Supplementary Methods

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 {Pickrell, 2010 #90;Gonzalez-Porta, 2012 #99}. A complete set of tissues was collected from three individuals from each of the species, enabling separate analysis of tissue-specific and individual variation. Animals of breeding age were chosen as gene expression is reported to be relatively stable in this age group for mammals {Somel, 2010 #100}. Only males were used, to enable analysis of testis, a tissue with an unusually large extent of alternative splicing {Yeo, 2004 #35}. 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.

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 {Parkhomchuk, 2009 #75}, 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.

Read mapping and analysis

The 80x80 libraries were mapped to their respective genomes (musmus9, rhemac2, ratnor4, bostau4, galgal3) using Tophat {Trapnell, 2009 #65} 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 {Trapnell, 2010 #93} 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 {Roberts, 2011 #92}. 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 (fig. S3). 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. Such transcripts might be involved in epigenetic regulation of overlapping sense gene loci {Magistri, 2012 #170}. We also identified at least several thousand putative novel exons in transcripts not overlapping known genes in each species.

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 intron-proximal 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 {Paten, 2008 #141}. 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 were restricted to exons that were detected in > 2 tissues in more than half of the individuals (requiring chicken). The locations of the conserved regions observed in Figure 3A were consistent with studies of exons with highly tissue-specific splicing {Sugnet, 2006 #95;Wang, 2008 #44}.

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 {Wang, 2008 #44}. 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 genomically could have potentially included 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 {Trapnell, 2010 #93}. These values, estimates made using MISO {Katz, 2010 #56}, and qRT-PCR measurements from 3 mouse tissues {Wang, 2012 #190} were all highly correlated with each other (*r* between 0.84 and 0.90, fig. S1).

In Figure 1 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 {Berretta, 2010 #94;Martinez, 2008 #98}. 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.

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 {Lindblad-Toh, 2011 #199}. 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. 4C), 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.

Tissue-specificity and splicing conservation calculations

For each orthologous event, a matrix of n rows by m columns was constructed, where each row repesented 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 vector's 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.

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. 3B, 5mers were ranked by "discrimination information", $f \log_2(f/g)$, 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 discimination information was plotted for lowly (species-specific), intermediately (rodent-specific), and broadly (all mammals) alternative exons in Fig. 3D. In Fig. 3C, 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. 3G, 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 {Wang, 2010 #179;Yeo, 2009 #180} (fig. S9).

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 {Stajich, 2002 #77} was used to submit and handle queries to Scansite {Obenauer, 2003 #58} of complete ORFs. Putative phosphorylation sites were cross-referenced with Phosphosite {Hornbeck, 2004 #144} to identify experimentally validated sites. To consider the tissue-specificity of sites, data from {Huttlin, 2010 #197} 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 family's counts and compared it with a similarly normalized expression vector composed of the sum of the kinases in each tissue. Increased post-translation modification was also observed in a set of tissue-specific human exons {Buljan, 2012 #178} identified based on a previous RNA-Seq analysis of human tissues {Wang, 2008 #44}.

Data analysis

Custom Python scripts were used for analyses, utilizing Numpy (numpy.scipy.org), Scipy (scipy.org), Pycogent{Knight, 2007 #188}, BioPerl{Stajich, 2002 #77}, Tabix{Li, 2011 #149}, Samtools{Li, 2009 #150},

Bedtools{Quinlan, 2010 #151}, FSA{Bradley, 2009 #66} and Matplotlib [http://www.citeulike.org/user/jabl/article/2878517].

Supplementary Figure Legends

Supplementary figure 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 {Wang, 2008 #44}, 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.

Supplementary figure 2

PSI values were calculated by Cufflinks, MISO, and compared to each other and to qRT-PCR measurements from a recent study {Wang, 2012 #190}. 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.

Supplementary figure 3

Heirarchical clustering of all samples by skipped exon PSI values is shown, as in Fig. 1B. Here, the minimum expression cutoff required of each event to be considered was raised to A) 10 FPKM or B) 15 FPKM.

Supplementary figure 4

Hierarchical clustering of samples by JSD based on FPKM for gene expression

of transcription factors (top) or splicing factors (middle). As in Figure 1A, 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 (genes listed in Supplementary Tables 3). 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.

Supplementary figure 5

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.

Supplementary figure 6

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.

Supplementary figure 7

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. 3C.

Supplementary figure 9

Preferential binding of TIA family splicing factors to broadly alternative exons. CLIP-Seq 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 CLIP-Seq 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.

Supplementary figure 9

Predicted phosphorylation site density in similar classes of exons as analyzed in Figure 4B in other species studied: A) chicken, B) cow, C) rhesus, D) rat, E) mouse

Supplementary figure 10

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).

Supplementary figure 11

The relative expression of kinases that are acid-directed, base-directed, or proline-directed and also evaluated in Scansite are plotted (y-axis) against the normalized distribution of spectral counts assigned to each family in {Huttlin, 2010 #197}.

Supplementary figure 12

Tissue-specificity of phosphorylation sites identified by {Huttlin, 2010 #197} 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

Supplementary figure 13

The splicing of mouse Drosha exon 5 is compared with the expression of cognate kinases A) Clk2, and B) Akt.

Supplementary figure 14

Network visualization of all kinase-exon relationships identified in the conserved, tissue-specific 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).

Supplementary figure 15

Heirarchical clustering by JSD of gene expression (FPKM) of all genes containing an exon analyzed in Fig. 1B. 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. 1A, restricting to the set of genes containing an exon that was alternative in all species.

Supplementary Table Legends

Supplementary table 1

Summary of sequence data generated in this study and comparison to annotations. Genes lists the total number of protein coding genes in Ensembl release 61 for each species. Genes detected is the subset of protein-coding genes with at least one annotated exon detected by Cufflinks. Annotated exons

is the total number of exons in protein-coding genes. Annotated exons detected is the number of annotated exons in protein-coding genes overlapping an exon detected in these data. Missed exons is the number of exons in existing Ensembl annotations that were not detected in these data. Cufflinks annotated exons are putative novel exons identified by Cufflinks that are not annotated in Ensembl protein-coding genes. Unannotated exons in spliced transcript overlapping known genes are defined as exons in spliced Cufflinks transcripts that overlap a known gene locus but are not connected to annotated exons in the gene. Unannotated spliced exons antisense to known genes are defined as unannotated exons in spliced transcripts that overlap a known gene in the anti-sense orientation but are not part of a known gene. Coverage was calculated by dividing the number of bases sequenced by 100 Mbp, an approximate estimate of the size of the total length of the protein-coding transcriptome in these organisms. Overall, ~98% of putative novel exons had AG at the 3' splice site, and a comparable number had GT or GC at the 5' splice site. (Excel table)

Supplementary table 2

Genes and exons analyzed in Figure 1 (Excel table)

Supplementary table 3

Transcription factor genes analyzed in Supplementary Figures 1A and 1B (Excel table), and splicing factor genes analyzed in Supplementary Figure 1C (Excel table)

Supplementary Table 4

Exons with alternative splicing inferred to be > 100 million years old (Excel table)

Supplementary Table 5

Frequencies of pentanucleotdes downstream of various classes of alternative exons relative to constitutive exons (Excel table)

Supplementary table 6

Gene Ontology analysis of species-specific and tissue-specific alternative exons (Excel table)

Supplementary table 7

Conserved, tissue-specific exons encoding putative phosphorylation sites, the associated kinases, and KSI values (Excel Table)

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.

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.) and by grants from the NIH to C. B. B.

Author Information

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

Supplementary References