Supporting Information

Drift and conservation of differential exon usage across tissues in primate species

Alejandro Reyes, Simon Anders, Robert J. Weatheritt, Toby J. Gibson, Lars M. Steinmetz, and Wolfgang Huber

Supplementary Methods

Raw data

Brawand and coworkers deposited the RNA-Seq reads of their study in NCBI's Gene Expression Omnibus (GEO) database under accession number GSE30352. We used the data from brain, cerebellum, kidney, heart and liver samples from primates. An overview of the samples used and their statistics is provided in Table S1.

RNA-seq data processing

Since some samples were sequenced in single-end and others in paired-end mode, and since different read lengths were used, we standardised the data from all samples to avoid potential biases that could be introduced by these differences, e. g. via alignment specificity and sensitivity. We discarded all second mates in paired-end runs and trimmed all remaining reads to 76 nucleotides, resulting in the same single read length throughout. We mapped the trimmed read sequences to their respective reference genomes from Ensembl [Supp1] release 64 using GSNAP 2012-11-01 [Supp2], allowing for novel junctions and uniquely mapped reads (see Table S1 for details).

Definition of one-to-one orthology between exons

To generate a table of one-to-one orthologous exons, we retrieved from Ensembl 64 the genomic sequence of all human genes with one-to-one orthologues in all the primate species in our study. For these genes, we retrieved all exons from all annotated transcripts [Supp3]. We aligned each of these individual human exons against the genomes of the other five primate species using GMAP 2012-11-01 [Supp4]. We also aligned them against the human reference genome to discard exons present in multiple copies. We allowed only unique matches with at least 90% sequence identity over the whole length of the exon, allowing at most 5% length difference due to insertions and deletions (indels). The exons that contained indels in protein coding exons modifying the reading frame were discarded. We took the intersection of exons that passed the above criteria in all the six species analysed. In addition, we discarded genes that contained only a single sequence-conserved exon. If an exon was annotated to have alternative start and/or end sites, we split the exon into multiple counting bins, which for the purpose of the differential expression analysis were treated as if they were separate exons, in the same manner as in DEXSeq (see Figure 1 in [Supp3]).

Parameter sensitivity analysis. SI Figure S5 and SI Table S2 present the results of a sensitivity analysis, in which we successively increased the stringency of the sequence similarity threshold of the considered exons. This analysis confirmed that our observations on inter-species variability of exon usage are not driven by technical biases caused by sequence differences between the genomes of the different species.

Inference

For each exon in each species-tissue combination, we calculated a relative exon usage coefficient (REUC), which we defined as the logarithm (base 2) of the ratio of the exon's usage in the species-tissue combination relative to the average over all species-tissue combinations. This approach is based on our *DEXSeq* method, described in [Supp3], and is described in detail below. For readers familiar with *DEXSeq*, we note that the approach deviated in three main aspects from [Supp3]:

1. The generalized linear model (GLM) that we fit for each exon incorporates as response two count values per sample, namely the number of reads mapped to the exon under consideration and the *sum* from all the other exons of the same gene. In [Supp3], each of the other exons was counted separately; the modification used here speeds up computation and simplifies interpretation. In newer versions of the *DEXSeq* package, this functionality is available via the *TRT* functions.
2. We employ an *empirical Bayes* shrinkage method to the fitted coefficients to suppress the high variance that fitted coefficients would otherwise have. A general treatment of this new feature will be subject of a forthcoming publication.
3. As discussed below, the likelihood ratio test usually employed for inference with GLMs was unsuitable for the specific application of inferring conservation of exon usage. Instead, we devised a custom test based on a covariance statistic.

Modelling with generalized linear models

We denote by k_{ij1} the number of uniquely mapping reads found to overlap with exon counting bin i in sample j and by k_{ij0} the sum of the read counts for all other exon counting bins in the same gene. We consider these quantities as realization of a random variable, K_{ijl}, which is determined by a generalized linear model (GLM) of the

negative binomial (NB) family with logarithmic link, i. e.

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

$$\log \mu_{ijl} = \sum_r x_{2j+l,\,r}\, \beta_{ir}. \quad (2)$$

Here, s_j is a size factor, accounting for the library sequencing depth for sample j ($j = 1, \ldots, m$) and estimated as described in [Supp3]. The GLM's model matrix x has $2m$ rows, as each of the m samples appears twice, once for the exon under consideration ($l = 0$) and once for the sum of all other exons ($l = 1$). Its columns are indexed by r and correspond to the regression coefficients, of which we have four different types, namely

- for each sample j, a sample coefficient β^{S}_{ij} that captures the overall expression of the gene in the given sample,
- an exon coefficient β^{E}_i capturing the average usage of the exon under consideration, i.e., the average contribution of the exon to the gene's read counts; this coefficient takes into account exon length and other, sequence-dependent exon properties,
- a sex coefficient β^{sex}_i to account for sex-dependence of the former, and finally
- for each species-tissue combination, the REUC β^{I}_{iut}, which estimates the ratio of the exon's usage in the combination of species u and tissue t relative to the average.

Hence, we can write out, equivalently but more explicitly, the product of the model matrix x and the model coefficients β in Equation (2) as follows.

$$\log \mu_{ijl} = \beta^{\mathrm{S}}_{ij} + (1-l)\beta^{\mathrm{E}}_i + (1-l)x^{\mathrm{sex}}_j \beta^{\mathrm{sex}}_i + (1-l)\beta^{\mathrm{I}}_{i,u(j),t(j)}, \quad (3)$$

where $x^{\mathrm{sex}}_j = -1/2$ if sample j is derived from a female individual and $x^{\mathrm{sex}}_j = +1/2$ in case of a male. $u(j)$ is the species and $t(j)$ the tissue from which sample j is taken, and l again indicates whether the count is modelled for exon i ($l = 0$) or for the sum of the other exons in the gene ($l = 1$). The design matrix x in Equation (2) contains zeroes and ones (expect for the column for the sex coefficient which contains $\pm 1/2$), and the structure of x can be read off from comparing Equation (3) with Equation (2).

Note that, while we performed the GLM fit using the natural logarithm as link function, all values for coeffcients given in this paper or shown in figures are on a $\log_2$ scale, to facilitate interpretation.

Fitting of GLM coefficients with shrinkage

When considered as estimators, GLM coefficients have a large sampling variance for exons with low read counts, and hence, the REUCs will tend to have larger absolute values for weakly used exons than for strongly expressed ones. This is disadvantageous as it hinders visualization and comparison in downstream analysis. Therefore, we devised a shrinkage approach that renders the REUCs approximately homoscedastic. Specifically, we imposed a common normal prior with mean zero and standard deviation σ_{P} on the coefficients β^{ES}_i and β^{I}_{ius}. The Bayesian posterior probablity can hence be written

$$\log p_i = \sum_{j,l} \ell\left(s_i \exp \sum_r x_{2j+l,\,r}\beta_{ir},\ \alpha_{il};\ k_{ijl}\right) - \frac{1}{\sigma^2_{\mathrm{P}}} \sum_{r \in R} \beta^2_{ir}, \quad (4)$$

where $\ell(\mu, \alpha; k)$ is the log-likelihood of the NB distribution with mean μ (and dispersion α, as discussed below), given the observed count k, and the sum over r in the ridge penality term runs over the set R of all those columns of the design matrix that correspond to a coefficient that is supposed to undergo shrinkage, i. e., all REUCs β^{I}_{iut} and the sex coefficient β^{sex}_i.

Instead of using the iterated reweighted least square (IRLS) method, we used the L-BFGS-B optimization algorithm to find maximum a posteriori estimates for the coefficients. For $\sigma_{\mathrm{P}} \to \infty$, this approach and the ordinary IRLS procedure are expected to give the same result. We determined a suitable value for the prior width σ_{P} in an empirical Bayes fashion, as described in the following. We first ran the fit with a large value, $\sigma_{\mathrm{P}} = 1000$, then calculated for each exon the average normalized counts and the sample standard deviation of the REUCs, $\mathrm{stdev}_{u,t}\left(\beta^I_{iut}\right)$, as a measure for the REUC's typical spread of values. See the left panel of Figure S1 for a scatter plot of the result: the spread of the coefficients was stronger for low-count exons, while above a certain threshold, roughly at 2^8, this dependence levelled out. Our aim was to shrink the coefficients of the low-count exons such that they did not scatter more than the high-count exons. Therefore, we computed the mean of the standard deviations of the exons with an average normalized count above 2^8 as a value for the prior σ_{P} and reran the fit to get final coefficient estimates. As the right panel of Figure S1 shows, this procedure succeeded in establishing approximate homoscedasticity.

The shrinkage approach has another advantage. Note that the design matrix of our GLM does not have full rank, because we fit coefficients for all species-tissue combinations. The conventional approach of leaving out one level from the tissue-species factor and let it be absorbed by the intercept, to give full rank to the design matrix, would have complicated the down-stream analysis, as the absorbed tissue-species combination would have become special. Hence, we took advantage of the fact that the ridge penality term causes the penalized likelihood to have a unique maximum, and so makes the model fully identifiable, despite the rank deficiency of the design matrix.

Estimation of dispersions

After the inital fit with the weak prior, we estimated dispersions. It turned out to be helpful to assume two different dispersion values α_{il} for each exon i, namely one for the response variables concerning the exon's counts ($l = 0$) and one for those with the sums of the other exons ($l = 1$). We fitted these dispersions by maximizing the Cox-Reid adjusted conditional likelihood found by keeping the coefficients β_{ir} from the initial fit fixed and maximizing with respect to $(\alpha_{i0}, \alpha_{i1})$. For the Cox-Reid adjustment [Supp5], we used the approach of [Supp6], as

in *DEXSeq*. We then reran the fit using the prior width σ_{P} (see above) and the dispersions α_{il} obtained from the first-pass fit to obtain final coefficient estimates β_{ir}.

TDU strength

The quantity termed *TDU strength* was calculated as

$$T_{iu} = \max_t \left(\beta^{\mathrm{I}}_{iut} - \overline{\beta}^{\mathrm{I}}{}_{iu\cdot}\right) \tag{5}$$

with $\overline{\beta}^{\mathrm{I}}_{iu\cdot} = \frac{1}{5}\sum_{t=1}^{5}\beta^{\mathrm{I}}_{iut}$.

ANOVA analysis of the REUCs

For the ANOVA analysis shown in Figure 2B, we fitted, for each exon, a linear model to its REUCs

$$\beta^{\mathrm{I}}_{iut} = \beta^{\mathrm{I0}}_{i} + \beta^{\mathrm{sp}}_{iu} + \beta^{\mathrm{ti}}_{it} + \epsilon_{iut}, \tag{6}$$

(with $\beta^{\mathrm{sp}}_{i,1} = \beta^{\mathrm{ti}}_{i,1} = 0$) using ordinary least square regression to mimimize $\sum_{ut}\epsilon^2_{iut}$.

The axes in Figure 2B refer to the variance explained by species and by tissue in the sense of a usual ANOVA table, i.e., the quantities $\sum_s \left(\beta^{\mathrm{sp}}_{is}\right)^2$ and $\sum_t \left(\beta^{\mathrm{ti}}_{it}\right)^2$. The axes tick marks indicate the variance values on the $\log_2$ scale, i.e., after division by $(\ln 2)^2$. Note that the axes in the figure have been "warped" using an asinh transformation, which provides a good compromise between a linear and a logarithmic axis scale, as both of these seemed unsuitable.

Testing for conservation

To assess conservation of the pattern of tissue-dependent usage of exon i between a pair of species u and u', we considered the sample covariance of the respective REUCs,

$$C_{iuu'} = \mathrm{cov}_t\left(\beta^{\mathrm{I}}_{iut},\ \beta^{\mathrm{I}}_{iu't}\right). \tag{7}$$

To construct an empirical null distribution for this quantity, we used the mirror method: we assumed that the null distribution is symmetric around zero, and that the mass of the alternative distribution at negative values was negligible (because an evolutionary process resulting in anti-correlation seems implausible), so that the empirical distribution of negative values could be used to estimate the empirical null by mirroring around zero [Supp7]. For the threshold $C_{\mathrm{th}} = 0.048$ (on the natural log scale; corresponding to $0.048/(\ln 2)^2 = 0.1$ on the $\log_2$ scale), we found that

$$\frac{|\{i : C_{iuu'} < -C_{\mathrm{th}}\}|}{|\{i : C_{iuu'} > C_{\mathrm{th}}\}|} < 0.1 \tag{8}$$

for all species pairs u, u'. Hence, for each species pair u, u', the exons with $C_{iuu'} > C_{\mathrm{th}}$ were called conserved between u and u' with a false discovery rate (FDR) of $< 10\%$.

Background sets

To avoid biases due to differences in inferential power, we compared in all enrichment analyses the set of strictly CTDU exons not to the set of all 1:1 othologous exons but instead to a set of "background" exons, taken from the whole set but chosen to match the strictly CTDU exons' empirical distribution of expression strength, exon length and variance across replicates. We used the R package *MatchIt* [Supp8] to construct these background sets.

Features

We extracted the coding exons from the human annotation file from Ensembl release 64, translated them into protein sequences and mapped them to the UniProt canonical protein database. The protein disordered region predictions were done using IUPRED [Supp9]. The UTR region coordinates were downloaded from Ensembl using the Ensembl API. The classification of exons into first, middle and last exons was also done using the Ensembl annotation. For the SI Tables S3 and S4 and SI Figures S7, S8, S9 and S10 only exons belonging to single categories were used. For example, we discarded exons that were first exons of a transcript but the middle exon of another transcripts.

Cis-regulatory region characterisation

We used SFmap [Supp10] to characterise the cis-regulatory regions of our set of conserved exons. In order to increase the confidence of their calls, SFmap takes into account (1) the propensity of splicing factor binding motifs of being clustered in the genome and (2) DNA sequence conservation. Hypothesis testing was done using the Wilcoxon test, and the 95% confidence intervals of Figure 3D were calculated by bootstrapping.

Stratification by exon function and position

To assess potential biases, or to detect additional trends, we repeated the analyses under both of the following two stratification schemes:

1. *by translation*: 5' untranslated, 3' untranslated, translated;
2. *by position*: first, middle, last.

Both of these stratifications were made based on Ensembl release 64 transcript models for human. Within each stratification scheme, exons which would be assigned into more than one category based on evidence from different transcripts were set aside (e. g. for the stratification by translation, this was the case for exons that are in an UTR in one transcript and in a coding region in another; for the stratification by position, exons were set aside that are the first in one transcript but a middle one in another, i. e. involve alternative transcription start sites). SI Figures S7–S10 and SI Tables S3 and S4 demonstrate that the conclusions drawn in the main text are not driven by a single category of exons, and that the effects that we describe are evident in all of the above exon categories.

Reprodicibility

We added a document with the reproducible code that was used to generate and analyse the REUCs as well as the the code used to generate the plots from the main text. Please see *SI Appendix, Dataset 1*. In this document, we also added plots for each gene that contained exons with strictly CTDU.

References

[Supp1] Flicek P, *et al.* (2011). Ensembl 2011. *Nucleic Acids Res*, 39:D800–D806. [link].

[Supp2] Wu TD, Nacu S (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26 (7):873–881. [link].

[Supp3] Anders S, Reyes A, Huber W (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22 (10):2008–2017. [link].

[Supp4] Wu TD, Watanabe CK (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21 (9):1859–1875. [link].

[Supp5] Cox DR, Reid N (1987). Parameter orthogonality and approximate conditional inference. *J Roy Statist Soc B*, 49 (1):1–39.

[Supp6] McCarthy DJ, Chen Y, Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res*, 40 (10):4288–4297. [link].

[Supp7] Muralidharan O. False discovery rate and empirical null methods. Master's thesis, Stanford University, Department of Mathematics.

[Supp8] Ho D, Imai K, King G, Stuart E (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.

[Supp9] Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21 (16):3433–3434. [link].

[Supp10] Paz I, Akerman M, Dror I, Kosti I, Mandel-Gutfreund Y (2010). SFmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res*, 38 (suppl 2):W281–W285. [link].

Supplementary Figures

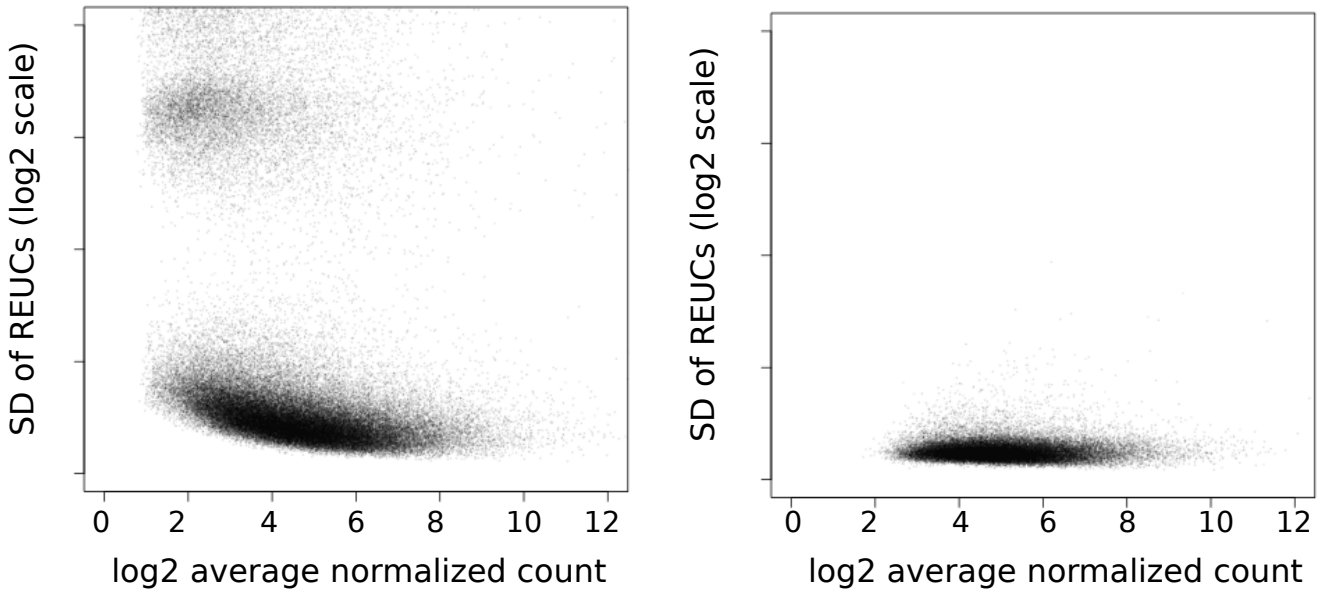

Figure S1: Standard deviation of REUCs versus average normalised count. The left panel shows the heteroskedasticity of the unshrunken coefficients, the right panel the approximate homoskedasticity of the shrunken coefficients.

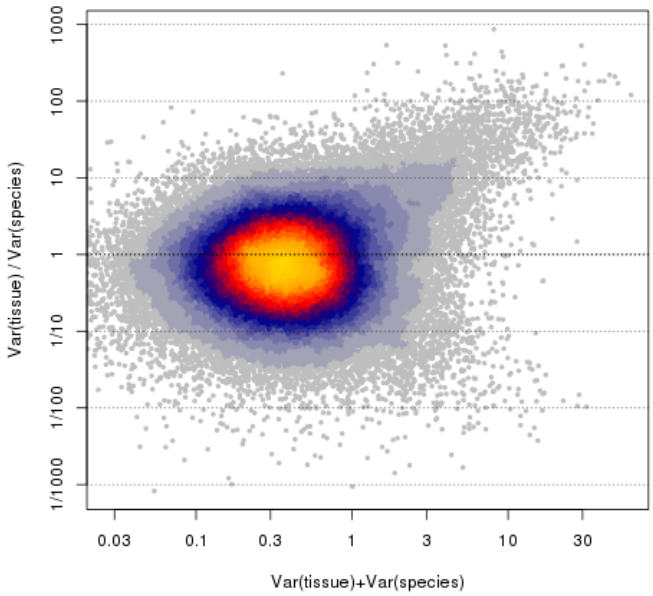

Figure S2: Comparison of REUC variance explained by tissue and by species, respectively. This plot provides an alternative visualization of the same data as in Fig. 2B.

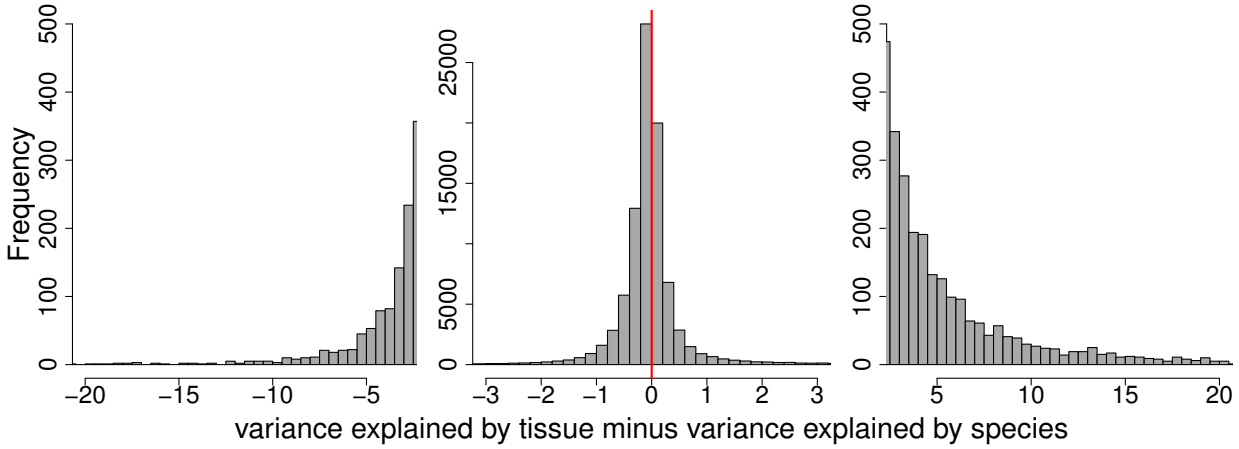

Figure S3: The central panel shows the histogram of the variance explained by tissue minus the variance explained by species. The panels at the left and right provide *zoom-in* views of the lower and upper distribution tails. The histogram shows that the centre of the distribution is located below 0, but that its upper tail is heavier than the lower one.

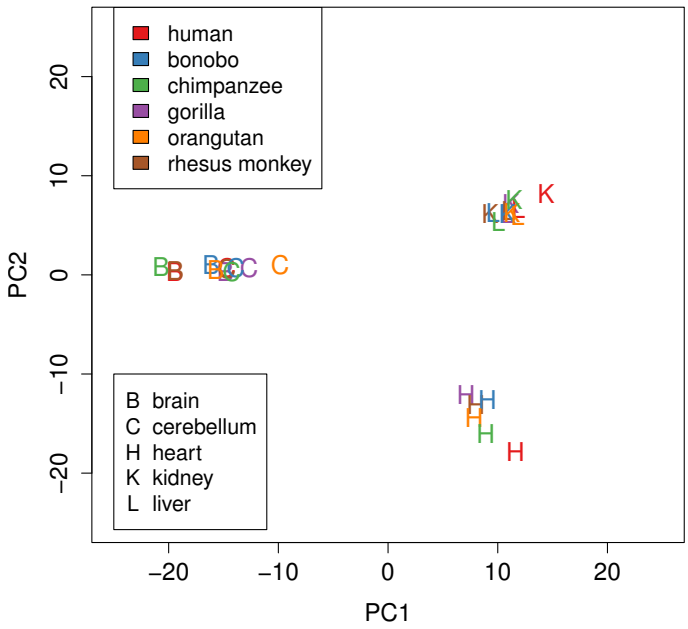

Figure S4: Principal component analysis of REUCs of the 1,292 strictly CTDU exons (i.e., with CTDU between all the species pairs).

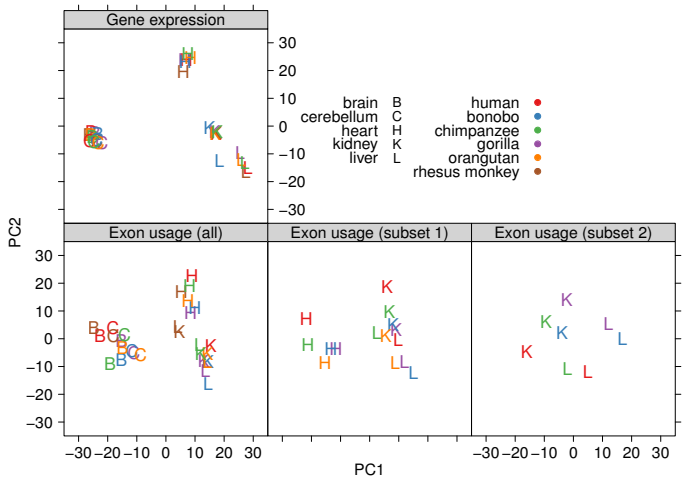

Figure S5: Principal component analysis, as in Figure 2A, but with more stringent requirements on sequence conservation: only exons with more than 95% sequence identity and without any indels across all six primate species were used. The PCA demonstrates that the inter-species variability that we observed is not mainly driven by possible technical biases caused by sequence differences between the species.

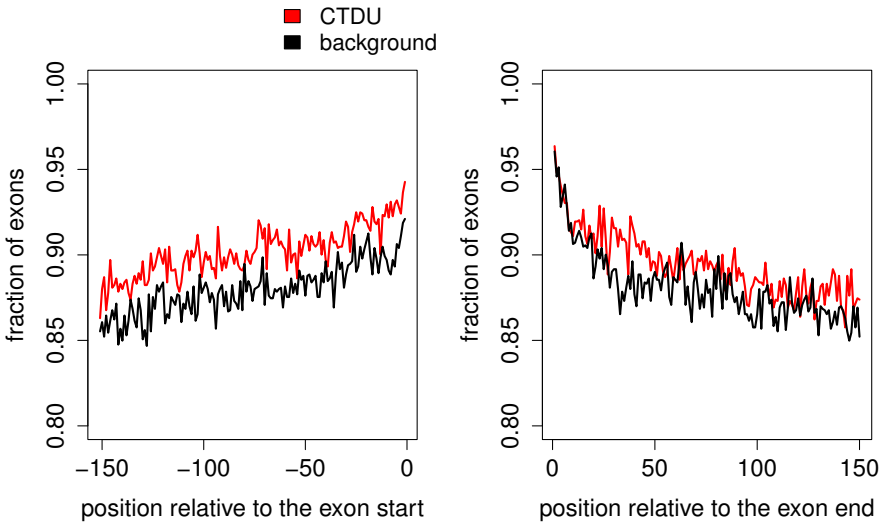

Figure S6: Sequence conservation in introns that flank strictly CTDU exons. For each position relative to the exon start or end, we calculated the fraction of strictly CDTU exons which show at this position in their flanking introns' sequences the same nucleotide in all species (red). The black line shows the same quantity for the set of background exons. Introns both in 5' and 3' of CTDU exons tend to be more conserved than the background (control) set of introns ($p = 1.9 \cdot 10^{-15}$ and $p = 5.7 \cdot 10^{-9}$, Wilcoxon rank sum test with continuity correction). This result is consistent with the notion that the need to maintain splicing-related cis-regulatory elements involved in tissue-dependent exon usage results in purifying selection of sequences within these introns.

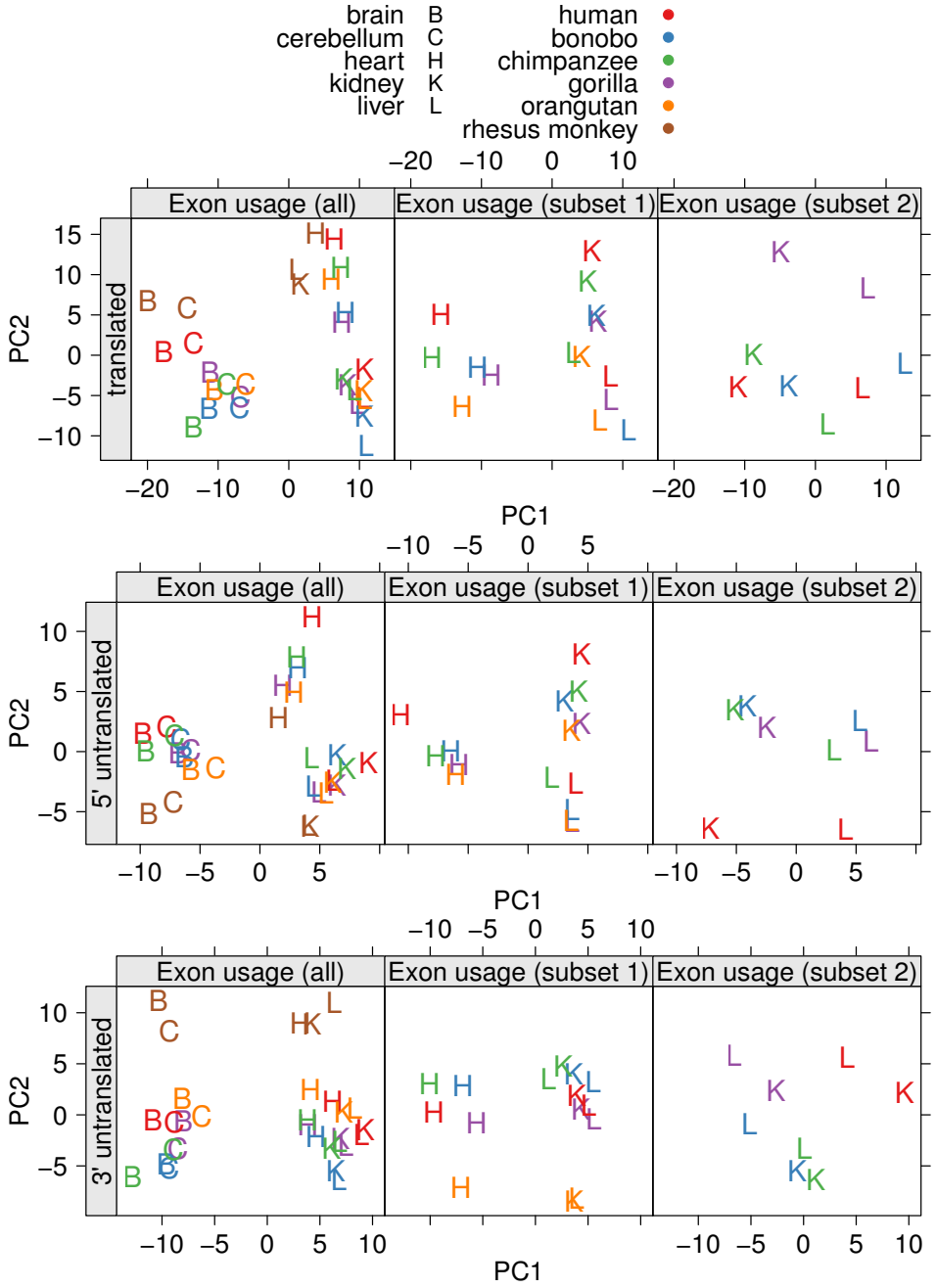

Figure S7: Principal component analyses, as in Figure 2A, but stratified by translational status (translated, 3'-untranslated, 5'-untranslated).

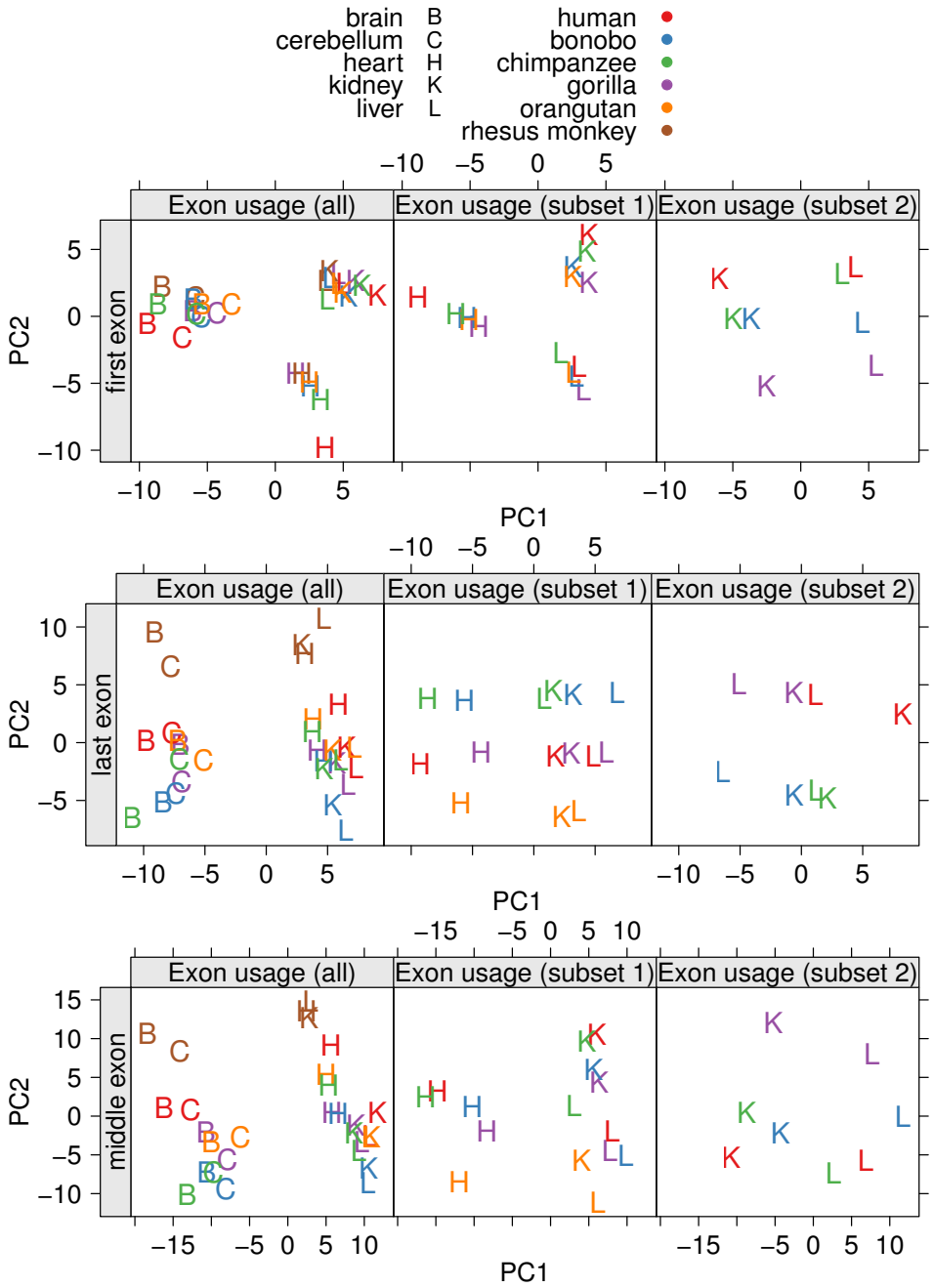

Figure S8: Principal component analyses, as in Figure 2A, but stratified by exon position in transcripts (first, middle, last).

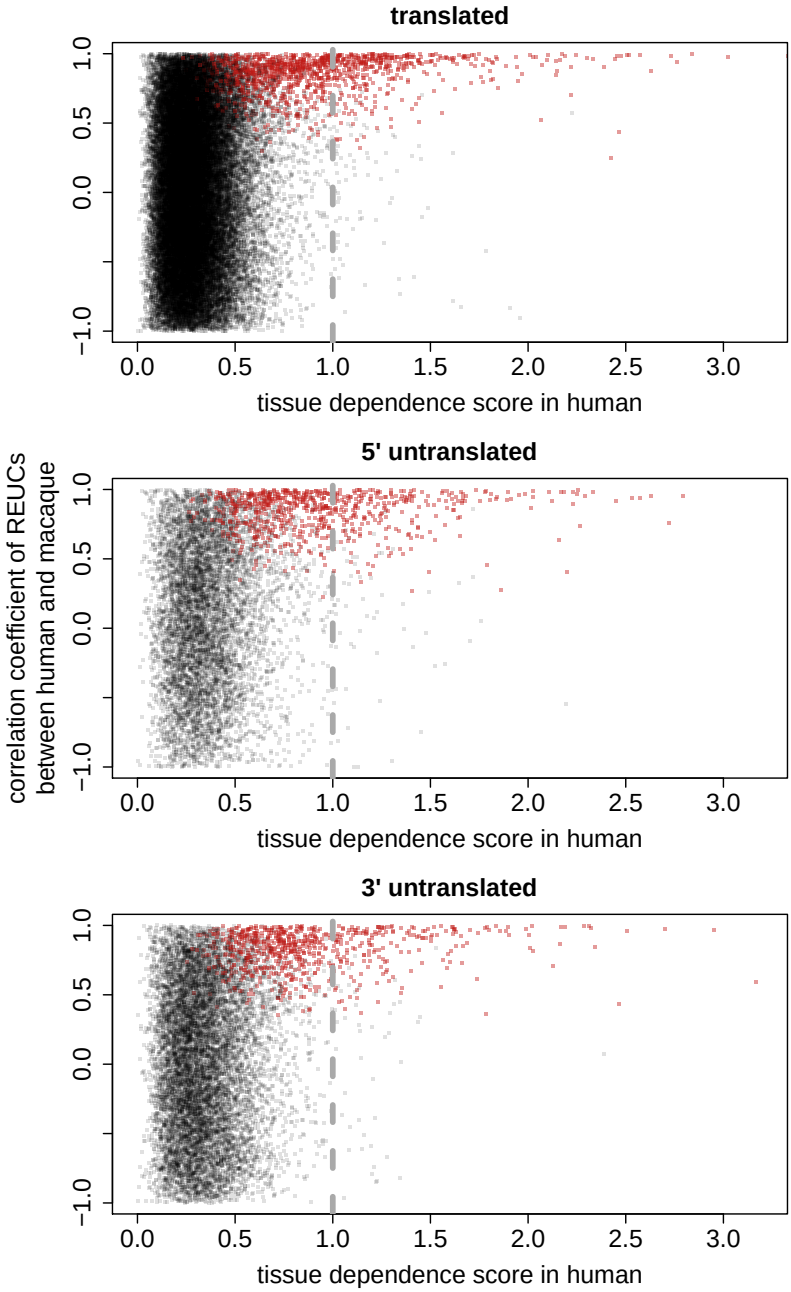

Figure S9: Tissue specificity and conservation of the regulation of differential usage of exons across tissues, as in Figure 2D, but stratified by translational status (translated, 3'-untranslated, 5'-untranslated).

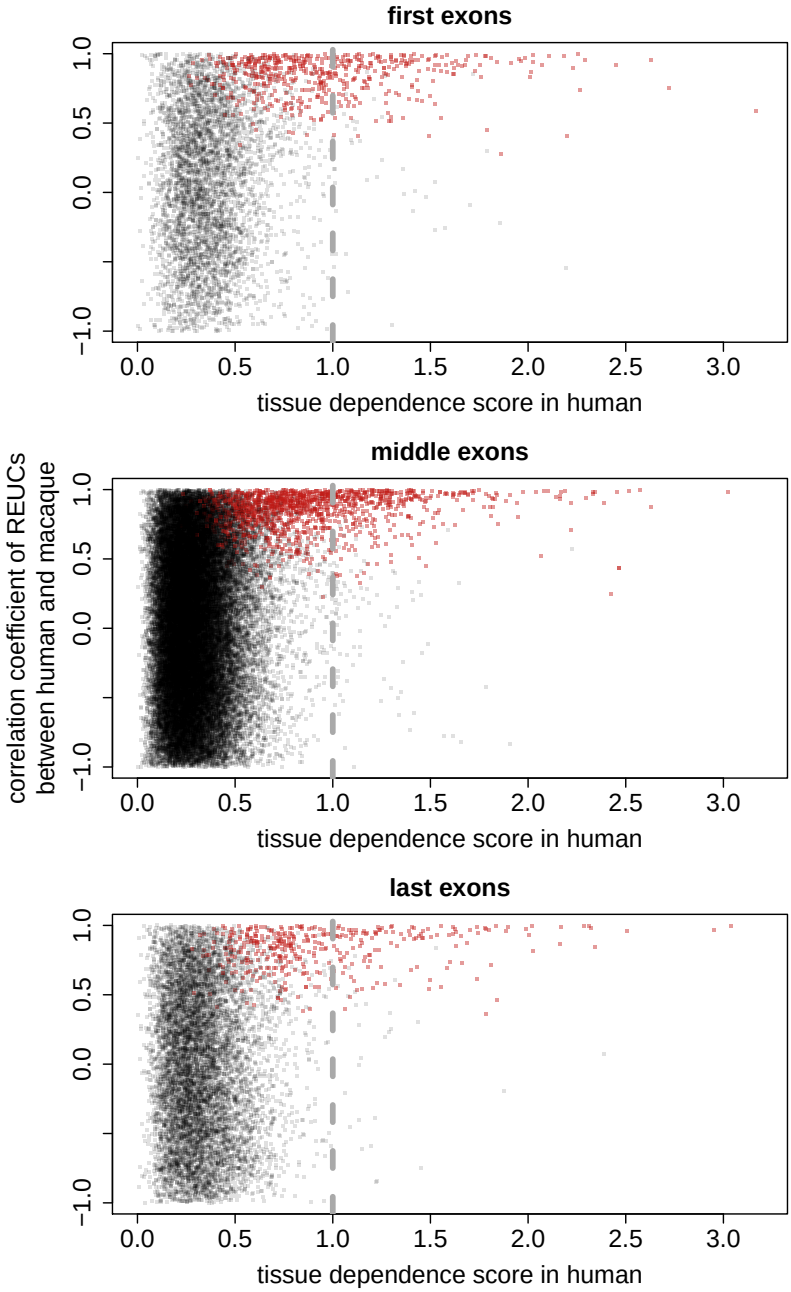

Figure S10: Tissue specificity and conservation of the regulation of differential usage of exons across tissues, as in Figure 2D, but stratified by exon position in transcripts (first, middle, last).

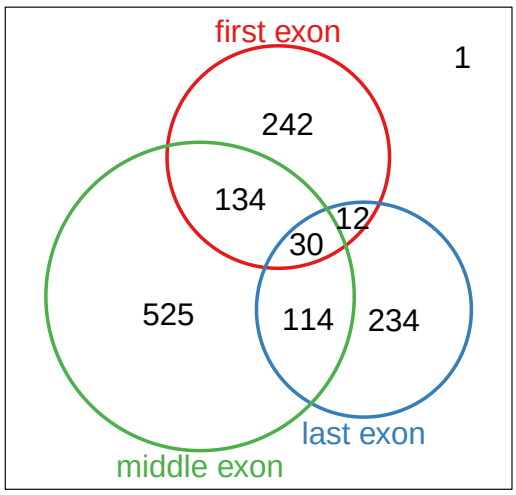

Figure S11: Venn diagram, showing the strictly CTDU exons's position in transcripts (first, middle, last).

Supplementary Tables

Table S1: Summary of samples used, number of fragments and fraction of uniquely aligned fragments.

	run_accession	species	tissue	sex	# of reads	fraction of unique alignments
1	SRR306838	human	brain	female	24,513,415	0.8
2	SRR306839	human	brain	male	18,850,030	0.65
3	SRR306840	human	brain	male	22,576,705	0.69
4	SRR306841	human	brain	male	24,325,223	0.66
5	SRR306842	human	brain	male	17,422,994	0.76
6	SRR306843	human	brain	male	7,913,181	0.52
7	SRR306844	human	cerebellum	female	32,698,558	0.75
8	SRR306845	human	cerebellum	male	46,755,221	0.66
9	SRR306845	human	cerebellum	male	46,755,221	0.66
10	SRR306847	human	heart	female	24,128,204	0.71
11	SRR306848	human	heart	male	30,896,351	0.66
12	SRR306848	human	heart	male	30,896,351	0.66
13	SRR306850	human	heart	male	25,197,713	0.69
14	SRR306851	human	kidney	female	22,493,518	0.77
15	SRR306852	human	kidney	male	20,684,752	0.73
16	SRR306853	human	kidney	male	31,386,619	0.66
17	SRR306854	human	liver	male	43,147,061	0.66
18	SRR306854	human	liver	male	43,147,061	0.66
19	SRR306856	human	liver	male	23,866,499	0.76
20	SRR306811	chimpanzee	brain	female	20,083,064	0.68
21	SRR306812	chimpanzee	brain	male	13,947,644	0.74
22	SRR306813	chimpanzee	brain	male	20,408,261	0.62
23	SRR306814	chimpanzee	brain	male	17,394,854	0.64
24	SRR306815	chimpanzee	brain	male	22,234,086	0.65
25	SRR306816	chimpanzee	brain	male	23,317,655	0.63
26	SRR306817	chimpanzee	cerebellum	female	32,043,112	0.72
27	SRR306818	chimpanzee	cerebellum	male	19,384,434	0.71
28	SRR306819	chimpanzee	heart	female	31,468,011	0.62
29	SRR306820	chimpanzee	heart	male	43,064,259	0.51
30	SRR306821	chimpanzee	kidney	female	25,454,775	0.71
31	SRR306822	chimpanzee	kidney	male	34,169,060	0.72
32	SRR306823	chimpanzee	liver	female	29,737,439	0.73
33	SRR306824	chimpanzee	liver	male	17,876,248	0.66
34	SRR306826	bonobo	brain	female	17,166,270	0.74
35	SRR306827	bonobo	brain	female	24,777,783	0.65
36	SRR306828	bonobo	brain	male	38,196,822	0.66
37	SRR306829	bonobo	cerebellum	female	30,345,120	0.69
38	SRR306830	bonobo	cerebellum	male	34,467,310	0.66
39	SRR306831	bonobo	heart	female	29,650,645	0.62
40	SRR306832	bonobo	heart	male	26,025,889	0.55
41	SRR306833	bonobo	kidney	female	30,139,364	0.65
42	SRR306834	bonobo	kidney	male	25,901,079	0.63
43	SRR306835	bonobo	liver	female	28,491,592	0.72
44	SRR306836	bonobo	liver	male	20,161,205	0.62
45	SRR306800	gorilla	brain	female	35,257,547	0.72
46	SRR306801	gorilla	brain	male	16,254,814	0.78
47	SRR306802	gorilla	cerebellum	female	28,305,051	0.74
48	SRR306803	gorilla	cerebellum	male	20,661,901	0.7
49	SRR306804	gorilla	heart	female	28,286,878	0.7
50	SRR306805	gorilla	heart	male	30,588,563	0.64
51	SRR306806	gorilla	kidney	female	19,804,877	0.73
52	SRR306807	gorilla	kidney	male	29,684,063	0.73
53	SRR306808	gorilla	liver	female	32,830,718	0.71
54	SRR306809	gorilla	liver	male	34,982,548	0.72
55	SRR306777	rhesus monkey	brain	female	19,068,947	0.68
56	SRR306778	rhesus monkey	brain	male	22,554,234	0.59
57	SRR306779	rhesus monkey	brain	male	21,461,283	0.76
58	SRR306780	rhesus monkey	cerebellum	female	25,528,147	0.6
59	SRR306781	rhesus monkey	cerebellum	male	21,141,815	0.67
60	SRR306782	rhesus monkey	heart	female	28,636,572	0.55
61	SRR306783	rhesus monkey	heart	male	20,815,484	0.58
62	SRR306784	rhesus monkey	kidney	female	17,581,272	0.59
63	SRR306785	rhesus monkey	kidney	male	24,115,366	0.44
64	SRR306786	rhesus monkey	liver	female	21,711,196	0.65
65	SRR306787	rhesus monkey	liver	male	32,224,651	0.69
66	SRR306787	rhesus monkey	liver	male	32,224,651	0.69
67	SRR306791	orangutan	brain	female	36,457,958	0.69
68	SRR306792	orangutan	brain	male	17,675,725	0.73
69	SRR306793	orangutan	cerebellum	female	20,807,820	0.71
70	SRR306794	orangutan	heart	female	36,798,263	0.63
71	SRR306795	orangutan	heart	male	31,482,282	0.64
72	SRR306796	orangutan	kidney	female	30,547,227	0.65
73	SRR306797	orangutan	kidney	male	30,043,284	0.68
74	SRR306798	orangutan	liver	female	21,355,541	0.74
75	SRR306799	orangutan	liver	male	35,683,453	0.71
76	total				2,018,132,789	0.67

Table S2: Decomposition of variance explained by species or tissues, using different choices of sequence similarity thresholds. This analysis demonstrates that our inter-species variability estimates are not mainly driven by sequence divergence effects. We considered exons with no insertions or deletions, and varied the threshold for sequency similarity from 90% (main text) in steps of 1% to 99% (this table). The first column lists the threshold value, the second column indicates the number of exons that passed the threshold, the third column indicates for what percentage of the exons the between-species variance was higher than the between-tissue variance. The fourth column indicates the number of exons that showed strong variance with respect to one or both of the contrasts (between-species, between-tissues). The fifth column shows the fraction of those exons in column 4 for which the between-species variance was larger than the between-tissue variance. These results indicate that irrespective of the similarity threshold, for the majority of exons, more of the variance is explained by species than by tissues, while for the exons with high variance, between-species differences are smaller than between-tissue effects.

% similarity threshold	# of exons with enough read counts for testing	% of exons where Var(species)>Var(tissues)	# of exons where Var(species)>0.75 or Var(tissues)>0.75	% of exons where Var(species)>0.75 or Var(tissues)>0.75 and Var(species)>Var(tissues)
91	91596	59.91	1442	19.00
92	88885	59.98	1396	19.20
93	84875	59.96	1330	19.10
94	79222	60.00	1258	19.08
95	71186	59.92	1137	18.56
96	59766	60.03	970	19.07
97	44706	60.02	718	18.38
98	25895	60.19	420	19.29
99	6260	60.40	105	15.24

Table S3: Columns 2 to 5 of this table are as in Table S2. Using the same sequence similarity threshold as in the main text (90%), the table shows a breakdown of the results for different categories of exons: translated or 3', 5' untranslated; or by position within transcripts.

exon type	# of exons with enough read counts for testing	% of exons where Var(species)>Var(tissues)	# of exons where Var(species)>0.75 or Var(tissues)>0.75	% of exons where Var(species)>0.75 or Var(tissues)>0.75 and Var(species)>Var(tissues)
translated	44903	60.34	493	14.40
5' untranslated	7503	57.55	272	22.43
3'untranslated	10799	60.15	311	25.40
middle exon	41631	59.47	479	16.08
first exon	4714	57.55	203	13.79
last exon	8739	62.92	224	28.57

Table S4: Breakdown of conserved tissue-dependent usage (CTDU) by categorisation of the exons by translation (5' untranslated, 3' untranslated, translated) and by position within transcripts (first, middle or last). The second column states the number of exons that resulted from this classification in the human genome, the third column shows the numbers of exons that also fulfilled our orthology criteria accross the six species. The fourth to eight columns indicate the fraction of exons among the sequence conserved exons that showed CTDU in comparisons between human and each of the other species; the ninth column states the fraction of exons that showed CTDU across all the species, refered to as "strictly conserved" in the main text.

exon type	total in human annotation	# sequence conserved exons	% of sequence conserved CTDU exons with regulation conservation between human and ...					strictly conserved
			bonobo	chimpanzee	gorilla	orangutan	rhesus monkey	
coding	173041	59582	2.92	2.53	2.36	2.09	2.08	0.81
5' untranslated	72287	14764	6.27	6.18	5.89	4.75	4.88	1.42
3' untranslated	64943	14441	9.04	6.60	7.05	5.27	4.94	1.75
middle exon	157897	52699	3.42	3.03	2.73	2.42	2.57	0.90
first exon	74506	10550	6.28	5.95	6.03	4.73	4.95	1.59
last exon	67428	11204	7.87	4.97	5.85	3.95	3.45	1.51