bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transcript isoform differences across human tissues are predominantly driven by alternative start and termination sites of transcription

Alejandro Reyes[1,2,3,*] **and Wolfgang Huber**[1,*]

1. European Molecular Biology Laboratory. Meyerhofstrasse 1, 69117 Heidelberg, Germany.
2. Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
3. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
* Corresponding authors: whuber@embl.de and alejandro.reyes.ds@gmail.com

Most human genes have multiple transcription start and polyadenylation sites, as well as alternatively spliced exons. While transcript isoform diversity contributes to shape cellular specificity, it is currently unclear what is the balance of contributions from alternative splicing compared to alternative start and termination sites of transcription. Here, we address this question by analyzing data from the Genotype-Tissue Expression Project. We found tissue-dependent usage of exons for around one-half of expressed genes. Although tissue-dependent splicing was frequent among untranslated exons, it explained less than half of the differences in exon usage across tissues, suggesting that most of these differences were driven by alternative transcription start and termination sites. Analysis of the FANTOM Project data confirmed widespread tissue-dependent usage of alternative transcriptional start sites. Our analysis highlights alternative initiation and termination sites of transcription as the main drivers of isoform diversity across tissues. We also show that most tissue-dependent splicing is unlikely to have consequences at the proteome level.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Alternative splicing, alternative promoter usage and alternative polyadenylation are interdependent molecular processes that enable genomes to synthesize multiple transcript isoforms per gene [1]. In mammalian genomes, it has been estimated that at least 70% of genes use multiple transcriptional sites, more than 50% of genes have alternative transcriptional start sites and nearly all genes undergo alternative splicing [2,3,4]. Hence, these molecular processes increase the repertoire of transcripts in mammalian genomes [5,6,7].

A variety of biological processes, including cell differentiation and cell proliferation, are regulated by the expression of alternative transcript isoforms. For example, the alternative usage of two promoters of the gene *RUNX1* is spatio-temporally modulated during mouse embryogenesis, and the disruption of this modulation leads to impaired development [8]. Similarly, the transcription factor *FOXP1* is differentially spliced between embryonic stem cells and differentiated cells and the regulation of this splicing switch is essential for the maintenance of ESC pluripotency and efficient reprogramming [9]. Mis-regulation of transcript isoform regulation has been observed in disease phenotypes, including cancer [10,11,12,13].

A large majority of alternatively spliced RNAs have been detected to be bound to ribosomes, suggesting that most transcript isoforms could be translated into protein products [14]. For dozens of genes, it has been experimentally demonstrated that alternative isoforms result in proteins with differences in the cellular localization, stability, DNA binding properties, lipid binding properties and enzymatic activities [15]. Alternative isoforms can behave like completely distinct proteins when considering their protein-protein interaction capabilities [16]. However, based on evidence from conservation of protein structures and functional features, it has been suggested that the majority of alternative transcript isoforms would be translated into proteins with disrupted structures and functions [17,18]. Such 'non-functional proteins' are not abundant at levels that can be detected with high confidence in large-scale proteomic experiments [19,20]. A large fraction of transcript isoforms might have functions at the RNA level rather than at the protein level. Indeed, the expression of multiple transcript isoforms of the same gene can enhance post-transcriptional regulation, resulting in fine tuning of mRNA levels [21,22] and translation control [23,24].

Although transcript isoform diversity contributes to phenotypic differences across cell-types, it is currently unclear to what extent it is driven by alternative splicing, and to what extent by alternative transcription initiation and polyadenylation sites. Here, we developed an analytic strategy to approach this question using data from the Genotype-Tissue Expression (*GTex*) Project V6 [25]. We show that transcript isoform regulation across tissues is widespread in the human genome and affects about one half of multi-exonic genes. We show that the majority of differences in exon usage cannot be explained by alternative splicing, suggesting that these differences are due to the alternative usage of start and termination sites of transcription. Integration of data from the *FANTOM* consortium [6] confirms prevalent tissue-dependent usage of alternative transcription start sites. Our results suggest that alternative splicing has more widespread repercussions on RNA products rather than on protein products and that alternative transcript start and polyadenylation sites play an under-explored role in defining cell-type specificity.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

RESULTS

The data

We used transcriptome data (RNA-seq) from the *V6* release of the Genotype-Tissue Expression (*GTEx*) project[25]. The dataset comprised 9,795 RNA-seq samples from 54 tissues out of a total of 551 humans. Since the dataset did not contain all tissues for each individual, we identified subsets of data that could be analyzed as fully crossed designs (i.e., containing all possible tissue-individual combinations). We mapped the sequenced fragments to the human reference genome version *GRCh38*, obtained from *ENSEMBL* release 84[26], using the aligner *STAR v2.4.2a*[27]. We excluded samples where the number of sequenced fragments was below 1,000,000 or where the percentage of uniquely mapping reads was below 60%. Using these filtering criteria, we defined three subsets of *GTEx* data:

- *Subset A* consisted of 8 brain cell-types (frontal cortex [BA9], nucleus accumbens, putamen, cortex, cerebellum, caudate, cerebellar hemisphere and hippocampus) across 30 individuals, comprising a total of 240 samples;
- *Subset B* included 9 tissues (skeletal muscle, thyroid, whole blood, lung, subcutaneous adipose, skin, tibial artery, tibial nerve and esophagus [mucosa]) from 34 individuals, i.e., a total of 306 samples;
- *Subset C* comprised 6 tissues (heart, aorta, esophagus [muscularis], colon and stomach) from 42 individuals (252 samples).

Subsets A, B and C were non-overlapping, and altogether our analysis employed a subset of 798 samples from the *GTEx* dataset.

Exon usage coefficients enable the study of transcript isoform regulation across tissues.

Next, we used *ENSEMBL* transcript annotations to define reduced gene models with non-overlapping exonic regions[28] (Online Methods). We obtained 499,667 non-overlapping exonic regions from the 35,048 multi-exonic genes, of which 412,116 were derived from 18,295 protein-coding genes. For each of the subsets A, B and C, we computed two measures of exon usage per exonic region: relative exon usage coefficients (*REUCs*)[29] and relative spliced-in coefficients (*RSICs*)[4]. The computation of these coefficients is illustrated in Figure 1. Both of them are measures of exon usage in a specific tissue in a particular individual relative to the average exon usage across all tissues and individuals (Online Methods). The *REUC* defines exon usage as the fraction of sequenced fragments that map to the exonic region among all fragments mapping to the rest of the exonic regions from the same gene. In contrast, the *RSIC* measures the fraction of sequenced fragments that map to the exonic region compared to the number of reads that support the skipping of that exonic region via alternative splicing (Figure 1C). Note that differences in exon usage due to alternative splicing are reflected in both *REUCs* and *RSICs* (Figure 1B and Figure 1C). Changes in exon usage due to alternative transcription initiation sites or alternative polyadenylation sites, which do not result in exon-exon junction reads, are only reflected in *REUCs* (Figure 1D).

As an example, we show the gene *5-Aminolevulinate Synthase 1* (*ALAS1*) in Figure 1D-E. *ALAS1* encodes an enzyme that performs the first catalytic reaction of the

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

biosynthetic pathway of heme. Heme is an organic cofactor that is essential for the proper function and differentiation of many cell types, including those of the hematopoietic, hepatic and nervous systems[30]. Induction of *ALAS1* has been associated with acute attacks of porphyria disease[31]. By exploring the *REUCs* for *ALAS1*, we found that a 5' untranslated exon was included more frequently in the transcripts generated from cerebellum and cerebellar cortex than in the other brain cell types (Figure 1D). The same pattern of tissue-dependent usage (*TDU*) was also evident from the *RSICs* (Figure S1), which indicates that the *TDU* pattern is a consequence of alternative splicing rather than of alternative transcriptional initiation or termination sites (Figure 1E). In cultured cells, it has been shown that *ALAS1* transcripts that include this 5' exon are resistant to heme-mediated decay and that their translation is inhibited[32]. The detected splicing pattern indicates that *ALAS1* might be post-transcriptionally regulated differently in cerebellum than in the rest of the brain.

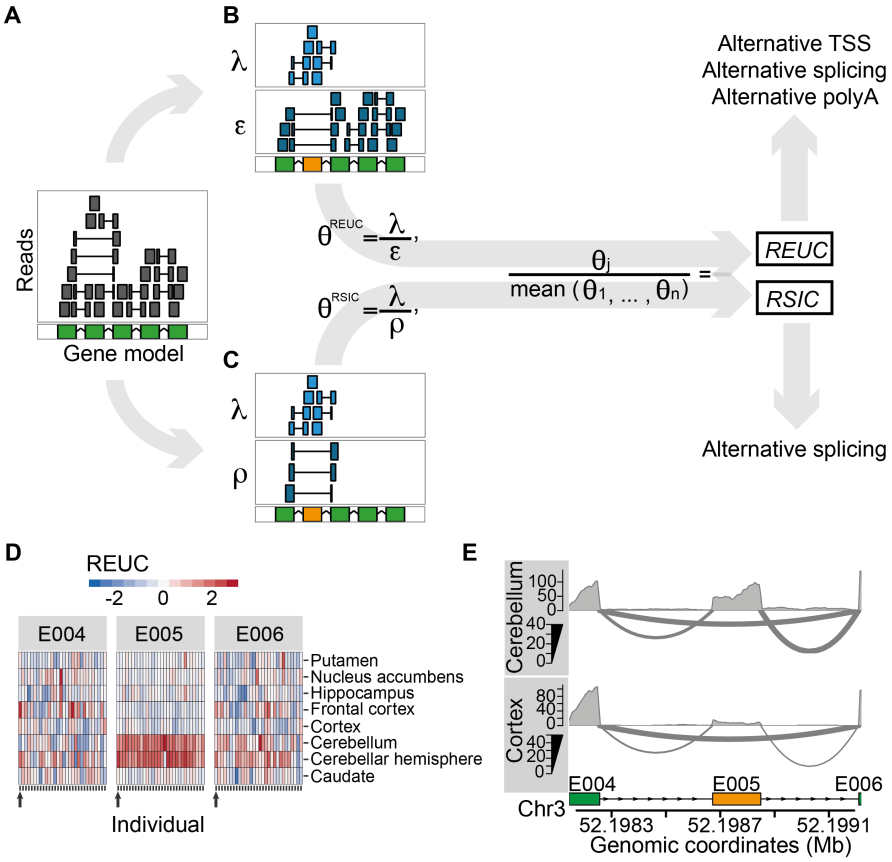

Figure 1: Caption in next page.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1: **Quantification of exon usage.** **(A)** Exemplary gene model in the reference genome (green) and alignments of RNA-seq reads (upper panel). Sequenced fragments whose alignments fall fully into an exonic region are shown by a grey box; alignments that map into two (or possibly more) exonic regions are shown by shorter grey boxes connected by a horizontal line. For a particular exon (highlighted in orange), we consider two strategies to quantify its usage that are illustrated in Panels **B** and **C** (see Online Methods for the formal description). The first strategy is illustrated in Panel **B**, where sequenced fragments are counted into two groups: the number of sequenced fragments that map fully or partially to the exon (λ) and the number of fragments that map to the rest of the exons (ϵ). θ^{REUC} is defined as the ratio between λ and ϵ, and the relative exon usage coefficient (*REUC*) for the exon in sample j is estimated as the ratio between θ^{REUC} in sample j with respect to the mean θ^{REUC} across the n samples. Panel **C** illustrates the second exon usage quantification strategy, where sequenced fragments are also counted into two groups: the number of sequenced fragments that map fully or partially to the exon (λ) and the number of sequenced fragments that align to exons both downstream and upstream of the exon under consideration (ρ). The latter group of sequenced fragments are derived from transcripts from which the exon was spliced out. θ^{RSIC} is now defined as the ratio between λ and ρ. The relative spliced-in coefficient (*RSIC*) for the exon in sample j is estimated as the ratio between θ^{RSIC} in sample j with respect to the mean θ^{RSIC} across the n samples. Notice that while differences in exon usage due to alternative splicing are reflected in both *REUCs* and *RSICs*, differences in exon usage due to alternative transcription or termination sites are only reflected in *REUCs*. **(D)** Heatmap representations of the *REUCs* for three exonic regions (*E004*, *E005* and *E006*) of the gene *ALAS1*, computed using *Subset A* of the *GTEx* data. The rows of the heatmaps correspond to the eight tissues, and each column corresponds to one individual. The horizontal colour patterns of exon *E005* indicate elevated inclusion of cerebellum and cerebellar cortex as compared to the rest of the brain cell-types. **(E)** RNA-seq samples from two cell-types (cortex and cerebellum) from individual *12ZZX* (also indicated by the arrows below each heatmap in Figure 1D) are displayed as sashimi plots[33]. The three exonic regions presented in Panel D are shown. The middle exon, *E005*, is an untranslated cassette exon (ENSEMBL identifier ENSE00002267562) that is spliced out more frequently in cortex than in cerebellum.

Tissue-dependent usage of exons is widespread in humans.

After observing further instances of tissue-dependent exon usage as in the *ALAS1* example, we asked how common this phenomenon was across the human genome. We defined a score based on the *REUCs* that measures the strength of tissue-dependent expression of an exonic region (Online Methods). Based on this, we considered an exonic region to be tissue-dependent if its differential usage pattern was statistically significant at a false discovery rate (FDR) of 10%, according to the *DEXSeq* method[28] and if it had a score larger than 1. We thus detected tissue-dependent usage (TDU) in at least one of the subsets of the *GTEx* data for 23% of the exonic regions (116,601 out of 524,219; Figure S3). Specifically, in *subset A* we retrieved 47,659 exonic regions of 9,839 genes with TDU, in *subset B* 76,562 exonic regions of 12,295 genes, and in *subset C* 30,719 exonic regions of 7,025 genes (Figure 2A and Figure S2). Remarkably, a

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

large fraction of genes detected as expressed were subject to transcript isoform regulation across tissues (Table S1). For example, out of the 18,805 multi-exonic genes that had an average of sequenced fragments of at least 10 in *subset A*, 51% (9,672) showed differential usage in at least one exonic region. Furthermore, the set of genes with TDU was enriched for protein-coding genes as compared to a background set of genes matched for expression strength and number of exonic regions (p-value $< 2.2 \cdot 10^{-16}$, odds-ratio = 3.4, Table S2).

We next investigated the nature of transcript isoform differences between tissues. For each gene containing exons with TDU, we estimated the fraction of exonic regions that were subject to TDU and the fraction of affected exonic nucleotides. TDU effects were localized to a relatively small fraction of exons for most genes (Figure 2B, Figure 2C and Figure S4). For instance, in *subset A* the percentage of exonic regions with TDU was below 25% for 70% (6,929) of the genes. In the same subset, the percentage of nucleotides affected by TDU was below 25% for 53% (5,248) of the genes (Table S3). The rest of the cases, where a larger fraction of the gene length was used in a tissue-dependent manner, were instances in which very different isoforms were expressed by different tissues. For example, all 27 exonic regions of the gene *Erythrocyte Membrane Protein Band 4.1 Like 4B* (*EPB41L4B*) were detected to be used in a tissue-dependent manner in *subset B* of the *GTEx* data: in tibial nerve and skeletal muscle, the first 16 exonic regions (counting from 5' to 3') had lower usage while the exonic regions located towards the end of the gene showed increased usage (Figure S5 and Figure S6). In this case, the observed pattern can be explained by the two annotated transcript isoforms of the gene: whereas most cell-types in *subset B* tend to express the short isoform, tibial nerve and skeletal muscle express the longer isoform more frequently (Figure 2D).

Our results indicate that differences in transcript isoform regulation across tissues is a widespread phenomenon across the human genome. It preferentially affects protein-coding genes. For hundreds of genes, different tissues express very distinct transcript isoforms at different levels.

Alternative transcriptional initiation and termination sites drive most transcript isoform differences between tissues.

In the previous section we saw the example of *EPB41L4B* for a transcript isoform switch that is not driven by alternative splicing, and instead, by the usage of an alternative polyadenylation site (here also referred to as transcription termination site) (Figure 2D). We asked more globally: among the thousands of instances of tissue-dependent exon usage, what fraction is driven by alternative splicing and what by alternative transcriptional start or termination sites?

For each exonic region, we searched for evidence of alternative splicing by counting, in each sample, the number of sequenced fragments that supported exon skipping (as sketched in Figure 1C). We found that only a minor fraction of exonic regions with tissue-dependent usage (TDU) had appreciable evidence of being spliced out from transcripts (Table S4). For instance, the mean of read counts supporting exon skipping was larger than 10 in only 30% (9,282) of the exonic regions with TDU in *subset C*. On the other hand, 58% (17,815) showed no or only weak evidence of being alternatively

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

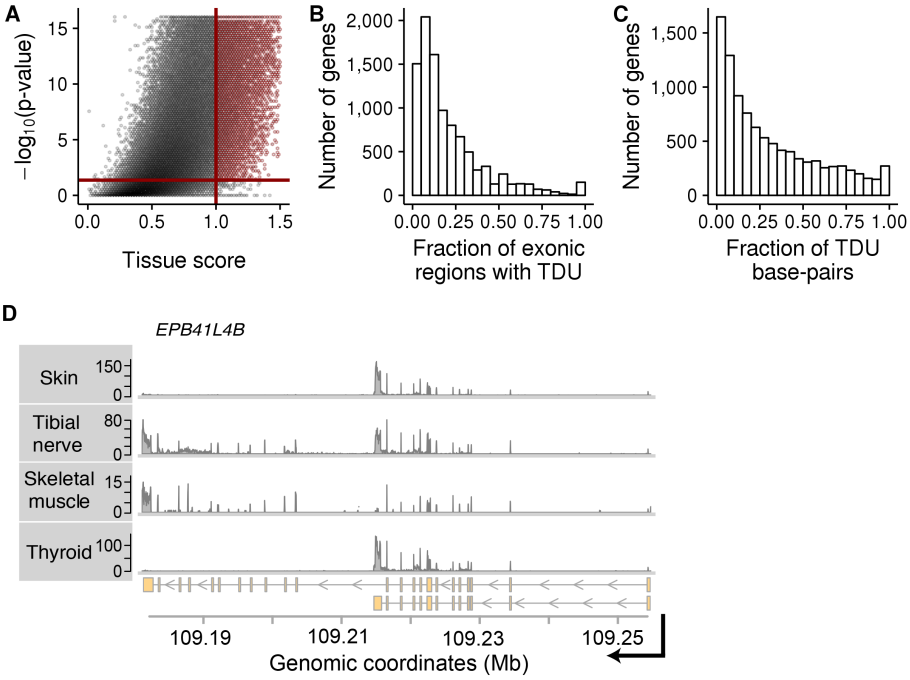

Figure 2: **Tissue-dependent exon usage is widespread in the human genome.** Panels **A**, **B** and **C** show data from *subset A* of the *GTEx* data. The same plots using data from subsets *B* and *C* can be found in Figure S2 and Figure S4. **(A)** Similar to a volcano-plot, this figure shows statistical significance (p-value on $-\log_{10}$ scale) versus effect size (tissue score) of our tissue-dependence test for each exonic region of the human genome. The solid red lines show the thresholds used in this study to call an exonic region tissue-dependent. The p-value threshold $4.28 \cdot 10^{-2}$ corresponds to an *adjusted p-value* of 0.1 according to the Benjamini-Hochberg method to control false discovery rate (FDR) for all p-values shown. **(B)** Histogram of the fraction of exonic regions within each gene that are subject to tissue-dependent usage (x-axis). The y-axis shows the number of genes. **(C)** Similar to Panel B, but expressed in terms of fraction of base-pairs within a gene affected by tissue-dependent usage. **(D)** Exemplary data from four tissues of individual *131XE*. Shown is RNA-Seq coverage (y-axis) plots along genomic coordinates (x-axis) of chromosome 9 in the locus of the gene *EBP41L4B*. The lower panel shows the transcript annotations for the gene *EPB41L4B*. Skin and thyroid express short isoforms of *EPB41L4B*, while tibial nerve and skeletal muscle express longer isoforms.

spliced (Figure 3A and Figure S7). We estimated that alternative splicing explains tissue-dependent transcript differences for at most 36% of the genes (Table S5).

As a second line of evidence, we quantitatively compared the relative exon usage and spliced-in coefficients (*REUCs* and *RSICs*, as defined above). For each exonic region and each subset of the *GTEx* data, we fit two analysis-of-variance models, one for the *REUCs* and one for the *RSICs*, using tissues and individuals as categorical

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

covariates. We determined the coefficient of partial determination (R^2) of the tissue covariate for each fit. A large value of R^2 in the *RSIC* fit indicates that the TDU is due to alternative splicing. Conversly, a large R^2 in the *REUC* fit indicates that the TDU is due to any of alternative splicing, alternative transcription initiation sites or alternative transcriptional termination sites. For the minority of exonic regions with TDU that also had strong evidence of alternative splicing, the *REUCs* and the *RSICs* were highly correlated, and their R^2 statistics were in good agreement, confirming that the TDU was due to alternative splicing (Figure 3B, Figure 3C and Figure S7). Nevertheless, for the majority of exonic regions with TDU, the TDU was consistent with alternative transcription initiation and termination sites.

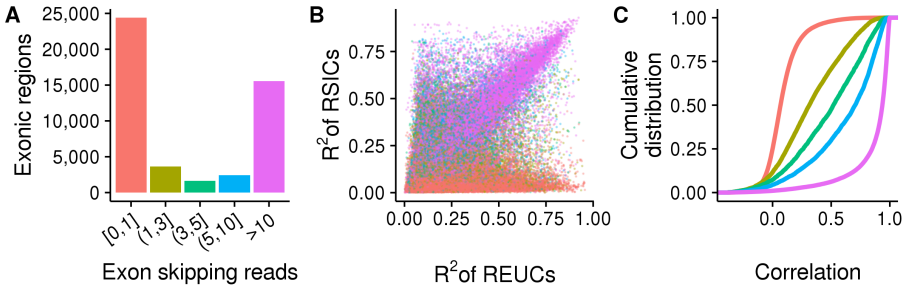

Figure 3: **Alternative splicing underlies only a minor fraction of exons with TDU, while the rest are consistent with alternative transcription start or stop sites.** The three panels show data from *subset A* of the *GTEx* data. Analogous plots using data from subsets *B* and *C* are shown in Figure S7. **(A)** The heights of the bars show the number of exonic regions with TDU, grouped according to the number of reads that support their splicing out from transcripts. Most exonic regions with TDU have either no or weak evidence of being spliced out from transcripts (bar colored in pink salmon). The bar colors serve also as color legends for Figure 3B and Figure 3C. **(B)** Each point represents one of the 47,659 exonic regions that were detected to be used in a tissue-dependent manner. The x-axis shows the fraction of *REUC* variance that is attributed to variance between tissues (R^2). Analogously, the y-axis shows the R^2 statistic for the *RSICs*. Exonic regions with strong evidence of being spliced out from transcripts (purple points) lay along the diagonal. **(C)** Cumulative distribution functions of the Pearson correlation coefficients between the REUCs and the RSICs are shown for exonic regions with TDU. The regions are stratified according to the number of sequenced fragments supporting their splicing out from transcripts. The *REUCs* and *RISCs* are highly correlated for the minor fraction of exons that have strong evidence of being spliced out from transcripts (purple line).

Analysis of CAGE data confirms prevalent tissue-dependent usage of alternative transcription start sites.

To further investigate the hypothesis that alternative splicing does not drive most transcript isoform diversity across tissues, we analyzed the Cap Analysis of Gene Expression (CAGE) data from the *FANTOM* consortium[6]. These data provide genome-wide quantitative information of transcriptional start sites (TSS) for many cell-types. For

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

each subset of the *GTEx* data, we generated a subset of *FANTOM* samples with the same composition of cell-types (as long as the samples existed and had replicates). For instance, based on the cell-types from *subset A* of the *GTEx* data, we selected a set of *FANTOM* samples consisting of caudates, cerebellums, cortexes, hippocampus and putamens. Then, for each of the three subsets of the *FANTOM* data, we tested each gene for changes in the relative usage of alternative TSS across cell-types. At a false discovery rate of 10%, we found 2,402, 6,763 and 2,778 genes with tissue-dependent usage of TSS across subsets *A*, *B* and *C*, respectively. Furthermore, the three lists of genes with differential TSS usage were in very good agreement with the counterpart lists of genes with TDU from the *GTEx* subsets (Table S6). When considering the genes with differential TSS usage across cell-types, 79% (1,904) of *subset A*, 80% (5,427) of *subset B* and 60% (1,657) of *subset C* also showed transcript isoform regulation in the corresponding *GTEx* subsets.

Figure 4 shows three examples of genes with TDU patterns that were explained by the usage of alternative TSS. In *subset A*, we found that the gene *Growth Arrest Specific 7* (*GAS7*) expressed tissue-specific isoforms (Figure S8). From the coverage of sequenced RNA fragments along the genome, we suspected that transcription initiated more upstream in cerebellum compared to cerebral cortex. The CAGE data revealed 5 major clusters of TSS for *GAS7*, of which 2 were strongly used in cerebellum and were practically absent from cerebral cortex (Figure 4A). The differential usage of these 2 TSS clusters explained the upstream transcription seen in cerebellum that was not observed in cerebral cortex. Similarly, by exploring the data for the gene *Keratin 8* (*KRT8*) in *subset B*, we found patterns of TDU that were very prominent in thyroid tissue compared to subcutaneous adipose tissue (Figure S9). These patterns of TDU were explained by the usage of a TSS located in the middle of the gene body that resulted in the expression of shorter transcript isoforms. This internal TSS of *KRT8* was used very frequently in thyroid tissue and was absent in subcutaneous adipose tissue (Figure 4B). We found the exact same pattern for the gene *Nebulette* (*NEBL*) in *subset C* of the data. For this gene, the usage of an internal TSS resulted in transcript isoforms that excluded several 5' exons. This internal TSS was used very frequently in heart tissue, while it was absent in pancreas tissue (Figure 4C, Figure S10).

Our integrative analysis of two orthogonal sources of data (independent samples, different technologies) confirms that there is an abundance of alternative TSSs that are used in a tissue-dependent manner and that result in TDU.

Tissue-dependent splicing of protein-coding exons is rare.

We asked which regions of genes were subject to tissue-dependent exon usage. We integrated information from the ENSEMBL and APPRIS databases to annotate each exonic region. Importantly, APPRIS uses information about protein structures, functional data and cross-species conservation to infer which transcript isoforms are likely to code for functional proteins. APPRIS flags transcript isoforms that are predicted to code for functional proteins as principal isoforms, while the rest of the transcripts are marked as non-principal isoforms[17]. Using these sources of information, we classified each exonic region into 5 categories: (1) exonic regions coding for principal isoforms, (2) exonic regions coding only for non-principal isoforms, (3) 5' untranslated exonic regions (5' UTR), (4) 3' untranslated exonic regions (3' UTR), and (5) untranslated ex-

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

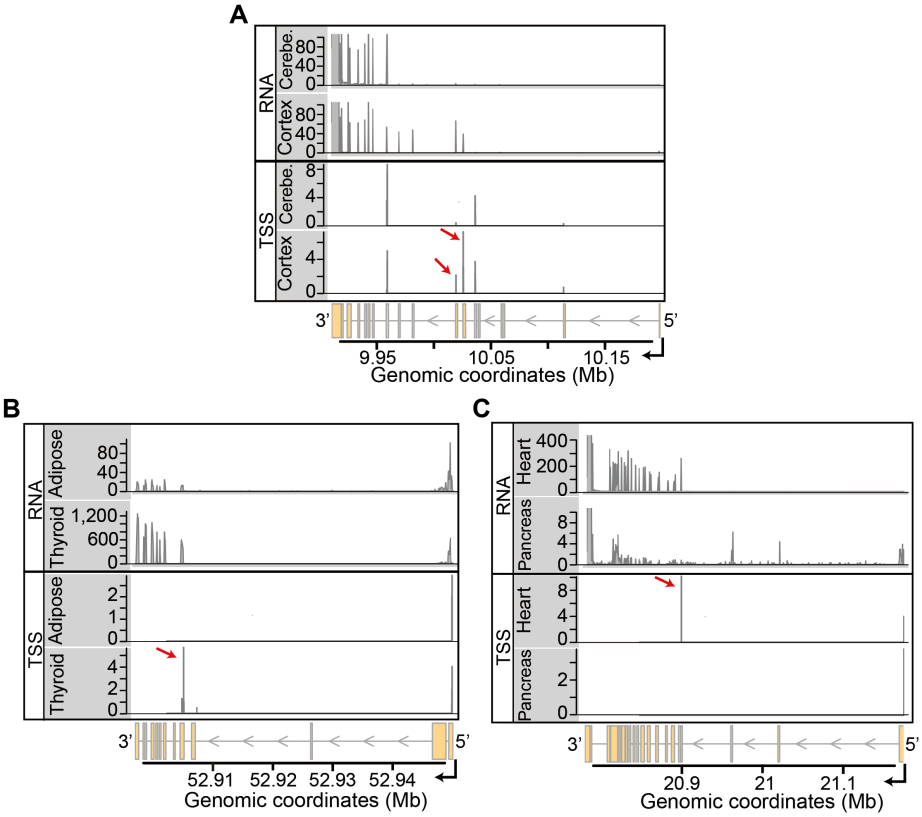

Figure 4: **Integration of RNA-seq and CAGE data.** Each panel displays an example of a gene where the usage of alternative transcription start sites explains the patterns of TDU. **(A)** Coverage tracks (y-axes) of RNA-seq and CAGE data for cerebral cortex and cerebellum are shown along the genomic coordinates (x-axis) of the locus of gene *GAS7*, located on chromosome 17. The upper two tracks show RNA-seq data from individual *12ZZX*. The lower two tracks show mean CAGE counts (on log_2 scale) for each annotated TSS. Cortex uses two transcription start site clusters (pointed by the red arrows) that are absent in cerebellum. The differential usage of these two TSS explains the upstream RNA-seq coverage seen in cortex. **(B)** Same as in Figure 4A, but showing data of thyroid and subcutaneous adipose tissue along the genomic coordinates of the *KRT8* locus on chromosome 12. The RNA-seq data corresponds to the individual *11EI6*. The internal TSS cluster that is indicated by the red arrow is strongly used in thyroid tissue, resulting in the expression of short transcript isoforms. **(C)** Same as in Figure 4A, but showing data of heart and pancreas along the genomic coordinates of the *NEBL* locus on chromosome 10. The RNA-seq data corresponds to the individual *ZF29*. In heart, the usage of an internal TSS (indicated by the red arrow) results in the expression of transcript isoforms that exclude several 5' exons of the gene.

ons belonging to non-coding processed transcripts. Then, for each subset of the *GTEx* data, we generated a background set of exons with the same distributions of both mean

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

counts and exon widths as the set of exonic regions with TDU.

We found that the proportions among the 5 exon categories were very different between exonic regions with tissue-dependent usage due to alternative splicing (TDU-AS), exonic regions with TDU but no evidence of alternative splicing (TDU-NAS) and the background sets of exons (p-value $< 2.2 \cdot 10^{-16}$, χ^2-test, Figure 5A, Figure S11 and Table S7). Specifically, exonic regions with TDU-AS were depleted among those coding for principal isoforms and enriched among exonic regions coding for non-principal isoforms and 3' UTRs. Our analysis also revealed that exons from non-coding processed transcripts, despite being weakly expressed, were alternatively spliced very frequently in a tissue-dependent manner (Figure 5, Figure S11 and Figure S12).

We also found that exonic regions with TDU-NAS showed a slight yet significant enrichment among 5' UTR exons compared to the background (p-value $< 1.2 \cdot 10^{-7}$, χ^2-test, Figure 5A and Figure S11). TDU-NAS occurred frequently among 3'UTR regions compared to the background (p-value $< 1.2 \cdot 10^{-7}$, χ^2-test). However, this enrichment among 3'UTR regions was observed only in subsets *B* and *C* of the *GTEx* data (Figure S11).

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

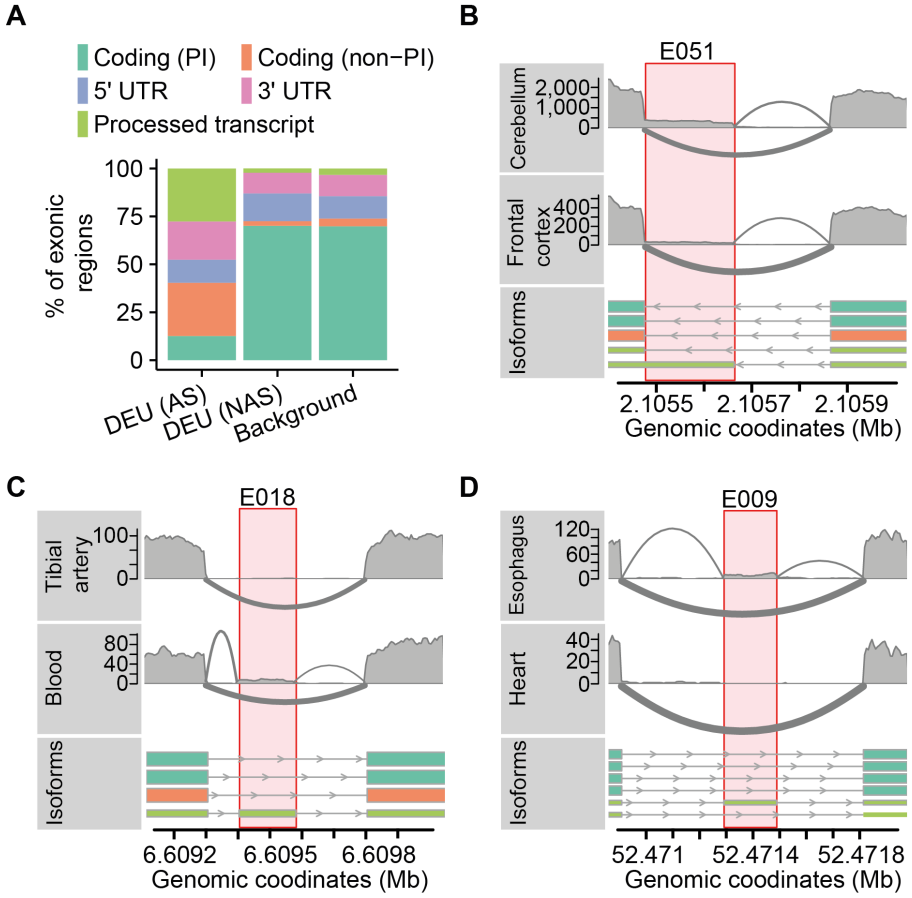

Figure 5: **Alternative splicing is infrequent among coding exons. (A)** The percentage of exonic regions (y-axis) is shown for three subsets of exons: (1) exonic regions with TDU due to alternative splicing [DEU (AS)], (2) exonic regions with TDU without evidence of alternative splicing [DEU (NAS)] and (3) a background set of exons matched for expression and exon width. Each color represents a different category of exons according to transcript biotypes: exons coding for principal transcript isoforms [Coding (PI)], exons coding for non-principal transcript isoforms [Coding (non-PI)], 5' UTRs, 3' UTRs and exons from non-coding processed transcripts [Processed transcripts]. **(B)** Sashimi plot representation of the RNA-seq data from frontal cortex and cerebellum of individual *WL46*. The lower data track shows the transcript isoforms of the gene *PKD1*. The transcripts are colored according to their biotype (the color legend is the same as in Figure 5A). The highlighted exon (*E051*) belongs to a non-coding transcript and is differentially spliced across tissues. **(C)** Same as in Figure 5B, but showing data from tibial artery and whole blood of the individual *ZTPG*. Transcripts from the gene *MAN2B2* along chromosome 4 are shown. The highlighted exon (*E018*) belongs to a non-coding transcript and is differentially spliced across tissues. **(D)** Same as in Figure 5B, but showing data from esophagus tissue (muscularis) and heart tissue (left ventricle) of the individual *111YS*. The lower track shows the transcripts annotated for gene *NISCH* along chromosome 4. The highlighted exon (*E009*) belongs to a non-coding transcript and is differentially spliced across tissues.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DISCUSSION

Our analysis highlights two aspects of transcript isoform diversity that have been underappreciated. First, alternative splicing is not the main driver of transcript isoform diversity across tissues. Instead, most transcript isoform differences are consistent with the alternative usage of transcription start and polyadenylation sites. Second, most tissue-dependent splicing is unlikely to have consequences for protein products, despite producing multiple isoforms at the transcriptome level.

Two previous studies reported similar observations about the contribution of alternative splicing to the diversity of transcript isoforms[34,35]. These studies, however, considered a very limited number of samples and did not study differences across cell-types. Here, we analyzed 798 *GTEx* transcriptomes covering 23 different cell-types. These data, together with the analytical approach illustrated in Figure 1A, enabled us to underline alternative transcription initiation and polyadenylation sites as the principal sources of transcript isoform diversity across cell-types. It has been suggested that gene expression is the main driver of cell-type specificity, with splicing playing a complementary role[36]. Our analysis suggest that alternative transcription and polyadenylation sites are also be essential to define phenotypic differences across tissues.

Transcriptome-wide studies have shown that, in a given cell-type, most genes express one major isoform at high levels, while the rest of the isoforms (i.e. the minor isoforms) are expressed at weaker levels[37,7]. Alternative splicing events were not detected with high confidence in a reanalysis of large-scale proteomic experiments[19]. Importantly, the protein isoforms detected in large-scale proteomic experiments are consistent with both the major RNA isoforms and the principal isoforms from the APPRIS database[18]. Extending and broadening these observations, our results suggest that most tissue-dependent splicing does not have consequences at the proteome level. We support this thesis with three main sources of evidence:

- First, tissue-dependent splicing is enriched among untranslated exons, particularly among exons from non-coding transcript isoforms.
- Second, tissue-specific splicing is depleted among exons coding for principal protein isoforms.
- Third, the exon categories where tissue-dependent splicing is more common are weakly expressed. Thus, the patterns of tissue-specific splicing are explained by tissue-specific expression of minor transcript isoforms.

The remaining open question is, if tissue-dependent splicing has little effect at the proteome level, what are its functions, if any, at the transcriptome level? A parsimonious answer may involve post-transcriptional regulation. Recent CRISPR-mediated interference screens identified 499 long non-coding RNAs that were essential for cellular growth, of which 89% showed growth-modifying phenotypes that were exclusive to one cell-type[38]. Similar screens at the transcript isoform level would be helpful to evaluate the essentiality of the thousands of non-coding and non-translated tissue-specific isoforms derived from protein-coding loci.

Alternative usage of promoters, alternative splicing of exons and alternative usage of polyadenylation sites are highly interleaved[1,39]. In the light of our results, it is conceivable that such coordination of decisions on the usage of transcription start sites, alternative exons and transcriptional termination is often cell-type specific. The un-

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

derstanding of such coordination will be facilitated by the development of sequencing technologies that provide quantitative measurements of full-length transcripts[40,41,42]. Furthermore, while alternative splicing may have limited effects for protein products, future research is needed to uncover to what extent tissue-dependent usage of alternative start and termination sites results in truncated versions of proteins.

Altogether, our results provide a step forward to the understanding of the human genome and its gene products, with wide-ranging implications for structural biology, protein interaction networks and drug design.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[1] de Klerk, E. & t Hoen, P. A. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends in Genetics* **31**, 128139 (2015). URL http://dx.doi.org/10.1016/j.tig.2015.01.001.

[2] Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**, 626635 (2006). URL http://dx.doi.org/10.1038/ng1789.

[3] Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology* **18**, 1830 (2016). URL http://dx.doi.org/10.1038/nrm.2016.116.

[4] Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470476 (2008). URL http://dx.doi.org/10.1038/nature07509.

[5] Ni, T. *et al.* Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* **14**, 615 (2013). URL http://dx.doi.org/10.1186/1471-2164-14-615.

[6] Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462470 (2014). URL http://dx.doi.org/10.1038/nature13182.

[7] Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101108 (2012). URL http://dx.doi.org/10.1038/nature11233.

[8] Pozner, A. *et al.* Developmentally regulated promoter-switch transcriptionally controls Runx1 function during embryonic hematopoiesis. *BMC Developmental Biology* **7**, 84 (2007). URL http://dx.doi.org/10.1186/1471-213X-7-84.

[9] Gabut, M. *et al.* An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* **147**, 132146 (2011). URL http://dx.doi.org/10.1016/j.cell.2011.08.023.

[10] Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600604 (2016). URL http://dx.doi.org/10.1126/science.aad9417.

[11] Xiong, H. Y. *et al.* The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 12548061254806 (2014). URL http://dx.doi.org/10.1126/science.1254806.

[12] Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673684 (2009). URL http://dx.doi.org/10.1016/j.cell.2009.06.016.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[13] Sendoel, A. *et al.* Translation from unconventional 5' start sites drives tumour initiation. *Nature* (2017). URL http://dx.doi.org/10.1038/nature21036.

[14] Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nature Structural & Molecular Biology* **23**, 11171123 (2016). URL http://dx.doi.org/10.1038/nsmb.3317.

[15] Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 130 (2013). URL http://dx.doi.org/10.1016/j.gene.2012.07.083.

[16] Yang, X. *et al.* Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805817 (2016). URL http://dx.doi.org/10.1016/j.cell.2016.01.029.

[17] Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research* **41**, D110D117 (2012). URL http://dx.doi.org/10.1093/nar/gks1058.

[18] Tress, M. L., Abascal, F. & Valencia, A. Alternative splicing may not be the key to proteome complexity. *Trends in Biochemical Sciences* (2016). URL http://dx.doi.org/10.1016/j.tibs.2016.08.008.

[19] Ezkurdia, I. *et al.* Most highly expressed protein-coding genes have a single dominant isoform. *Journal of Proteome Research* **14**, 18801887 (2015). URL http://dx.doi.org/10.1021/pr501286b.

[20] Abascal, F. *et al.* Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLOS Computational Biology* **11**, e1004325 (2015). URL http://dx.doi.org/10.1371/journal.pcbi.1004325.

[21] Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development* **27**, 23802396 (2013). URL http://dx.doi.org/10.1101/gad.229328.113.

[22] Gupta, I. *et al.* Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Molecular Systems Biology* **10**, 719719 (2014). URL http://dx.doi.org/10.1002/msb.135068.

[23] Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5** (2016). URL http://dx.doi.org/10.7554/eLife.10921.

[24] Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoformspecific translational regulation via alternative transcription start sites in mammals. *Molecular Systems Biology* **12**, 875 (2016). URL http://dx.doi.org/10.15252/msb.20166941.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[25] Lonsdale, J. *et al.* The genotype-tissue expression (GTEx) project. *Nat Genet* **45**, 580585 (2013). URL http://dx.doi.org/10.1038/ng.2653.

[26] Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Research* **44**, D710D716 (2015). URL http://dx.doi.org/10.1093/nar/gkv1157.

[27] Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 1521 (2012). URL http://dx.doi.org/10.1093/bioinformatics/bts635.

[28] Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Research* **22**, 20082017 (2012). URL http://dx.doi.org/10.1101/gr.133744.111.

[29] Reyes, A. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proceedings of the National Academy of Sciences* **110**, 1537715382 (2013). URL http://dx.doi.org/10.1073/pnas.1307202110.

[30] Mense, S. M. & Zhang, L. Heme: a versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to map kinases. *Cell Research* **16**, 681692 (2006). URL http://dx.doi.org/10.1038/sj.cr.7310086.

[31] Balwani, M. & Desnick, R. J. The porphyrias: advances in diagnosis and treatment. *Blood* **120**, 44964504 (2012). URL http://dx.doi.org/10.1182/blood-2012-05-423186.

[32] Roberts, A. G., Redding, S. J. & Llewellyn, D. H. An alternatively-spliced exon in the 5'-UTR of human ALAS1 mRNA inhibits translation and renders it resistant to haem-mediated decay. *FEBS Letters* **579**, 10611066 (2005). URL http://dx.doi.org/10.1016/j.febslet.2004.12.080.

[33] Katz, Y. *et al.* Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**, 24002402 (2015). URL http://dx.doi.org/10.1093/bioinformatics/btv034.

[34] Shabalina, S. A., Ogurtsov, A. Y., Spiridonov, N. A. & Koonin, E. V. Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Research* **42**, 71327144 (2014). URL http://dx.doi.org/10.1093/nar/gku342.

[35] Pal, S. *et al.* Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Research* **21**, 12601272 (2011). URL http://dx.doi.org/10.1101/gr.120535.111.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[36] Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660665 (2015). URL http://dx.doi.org/10.1126/science.aaa0355.

[37] Gonzlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* **14**, R70 (2013). URL http://dx.doi.org/10.1186/gb-2013-14-7-r70.

[38] Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111 (2016). URL http://dx.doi.org/10.1126/science.aah7111.

[39] Han, H. *et al.* Multilayered control of alternative splicing regulatory networks by transcription factors. *Molecular Cell* **65**, 539553.e7 (2017). URL http://dx.doi.org/10.1016/j.molcel.2017.01.011.

[40] Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology* **33**, 736742 (2015). URL http://dx.doi.org/10.1038/nbt.3242.

[41] Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127131 (2013). URL http://dx.doi.org/10.1038/nature12121.

[42] Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *BioRχiv* (2016). URL http://dx.doi.org/10.1101/068809.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Online Methods

Data processing and sample selection

We downloaded and decrypted the *GTEx* samples using the *Short Read Archive Toolkit* software. We used genomic and annotation files of the human reference genome version *GRCh38* as provided by the release number 68 of *ENSEMBL*[26]. To avoid any mapping bias, we standardized the read length of all samples. Since most samples consisted of sequenced fragments of 76 bp, we trimmed the sequenced fragments to 76 bp for samples with longer read lengths and excluded the samples with read lengths below 76 bp. Next, we mapped the resulting sequencing fragments to the human reference genome using *STAR v2.4.2a*[27]. We provided the aligner with annotated exon-exon junctions and followed the recommended "2-pass alignment" pipeline to optimize mapping accuracy. We excluded the samples with less than 1,000,000 sequenced fragments as well as those samples where less that 60% of the sequenced fragments could be assigned to a unique position in the reference genome. Since the *GTEx* data did not contain the samples for all tissues of each individual, we defined three large subsets of samples that would enable us to analyze each subset as a fully-crossed design (containing all tissue-individual combinations) while at the same time keeping as many different individuals and tissues as possible. A description of these subsets, which comprised a total of 798 samples, can be found in the main text.

Based on the transcript isoform annotations, we defined reduced models with non-overlapping exonic regions using the *HTSeq*[43] python scripts from the *DEXSeq* package. Importantly, reduced gene models enabled us to unambiguously assign sequenced fragments to exonic regions. For each of the 798 samples, we tabulated the sequenced fragments to each exonic region. Only reads mapping uniquely to the reference genome were considered for further analysis.

Relative exon usage coefficients

We model the counts using generalized models of the Negative Binomial family for each subset of the *GTEx* data[28,29]. We denote k_{ij1} as the number of sequenced fragments mapping to exonic region i in sample j. When estimating *REUCs*, k_{ij0} denotes the sum of sequenced fragments mapping to exonic regions of the same gene as exonic region i but excluding exonic region i (Figure 1B). k_{ij0} and k_{ij1} are realizations of a random variable K_{ijl} that is assumed to follow a Negative Binomial distribution,

$$K_{ijl} \sim \mathrm{NB}(\text{mean} = s_j \mu_{ijl}; \text{dispersion} = \alpha_{il}), \tag{1}$$

where s_j is a scaling factor that accounts for between-sample differences in sequencing depth and α_{il} is the dispersion parameter that describes the spread of the count data distribution. s_{ij} is estimated using the *DESeq* method[44] and α_{il} is estimated as in the *DEXSeq* method[28]. The mean μ_{ijl} is predicted by the model

$$\log \mu_{ijl} = \beta^S_{ij} + l\beta^E_i + l x^{\text{sex}}_j \beta^{\text{sex}}_i + l\beta^{\text{REUC}}_{i,u(j),t(j)}, \tag{2}$$

where $l = 1$ when referring to the exonic region i and $l = 0$ when referring to the counts from the rest of the exons of the same gene. The coefficients of the model are explained as follows:

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- β_j^S estimates overall gene expression effects on sample j.
- Since β_i^E is only included when $l = 1$, it estimates the mean across samples of the logarithmic ratio between the counts from exon i with respect to the counts of the rest of the exons of the same gene (i.e. K_{ij1}/K_{ij0}). Therefore, this coefficient is a measure of the average exon usage across all samples.
- β_i^{sex} captures sex-dependent differences in exon usage. In the GLM model matrix, x_{ij} takes the value of $-1/2$ if sample j is from a male individual and $1/2$ if sample j is from a female individual. Thus, this coefficient estimates the logarithmic fold change of the usage of exonic region i for each sex with respect to the average exon usage.
- The *Relative Exon Usage Coefficient* (*REUC*), $\beta^{\text{REUC}}_{i,u_{(j)},t(j)}$, is the interaction coefficient between individual $u_{(j)}$ and tissue $t_{(j)}$ from which sample j was taken. For exonic region i, the coefficient $\beta^{\text{REUC}}_{i,u_{(j)},t_{(j)}}$ thus estimates the logarithmic fold change in exon usage for each individual-tissue combination with respect to the average exon usage.

The *REUCs* are subjected to a Bayesian shrinkage procedure in order to reduce the mean-variance dependencies[29, 45].

Relative spliced-in coefficients

To estimate *Relative Spliced-In Coefficients* (*RSICs*) we use Equation 1 and Equation 2 to model a modified read counting scheme (Figure 1C). k_{ij1} remains the same as for the *REUCs* fit but k_{ij0} (i.e. $l = 0$) now denotes the number of sequenced fragments supporting the splicing out from transcripts of exonic region i (Figure 1C). For exonic region i, the coefficient β_i^E from Equation 2 now measures the mean across samples of the logarithmic ratio between the number of reads supporting the splice in of exonic region i and the number of reads supporting the splice out of exonic region i (i.e. the average spliced-in (*SI*) coefficient). The coefficient β_i^{sex} for exonic region i now measures the change of *SI* between each sex with respect to the average *SI*. The *RSIC* for exon i, $\beta^{\text{RSIC}}_{i,u_{(j)},t(j)}$, measures the logarithmic fold change in the exon's *SI* for each tissue-individual combination with respect to the average *SI*. As for the *REUCs*, the *RSICs* are also subjected to the Bayesian shrinkage procedure to eliminate the mean-variance trend[29].

Changes in exon usage driven by alternative splicing are reflected in both *REUCs* and *RSICs*. Changes in exon usage due to alternative initiation or termination sites of transcription, which do not result in exon-exon junction reads, are only reflected by *RSICs*.

Estimation of tissue-dependance score

For each exonic region on each subset of the data, we estimated a score based on the *REUCs* to measure to what extent the usage of each exonic region was tissue-dependent. First, the *REUCs* of a given exonic region i were expressed as the number of standard deviations away from the median of the exon's *REUCs*,

$$Z_{iut} = \frac{\beta^{\text{REUC}}_{iut} - \underset{u,t}{\text{Median}}(\beta^{\text{REUC}}_{iut})}{\underset{u,t}{\sigma}(\beta^{\text{REUC}}_{iut})}. \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Then, the tissue-dependence score for exonic region i was defined by

$$T_i = \max_t \left\{ \left| \frac{1}{m} \sum_{u=1}^{m} Z_{iut} \right| \right\}, \tag{4}$$

with m being the number of individuals on the data subset.

Analysis of variance of *REUCs* and *RSICs*

For each exonic region on each subset of the data, we fitted an analysis of variance model,

$$\beta_{iut}^{\text{REUC}} = \beta_i^0 + \beta_{iu}^{\text{Individual}} + \beta_{it}^{\text{Tissue}} + \epsilon_{iut}, \tag{5}$$

using ordinary least squares regression to minimize the residual sum of squares (RSS),

$$\text{RSS}_i = \sum_{u,t} \epsilon_{iut}^2 = \sum_{u,t} (\beta_{iut}^{\text{REUC}} - \hat{\beta}_{iut}^{\text{REUC}})^2, \tag{6}$$

where $\hat{\beta}_{iut}^{REUC}$ are the *REUC* values predicted by the model. In order to estimate the coefficient of partial determination (R^2) for the tissue predictor (i.e. the proportion of total variance that can be attributed to variance across tissues), we fitted a reduced model lacking the $\beta_{it}^{\text{Tissue}}$ term,

$$\beta_{iut}^{\text{REUC}} = \beta_i^0 + \beta_{iu}^{\text{Individual}} + \epsilon_{iut}. \tag{7}$$

The R^2 for a given exon i is then calculated by,

$$R_i^2 = 1 - \frac{\text{RSS}_i(\text{full})}{\text{RSS}_i(\text{reduced})}, \tag{8}$$

where $\text{RSS}_i(\text{full})$ is the RSS from the full model (i.e. Equation 5) and $\text{RSS}_i(\text{reduced})$ is the RSS from the reduced model (i.e. Equation 7). The same procedure was followed to estimate R^2 on the *RSICs* but using $\beta_{iut}^{\text{RSIC}}$ as the response variable in Equation 5 and in Equation 7.

Genomic analysis

To test for over-representation of features among the genes with TDU, we used the *R CRAN* package *MatchIt*[46] to generate background sets of genes with the same distribution of expression strength and number of exonic regions as the genes with TDU. Genes were classified according to *ENSEMBL* annotations and we used a χ^2-test for differences between genes with TDU and the background set of genes. Gene biotypes were retrieved from *ENSEMBL* using the *Bioconductor*[47] package *biomaRt*[48]. For enrichment of features among exons with tissue-dependent usage (*TDU*), we also used *MatchIt* to generate background sets of exons with the same distribution of expression strength and exon widths. We tested for differences between exons with *TDU* and the background set of exons using a χ^2-test.

Operations on genomic ranges were done using the *Bioconductor* package *GenomicRanges*[49]. Data visualizations and graphics were generated using the *Bioconductor* packages *ggplot2*[50] and *Gviz*[51].

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Reproducibility

The *Bioconductor* package *HumanTissuesDEU* contains the *R* data objects and code needed to reproduce the analysis and figures presented in this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/127894; this version posted April 17, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

[43] Anders, S., Pyl, P. T. & Huber, W. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166169 (2014). URL http://dx.doi.org/10.1093/bioinformatics/btu638.

[44] Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11**, R106 (2010). URL http://dx.doi.org/10.1186/gb-2010-11-10-r106.

[45] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15** (2014). URL http://dx.doi.org/10.1186/s13059-014-0550-8.

[46] Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software* **42**, 1–28 (2011). URL http://www.jstatsoft.org/v42/i08/.

[47] Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**, 115121 (2015). URL http://dx.doi.org/10.1038/nmeth.3252.

[48] Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols* **4**, 11841191 (2009). URL http://dx.doi.org/10.1038/nprot.2009.97.

[49] Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Computational Biology* **9**, e1003118 (2013). URL http://dx.doi.org/10.1371/journal.pcbi.1003118.

[50] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2009). URL http://ggplot2.org.

[51] Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and Bioconductor. *Statistical Genomics* 335351 (2016). URL http://dx.doi.org/10.1007/978-1-4939-3578-9_16.