ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

The mRNA and protein isoforms produced by alternative processing of primary RNA transcripts may differ in structure, function, localization or other properties[1,2]. Alternative splicing in particular is known to affect more than half of all human genes, and has been proposed as a primary driver of the evolution of phenotypic complexity in mammals[3,4]. However, assessment of the extent of differences in mRNA isoform expression between tissues has presented substantial technical challenges[5]. Studies using expressed sequence tags have yielded relatively low estimates of tissue specificity, but have limited statistical power to detect differences in isoform levels[6–8]. Microarray analyses have achieved more consistent coverage of tissues[9], but are constrained in their ability to distinguish closely related mRNA isoforms. High-throughput sequencing technologies have the potential to circumvent these limitations by generating high average coverage of mRNAs across tissues while using direct sequencing rather than hybridization to distinguish and quantify mRNA isoforms[10,11].

Tissue-specific alternative splicing is usually regulated by a combination of tissue-specific and ubiquitously expressed RNA-binding factors that interact with *cis*-acting RNA elements to influence spliceosome assembly at nearby splice sites[1,2]. Many factors can both activate and repress splicing in different contexts, with activity often summarizable by an 'RNA map' describing dependence on the location of binding relative to that of core spliceosomal components[12,13].

A digital inventory of mRNA isoforms

To assess gene and alternative mRNA isoform expression, the mRNA-Seq protocol (Supplementary Methods) was used to amplify and sequence between 12 million and 29 million 32-base-pair (bp) cDNA fragments from ten diverse human tissues and five mammary epithelial or breast cancer cell lines, generating over 400 million reads in total (Supplementary Fig. 1a). Tissue samples were derived from single anonymous unrelated individuals of both sexes; for one tissue, cerebellar cortex, samples from six unrelated men were analysed to assess variation between individuals (Supplementary Table 1). In total, ~60% of reads mapped uniquely to the genome, allowing up to 2 mismatches, and an additional 4% mapped uniquely to splice junctions. Thus, about two-thirds of reads could be assigned unambiguously to individual genes; the frequency of mapping to incorrect genomic locations was estimated to be ~0.1% (Supplementary Table 2).

Read density (coverage) was over 100-fold higher in exons than in introns or intergenic regions (Supplementary Fig. 1c), and only ~3% of reads mapped to ribosomal RNA genes, indicating that most reads derived from mature mRNA. Comparison of relative mRNA-Seq read densities to published quantitative polymerase chain reaction with reverse transcription (RT–PCR) measurements for 787 genes in two reference RNA samples[14] yielded a nearly linear relationship across ~5 orders of magnitude (Supplementary Fig. 1d), indicating that mRNA-Seq read counts give accurate relative gene expression measurements across a very broad dynamic range[10].

Alternative splicing is nearly universal

The mRNA-Seq data were used to assess the expression of alternative transcript isoforms in human genes, as illustrated for the mitochondrial phosphate transporter gene *SLC25A3* in Fig. 1a. Exons 3A and 3B of this gene are 'mutually exclusive exons' (MXEs), meaning that transcripts from this gene contain one or the other of these exons, but not both. Much greater read coverage of exon 3A was seen in heart and skeletal muscle, with almost exclusive coverage of exon 3B in

[1]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [3]Department of Cell and Molecular Biology, Karolinska Institutet, 171 77 Stockholm, Sweden. [4]Illumina Inc., 25861 Industrial Boulevard, Hayward, California 94545, USA. [5]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA. [6]National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, New Mexico 87505, USA.
*These authors contributed equally to this work.

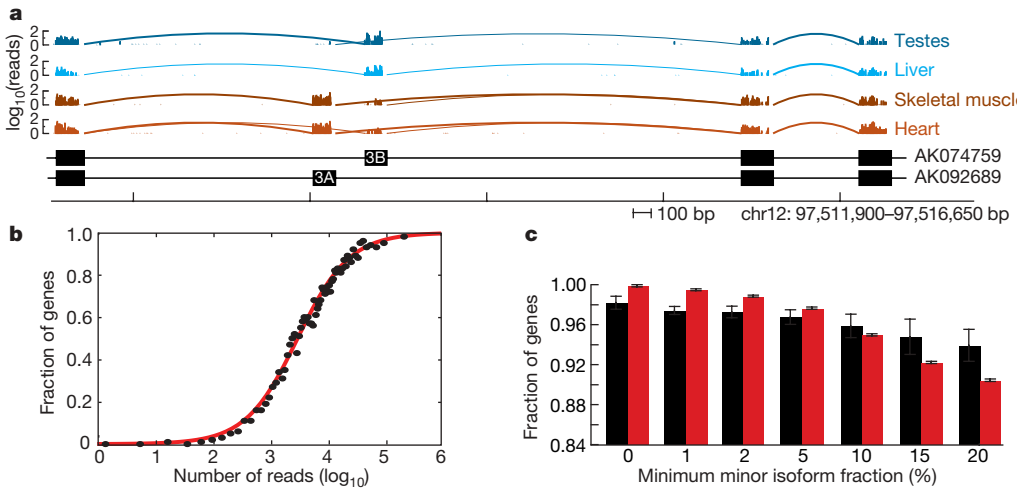


Figure 1 | Frequency and relative abundance of alternative splicing isoforms in human genes. **a**, mRNA-Seq reads mapping to a portion of the *SLC25A3* gene locus. The number of mapped reads starting at each nucleotide position is displayed ($\log_{10}$) for the tissues listed at the right. Arcs represent junctions detected by splice junction reads. Bottom: exon/intron structures of representative transcripts containing mutually exclusive exons 3A and 3B (GenBank accession numbers shown at the right). **b**, Mean fraction of multi-exon genes with detected alternative splicing in bins of 500 genes, grouped by total read count per gene. A gene was considered as alternatively spliced if splice junction reads joining the same 5′ splice site (5′SS) to different 3′ splice sites (3′SS) (with at least two independently mapping reads supporting each junction), or joining the same 3′SS to different 5′SS, were observed. The true extent of alternative splicing was estimated from the upper asymptote of the best-fit sigmoid curve (red curve). Circles show the fraction of alternatively spliced genes. **c**, Frequency of alternative splicing in the top bin (black bars) and after estimation (as in **b**, red bars), considering only events with relative expression of less abundant (minor) splice variant exceeding a given threshold. Error bars, s.e.m.

testes and liver (as well as in other tissues studied), consistent with the predominant heart and muscle symptoms of exon 3A mutation[15].

The genome-wide extent of alternative splicing was assessed by searching against known and putative splicing junctions using stringent criteria that required each alternative isoform to be supported by multiple independent splice junction reads with different alignment start positions. Binning the multi-exon genes in the RefSeq database (94% of all RefSeq genes) by read coverage and fitting to a sigmoid curve enabled estimation of the asymptotic fraction of alternatively spliced genes in this set as ~98% when excluding cell line data (Supplementary Fig. 2) and ~100% when using all samples (Fig. 1b). This analysis indicated that alternative splicing is essentially universal in human multi-exon genes, which comprise 94% of genes overall, with the important qualification that a portion of detected alternative splicing events may represent allele-specific splicing[16,17].

Some of these events may involve exclusively low frequency alternatively spliced isoforms. However, 92% of multi-exon genes were estimated to undergo alternative splicing when considering only events for which the relative frequency of the minor (less abundant) isoform exceeded 15% in one or more samples (Fig. 1c). Thus, 0.92×0.94 or ~86% of human genes were estimated to produce appreciable levels of two or more distinct populations of mRNA isoforms. Conversely, no evidence of alternative splicing was detected in the 6% of RefSeq genes annotated as consisting of a single exon, even when searching against junctions between predicted exons in these genes.

New exons and splice junctions not previously seen in transcript databases were identified by mapping the reads against predicted exons and junctions. This approach yielded a set of 1,413 high-confidence new exons (Supplementary Table 3), with an estimated false discovery rate (FDR) of <1.5% (Supplementary Information), and thousands of putative new splice junctions (not shown). Thus, mRNA-Seq has strong potential for discovery of new exons, although very substantial read depth is required to efficiently detect low-abundance isoforms (Supplementary Fig. 3).

Tissue-specific isoform expression

To explore the extent of tissue regulation of alternative transcripts, we examined eight common types of 'alternative transcript events'[1,2], each capable of producing multiple mRNA isoforms from human genes through alternative splicing, alternative cleavage and polyadenylation (APA) and/or alternative promoter usage (Fig. 2). Event types considered included skipped exons and retained introns, in which a single exon or intron is alternatively included or spliced out of the mature message, and MXEs, described previously. Also included were alternative 5′ splice site (A5SS) and alternative 3′ splice site (A3SS) events, which are particularly difficult to interrogate by microarray analysis because the variably included region is often quite small. Tandem 3′ untranslated regions (UTRs) and alternative last exons (ALEs), in which alternative use of a pair of polyadenylation sites results in shorter or longer 3′ UTR isoforms or in distinct terminal exons, respectively, were also considered. Finally, we considered alternative first exons (AFEs), in which alternative promoter use results in mRNA isoforms with distinct 5′ UTRs.

For each of these event types, reads deriving from specific regions can support the expression of one alternative isoform or the other (Fig. 2). The 'inclusion ratio', defined as the ratio of the number of 'inclusion' (blue) reads to inclusion plus 'exclusion' (red) reads, can be used to detect changes in the proportions of the corresponding mRNA isoforms. The fraction of mRNAs that contain an exon—the 'per cent spliced in' (PSI or Ψ) value—can be estimated as the ratio of the density of inclusion reads (that is, reads per position in regions supporting the inclusion isoform) to the sum of the densities of inclusion and exclusion reads.

To assess tissue-regulated alternative splicing, a comprehensive set of ~105,000 events of these eight types was derived on the basis of available human cDNA and expressed sequence tag data. Reads supporting both alternative isoforms were observed for more than one-third of these events (Fig. 2), and the extent of tissue-specific regulation of these events was assessed by comparison of the inclusion ratio in each tissue relative to the other tissues, requiring a minimum of a 10% absolute change in inclusion ratio (Supplementary Fig. 4). Naturally, transcripts or isoforms identified as being differentially expressed between tissues will reflect the combined effects of cell-type-specific differences in transcript levels, variation in the relative abundances of cell types between tissues, and variations between the individuals from whom the tissues derived.

Notably, a high frequency of tissue-specific regulation was observed for each of the eight event types, including over 60% of the analysed skipped exon, A5SS, A3SS and tandem 3′ UTR events

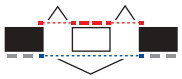
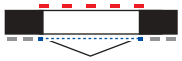
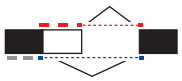
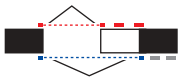
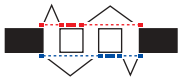
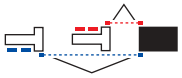
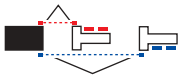
Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1–4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue-regulated (Fisher's exact test) at an FDR of 5% (assuming negligible technical variation[10]). Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, 'common reads' (grey bars) were used in lieu of (for tandem 3′ UTR events) or in addition to 'exclusion' reads for detection of changes in isoform levels between tissues.

(Fig. 2 and Supplementary Table 4). In all, a set of over 22,000 tissue-specific alternative transcript events was identified, far exceeding previous sets of tissue-specific alternative splicing events that have typically numbered in the hundreds to low thousands[6–9,18,19]. Tissue-regulated skipped exon and MXE events are listed in Supplementary Tables 5 and 6, respectively. Binning events by expression level commonly yielded sigmoid curves for the fraction of tissue-regulated events of each type, enabling estimation of the true frequency of tissue regulation for each event type (Supplementary Figs 5 and 6). These estimates, ranging from 52% to 80% (Fig. 2), indicated that most alternative splicing events are regulated between tissues, providing an important element of support for the hypothesis that alternative splicing is a principal contributor to the evolution of phenotypic complexity in mammals.

Individual-specific isoform expression

To assess the extent of alternative splicing isoform variation between individuals in comparison to tissue-regulated alternative splicing, the correlations among the vectors of inclusion ratios for all expressed skipped exons between pairs of samples were determined (Fig. 3); this was performed similarly for other event types (not shown). In this analysis, strong clustering of the six cerebellar cortex samples was observed, with generally higher correlations among these samples than between pairs representing distinct tissues. Strong clustering of the five cell lines was also observed. This probably results from a combination of factors, including the common mammary epithelial origin of the cell lines studied, similar adaptations to culture conditions, and the high diversity of the tissues chosen.

The extent of variation in alternative isoform expression between individuals was also addressed by determining the number of differentially expressed exons among the six cerebellar cortex samples. Using the same approach as in Fig. 2, between ~10% and 30% of alternative transcript events showed individual-specific variation, depending on the event type (Supplementary Fig. 7), providing updated estimates of the scope of mRNA isoform variation between individuals[16]. These numbers are higher than estimates based on microarray analyses[20], but are in general agreement with an integrated analysis of multiple data types that estimated that ~21% of alternatively spliced genes are affected by polymorphisms that alter the relative abundances of alternative isoforms[17]. However, these frequencies are still below the 47–74% of events that showed variation among the ten tissues (Fig. 2), and approximately twofold to threefold less than the frequencies observed in comparisons among subsets of six tissues (Supplementary Fig. 7), indicating that, although inter-individual variation is fairly common, it is still substantially less frequent than variation between tissues. Thus, most of the differences observed between tissue samples are likely to represent tissue-specific rather than individual-specific variation.

Switch-like alternatively spliced exons

The quantitative nature of the mRNA-Seq approach allowed assessment of both subtle and switch-like alternative splicing events. By

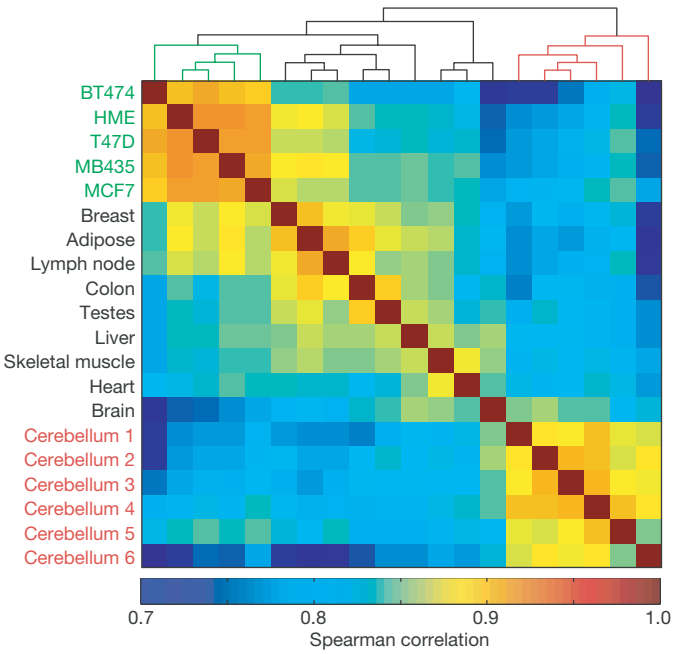


Figure 3 | The extent of individual-specific differences in alternative isoform expression. Spearman correlations of inclusion ratios for skipped exons in human tissues and cell lines (see Methods Summary). Correlations were computed separately for each pair of tissues and cell lines, and clustered according to similarity using average linkage hierarchical clustering.

comparing inclusion levels of skipped exons between tissues, a class of 'switch-like' exons was observed that had markedly different inclusion levels between different tissues (shown for heart versus nine other tissues in Fig. 4a). The examples shown in colour in Fig. 4a (for example, *TPM1* exon 2, with Ψ of 2% in heart and 95% in skeletal muscle, and the *SLC25A3* MXE pair shown in Fig. 1a) underscore the flexibility of the splicing regulatory machinery, with a sizeable number of exons being recognized predominantly as exons in one tissue and predominantly as introns in another tissue, even for developmentally related pairs of tissues such as heart and skeletal muscle.

To characterize functional features of such switch-like exons, skipped exons and MXEs were divided into groups depending on their 'switch score', defined as the maximum pairwise Ψ difference between tissues. Switch scores for pairs of MXEs were shifted towards higher values relative to skipped exons ($P = 3.7 \times 10^{-5}$, Kolmogorov–Smirnov test; Fig. 4b), suggesting that MXEs are more often involved in regulating highly tissue-specific functions. Preservation of the reading frame in both isoforms was observed more commonly for exons with higher switch scores both for skipped exons, consistent with ref. 19, and to an even greater extent for MXEs (Fig. 4c). Thus, switch-like regulation seems to be used preferentially to express distinct 'full-length' protein isoforms in different tissues rather than as a means to switch off genes through production of truncated proteins or of messages subject to nonsense-mediated mRNA decay[21]. Indeed, genes containing skipped exons with high switch scores were enriched for Gene Ontology functional categories including 'developmental processes', 'cell communication', 'signal transduction' and 'regulation of metabolism' that are likely to contribute to fundamental differences in the biology of different human tissues (Supplementary Table 7).

Notably, skipped exons with switch scores exceeding 0.5 showed higher sequence conservation in the regulated exon itself[19] and in portions of the flanking introns than exons with lower switch scores (Fig. 4d). This observation suggested that such exons are of unusual biological importance and that switch-like regulation between tissues requires the presence of additional splicing regulatory sequence information, particularly in adjacent intronic regions.

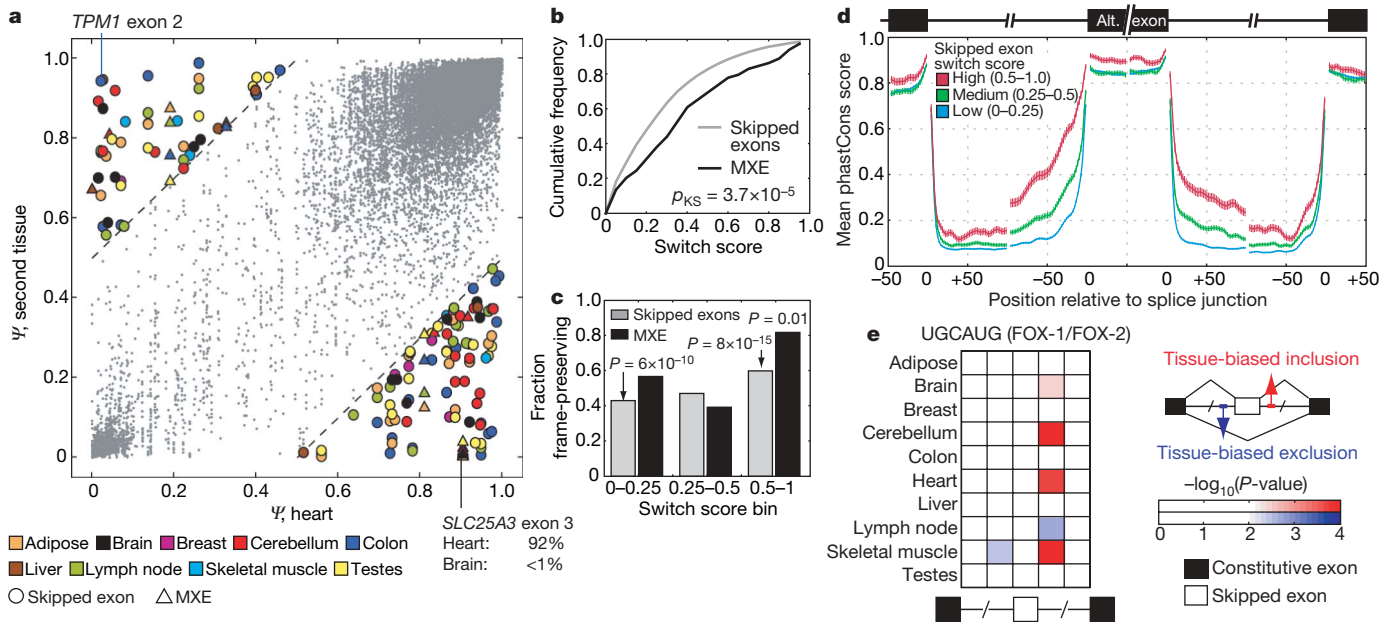


Figure 4 | Conservation and function of switch-like alternative splicing exons. **a**, Scatter plot showing Ψ values of skipped exons and MXEs for which switch score was determined on the basis of comparison of heart (*x*-axis) with a second tissue (*y*-axis). Exons with a switch score >0.5 are shown as filled symbols; others are shown as small grey dots. **b**, Cumulative distribution functions of switch scores for skipped exons and MXE pairs (*P*-value based on Kolmogorov–Smirnov test). **c**, Reading frame preservation of skipped exons and MXEs grouped by switch score. Skipped exons with lengths divisible by 3 and MXE pairs with lengths differing by 0 or a multiple of 3 were considered to preserve the reading frame. *P*-values are based on Fisher's exact test. **d**, Conservation in skipped exon and flanking intron regions grouped by skipped exon switch score. The mean per-position phastCons score (from alignment of four mammalian genomes) and s.e.m. are shown. **e**, Enrichment of UGCAUG motifs near tissue-regulated skipped exons. Coloured squares represent $-\log_{10}$(*P*-value) for the enrichment of UGCAUG counts relative to cohorts of control hexanucleotides in regions surrounding skipped exons with significantly increased (red) or decreased (blue) inclusion in each tissue with respect to other tissues.

FOX-1 and FOX-2 activity map

Among the best-characterized tissue-specific splicing factors are the FOX-1 (also known as A2BP1) and FOX-2 (RBM9) proteins, which bind RNA *cis*-elements that contain UGCAUG hexanucleotides or closely related sequences[22–24]. Analysis of UGCAUG frequencies revealed substantial enrichment in the intron immediately downstream of exons with increased inclusion in heart, skeletal muscle, brain and cerebellar cortex (Fig. 4e)—tissues where FOX proteins are highly expressed, suggesting common splicing activation activity in this location[22–24]. Enrichment of UGCAUG hexanucleotides was also noted upstream of exons that had reduced inclusion in skeletal muscle, suggesting possible repressive activity in this context. This example illustrates the power of these expanded tissue-specific exon sets for inference of 'tissue RNA maps', summarizing both the location-dependent activity and tissue specificity of splicing regulatory elements.

Applying a similar approach to analyse enrichment of all hexanucleotides in regions adjacent to tissue-specific exons identified 362 motif/tissue enrichment patterns (at an estimated 17% FDR), representing hexanucleotides that showed significant enrichment adjacent to exons with increased or decreased inclusion in specific cell lines or tissues (Supplementary Table 8). Enrichment of UGCAUG downstream of exons with high inclusion in skeletal muscle appears as the third most significant motif/tissue pair, after enrichment of UCUCUC and CUCUCU (resembling the binding motifs of PTBP1 (also known as PTB) and PTBP2 (nPTB)[25]) upstream of exons with increased inclusion in cerebellar cortex. The remaining motif/tissue pairs contained a variety of known regulatory elements, including ACUAAC (see later), as well as putative new regulatory motifs.

Coordination of splicing and polyadenylation

Tandem 3′ UTR events showed an even higher frequency of tissue-regulated expression than skipped exons or other alternative splicing events studied (Fig. 2), yet little is known about how tissue regulation of tandem UTRs is accomplished (for example, whether through APA or through the differential stability of alternative UTR isoforms). By grouping tandem 3′ UTRs by switch score, the most switch-like events showed increased sequence conservation relative to events with lower switch scores in the vicinity of and upstream of the proximal (5′) polyadenylation signal (PAS), and also upstream of the distal (3′) PAS (Fig. 5a). Whereas *cis*-regulatory elements contributing to differential stability should be located predominantly in the region unique to the long UTR isoform, APA could be regulated by elements located near to either or both PASs. The observation of increased conservation around and upstream of the proximal PAS in switch-like tandem UTRs therefore supports a primary role for regulation at the level of APA.

In assessing the spectrum of *cis*-elements that may drive tissue regulation of tandem 3′ UTRs, a set of heptanucleotides was identified that showed high conservation in the extension region of tandem 3′ UTRs (Fig. 5a, inset), with signal:background ratios in four mammals[26] exceeding 2:1. As expected, this set included the extended (seven-base) seed matches to a number of conserved mammalian microRNAs (miRNAs)[26–28]. Surprisingly, it also included all eight of the heptanucleotides that contain the FOX-1/FOX-2 consensus binding motif, UGCAUG: all such heptanucleotides had signal/background ratios above 2.5:1, exceeding the signal/background ratio observed for seed matches to important miRNAs such as miR-7 and miR-181 (inset, Supplementary Table 9). Strong conservation of UGCAUG motifs in this location (>1 kilobase on average from the nearest splice site) would not be expected on the basis of the canonical splicing regulatory activity of FOX-1/FOX-2 proteins. Instead, the high conservation observed in extended 3′ UTR regions suggests that these factors (or others with identical RNA-binding specificity) have additional 3′ UTR-related roles, for example, in APA or in mRNA localization and/or translation.

To investigate possible connections between tissue-specific regulation of alternative splicing and APA further, global patterns of tissue-specific alternative isoform expression were compared. By applying singular value decomposition (SVD) (Supplementary Methods) to the vectors of inclusion ratios across samples for each alternative splicing and APA event type separately, a strong and consistent separation of the breast cell lines (four cancer-derived and one immortalized cell line) from all tissue samples was observed (Fig. 5c,

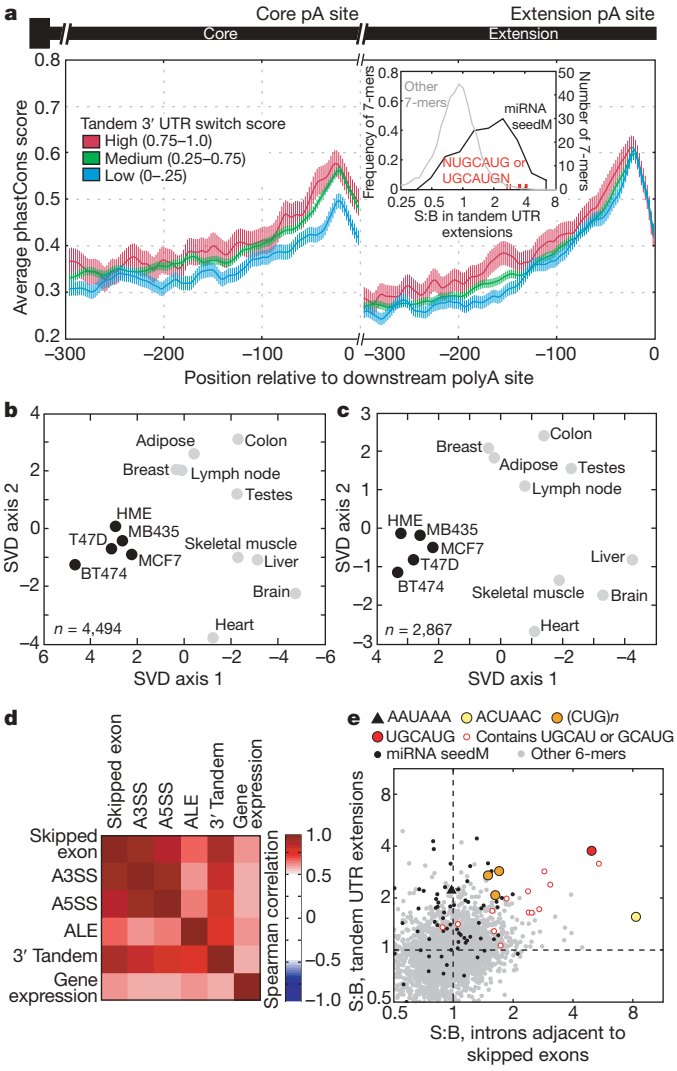


Figure 5 | Evidence for coordination between splicing and polyadenylation. **a**, Mean and s.e.m. of per-position phastCons score in the region 300 bp upstream of proximal and distal cleavage sites for tandem 3′ UTRs grouped by switch score. Inset, increased conservation of FOX-1/FOX-2 motifs in tandem 3′ UTR extension regions. All non-CpG-containing heptanucleotides (grey line), miRNA seed matches (black), and 7-mers containing UGCAUG (red) are shown. **b**, SVD analysis of skipped exon inclusion ratio values across tissues and cell lines for skipped exons meeting minimum read coverage criteria in each of the 14 samples. Projections are shown in the dimensions corresponding to the two leading eigenvalues, which accounted for 25% of the variance. **c**, SVD analysis of tandem UTR inclusion ratio values (as in **b**). **d**, SVD analysis was conducted for the 14 samples on the basis of inclusion ratio values for the five indicated alternative transcript event types or on the basis of gene expression values. Spearman correlations between corresponding pairwise distances in projections of the sort shown in **b** and **c** are shown. **e**, Signal:background (S:B) ratios of non-CpG-containing hexanucleotides in introns flanking skipped exons (*x*-axis) and in extended 3′ UTR regions (*y*-axis). The canonical PAS hexanucleotide AAUAAA (black triangle), hexanucleotides corresponding to seed matches to conserved mammalian miRNAs (black dots), hexanucleotides corresponding to binding motifs for the indicated splicing or 3′ UTR-binding factors (coloured), and other hexanucleotides (small gray dots) are shown.

d). This separation implied the existence of a systematic difference in RNA processing regulation between cell lines and tissues that held for all types of alternative events studied. For most alternative splicing and APA events, SVD analysis yielded similar groupings of tissues, for example, with heart, skeletal muscle, brain and liver consistently clustered (Supplementary Fig. 8). Consistent with this observation, pairwise distances between SVD projections for different types of alternative splicing events, for example, skipped exons, A5SS and A3SS events, were all highly correlated (Fig. 5e), suggesting similarities in the regulatory control of these types of events[1,2,13]. More surprisingly, distances between SVD projections for tandem 3′ UTR events also correlated highly with distances for events controlled purely at the level of splicing such as skipped exons (Fig. 5e). This observation raised the possibility that splicing and polyadenylation may be coordinately regulated across human tissues.

To explore possible regulatory connections between splicing and polyadenylation regulation (for example, refs 29–32), the conservation of hexanucleotides adjacent to conserved alternative splicing and APA events was compared. Whereas canonical 3′ UTR regulatory motifs such as the consensus PAS hexanucleotide AAUAAA and various miRNA seed matches showed high signal:background ratios, often 1.5:1 or higher, in extended 3′ UTR regions, these motifs generally had signal/background ratios close to 1:1 in alternatively spliced introns. However, a distinct subset of motifs with high signal:background ratios in both UTRs and introns was also observed, several of which corresponded to well-known splicing-related motifs (Fig. 5h and Supplementary Table 9). This set included not only the FOX-1/FOX-2 motif UGCAUG and variations, consistent with the heptanucleotide analysis of Fig. 5a, but also permutations of $(CUG)_n$, which represent putative substrates of the bruno-like (BRUNOL, also known as CELF) and muscleblind-like (MBNL) families of muscle- and brain-specific splicing factors[33]. The highly significant signal:background ratio in both 3′ UTRs and introns suggested that these well-known splicing-related motifs also commonly have 3′ UTR-related roles—for example, control of APA or of mRNA stability, localization or translation—as recently demonstrated for the NOVA family of splicing factors[34].

The hexanucleotide ACUAAC, an excellent match to the consensus binding motifs of STAR family RNA-binding factors, in particular quaking homologue (QKI)[35], was also notable. Not only did ACUAAC have significant signal:background ratio in 3′ UTRs, as expected from the known role of QKI in control of mRNA stability[36], but it also showed an extremely high signal:background ratio in introns, exceeding 7:1. This extreme conservation suggested a common and important function in splicing regulation—a role that has been suggested but not yet directly demonstrated[9,37]. Motif enrichment analyses also suggested a possible role in brain-specific APA regulation (Supplementary Fig. 9).

Discussion

We conclude that the coordination between tissue-specific alternative splicing and APA events implied by the correlated patterns of tissue bias observed in Fig. 5 may be mediated at least in part by tissue-specific RNA-binding factors that have roles in regulation of both of these RNA processing steps. Such factors may include both canonical tissue-specific splicing factors (for example, of the FOX-1/FOX-2 and CELF families), moonlighting in 3′ UTR-related roles, and also canonical UTR-binding factors such as QKI. Such functional duality has the potential to enable tightly coordinated regulation of polyadenylation and splicing, ensuring that the appropriate UTR regulatory sequences are expressed in conjunction with the coding regions for the relevant tissue-specific protein isoforms.

METHODS SUMMARY

Tissues and cell lines. Tissue samples from individual unrelated anonymous donors (Supplementary Table 1) were obtained from Ambion for the following tissue types: adipose, whole brain, breast, colon, heart, liver, lymph node, skeletal muscle and testes. Cerebellar cortex samples were obtained from six anonymous unrelated donors, according to NIH guidelines for confidentiality and privacy using protocols described previously[38]. HME is a human mammary epithelial cell line immortalized with human TERT[39]. The other cell lines are all breast cancer cell lines derived from invasive ductal carcinomas (ATCC). MCF-7, BT474 and T47D are oestrogen-receptor- and progesterone-receptor-positive; MDA-MD435 is negative for both.

Library preparation for Illumina sequencing. Poly-T capture beads were used to isolate mRNA from 10 μg of total RNA. First-strand cDNA was generated using random hexamer-primed reverse transcription, and subsequently used to generate second-strand cDNA using RNase H and DNA polymerase. Sequencing adaptors were ligated using the Illumina Genomic DNA sample prep kit. Fragments ~200 bp long were isolated by gel electrophoresis, amplified by 16 cycles of PCR, and sequenced on the Illumina Genome Analyser, as described[40,41].

Computational analyses of mRNA-Seq read data. Computational and statistical methods used in analysis of the read data are described in the Supplementary Methods. High-confidence new exons were required to be supported by at least one splice juction read involving each splice site, and at least one exon body read; putative new splice junctions required splice junction read support only[42].

Received 23 June; accepted 3 October 2008.
Published online 2 November 2008; corrected 27 November 2008 (details online).

1. Black, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72,** 291–336 (2003).
2. Matlin, A. J., Clark, F. & Smith, C. W. Understanding alternative splicing: towards a cellular code. *Nature Rev. Mol. Cell Biol.* **6,** 386–398 (2005).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
4. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141–2144 (2003).
5. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126,** 37–47 (2006).
6. Xu, Q., Modrek, B. & Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30,** 3754–3766 (2002).
7. Gupta, S., Zink, D., Korn, B., Vingron, M. & Haas, S. A. Strengths and weaknesses of EST-based prediction of tissue-specific alternative splicing. *BMC Genomics* **5,** 72 (2004).
8. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5,** R74 (2004).
9. Sugnet, C. W. *et al.* Unusual intron conservation near tissue-regulated exons found by splicing microarrays. *PLOS Comput. Biol.* **2,** e4 (2006).
10. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
11. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321,** 956–960 (2008).
12. Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444,** 580–586 (2006).
13. Wang, Z., Xiao, X., Van Nostrand, E. & Burge, C. B. General and specific functions of exonic splicing silencers in splicing control. *Mol. Cell* **23,** 61–70 (2006).
14. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnol.* **24,** 1151–1161 (2006).
15. Mayr, J. A. *et al.* Mitochondrial phosphate-carrier deficiency: a novel disorder of oxidative phosphorylation. *Am. J. Hum. Genet.* **80,** 478–484 (2007).
16. Graveley, B. R. The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. *Trends Genet.* **24,** 5–7 (2008).
17. Nembaware, V., Wolfe, K. H., Bettoni, F., Kelso, J. & Seoighe, C. Allele-specific transcript isoforms in human. *FEBS Lett.* **577,** 233–238 (2004).
18. Pan, Q. *et al.* Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Mol. Cell* **16,** 929–941 (2004).
19. Xing, Y. & Lee, C. J. Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet.* **1,** e34 (2005).
20. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nature Genet.* **40,** 225–231 (2008).
21. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA* **100,** 189–192 (2003).
22. Underwood, J. G., Boutz, P. L., Dougherty, J. D., Stoilov, P. & Black, D. L. Homologues of the *Caenorhabditis elegans* Fox-1 protein are neuronal splicing regulators in mammals. *Mol. Cell. Biol.* **25,** 10005–10016 (2005).
23. Auweter, S. D. *et al.* Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25,** 163–173 (2006).
24. Nakahata, S. & Kawamoto, S. Tissue-dependent isoforms of mammalian Fox-1 homologs are associated with tissue-specific splicing activities. *Nucleic Acids Res.* **33,** 2078–2089 (2005).
25. Oberstrass, F. C. *et al.* Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309,** 2054–2057 (2005).

26. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
27. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434,** 338–345 (2005).
28. Majoros, W. H. & Ohler, U. Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics* **8,** 152 (2007).
29. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416,** 499–506 (2002).
30. McCracken, S., Lambermon, M. & Blencowe, B. J. SRm160 splicing coactivator promotes transcript 3′-end cleavage. *Mol. Cell. Biol.* **22,** 148–160 (2002).
31. Castelo-Branco, P. *et al.* Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell. Biol.* **24,** 4174–4183 (2004).
32. Zhang, L., Lee, J. E., Wilusz, J. & Wilusz, C. J. The RNA-binding protein CUGBP1 regulates stability of tumor necrosis factor mRNA in muscle cells: implications for myotonic dystrophy. *J. Biol. Chem.* **283,** 22457–22463 (2008).
33. Ladd, A. N. & Cooper, T. A. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.* **3,** reviews0008.1–reviews0008.16 (2002).
34. Licatalosi, D. *et al.* Mechanisms of alternative mRNA processing in the brain revealed by HITS-CLIP. *Nature* doi:10.1038/nature07488 (this issue).
35. Galarneau, A. & Richard, S. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nature Struct. Mol. Biol.* **12,** 691–698 (2005).
36. Kim, H. H. & Gorospe, M. GU-rich RNA: expanding CUGBP1 function, broadening mRNA turnover. *Mol. Cell* **29,** 151–152 (2008).
37. Wu, J. I., Reed, R. B., Grabowski, P. J. & Artzt, K. Function of quaking in myelination: regulation of alternative splicing. *Proc. Natl Acad. Sci. USA* **99,** 4233–4238 (2002).
38. Paz, R. D. *et al.* Increased expression of activity-dependent genes in cerebellar glutamatergic neurons of patients with schizophrenia. *Am. J. Psychiatry* **163,** 1829–1831 (2006).
39. Elenbaas, B. *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. *Genes Dev.* **15,** 50–65 (2001).
40. Schroth, G. P., Luo, S. & Khrebtukova, I. Transcriptome analysis using high-throughput DNA sequencing. *Methods Mol. Biol.* (in the press).
41. Illumina, Inc. Transcriptome Analysis: mRNA-Seq. ⟨http://www.illumina.com/pages.ilmn?ID=291⟩ (2008).
42. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human genome by next generation sequencing. *Nature Genet.* (in the press).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. Anderson, D. Black, B. Friedman, and members of the Burge laboratory for comments on the manuscript, N. Spies for analyses, J. Mudge, G. D. May, N. A. Miller, E. Vermaas, T. Kerelska, J. Yan and V. Quijano for assistance in generating the mRNA-Seq data, and R. C. Roberts and N. Perrone-Bizzozero for supplying cerebellar cortex RNA samples. This research was supported by an NIH training grant (E.T.W.), and by grants from the Knut & Alice Wallenberg Foundation and the Swedish Foundation for Strategic Research (R.S.) and from the NIH (C.B.B.).

Author Contributions E.W. and R.S. designed and performed the computational analyses of sequencing reads, prepared figures, tables and methods and contributed to manuscript text. S.L. developed protocols and created libraries, L.Z. contributed to sequencing development, and I.K., S.L. and L.Z. did primary data analysis. G.P.S. contributed to study design and manuscript preparation. C.M. and S.F.K. provided RNA samples and contributed to manuscript preparation. C.B.B. designed the study and prepared the manuscript, with input from other authors.

Author Information The reported sequence read data have been deposited to the Short Read Archive section of GEO at NCBI under accession numbers GSE12946 and SRA002355.1. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to C.B.B. (cburge@mit.edu).