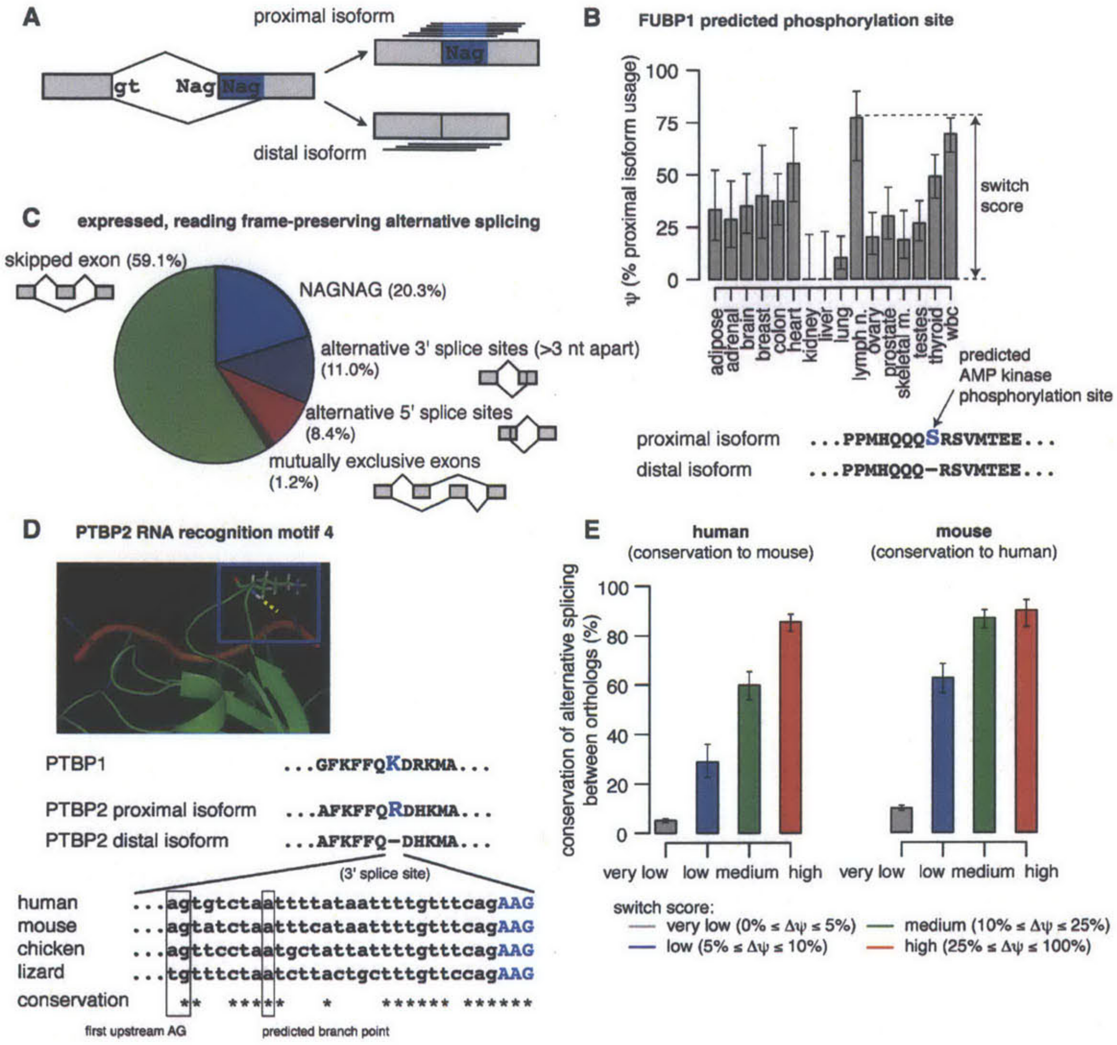


Figure 2-1:

(a) Short reads were aligned to the intron-proximal and intron-distal splice junctions of NAGNAG splicing events in order to estimate isoform ratios.

(b) Estimated proximal isoform usage psi for a NAGNAG which inserts/deletes a predicted phosphorylation site in far upstream element binding protein 1 (FUBP1). Phosphorylation site and corresponding kinase were predicted by Scansite (Scansite z-score -3.02.84) (55). Error bars indicate the 95% binomial confidence interval.

(c) Number of reading frame-preserving alternative splicing events in protein-coding regions, with both isoforms expressed at ge 5 in at least one tissue.

(d) A NAGNAG which inserts/deletes an arginine in RNA recognition motif 4 (RRM4) of the splicing factor PTBP2 is deeply conserved. Alignment of orthologous 3' splice site sequences shown below the NMR structure (PDB accession 2ADC, displayed with PyMOL) of the highly homologous PTBP1 protein (green) complexed with RNA (red) (33). Boxed is K489 of PTBP1, which is homologous to the arginine shown in PTBP2, and hydrogen bonds to the RNA backbone (dotted yellow line). Putative branch point based on location of the first upstream AG, the sequence motif identified in (56), and the pattern of sequence conservation.

(e) Conservation of alternative splicing between orthologous human and mouse NAGNAGs increases with tissue specificity. NAGNAGs that were alternatively spliced in human (left) and mouse (right) were grouped by switch score — defined as the maximum psi difference between tissues — as indicated by colors, and the fraction of orthologs which were alternatively spliced in the other species is shown. Error bars indicate 95% binomial confidence intervals.

is considered an increase of 25%, not 250%, and the largest difference in ψ between tissues is defined as the "switch score" (1). Other genes containing NAGNAGs with switch scores of 50% or more included HOXD8, CAMK2B, ATRX, CAPRIN2, and MLLT4. Technical replicates — sequencing of the same RNA-Seq libraries with 75 bp single-end reads at depths of ~50 million reads per tissue — yielded similar estimates of NAGNAG abundance and regulation.

2.3.2 Regulated NAGNAGs are selectively conserved between primates and rodents

Regulation that contributes to fitness is expected to be evolutionarily conserved. A previous study reported the existence of selection against alternatively spliced NAGNAGs in coding sequences (24). Nevertheless, some NAGNAGs are quite deeply conserved, e.g., a NAGNAG that generates an arginine insertion/deletion in a RNA-binding domain of the splicing factor PTBP2 (also known as nPTB or brPTB). Both isoforms of this NAGNAG event are observed in ESTs from human, mouse and chicken, and the potential for alternative splicing is conserved at the sequence level to lizard (Fig. 2-1d). Consistent with this example, a previous analysis of EST databases suggested that a subset of alternatively spliced NAGNAGs are under purifying selection in vertebrates (25). We systematically assessed the global conservation of NAGNAG isoform levels using RNA-Seq data generated from eight mouse tissues (brain, colon, kidney, liver, lung, skeletal muscle, spleen, and testes). Restricting to the set of NAGNAGs which were alternatively spliced in human (both isoforms expressed at $\geq$ 5% in at least one tissue), we found that NAGNAGs which were strongly regulated were approximately ten times more likely than unregulated NAGNAGs to exhibit alternative splicing in their mouse orthologs, and vice versa (Fig. 1e). This large and consistent increase in conservation of alternative splicing with increasing switch score suggests that regulated NAGNAGs are much more likely to contribute to organismal fitness, and therefore to be selectively maintained, than are alternatively spliced events which do not exhibit tissue specificity. If NAGNAG alternative splicing were selectively neutral, then we would not expect to see a correlation between the observed degree of tissue specificity in one species and conservation of alternative splicing in the other species.

NAGNAG isoform levels were very well correlated between biological replicates, consisting of individual mice of strains C57BL/6J and DBA/2J, whose genomes differ to an extent similar to that of unrelated humans ($r = 0.96$, Fig. 2-2a), demonstrating the robustness and reproducibility of these RNA-Seq-based estimates of NAGNAG ψ values. Similar numbers of alternatively spliced NAGNAGs were detected in mouse as in human, with 28% of alternatively spliced NAGNAGs in mouse exhibiting evidence of tissue-specific regulation and 8% being strongly regulated across the eight tissues studied. Many orthologous NAGNAGs in human and mouse exhibited tissue-specific regulation in both species, e.g. NAGNAGs in FUBP1, CAMK2B, CAPRIN2, and ATRX. The higher fraction of regulated NAGNAGs detected in the human data probably results from a combination of factors, including the greater number of tissues sampled (Supplementary Fig. 2-3), the diverse genetic backgrounds of the human samples, and intrinsically higher read coverage variability in the human RNA-Seq data used. Comparing technical replicates of human tissues, which capture variability in sequencing, we estimated false discovery rates (FDRs) for discovering strongly regulated NAGNAGs ranging from ~0.8% to ~13.3%, with a mean FDR of 4.4% (Supplementary Fig. 2-4). In contrast, comparing biological replicates of mouse tissues, which capture all major sources of variability (tissue collection, library preparation, sequencing, and individual-specific splicing differences), we estimated FDRs ranging from 0.6% to 1.9%, with a mean of 1.1% (Supplementary Fig. 3). Using these estimated FDRs, and extrapolating the mouse data to 16 tissues (Supplementary Fig. 2-3), we estimated that between 12% and 37% of NAGNAGs are strongly regulated across tissues in mammals, making strong regulation a fairly common occurrence — though somewhat less common than for other types of splicing events. The relatively small differences between samples of the same tissue from mice whose

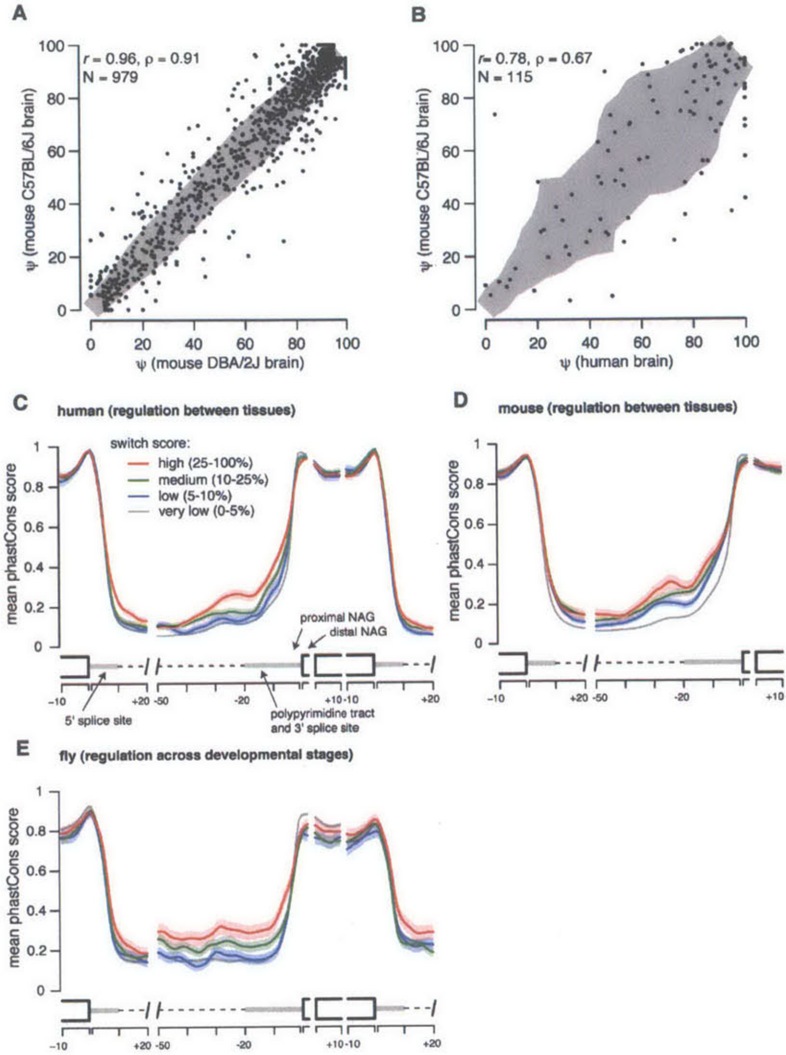


Figure 2-2:

(a) NAGNAG ψ estimates are highly consistent in brain RNA-Seq data from the mouse strains DBA/2J and C57BL/6J. Only NAGNAGs with both isoforms expressed at $\geq 5\%$ in either strain are shown. The 75th percentile of the deviation from the line $y = x$ is shown in gray.

(b) NAGNAG ψ estimates are quantitatively conserved between human and mouse brain. Only NAGNAGs with both isoforms expressed at $\geq 5\%$ in either species and satisfying proximal 3' splice site score - distal 3' splice site score ≤ 0.5 bits are plotted (splice sites scored by MaxEnt model (36)). Deviation from $y = x$ shown as in (a).

(c) Sequence conservation of human NAGNAGs, where all NAGNAGs are aligned by their 3' splice site junctions and grouped by switch score. Mean (solid line) and standard error of the mean (shaded area about solid line) of phastCons score (50) shown by position (averaged over a 2 nt sliding window) for each switch score category. Analysis restricted to human NAGNAGs for which the two AGs were conserved at the sequence level in mouse.

(d) As in (c), but grouped by switch score in mouse and restricted to mouse NAGNAGs for which the two AGs were conserved at the sequence level in human.

(e) As in (d), but for NAGNAGs in *Drosophila melanogaster*, with switch score defined across developmental stages rather than between tissues. Analysis restricted to *D. melanogaster* NAGNAGs for which the two AGs were conserved at the sequence level in D. yakuba.

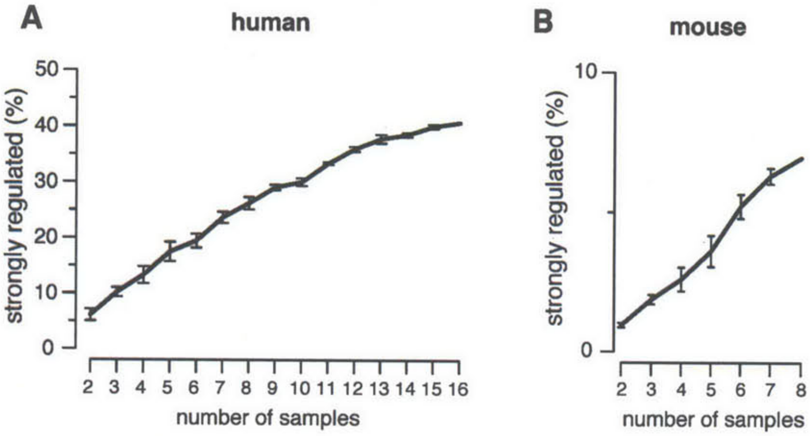


Figure 2-3:

(a) Human

(b) Mouse

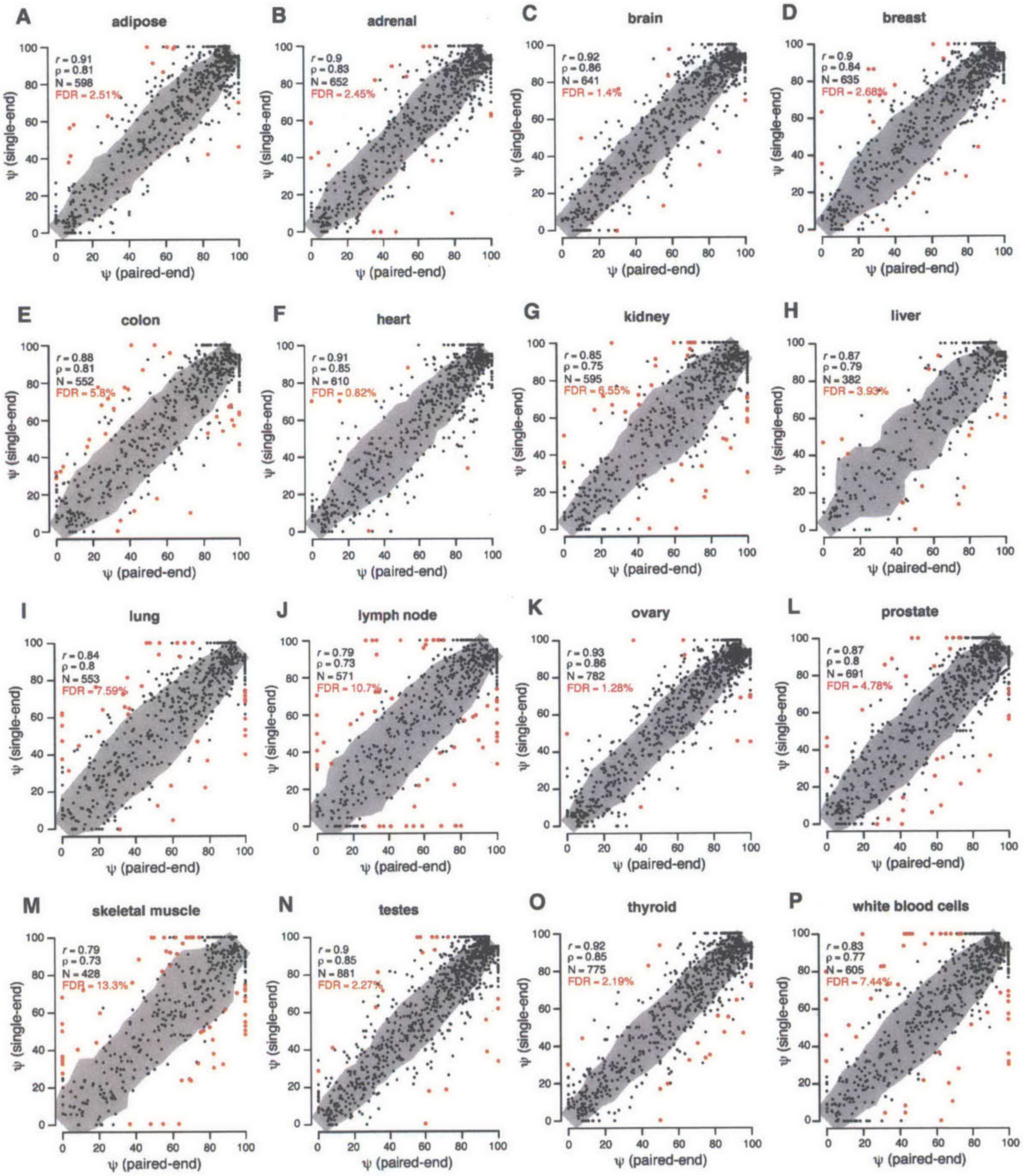


Figure 2-4:
Single-end (75 bp) and paired-end (250 bp) sequencing of the same human libraries captures sequencing variability.

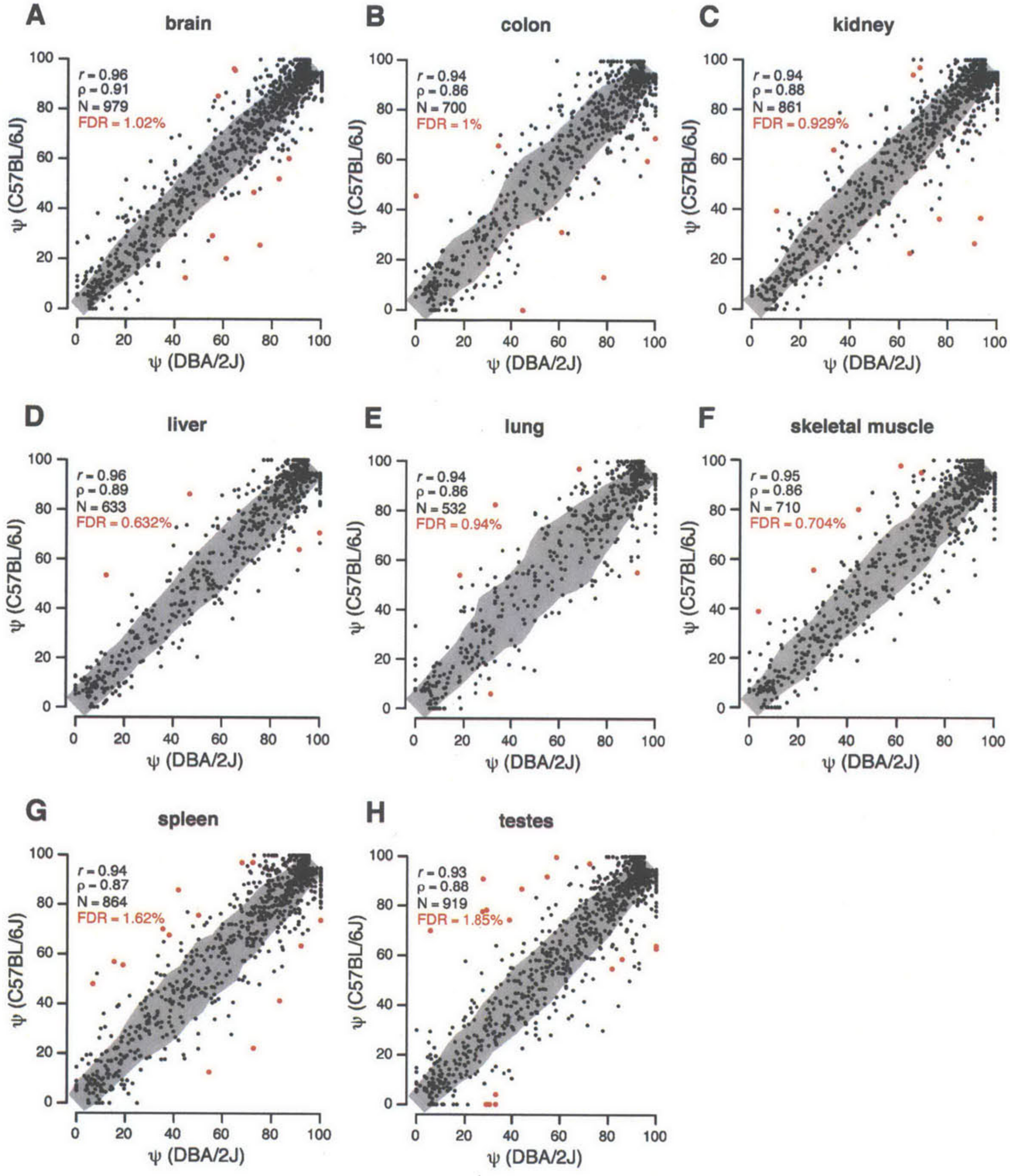


Figure 2-5:
Sequencing of mouse libraries created from two different individuals captures all major sources of variability, including library preparation (2x36 bp versus 2x80 bp), sequencing, sample collection, and individual-specific splicing (C57BL/6J versus DBA/2J).

genomes differed to an extent comparable to that of unrelated humans (Fig. 2-2a) suggested that inter-individual variation contributed less than other sources of variation (e.g., tissue-specific differences) to the variations observed between the human libraries.

Orthologous human and mouse NAGNAGs exhibited high quantitative conservation of isoform levels. This was particularly true when the difference between the proximal and distal 3' splice site scores — using a method that scores the strength of the polypyrimidine tract and AG region — was conserved (Spearmans $\rho = 0.67$, Fig. 2-2b). The correlation decreased somewhat in cases where the differences in 3' splice site scores were less conserved ($\rho = 0.54$, p = 0.013 for test of equality of correlation using the Fisher transformation; Supplementary Fig. 2-6), suggesting that changes in relative 3' splice site strength may contribute to species-specific differences in NAGNAG splicing. Notably, many NAGNAGs with diverged splice site scores were alternatively spliced in one species but constitutively spliced in the other, suggesting relatively rapid evolution of 3' splice site positions.

2.3.3 Regulated NAGNAGs have conserved upstream intronic sequence

To better understand how NAGNAG splicing is regulated, and which sequence regions might be involved, we examined sequence conservation of flanking intronic and exonic regions for NAGNAGs grouped by switch score using alignments of the genomes of placental mammals. Tissue-specific NAGNAGs exhibited markedly increased sequence conservation in the upstream intron (Fig. 2c-d), with little or no increase in other analyzed regions. The consistent increase in conservation in the upstream intron with increasing switch score provides further evidence that these regulated NAGNAGs contribute to organismal fitness, and is consistent with previous observations that alternatively spliced NAGNAGs have higher upstream sequence conservation than constitutive 3' splice sites (26). Enumerating NAGNAGs in introns of the fly

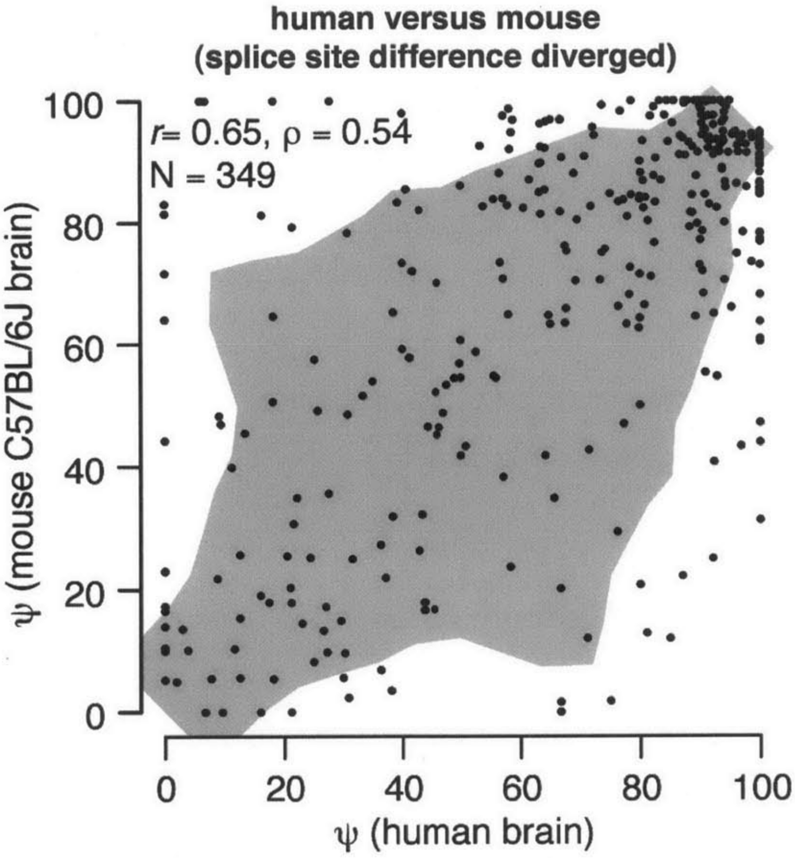


Figure 2-6:
As Figure 2B, but for NAGNAGs with | $proximalsplicesitescore - distalsplicesitescore$ | >0.5.

Drosophila melanogaster, and comparing isoform usage across thirty developmental time points (embryo to adult) using RNA-Seq data from the modENCODE consortium (2), we identified over five hundred NAGNAGs in coding regions of *Drosophila* genes where both isoforms were expressed at $\geq 5\%$ in at least one developmental time point. Of these, 14% were developmentally regulated, with 5% being strongly regulated as defined above. As in mammals, more highly regulated fly NAGNAGs were associated with increased sequence conservation within and upstream of the 3' splice site (Fig. 2-2e). The consistent location of the sequence conservation signal for regulated NAGNAGs in mammalian and insect genomes (Fig. 2-2c-e) suggested that the region ~50 bp upstream of the NAGNAG motif, encompassing the competing 3' splice sites themselves, may contain most of the regulatory information that governs NAGNAG alternative splicing. The extensive tissue-specific regulation observed in mammals and developmental regulation seen in flies may indicate that regulated NAGNAG alternative splicing is widespread in metazoans.

2.3.4 Splice site score difference explains mean NAGNAG isoform expression

The increased divergence in isoform usage observed for NAGNAGs that had undergone divergence in 3' splice site score difference (Fig. 2-2b, Supplementary Fig. 2-6) suggested that relative splice site strength is a major determinant of NAGNAG quantitative isoform usage. Supporting this hypothesis, previous EST-based analyses have demonstrated that splice site strength impacts whether or not a NAGNAG will be alternatively spliced (21, 27). To explore the relationship between splice site strength and quantitative isoform levels, rather than simply the presence or absence of alternative splicing, we created a biophysical model wherein the probabilities of using the proximal and distal splice sites are proportional to $Q \cdot e^{Bi(proximalscore)}$ and $e^{Bi(proximalscore)}$, respectively, where the parameter Q determines the inherent preference for using the intron-proximal splice site and B is a scaling factor for the splice site scores. This simple model, containing just two free parameters, accurately predicted mean isoform usage across human tissues (Fig. 2-8a), suggesting that relative 3' splice site strength is the primary determinant of basal NAGNAG isoform levels. The fitted value $Q = 0.55$ provides a quantitative measurement of preference for the proximal splice site in NAGNAG 3' splice site recognition, predicting that the distal splice site of a NAGNAG must typically be $-log(Q)/B = -log(0.55)/0.58 = 1$ bit stronger than the proximal splice site in order to be spliced with equal efficiency. Analysis of mouse NAGNAGs yielded similar values of the Q and B parameters (Supplementary Fig. 2-7), supporting the robustness of these estimates. This preference for the proximal site was obvious even after controlling for the identity of the -3 bases (the Ns of the NAGNAG) (Fig. 2-8b), which are known to be important determinants of NAGNAG isoform choice (18, 26, 27). Preference for the proximal splice site is consistent with models of 3' splice site recognition that involve scanning or diffusion from an upstream branch point (28, 29).

While the mean ψ value was accurately predicted by our model, the variability around the mean was substantially higher than expected based on measurement noise (Fig. 2-8a). This observation is consistent with the concept that splice site strength determines the basal levels of the two NAGNAG isoforms, but the presence of regulatory sequence elements not captured by the 3' splice site score, and variation in the levels of associated trans-acting factors, modulates the isoform ratios that occur in different tissues.

2.3.5 Specific sequence features associated with basal and regulated NAGNAG splicing

The observed regulation of NAGNAGs implies that features outside of splice site strength and the -3 base must also be involved in determining isoform usage. For example, the NAGNAG in the splicing factor PTBP2 (Fig. 2-1d) represents an exception to the pattern observed above: the -3 bases (CAGAAG) predict predominant

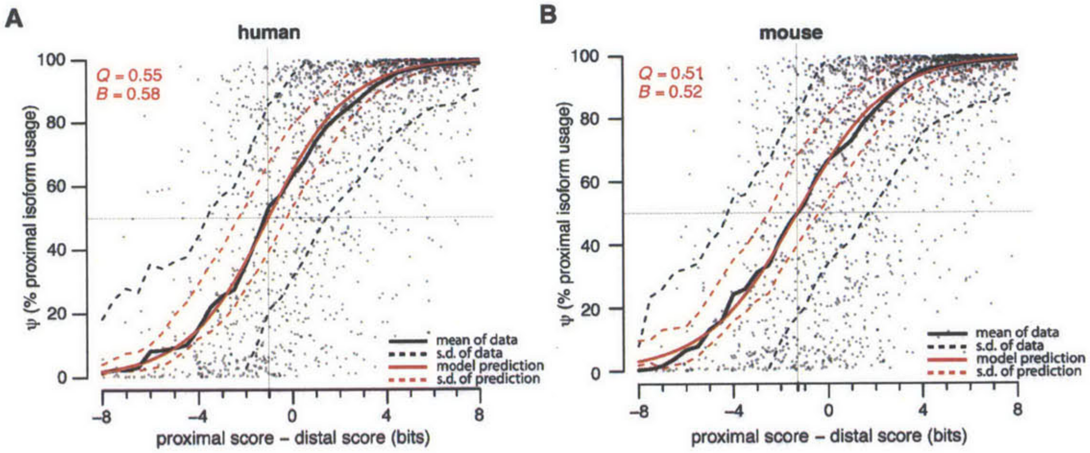


Figure 2-7:

(a) Human (identical to Figure 2-8a)

(b) Mouse

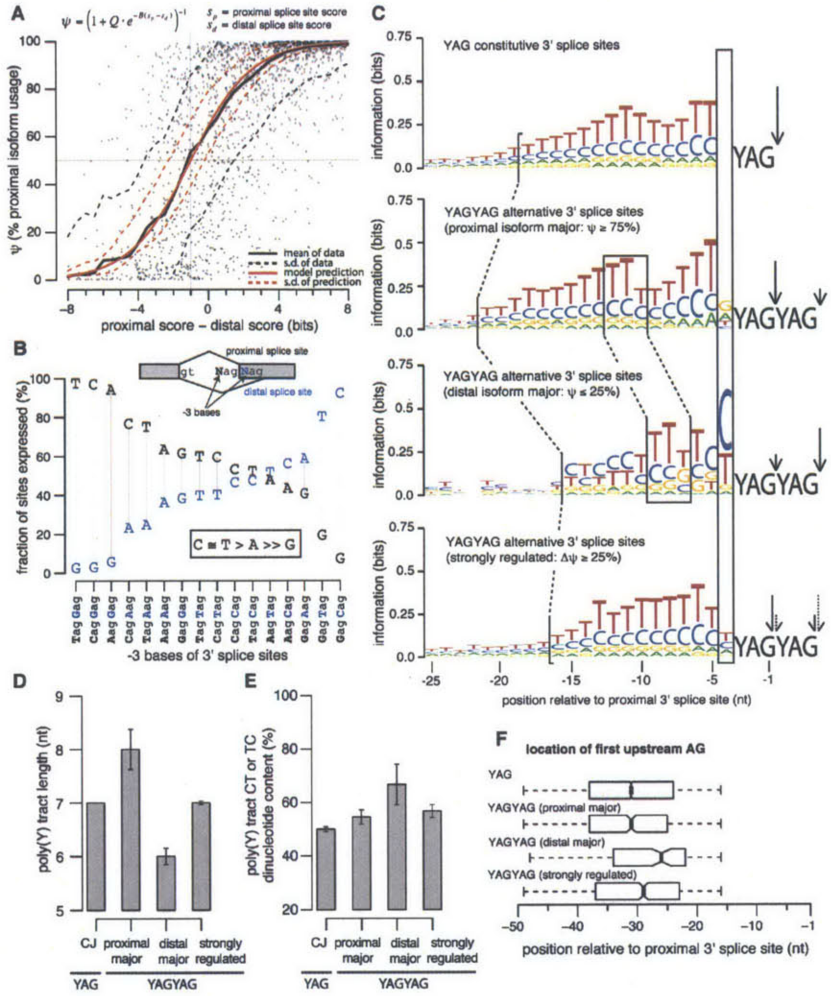


Figure 2-8:

(a) A simple biophysical model of NAGNAG splicing accurately models mean isoform usage across tissues as a function of difference in 3' splice site score. Each point represents a single human NAGNAG, and the solid and dashed black lines show the mean ψ (across values for individual NAGNAGs with similar splice site score difference, with sliding window of 3.25 bits) and the standard deviation about the mean. The solid red line shows the prediction based on the model for parameters $Q = 0.55$ and $B = 0.58$, and the dashed red line indicates the standard deviation about the model mean expected from measurement error. The horizontal and vertical dashed lines indicate the splice site score difference (approximately 1 bit) at $\psi = 50\%$.

(b) The -3 bases largely determine whether a NAGNAG is alternatively spliced. We grouped NAGNAGs in the human genome according to their -3 bases and computed the fraction of each group which expressed the proximal (black) or distal (blue) isoform at $\geq 5\%$ in at least one tissue.

(c) Constitutive 3' splice sites (top, YAG), YAGYAGs which express the proximal isoform at $\geq 75\%$ in all tissues (middle, YAGYAG proximal major), YAGYAGs which express the distal isoform at $\geq 75\%$ in all tissues (middle, YAGYAG distal major), and strongly regulated YAGYAGs (bottom, YAGYAG strongly regulated) all exhibit distinct upstream sequence preferences. The x-axis shows the position relative to the 3' splice site (YAG) or proximal 3' splice site (YAGYAG), and arrows indicate the 3' splice site that is predominantly used. Figure was created with WebLogo (53). Human and mouse YAGYAGs were grouped together to increase the statistical signal for (c-f).

(d) Distal major YAGYAGs have shorter polypyrimidine tracts (p <0.001 relative to proximal major class, Kolmogorov-Smirnov test). Plot shows median length of the polypyrimidine tract, estimated as the first stretch of ≥ 5 consecutive pyrimidines upstream of the -3 position. Error bars indicate the standard deviation of the median, estimated by bootstrapping (the error bars for CJ were too small to be visible).

(e) Distal major YAGYAGs have higher CT and TC dinucleotide content (p <0.005 relative to proximal major class, Kolmogorov-Smirnov test). Median CT and TC dinucleotide content of the polypyrimidine tract, computed as the fraction of the polypyrimidine tract composed of CT dinucleotides, with an optional T at the beginning or C at the end. Error bars indicate the standard deviation of the median, estimated by bootstrapping.

(f) The AG exclusion zone (57) is more distally located in distal major YAGYAGs (p <0.001 relative to proximal major class, Kolmogorov-Smirnov test). Position of the first AG dinucleotide upstream of the -15 position is proximal splice site usage, since C is strongly favored over A and is also proximal, but roughly equal proportions of both isoforms are expressed across all tissues studied (Supplementary Fig. 2-9). This observation led us to wonder whether other aspects of this 3' splice site, e.g., the relatively short and distally located polypyrimidine tract and the relatively distal location of the putative branch point (Fig. 2-1d) might favor use of the distal NAG in this and other cases.

While many analyses support the importance of the -3 base combination in NAGNAG alternative splicing (18, 26, 27), there is less consensus in the literature about the relevance of other major elements of the 3' splice site, including the polypyrimidine tract and branch site. Molecular genetics experiments demonstrated that mutating sequences near the polypyrimidine tract and branch site influenced alternative splicing of specific NAGNAGs (30, 31), but two computational studies that used machine-learning approaches (27, 32) concluded that neither of these elements significantly influenced NAGNAG splicing globally. Notably, the experimental studies (30, 31) measured quantitative isoform ratios, as we do in this study, while the machine-learning studies (27, 32) simply classified NAGNAGs as constitutively or alternatively spliced.

In order to dissect features that impact NAGNAG isoform choice, controlling for the effect of the -3 bases, we considered the large class of NAGNAGs with favored (C or T) nucleotides at both -3 bases (YAGYAGs). We found that exons that predominantly used the proximal splice site ("proximal major" YAGYAGs) had substantially distinct nucleotide preferences from those that predominantly used the distal site ("distal major" YAGYAGs) (Fig. 2-8c), consistent with the experimental results of Tsai et al. (30, 31), who found that modifying the sequence upstream of the 3' splice site influenced NAGNAG splicing. For example, distal major YAGYAGs tended to have shorter, more distal, polypyrimidine tracts than proximal major YAGYAGs (Fig. 2-8d), implicating polypyrimidine tract length and location in control of NAGNAG splicing. The proportion of CT/TC dinucleotides in the polypyrimidine tract was ~25% higher for distal major YAGYAGs (Fig. 2-8e), suggesting the possible in-

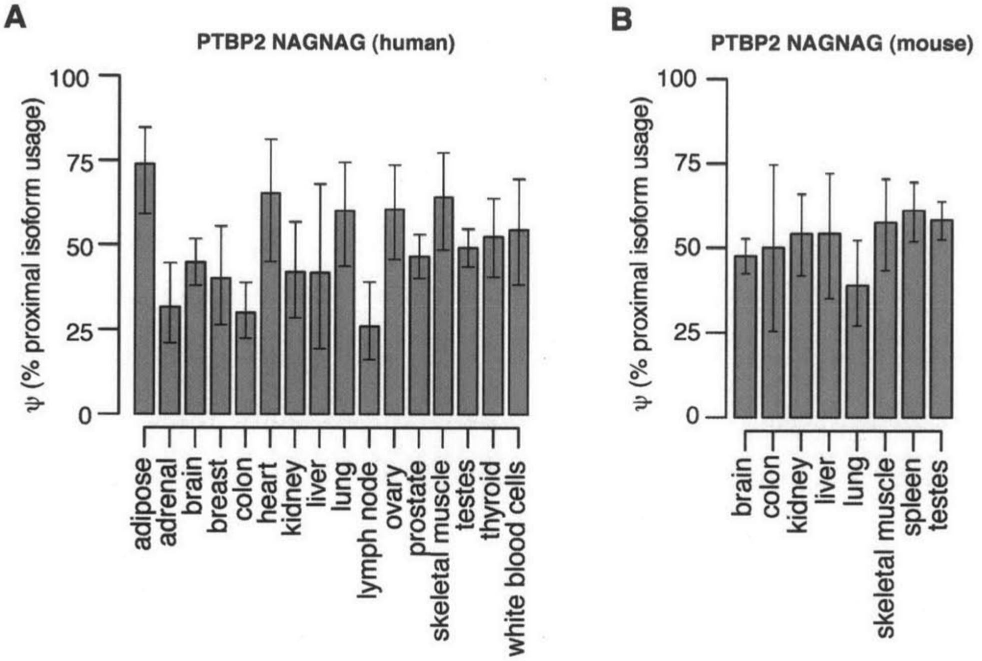


Figure 2-9:
Isoform usage of the NAGNAG in the PTBP2 gene illustrated in Figure 1D.

(a) Human

(b) Mouse

volvement of CU/UC-binding factors such as those of the PTB family (33) — some of which are tissue-specifically expressed — in promoting use of distal NAGs. The location of the first upstream AG was also shifted several bases downstream in distal major YAGYAGs compared to other 3' splice sites (Fig. 2-8f), suggesting that the branch site is located further downstream in this class and that use of a distally located branch site favors use of the distal YAG, perhaps because the distance to the 3' splice site is more optimal.

Strongly regulated YAGYAGs had features that were intermediate between the extremes found for proximal major and distal major YAGYAGs, such as polypyrimidine tracts of intermediate length (Fig. 2-8d), suggesting that the presence of intermediate features facilitates regulation. Increased regulation was also associated with reduced 3' splice site strength and greater similarity in strength between the competing sites (Supplementary Fig. 2-10), consistent with previous studies of other types of alternative splicing (34).

The -42 base, four nucleotides upstream of the 3' splice site, is not generally considered to be important in splicing (with rare exceptions (35)). This position contains little or no information in alignments of constitutive 3' splice sites (36), although a previous machine-learning analysis of features distinguishing between constitutively and alternatively spliced NAGNAGs included the -4 base in their classifier (27). Our quantitative analysis strongly supported a special role in NAGNAG regulation for this canonically unimportant position. For distal major YAGYAGs, the -4 position (here referring to the position four nucleotides upstream of the intron-proximal splice site) had the highest information content of any position upstream of the YAGYAG (Fig. 2-8c); furthermore, the -4 base was more conserved in distal major and strongly regulated YAGYAGs than for other classes of 3' splice sites (Supplementary Fig. 2-11).

Of the observations in Figure 3, the two that seemed most compelling were the preference for pyrimidines at the -4 position and the more distal positioning of branch

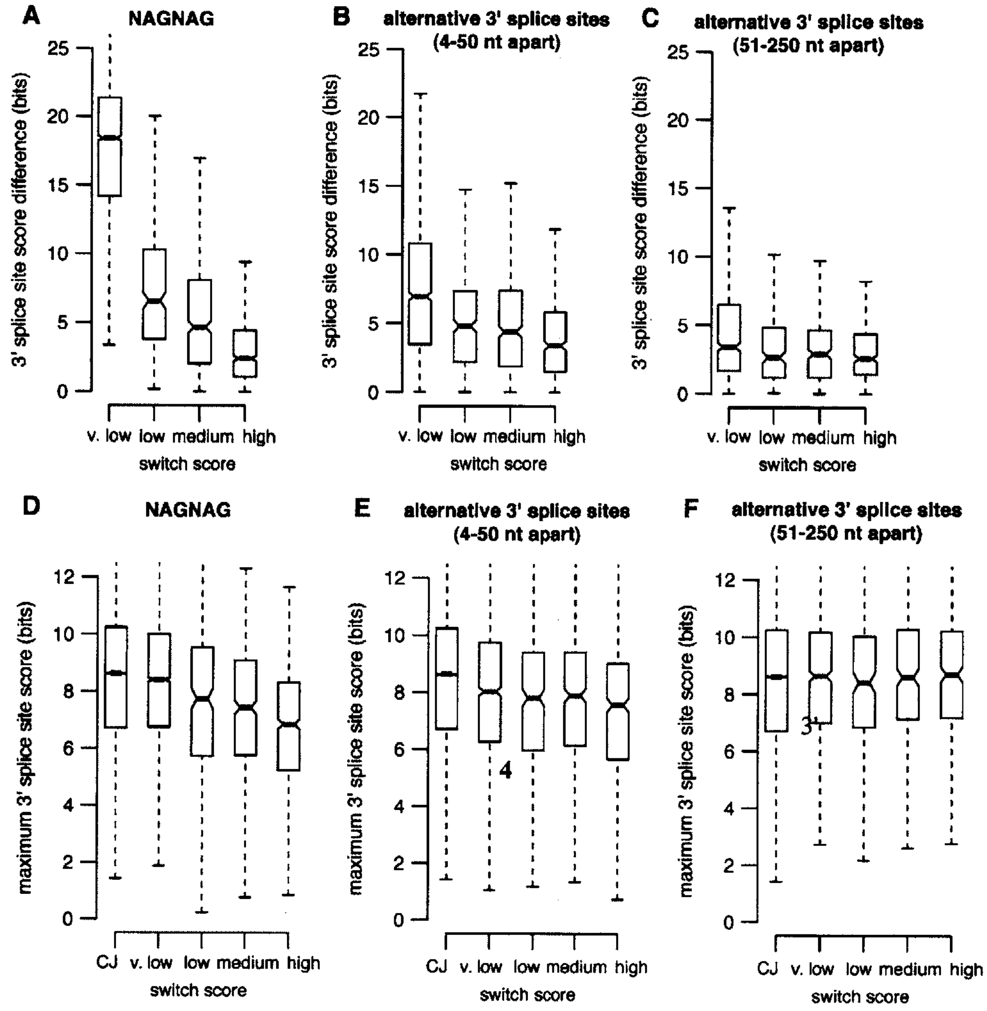


Figure 2-10:

(a) The splice site scores of regulated NAGNAG 3 splice sites tended to be far more similar to one another than those of unregulated events, suggesting that regulation is easier to achieve when the intrinsic strengths of the sites are evenly matched.

(b) This trend was much weaker for more distant alternative 3 splice site events.

(c) As in (B), but with a longer distance

(d) The 3 splice site scores of tissue-regulated NAGNAGs also tended to be somewhat weaker than for unregulated NAGNAGs or constitutive 3 splice sites. This observation suggested that weaker splice sites are more easily regulated, consistent with previous studies of other types of alternative splicing.

(e) This trend for regulated events to be associated with weaker splice site scores was observed to a much lesser extent for alternative 3 splice sites separated by longer distances, suggesting that splicing regulatory elements may more readily exert differential effects on more widely spaced 3 splice sites, making matching of splice site scores less critical for achieving regulation for this class than it is for NAGNAGs. For example, we have previously shown that most exonic splicing silencer (ESS) elements inhibit the intron-proximal site when situated between competing 3 splice sites, an arrangement that requires separation of the competing sites by sufficient space to accommodate the ESS, and so does not apply to NAGNAGs. "v. low" indicates "very low," and "CJ" indicates the 3 splice sites of constitutive junctions.

(f) As in (E), but a longer distance

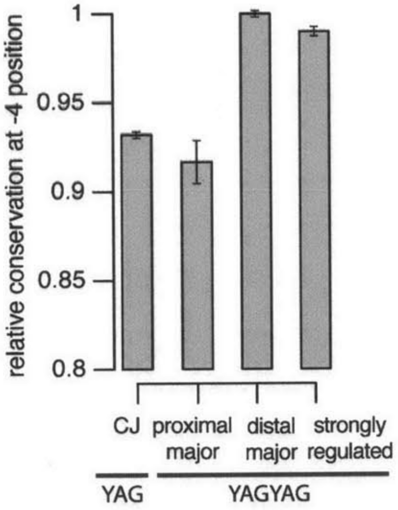


Figure 2-11:
Plot shows median relative conservation at the -4 position, computed as (phastCons score at -4 position/phastCons score at -3 position). "CJ" indicates the 3 splice sites of constitutive junctions. Error bars indicate the standard error of the median, estimated by bootstrapping.

points in YAGYAGs that favored the distal splice site. To test the predicted role of the -4 base in regulation of NAGNAG splicing, we used a minigene reporter based on the NAGNAG in PTBP2, whose splicing alters an exon coding for the RRM4 RNA binding domain (Fig. 2-1d, 4a). As predicted based on the data in Fig. 2-8c, mutation of the -4 base (T in the wildtype) to A or G resulted in a substantial shift in splicing toward use of the proximal NAG, while mutation to C had no effect (Fig. 2-12b). These observations confirm that presence of a pyrimidine at the -4 position favors use of the distal NAG, even though no sequence preference was observed at this position in constitutive splice sites (Fig. 2-8c). Presence of a pyrimidine at the -4 position of a NAGNAG might function to shift the location of binding of U2AF65 downstream by a base or more from its normal position, which might then result in preferential binding of U2AF35 to the downstream NAG, though this will require further study. We also tested the role of the branch point in NAGNAG splicing by manipulating the branch site to 3' splice site distance in this reporter, either in a context in which the inferred native branch point sequence (BPS) was intact or in a context in which the native BPS had been replaced by the previously mapped BPS of IGF2BP1 intron 11 (Fig. 2-12a). With the native BPS present, an increase of just 4 bases in the BPS-3' splice site distance was sufficient to cause a substantial shift in splicing towards the proximal NAG, with little or no additional shift resulting from addition of 3 more bases (Fig. 2-12c). In the context of the exogenous IGF2BP1 BPS, a somewhat higher basal level of proximal splice site usage was reduced by deletion of 6 bases, with deletion of 3 bases producing a modest change (Fig. 2-12d). These data indicate that the BPS plays a significant role in NAGNAG splicing, and confirm that shorter BPS-3' splice site distances can shift splicing toward the proximal NAG.

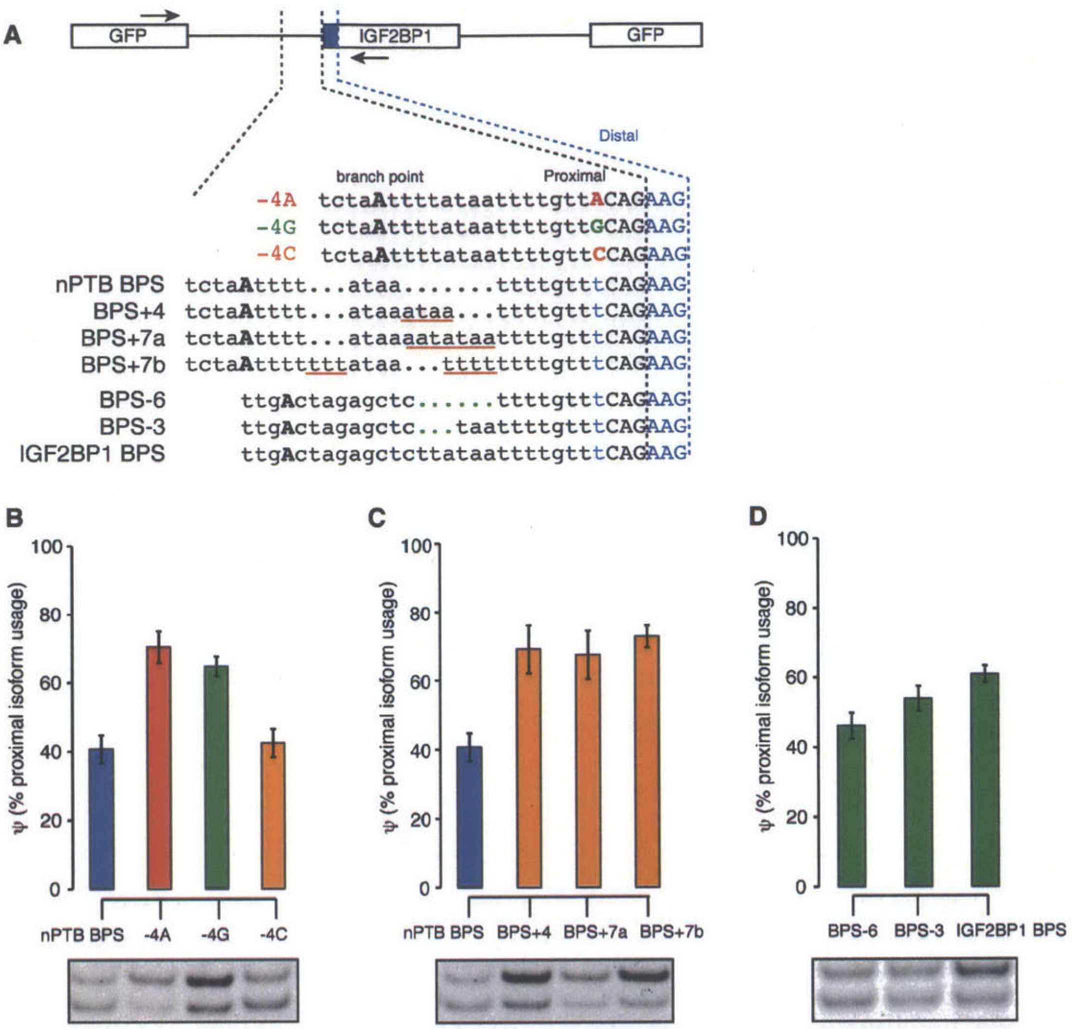


Figure 2-12:

(a) Illustration of NAGNAG minigene constructs, designed to test the roles of the branch point to 3' splice site distance and of the -4 base in NAGNAG splicing. A short segment of intronic sequence spanning the branch point to the 3' splice site of the PTBP2 NAGNAG was cloned upstream of the IGF2BP1 exon. To confirm the importance of a pyrimidine at the -4 position for distal NAG use, the effects of all four nucleotides at the -4 position were tested. The branch point to 3' splice site distance was varied by introducing nucleotides (underlined in orange) in constructs containing the PTBP2 branch point sequence, or by removing nucleotides (indicated by green dots) in constructs containing the IGF2BP1 branch point sequence. Locations of RT-PCR primers are indicated by arrows.

(b) Proximal isoform expression increased dramatically after the introduction of a purine at the -4 position. Splicing was monitored after minigene transfection into HEK293T cells by RT-PCR. Mean and standard deviation of at least 3 independent transfections are shown. A representative gel is shown below (top and bottom bands represent proximal and distal isoforms, respectively).

(c) as in (b), but varying the branch point to 3' splice site distance in the context of the native nPTB branch point sequence. The distance was increased by insertion of 4 or 7 nucleotides of sequence of varying purine/pyrimidine composition as shown in (a).

(d) as in (c), but decreasing the branch point to 3' splice site distance in the context of the exogenous IGF2BP1 BPS by deletion of 3 or 6 bases as shown in (a).

2.3.6 NAGNAGs accelerate protein evolution at exon-exon boundaries

Together, our analyses of proximal/distal major splicing suggested that NAGNAG 3' splice sites afford broad scope for evolutionary tuning of isoform ratios, even in cases where the sequence of the second NAG is constrained by selection on the encoded amino acid. For example, mutations affecting the upstream -3 and -4 bases, the polypyrimidine tract, or the location of the branch site could all potentially modulate the ratio of the two isoforms across a range from predominantly proximal to predominantly distal isoform usage, which might facilitate evolutionary addition and deletion of single codons at 3' splice junctions. A previous study observed reduced frequencies of amino acid substitutions near exon-exon junctions relative to the centers of exons, presumably resulting from purifying selection acting on exonic splicing enhancer motifs (10, 11). By contrast, when we examined exon length changes in alignments of orthologous human and mouse coding exons (Fig. 2-13a), we observed a striking 18.5-fold enrichment for gain/loss of exonic sequence at 3' splice sites relative to flanking positions (Fig. 2-13b; assignment of gaps is illustrated in example alignments in Supplementary Fig. 2-14). No particular enrichment for gain/loss of exonic sequence was observed at the 5' splice site, suggesting that increased addition/deletion of exonic sequence is associated with properties of the 3' splice site itself, rather than being a generic feature of exon boundaries. This pattern was not changed when restricting to constitutive splice junctions (Supplementary Fig. 2-15). A majority of the changes plotted in Fig. 2-13b involved gain/loss of precisely 3 bases, and restricting to changes of exactly this size yielded a similar degree of enrichment at the 3' splice site (Fig. 2-13c).

While gain/loss of exonic sequence is normally attributed to insertions or deletions ("indels") in the genome, the increased frequency of changes at the 3' splice site suggested a prominent role for an alternative mechanism involving genomic substitutions that give rise to 3 base shifts in exon boundaries without insertion or deletion

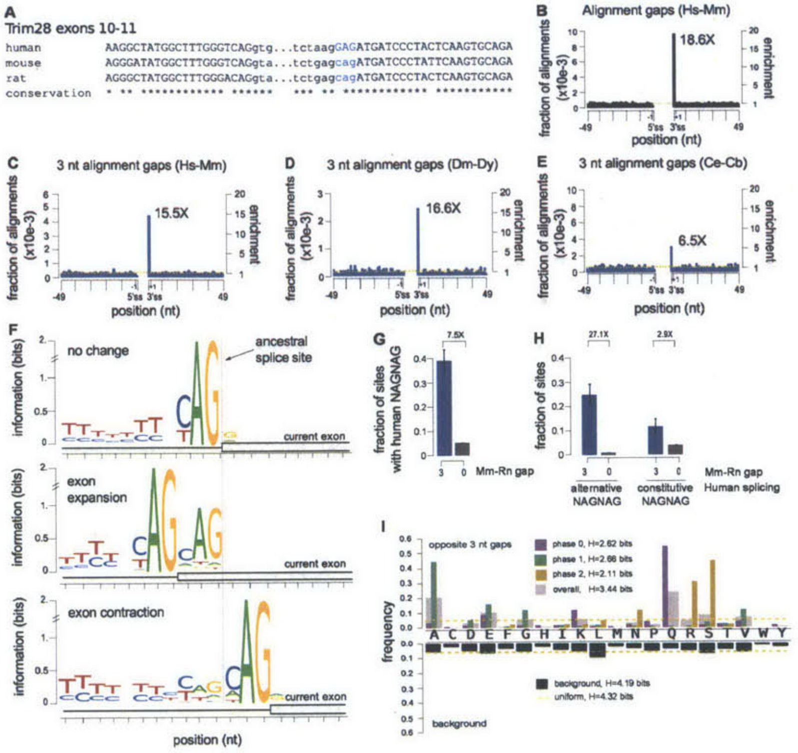


Figure 2-13:

(a) Alignment of portions of exons 10 and 11 of TRIM28 gene from 3 mammals, illustrating a shift in the upstream boundary of exon 11 between human and rodents. Exonic sequence shown in capitals; intronic sequence in lower case.

(b) Gain/loss of exonic sequence between human and mouse occurs preferentially at 3' splice sites (p <10-6, permutation test). The fraction of aligned orthologous human and mouse exons with gaps at each position is shown; the background level (mean fraction across the indicated region excluding the 3' splice site) is shown by the dotted yellow line; the right-hand axis shows enrichment relative to this background.

(c) As in (b), but restricted to gaps of length 3 bp. Preferential occurrence at 3' splice sites was highly significant (p <10-6, permutation test).

(d) Similar to (c), but based on alignments of orthologous *D. melanogaster* and D. yakuba exons. Preferential occurrence at 3' splice sites was highly significant (p <10-6, permutation test).

(e) Similar to (c), but based on alignments of orthologous *C. elegans* and *C. briggsae* exons. Preferential occurrence at 3' splice sites was highly significant (p <10-6, permutation test).

(f) Residual NAG motif at exons whose boundaries changed in the rodent lineage. Orthologous mouse and rat exons were classified as unchanged (top), expanded by 3 nt (middle), or contracted by 3 nt (bottom) based on comparison to an outgroup (human, cow, chicken, or *Xenopus laevis*), aligned to the inferred location of the ancestral 3' splice site (dotted line). Information content of each position is shown relative to a uniform background composition.

(g) Exons whose 3' splice site boundaries differ by 3 nt between rat and mouse are 7.5 times as likely to have a NAGNAG in the human ortholog as exons whose boundaries did not change (p-value for difference <10-24 by Fishers exact test). Error bars indicate the 95% binomial confidence interval.

(h) Rodent exons orthologous to alternatively spliced human NAGNAG exons (left) are much more likely to exhibit 3 bp exon boundary changes than those orthologous to constitutively spliced human NAGNAGs (right) (p-value for difference <10-10 by Fishers exact test). Blue and gray bars in (h) represent subsets of blue and gray bars in (g), respectively. Error bars indicate the 95

(i) Frequency of encoded amino acids that occur opposite gaps at the 3' splice site in alignments of human and mouse exons is plotted above, overall (pink) and separately by the phase of the upstream intron (i.e. the number of bases, if any, in the last incomplete codon of the upstream exon); amino acid frequency at background positions (4 codons downstream of the 3' splice site) are shown below. The Shannon entropy (a measure of randomness) of each amino acid frequency distribution is also shown.

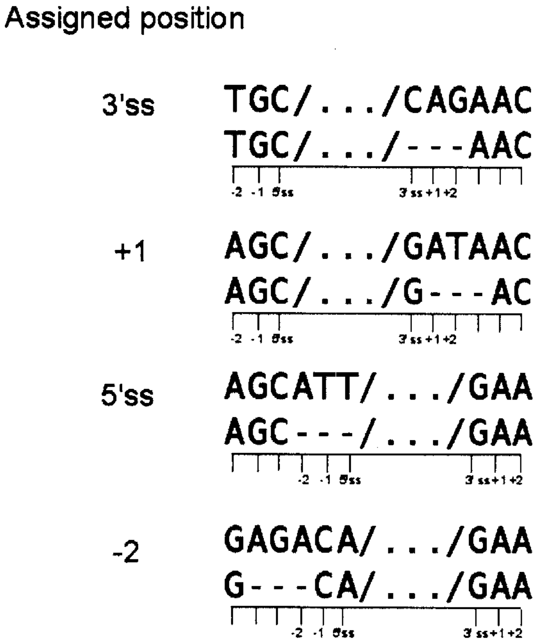


Figure 2-14:
Examples shown in the figure illustrate the numbering system used for assessing gap positions relative to the 5 and 3 splice sites. The splice sites are numbered 0, and gap position is numbered relative to the nearest splice site. Gaps that could not be unambiguously assigned to one splice site were very rare and their inclusion or exclusion did not affect our conclusions.

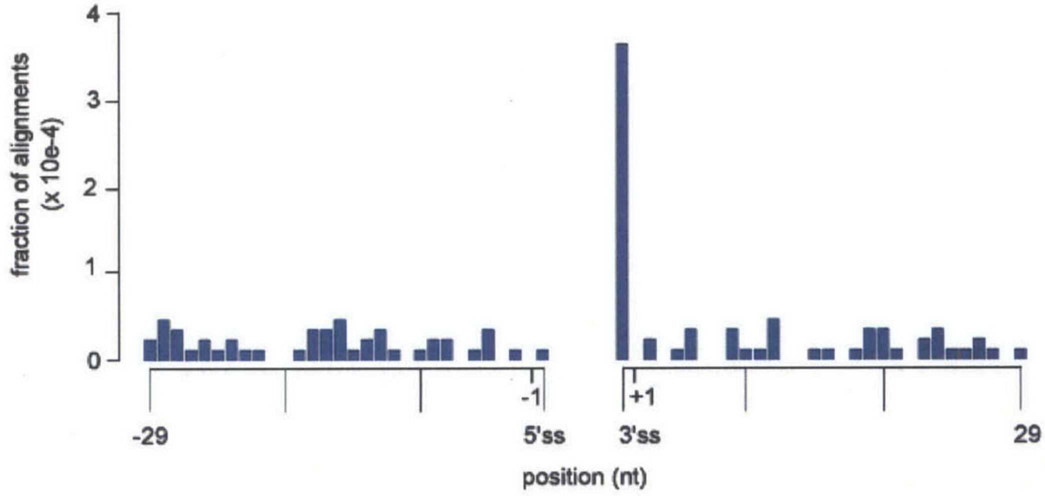


Figure 2-15:
We restricted our analysis in Figure 5B to exons containing NAGNAGs which were constitutively spliced (ψ <5% or ψ >95% across all tissues) in both human and mouse. We observed qualitatively similar patterns of specific enrichment of gaps at the 3 splice site, suggesting that the signal observed in Figure 5B was not due to unannotated alternative splicing of NAGNAGs.

of genomic DNA. For example, creation of a NAG motif immediately upstream of a 3' splice site NAG by mutation would be expected to commonly shift splicing upstream by 3 bases (resulting in exonization of 3 bases of intron) or generate an alternatively spliced NAGNAG that could subsequently lose splicing at the downstream NAG through mutation. Alternatively, a mutation creating an immediately downstream NAG — or a mutation that weakened the upstream NAG relative to a pre-existing downstream NAG — could result in either alternative splicing or loss of 3 bases of exonic sequence. Both of these scenarios could arise frequently by single base substitutions, which occur at a rate that is an order of magnitude higher than the rate of genomic indels (37).

Consistent with this substitution/exaptation model and the finding that many NAGNAGs are alternatively spliced in the *Drosophila* lineage, we observed similar enrichment for gain/loss of 3 bp of exonic sequence at the 3' splice site when comparing orthologous *D. melanogaster* and *D. yakuba* coding exons (Fig. 2-13d). Notably, the enrichment of 3 base gaps at the 3' splice site was three-fold weaker in comparisons of *Caenorhabditis elegans* and *C. briggsae* exons (Fig. 2-13e). NAGNAG alternative splicing is reported to occur rarely in nematodes due to a highly constrained 3' splice site motif (15). We confirmed the rarity of NAGNAG alternative splicing in *C. elegans* using RNA-Seq data from 14 developmental time points and conditions generated by the modENCODE consortium. Enumerating NAGNAGs in introns of *C. elegans* coding genes, we detected alternative splicing (both isoforms expressed at 5% in at least one developmental time point) for only 18% of NAGNAGs with favorable pyrimidine bases at both -3 positions based on RNA-Seq read depths slightly below those used in human. By contrast, 50-85% of human, mouse, and *Drosophila* YAGYAGs were detected as alternatively spliced, suggesting that NAGNAG alternative splicing is substantially rarer in worms than in other metazoans. This decrease in abundance mirrors the three-fold weaker enrichment of 3 base gaps at 3' splice sites observed in worms (Fig. 5e).

Sequence motif analyses further implicated NAGNAG splicing in the exon length changes observed at exon boundaries. Classifying the borders of orthologous mouse and rat exons as unchanged, expanded, or contracted (comparing to human, cow, chicken, and/or *Xenopus laevis* as outgroups), we observed evidence of residual NAG-NAG motifs in exons with altered boundaries (Fig. 2-13f). Specifically, exons expanded in mouse or rat exhibited a consensus NAG at exonic positions +1 to +3, and contracted exons exhibited a consensus NAG at intronic positions -6 to -4. The presence of this residual sequence motif provides further evidence that a substantial portion of exon length changes observed between orthologous mammalian exons derive from splicing-mediated shifts in exon boundaries rather than genomic indels. Likely because of subsequent selection to optimize the polypyrimidine tract, the residual NAG signal was weaker for contracted than for expanded exons.

Consistent with these findings, we observed a strong association between gain/loss of 3 bases in the rodent lineage and presence of a NAGNAG in orthologous human exons. Exons that expanded or contracted in rodents were 7.5-fold more likely to have a NAGNAG in the orthologous human exon than were exons with unchanged boundaries (Fig. 2-13g). Further subdividing these exons according to the splicing pattern of the NAGNAG in human, we observed that rodent exons orthologous to alternatively spliced human NAGNAGs were ~9 times more likely to have gained/lost exonic sequence than those orthologous to constitutively spliced human NAGNAGs (Fig. 2-13h). These analyses implicate NAGNAG alternative splicing as a very common evolutionary intermediate in the gain and loss of single codons from exons.

This model, where frequent alternative splicing at the 3' splice site leads to gain/loss of exonic sequence, is expected to play out very differently at 5' splice sites. Competing 5' splice sites are most frequently 4 bp apart (22), resulting in a frame-shift which is likely to render one of the protein products non-functional and potentially target the mRNA for nonsense-mediated decay. Although common, competing 5' splice sites separated by 4 bp are therefore unlikely to lead to accelerated exon length changes and we observed no significant increase in exon length changes at the 5' splice site (Fig. 2-13a).

2.3.7 NAGNAG-accelerated protein evolution is highly biased

Most three base changes to mRNAs minimally affect RNA-level properties such as message stability. However, insertion/deletion of a single amino acid residue can have a profound impact on protein function. For example, deletion of a single codon can alter protein degradation, subcellular localization, DNA binding affinity or other protein properties (38, 39), can cause diseases including cystic fibrosis and Tay-Sachs disease (40, 41), and can even rescue a disease-related phenotype (42). Insertion or deletion of a codon in a protein structural motif with a periodic hydrogen bonded structure such as a beta sheet or coiled coil domain might have a disproportionate effect on protein structure by altering the hydrogen bonding of a large number of downstream residues. Considering the spectrum of codons that occurred opposite 3 base gaps at the beginnings of exons (corresponding to the peak in Fig. 2-13c), we observed a highly non-random distribution that strongly favored glutamine, alanine, glutamate and serine and disfavored most other residues including cysteine, phenylalanine and histidine relative to the background. Distinct and far stronger biases were observed when grouping introns by "phase" (position relative to the reading frame) (15). These biases occurred in a pattern consistent with frequent origin via exaptation of NAGNAGs (Fig. 2-13i), whose codon-level effects are largely determined by intron phase. For example, glutamine (mostly coded by CAG) was the most commonly added residue at the end of "phase 0" introns, for which the first 3 bases of the downstream exon form a codon. Serine (mostly AGY) and arginine (mostly AGR) were the most commonly added residues at the boundaries of phase 2 introns, for which the AG of an added NAG would form the first two bases of a codon. These biases contributed to a strong enrichment observed for gain/loss of predicted phosphorylation sites at 3' splice sites (Supplementary Fig. 2-16). Together, the analyses in Fig. 2-13 demonstrate that gain and loss of residues along proteins occurs in a strongly biased manner, with a highly accelerated rate and biased codon spectrum at the beginnings of exons that is likely driven by genomic substitutions that alter

NAGNAG motifs or their splicing patterns. These observations suggest that the evolutionary trajectories of proteins in metazoans are shaped to a surprising extent by the specific locations and phases of introns that interrupt their coding sequences.

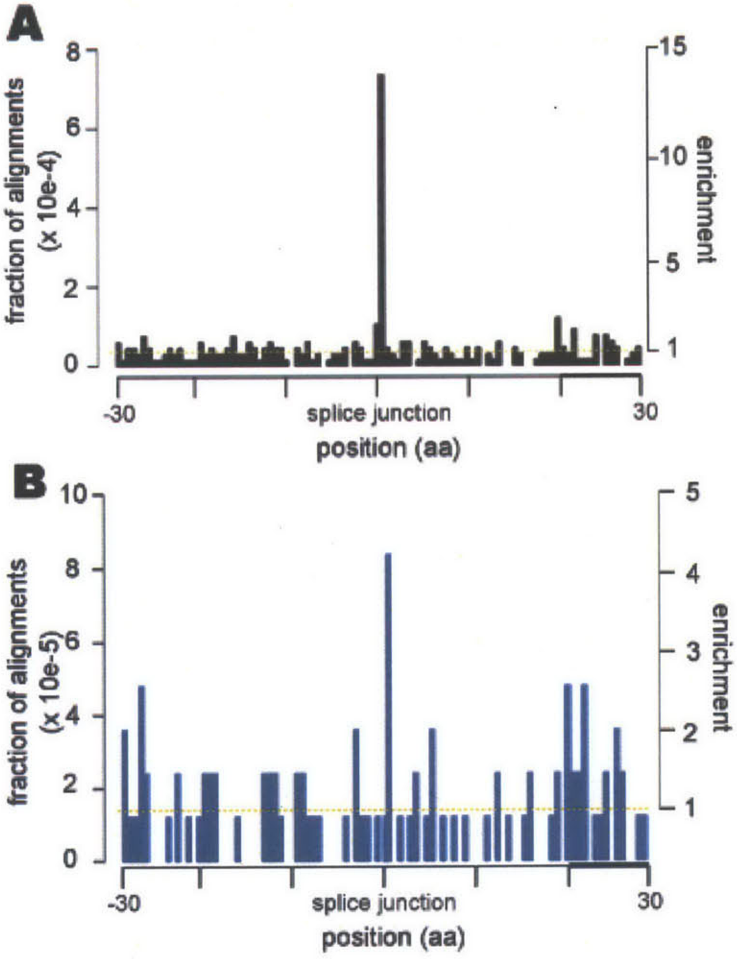


Figure 2-16:
The distribution of alignment gaps containing one or more predicted phosphorylation sites is shown for

(a) all gaps

(b) gaps of three bases

2.4 Methods

2.4.1 Accession codes

Mapped sequence reads from the human and mouse RNA-Seq experiments are located in NCBIs GEO database (accession number GSE30017). The complete Body Map 2.0 sequence data are in the ENA archive with accession number ERP000546 (available at http://www.ebi.ac.uk/ena/data/view/ERP000546). These data are also accessible from ArrayExpress (ArrayExpress accession: E-MTAB-513). The Body Map 2.0 data were generated by the Expression Applications R&D group at Illumina using the standard (polyA-selected) Illumina RNA-Seq protocol from total RNA obtained commercially (Ambion) using the HiSeq 2000 system. We downloaded *D. melanogaster* (Developmental Stage Timecourse Transcriptional Profiling with RNA-Seq) and *C. elegans* (Global Identification of Transcribed Regions of the *C. elegans* Genome) RNA-Seq data from the modMINE (http://intermine.modencode.org/) website of the modENCODE consortium. For the *C. elegans* data, we restricted to 36 bp reads for consistency with other analyses.

2.4.2 Splicing events

We used the set of splicing events from (1) to identify skipped exons, alternative 3' splice sites (>3 nt apart), alternative 5' splice sites, and mutually exclusive exons in the human (GRCh37, or hg19) and mouse (NCBIM37, or mm9) genomes (Fig. 2-1c). We enumerated all possible NAGNAGs in the human genome by finding all 3' splice sites in these alternative splicing events and the Ensembl (43) and UCSC (44) annotation databases and then searching for NAGNAG motifs. We classified splice junctions as constitutive if they did not overlap any alternative splicing event present in the databases described above.

2.4.3 Mouse tissues and RNA-Seq library preparation

Mouse tissues from a 10-week-old male were extracted immediately after death and stored in RNAlater per the manufacturers instructions (Ambion). Tissue was lysed in Trizol and RNA was extracted with Qiagen miRNeasy mini columns. Using 5 ug of total RNA, we performed polyA selection and prepared strand-specific libraries for Illumina sequencing following the strand-specific dUTP protocol (45) and using the SPRIworks Fragment library system (Beckman Coulter). We obtained final insert sizes of approximately 160 bp. We sequenced these libraries using the Illumina HiSeq 2000 and the GAIIx machines.

2.4.4 RNA-Seq read analysis

For each NAGNAG, we extracted the sequence flanking the proximal and distal 3' splice sites and used Bowtie (46) version 0.12.7 to map reads to these two sequences. We required that short reads have at least 6 nt on either side of the splice junction (an overhang of 6 nt), and furthermore that there be no mismatches within the overhang region. In order to eliminate errors in read mapping due to non-unique splice junctions, we restricted the set of NAGNAGs enumerated across the genome to the subset of NAGNAGs for which all 36-mers mapping to either splice site did not map to the genome or any other splice junction (we used 36-mers because they were the shortest reads analyzed in our experiments). We then computed ψ values as (number of reads mapping to the proximal splice junction) / (number of reads mapping to either the proximal or distal splice junction). For all bioinformatics analyses, we only analyzed the subset of tissues for which a particular NAGNAG had a total of at least 10 reads in order to control for variation in junction coverage due to gene expression differences. We experimented with requiring different levels of junction coverage (10-100 reads per NAGNAG) and confirmed that our conclusions were insensitive to the chosen cutoff. We identified alternatively spliced events as those for which both isoforms were expressed at $\geq 5\%$ in at least one sample (restricting to tissues for which a particular NAGNAG had ≥ 10 reads), and identified regulated events as those with $p \leq 0.01$ by the proportion or z-test (prop.test in R [http://www.R-project.org/]). As described in the text, when computing the fraction of regulated NAGNAGs, we only considered NAGNAGs which were alternative spliced by these criteria (both isoforms expressed at $\geq 5\%$ in at least one sample).

For Fig. 2-1c, we re-mapped the reads using TopHat (47) version 1.1.4 and restricted to uniquely mapping reads with an overhang of 6 nt and no mismatches in the overhang region. Using only reads mapping to the two 3' (skipped exons, NAGNAGs, alternative 3' splice sites, and mutually exclusive exons) or 5' (alternative 5' splice sites) splice sites of each event, we computed ψ values and identified alternative spliced and regulated events as described above.

2.4.5 False discovery rates

We estimated false-discovery rates as the fraction of events which were differentially expressed between technical (human) or biological (mouse) replicates identified using the procedure described above for regulated events. Briefly, for each tissue and pair of replicates, we restricted to the set of NAGNAGs which were alternatively spliced in at least one of the replicates and computed the fraction of these NAGNAGs which were differentially expressed with $p \leq 0.01$ between the replicates. We estimated mean FDRs for human (4.4%) and mouse (1.1%) by taking a weighted average over tissues, where we weighted the FDR computed for each tissue by the number of alternatively spliced NAGNAGs analyzed for that tissue. The fraction of strongly regulated NAGNAGs increased essentially linearly with the number of tissues considered for both human and mouse (Supplementary Fig. 2-3). We expect this trend to continue as the number of mouse tissues increases, as it does for the human data. Accordingly extrapolating the mouse data to 16 tissues with a linear fit and subtracting the mean FDR of 1.1%, we estimated that at least 12% of alternatively spliced mouse NAGNAGs are strongly regulated, providing a lower bound on the fraction of strongly regulated

NAGNAGs in mammals. We used the human data to compute a corresponding upper bound of 37% by subtracting the mean FDR of 4.4% from the observed fraction of strongly regulated NAGNAGs (Supplementary Fig. 2-3).

2.4.6 Boltzmann model

For each NAGNAG event, the probabilities of using the proximal and distal splice sites are proportional to $Q \cdot e^{Bi(s_p)}$ and $e^{Bi(s_d)}$, where s_p and s_d are the proximal and distal splice site scores. The probability of using the proximal splice site is therefore $[1+Q \cdot e^{-B(s_p-s_d)}]^{-1}$. We fit the parameters Q and B as follows: For each NAGNAG, we computed the mean ψ (averaging over tissues). We then binned NAGNAGs according their splice site score differences, using a bin size of 3.25 bits and a bin increment of 0.5 bits, and computed the median ψ for each bin. We fit a straight line to the six bins flanking the point where $\psi = 50\%$ and estimated the parameters as $Q = 0.55$ and $B = 0.58$ based on a first-order Taylor expansion.

2.4.7 Ortholog identification and sequence conservation analysis

We performed a whole-genome alignment of human and mouse using Mercator:

http://www.biostat.wisc.edu/ cdewey/mercator/

and FSA (48), and identified orthologous NAGNAGs as those for which both the 5' splice site and competing 3' splice sites were orthologous according to the corresponding sequence alignment. For the *Drosophila* analysis, we used a previously described *D. melanogaster-D. yakuba* whole-genome alignment (49). For all sequence conservation analyses, we downloadedded phastCons scores (50) from the UCSC annotation databases (44). We used phastCons46 (placental mammals) for human, phastCons30way (placental mammals) for mouse, and phastConst15way for

D. melanogaster.

2.4.8 Minigene assays

Segments of PTBP2 intronic sequence containing the NAGNAG were cloned into a modular splicing reporter (51) upstream of the IGF2BP1 exon using SacI and XhoI restriction enzyme sites. Forward and reverse oligonucleotides (below) were mixed in equimolar ratios, annealed, and double-digested with SacI and XhoI, or in some cases the oligonucleotides were ordered with desired restriction site overhangs, and ligated into the pGM4G9 minigene. For constructs analyzing the effects of distance to the native PTBP2 branch point, the vector (IGF2BP1) branch point sequence was first mutated by site-directed mutagenesis (TCATTGA was deleted, immediately upstream from the SacI restriction site) prior to insertion of the PTBP2 3' splice site. All minigene reporters (0.5 ug) were transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen). RNA was isolated 18-24 hours post-transfection with RNeasy Mini Kits (Qiagen). RT-PCR was performed with a fluorescent primer (NAGNAG_Forward: 5' 6FAM- TCTTCAAGTCCGCCATGC and NAGNAG_reverse: 5' AGTCAGGTGTTTCGGGTGGT). The proximal (63 nucleotides) and distal (60 nucleotides) isoforms were resolved on a 10TBE gel and detected with a Typhoon 9000 scanner (GE Healthcare). Proximal and distal isoforms were quantified with ImageJ software.

Primers:

PTB2_For: cagtgtctaattttataattttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTc

PTB2_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaattataaaattagacactgagct

BPS+4_For: cagtgtctaattttataaataattttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTc

BPS+4_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaattatttataaaattagacactgagct

BPS+7a_For: cagtgtctaattttataaataaatattttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTc

BPS+7a_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaatatttatttataaaattagacact gagct

BPS+7b_For: cagtgtctaatttttttataattttttttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTC

BPS+7b_Rev: TcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaaaaaattataaaaaaaattagacactgagct

-4A_For: cagtgtctaattttataattttgttacagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTC

-4_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgtaacaaaattataaaattagacactgagct

-4G_For: cagtgtctaattttataattttgttgcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTc

-4G_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgcaacaaaattataaaattagacactgagct

-4C_For: cagtgtctaattttataattttgttccagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTc

-4C_Rev: tcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctggaacaaaattataaaattagacactgagct

IGF2BP1BPS_For: gcgagctcttataattttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTctcgagcgg

IGF2BP1BPS_Rev: ccgctcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaattataagagctcgc

BPS-3_For: gcgagctctaattttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTctcgagcgg

BPS-3_Rev: ccgctcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaattagagctcgc

BPS-6_For: gcgagctcttttgtttcagAAGATTGCACCACCCGAAACACCTGACTCCAAAGTTCGTATGGTTctcgagcgg

BPS-6_Rev: ccgctcgagAACCATACGAACTTTGGAGTCAGGTGTTTCGGGTGGTGCAATCTTctgaaacaaaagagctcgc

2.4.9 Evolutionary analysis

We restricted all analyses to "singleton orthologs," genes without paralogs and with unambiguous orthology assignments in all species considered for each analysis, annotated in Ensembl (43) and queried with PyCogent (52). For each gene, we required that the longest annotated coding sequence have the same number of exons in all species, and performed all subsequent analyses using this longest coding sequence. For each longest coding sequence, we extracted pairs of consecutive exons, concatenated them, and then aligned them to their corresponding orthologous sequences using FSA (48). In order to control for alignment error, we required that alignment sequence identity be greater than 70% and that the total inserted sequence be no longer than 20% of the length of the shortest exon. Furthermore, if gaps in an alignment could be moved to lie at exon-exon boundaries rather than within exonic sequence while preserving the alignment quality (number of exact matches), then we modified the alignment accordingly, as FSA is unaware of exon structures. This modification affected only a small fraction of alignments, and our results in Fig. 2-13 are unchanged without this modification.

We classified orthologous mouse and rat exons as unchanged, expanded, or contracted based on comparison with an outgroup (human, cow, chicken, *Xenopus laevis*, or *Danio rerio*, in that order, until an informative comparison was found). For each exon in each class, we extracted the corresponding intronic sequence and created a sequence logo using WebLogo (Fig. 2-13f-h) (53).

For analyses of amino acid sequences in Fig. 2-13i, we compared the amino acids gained or lost in alignments with 3 nt gaps at the 3' splice site. If the next gain/loss was a single amino acid (for example, if the human peptide was SR and the mouse peptide was R), then we counted only the single amino acid which was inserted (S); if the gain/loss was two amino acids (for example, if the human peptide was SR and the mouse peptide was K), then we counted both amino acids which were inserted (SR).

For Supplementary Fig. 2-16, we used a BioPerl module (54) to query Scansite (55) to predict phosphorylation sites (medium stringency) in the translated longest annotated coding sequence, and plotted the location of predicted phosphorylation sites which were gained/lost in human and mouse.

Unless otherwise described, all plots in Fig. 2-13 were created with matplotlib (http://matplotlib.sourceforge.net/).

2.5 Acknowledgements

We thank Caitlin Russell and Daniel Treacy for expert technical assistance, the staff at the MIT Biomicro Center for help with Illumina sequencing, and Andy Berglund, Wendy Gilbert, Richard Padgett, Phillip Sharp and members of the Burge lab for helpful discussions and comments on this manuscript. This work was supported by a Damon Runyon Cancer Research Foundation Postdoctoral Fellowship (RKB, DRG 2032-09), by a major equipment grant from the National Science Foundation (no. 0821391) and by grants from the US National Institutes of Health (CBB).

2.6 Author Contributions

Project conception: RKB and CBB. Bioinformatic analyses of regulation and conservation (Fig. 1-3): RKB. Splicing reporter assays: NL and RKB. Bioinformatic analyses related to exon evolution (Fig. 5): JM. Figure preparation: RKB, NL and JM. Manuscript preparation: RKB and CBB.

2.7 Competing Financial Interests

The authors declare no competing financial interests.

2.8 References

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

[2] Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, et al. (2010) The developmental transcriptome of Drosophila melanogaster. Nature 471: 473-479.

[3] Cascino I, Papoff G, De Maria R, Testi R, Ruberti G (1996) Fas/Apo-1 (CD95) receptor lacking the intracytoplasmic signaling domain protects tumor cells from

Fas-mediated apoptosis. J Immunol 156: 13-17.

[4] Christofk HR, Vander Heiden MG, Wu N, Asara JM, Cantley LC (2008) Pyruvate kinase M2 is a phosphotyrosine-binding protein. Nature 452: 181-186.

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

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

[7] Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007-1013.

[8] 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.

[9] Zhang XH, Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18: 1241-1250.

[10] Fairbrother WG, Holste D, Burge CB, Sharp PA (2004) Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2: E268.

[11] Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD (2007) Splicing and the evolution of proteins in mammals. PLoS Biol 5: e14.

[12] Boise LH, Gonzalez-Garcia M, Postema CE, Ding L, Lindsten T, et al. (1993) bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death. Cell 74: 597-608.

[13] Wang Z, Xiao X, Van Nostrand E, Burge CB (2006) General and specific functions of exonic splicing silencers in splicing control. Mol Cell 23: 61-70.

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

[15] Hiller M, Huse K, Szafranski K, Jahn N, Hampe J, et al. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. Nat Genet 36: 1255-1257.

[16] Iida K, Shionyu M, Suso Y (2008) Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals. Mol Biol Evol 25: 709-718.

[17] Schindler S, Szafranski K, Hiller M, Ali GS, Palusa SG, et al. (2008) Alternative splicing at NAGNAG acceptors in Arabidopsis thaliana SR and SR-related protein-coding genes. BMC Genomics 9: 159.

[18] Daines B, Wang H, Wang L, Li Y, Han Y, et al. (2011) The Drosophila melanogaster transcriptome by paired-end RNA sequencing. Genome Res 21: 315-324.

[19] Tsai KW, Lin WC (2006) Quantitative analysis of wobble splicing indicates that it is not tissue specific. Genomics 88: 855-864.

[20] Tadokoro K, Yamazaki-Inoue M, Tachibana M, Fujishiro M, Nagao K, et al. (2005) Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. Journal of human genetics 50: 382-394.

[21] Chern TM, van Nimwegen E, Kai C, Kawai J, Carninci P, et al. (2006) A simple physical model predicts small exon length variations. PLoS Genet 2: e45.

[22] Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ (2006) Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA 12: 2047-2056.

[23] Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7: 1009-1015.

[24] Hiller M, Szafranski K, Huse K, Backofen R, Platzer M (2008) Selection against tandem splice sites affecting structured protein regions. BMC Evol Biol 8: 89.

[25] Hiller M, Szafranski K, Sinha R, Huse K, Nikolajewa S, et al. (2008) Assessing the fraction of short-distance tandem splice sites under purifying selection. RNA 14: 616-629.

[26] Akerman M, Mandel-Gutfreund Y (2006) Alternative splicing regulation at tandem 3' splice sites. Nucleic Acids Res 34: 23-31.

[27] Sinha R, Nikolajewa S, Szafranski K, Hiller M, Jahn N, et al. (2009) Accurate prediction of NAGNAG alternative splicing. Nucleic Acids Res 37: 3569-3579.

[28] Smith CW, Porro EB, Patton JG, Nadal-Ginard B (1989) Scanning from an independently specified branch point defines the 3' splice site of mammalian introns. Nature 342: 243-247.

[29] Smith CW, Chu TT, Nadal-Ginard B (1993) Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. Mol Cell Biol 13: 4939-4952.

[30] Tsai KW, Tarn WY, Lin WC (2007) Wobble splicing reveals the role of the branch point sequence-to-NAGNAG region in 3' tandem splice site selection. Mol Cell Biol 27: 5835-5848.

[31] Tsai KW, Chan WC, Hsu CN, Lin WC (2010) Sequence features involved in the mechanism of 3' splice junction wobbling. BMC Mol Biol 11: 34.

[32] Akerman M, Mandel-Gutfreund Y (2007) Does distance matter? Variations in alternative 3' splicing regulation. Nucleic Acids Res 35: 5487-5498.

[33] Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, et al. (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science 309: 2054-2057.

[34] Baek D, Green P (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proc Natl Acad Sci U S A 102: 12813-12818.

[35] Corrionero A, Raker VA, Izquierdo JM, Valcarcel J (2011) Strict 3' splice site sequence requirements for U2 snRNP recruitment after U2AF binding underlie a genetic defect leading to autoimmune disease. RNA 17: 401-411.

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

[37] Silva JC, Kondrashov AS (2002) Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends Genet 18: 544-547.

[38] Vogan KJ, Underhill DA, Gros P (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. Mol Cell Biol 16: 6677-6686.

[39] Tsai KW, Tseng HC, Lin WC (2008) Two wobble-splicing events affect ING4 protein subnuclear localization and degradation. Exp Cell Res 314: 3130-3141.

[40] Consortium TCFGA (1990) Worldwide survey of the delta F508 mutation–report from the cystic fibrosis genetic analysis consortium. Am J Hum Genet 47: 354-359.

[41] Navon R, Proia RL (1991) Tay-Sachs disease in Moroccan Jews: deletion of a phenylalanine in the alpha-subunit of beta-hexosaminidase. Am J Hum Genet 48: 412-419.

[42] Hinzpeter A, Aissat A, Sondo E, Costa C, Arous N, et al. (2010) Alternative splicing at a NAGNAG acceptor site as a novel phenotype modifier. PLoS Genet 6.

[43] Flicek P, Amode MR, Barrell D, Beal K, Brent S, et al. (2011) Ensembl 2011. Nucleic Acids Res 39: D800-806.

[44] Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39: D876-882.

[45] Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37: e123.

[46] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

[47] Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25: 1105-1111.

[48] Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5: e1000392.

[49] Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, et al. (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. PLoS biology 8: e1000343.

[50] 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.

[51] 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. Nature structural & molecular biology 16: 1094-1100.

[52] Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, et al. (2007) PyCogent: a toolkit for making sense from sequence. Genome Biol 8: R171.

[53] Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188-1190.

[54] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.

[55] Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31: 3635-3641.

[56] Reed R, Maniatis T (1985) Intron sequences involved in lariat formation during pre-mRNA splicing. Cell 41: 95-105.

[57] Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, et al. (2006) A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. Genome Biol 7: R1.

Chapter 3

Evolutionary dynamics of gene and isoform regulation in mammalian tissues

As published in Science:

Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B. Burge. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. Science. 21 December 2012: 338 (6114), 1593-1599. doi:10.1126/science.1228186

Author Contributions:

J.M. and C.B.B. designed the study and wrote the manuscript. J.M. collected tissue samples, extracted RNA, conducted computational analyses and prepared figures. C.R. prepared RNA-Seq libraries and developed protocols. P.C. contributed computational analyses.

3.1 Abstract

Most mammalian genes produce multiple distinct mRNAs through alternative splicing, but the extent of splicing conservation is not clear. To assess tissue-specific transcriptome variation across mammals, we sequenced cDNA from 9 tissues from 4 mammals and one bird in biological triplicate, at unprecedented depth. We find that while tissue-specific gene expression programs are largely conserved, alternative splicing is well conserved in only a subset of tissues and is frequently lineage-specific. Thousands of novel, lineage-specific and conserved alternative exons were identified; widely conserved alternative exons had signatures of binding by MBNL, PTB, RBFOX, STAR and TIA family splicing factors, implicating them as ancestral mammalian splicing regulators. Our data also indicate that alternative splicing often alters protein phosphorylatability, delimiting the scope of kinase signaling.

3.2 Results and Discussion

Alternative pre-mRNA processing can result in mRNA isoforms that encode distinct protein products, or may differ exclusively in untranslated regions, potentially affecting mRNA stability, localization or translation (1). It can also produce nonfunctional mRNAs that are targets of nonsense-mediated mRNA decay (NMD), serving to control gene expression (2). Most human alternative splicing is tissue-regulated (3, 4), but the extent to which tissue-specific splicing patterns are conserved across mammalian species has not yet been comprehensively studied.

To address outstanding questions about the conservation and functional significance of tissue-specific splicing, we conducted transcriptome sequencing (RNASeq) analysis of 9 tissues from 5 vertebrates, consisting of 4 mammals and one bird. The species, chosen based on the quality of their genomes (all high coverage finished or draft genomes) and their evolutionary relationships, include the rodents mouse and rat, the rhesus macaque, a non-rodent/non-primate boroeutherian, cow, and chicken as an outgroup. These relationships allow for the evaluation of transcriptome changes between species with divergence times ranging from <30 million years to >300 million years (Fig. 3-2A). Our sequencing strategy used paired-end short or long read sequencing of polyA-selected RNA. In total, we generated over 16 billion reads (>8 billion read pairs) totaling over 1 trillion bases (3, 5). The data were mapped to the relevant genomes using software that can identify novel splice junctions and isoforms (6).

To assess coverage of genes, these de novo annotations were compared with existing Ensembl annotations. We detected over 211,000 (97%) of the ~217,000 annotated exons in mouse, and similarly high fractions in most other species, including more than 99% of exons in chicken. We estimated that nearly all multi-exonic genes in the species studied are alternatively spliced (Fig. 3-1) (3).

3.2.1 All tissues have conserved expression signatures

To explore the expression relationships between the samples, we used hierarchical clustering based on Jensen-Shannon divergence (JSD) distances between the expression of orthologous genes. A clear pattern emerged in the resulting dendrogram (Fig. 3-2A). Samples of the same tissue from different individuals of the same species were invariably the most similar, followed by samples from the same tissue from other species, with few exceptions. This tissue-dominated clustering pattern indicates that most tissues possess a conserved gene expression signature and suggests that conserved gene expression differences underlie tissue identity in mammals (5, 7). Since gene expression varies by cell type, some observed differences could reflect changes in cell type composition. The most notable exceptions to tissue dominance were that some chicken muscle samples clustered with chicken heart rather than mammalian muscle, and that chicken lung, colon and spleen samples clustered with each other rather than with their mammalian counterparts. These exceptions suggest that species-specific divergence in expression begins to exceed conserved tissue-specific differences

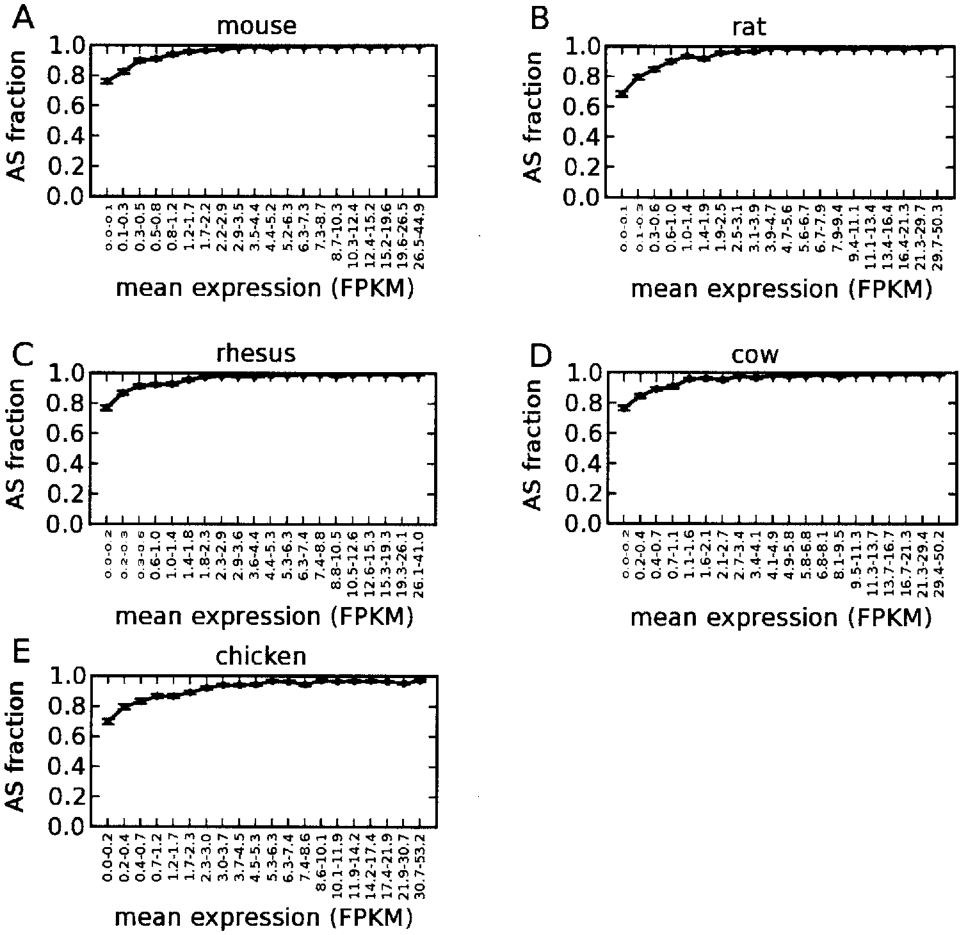


Figure 3-1:
Genes are binned by average expression across tissues and the fraction of genes in each bin that are alternatively spliced ± binomial standard deviation are plotted. The AS fraction is determined as in (3), with a stricter requirement of 5 unique reads supporting both junctions.

(a) AS fraction in mouse.

(b) AS fraction in rat.

(c) AS fraction in rhesus.

(d) AS fraction in cow.

(e) AS fraction in chicken.

at a phylogenetic distance of 300 MY, corresponding to the split between birds and mammals.

3.2.2 Some tissues have conserved splicing signatures

To understand the splicing relationships between the samples, we performed an analogous clustering analysis using the percent spliced in (PSI or ψ) values of the exons that were alternatively spliced in all species containing them. PSI values, the fraction of a genes mRNAs that contain the exon, were calculated from transcript abundance measurements (8) (Fig. 3-4), and were clustered using the same metric (Fig. 3-2B). Samples of the same tissue from individuals of the same species almost invariably clustered together. However, at larger distances a more complex pattern emerged. Tissue-dominated clustering was observed for brain and for the combination of heart and muscle, indicating that these tissues have strong splicing signatures conserved between mammals and chicken, and the rodent testis samples also clustered together. By contrast, samples from the remaining tissues (colon, kidney, liver, lung, spleen) exhibited species-dominated clustering, forming distinct clusters by species rather than by tissue. This trend suggests that alternative splicing patterns specific to this latter group of tissues are less pronounced or less conserved than those of brain, testis, heart and muscle (Fig. 3-5). The greater prominence of species-dominated clustering of PSI values suggests that exon splicing is more often affected by lineage-specific changes in cis-regulatory elements (9) and/or trans-acting factors than is gene expression (6). Lineage-specific changes in splicing factor expression may have contributed to the tendency of splicing patterns to cluster by species more often than by tissue (Fig. 3-6).

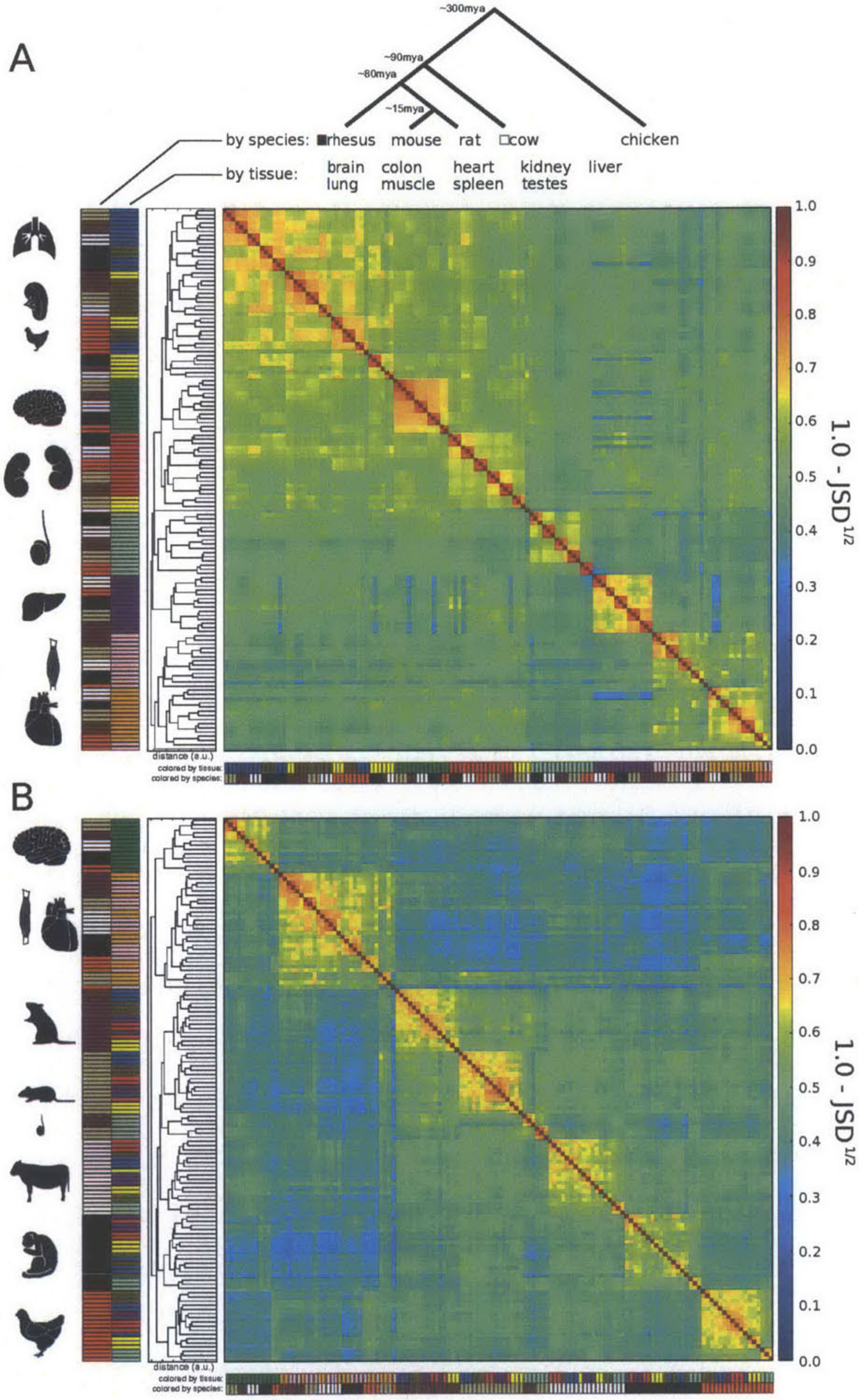


Figure 3-2:

(a) Clustering of samples based on expression values (FPKM) of singleton orthologous genes present in all 5 species (n=7713). Average linkage hierarchical clustering was used with distance between samples measured by the square root of the Jensen-Shannon Divergence (JSD) between the vectors of expression values.

(b) Clustering of samples based on PSI values of exons in singleton orthologous genes conserved to chicken, with alternative splicing detected in all individuals analyzed (n=489). Clustered as in (A). When the set of genes used in this analysis was clustered by gene expression rather than PSI values, tissue-dominated clustering was observed, as in (A) (fig. S15).

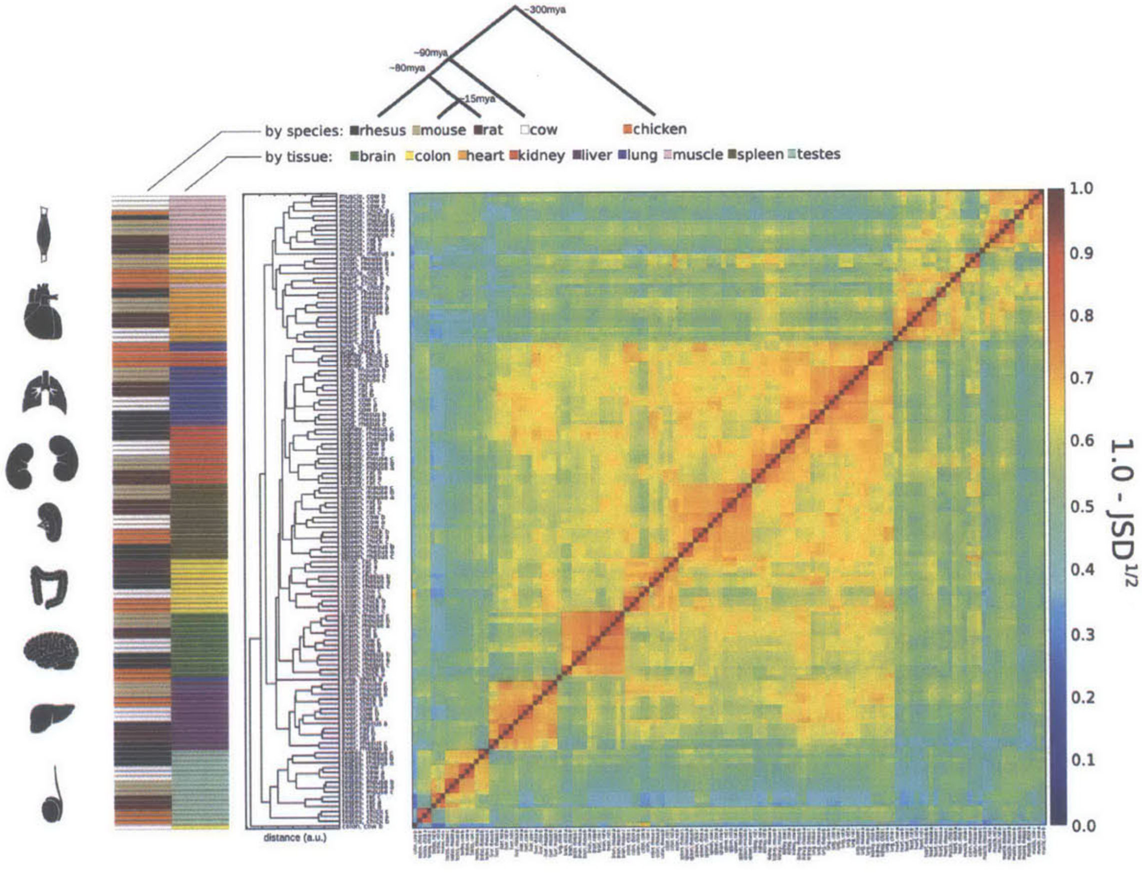


Figure 3-3:
Hierarchical clustering by JSD of gene expression (FPKM) of all genes containing an exon analyzed in Fig. 3-2B. To determine if the species-dominated clustering of skipped exon PSI values was due to changes in the expression of the subset of genes containing exons that are alternative in all species, we repeated the analysis in Fig. 3-2A, restricting to the set of genes containing an exon that was alternative in all species.

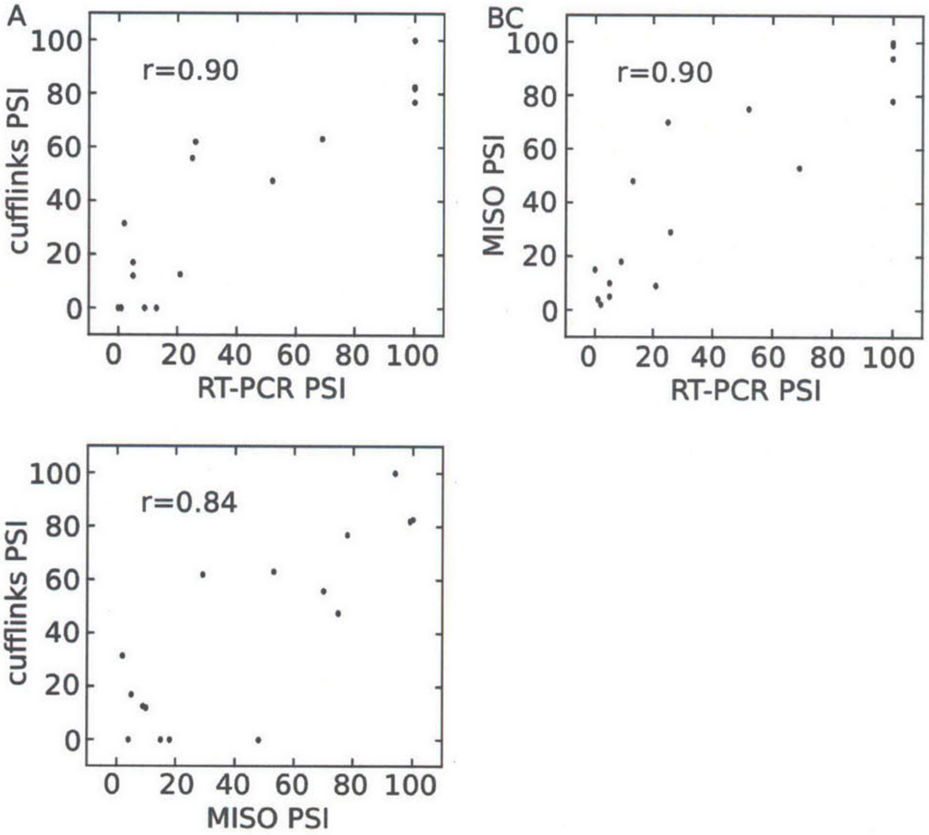


Figure 3-4:
PSI values were calculated by Cufflinks, MISO, and compared to each other and to qRT-PCR measurements from a recent study (19). The estimates from each method were compared as follows:

(a) Cufflinks compared with qRT-PCR

(b) MISO compared with qRT-PCR

(c) Cufflinks compared with MISO.

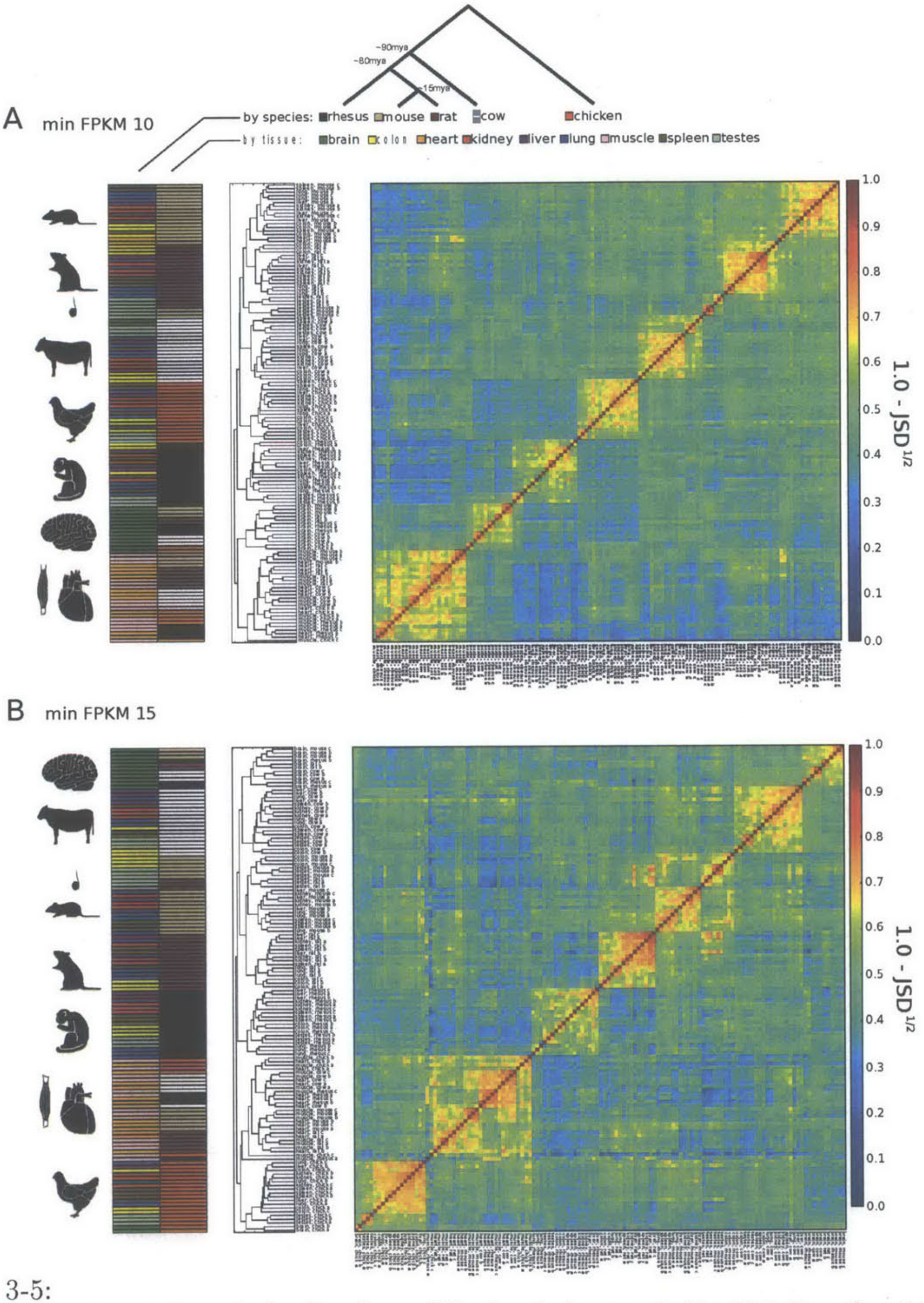


Figure 3-5:
Hierarchical clustering of all samples by skipped exon PSI values is shown, as in Fig. 3-2B. Here, the minimum expression cutoff required of each event to be considered was raised to

(a) 10 FPKM, or

(b) 15 FPKM

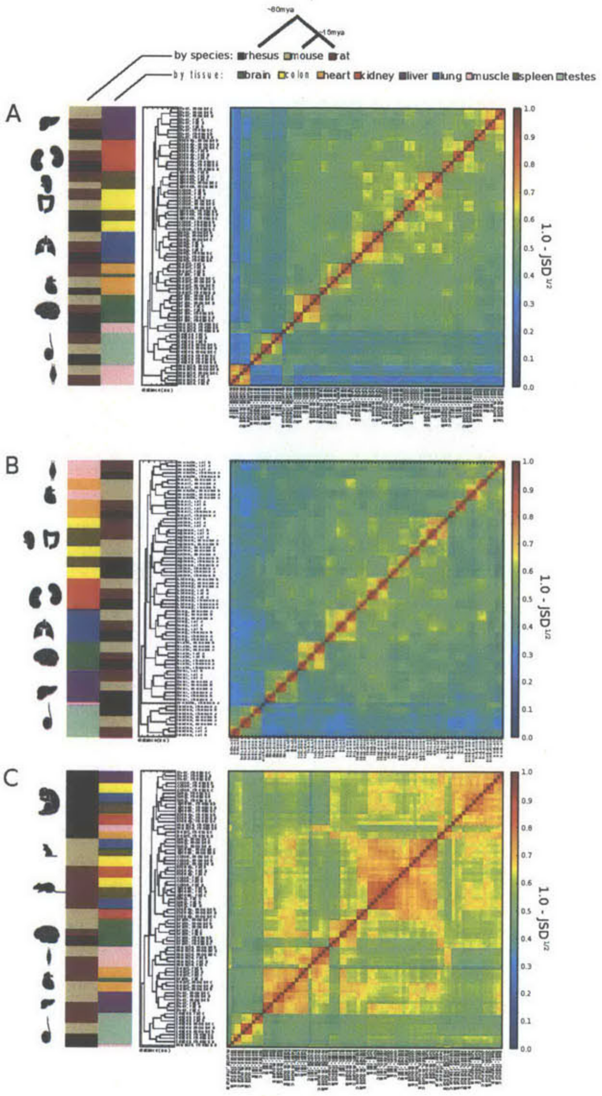


Figure 3-6:

Hierarchical clustering of samples by JSD based on FPKM for gene expression of transcription factors (top) or splicing factors (middle). As in Figure 3-2A, only singleton orthologs in mouse-rat-rhesus were used in the analysis. The species analyzed here were restricted to mouse-rat-rhesus in order to minimize the number of duplication and thus increase the number of singleton orthologs in each category. To explore the potential contributions to these patterns of variation in the expression of trans-acting factors, we performed clustering of the expression values of genes encoding transcription factors and splicing factors. While both types of regulatory factors tended to cluster primarily by tissue at small distances, at greater distances splicing factors were somewhat more likely to cluster by species, at least for the set of singleton orthologous genes studied, independent of whether the transcription factor genes were subsampled to match the number of splicing factor genes or not (bottom). This observation suggests that lineage-specific changes in splicing factor expression may have contributed to the tendency of splicing patterns to cluster by species more often than by tissue. Taken together, the clustering analyses above are consistent with a model in which gene expression differences often drive conserved differences in morphology and physiology between tissues, while splicing differences are more often species-specific, potentially contributing more frequently to phenotypic differences between species.

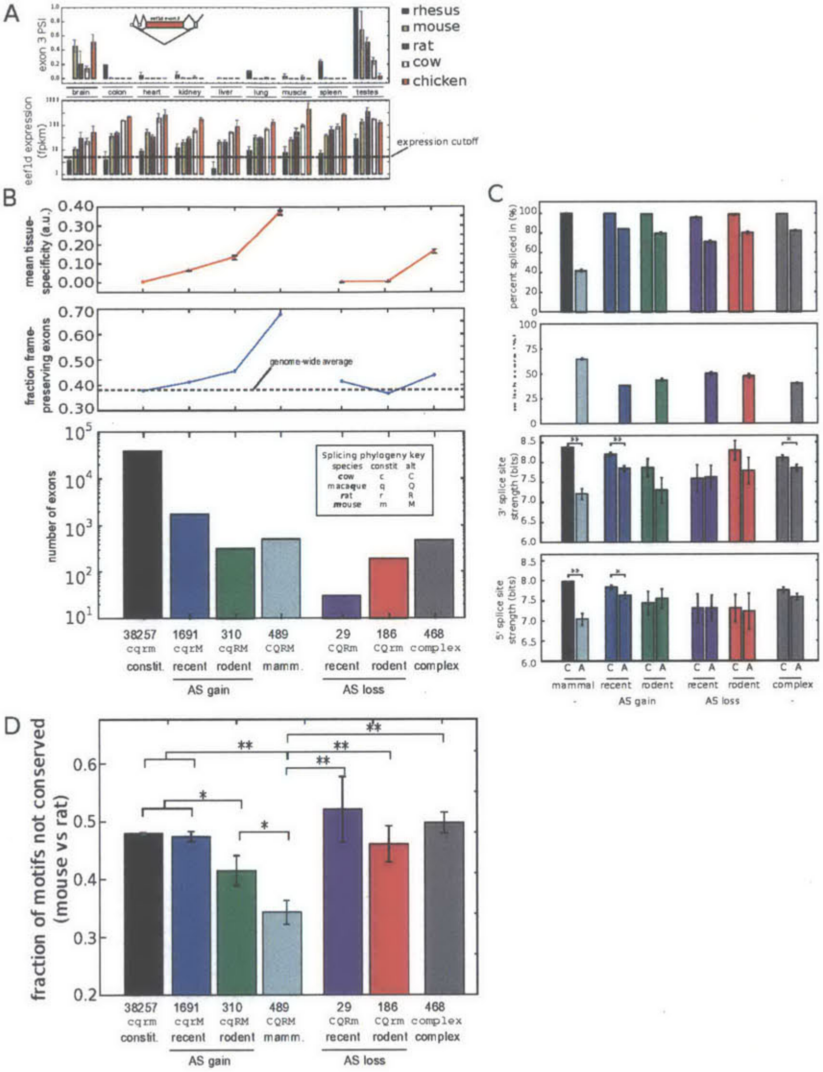


Figure 3-7:

(a) Above: PSI values for eef1d exon 3 across tissues and species analyzed. Below: Eef1d gene expression values. (Mean ± SD of 3 biological replicates). PSI values were calculated only for tissues with FPKM ≥ 5. Inset: exon structure of 5' end of eef1d gene (ENSMUSG00000055762).

(b) Below: Number of internal exons binned by the age of the inferred alternative splicing based on occurrence in ≥ 2 individuals. Above: The fraction of exons with length divisible by 3 and the mean and SEM of the tissue-specificity.

(c) Top: mean ± SEM of PSI values of exons binned by the phylogenetic extent of alternative splicing as in (B). Middle: mean ± SEM of 3' splice site scores of exons in each bin. Bottom: mean ± SEM of 5' splice site scores. Splice sites were scored using the MaxEnt model (31). *Indicates t-test p-value <0.05. **Indicates P <0.001.

(d) Fraction of regulatory 5mers in the downstream intron that differed between mouse and rat in exons binned by the phylogenetic extent of alternative splicing as in (B) (Mean ± SEM). *Indicates t-test P <0.05 (t-test). **P <0.001.

3.2.3 Features of conserved, tissue-specific alternative exons

A subset of several hundred alternative exons exhibited highly conserved tissue-specific splicing patterns. The gene for eukaryotic translation elongation factor 1 delta (*$EEF1\delta$*) (Fig. 3-7A) and many other examples in our data demonstrate that highly tissue-specific patterns of splicing can be conserved for hundreds of millions of years (9).

To assess the phylogenetic distribution of alternative splicing events across mammals, we grouped exons by the inferred age of alternative splicing, defined as PSI <97%. Out of ~48,000 internal exons with clear orthologs in chicken and at least two mammals, we identified exons alternatively spliced in a species- or rodent-specific manner as well as ~500 broadly alternative exons with alternative splicing observed in all mammals studied (Fig. 3-7B). Conversely, we identified exons that were constitutively spliced in a lineage-specific manner (and alternatively spliced elsewhere), representing losses of alternative splicing. Using data from the Illumina human Body Map 2.0 dataset, rhesus-specific alternative exons were twice as likely to be detected as alternatively spliced in human as were exons with exon skipping detected only in a single rodent (Fig. 3-8), consistent with the closer phylogenetic relationship of human to rhesus than to mouse or rat. In addition, more than 500 exons were identified whose phylogenetic splicing patterns imply multiple changes between constitutive and alternative splicing during mammalian evolution, suggesting frequent inter-conversion between constitutive and alternative splicing (10).

We observed a monotonic increase in tissue specificity within mouse as the phylogenetic breadth of alternative splicing increased from 1 to 4 mammals (Fig. 3-7B,C). The fraction of exons that preserved reading frame in both inclusion and exclusion isoforms also increased from ~40% to ~70% with increasing phylogenetic breadth of alternative splicing. These patterns suggest that more broadly occurring (ancient) alternative splicing events function primarily to generate distinct protein isoforms, which are often tissue-specific (11). On the other hand, while more lineage-restricted

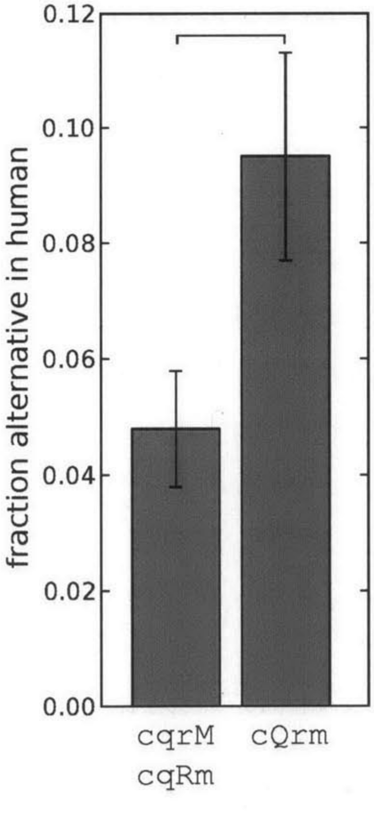


Figure 3-8:
The fraction of alternative exons that were detected alternatively spliced in either mouse/rat or rhesus that was detected as alternatively spliced in publicly available human data (Illumina human Body Map 2) are shown ± binomial standard deviation. * indicates p <0.01 by binomial proportion test.

(recently evolved) alternative splicing events contribute more often to regulation involving reading frame disruption, which may yield truncated or nonfunctional mRNAs or proteins or serve to down-regulate expression, usually in a less tissue-specific manner.

3.2.4 Splice site changes may convert alternative to constitutive splicing

Exons that recently converted from constitutive to alternative splicing had significantly weaker 3' and 5' splice sites in the alternative splicing species than in their constitutively spliced orthologs (Fig. 3-7C) (10). However, recent conversion from alternative to constitutive splicing was not associated with significant changes in 5' or 3' splice site strength, suggesting involvement of other events such as loss of negative-acting cis-regulatory elements. Constitutive exons that converted to alternative splicing in other species tended to have weaker splice sites than maintained constitutive exons (P ¡ 0.01, rank-sum test), suggesting that exons with weaker splice sites may be predisposed to convert to alternative splicing. We found that exons with nearby G-runs (often bound by hnRNP H family proteins) were 25-60% more likely to have converted from constitutive to alternative splicing (Fig. 3-9) (12).

Exons alternatively spliced in all mammals tended to have the weakest 5' and 3' splice sites, approximately 1 bit weaker than maintained constitutively spliced exons (Fig. 3-7C) (13). These exons had mean PSI values that were closer to 50% than other exon groups (Fig. 3-7C), suggesting that weaker splice sites may have evolved in these exons to enable a broader range of exon inclusion levels.

Splicing cis-regulatory elements located adjacent to (or within) alternative exons often confer regulation through interaction with cell type- or condition-specific protein factors (14). Using a large set of intronic splicing regulatory elements (ISRE) motifs recognized by both tissue-specific and broadly expressed splicing factors derived from

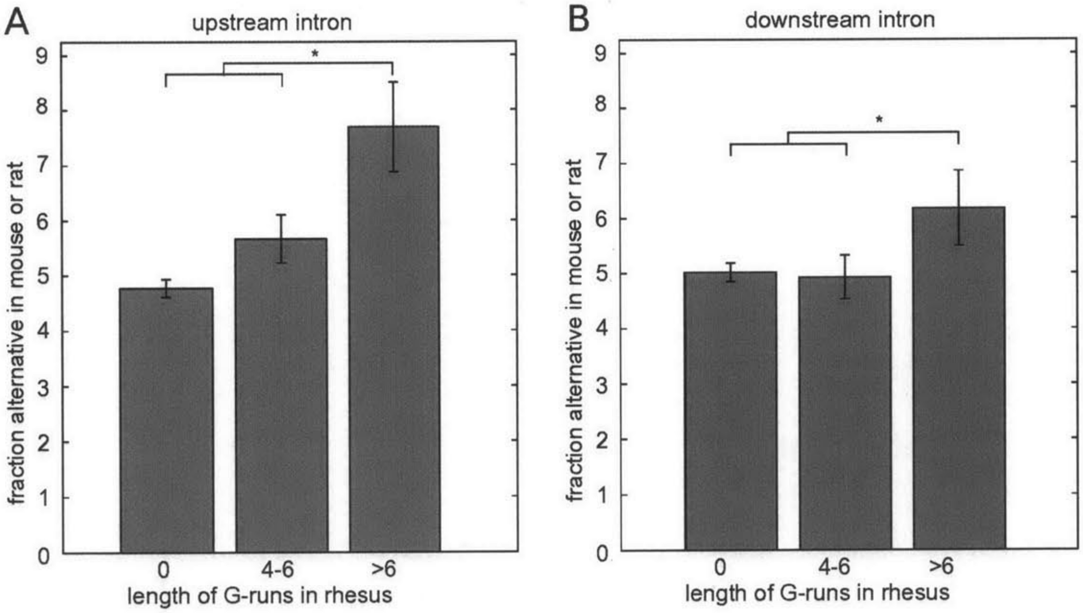


Figure 3-9:
The fraction of exons that are constitutive in rhesus with indicated numbers of Gs in G-runs in the

(a) upstream, or

(b) downstream

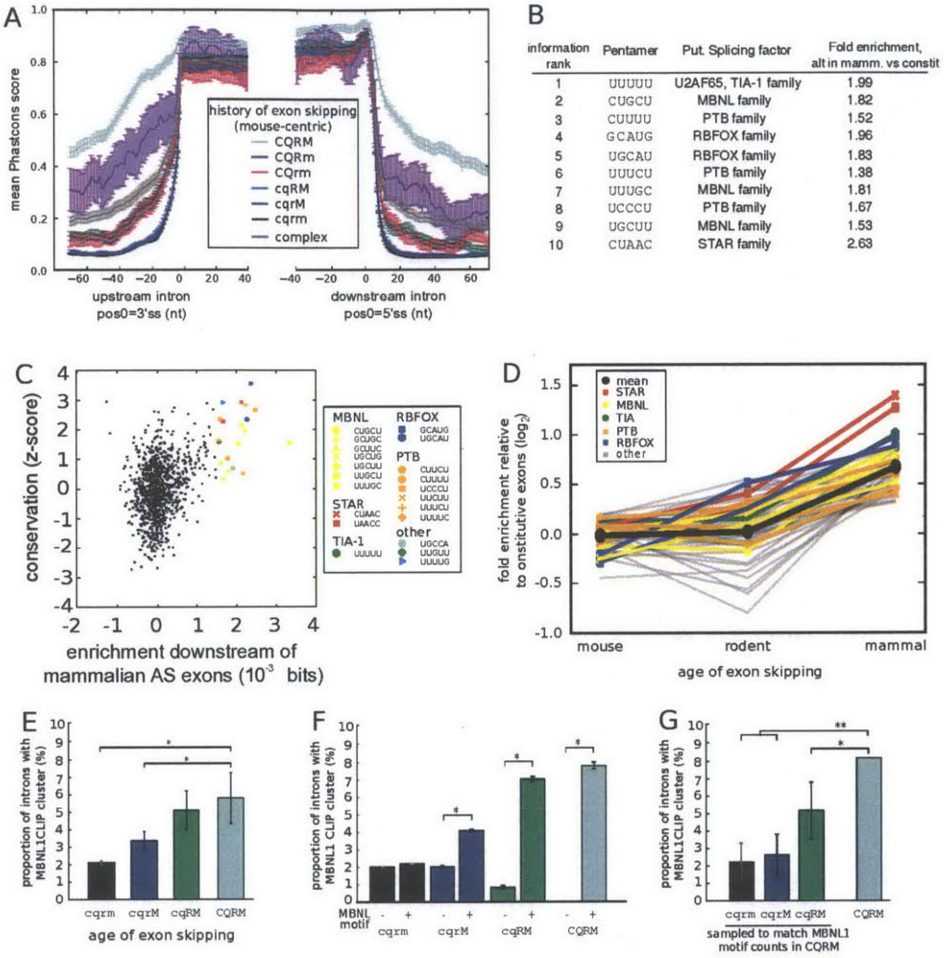
intron that are alternatively-spliced in mouse are shown ± the binomial standard deviation. * indicates $p < 0.05$ by binomial proportion test in both panels.

(15, 16), reduced motif turnover was observed in exons alternatively spliced in multiple species relative to constitutive or recently converted alternative exons (Fig. 3-7D) (11, 17). Exons that converted from alternative to constitutive splicing in one or both rodents showed substantially increased turnover of ISREs than mammalian-wide alternative exons (Fig. 3-7D), suggesting that mutations affecting ISREs may contribute to these conversions.

3.2.5 Tissue-specific regulatory motifs accumulate in broadly alternative exons

Using vertebrate whole-genome alignments, strong sequence conservation of only the exon and core splice site motifs was observed in broadly constitutive exons and exons that recently acquired alternative splicing. However, increased sequence conservation both within the exon and extending at least 70 bases into the intron on either side was observed with increasing phylogenetic breadth of alternative splicing (Fig. 3-10A), suggesting the occurrence of purifying selection on adjacent ISREs and providing support for the reliability of these exon classifications.

To assess the nature of potential regulatory elements present in introns adjacent to alternative exons, we ranked pentanucleotides (5mers) based on their relative frequency of occurrence in introns downstream of broadly alternative exons relative to constitutive exons using an information criterion (6). Among the top ten 5mers in this ranking were perfect or near-perfect matches to consensus motifs for tissue-specific splicing regulatory factors, including those of the MBNL, PTB, RBFOX, STAR and TIA families of splicing factors (18) (Fig. 3-10B). Presence of motifs associated with almost all of these splicing factor families was conserved downstream of broadly alternative exons more than two standard deviations more often than control motifs (Fig. 3-10C), implying strong selection to maintain their presence. Pronounced enrichment of these motifs was restricted primarily to exons with broad alternative splicing ($\geq$ 4 species), with only modest enrichment downstream of rodent-specific



information rank	Pentamer	Put. Splicing factor	Fold enrichment, alt in mamm. vs constit
1	UUUUU	U2AF65, TIA-1 family	1.99
2	CUGCU	MBNL family	1.82
3	CUUUU	PTB family	1.52
4	GCAUG	RBFOX family	1.96
5	UGCAU	RBFOX family	1.83
6	UUUCU	PTB family	1.38
7	UUUGC	MBNL family	1.81
8	UCCCU	PTB family	1.67
9	UGCUU	MBNL family	1.53
10	CUAAC	STAR family	2.63

Figure 3-10:

(a) Mean ± SEM of Phastcons scores (using the placental mammals alignment, with mouse coordinates) in exons and flanking introns grouped by phylogenetic pattern of alternative splicing. Splicing pattern indicated by letters adjacent to colored bars, as in Fig. 3-7B.

(b) Top ten 5mers in broadly alternative exons relative to constitutive exons ranked by discrimination information (6).

(c) The conservation of all 5mers (6) compared with their discrimination information. All 5mers with discrimination information ≥ 0.001 bits are highlighted. UUUUU was an outlier in enrichment (0.011 bits) and is not shown.

(d) Fold enrichment (log_2) relative to constitutive exons in downstream region was plotted for 5mers with discrimination information ≥ 0.001 bits for exons grouped by phylogenetic breadth of alternative splicing. 5mers associated with known splicing regulators are shown in color, with mean of all 5mers in black.

(e) The fraction of introns containing MBNL1 CLIP-Seq clusters (19) was assessed in introns adjacent to exons with different phylogenetic patterns splicing, as in (A).

(f) As in (E), but grouped by presence/absence of a MBNL1 motif. The mean fraction ± SEM of 1000 bootstrap samples is shown. *P <0.01 (binomial test).

(g) As in (E), but with exons sampled from each set to match the MBNL motif counts in the CQRM set. Mean ± SD of 1000 samplings is shown for the first 3 groups; observed mean is shown for the CQRM set. *P <0.05, **P <0.001.

alternative exons and little to no enrichment near mouse-specific alternative exons (Fig. 3-10D, Fig. 3-11). These observations suggested that exons with more ancient alternative splicing — which are more often tissue-specific (Fig. 3-7B) — are more reliant for their regulation on a distinct subset of ISRE motifs corresponding to the tissue-specific factors listed above (MBNL, RBFOX, etc.).

To explore this hypothesis, we analyzed cross-linking / immunoprecipitation sequencing (CLIP-Seq) data to assess the transcriptome-wide binding of the mouse splicing factor muscleblind-like 1 (MBNL1) (19). Greater phylogenetic breadth of alternative splicing was associated with ~3-fold increased frequency of in vivo MBNL1 binding (Fig. 3-10E). Presence of a MBNL motif was associated with increased binding near alternative but not constitutive exons (Fig. 3-10F), suggesting that motif presence is necessary but not sufficient for strong binding in vivo. As a group, broadly alternative exons have somewhat higher density of MBNL motifs (Fig. 3-10B), but increased frequency of MBNL binding was observed even when comparing to subsets of constitutive or more narrowly alternative exons with identical MBNL motif counts (Fig. 3-10G). These observations suggest that broadly alternative exons have evolved features beyond motif abundance (such as favorable RNA structural features) to enhance binding of MBNL family splicing regulators. This phenomenon may extend to other factors (Fig. 3-12).

3.2.6 Alternative splicing alters phosphorylation potential

Exons whose presence was widely conserved (at least 4 out of 5 species) were classified based on the tissue specificity and evolutionary conservation of their splicing patterns using JSD-based metrics (6) into constitutive exons and four groups of alternative exons grouped by the degree of tissue-specificity and evolutionary conservation of their splicing patterns. Functional analysis of species-specific alternative exons yielded few significant biases. However, analysis of the tissue-specific conserved group using DAVID (20, 21) identified a number of significantly enriched keywords and Gene

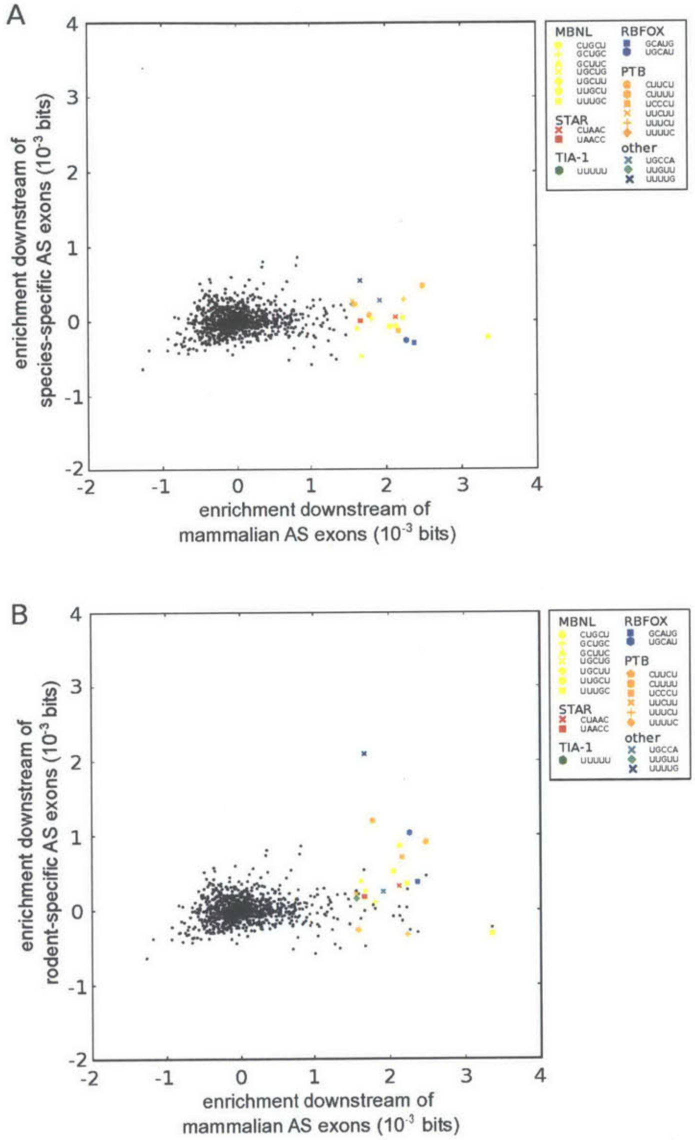


Figure 3-11:
Discrimination information content (bits) of 5mers in

(a) mouse-specific alternative exons, or

(b) rodent-specific alternative exons

are plotted against broadly alternative exons. All densities are calculated using only mouse introns. 5mers highlighted match those highlighted in Fig. 3-10C.

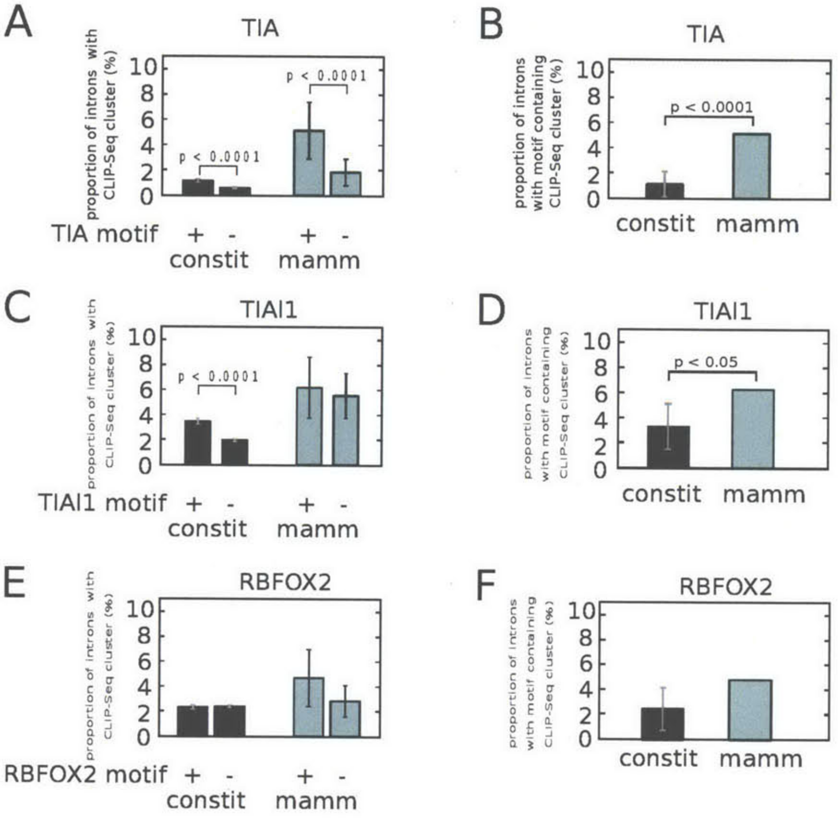


Figure 3-12:
Preferential binding of TIA family splicing factors to broadly alternative exons. CLIPSeq data for TIA-1 and TIAL1 was from (25); RBFOX2 CLIP-Seq data was from (26). The CLIP-Seq data analyzed in this figure were from human rather than from an organism analyzed in this study, so the most relevant sets of exons for comparison were exons that were either constitutive across mammals or alternatively spliced in all mammals.

(a) Frequency of TIA-1 CLIP-Seq clusters in introns with and without a TIA-1 motif. P-value calculated by test of binomial proportions (mean 1 binomial SD shown).

(b) Frequency of TIA-1 CLIP-Seq clusters in introns near constitutive exons resampled to match the distribution of motif counts near mammalian alternative exons. P-value calculated by bootstrap sampling (mean 1 SD of 1000 samplings with replacement is shown).

(c) As in A, but shows frequency of TIAL1 CLIP-Seq clusters in introns with and without a TIAL1 motif.

(d) As in B, but shows frequency of TIAL1 CLIPSeq clusters in introns near constitutive exons resampled to match the distribution of motif counts near mammalian alternative exons.

(e) As in A, but shows frequency of RBFOX2 CLIP-Seq clusters in introns with and without an RBFOX2 motif.

(f) As in B, but shows frequency of RBFOX2 CLIP-Seq clusters in introns near constitutive exons resampled to match the distribution of motif counts near mammalian alternative exons.

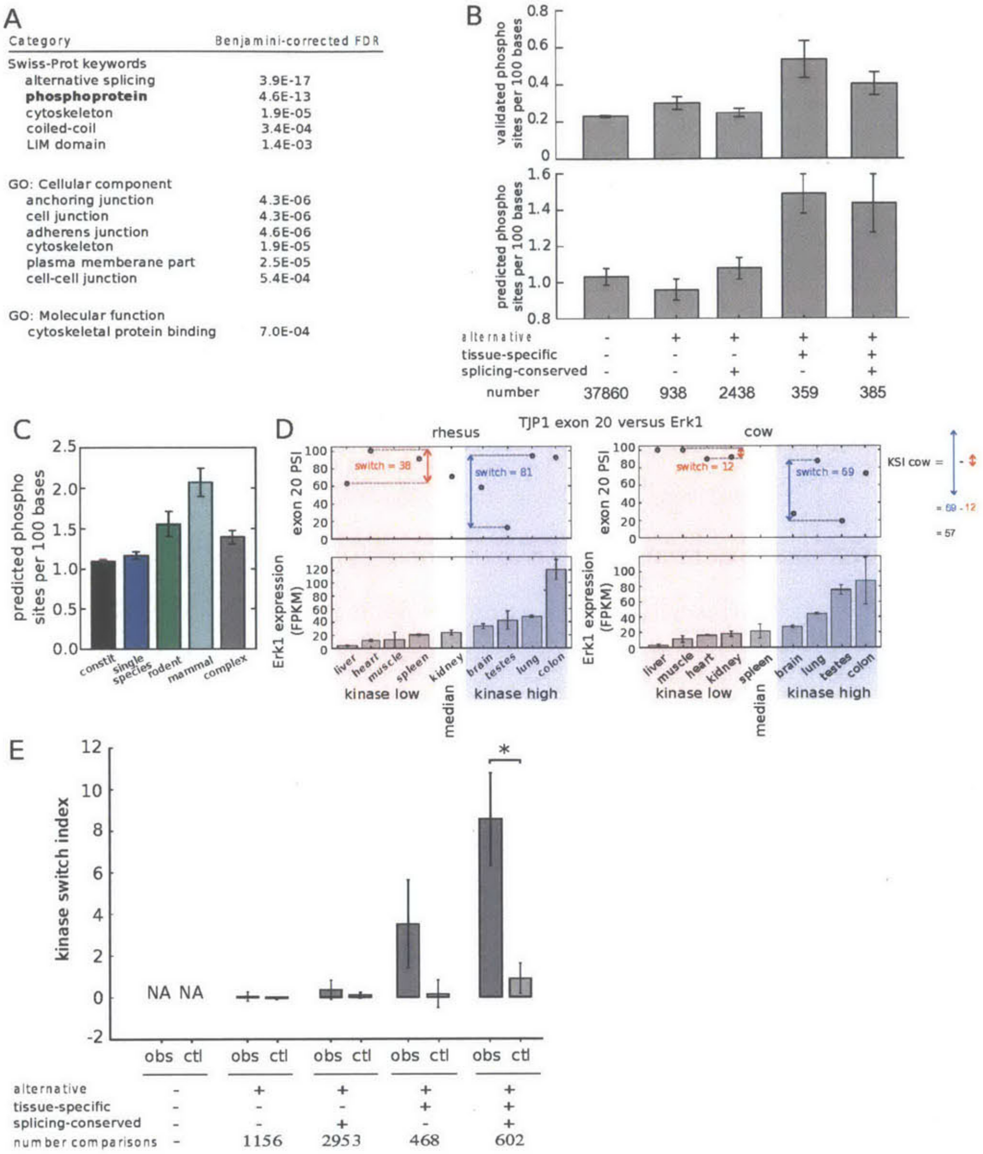


Figure 3-13:

(a) GO analysis of genes containing tissue-regulated exons whose splicing is conserved.

(b) Density of Phosphosite phosphorylation sites (top) or Scansite predicted phosphorylation sites (bottom) in exons grouped by alternative splicing status, tissue-specificity and splicing pattern conservation (6). Mean ± SEM is shown.

(c) Mean Scansite predicted phosphorylation site density in exons grouped by phylogenetic breadth of splicing.

(d) TJP1 exon 20 splicing has a higher switch score in tissues where Erk1 is expressed above median levels (shaded blue) than where it is expressed below median (shaded pink); KSI is defined as the difference between these switch scores. PSI value not calculated if TJP1 expression fell below a cutoff (e.g., cow spleen).

(e) Mean KSI values for kinase-exon pairs involving the sets of exons as in (B). Mean ± SEM is shown. Observed values (obs) were compared with controls in which PSI values in different tissues were randomly permuted (ctl). Comparisons marked (*) were significant by Mann-Whitney U test (P <0.005).

Ontology categories, including several related to cell-cell junctions and cytoskeleton (Fig. 3-13A), suggesting that these splicing events may contribute to differences in cell structure, cell motility and tissue architecture (22). The most enriched keyword, alternative splicing, reflects simply the abundant alternative splicing of this set of genes; the next most significantly enriched keyword was phosphoprotein.

To explore this connection to phosphorylation, we used Scansite (23) to predict phosphorylation sites in peptides encoded by different subsets of exons. Tissue-specific alternatively spliced exons, including both the conserved and non-conserved subsets, contained about 40% more predicted phosphorylation sites than other classes of exons (Fig. 3-13B and Fig. 3-14). A comparable degree of enrichment for phosphorylation sites was observed in these exons using the curated Phosphosite database (24) of experimentally determined phosphorylation sites (Fig. 3-13B). Phosphorylation site density in exons was correlated with phylogenetic breadth of alternative splicing, (Fig. 3-13C, Fig. 3-15). These observations suggest that tissue-specific alternative splicing is often used to alter the potential for protein phosphorylation, which can alter protein stability, enzymatic activity, subcellular localization and other properties.

Exon 20 of the mouse tight junction protein 1 (TJP1) gene exhibits strongly tissue-specific alternative splicing and encodes a peptide containing an established phosphorylation site (25) that is predicted to be phosphorylated by ERK1 (aka MAPK3). In rhesus, ERK1 expression was above its median value in colon, lung, testis and brain (therefore referred to as kinase high tissues) relative to liver, heart, muscle and spleen (kinase low tissues). The PSI value of TJP1 exon 20 was much more variable in the kinase high tissues, ranging from 9% in testis to 90% in in colon — a switch score of 81% — than in the kinase low tissues, where it had a switch score of 38%. Similar trends were observed in cow (Fig. 3-13D), and in rat and mouse (not shown).

To explore this phenomenon, we analyzed the splicing patterns of exons that con-

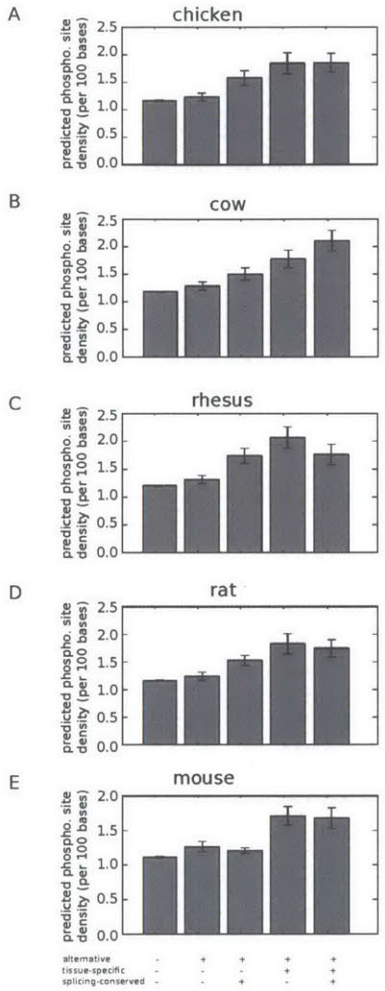


Figure 3-14:
Predicted phosphorylation site density in similar classes of exons as analyzed in Figure 3-13B in other species studied:

(a) chicken

(b) cow

(c) rhesus

(d) rat

(e) mouse

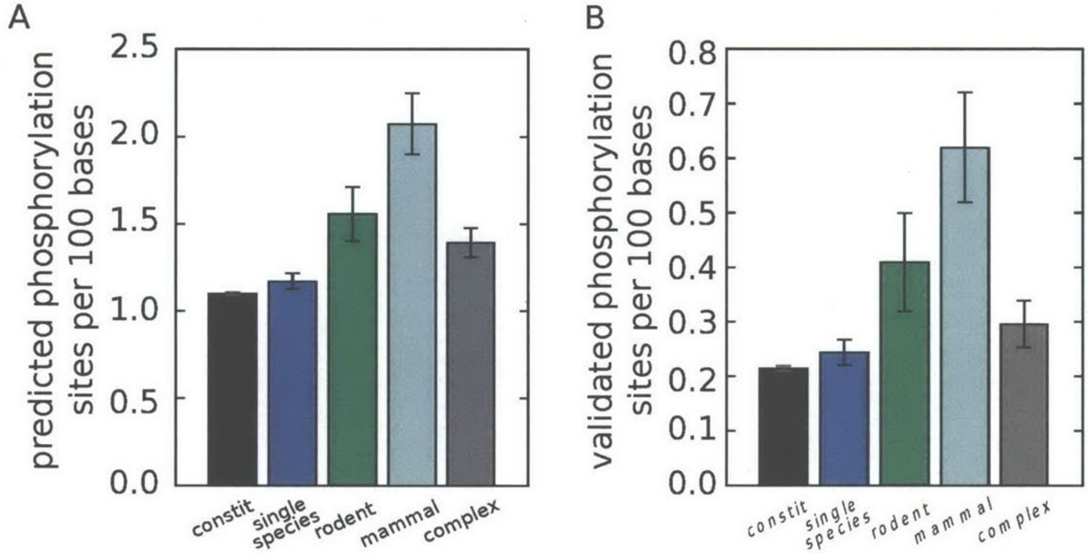


Figure 3-15:

(a) Predicted phosphorylation site density in exons binned by how long they have been alternatively spliced.

(b) Experimentally-determined phosphorylation site density in exons binned as in (A).

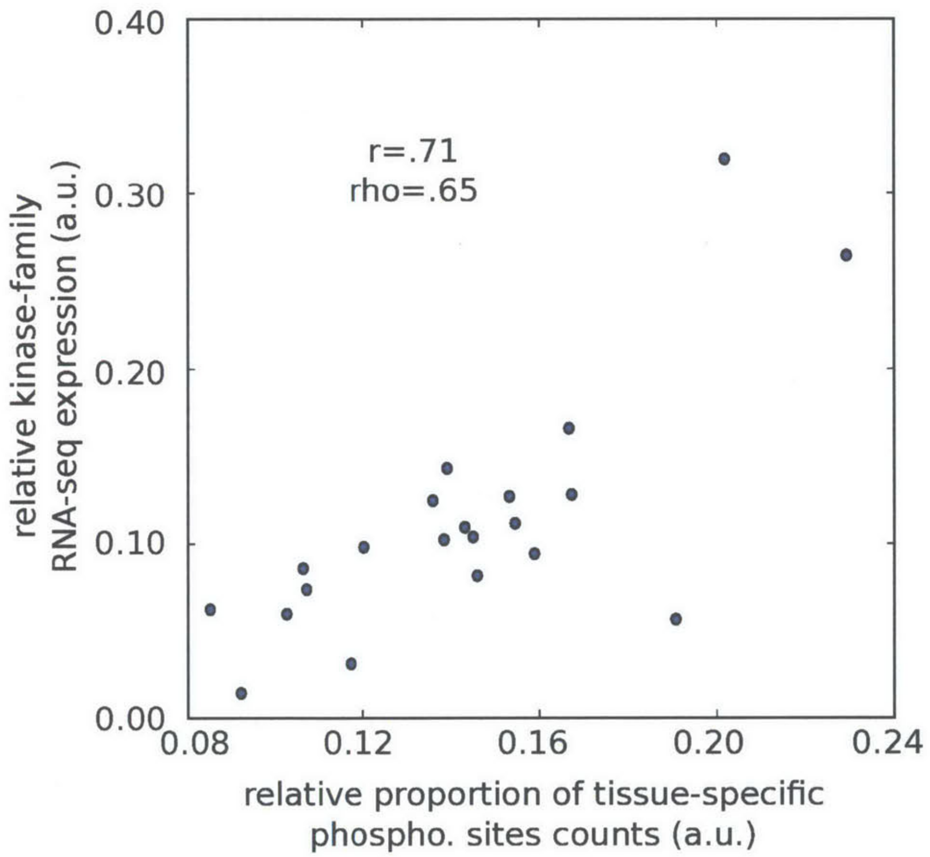


Figure 3-16:
The relative expression of kinases that are acid-directed, base-directed, or prolinedirected and also evaluated in Scansite are plotted (y-axis) against the normalized distribution of spectral counts assigned to each family in (44).

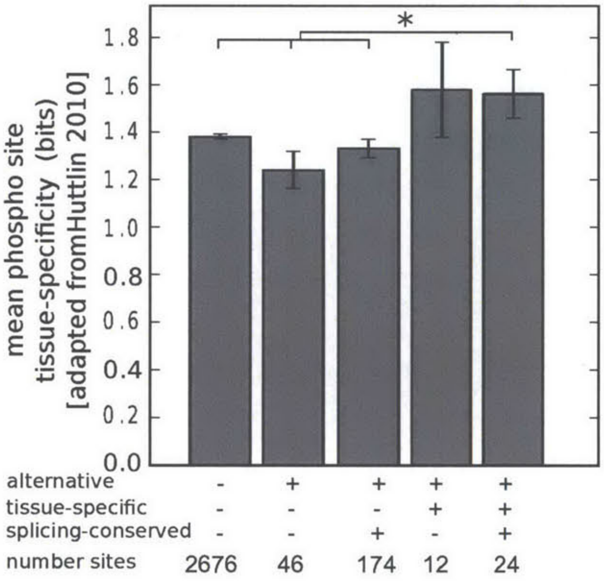
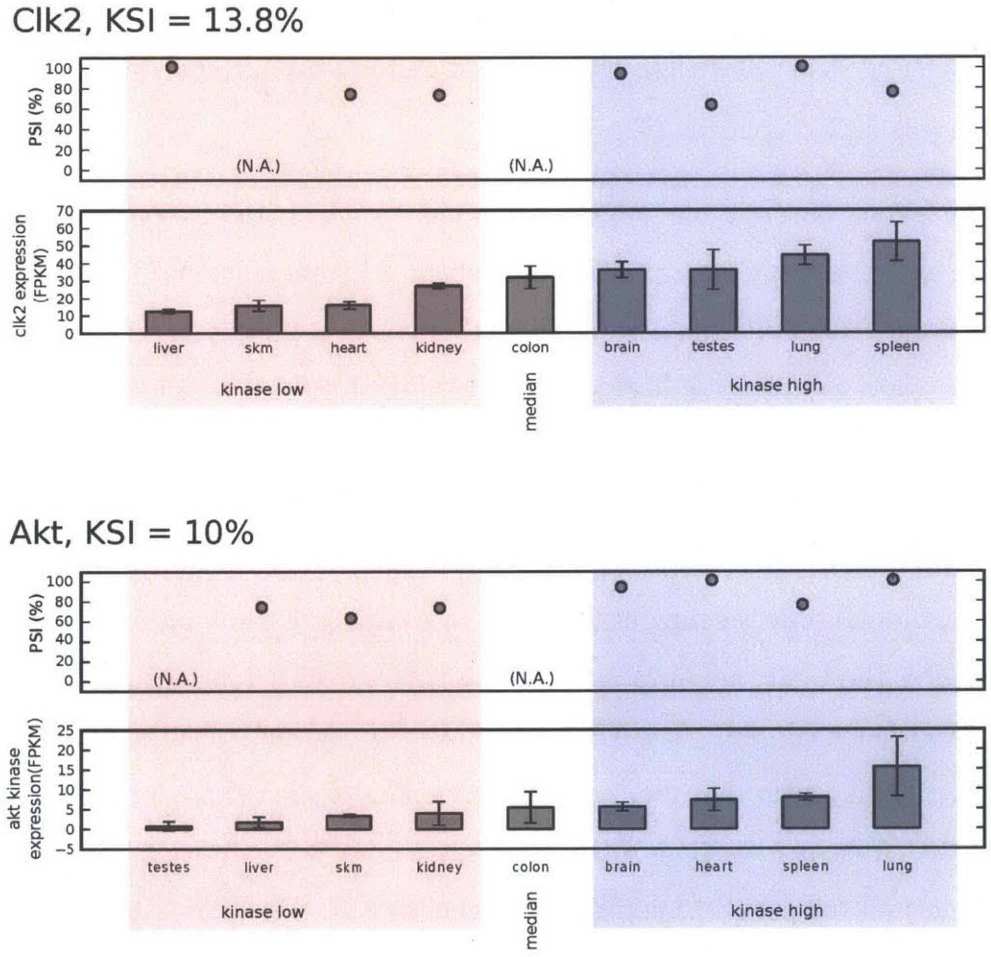


Figure 3-17:
Tissue-specificity of phosphorylation sites identified by (44) were calculated as described by the authors using spectral counts and the mean tissue-specificity ± SEM of exons binned by various filters are shown. * indicates p <0.05 by Mann-Whitney U test

tained predicted phosphorylation sites in relation to the expression of the associated kinases. To characterize the relationship between each exon-kinase pair, the kinase switch index (KSI) was defined as the switch score in kinase high tissues minus the switch score in kinase low tissues (see example in Fig. 3-13D). We used RNA-Seq estimates of kinase expression, which were reasonably well correlated with in vivo kinase activity patterns ($r = 0.71$, Fig. 3-16). We observed that phosphorylation of sites within conserved, tissue-regulated exons is more tissue-specific than in other sets of exons (Fig. 3-17) and that these exons exhibit substantially elevated KSI values relative to shuffled controls (Fig. 3-13D).

The above observations suggest a model in which tissue-regulated alternative splicing delimits the scope of tissues in which a kinase can phosphorylate a target. For example, the TJP1 protein mentioned above is a cytoplasmic constituent of the tight junction complex implicated in the timing of tight junction formation, and its phosphorylation is involved in tight junction dynamics (26, 27). The testes display unique tight junction biology in that tight junctions regularly dissolve and reform to permit passage of preleptotene spermatocytes (28). The specific exclusion of exon 20 in the mammalian testis may allow TJP1 to escape phosphorylation that would otherwise alter the tight junction association kinetics required for normal testis function (29). Another example, with KSI value closer to the mean value, is an alternative exon in Drosha, a protein required for processing of microRNA primary transcripts (Fig. 3-18). Phosphorylation of Drosha confers nuclear localization, which is required for its normal function in microRNA biogenesis (30). Therefore, splicing of Drosha may be used to alter the level of phosphorylatable Drosha protein, potentially influencing microRNA abundance in different cells or tissues.

We identified a large number of other exon-kinase pairs with elevated KSIs, including prominent kinases involved in development, cell cycle and cancer (e.g., Akt1, Clk2, PKC, Src) and targets with important roles in tissue biology such as Atf2, Enah, and Vegfa (Fig. 3-19). Their elevated KSIs suggest that splicing is often used to focus the scope of signaling networks, connecting specific kinases to specific targets



Clk2, KSI = 13.8%

Akt, KSI = 10%

Figure 3-18:
The splicing of mouse Drosha exon 5 is compared with the expression of cognate kinases

(a) Clk2

(b) Akt

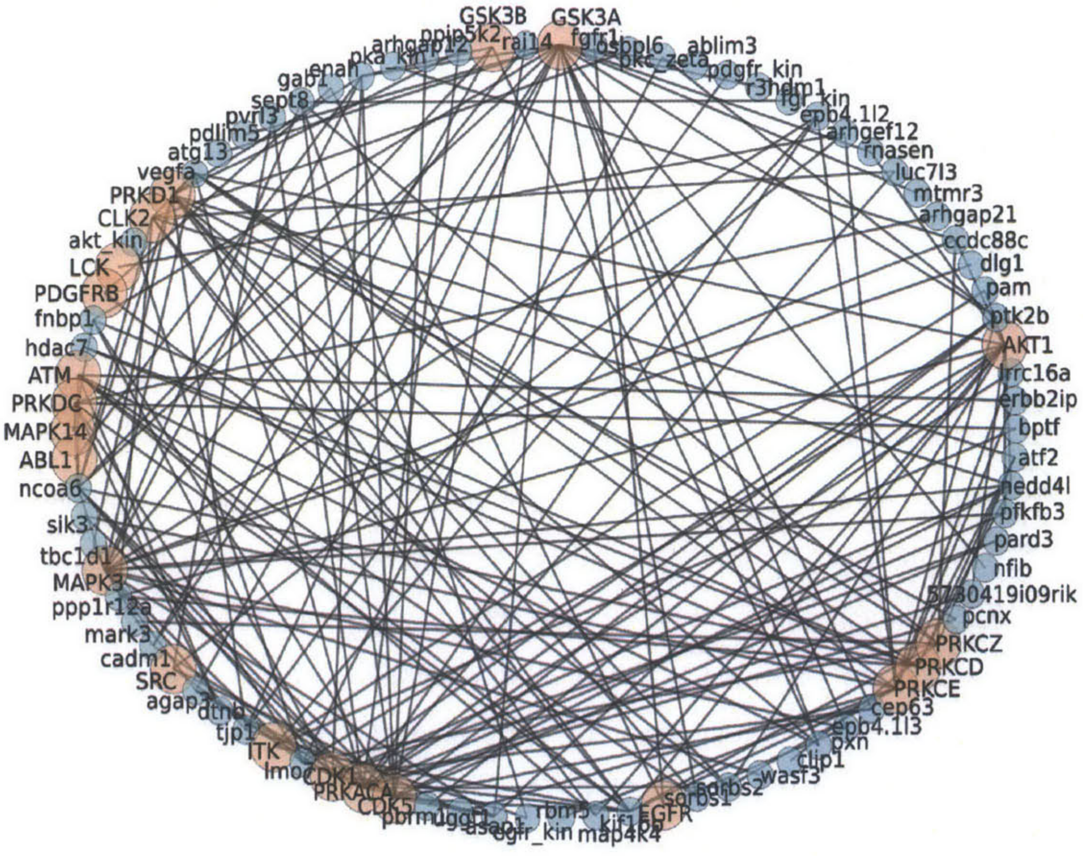


Figure 3-19:
Network visualization of all kinase-exon relationships identified in the conserved, tissuespecific category of exons with a KSI ¿= 5%. Edges connect kinases (orange circles) with target genes that contain alternative exons with putative phosphorylation site(s) for that given kinase (blue circles).

in a more cell- or tissue-restricted fashion than would occur from expression alone.

Taken together, the analyses described here reveal two disparate facets of mammalian alternative splicing. We identify a core of ~500 exons with conserved alternative splicing in mammals and high sequence conservation. These exons often encode phosphorylation sites and their tissue-specific splicing is likely to have substantial impacts on the outputs of signaling networks. Conversely, we observe extensive variation in the splicing of these exons between species, often exceeding intra-species differences between tissues, suggesting that changes in splicing patterns often contribute to evolutionary rewiring of signaling networks.

3.3 Acknowledgements

We thank Alex Robertson for analysis of coding potential of alternative isoforms, Eric Wang for help with the Mbnl1 CLIP analysis, Daniel Treacy for assistance with library preparation, Sean McGeary, David Page, Athma Pai, Charles Lin, and members of the Burge lab for comments on the manuscript, and the MIT BioMicro Center for assistance with sequencing. This work was supported by a Broad Institute SPARC grant (C. B. B.), by an NIH training grant (J. M.), by grant number 0821391 from the NSF, and by grants from the NIH to C. B. B.

3.4 Author Information

Sequence data associated with this manuscript have been submitted to NCBI GEO (accession number GSE41637).

3.5 Supplementary Methods

3.5.1 Sample collection

The tissues were chosen to represent a broad spectrum of the organs that are present in birds and mammals and are derived from all three germ layers. Multiple unrelated individuals from each species were sequenced to help distinguish species-specific alternative splicing from individual variation. A complete set of tissues was collected from three individuals from each of the species, enabling separate analysis of tissuespecific and individual variation. Animals of breeding age were chosen as gene expression is reported to be relatively stable in this age group for mammals (32). Only males were used, to enable analysis of testis, a tissue with an unusually large extent of alternative splicing (33). Mouse and rat strains included two inbred strains that differed from each other by approximately 1 single-nucleotide polymorphism (SNP) per kilobase (kb) of genomic sequence (roughly comparable to the similarity between unrelated humans), and one individual from an outbred line. Tissues were isolated from freshly sacrificed animals (10-30 minutes from time of death to tissue fixation) from the following areas of the organs: visual cortex, encompassing both grey and white matter (brain); left ventricle, transmural (heart); right quadriceps (skeletal muscle); right middle lobe (mammals, lung) or right lung excluding air sac (chicken, lung); inner edge of spleen, avoiding blood vessels; ascending colon; right kidney, from the lower pole encompassing both cortex and medulla; right testis, transverse section. Tissues were washed twice in cold PBS and stored in RNALater (Ambion) per manufacturer's instructions.

3.5.2 Library construction

RNA was isolated in Trizol, purified on miRNEasy columns (Qiagen), and treated with on-column DNAse digestion (Qiagen). RNA quality was assessed on Agilent Bioanalyzer. Libraries were constructed using the dUTP strand-specific method (34), with the second-day reactions performed on the Spriworks SPRI-TE machine (Beckman Coulter). Fragments between 200-400 nt were chosen, and 300 nt fragments were further size-selected in 2% agarose gel. Multiplexed barcodes were added during final PCR amplification, and sequencing was performed on Illumina GAIIx or Illumina HiSeq instruments. Paired-end short read sequencing (2 x 36-50 bases) of two individuals of each species was combined with paired-end long read sequencing (2 x 80 bases) of the third individual to facilitate genome annotation and transcript discovery.

3.5.3 Read mapping and analysis

The 2x80 base libraries were mapped to their respective genomes (musmus9, rhemac2, ratnor4, bostau4, galgal3) using Tophat (35) version 1.1.4 and Ensembl Release 61 (Ensembl) annotations. The junctions from all of these libraries within a species were combined and used as input in a second round of mapping, wherein all of the libraries were mapped using the same set of defined junctions. Cufflinks (8) version 1.0.2 was used to identify novel transcripts in each of the 2 x 80 base libraries. The set of transcripts from each library, along with the existing Ensembl annotations, were compiled into a single set of annotations for each species using Cuffcompare (36). Cufflinks was then used on each library to quantitate the same set of transcripts. For each transcript, a translation start site was assigned by using an annotated start site if one was contained in the transcript, or the longest coding region. Cufflinks was used to estimate transcript abundance in each library (in standard FPKM units), and these values were used as the basis for splicing estimates or summed to obtain gene expression values.

These data can be used to substantially extend existing annotations for these species. Based on Cufflinks analysis of the long-read data alone, in each species we identified tens thousands of unannotated exons in known genes. We expect that these data will provide a rich resource for characterization of novel exons and transcripts in mammals and studies of their function and evolution. In addition to unannotated sense-oriented exons in known transcripts, we identified several thousand unannotated "antisense exons" occurring in spliced transcripts overlapping known genes in antisense orientation. We also identified at least several thousand putative novel exons in transcripts not overlapping known genes in each species.

3.5.4 Orthologous exon assignments

Exons with multiple 5' and 3' ends were collapsed into a single exon for the purposes of this study. Therefore, the set of exons was determined by finding the longest exonic region in transcripts with Cufflinks class codes of "j" or "=" by taking the most intronproximal 5' and 3' splice sites that do not include retained intron(s). Orthologous exons were identified by finding annotated exons that overlapped with the query exonic region in Ensembl Pecan 19 amniota genome alignments (37). To simplify the analysis, exon groups with multiple overlapping exons in any species were excluded. Exons were considered lost" in a species if there was no syntenic region in that species or if no exon overlapping the syntenic region was identified and spliced into transcripts identified herein with a PSI ≥ 0. All analyses except those in Figure 3-10 were restricted to exons that were detected in >2 tissues in more than half of the individuals (requiring chicken).

3.5.5 Gene expression and PSI calculation

Expression of transcripts with Cufflinks class codes of "c", "j", "=", "e", or "o" were summed to determine an overall gene-level expression estimate. Genes were considered to be alternatively spliced if they contained a 5' splice site that was spliced to 2 unique 3' splice sites or a 3' splice site that was spliced to 2 unique 5' splice sites, where each junction was supported by 5 unique reads, similar to (3). Skipped exons (SE) were identified by looking for exons that were excluded in ≥ 1 detected transcript with a class code of "j" or "=" with genomic coordinates sufficient to have potentially included the exon. The FPKM of transcripts that included the exon was divided by that of transcripts that spanned the location of the exon to calculate a skipped exon PSI. This will avoid artificially deflated PSI values when a transcript exhibits an internal transcription initiation site. Exon inclusion levels were calculated from transcript abundance measurements using Cufflinks (8). These values, estimates made using MISO (38), and qRT-PCR measurements from 3 mouse tissues (19) were all highly correlated with each other (r between 0.84 and 0.90, Fig. 3-4). In Figure 3-2 and elsewhere, samples were compared using Jensen-Shannon divergence between expression values or PSI values. JSD is a symmetrical information theoretic measure of distance between distributions whose square root is proportional to the Fisher information metric and which has several desirable properties relative to correlation-based measures (39, 40). The observation that samples of the same tissue from different individuals of the same species were highly similar helps to validate the consistency of tissue and library preparation, and the values obtained (JSD between 0.05-0.40, correlation values between 0.80-0.99) provide measures of the extent of variation between unrelated individual mammals and birds.

3.5.6 Phylogenic inference of age of alternative splicing

Parsimony was used to infer the age of an alternative spicing event. An exon alternatively spliced exclusively in mouse, rat, or rhesus was considered to be a species-specific gain of alternative spicing. Rhesus-specific events were grouped with mouse- and rat-specific events despite their different ages in absolute years because the substitution distance to rhesus from its most recent common ancestor in this species tree is similar to that of either mouse or rat (41). Observing alternative splicing in mouse and rat, but not rhesus and cow was considered rodent-specific. Observed alternative splicing in mouse, rat, rhesus, and cow was considered as alternatively spliced in mammals. Alternative spicing in cow and any two of mouse, rat, and rhesus was considered to be a species-specific loss of alternative splicing. Alternative splicing in cow and rhesus but not mouse or rat was considered to represent rodent-specific loss of alternative splicing. Exons that did not fall into these classifications were considered complex. Exons with more recently evolved alternative splicing tended to have mean PSI values near 80% (Fig. 3-13C), suggesting that the exon inclusion isoform may be present at sufficient levels to confer most of the ancestral protein function. In their orthologs where constitutive splicing was observed (defined as PSI >97% in all tissues in at least 2 out of 3 individuals), mean PSI values were slightly lower than for broadly constitutive exons. This difference reflects frequent alternative splicing in the third individual and suggests that conversion to constitutive splicing is often incomplete, but may persist as a polymorphic trait in the population. An important caveat to the inferences of phylogenetic breadth of splicing made here is that we have examined only 9 tissues in adult males, potentially neglecting splicing events that occur exclusively in other tissues, in females, or at other developmental stages.

3.5.7 Tissue-specificity and splicing conservation calculations

For each orthologous event, a matrix of n rows by m columns was constructed, where each row represented an individual and each column represented a tissue. Tissue-specificity was calculated within each row, yielding a vector with n dimensions, while splicing conservation was calculated within each column, yielding a vector with m dimensions. In each case, the square root of the Jensen-Shannon Divergence of each row (for tissue-specificity) or column (for splicing conservation) vector relative to a uniform vector with all values set at the vectors empirical mean was calculated. The JSD was then recalculated for the matrix consisting of 1 - PSI values, considering that exon exclusion in a single sample is as informative as inclusion in a single sample, and the values were averaged. The vector of divergences was then averaged, yielding a pair of values for each event, representing the tissue-specificity and splicing conservation of each event. Tissue-specific exons would be expected to have higher divergences while splicing-conserved exons would have lower values. These measures are defined for constitutive exons, as well as for exons that are only alternative in a subset of samples.

3.5.8 5mer motif analysis

The frequencies of each 5mer in intronic bases 10 to 130 relative to the 5' splice site. We focused on the region near the 5' splice site rather than the 3' splice site, to avoid complications relating to the (typically unknown) location of the branch point sequence; similar results were obtained using the upstream intronic region (not shown). In Fig. 3-10B, 5mers were ranked by "discrimination information", $flog_2(f/g)$, where f is the frequency in the window between 10 and 130 bases downstream of the 5' splice site in broadly alternative exons and g is the frequency downstream of constitutive exons. The fold enrichment for exons above a minimum discrimination information was plotted for lowly (species-specific), intermediately (rodent-specific), and broadly (all mammals) alternative exons in Fig. 3-10D. In Fig. 3-10C, conservation was calculated as a z-score, comparing the mean branch length over which presence of the 5mer is conserved to the mean branch lengths for a set of 5mers matched for A+T content and CpG dinucleotide content. In Fig. 3-10G, neither the relative proportions of different MBNL motifs nor the relative proportions of nucleotides immediately flanking each MBNL motif differed between the sets of exons analyzed (Chi-square test, corrected for the number of comparisons). Analysis of available CLIP-Seq data indicated a similar pattern of increased binding to broadly alternative exons beyond that expected from motif abundance for the human TIA-1 and TIAL1 splicing factors as well (42, 43) (fig. S8).

3.5.9 Analysis of phosphorylation

Coding sequences were determined for each transcript by first determining if there was an annotated coding start for the gene contained within. If one was found, it was used in downstream analyses. Otherwise, the longest open reading frame (ORF) was determined and used. To identify predicted phosphorylation sites, Bioperl was used to submit and handle queries to Scansite (23) of complete ORFs. Putative phosphorylation sites were cross-referenced with Phosphosite (24) to identify experimentally validated sites. To consider the tissue-specificity of sites, data from (44) were used. Similar to the authors, we calculated the entropy of spectral counts corresponding to individual phosphopeptides over the distribution of tissues as a measure of tissue-specificity, restricting to the tissues that overlap with our study. To compare kinase expression with kinase activity, we normalized the distribution of spectral counts assigned to each family of kinase that is also predicted by Scansite to the sum-total of the familys counts and compared it with a similarly normalized expression vector composed of the sum of the kinases in each tissue. Increased post-translational modification was also observed in a set of tissue-specific human exons (45) identified based on a previous RNA-Seq analysis of human tissues (3).

3.5.10 Data analysis

Custom Python scripts were used for analyses, utilizing Numpy (numpy.scipy.org), Scipy (scipy.org), Pycogent (46), BioPerl, Tabix, Samtools, Bedtools, FSA and Matplotlib [http://www.citeulike.org/user/jabl/article/2878517].

3.6 References

[1] S. Stamm et al. Gene 344, 1 (Jan 3, 2005).

[2] L. F. Lareau, A. N. Brooks, D. A. Soergel, Q. Meng, S. E. Brenner. Advances in experimental medicine and biology 623, 190 (2007).

[3] E. T. Wang et al. Nature 456, 470 (Nov 27, 2008).

[4] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe. Nature genetics 40, 1413 (Dec, 2008).

[5] D. Brawand et al. Nature 478, 343 (Oct 20, 2011).

[6] Materials and methods are available as supplementary material on Science Online.

[7] E. T. Chan et al. Journal of biology 8, 33 (2009).

[8] C. Trapnell et al. Nature biotechnology 28, 511 (May, 2010).

[9] N. Jelen, J. Ule, M. Zivin, R. B. Darnell. PLoS genetics 3, 1838 (Oct, 2007).

[10] G. Lev-Maor et al. PLoS genetics 3, e203 (Nov, 2007).

[11] G. W. Yeo, E. Van Nostrand, D. Holste, T. Poggio, C. B. Burge. Proc Natl Acad Sci U S A 102, 2850 (Feb 22, 2005).

[12] X. Xiao et al. Nature structural & molecular biology 16, 1094 (Oct, 2009).

[13] D. Baek, P. Green. Proc Natl Acad Sci U S A 102, 12813 (Sep 6, 2005).

[14] A. J. Matlin, F. Clark, C. W. Smith. Nat Rev Mol Cell Biol 6, 386 (May, 2005).

[15] S. C. Huelga et al. Cell reports 1, 167 (Feb 23, 2012).

[16] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, T. R. Hughes. Nucleic acids research 39, D301 (Jan, 2011).

[17] R. Sorek, G. Ast. Genome research 13, 1631 (Jul, 2003).

[18] A. N. Ladd, T. A. Cooper. Genome biology 3, reviews0008 (Oct 23, 2002).

[19] E. T. Wang et al. Cell 150, 710 (Aug 17, 2012).

[20] W. Huang da, B. T. Sherman, R. A. Lempicki. Nature protocols 4, 44 (2009).

[21] W. Huang da, B. T. Sherman, R. A. Lempicki. Nucleic acids research 37, 1 (Jan, 2009).

[22] I. M. Shapiro et al. PLoS genetics 7, e1002218 (Aug, 2011).

[23] J. C. Obenauer, L. C. Cantley, M. B. Yaffe. Nucleic Acids Res 31, 3635 (Jul 1, 2003).

[24] P. V. Hornbeck, I. Chabra, J. M. Kornhauser, E. Skrzypek, B. Zhang. Proteomics 4, 1551 (Jun, 2004).

[25] N. Dephoure et al. Proceedings of the National Academy of Sciences of the United States of America 105, 10762 (Aug 5, 2008).

[26] G. Samak, S. Aggarwal, R. K. Rao. American journal of physiology. Gastrointestinal and liver physiology 301, G50 (Jul, 2011).

[27] E. Sabath et al. Journal of cell science 121, 814 (Mar 15, 2008).

[28] D. D. Mruk, C. Y. Cheng. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365, 1621 (May 27, 2010).

[29] S. Aggarwal, T. Suzuki, W. L. Taylor, A. Bhargava, R. K. Rao. The Biochemical journal 433, 51 (Jan 1, 2011).

[30] X. Tang, Y. Zhang, L. Tucker, B. Ramratnam. Nucleic acids research 38, 6610 (Oct, 2010).

[31] G. Yeo, C. B. Burge. J Comput Biol 11, 377 (2004).

[32] M. Somel et al. Genome research 20, 1207 (Sep, 2010).

[33] G. Yeo, D. Holste, G. Kreiman, C. B. Burge. Genome Biol 5, R74 (2004).

[34] D. Parkhomchuk et al. Nucleic Acids Res 37, e123 (Oct, 2009).

[35] C. Trapnell, L. Pachter, S. L. Salzberg. Bioinformatics 25, 1105 (May 1, 2009).

[36] A. Roberts, H. Pimentel, C. Trapnell, L. Pachter. Bioinformatics 27, 2325 (Sep 1, 2011).

[37] B. Paten, J. Herrero, K. Beal, S. Fitzgerald, E. Birney. Genome research 18, 1814 (Nov, 2008).

[38] Y. Katz, E. T. Wang, E. M. Airoldi, C. B. Burge. Nat Methods 7, 1009 (Dec, 2010).

[39] R. Berretta, P. Moscato. PloS one 5, e12262 (2010).

[40] O. Martinez, M. H. Reyes-Valdes. Proceedings of the National Academy of Sciences of the United States of America 105, 9709 (Jul 15, 2008).

[41] K. Lindblad-Toh et al. Nature 478, 476 (Oct 27, 2011).

[42] Z. Wang et al. PLoS biology 8, e1000530 (2010).

[43] G. W. Yeo et al. Nature structural & molecular biology 16, 130 (Feb, 2009).

[44] E. L. Huttlin et al. Cell 143, 1174 (Dec 23, 2010).

[45] M. Buljan et al. Molecular cell 46, 871 (Jun 29, 2012).

[46] R. Knight et al. Genome biology 8, R171 (2007).

Chapter 4

Origins and impacts of novel exons in mammalian genes

Author contributions:

J.M. and C.B.B. designed the study and wrote the manuscript. J.M. conducted computational analyses and prepared figures. P.C. conducted computational analyses under supervision of J.M. and prepared figures.

4.1 Abstract

Though the exon-intron boundaries are generally well conserved within mammals (1), there are many splicing differences between species that result in lineage-specific exons (49, 50, 2, 62, 39, 44). To assess species-specific exons in mammals, we analyzed cDNA from 9 tissues from 4 mammals and one bird in biological triplicate. We find that species-specific exons frequently arise from pre-existing intronic sequence. We further find that changes in intron length are associated with splicing of new exons, and that these changes may impact splicing through altering nucleosome positioning

and RNA PolII kinetics. Our data also indicate that since the splicing of novel exons often increases gene expression, changes in splicing between species may explain some species-specific changes in gene expression.

4.2 Results and Discussion

4.2.1 Identification of recently created exons

We recently studied the splicing of ancient (mammalian-wide) alternatively spliced exons using polyA-selected RNA-seq analysis of 9 diverse organs and tissues from 4 mammals and one bird, in biological triplicate, yielding ~1000-fold aggregate coverage of each transcriptome (38). Here, we used these data (54) in combination with whole-genome alignments (41, 7, 46, 22) to classify exons as species-specific, lineage-specific (e.g., unique to rodents or to mammals), or ancient (present in both mammals and birds) at both the genomic sequence level ("genomic age") and at the level of expression ("splicing age") (Fig. 4-1A). Using the principle of parsimony, we assigned both genomic and splicing ages to ~60,000 internal exons, restricting our analysis to unduplicated protein-coding genes in these species to facilitate read mapping and orthology assignment. Throughout our analyses, genomic age reflects the duration over which sequences similar to the exon are present in the genome, while splicing age reflects the duration over which these sequences are represented as spliced exon in RNA-seq data.

Approximately 85% of exons in the analyzed genes predated the split between bird and mammals (~300 million years ago, Mya) in their splicing age, designated `MRQCG` using a one-letter code of organisms (Fig. 4-1B). However, we also found many occurrences of the creation of novel exons during mammalian evolution. For example, 1089 mouse exons were classified as mouse-specific (designated `M----`), as they were detected in RNA-seq data from mouse but from no other species (Fig. 4-1B, also (39, 62, 2)). These exons were assigned an age of <25 My, corresponding to the time

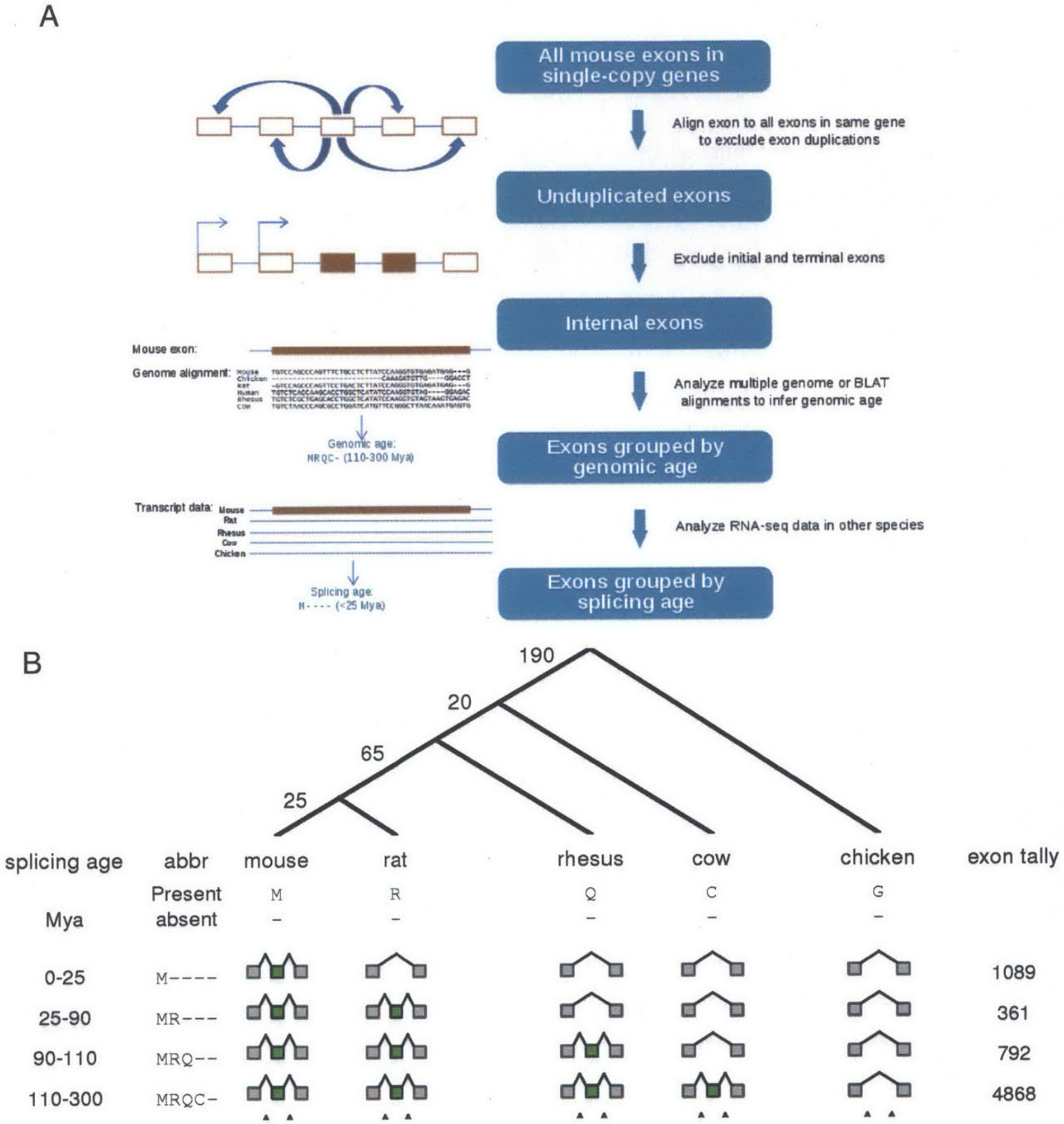


Figure 4-1:

(a) A schematic diagramming our bioinformatic pipeline to identify novel exons (Methods). We considered every exon in the target species (here, mouse) and aligned it to other exons in the same gene to filter out exons arising from exon duplication (14). We filtered out terminal exons to focus our analyses on splicing. We used multiple alignments between species studied here to assign an orthologous region to each exon in other species and used parsimony to interpret the pattern of genomic presence or absence as the genomic age. We then looked for an overlapping exon in the orthologous region to determine if the mouse exon was spliced in a given species, and interpreted this pattern of presence or absence of splicing as a splicing age.

(b) Top: a phylogenetic tree presenting the main species used for dating exons and the branch lengths in millions of years. Bottom: exons of increasing evolutionary splicing age, their pattern of presence or absence in various species, and the number of each class of exons identified.

of divergence between mouse and rat. We identified ~7000 mouse exons whose splicing was restricted to particular mammalian lineages (Fig. 4-1). Overall, presence of one or more exons were detected in mouse and not primates in about 17% of the ~6300 genes analyzed that passed filters. To ask whether species-specific exons occurred at a similar frequency in human, we compared our data to corresponding tissue data from the Illumina Human Body Map 2.0 dataset (6). Although the human data had lower sequencing depth, we identified similar numbers of exons at each splicing age (Fig. 4-2), including about 2300 exons found in human but not mouse in about 25% of analyzed human genes. Together, these observations indicate that, when comparing a human gene to its ortholog in the most commonly used mammalian model, the mouse, almost ~40% of ortholog pairs will differ by presence/absence of a lineage-specific exon. The prevalence of species-specific exons could contribute to widespread functional differences between human and mouse orthologs, complicating extrapolations from mouse models.

To assess whether species-specific exons have defined tissue-specific splicing patterns, we performed clustering of exons and tissue samples based on the tissue-specific splicing patterns of mouse-specific alternative exons. This analysis revealed robust clustering by tissue of origin across the three mouse strains analyzed (Fig. 4-4F). The only deviation from this pattern was some overlap between cardiac and skeletal muscle, consistent with similarity between the splicing programs of these developmentally related tissues, as seen when considering ancient alternative exons (38). A substantial fraction of novel exons showed predominant inclusion in testis (Fig. 4-4F), similar to the testes-biased expression observed for novel, species-specific, genes (31, 20). Similarly, we found that genes containing novel constitutive exons are enriched for testis expression (Fig. 4-3). These observations suggest a potential role of germ cell transcription in exon creation, consistent with previous studies indicating that increased transcription in germ cells can increase the frequency of mutations (43), presumably including those that give rise to novel exons.

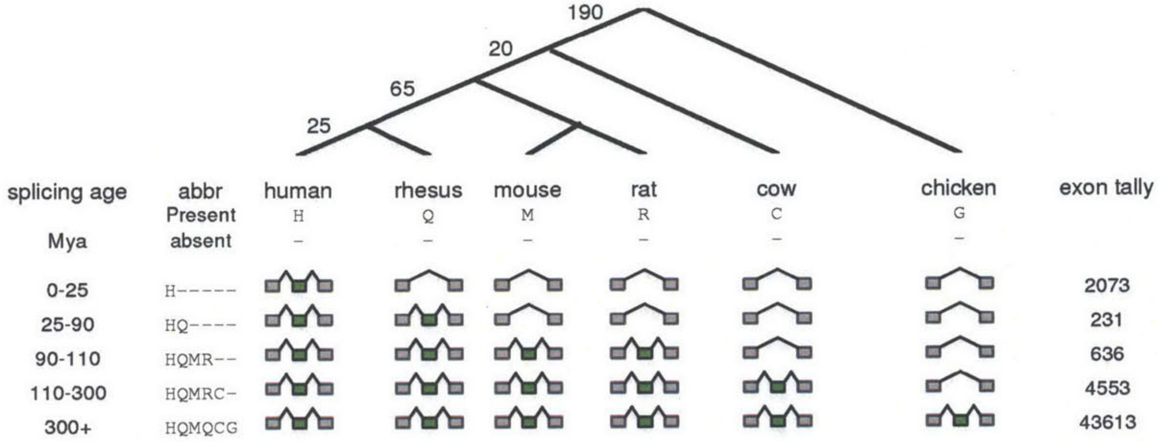


Figure 4-2:
Top: a phylogenetic tree presenting the main species used for dating exons and the branch lengths in millions of years. Bottom: human exons of increasing evolutionary splicing age, their pattern of presence or absence in various species, and the number of each class of exons identified.

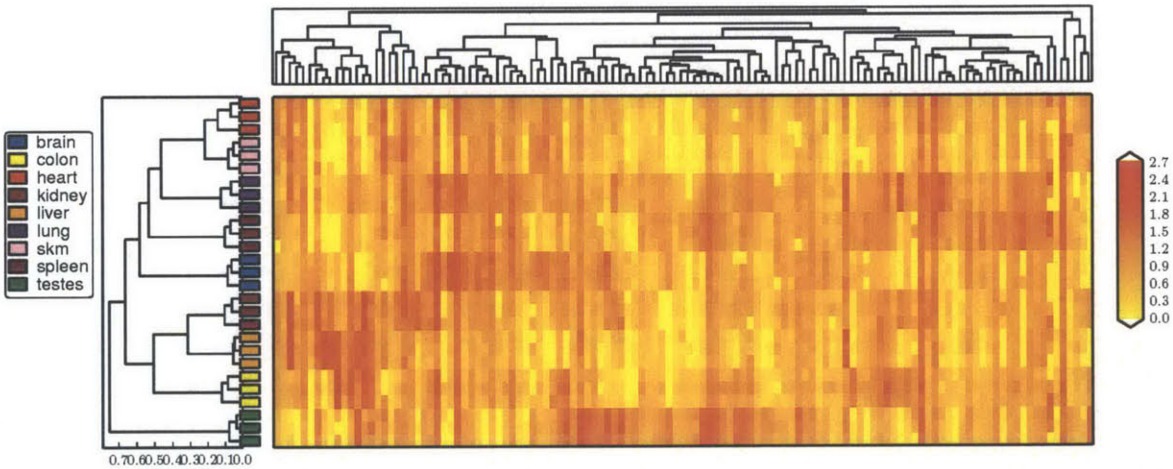


Figure 4-3:
Clustering of genes containing a mouse-specific constitutive exon based on gene expression. Heatmap is shown on a log scale, but clustering is done on untransformed values.

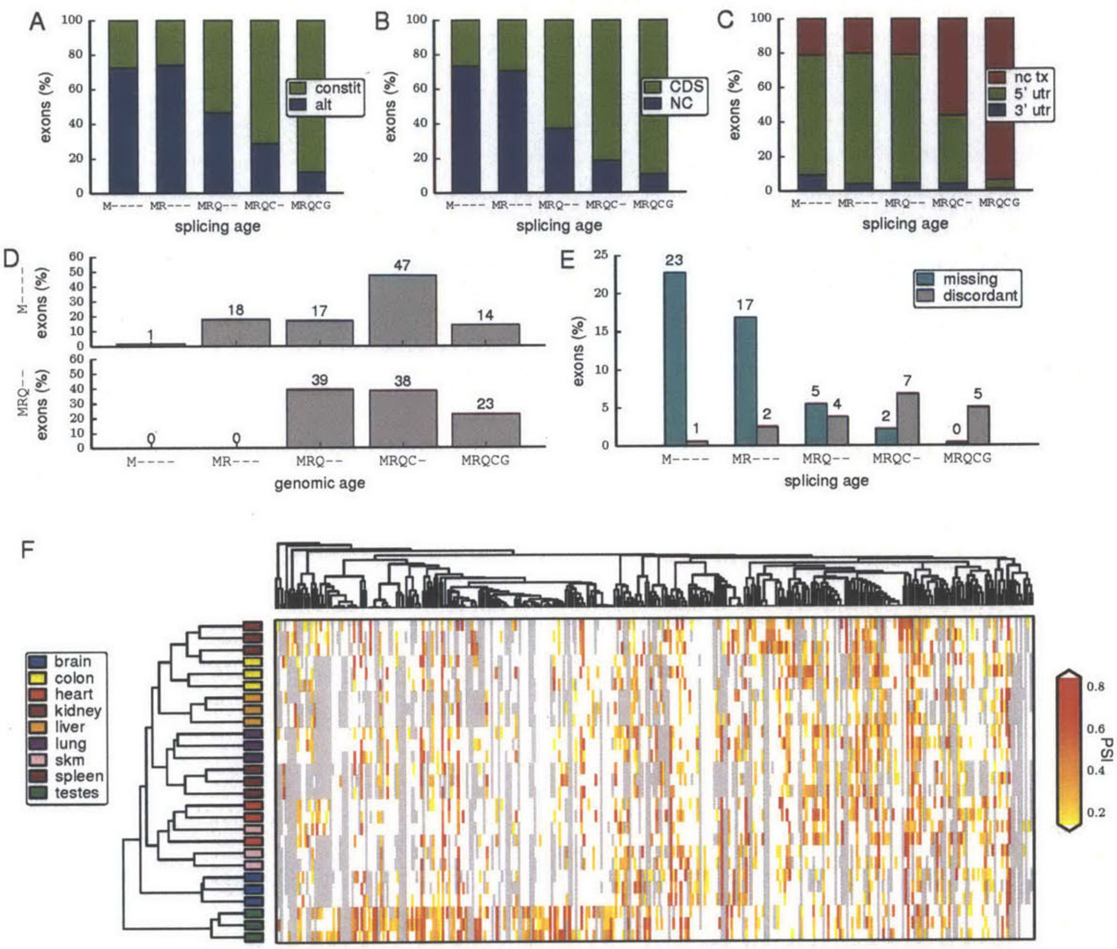


Figure 4-4:

(a) The proportion of exons of various ages that are alternatively or constitutively spliced.

(b) The proportion exons of various ages that contain coding sequence.

(c) The proportion of non-coding exons in various transcript regions within exons of various ages.

(d) The distributions of genomic ages of `M----` or `MRQ--` exons are plotted.

(e) The proportion of mouse exons of various ages that are missing in one individual or where the splicing status (skipped or constitutive) in one individual disagrees with the other two mice.

(f) Average-linkage hierarchical agglomerative clustering of samples (vertical axis) or exons (horizontal axis) based solely upon PSI values of mouse-specific exons. The tissue of origin of each sample is colored according to the key at left and the PSI value is visualized in the heatmap (center). Clustering is based upon tissues with gene expression $\geq 5FPKM$, but splicing is visualized in tissues with expression $\geq 2FPKM$ to enhance visualization. Tissues meeting the expression filters but where the exon completely excluded (PSI = 0) are white.

4.2.2 Unique properties of species-specific exons

In many respects, exons of different evolutionary ages had dramatically different properties. While constitutive splicing was the norm for ancient (MRQCG) exons in these data, the vast majority of species-specific exons were alternatively excluded (skipped) in at least one tissue (Fig. 4-4A, (39)). Similarly, ancient exons were mostly located within the open reading frame (ORF), while most species-specific exons were located in non-coding regions (Fig. 4-4B). New exons occurred with much higher frequency in 5' untranslated regions (UTRs) than in 3' UTRs (Fig. 4-4C, also (9, 47). Various factors may contribute to the bias for 5' UTRs, including the greater length of first introns relative to later introns (25, 44), the low frequency of 3' UTR introns (15) and the increased potential for some new 3' UTR exons to trigger mRNA decay via the nonsense-mediated mRNA decay (NMD) pathway. By contrast, the non-coding subset of ancient exons were located predominantly in non-coding transcripts in otherwise coding loci (Fig. 4-4C).

4.2.3 Most species-specific exons arise by exaptation of intronic sequences

Our classification pipeline used sequence similarity filters to exclude species-specific new exons that arose from internal gene duplications, a class that has been well studied previously (14). Therefore, the novel exons studied here must have arisen by insertion of novel sequence into an intron (30) or by exaptation of pre-existing intronic sequence (8). To compare the relative contributions of these two mechanisms, we analyzed the genomic age of each recently created exon. Approximately 1% of mouse-specific exons arose in sequence found in only mouse while nearly 75% of these exons derive from sequence that predates the rodent-primate split (despite being recognized exclusively as intronic in the other species studied) and the remainder were alignable to rat only (Fig. 4-4D). Rodent-specific and rodent/primate-specific exons also could often be aligned to cow or chicken (Fig. 4-4D and data not shown). We hypothesized that since M---- exons recently acquired splicing activity, their splicing pattern may be polymorphic. Using our RNA-seq data derived from three different mouse strains, we observed that nearly a quarter of mouse-specific exons were detected in just 2 of the 3 strains, while for ancient exons nearly 99% were detected in all 3 mouse strains. This observation indicates that that novel mouse exons might be much more frequently variable in their splicing than ancient exons, suggesting that extended periods of population level variation (likely in the millions of years) may precede fixation of novel exons in the mouse strains analyzed here.

We next sought to identify features associated with new exon creation. We observed that more than 60% of new internal exons in mouse are derived from unique intronic sequence. In most cases, these exons aligned to sequences in the orthologous intron in rat (Fig. 4-5A). Using a similar approach to identifying novel exons in human — with criteria designed to allow mapping to repetitive elements (Methods) — yielded a similarly high proportion of unique mapping (about 54%) (Fig. 4-6A). Alu elements, a class of primate-specific SINE repeats, have previously been implicated as a major source of new exons in primates (30, 50). Here, we found that about 19% of exons we classified as human-specific overlap with Alus (Fig. 4-6), and a similar proportion of mouse-specific exons (about 18%) overlap with rodent SINEs, which are also thought to derive from 7SL RNA (Fig. 3B). Thus, our analysis indicated that SINEs contribute to new exon creation to a similar extent in rodents as Alus do in primates. Although these proportions exceeded the genomic background frequencies of SINEs in these species (Fig. 4-5C; Fig. 4-6B), the proportion of species-specific exons derived from unique genomic sequence was 2- to 3-fold higher than SINEs in both organisms, contrasting with previous suggestions that Alu elements may be a predominant source of new exons in primates. The differences in conclusions likely result from differences in data sources (RNA-seq versus EST) and analytical procedures, making our analysis less biased and more sensitive to detection of low-abundance isoforms (60). Other types of repetitive elements (LINEs, LTRs and others) together contribute a similar proportion of species-specific exons as SINEs in both human and

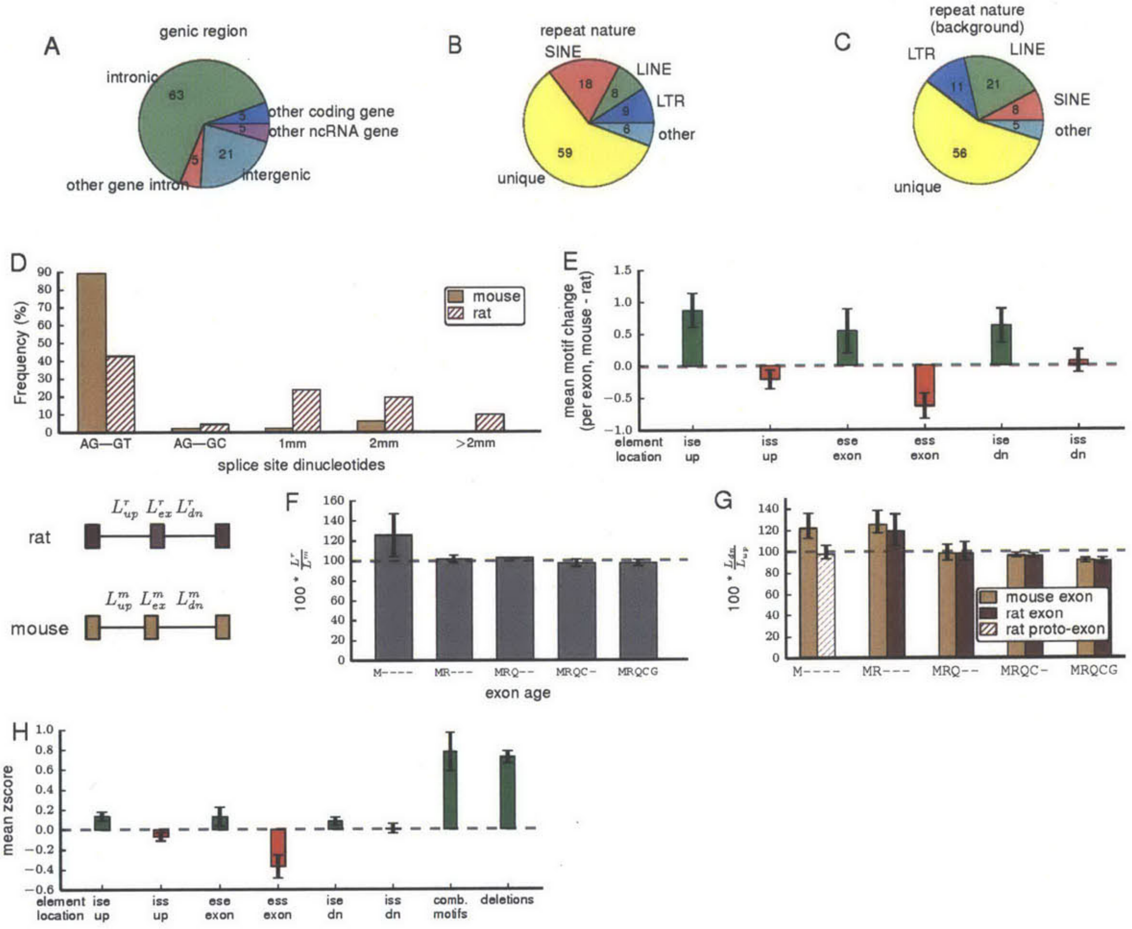


Figure 4-5:

(a) Proportions of new exons that map to various genomic regions in rat.

(b) Proportions of new mouse exons that belong to various repeat categories.

(c) Proportions of genome that belong to various repeat categories.

(d) The proportion of new mouse exons with a given splice site dinucleotide sequence in mouse and rat.

(e) The change in splicing regulatory element number in various regions in and around a new exon associated with its creation (mean ± SEM).

(f) The change in length of the entire intron region between rat and mouse. The length in rat is plotted as a percentage of the length in mouse (mean ± SEM).

(g) The relative length of the downstream intron as a percentage of the upstream intron (mouse) or the downstream aligned intron/region as a percentage of the upstream aligned intron/region (rat) (mean ± SEM). The rat bar in the M---- class is hatched to represent the fact that it is not an exon

(h) The magnitude of each change associated with splicing of M---- exons is converted into a z-score based upon the distribution of such changes between mouse and rat in MRQCG exons. Changes that are expected to promoter splicing are colored in green and changes that are expected to block splicing are red. For the purpose of summing (penultimate bar), the sign of changes that inhibit splicing was reversed so that positive z-scores promote splicing in all cases.

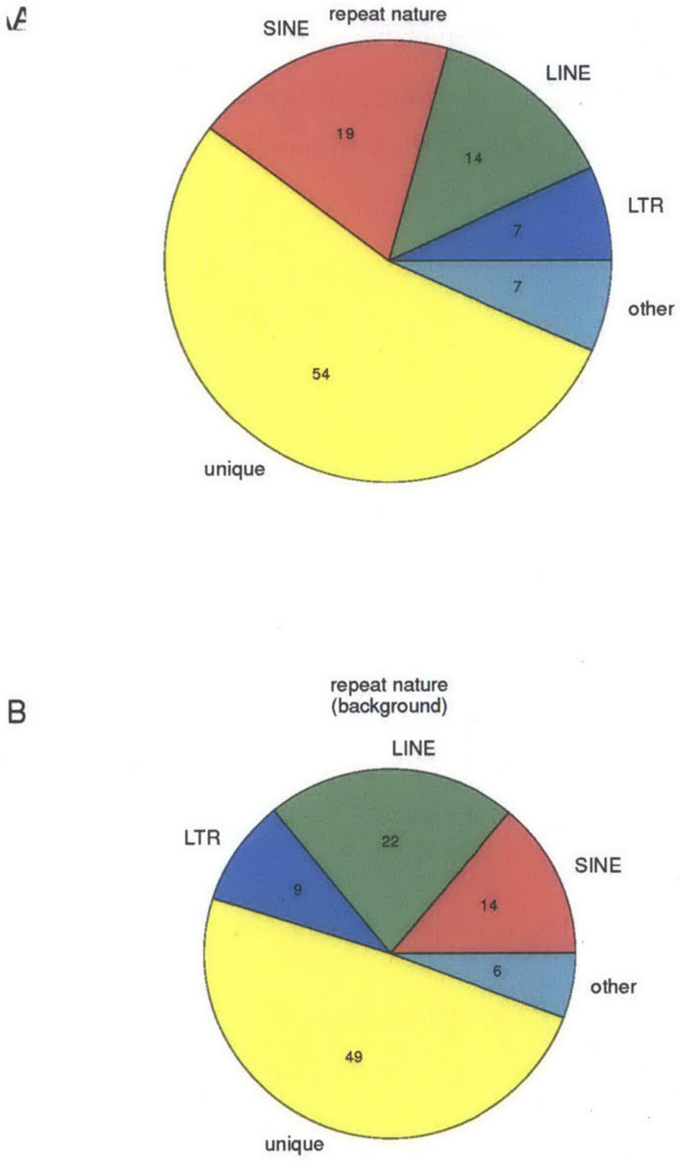


Figure 4-6:

(a) Proportions of new human exons that belong to various repeat categories.

(b) Proportions of human genome that belong to various repeat categories.

mouse (Fig. 4-5B, Fig. 4-6A).

Mutations that create or disrupt splice site motifs frequently cause changes in splicing patterns over evolutionary time periods (30, 6). While the vast majority of mouse-specific exons contained consensus splice site motifs (GT or GC at the 5' splice site, AG at the 3' splice site), about half of homologous "proto-exon" sequences in rat lacked these minimal splicing motifs (Fig. 4-5D). This observation suggests that mutations that create minimal splice sites may contribute to up to ~60% of exon creation events. Further, about 40% of `M----` exons have minimal splice sites in rat without evidence of their splicing detected in rat tissues, implicating other types of changes in their creation.

Motifs present in the body of an exon or in the adjacent introns can enhance or suppress exon inclusion (36). We found that mouse-specific exons contain a higher density of exonic splicing enhancer (ESE) motifs and a lower density of exonic splicing silencer (ESS) motifs than their associated rat proto-exons (Fig. 4-5E). Thus, both the gain of enhancing motifs and the loss of silencing motifs may contribute to creation and/or maintenance of novel exons. Changes in flanking sequences may also contribute to exon creation, as a higher density of intronic splicing enhancer (ISE) motifs was observed adjacent to mouse-specific exons relative to homologous rat sequences; no difference in intronic splicing silencers (ISSs) was observed (Fig. 4-5E).

Intron length is correlated with a number of splicing properties, including lower levels of exon inclusion (61). We therefore asked whether changes in intron length might be associated with splicing of novel exons. Notably, we found that the distance between the exons flanking `M----` exons was shorter on average than the distance between the homologous exons in rat (rat distance exceeded mouse by 1.3-fold on average; interquartile range: 0.9-fold to 1.7-fold; Fig. 4-5F). The distance between the exons flanking `-R---` exons was even more biased toward shorter lengths than the distance between the homologous mouse exons (mouse distances exceeded rat by 1.7-fold on average; interquartile range 1.0 to 2.7-fold; Fig. 4-7). These observations suggest that substantial changes in intron length may often contribute to exon creation. Comparing the lengths of introns flanking each species-specific exon to each other, we observed that the intron downstream of `M----` exons was 1.3-fold longer on average than the upstream intron, compared to no difference for the homologous regions in rat (Fig. 4-5G). When examining rodent-specific exons, we observed a similar bias towards shorter upstream intron in both mouse and rat. Comparison to an outgroup (macaque) indicated that the differences in flanking intron length in the rodent lineages that acquired new exons most often reflect upstream deletions rather than downstream insertions (Fig. 4-8). Older groups of exons showed no such bias, suggesting that exons may acquire tolerance for expansion of the upstream intron over time, as their splice sites strengthen (Fig. 4-5F, Fig. 4-8). Together, these data suggest that deletions upstream of proto-exons may favor creation or maintenance of novel exons. While shortening of an upstream intron has been associated with enhancement of exon inclusion in a mini-gene context (12), the generality of this effect and its evolutionary impact have not been previously explored.

Having identified and evaluated a number of changes associated with new exon splicing, we then asked about the relative contributions of each. To compare the magnitudes of these different types of changes using a standard scale, we converted them all to z-scores, using the standard deviation of each type of change observed between mouse and rat in ancient (`MRQCG`) exons. We observed relatively small z-scores (<0.4) associated with any one cis-motif change, while upstream intronic deletions had an average z-score of ~0.75, comparable to the sum of the z-scores of all cis-motifs analyzed (Fig. 4-5H). This observation suggests that upstream intronic deletions may contribute to creation of novel exons to a substantial extent, comparable to that of changes in known classes of splicing regulatory elements.

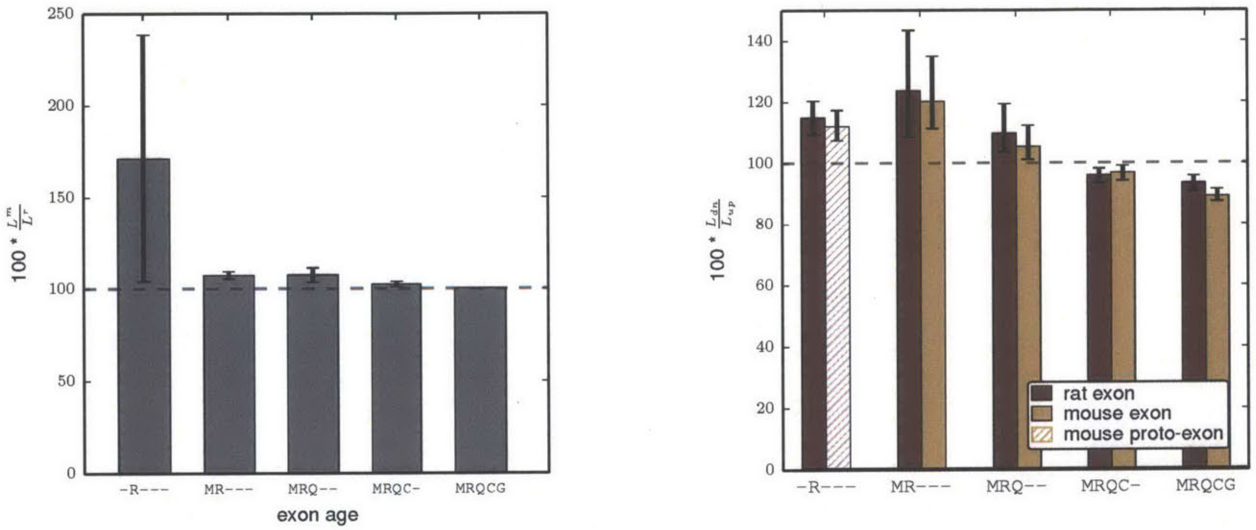


Figure 4-7:
As in Figure 4-5E and 4-5F, but in a rat-centric manner.

left The change in length of the entire intron region between rat and mouse. The length in rhesus is plotted as a percentage of the length in mouse (mean ± SEM).

right The relative length of the downstream intron as a percentage of the upstream intron (rat) or the downstream aligned intron/region as a percentage of the upstream aligned intron/region (rat) (mean ± SEM). The mouse bar in the `-R---` class is hatched to represent the fact that it is not an exon

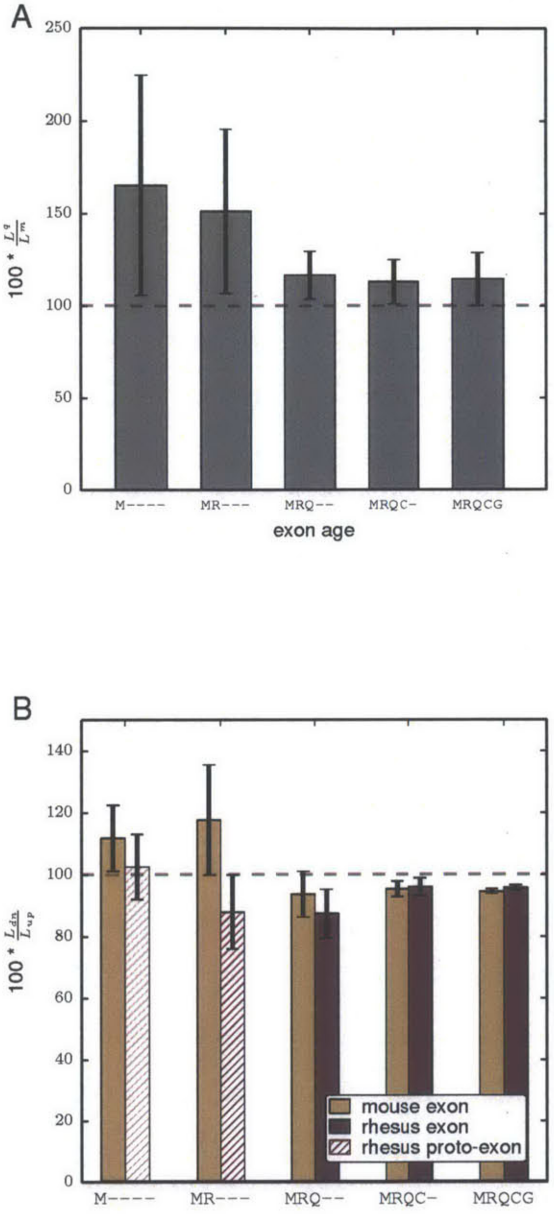


Figure 4-8:
Outgroup analysis showing that indels observed in Figure 4-5 associated with exon creation are generally deletions.

(a) The change in length of the entire intron region between rhesus and mouse. The length in rhesus is plotted as a percentage of the length in mouse (mean ± SEM).

(b) The relative length of the downstream intron as a percentage of the upstream intron (mouse) or the downstream aligned intron/region as a percentage of the upstream aligned intron/region (rhesus) (mean ± SEM). The rhesus bar in the M---- class is hatched to represent the fact that it is not an exon

4.2.4 Upstream indels are associated with increased nucleosome occupancy and RNA PolII pausing over novel exons

Intron length can impact splicing in multiple ways. Shortening introns can promote exon inclusion in splicing reporter experiments, with a larger effect observed in the upstream intron, possibly by promoting intron definition (4, 12). Lengthening of the downstream intron can impact splicing through effects on the dynamics of transcription (10, 18, 5). Intronic length changes might also promote exon creation by impacting nucleosome positioning. Nucleosomes are generally positioned near the centers of internal exons and exon-associated nucleosomes have higher density of the H3K36me3 modification (51, 45, 53). There are reports that histone modifications can impact recruitment of splicing factors, suggesting functional links between nucleosomes and splicing (34). We used micrococcal nuclease (MNase) sequencing data from digestion of chromatin from mouse embryonic stem cells to identify nucleosome-protected regions in the vicinity of mouse-specific exons. We observed a stronger enrichment for nucleosome positioning over mouse-specific exons which had deletions in the upstream intron (relative to rat) compared to ancient exons, mouse-specific exons without upstream deletions, and mouse regions orthologous to rat-specific exons (Fig. 4-9A). This association suggested a connection between upstream deletions and changes in nucleosome positions. While indels in either the upstream or downstream intron could potentially impact nucleosome positions, deletions in the upstream intron are more likely to favor intron definition and exon inclusion based on previous studies (12). Therefore, the bias for upstream deletions seems mostly likely to result from effects on intron definition, with a possible secondary contribution from effects on nucleosome positioning. Because the relationship between nucleosome positioning and splicing is less understood, we chose to further explore this potential connection.

It has been proposed that nucleosomes can function as molecular "speed bumps" to slow down RNA polymerases as they transcribe through exons (5). This effect may

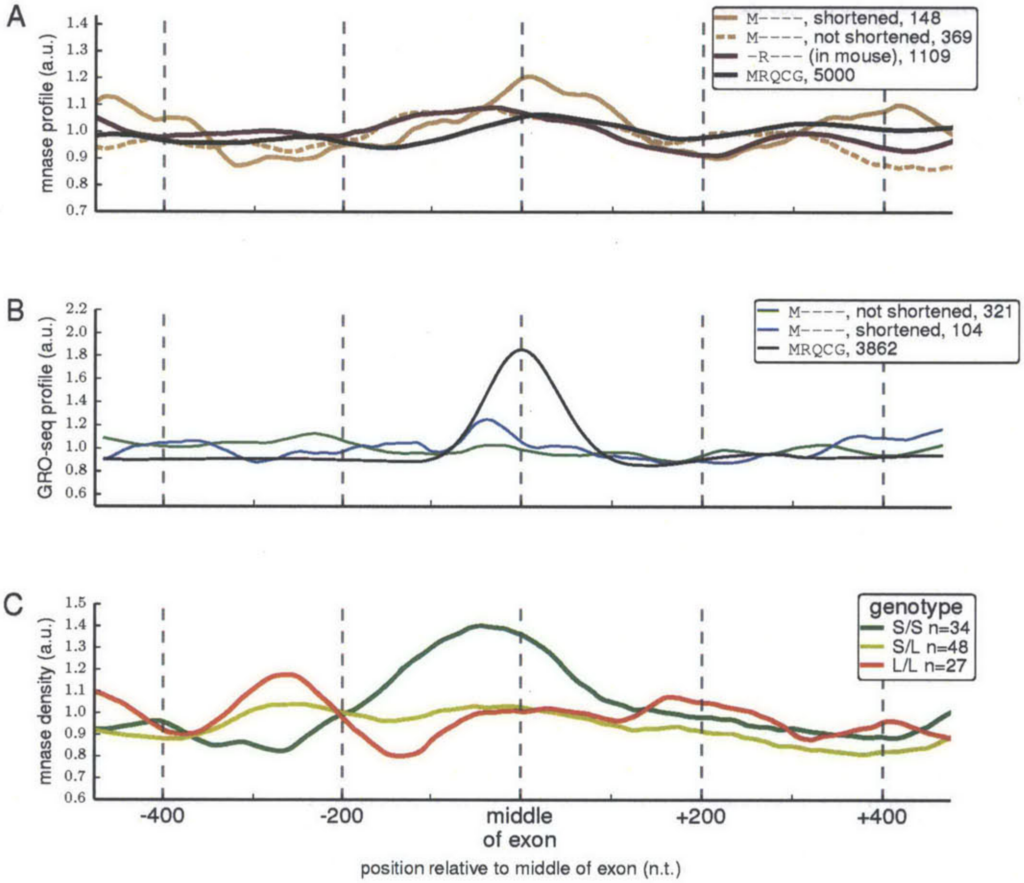


Figure 4-9:

(a) Nucleosome positioning (measured by protection from MNase treatment) around various sets of exons.

(b) Global run-on sequencing (GRO-seq) showing the position of elongating RNA PolII by sequencing nascent RNA around various sets of exons.

(c) Nucleosome positioning (measured by protection from MNase treatment) around exons with a structural sQTL in the upstream intron binned by sQTL genotype.

contribute to exon inclusion by increasing the time available for splicing machinery associated with the C-terminal domain (CTD) of RNA polymerase II to associate with the exon and commit it to splicing (16, 63). Mutations that slow down RNA polymerase elongation are reported to enhance recognition of exons with weak splice sites (18). We hypothesized that changes promoting stronger nucleosome positioning over novel exons might slow polymerase elongation and thereby act to promote splicing. To test this hypothesis, we used available global run-on-sequencing (GRO-seq) data, which detects nascent transcripts. Using data from (21) in mouse macrophages, we observed a strong GRO-seq peak over ancient MRQCG exons, approximately 93% over background (Fig. 4-9B), suggesting that polymerases slow down by almost a factor of two while transcribing through these exons. To our knowledge this is the first analysis showing that wildtype mammalian polymerases slow substantially over the bodies of exons. When considering mouse-specific exons with an upstream intronic deletion, we observed a GRO-seq peak about 37% above background while mouse-specific exons without upstream deletions lacked an appreciable peak. This difference might reflect increased polymerase pausing in mouse-specific exons with upstream deletions that result from the more strongly positioned nucleosomes observed in this group. Therefore, we propose that changes in chromatin and polymerase dynamics may represent an additional contributor to creation or maintenance of novel exons.

Recent studies looking at the genetic basis for gene expression variation have also identified thousands of genetic variants associated with altered levels of splicing between human individuals (42, 26, 29). We reasoned that we may observe a similar effect of structural variants affecting splicing regulation. Specifically, we hypothesized that splicing quantitative trait loci (defined as genetic variants associated with variation in splicing, "sQTLs") that are structural variants located in the intron upstream of their associated splicing events (similar to the changes we infer to have occurred associated with exon creation) might affect nucleosome localization over the exons they are associated with. To test this hypothesis, we used sQTLs identified in genotyped human lymphoblastoid cell lines studied by the GEUVADIS Consortium (29) and

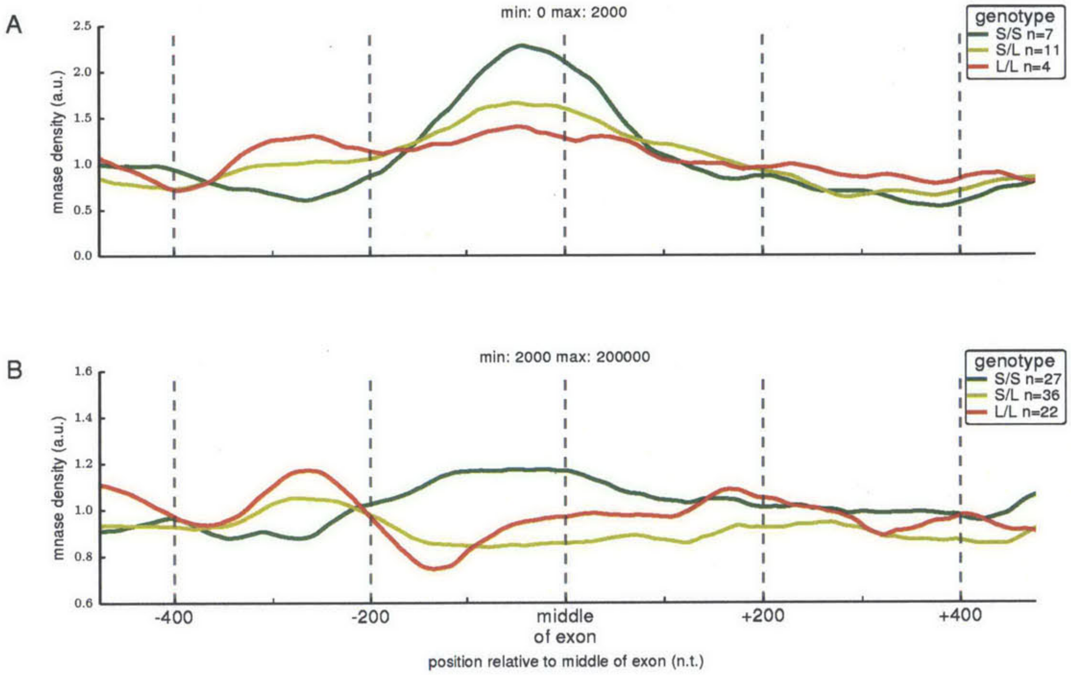


Figure 4-10:

(a) Nucleosome positioning (measured by protection from MNase treatment) around exons with a structural sQTL <2kb from the target exon in the upstream intron binned by sQTL genotype.

(b) Nucleosome positioning (measured by protection from MNase treatment) around exons with a structural sQTL >2kb from the target exon in the upstream intron binned by sQTL genotype.

MNase-seq data from a subset of these individuals (13). This allowed us to evaluate the nucleosome positioning across individuals with either the reference allele or the alternate longer or shorter intronic alleles. We considered structural variant sQTLs of 4 or more nucleotides in the upstream intron (corresponding to approximately half the period of nucleosome-associated dinucleotide frequency (13)), corresponding to about 50 events for which we also had MNase data representing at least two genotypes. As both insertions and deletions would be expected to alter nucleosome occupancy, we compared MNase data for individuals containing the shorter intronic allele to MNase data for individuals containing the longer allele and found that, in aggregate, the shorter variant was associated with stronger nucleosome positioning over the exon (Fig. 4-9C). We observed an effect in which stronger differences in nucleosome density were associated with indel sQTLs located closer to the affected exon, as expected if these indels directly impact nucleosome placement (Fig. 4-10). Together, the data in Figure 4-9 implicate upstream intronic indels in changes in nucleosome positioning and splicing between both species and individuals.

4.2.5 New exon splicing is associated with species-specific increases in expression

We next asked what effects new exons have on the genes in which they arise. Since the majority of species-specific exons we identified were non-coding (Fig. 4-4B), we examined gene expression. Intron-mediated enhancement is a well-characterized, though incompletely understood, phenomenon in which introduction of a (heterologous) intron or exon into a gene or mini-gene often leads to higher expression of the gene (35, 40, 19). During evolution, creation of a novel exon in an intron will increment the number of introns and exons in a transcript. Here, we observed significantly higher expression (in mouse) of genes containing mouse-specific exons relative to their orthologs in corresponding tissues in rat (Fig. 4-11A). This effect was specific to those mouse tissues where the new exon was included, consistent with a positive effect of splicing on steady state expression levels (Fig. 4-11A). The inclusion of a new exon was associated with an average increase in gene expression of about 10% (Fig. 4-11A, inset).

An alternative way to measure the effect of the splicing of a new exon on expression is to compare expression in tissues where the exon is included to expression in tissues where the exon is excluded (the "exon-associated expression index", EEI) in the species containing the exon to the EEI calculated in a species where the exon is not present. Under the null hypothesis that the splicing of the new exon does not affect gene expression, the ratio of these EEI values ($EER_{mouse} = EEI_{mouse}/EEI_{rat}$) should be distributed symmetrically around one. This approach controls for a number of technical factors that could impact estimation of expression levels in different species. Comparing EER values for genes containing mouse-specific exons or rat-specific exons to shuffled controls (Fig. 4-11B), we observed significantly elevated ratios (~1.1) in both cases, consistent with the 10% increase in expression observed above (Fig. 4-11A), and further supporting the alternative hypothesis that splicing in of new exons enhances gene expression.

These observations suggest a widespread impact of splicing on gene expression. To further explore this phenomenon, we considered exons whose presence in the transcriptome is ancient, but that have recently acquired skipping in mouse (38). We observed that the species-specific skipping in mouse of these exons was associated with lower gene expression (relative to rat), suggesting that the relative loss of one splicing reaction in these genes in mouse may therefore me lowering its gene expression in that species (Fig. 4-11C). Furthermore, we observed a dose-dependent effect, where inclusion of these exons with low PSI values was associated with a larger decrease in gene expression than higher PSI values (Fig. 4-11D). Additionally, this effect is dominated by internal exons located in the 5' end of the gene, with virtually no change in gene expression associated with mouse-specific skipping of internal exons located in the 3' end of the gene (Fig. 4-11E). These results suggest that the act of splicing has genome-wide effects on gene expression and that the splicing of new

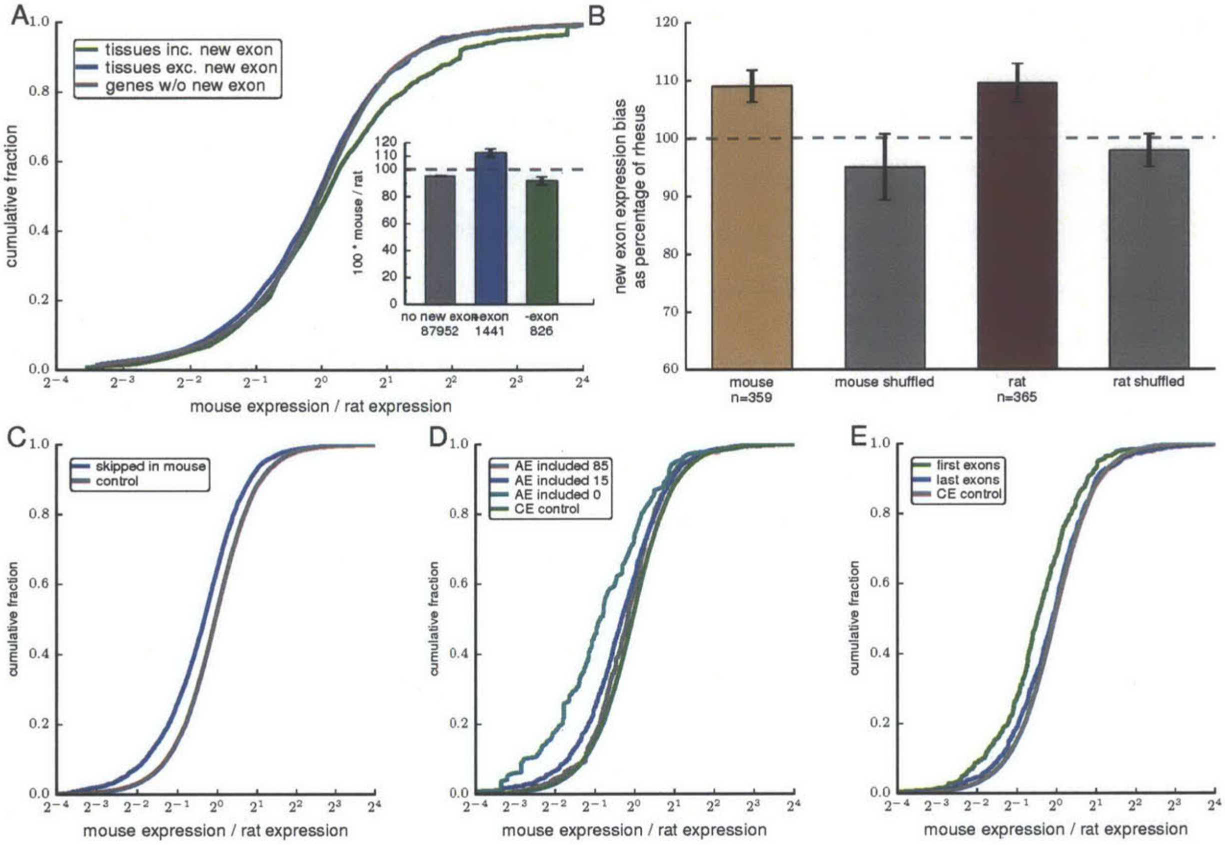


Figure 4-11:

(a) Fold change in gene expression between mouse and rat. Inset: mean ± SEM of displayed distributions.

(b) The mean expression in tissues where a novel exon is included is divided by the mean expression in tissues not containing the exon. This ratio is calculated in a related species with matched tissues, and the ratio of these two values is plotted (mean ± SEM).

(c) Fold change in gene expression between mouse and rat in genes where an old exon has become skipped in mouse.

(d) Fold change in gene expression between mouse and rat in genes where an old exon has become skipped in mouse, binned by the PSI of the exon in the tissue.

(e) Fold change in gene expression between mouse and rat in genes where an old exon has become skipped in mouse, binned by location of the exon within the gene.

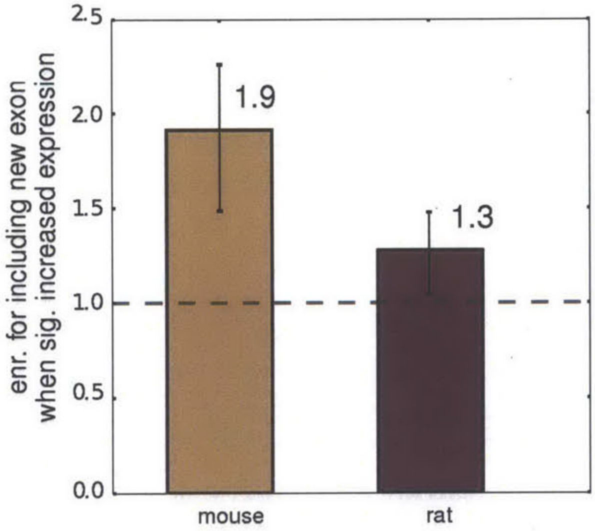


Figure 4-12:
The fraction of genes with a significant expression change at an FDR of 0.1% that contain a novel exon is compared to the fraction of genes without a significant expression change that contain a novel exon in either mouse or rat.

exons therefore can increase gene expression in a tissue- and species-specific manner. We note that such observations are subject to potential detection biases (58). However, the detection bias when considering species-specific presence and skipping of exons would both require higher expression (compare with the lower expression observed associated with species-specific skipping of old exons in Fig. 4-11C-E). We can therefore conclude that splicing has a global role in promoting gene expression and that changes in splicing are potentially a major cause of changes in gene expression between species. Supporting this, we found that species-specific increases in gene expression were enriched in genes containing species-specific exons (Fig. 4-12), suggesting that new exons may be a major contributor to changes in gene expression between species.

4.3 Methods

RNA-seq and genome builds Data from from mouse, rat, rhesus, cow, and chicken were processed as in (38) using TopHat v1.1.4 (54) and Cufflinks v1.0.2 (55). Reads were initially allowed to map to up to 20 positions in the genome by TopHat to further assist in the detection of repeat-associated exons. Mouse data were mapped to mm9, rat data to rn4, rhesus data to rhemac2, cow data to bostau4, and chicken data to galgal3.

Assignment of ages to exons Exons from each species (mouse, rat, rhesus, cow, chicken) from (38) were used in this analysis. As done in that study, we only considered single copy genes. We flagged and removed terminal exons and focused only on internal exons from these genes. We filtered internal exon duplications (14) by aligning each exon to other exons in the same gene. Aligned regions in other species for each query exon were collected based on whole genome alignments generated by PECAN and EPO (41) and pairwise alignments from BLASTZ (?). In addition, we further attempted to align exons without a genomic aligned region not expressed in chicken to the genome of each species using BLAT (22) to reduce a false negative rate of finding an aligned region, taking the best matching region and requiring a minimum of 80% identity for alignment to rat, 66% identity for alignment to rhesus, 65% identity for alignment to cow, and 54% identity for alignment to chicken. These thresholds were calculated by taking value 3 standard deviations below the average percentage identity of exons between the query species (mouse) and the other species in question.

An exons genomic age was defined based solely on the pattern of species with genomic regions aligned to the query exon. We interpret this pattern using parsimony (39, 62, 2), which is reasonable given the small number of events (presence or absence of an aligned region). We consider the minimum number of changes that can explain the pattern of aligned regions when compared with a precomputed species tree. We only consider unambiguous age assignments (i.e. if there are two equally compatible interpretations that would yield different ages, then we do not consider the exon in these analyses). An exons splicing age was assigned in a similar manner to the genomic age, only it was based the pattern of presence or absence of an expressed region (i.e. an exon) in the orthologous gene overlapping the genomic aligned region.

For example, a mouse exons genomic age was assigned to 0-25 (new), 25-90, 90-110, 110-300 and 300+ if there were aligned regions in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken.Similarly, its splicing age was assigned to similar categories if there were aligned regions expressed (i.e. exons included in processed transcripts) in rat, rat/rhesus, rat/rhesus/cow and rat/rhesus/cow/chicken (Fig. 4-1B).

Note: we only considered exons detected in the previous RNA-seq study (38). This was done to mitigate the effects of prior transcript annotation quality on our results since, for instance, mouse and rhesus annotations (by proxy from converting human annotations) would be expected to be much better than cow or rat. This approach will miss annotated exons only included in embryonic tissues, for instance, but those would likely have been incorrectly assigned to the novel, recently created, exon category due to the possibility of their not being found in other species because we dont have comparable data.

Basic exon properties Exons with PSI >0 and PSI <97 (where PSI represents the Percent Spliced In, or the percentage of transcripts in a particular tissue estimated to include the exon in question (58)) in at least 1 tissue were categorized as skipped exons (SE) while exons with PSI >97 in all expressed tissues were defined as constitutive exons (CE) for each individual. We required an exon be evaluated in 3 or more tissues for this classification, since the probability of detecting exon skipping increases with the number of tissues considered. In Fig. 4-4A, the proportion of exons that are skipped or constitutive were calculated by SE/(SE+CE) and CE/(SE+CE) respectively, where SE and CE represents the number of alternative spliced exons and the number of constitutive exons.

Transcripts' open-reading frames (ORFs) were annotated as in (38). Briefly, if a transcript contained an annotated translation start site, then the longest ORF originating from that site was used. If no such site was contained in the transcript, then the longest ORF 100 amino acids or longer was used. If none existed, then the transcript was considered non-coding. Exons that can map to transcripts' ORF region, upstream and downstream region of transcript ORF, and regions in transcripts without ORF were categorized as coding exons, 5'UTR and 3' UTR exons and non-coding exons, respectively. In Fig. 4-4B, the proportion of coding exons were calculated by coding exons / total, where coding is the number of coding exons and total is total counts of exons at each age.

Genomic sources of new exons We traced the origins of new exons by allocating the genomic locations of aligned regions in the closest species (for example, in mouse, we used rat as its closest species). In Fig. 4-5A, exons were categorized into "intronic", "intergenic", "other coding gene", "other intron" and "other ncRNA gene" if their aligned regions in the closest species are located in the intronic regions of the same gene, intergenic regions which does not overlap any gene, exonic regions of other genes, intronic regions of other genes and other regions of ncRNA, respectively.

The origin of new exons were also categorized based on the repeated sequences. The RepeatMasker (48) track was downloaded from the UCSC browser and used to identify repeats overlapping each exon. Exons were categorized as containing SINEs, LINEs, LTRs, or other repeats (poorly sampled categories with low counts). Exons not overlapping any repeats were assigned to the "unique" group in Fig. 4-5B.

Splice site and splicing regulatory element analysis The dinucleotide frequencies of the intronic 5' and 3' splice sites of mouse new exons and their aligned regions in rat were compared in Fig. 4-5C. In Fig. 4-5D, exonic splicing enhancers (ESEs) from (11), exonic splicing silencers (ESSes) from (59), intronic splicing enhancers (ISEs) from (56), and intronic splicing silencers (ISSes) from (57) were used. The 100nt of intronic sequence upstream and downstream of each exon in mouse or the aligned region in rat was considered for searching for intronic splicing regulatory elements. The entire exon was searched for exonic splicing regulatory elements. To control for differences in exon length, the average frequency of such changes were multiplied by the average new exon length to arrive at the average change per exon.

Intron length analyses For each exon age, the lengths of each mouse exon and its upstream and downstream introns were compared to the corresponding sum in rat by summing the lengths of the rat exon (or aligned region for mouse-specific exons) and the surrounding introns (Fig. 4-5E). For Fig. 4-5F, the length downstream mouse intron was divided by the length of the upstream mouse intron. A similar ratio was calculated in rat, where the downstream intron (or the remainder of the intron downstream of the aligned exon region for mouse-specific exons) was divided by the upstream intron (or upstream remainder of the intron).

Z-score conversion for comparisons For each change considered (changes in ISEs, ISSes, ESEs, ESSes, or deletions), the empirical distribution of such changes in the ancient set of exons `MRQCG` was determined. The mean and standard deviation of this distribution was calculated. For each new exon, Each change was then calculated for each new exon and converted to a z-score using the values calculated in the ancient group.

Nucleosome localization and GRO-seq analyses We downloaded the MNase-seq data from (52) from GEO (accession GSE40910). We mapped the reads with Bowtie v0.12.7 (27) to mm9. We considered ancient (`MRQCG`) exons, new mouse exons with no upstream intron deletion, new mouse exons with an upstream intron deletion, and the orthologous region of new rat exons. We used pysam v0.7.7 and samtools v0.1.16 (33) to count the number of reads in a 1kb window of each exon. Each exons profile was internally normalized, and the average profile of each set of exons was smoothed with a sliding window and plotted, centered on the exon midpoint or 3' splice site.

We downloaded the GRO-seq data from (21) from GEO (accession GSE48759). We combined the various samples to increase our power. While the transcriptional level of a particular gene in each condition may be different, since we focused on internal exons and internally normalize each region, this should not affect our results. These data were then processed in the same manner as the MNase-seq data.

To investigate the impact of intronic structural variants on nucleosome localization (Fig. 4-11C), we downloaded the following files:

- the sQTL table EUR373.exon.cis.FDR5.all.rs137.txt.gz from the GEUVADIS consortium (29),
- Gencode v12 (17) matching the annotations used in the GEUVADIS study,
- MNase-seq data from individuals included in the GEUVADIS study from (13),

- Genotype data for these individuals from the 1000 Genomes Project (23) tables ALL.chr**.phase1_release_v3.20101123.snps_indels_svs.genotypes.vcf.gz, where ** represent different chromosomes,
- GRCh37.remap.all.germline.gvf from (28) for determining variant lengths.

The MNase-seq data were processed as described above. We filtered out all SNV sQTLs, as well as any indel or structural variant that was smaller than 5 bp. We further filtered this list such that the sQTL was wholly contained within the upstream or downstream intron. We then further filtered the sQTLs considered such that all individuals analyzed did not contain the same genotype for that particular variant. We then compiled the MNase profiles of individuals with genotypes representing shorter upstream introns (reference allele for upstream insertions and variant allele for upstream deletions) and longer upstream introns (reference allele for upstream deletions and variant allele for upstream insertions) and processed and plotted as done previously.

New exon inclusion and species-specific expression changes Gene expression in mouse was compared to gene expression in rat by taking the ratio of mouse / rat using gene expression from (38). We considered the following cases: 1) genes with a new exon where the new exon is included in the tissue in question, 2) genes with a new exon where the new exon is not included in the tissue in question, and 3) genes with no new exon in either mouse or rat.

The intra-species expression ratio (Fig. 4-11B) is calculated by averaging a gene's expression in mouse in the tissues where the exon is included and dividing that by the mean expression in tissues where the exon is not included. This ratio was then calculated in rat, matching the tissues in the fore- and background, and the ratio of these two values was analyzed. As a control, the tissue labels were shuffled and the statistic was recalculated.

The analysis identifying an enrichment for new exons in genes containing expression changes (Fig. 4-11C) was conducted as follows. For each gene, we constructed a set of constitutive exons in each species containing no alternatively spliced segments. For each tissue in mouse and rat, we counted the number of reads overlapping each region using pysam and adjusted the raw counts for differences in length considered between species, down-sampling to match the shorter length. We then applied DESeq (3) and identified genes with higher expression within the species being studied with an adjusted FDR in a single tissue of 0.01%. Correcting for multiple testing across the 9 tissues yields a conservative overall FDR of approximately 0.1%. We then divided the fraction of genes with significantly elevated expression that contain a novel exon to the overall fraction of genes that contain a novel exon.

Chi-square test on contingency table of two variables, expression change (gain or no gain) and exon status (new exon or not a new exon) was analyzed for each tissue using genes with alternative spliced new exons in mouse and rat orthologs. A new-exon-gene with expression higher in mouse than rat was assigned to “gain” category and new exon with PSI = 0 was assigned to “not a new exon” category in a certain tissue.

Software versions The analyses were conducted in Python v2.7.2 using Scipy v0.13.2 (scipy.org), Numpy v1.8.0 (numpy.scipy.org), Matplotlib v1.3.1

$$http://www.citepulike.org/user/jabl/article/2878517$$

, pycogent v1.5.1 (24), Tabix (32), Samtools (33), FSA (7), and pandas v0.10.0(37).

4.4 Author Contributions

J.M. and C.B.B. designed the study, interpreted the data, and wrote the manuscript. J.M. and P.C. conducted computational analyses. J.M. prepared figures. S.H. contributed to study design.

Bibliography

[1] Josep F. Abril, Robert Castelo, and Roderic Guig. Comparison of splice sites in mammals and chicken. *Genome Res*, 15(1):111–119, Jan 2005.

[2] Alexander V. Alekseyenko, Namshin Kim, and Christopher J. Lee. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, 13(5):661–670, May 2007.

[3] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[4] M. V. Bell, A. E. Cowper, M. P. Lefranc, J. I. Bell, and G. R. Screaton. Influence of intron length on alternative splicing of cd44. *Mol Cell Biol*, 18(10):5930–5941, Oct 1998.

[5] David L. Bentley. Coupling mrna processing with transcription in time and space. *Nat Rev Genet*, 15(3):163–175, Mar 2014.

[6] Robert K. Bradley, Jason Merkin, Nicole J. Lambert, and Christopher B. Burge. Alternative splicing of rna triplets is often regulated and accelerates proteome evolution. *PLoS Biol*, 10(1):e1001229, Jan 2012.

[7] Robert K. Bradley, Adam Roberts, Michael Smoot, Sudeep Juvekar, Jaeyoung Do, Colin Dewey, Ian Holmes, and Lior Pachter. Fast statistical alignment. *PLoS Comput Biol*, 5(5):e1000392, May 2009.

[8] J. Brosius and S. J. Gould. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk dna". *Proc Natl Acad Sci U S A*, 89(22):10706–10710, Nov 1992.

[9] Andr Corvelo and Eduardo Eyras. Exon creation and establishment in human genes. *Genome Biol*, 9(9):R141, 2008.

[10] Manuel de la Mata, Claudio R. Alonso, Sebastin Kadener, Juan P. Fededa, Matas Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R. Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Mol Cell*, 12(2):525–532, Aug 2003.

[11] William G. Fairbrother, Ru-Fang Yeh, Phillip A. Sharp, and Christopher B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–1013, Aug 2002.

[12] Kristi L. Fox-Walsh, Yimeng Dou, Bianca J. Lam, She-Pin Hung, Pierre F. Baldi, and Klemens J. Hertel. The architecture of pre-mrnas affects mechanisms of splice-site pairing. *Proc Natl Acad Sci U S A*, 102(45):16176–16181, Nov 2005.

[13] Daniel J. Gaffney, Graham McVicker, Athma A. Pai, Yvonne N. Fondufe-Mittendorf, Noah Lewellen, Katelyn Michelini, Jonathan Widom, Yoav Gilad, and Jonathan K. Pritchard. Controls of nucleosome positioning in the human genome. *PLoS Genet*, 8(11):e1003036, 2012.

[14] Xiang Gao and Michael Lynch. Ubiquitous internal gene duplication and intron creation in eukaryotes. *Proc Natl Acad Sci U S A*, 106(49):20818–20823, Dec 2009.

[15] Corinna Giorgi, Gene W. Yeo, Martha E. Stone, Donald B. Katz, Christopher Burge, Gina Turrigiano, and Melissa J. Moore. The ejc factor eif4aiii modulates synaptic strength and neuronal protein expression. *Cell*, 130(1):179–191, Jul 2007.

[16] Felizza Q. Gunderson, Evan C. Merkhofer, and Tracy L. Johnson. Dynamic histone acetylation is critical for cotranscriptional spliceosome assembly and spliceosomal rearrangements. *Proc Natl Acad Sci U S A*, 108(5):2004–2009, Feb 2011.

[17] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cdric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guig, and Tim J. Hubbard. Gencode: the reference human genome annotation for the encode project. *Genome Res*, 22(9):1760–1774, Sep 2012.

[18] Joanna Y. Ip, Dominic Schmidt, Qun Pan, Arun K. Ramani, Andrew G. Fraser, Duncan T. Odom, and Benjamin J. Blencowe. Global impact of rna polymerase ii elongation inhibition on alternative splicing regulation. *Genome Res*, 21(3):390–401, Mar 2011.

[19] J. J. Jonsson, M. D. Foresman, N. Wilson, and R. S. McIvor. Intron requirement for expression of the human purine nucleoside phosphorylase gene. *Nucleic Acids Res*, 20(12):3191–3198, Jun 1992.

[20] Henrik Kaessmann, Nicolas Vinckenbosch, and Manyuan Long. Rna-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet*, 10(1):19–31, Jan 2009.

[21] Minna U. Kaikkonen, Nathanael J. Spann, Sven Heinz, Casey E. Romanoski, Karmel A. Allison, Joshua D. Stender, Hyun B. Chun, David F. Tough, Rab K. Prinjha, Christopher Benner, and Christopher K. Glass. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. *Mol Cell*, 51(3):310–325, Aug 2013.

[22] W James Kent. Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[23] Ekta Khurana, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, Lucas Lochovsky, Jieming Chen, Arif Harmanci, Jishnu Das, Alexej Abyzov, Suganthi Balasubramanian, Kathryn Beal, Dimple Chakravarty, Daniel Challis, Yuan Chen, Declan Clarke, Laura Clarke, Fiona Cunningham, Uday S. Evani, Paul Flicek, Robert Fragoza, Erik Garrison, Richard Gibbs, Zeynep H. Gms, Javier Herrero, Naoki Kitabayashi, Yong Kong, Kasper Lage, Vaja Liluashvili, Steven M. Lipkin, Daniel G. MacArthur, Gabor Marth, Donna Muzny, Tune H. Pers, Graham R S. Ritchie, Jeffrey A. Rosenfeld, Cristina Sisu, Xiaomu Wei, Michael Wilson, Yali Xue, Fuli Yu, 1000 Genomes Project Consortium , Emmanouil T. Dermitzakis, Haiyuan Yu, Mark A. Rubin, Chris Tyler-Smith, and Mark Gerstein. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, 342(6154):1235587, Oct 2013.

[24] Rob Knight, Peter Maxwell, Amanda Birmingham, Jason Carnes, J Gregory Caporaso, Brett C. Easton, Michael Eaton, Micah Hamady, Helen Lindsay, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Michael Robeson, Raymond Sammut, Sandra Smit, Matthew J. Wakefield, Jeremy Widmann, Shandy Wikman, Stephanie Wilson, Hua Ying, and Gavin A. Huttley. Pycogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171, 2007.

[25] E. V. Kriventseva and M. S. Gelfand. Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. *J Biomol Struct Dyn*, 17(2):281–288, Oct 1999.

[26] Emilie Lalonde, Kevin C H. Ha, Zibo Wang, Amandine Bemmo, Claudia L. Kleinman, Tony Kwan, Tomi Pastinen, and Jacek Majewski. Rna sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res*, 21(4):545–554, Apr 2011.

[27] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[28] Ilkka Lappalainen, John Lopez, Lisa Skipper, Timothy Hefferon, J Dylan Spalding, John Garner, Chao Chen, Michael Maguire, Matt Corbett, George Zhou, Justin Paschall, Victor Ananiev, Paul Flicek, and Deanna M. Church. Dbvar and dgva: public archives for genomic structural variation. *Nucleic Acids Res*, 41(Database issue):D936–D941, Jan 2013.

[29] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedlnder, Peter A C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg, Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy,

Paul Flicek, Tim M. Strom, Geuvadis Consortium , Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Angel Carracedo, Stylianos E. Antonarakis, Robert Hsler, Ann-Christine Syvnen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guig, Ivo G. Gut, Xavier Estivill, Emmanouil T. Dermitzakis, and Geuvadis Consortium . Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, Sep 2013.

[30] Galit Lev-Maor, Rotem Sorek, Noam Shomron, and Gil Ast. The birth of an alternatively spliced exon: 3' splice-site selection in alu exons. *Science*, 300(5623):1288–1291, May 2003.

[31] Mia T. Levine, Corbin D. Jones, Andrew D. Kern, Heather A. Lindfors, and David J. Begun. Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*, 103(26):9935–9939, Jun 2006.

[32] Heng Li. Tabix: fast retrieval of sequence features from generic tab-delimited files. *Bioinformatics*, 27(5):718–719, Mar 2011.

[33] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup . The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[34] Reini F. Luco, Qun Pan, Kaoru Tominaga, Benjamin J. Blencowe, Olivia M. Pereira-Smith, and Tom Misteli. Regulation of alternative splicing by histone modifications. *Science*, 327(5968):996–1000, Feb 2010.

[35] D. Mascarenhas, I. J. Mettler, D. A. Pierce, and H. W. Lowe. Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol Biol*, 15(6):913–920, Dec 1990.

[36] Arianne J. Matlin, Francis Clark, and Christopher W J. Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–398, May 2005.

[37] Wes McKinney. Data structures for statistical computing in python. In Stefan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[38] Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B. Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, Dec 2012.

[39] Barmak Modrek and Christopher J. Lee. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat Genet*, 34(2):177–180, Jun 2003.

[40] R. D. Palmiter, E. P. Sandgren, M. R. Avarbock, D. D. Allen, and R. L. Brinster. Heterologous introns can enhance expression of transgenes in mice. *Proc Natl Acad Sci U S A*, 88(2):478–482, Jan 1991.

[41] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res*, 18(11):1814–1828, Nov 2008.

[42] Joseph K. Pickrell, John C. Marioni, Athma A. Pai, Jacob F. Degner, Barbara E. Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, Apr 2010.

[43] Paz Polak and Peter F. Arndt. Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res*, 18(8):1216–1223, Aug 2008.

[44] Meenakshi Roy, Namshin Kim, Yi Xing, and Christopher Lee. The effect of intron length on exon creation ratios during the evolution of mammalian genomes. *RNA*, 14(11):2261–2273, Nov 2008.

[45] Schraga Schwartz, Eran Meshorer, and Gil Ast. Chromatin organization marks exon-intron structure. *Nat Struct Mol Biol*, 16(9):990–995, Sep 2009.

[46] Scott Schwartz, W James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. Human-mouse alignments with blastz. *Genome Res*, 13(1):103–107, Jan 2003.

[47] Shihao Shen, Lan Lin, James J. Cai, Peng Jiang, Elizabeth J. Kenkel, Mallory R. Stroik, Seiko Sato, Beverly L. Davidson, and Yi Xing. Widespread establishment and regulatory impact of alu exons in human genes. *Proc Natl Acad Sci U S A*, 108(7):2837–2842, Feb 2011.

[48] Hubley R & Green P Smit, AFA. Repeatmasker open-3.0. 1996-2010 ¡http://www.repeatmasker.org¿.

[49] Rotem Sorek. The birth of new exons: mechanisms and evolutionary consequences. *RNA*, 13(10):1603–1608, Oct 2007.

[50] Rotem Sorek, Gil Ast, and Dan Graur. Alu-containing exons are alternatively spliced. *Genome Res*, 12(7):1060–1067, Jul 2002.

[51] Noah Spies, Cydney B. Nielsen, Richard A. Padgett, and Christopher B. Burge. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell*, 36(2):245–254, Oct 2009.

[52] Vladimir B. Teif, Yevhen Vainshtein, Mawen Caudron-Herger, Jan-Philipp Mallm, Caroline Marth, Thomas Hfer, and Karsten Rippe. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol*, 19(11):1185–1192, Nov 2012.

[53] Hagen Tilgner, Christoforos Nikolaou, Sonja Althammer, Michael Sammeth, Miguel Beato, Juan Valcrcel, and Roderic Guig. Nucleosome positioning as a determinant of exon recognition. *Nat Struct Mol Biol*, 16(9):996–1001, Sep 2009.

[54] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[55] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[56] Eric T. Wang, Neal A L. Cody, Sonali Jog, Michela Biancolella, Thomas T. Wang, Daniel J. Treacy, Shujun Luo, Gary P. Schroth, David E. Housman, Sita Reddy, Eric Lcuyer, and Christopher B. Burge. Transcriptome-wide regulation of pre-mrna splicing and mrna localization by muscleblind proteins. *Cell*, 150(4):710–724, Aug 2012.

[57] Yang Wang, Xinshu Xiao, Jianming Zhang, Rajarshi Choudhury, Alex Robertson, Kai Li, Meng Ma, Christopher B. Burge, and Zefeng Wang. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol*, 20(1):36–45, Jan 2013.

[58] Zefeng Wang and Christopher B. Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008.

[59] Zefeng Wang, Michael E. Rolish, Gene Yeo, Vivian Tung, Matthew Mawson, and Christopher B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–845, Dec 2004.

[60] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[61] Gene W. Yeo, Eric Van Nostrand, Dirk Holste, Tomaso Poggio, and Christopher B. Burge. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A*, 102(8):2850–2855, Feb 2005.

[62] Xiang H-F. Zhang and Lawrence A. Chasin. Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons. *Proc Natl Acad Sci U S A*, 103(36):13427–13432, Sep 2006.

[63] Hua-Lin Zhou, Melissa N. Hinman, Victoria A. Barron, Cuiyu Geng, Guangjin Zhou, Guangbin Luo, Ruth E. Siegel, and Hua Lou. Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an rna-dependent manner. *Proc Natl Acad Sci U S A*, 108(36):E627–E635, Sep 2011.

Chapter 5

Conclusion

5.1 Summary

This thesis applied functional genomics and high-throughput sequencing to investigate the evolution of splicing and alternative splicing in mammals. Chapter 2 describes a targeted investigation into the evolution of a class of alternative splicing, tandem 3' splice sites or NAGNAGs. Through the use of comparative genomics, I found that NAGNAG turnover is very frequently implicated in the gain or loss of coding sequence at exon boundaries. Chapter 3 describes the evolution of ancient exons, defined here as being present in chicken and mammals (~300mya). I describe the generation of a large RNA-seq dataset used in the analyses, and detail how the tissue-specific splicing patterns of these exons are often poorly conserved and frequently lineage-specific. I identified a subset of exons with conserved tissue-regulated splicing patterns that alter the protein's phosphorylation potential in a subset of tissues. I further identified unique regulatory properties in ancient alternative exons. Chapter 4 discusses the sources of novel exons and the evolution of gene structures. I observe that most new exons arise from pre-existing non-repetitive intronic sequences and detail associated genomic changes that are implicated in this process. I further suggest that one major impact of these new exons is to increase the gene's level of expression generally or in specific tissues.

5.2 Future directions

5.2.1 Ancient exon regulation

One surprising result from this work is that ancient alternative exons are better bound by some splicing splicing factors (explored most deeply with regards to Muscleblind) than more lineage-specific alternative exons or constitutive exons. We found this phenomenon to be robust to differences in expression and motif count, suggesting a potentially biological explanation. A better understanding of the cause for this preferential binding will lead to a better understanding of the biology and function of these splicing factors. This is now being explored *in vitro* using Bind-N-Seq, a technique that applies *in vitro* binding of a target protein to an RNA oligo pool followed by high-throughput sequencing to experimentally determine K_d values for potentially all k-mers up to a ~8-10. The ability to recapitulate the phenomenon *in vitro* with purified protein and RNA would suggest something intrinsic to the protein and RNA, as opposed to something like negative cooperativity or competitive binding with other proteins, is conferring this property. We are also testing the effect of secondary structure by applying SHAPE-seq to experimentally interrogate the secondary structure propensities of the motifs' neighborhoods within introns of different ages of alternative exons. Once this phenomenon has been teased apart and understood with regards to Muscleblind, how other proteins effect a similar pattern of binding (or why they don't) can then investigated. This will lead to a better understanding of the so-called "splicing code."

5.2.2 More ideal data sources as future technologies mature

This thesis analyzed levels of mRNA from tissues to infer splicing levels and patterns. However, we analyze the mRNA to understand the protein isoforms and levels since it is generally the protein product that is biologically active. While RNA levels are significantly correlated with protein levels, this correlation is only modest. Proteomic technology is progressing rapidly, but is not yet able to globally analyze splicing. It will therefore be informative in the future to interrogate at the protein level the evolution of alternative splicing.

In studying RNA collected from tissue lysates, the RNA pools in different cell types are therefore mixed. This mixing can confound analyses if the relative amounts of each cell type change or if there exist cell-types not found in some species. Therefore, another promising direction is to conduct RNA-seq analyses on purified cell populations or, preferably, on single cells from homogeneous populations purified from various tissues in different species. Single-cell sequencing will allow for the comparison of truly orthologous cell types and remove potential confounding effects from changes in cell-type composition of tissues between species.

5.2.3 Evolutionary pressures on splicing

A number of evolutionary pressures have been described over the years, with perhaps the simplest being positive and negative selection. Negative or stabilizing pressure acts to prevent changes due to negative or deleterious alleles because a process or molecule is vital to the fitness of an organism. Positive selection occurs when there exist pressures, environmental or otherwise, that select for a particular trait or characteristic. It has been recently suggested that gene expression has been generally evolving under stabilizing pressure with largely neutral variation within mammals. More complex evolutionary models (that are being observed with increasing frequency) include ideas such as polygenic adaptations, where small genetically-mediated changes at multiple sites can have an additive effect that is physiologically relevant while no individual change is significant. In this thesis, we described the patterns of splicing evolution. We did not, however, test these models for the predominant mode or modes of splicing evolution and therefore additional work will be necessary in this area.

5.2.4 Interpretation of disease variants

In chapter 3 of this thesis, I describe a connection between alternative exons and phosphorylation potential of proteins. In chapter 4, I describe evidence suggesting that intronic indels alter nucleosome positioning. I propose and provide evidence that when located in the intron upstream of an alternative exon deletions tend to promote the exon's inclusion. These two observations can be combined to potentially aid in the interpretation of genomic variants, particularly as they pertain to signaling pathways and potentially disease. Genomic variants located in introns are often ignored because it is difficult to interpret how they can exert an impact (contrasted with coding variants, where missense and nonsense variants can be much more straight-forward to interpret). Intronic indels that alter splicing of exons containing phosphorylation sites for a particular signaling pathway can lead to perturbations within that pathway, potentially without any detected coding mutations. For instance, if an exon's inclusion is increased from 10% to 80%, then the protein would have switched from being mostly unphosphorylatable to mostly phosphorylatable. An upstream kinase can then phosphorylate that protein, leading to altered levels of signaling. As the number of both healthy and sick individuals with their whole genome sequences and complementary datasets (such as RNA-seq or mass-spectrometry increases) generated increases, it will be possible to look at this potential relationship as it pertains to human disease.