www.sciencemag.org/content/348/6235/660/suppl/DC1

Supplementary Materials for

The human transcriptome across tissues and individuals

Marta Melé, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M Goldmann, Dmitri D. Pervouchine, Timothy J. Sullivan, Rory Johnson, Ayellet V. Segrè, Sarah Djebali, Anastasia Niarchou, The GTEx Consortium, Fred A. Wright, Tuuli Lappalaienen, Miquel Calvo, Gad Getz, Emmanouil Dermitzakis, Kristin G. Ardlie,* Roderic Guigó*

*Corresponding author. E-mail: kardlie@broadinstitute.org (K.G.A.); roderic.guigo@crg.cat (R.G.)

Published 8 May 2015, *Science* **348**, 660 (2015)
DOI: 10.1126/science.aaa0355

This PDF file includes:

Materials and Methods
Figs. S1 to S25
Tables S1 to S20
Captions for data tables S4 to S7 and S9 to S18
References

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/content/348/6235/660/suppl/DC1)

Data tables S4 to S7 and S9 to S18 (Excel)

Supplementary Material for "The human transcriptome across tissues and individuals"

Marta Melé, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M Goldmann, Dmitri D. Pervouchine, Timothy J. Sullivan, Rory Johnson, Ayellet V. Segré, Sarah Djebali, Anastasia Niarchou, The GTEx Consortium, Fred A. Wright, Tuuli Lappalaienen, Miquel Calvo, Gad Getz, Emmanouil Dermitzakis, Kristin G. Ardlie and Roderic Guigo.

Table of Contents

Materials and Methods

1 Sample collection and processing

Biospecimen collection and processing of materials for nucleic acids is described in detail in (*6*) . All samples analyzed here are those defined in the GTEx Pilot analysis set (see Table S1). Excluded samples and the reasons for their exclusions are described in the "sample attributes" file that accompanies the data released to dbGaP. Following processing and QC, RNA samples were available from blood, from cell lines (LCL and Fibroblast), and from PAXgene preserved and Frozen tissue samples. All samples that met the criteria of having an RNA Integrity Number (RIN) value of 6.0 or higher and at least 1 μg of total RNA, were included and batched for RNA sequencing. To the extent possible, based on sample availability, batches for library construction were designed to include a range of samples from different tissue types and spanning multiple donors, to minimize both donor and/or tissue-specific batch effects. A set of 9 tissues was prioritized for sequencing from as many donors as possible to increase statistical power for eQTL analyses. The 9 tissues were: adipose (subcutaneous), tibial artery, heart (left ventricle), lung, muscle (skeletal), tibial nerve, skin (sun exposed), thyroid, and whole blood (Table S1). These tissues were selected based on abundance (they were routinely sampled and received), and they generally tended to meet RNA QC criteria. For the donors from whom brain tissue samples were available, all RNA samples that met QC were included for sequencing, so as to sample a broader array of tissues on some donors, along with those sampled more deeply on most donors. One control sample (K-562) underwent library construction and sequencing with each sample batch, and a set of samples were run in triplicate across separate sequencing runs. Unless otherwise specified, for most of the analyses presented here, samples were combined at the tissue level (e.g. adipose, subcutaneous was combined with adipose, visceral and is simply denoted as "adipose") to increase sample sizes for the less deeply sampled tissues (see Table S1).

2 RNA sequencing, expression quantification, and quality control

2.1 RNA Sequencing and QC

Library preparation and sequencing, as well as the data QC pipeline are described in detail in (*6*). Briefly, RNA samples meeting QC criteria were sequenced using a standard non-strand specific protocol with poly-A selection of mRNA (the Illumina Tru Seq™ protocol as implemented using a large scale automated protocol at the Broad Institute - Illumina: TruSeq Protocol Info). Sequencing was performed on Illumina HiSeq 2000 instruments, with sequence coverage to a minimum of 50M reads (corresponding to a minimum of 25M 76bp paired-end reads).

RNA-seq data were aligned with Tophat version v1.4.1 (*24*) to the UCSC human genome release version hg19 (Genome Reference Consortium GRCh37). Gencode version 12 (*9*) was used as the transcript model reference for the alignment as well as for all gene and isoform quantifications. Gencode annotated a total of 53,934 genes, which includes 20,110 protein coding genes, 11,790 long noncoding RNA's (lncRNA's), and 12,648 pseudogenes. Expression levels were produced at the gene and exon level in RPKM units (RPKM = reads per kilobase per million mapped reads [controlling for gene length and sequencing depth]), (*25*) using RNA-SeQC (*8*). Exon coordinates per gene were derived from the Gencode GTF using an isoform collapsing procedure: exons labeled as 'retained_intron' were excluded; overlapping intervals were merged; intervals associated with multiple genes were discarded; and a final gene level model was produced in GTF format.

To produce gene and exon level read count and gene level RPKM values, reads were filtered based on the following: (1) reads must be uniquely mapped (for Tophat this equates to mapping quality equal to 255); (2) reads must have proper pairs; (3) alignment distance must be <=6; (4) reads must be contained 100% within exon boundaries. Reads overlapping introns were not counted. For exon read counts, if a read overlapped multiple exons, then a fractional value equal to the portion of the read contained within that exon was allotted.

Several additional quality control metrics were applied to RNA-seq samples to determine inclusion in the final GTEx consortium pilot analysis set. All samples with fewer than 10 million mapped reads were removed, and sample outliers were identified using a correlation-based statistic, and sex incompatibility checks using the methods of (*26*) and (*6*) for more details. For all processing replicates (the same sample sequenced twice), only the sample with the greater number of reads was retained for inclusion in the final analysis set. Samples derived from the two individuals with Klinefelter's Syndrome (which failed the sex-specific expression check) and from one individual with multiple tissues that were D statistic outliers were also excluded.

Transcript isoform reconstruction was performed using the Flux Capacitor (version 1.2.3, http://flux.sammeth.net) to quantify the expression of multiple transcriptional elements. Flux quantifications distinguished 3 transcriptional elements: (1) ***splice junctions*** (gtf-feature "SJ"): all read-pairs compatible with the annotation are considered, and those aligning immediately up- and downstream of an annotated intron are considered to quantify the corresponding splice junction; (2***) introns*** (gtf-feature "intron"): all read mappings of which one mate agrees with the reference annotation and of which the other mate falls in a region that is not overlapping with any Gencode exon are considered to quantify the retention of the corresponding intronic region; and (3) ***transcripts*** (gtf-feature "transcript"). Following the deconvolution strategy described in (*27*), all read-pairs that comply with the reference annotation are represented as a system of linear equations. Based on read counts obtained by the deconvolution the RPKM measurement is then computed (*25*). More details are described in (*6*).

2.2 Comparison of gene expression array and RNA-seq data

To enable a benchmark comparison between RNA expression as measured by gene arrays and by RNA sequencing, the project ran the first approximately 1000 samples on both platforms. For the gene arrays mRNA expression data was obtained using the Affymetrix Human HT exon expression array, HuGene-1.1-ST-v1 96HTA, according to the manufacturer's specifications. Array preparation and scanning was performed by the Genetics Analysis Platform at the Broad Institute. Gene-centric expression values were obtained using updated probe set definition files (CDF files) from Brainarray (*28*); and background correction was accomplished using RMA (Robust Multichip Average) (*29*) and quantile normalization (*30*).

Following QC, there were a total of 835 RNA samples that were run on the Affymetrix Human Gene 1.1 ST array. Expression values for these samples, for 22,704 genes, were summarized using Bioconductor (http://www.bioconductor.org/). RNA-Seq data were also available for 736 of these samples. In total there were 22,273 genes which could be mapped unambiguously between the two platforms, and for which we could compare gene expression. We examined the correlation of gene expression between the platforms and variation in signal intensity among different expression classes. Considering all samples together the Pearson Correlation of gene expression between the two platforms was 0.829. Figure S3A show a correlation that is linear for moderately expressed genes, but somewhat sigmoidal when considering the extremes. RNA-seq tends to retain dynamic range at the high end of the expression spectrum. Conversely, the lowest expressed genes tend to show very little signal, while Affymetrix intensities do show variability at that end. Figure S3B indicates that RNA-seq exhibits the lowest amount of variation for highly expressed genes but the greater variation for the low expressed genes. The variation in Affymetrix signals tends to vary moderately with little impact of expression level on signal variation. This is consistent with previous studies (*31, 32*).

2.3 Analysis of ischemic time effects on GTEx gene expression data

Considering that the GTEx tissue samples are acquired post-mortem, we wanted to investigate to what extent they are representative of the same tissue samples acquired from living tissues. Acknowledging that differences will always be observed when comparing datasets across platforms and projects, we asked the general question of whether expression differences observed between GTEx samples and those of other living tissue dataset projects were comparable to, or greater than, the differences also observed between any two of the living tissue datasets. To compare GTEx tissue expression data (obtained from deceased donors) to the gene expression patterns from similar samples collected from living donors, we downloaded raw expression data from 609 samples available from the Gene Expression Omnibus (GEO) and compared their gene expression signatures to 798 GTEx samples (see Table S2). The GEO samples spanned 35 distinct collections and 5 different Affymetrix platforms (Table S2). A total of 8 tissue sites that overlapped with GTEx tissue sites were represented. All GEO samples were selected to be 'normal' (or adjacent-normal) i.e. non-diseased samples, so as to be as similar to GTEx samples as possible. Initial probe-set mapping, background correction, normalization, and calculation of gene-centric expression values for GEO samples were performed as described above. The GTEx data set was pre-processed as follows: (1) Removed non-protein coding genes, (2) Removed genes with low expression (RPKM<10) and low variation (Fold Change < 2 and Delta < 10) across samples, (3) Applied ceiling to extremely high RPKM values (RPKM ≥10,000), (4) Rank normalized expression values. All GEO data sets were column-rank normalized.

As a preliminary analysis we calculated the spearman correlation coefficients between the GTEx samples and the GEO samples, matched by tissue. Choosing the same number of GTEx and GEO samples in each tissue group, we computed a correlation matrix, producing a value for each combination of samples from GTEx with samples from GEO. For each tissue we recorded the median values of the correlation matrix. The median of values for the five tissues investigated was 0.723. We then repeated the analysis comparing results from among the different GEO data sources over five tissues. This produced the nearly identical level of correlation of 0.718.

Given that many factors can contribute to discrepancies across platforms and projects, we investigated whether factorization methods would be more powerful for finding the strongest distinct biological signals in the GTEx and external data and then performing the comparison utilizing these factors. This was done by employing the metagene projection methodology described by Tamayo and colleagues (*12, 33*). This methodology is ideal for cross platform comparison of gene expression data sets because it compares linear combinations of gene expression values, or *metagenes*, rather than individual gene expression values.

Prior to performing the metagene projection, we used consensus non-negative matrix factorization (NMF) to determine the optimal number of latent factors, or *metagenes*, to extract from the GTEx gene expression dataset. Evaluation of the values of *k* from 4-13 revealed that *k*=7 was the optimum number of metagenes to extract that best defined the 8 tissues in the GTEx dataset (*34*). The number of clusters was determined objectively by the method, using consensus clustering and maximizing the cophenetic correlation coefficient, which does not force a structure on the data. A support vector machine (SVM) classifier was trained on the GTEx data set: (1) The GTEx data set was split into train and test data sets (2) metagene class weights were derived for each tissue type for the training data set, and (3) the tissue types of the test data set were then predicted. Results are summarized in Figure 1C and Table S3A. To assess how well the GTEx samples represent the gene expression patterns of samples collected from living donors, each GEO data set was then projected into the metagene space via a pseudo inverse projection. Similar to principal component analysis, this procedure reduces the entire gene expression data set to *k*=7 metagenes. The tissue type of each sample in the GEO data sets was then predicted using the previously developed SVM, which was highly accurate in predicting the tissue type of the samples collected surgically. Results are summarized in Figure 1C and Table S3B.

3 Gene expression analysis

3.1 Expression data summary statistics

For each RNA-seq experiment, mapped read statistics were computed using an in-house script based on the bam file. Briefly, the primary alignments were partitioned into the following categories: exonic, intronic, exonic-intronic, intergenic and other. Mappings were also further subdivided into contiguous mappings and split mappings. Figure S1 shows the proportion of mapped reads in the different genomic categories across all the tissues.

We considered a gene to be detected/expressed if it had normalized expression value greater than 0.1 RPKMs (RPKM > 0.1). The consortium chose 0.1 RPKM as a threshold for the GTEx pilot data set (and all related analyses) because it corresponded to about 5 reads in most genes. Considering RNASeq to be a Poisson-like sampling exercise this cutoff reduced the sampling noise drastically while remaining broadly inclusive. This corresponds to about 88% of protein coding genes, and 71% of lncRNAs being expressed at > 0.1 RPKM in at least one sample. In other words, ~12% of protein coding genes and 29% of lncRNAs are expressed between 0 and 0.1 RPKM across all samples. The 0.1 threshold allowed us to investigate expression of lncRNAs, which are typically expressed at lower levels. Indeed, while 31% of lncRNAs are detected on average per sample at the 0.1 threshold, only 13% are detected at more than 1 RPKM. Thus, a substantial number of lncRNAs are detected in the > 0.1 and < 1 range. Also, as noted by our comparison with microarrays (Figure S2), at the 0.1 threshold for RNASeq data, there is still substantial variation detected by microarrays.

The number of expressed genes across all the samples was categorized for all genes, protein-coding according to the Gencode annotation (*9*). Long non-coding RNAs (lncRNAs) were classified according to gene group described in Table S20 as in (*35*).

3.2 Selection of samples and tissues for analyses

As described above (Section 1) the distinctly sampled tissue sites described in Table S1 were analyzed separately in some instances, or were combined at the tissue level (e.g. “adipose, subcutaneous” was combined with” adipose, visceral” and simply denoted as “adipose”) dependent on the analysis being done, to increase sample sizes for the less deeply sampled tissues. For the analyses were larger sample sizes were required we focused on the nine tissues that were used for eQTL analysis in the main paper (*6*), combining samples at the tissue level. These are denoted here simply as: adipose, artery, heart, lung, muscle, nerve, skin, thyroid and blood and we refer to them as the 9 main combined tissues across the manuscript. Brain is the tissue with the largest ‘combined’ sample size but due to the heterogeneity of its sub-sampled regions, and unless stated otherwise, it was only used in analyses that specifically considered its different sub-regions. For some analyses we have included tissues other than the primary 9 combined tissues, in cases where these were relevant, for example including testis in the analysis of tissue specificity of splicing given its predominant role in the earlier analysis of tissue specificity. Overall, for differential expression analysis we considered the twenty tissues with ten or mores samples.

3.3 Tissue Clustering and multi-dimensional scaling

We explored gene expression similarity between tissues and across samples, by performing hierarchical clustering (HC) using different settings. RPKM values were used in log2-transformed (log2(1+rpkm)) scale. Distance between samples being defined as *distance = 1 – correlation*. Pearson was used as the correlation measure, although Pearson and Spearman correlations showed similar results (results not shown). Average linkage method was used for all the tested settings. All the genes from the annotation were considered. To create a genealogy of tissues (Figure 1B), we calculated the centroid expression by obtaining the median expression across all the samples of a given tissue. HC was then performed as described above. Multidimensional scaling was performed to represent the distances among samples in a parsimonious way. We used the isoMDS function from R, with the distance being defined as for the HC analysis (*36*).

3.4 Transcriptome complexity analysis

We calculated the average contribution of each gene to the total transcriptional output of a tissue (*37*), following the procedure below:

1. We calculated the average expression of each gene across all samples of the same tissue
2. For each tissue, we sorted the average expression values in decreasing order and divided each value by the sum of all average expression values. This measures how much each gene contributes to the overall transcriptional load of that tissue.
3. We plotted (figure 1D) the cumulative distribution of the contribution of each gene to the total transcriptional output.
4. Standard deviations were also divided by the sum of all mean RPKM values and depicted as error bands.

Low complexity tissues will be those tissues in which a low number of genes contribute to a large fraction of the transcriptional output whereas high complexity tissues will have many genes equally contributing to the total transcriptional output. We calculated transcriptome complexity for all genes and all tissues with more than one sample (Figure 1D). Genes are classified according to gene group described in Table S20, except for genes encoded in the mitochondria that are treated as a separate group. The top hundred most expressed genes in each tissue can be found in Table S4.

3.5 Differential gene expression across tissues and tissue specificity analysis

In this section we used the twenty tissues with at least ten or more samples. For statistical tests across the manuscript we used a False Discovery Rate (FDR) implemented through Benjamini-Hochberg (BH) (*38*) implemented in the R package *multtest (39)* or Q-value estimation (*40*) where 1- Q-value =FDR. We used for the majority of the analysis a threshold of 0.05. In some specific analysis we required a more stringent FDR threshold of 0.01.

3.5.1 Pairwise differential gene expression analysis

Differential expression was performed with NOISeq (*41*) and DEseq2 (*42*). We used the noiseqbio function with q > 0.95 (q = 1-FDR thus FDR=0.05) as cut-off for statistical significance and FDR=0.05 in DESeq2. For NOISeq we used as input RPKM normalized values and for DESeq2 we used the read counts after TMM normalization (*43*). All pairwise combinations between all tissues were tested. We took a conservative approach and defined genes differentially expressed as those found to be in common between the two methods and passing the thresholds above (see figure S8).

3.5.2 Tissue preferential gene expression analysis

The analyses were performed in a similar manner as the differential gene expression between tissues, but in this analysis, due to scalability issues, only the NOISeq method was used. We used the noiseqbio function with q > 0.99 (FDR=0.01) and log2 fold change >= 4 to call tissue preferential gene expression after comparing the samples from a given tissue to those samples that did not belong to the tissue (Figure S9 and Table S5).

3.5.3 Tissue exclusivity analysis

To find tissue exclusive genes, we calculated the phi correlation coefficient on the contingency table generated by dividing the samples based on two conditions. The first condition selected samples that were either expressed (RPKM > 0.1) or were not expressed (RPKM < 0.1). We selected a threshold of 0.1 RPKM to conservatively select tissue specific genes that had substantial expression. The second condition selected samples coming either from the tested tissue or from all other tissues. We calculated the phi-correlation coefficient using function phi of the R package psych (*44*). Phi correlation coefficient measures association between two binary variables. We defined those genes with phi values higher or equal than 0.95 or lower or equal than -0.95 as tissue exclusive genes (Figure S10 and S11, and Table S6). To assess whether a threshold of expression > 0.1 to define expressed genes could affect our results; we run the same analysis using different thresholds (Table S7).

3.6 Repeat elements analysis

3.6.1 Repeat elements expression analysis

We used the RepeatMasker annotation (http://www.repeatmasker.org/) to define repeats and we removed those repeats overlapping coding regions (CDS in Gencode v12 annotation). We counted the number of reads overlapping each annotated repeat instance in the 1,486 samples from the 10 tissues with highest sample size (the 9 main combined tissues plus brain considering all samples together). We normalized read counts using TMM normalization (*43*) package to correct for differences in library sizes.

To discriminate true expression from noise, we used expression profiles from those repeats that were most likely not expressed. For each tissue, we selected those repeat instances that had no reads mapped in at least half of the samples. Then, expression values in the other half of the samples were used as a proxy of noise expression distribution. We used the 99th percentile of this expression distribution as a threshold so that a repeat was considered expressed in a tissue if the median expression across the samples was above the noise threshold. Using this approach, from the 5,285,549 annotated repeats, 209,541 were expressed in at least one tissue.

We ran hierarchical clustering based on average linkage criterion on the read counts of the 209,541 repeats using the same settings as for gene expression. The clustering recapitulated tissue classification. This clustering held when focusing on the subset of repeats (62,539) located more than 3kbp away from a gene, suggesting that expression likely originates from actual tissue-specific patterns as opposed to by-products of gene expression (Fig S12A).

3.6.2 Expression correlation analysis of repeat element and nearby genes

Because some repeats may induce expression of the gene nearby by recruiting relevant transcription factors (*45*), we compared expression of each repeat to its nearest gene (upstream or downstream). We computed Pearson correlation between repeat and gene in each tissue separately. We assessed significance by computing correlations between randomly chosen pairs of repeat/gene. A p-value for positive (negative) correlation was computed from the number of control pairs with higher (lower) correlation. Finally, to correct for multiple testing we applied False Discovery Rate control using BH algorithm (*38*). While no significant negative correlation was found, thousands of repeats expression were significantly positively associated with gene expression.

Additionally, if a particular repeat is indeed affected by some regulatory processes, it is likely to show similar expression patterns at the family level. This could be explained because repeat from the same repeat family share extensive sequence similarity and hence could be regulated by the same or similar factors. For each repeat family, we computed the average of the Pearson correlation between all possible pairs of repeats from the same family. We assessed significance by computing the average Pearson correlation between repeat pairs selected from a group of repeats of size equal to the studied family that had been randomly chosen. We then corrected for multiple testing (FDR <0.05) using BH method. In total, 3966 repeats showed significant correlation (FDR <0.05) between their expression and the gene nearby as well as significant family co-expression (FDR <0.05) in at least one tissue. These instances are potentially implicated in the regulatory processes controlling gene expression. Moreover, 276 of these repeats are located upstream of the gene, far enough (3kbp) not to be confounded by gene expression by-products (see last column of Table S8).

Multiple mapping reads could cause spurious correlations between repeats of the same family. Moreover, we did not use mapping quality filter because we wanted to use the maximum number of reads for the analysis. In order to test the existence of multiple mapping biases in our analysis, we explored the relation between expression correlation between two repeats of the same family and their sequence similarity. If there is no significant correlation, we can dismiss the effect of mapping bias. Figure S13A shows an example of a small repeat family with no clear relationship between pairwise sequence similarities between pairs of repeats and correlated expression. Pearson correlation was used to calculate expression correlation across samples. Sequence similarity was calculated as the proportion of matched nucleotides when aligning both repeats sequences (global-local alignment with gap opening and extension penalties of -10 and -4 respectively).

Then, for each repeat family, we selected the pair of repeats with the highest expression correlation and computed their sequence similarity. We also replaced one of the paired repeats by another repeat of similar size from the same family but with no expression correlation and computed their sequence similarity. Figure S13B, represents the distribution of the difference in sequence similarity between the coexpressed pair of repeats and the non-coexpressed pair. This distribution is nicely centered in zero meaning that, on average, two co-expressed repeats in a family are as similar as two non-coexpressed ones. Therefore, for the majority of families, co-expression is not due to mapping artifact. Figure S13B also shows the difference in expression correlation between the coexpressed pair of repeats and the non-coexpressed pair.

3.6.3 Estimation of repeat element effect on lncRNA expression

The set of Gencode 19 lncRNA transcripts (*35*) were intersected with RepeatMasker repeats using a custom script based on Bedtools IntersectBed (*46*). Those transcripts whose annotated transcription start site fell within an annotated repeat were defined to be repeat-promoted and selected for further study. Analysis was carried out at the level of Repeat Class. For each tissue, the mean expression of the repeat-promoted lncRNAs was computed. The top 500 most expressed lncRNAs were selected and classified according to Repeat Class. Then, a contingency table crossing tissue and Repeat Class was obtained, where cells had the observed frequencies of each Repeat Class at each tissue. To investigate the relationship between repeat promoter and tissue expression we carried out Correspondence Analysis. This technique can represent a contingency table as a map of points representing the rows and columns of the table where those variables that are correlated will appear closer in the plot (Figure S12C). The ca library from R package (*47*) was used to carry out the Correspondance Analysis.

4 Analysis of the contribution of tissue, individual, sex, ethnicity and age to gene expression variation

4.1 Estimation of tissue and individual contribution to gene expression variation

To assess the contribution of tissue and individual to gene expression variation, we used a linear mixed model (LMM). Gene expression was modeled as a function of tissue and individual (considered as random factors). The LMM was implemented in the R package lme4 (*48*). Genes not expressed (RPKM > 0) in any of the samples were excluded from the analysis (overall 31,059 genes, 11,508 lncRNAs and 19,550 protein coding, were analyzed). We used log2 (RPKMs) to normalize the data and pseudocounts to deal with zero expression values. To obtain the variance components, we divided the restricted maximum likelihood (REML) estimators for the random effects of tissue, individual and residual variance by their sum (Table S9).

To compare protein coding genes and lncRNAs with similar expression levels, we computed the median of the log2 expression across samples (1641 samples) for each gene. We visually examined the scatter plot of the contribution to expression variation of tissue plus individual versus median expression. We observed that there was a correlation between them and that this correlation reached a plateau at around median expression greater than 2.5 RPKMs. We then selected those lncRNAs and protein coding genes with higher median expression than 2.5 and calculated the average contribution of individual and tissue to gene expression variation.

4.2 Sex, ethnicity and age differential gene expression analysis

We investigated the effect of sex, ethnicity and age in gene expression extending the LMM model above to incorporate sex, ethnicity and age as covariates together with individual and tissue. Available methods such as DESeq (*49*) or NOISeq (*41*) can not analyze mixed models. We considered individuals as block random effects (note that individuals are not necessarily the same for all tissues; some individuals may provide a subset of tissues). We used the function lme of the nlme (*50*) package of R. We can write the model as

$$y_{ijk} = t_i + g_j + b_k + e_{ijk}$$

i=1,...,T; T = #tissues
j=1,...,L; L= #levels of the factor
k=1,...,I; I= #individuals

Where y_{ijk} represents the *ijk*th observation (log gene expression) on the *i*th tissue *j*th level of the factor and *k*th individual, e_{ijk} represents the random error present in the *ijk*th observation on the *i*th tissue *j*th level of the factor and *k*th individual. Tissue is a fixed effect, sex and ethnicity are fixed factors and age is a covariate, and Individual is the random effect of the mixed model. We use t_i to denote the tissue effect, g_j to denote the factor (sex/ethnicity/age) effect and b_k the individual effect. Individual random effects b_k are assumed to be normally and independently distributed with $\mathrm{var}(\mathbf{b}) = \sigma_b^2\mathbf{I}$. The errors e_{ijk} are assumed to be normally and independently distributed with $\mathrm{var}(\mathbf{e}) = \sigma^2\mathbf{I}$.

RPKM values were log2 normalized. A minimum amount of variation across individuals and tissues was necessary to fit our model. Therefore, we selected those genes that had expression higher than 0.1 RPKMs in at least 5% of the samples and focused on protein coding genes and lncRNAs. All tissues were included in this analysis.

All analysis were corrected for multiple testing at FDR < 0.05 using BH method .To run functional enrichment analysis we used the program DAVID (*51*) with default settings. GWAS hits were seek using the GWAS table downloaded UCSC (*52*). To find genes laying in regions reported to be under positive selection in Europeans or African Americans we used information from (*53, 54*) of regions reported to be selected either in Bantu or Yoruba populations or populations of European ancestry.

To assess whether the genes that decreased expression with age had significantly more SNP hits for Parkinson and Alzheimer GWAS than the rest of genes in the genome we run a fisher exact test (p<0.05).

Genes differentially expressed by sex, ethnicity and age can be found at tables S10, S11 and S12 respectively.

4.3 Sex and ethnicity differential gene expression analysis tissue by tissue

We used DESeq (*49*) to perform differential expression analysis by gender and ethnicity within each tissue separately. We analyzed those tissues with at least 10 samples per condition. In the case of brain, we analyzed it region by region. All analysis were corrected for multiple testing at FDR < 0.05 using BH method (*38*). To run functional enrichment analysis we used the program DAVID (*51*) using default settings. GWAS hits were seek using the GWAS table downloaded from UCSC (*52*). To assess whether skin had significant more ethnicity differentially expressed genes than other tissues we run a binomial test (p<0.05). Sex and ethnicity differentially expressed genes in each tissue can be found in tables S13 and S14 respectively.

4.4 Gender differential co-expression network analysis

We explored differential analysis of co-expression networks (*55*) between males and females using gene expression. Data from the nine main combined tissues (with n>60 per tissue), plus breast was used in a total of 381 female and 637 males samples.

The purpose of constructing the coexpression networks is to discover groups of genes that are functionally related (either perform similar functions or are part of the same biological process). By comparing the topology of equivalent modules across data sets, i. e. males vs females, we can identify biological processes or parts of processes that are performed differently between data sets; moreover, by examining differences in hub genes we also get information on the genes driving the observed differences. (The criterion for grouping genes has been shown to produce biologically meaningful modules and by adopting a systems-level view - modules rather than individual genes - we significantly alleviate the multiple testing issues).

We built co-expression networks independently for each gender and identified 42 modules in males and 46 in females. We matched them based on the number of overlapping genes using Fisher's exact test and found 39 modules in common between males and females. For each module, we assessed the preservation of its topology across the two groups using a measure of intramodular connectivity similarity (correlation of kME, see below). We found that in 36 out of 39 modules the network topology is similar in both data sets. Figure S16 shows examples of modules that were identified in both sets but exhibit different topologies, as well as modules that were only identified in one data set (either in males or in females).

4.4.1 Network Construction

Networks for each dataset were constructed using weighted correlation network analysis (*55, 56*). For each dataset, we computed an adjacency matrix:

$$a_{ij} = \left|corr\left(e_i, e_j\right)\right|^{\beta}$$

where e_n is the expression of gene n, $corr$ is the Pearson correlation and β is the soft-thresholding power (*56*). We set $\beta = 6$, the default value for unsigned networks (there is no motivation for applying the scale-free topology criterion here, as our samples comprise several different tissues). The adjacency matrix was then transformed to a similarity measure appropriate for clustering, the Topological Overlap Matrix (TOM):

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min\{\sum_u a_{iu}, \sum_u a_{ju}\} + 1 - a_{ij}}$$

The TOM is a measure of the interconnectedness of genes i and j (specifically, their shared neighbors). It has been shown to produce biologically meaningful and tighter modules than the correlation of expression alone (*57*).

4.4.2 Module Detection

The TOM was transformed to a dissimilarity measure (1-TOM) and genes were clustered using average linkage hierarchical clustering. Modules were derived from the dendrogram using the Dynamic Tree Cut package (*56*), which eliminates the need for manual choice of a cutoff height – something that could result in poor module definition in complicated dendrograms. The minimum module size was set to 30, in accordance with common practice.

The expression profile of each module is summarized by its eigengene, which is the first principal component.

4.4.3 Intramodular connectivity - hub preservation using kME

kME is calculated as the Pearson correlation of expression of a gene with the module eigengene and it is a measure of intramodular connectivity (*58*). Hub genes exhibit a high kME. In order to assess hub preservation of a module across datasets, we compute the Pearson correlation of kME of the genes comprising the module of interest. If the correlation is high (implying preservation of intramodular connectivity), genes with high kME in all datasets retain their hub status.

4.4.4 Module comparison

In order to compare male to female modules, we first identified similar modules in terms of gene content between male and female networks. Then, for each gene in a given module, we computed kME (the Pearson correlation of its expression with the module eigengene). This was done separately for males and females. Then, we computed the correlation of kME between the two datasets (the kME correlation between the male and the female module). High correlation of kME suggests preservation of connection patterns for the module genes.

We then refer to modules that are common across data sets but whose network topology is not preserved, for those modules where the correlation of kME in the two data sets is below 0.7. Male- or female-specific modules are those that only occur in one of the data set and have no counterpart in the other. For such modules we also plotted the connections of that same group of genes in the other data set to highlight the differences in connection patterns (Figure S16A, B).

5 Alternative splicing analysis

5.1 Splice junction analysis

5.1.1 Detection of annotated and novel splice junctions

We selected split-mapped reads (reads that do not map contiguously to the genome but when split in two parts can be mapped independently) from TopHat mappings (*24*) to identify exon-exon junctions. By default TopHat reports "GT-AG", "GC-AG" and "AT-AC" introns. Split-mapped reads were clustered and each splice junction (SJ) was annotated with the number of supporting reads and the tissues/samples where it occurred. Reads from all the samples were also pooled to obtain a transcriptome wide set of splice-junctions. Comparison with introns from the Gencode annotation allowed classifying SJ as known/annotated (if already present in the annotation), novel SJ with one known splice site (if one splice site is present in the annotation and the other is unknown), novel SJ with two known splice sites (if both splice sites are present in the annotation but the intron between them is not annotated) or novel SJ with unknown splice sites (if both splice sites are not in the annotation) (Figure S17).

We defined two sets of splice junctions: a) high confidence set SJ and b) a less filtered set of SJ. To identify a set a) of strongly supported highly reliable splice junctions, we employed a stringent threshold of entropy > 3 computed on the distribution of split-points in reads aligned to each splice junction. Using this threshold, we identified 329,984 splice junctions, of which about 25% (87.005) are novel. Strongly supported novel junctions tend to be more tissue specific, detected in fewer number of samples (Figure S17A). A set of less stringently selected SJs b) was defined by selecting those SJ that are supported by at least 3 split-mapped reads when pooling all the samples of a given tissues. This set support millions of novel splice junctions in the human genome (Figure S17B).

To find one-to-one orthologous mouse to human splice junctions, human splice sites were projected onto the mouse genome by a per-nucleotide lift-over procedure (*59*) using filtered pairwise whole-genome chain alignments (*60*). Similarly, mouse splice sites were projected to the human genome. Splice sites that were mapped uniquely and bijectively (i.e., the human-to-mouse and mouse-to-human projections were mutually inverse as functions) were said to be one-to-one orthologs. A human splice junction was said to be one-to-one orthologous to a mouse splice junction if the corresponding splice sites were orthologous (as defined above).

5.1.2 Proportion of genes with novel splice junctions

Given that different tissues express different number of genes, we calculated the proportion of genes with unannotated splicing events per sample. For each sample, we divided the number of genes with at least one unannotated splice junction detected by the number of genes where at least one annotated splice junction was detected. This ratio corresponded to the proportion of genes with unannotated alternative splicing events over all genes spliced per sample (Figure S22).

5.2 Exon inclusion based clustering

5.2.1 Percent spliced in (PSI) estimation

Exon inclusion levels were calculated for all internal exons of genes with three or more exons. We calculated the 'Percent Spliced-in' (PSI) (*61*) as in (*62*). The PSI measure for each exon is defined as the ratio between the reads that support the inclusion of the exon and the sum of the reads that support the inclusion plus the exclusion of the exon. PSI values range between 0 and 1, where 1 represents full inclusion of the exon and 0 full exclusion. For an internal exon C and its neighbor exons A1 and A2, *Inc* corresponds to reads that support the junction A1-C and *Inc'* the junction C-A2. *Exc* reads support the junction A1-A2. The PSI formula is then defined as PSI = avg(Inc,Inc') / (avg(Inc,Inc') + Exc). Only exons supported by a sufficient number of reads, Inc + Inc' + Exc >= 10, were considered.

5.2.2 Correlation analysis between expression patterns of RNA binding proteins and splicing patterns of all genes across samples

We selected a set of 67 human curated RNA-binding (RBP) splicing regulatory proteins from the SpliceAid-F database (*63*) to analyze the relation between the expression levels of these genes across tissues and the differential splicing patterns found across tissues. Splicing hierarchical clustering was performed for all samples based on PSI values (selected 54,330 exons with PSI values in more than 90% of the samples). We used the same settings as the HC performed in expression clustering, i.e. distance= 1 – Pearson correlation, and average linkage clustering method. The "na.or.complete" parameter was used to handle missing values. We then plotted the normalized gene expression of the 67 RBPs according to the order of the samples derived from the splicing clustering (Figure S18).

5.2.3 Exon differential and preferential inclusion analysis

Differential and preferential exon inclusion was performed using PSI values. For differential exon inclusion, tissues were compared in a pairwise manner (Figure S19A). For preferential exon inclusion analysis, we compared exon inclusion values in one tissue with the remaining tissues (Figure S19B, Table S15). We applied the Wilcoxon test in R (*64*), with p-values corrected by the BH method (*38*). Tissues with ten or more samples were selected for this analysis. Following the methodology in (*65*) for differential exon inclusion, exons are considered differentially included if FDR < 0.01 and absolute difference in median PSI between groups > 0.1.

5.3 Differential exon inclusion

5.3.1 Tissue exclusivity analysis of exons

We analyzed 20,219 exons in which there was PSI variation across samples (at least two PSI values were different). For each exon and for each (tested) tissue, we computed a two-way contingency table calculating the sample frequency depending on two conditions. The first condition selected samples that either had a PSI value > 0.8 or had a PSI value < 0.5. The second condition selected samples coming either from the tested tissue or from all other tissues. Then, we calculated the phi-correlation coefficient using function phi of the R package psych (*44*). Phi correlation coefficient measures association between two binary variables. We selected those exons with phi values higher than 0.95 or lower than -0.95 for further study. Tissues analyzed included the main 9 combined tissues plus brain (given its differential splicing pattern) and testis (given its high number of tissue specific genes) (Figure S20, Table S16).

5.3.2 Tissue exclusivity analysis in microexons

We selected multi-split alignments (i.e., the alignments that were split at least twice) requiring that (1) each split had the canonical GT/AG splice sites and (2) each split was confirmed by at least two staggered reads (not necessarily multi-split). The sum of at least five reads supporting inclusion and exclusion was required to compute PSI in each sample. The presence call for a short exon was made if PSI value could be computed in at least 25% of samples. In total, we detected 335 exons shorter than 16nt (referred to as microexons), of which 28 were not annotated in the latest version of Gencode (v19). To assess whether some microexons were preferentially expressed in specific tissues, we computed the phi statistic for each tissue and microexon from a 2x2 contingency table depending on two conditions. The first condition selected samples that either had a PSI value > 0.8 or had a PSI value < 0.5; samples with intermediate PSI values were discarded. The second condition selected samples coming either from the tested tissue or from all other tissues. Only the twenty tissues with at least ten samples were used for this analysis. Functional enrichment for genes containing microexons was calculated using the R package GOstat (*66*). To test whether microexons were more included in brain compared to the other tissues, we run Wilcoxon test on the phi distributions in brain vs. non-brain tissues for microexons (length<15nt) and the same for other small exons (50< length (nt) <80). The phi distribution values for microexons across tissues can be seen in Figure 3B and for short exons in Figure S21.

5.3.3 Contribution of tissue and individual to exon inclusion variation

In a manner parallel to that for gene expression, we assessed the contribution of tissue and individual to exon inclusion (PSI) variation. We used Percent spliced in (PSI) values as calculated in section "Percent spliced in" (PSI) estimation and selected the 20,219 exons for which PSI values varied at least between two samples.

To assess the contribution of tissue and individual to exon inclusion variation, we used a linear mixed model (LMM). Exon inclusion was modeled as a function of tissue and individual considered as random factors. The LMM was implemented in the R package lme4 (*48*). The restricted maximum likelihood (REML) estimators for the random effects of tissue, individual and residual variance were normalized by their sum to give the variance components (Figure S23).

5.3.4 Effect of ischemic time on splicing and expression

We performed hierarchical clustering (distance=1-Pearson correlation and average clustering method) on the samples of the nine main combined using exon inclusion levels (PSI). Across the different tissues, we detected two distinct clusters of individuals (a larger and smaller sub-cluster). For each tissue, we then split the dendrogram in order to retrieve these two distinct clusters (using cutree function in R). Next, for the different tissues we intersected the respective larger and smaller cluster samples. We then retrieved the identifiers of the individuals that are common across the larger and smaller clusters in the different tissues. In the smaller sub-cluster we found a set of samples originating from 17 individuals that are common to the corresponding sub-cluster in at least 5 tissues. We colored the samples corresponding to these individuals in red and the remaining samples in black. We also performed hierarchical clustering for gene expression in these same nine main combined tissues (clustering was performed for each tissue using the same settings as for the MDS analysis, see Tissue clustering section). Finally, we plotted the distribution of the ischemic time (in minutes) between the "red" and "black" group of individuals (Figure S24).

We additionally investigated if this effect had a bias in some particular part of the transcript, for example if there might be an effect due to partial transcript degradation of exons located in the 3' part of the gene that are more affected. We ranked exons by their differential splicing between these two groups of individuals and classified then according to the relative distance (0 to 1) to the start of the transcript. However, we found that differential spliced exons have no particular enrichment along the transcripts, suggesting no evident effect of 3' degradation.

This data suggest that even though we do not see a substantial impact of ischemic time on the gene expression, there could be a larger impact on gene splicing. We will continue these analyses, as larger sample numbers are available across both tissues, and ischemic time points.

5.3.5 Analysis of the contribution of tissue and individual to splicing variation

In a manner parallel to that for gene expression, we assessed the contribution of tissue and individual to splicing variation. We analyzed 10,597 protein-coding genes with more than one isoform and for which the sum of transcript expressions was greater than zero in all samples. We used ten tissues (9 main combined tissues + brain, considering all samples together). We used only individuals for which data existed for all these tissues. This resulted in 38 individuals. We used quantifications of abundances of transcript isoforms. Because the splicing of a gene is, thus, represented as a multivariate distribution, we developed an approach to estimate the components of the splicing variability in each gene based on an orthogonal decomposition of the gene's total sum of squares. In detail:

Let y_{ijk} denote the square root relative abundance of the isoform *i* in the tissue *j* in the individual *k*. Let J (=10) denote the number of tissues and let K (=38) denote the number of individuals. The square root transformation used to measure the variability follows the approaches in (*67*) and in (*68*). Equivalent expressions can be derived by using x_{ijk}, the square root absolute abundance of the *i* isoform in the tissue *j* of the *k* individual.
The total sum of squares for a specific gene, *SST*, is given by adding the total sum of squares of all the transcripts in the gene. Following the classical ANOVA decomposition of variability sources, the total sum of squares of each transcript, *SST_i*, decomposes as the sum of the sum of squares among tissues, *SSTs_i*, plus the sum of squares among individuals, *SSI_i*, plus the residual sum of squares, *SSR_i*. That is:

$$SST = \sum_{i}^{I} SST_i = \sum_{i}^{I} \left\{ SSTs_i + SSI_i + SSR_i \right\} = \sum_{i}^{I} SSTs_i + \sum_{i}^{I} SSI_i + \sum_{i}^{I} SSR_i$$

If $\bar{y}_{ij.}$ denotes the mean relative expression of the isoform *i* in the tissue *j*, $\bar{y}_{i..}$ the mean of the isoform *i*, and $\bar{y}_{i.k}$ denotes the mean expression of the isoform *i* in the individual *k,* the sources of variation can be expressed as:

$$SSTs_i = K \sum_{j}^{J} (\bar{y}_{ij.} - \bar{y}_{i..})^2 ,$$

$$SSI_i = J\sum_{k}^{K}(\bar{y}_{i.k} - \bar{y}_{i..})^2,$$

$$SSR_i = \sum_{j}^{J}\sum_{k}^{K}(y_{ijk} - \bar{y}_{ij.} - \bar{y}_{i.k} + \bar{y}_{i..})^2.$$

Finally, the relative variability components are estimated by

$$\frac{\sum_{i}^{I} SSTs_i}{SST}, \quad \frac{\sum_{i}^{I} SSI_i}{SST}, \quad \frac{\sum_{i}^{I} SSR_i}{SST}$$

which are, respectively, the relative variability among tissues, the relative variability among individuals and the relative residual variability respect to the total variability of the gene.

The contribution of individual and tissue to splicing variation across genes can be seen in Figure 3D and for each gene values can be found in Table S17.

5.4 Contribution of gene expression and alternative splicing to transcript abundance variation

Gene expression contribution to transcript abundance variation was computed following the methodology developed in (*67*). In a nutshell, for each gene, samples are represented in a multidimensional space using their transcript abundances as coordinate. The contribution of gene expression in the transcript abundance variation is computed by the variation after projecting the samples into a model of constant splicing (a line in the multidimensional space) divided by the total variation without projection (see below). If this ratio is close to 1 so that projecting into the "no splicing" model did not reduce transcript variation, the main contributor to transcript abundance variation would be gene expression. Conversely, if the ratio is close to 0, alternative splicing would be responsible for most of the variation in transcript abundance. Additionally, we implemented two improvements on the version described in (*67*). First, the effect of outlier samples is mitigated by means of a bootstraping approach. Second the contribution of gene expression to transcript abundance when the major isoform is extremely abundant can be overestimated. Here we reduced this effect by rescaling transcript abundances using square-root transformation. Each tissue was analyzed separately.

We have extended the methodology of (*67*) to include the analysis of a between source of variation in the "no-splicing" model. The contribution of gene expression in transcript abundance variation within a tissue was generalized for the multiple-tissue design when samples from different tissues are studied together. Precisely, we asked how much of the transcript variation attributed to tissue is due to changes in gene expression. In practice we compared the proportion of variation explained by the tissue classification after and before projecting the samples into the “no splicing” model. The proportion of variance explained by tissue classification was derived from classical ANOVA decomposition. The “no splicing” model was represented by a line in the multi-dimensional space formed by the different transcripts abundances. As in the within-tissue analysis, a value around 1 means that the projection didn't affect the estimate of variance explained, supporting a full contribution of gene expression. A ratio around 0 means that the variance explained was greatly reduced after projection, supporting a major contribution of alternative splicing. We then asked the same question about isoform abundance variation between individuals. Here the samples were grouped according to their individual of origin to assess the contribution of gene expression in the variation attributed to individual variation. To avoid bias from inconsistent sample availability, we used only samples from the nine main tissues and individuals with samples available for all these tissues.

More precisely, if x_{ijk} stands for the square root absolute abundance of the *i* isoform in the tissue *j* of the *k* individual, the multiplicative model defined in (*67*) can be expressed here giving the expression level for each transcript as the product of a global expression parameter λ_{jk} of the tissue of this individual multiplied by the relative expression level p_i of each transcript:

$$\vec{x}_{jk} = \begin{pmatrix} x_{1jk} \\ x_{2jk} \\ \vdots \\ x_{Ijk} \end{pmatrix} = \lambda_{jk} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_I \end{pmatrix}$$

The model assumes therefore a constant splicing ratio over all the individuals and tissues. Let define the sample size of each tissue by K_j, the total number of samples by $K = \sum K_j$ and X=[x_{ijk}] as the matrix of counts organized in *I* rows and *K* columns. In (*67*) it is shown that using the least square criteria to fit the above model, with the restrictions $\lambda_{jk} \geq 0$ and $p_i \geq 0$, allow to obtain the projections of the samples in the constant splicing model subspace expressed in terms of the first left singular vector u_1 of *X*:

$$\vec{z}_{jk} = \begin{pmatrix} z_{1jk} \\ z_{2jk} \\ \vdots \\ z_{Ijk} \end{pmatrix} = u_1^t \vec{x}_{jk}$$

The line of $\mathbb{R}^I$ defined with the above formula minimizes the distance between the original and the *K* projected points. V_{ls} is defined in (*67*) as the variability explained by the multiplicative model. V_{ls} can be expressed in terms of $\sum$, the sample covariance matrix of the *I*-dimensional vectors of splicing counts:

$$V_{ls} = u_1^t \Sigma u_1$$

The total variation V_T of *X* is defined as the sum of the variances of the alternative splice forms across the K_j individuals of the *J* tissues. Similar to the *Variability decomposition splicing* section above, V_T can be decomposed in a tissue variability term (between) and a residual (within) variability term:

$$SS_T = KV_T = K\,\mathrm{tr}(\Sigma) = SS_B + SS_W$$

If $\sum_j$ are the *J* sample covariance matrices on the different groups/tissues, the between term can be expressed as:

$$SS_B = K\mathrm{tr}(\Sigma) - \sum_{j=1}^{J} K_j \mathrm{tr}(\Sigma_{\mathrm{j}}) = \sum_{i=1}^{I}\sum_{j=1}^{J} K_j \left(\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot}\right)^2$$

Considering again the constant splicing model, if $\mathrm{SS}_{\mathrm{Tls}}$ is the total variability of the projected samples and $\sum_Z$ the sample covariance matrix of the projected points $\vec{z}_{jk}$, we have:

$$SS_{Tls} = KV_{ls} = K\mathrm{tr}(\Sigma_{\mathrm{Z}}) = Ku_1^t \Sigma u_1$$

Defining by $\sum_{Zj}$ the J sample covariance matrices of the projected points on the different groups, the within tissue variability of the projected points $\mathrm{SS}_{\mathrm{Wls}}$ is:

$$SS_{Wls} = \sum_{j=1}^{J} K_j \mathrm{tr}(\Sigma_{\mathrm{Zj}}) = \sum_{j=1}^{J} K_j u_1^t \Sigma_j u_1$$

The between tissues variability of the projected points $\mathrm{SS}_{\mathrm{Bls}}$ can be computed by any of the following expressions:

$$SS_{Bls} = K\mathrm{tr}(\Sigma_{\mathrm{Z}}) - \sum_{j=1}^{J} K_j \mathrm{tr}(\Sigma_{\mathrm{Zj}}) = Ku_1^t \Sigma u_1 - \sum_{j=1}^{J} K_j u_1^t \Sigma_j u_1 = \sum_{i=1}^{I}\sum_{j=1}^{J} K_j \left(\bar{z}_{ij\cdot} - \bar{z}_{i\cdot\cdot}\right)^2$$

Finally, the ratio of between variability explained by the constant splicing ratio model can be obtained, for instance, by:

$$\frac{SS_{Bls}}{SS_B} = \frac{K\mathrm{tr}(\Sigma_{\mathrm{Z}}) - \sum_{j=1}^{J} K_j \mathrm{tr}(\Sigma_{\mathrm{Zj}})}{K\mathrm{tr}(\Sigma) - \sum_{j=1}^{J} K_j \mathrm{tr}(\Sigma_{\mathrm{j}})}$$

A brief summary of the relation between the different sums of squares is:

$$SS_T = SS_B + SS_W$$

$$SS_B = SS_{Bls} + SS_{Wls}$$

$$\frac{SS_{Bls}}{SS_T} \le \frac{SS_B}{SS_T}$$

The contribution of gene expression to variation in isoform abundance within each tissue can be found in Figure S25A and the contribution of gene expression to the between-individual and between-tissue variation in isoform abundance in figure 3E.

We investigated the reproducibility of our results with Cufflinks, a different transcript abundance quantifier (*69*). Cufflinks quantifications were available for a subset of 876 samples, of which we could use 133, corresponding to 19 individuals with samples across the same 7 tissues. This design approaches as much as possible the one used on Flux Capacitor's quantifications (38 individuals across 10 tissues) in terms of tissues and individuals numbers. We then performed the same analysis on the same genes and compared the distribution of the contribution of gene expression to the between-individual and between-tissue variation in isoform abundance. The results are remarkably consistent (Fig 25B) even though samples could not be fully matched. The average contribution of gene expression is 0.82 and 0.54 for the between-tissue and between-individual variation using Cufflinks against 0.84 and 0.45 for Flux Capacitor analysis. These results support the robustness of our conclusions.

5.5 Major isoform switch detection across tissues

Differences in transcript usage across tissues might be subtle or they could imply a change of major isoform when the most expressed isoform from a gene changes from tissue to tissue. A transcript was identified as the major isoform if it was the most expressed transcript consistently in at least 80% of the tissue samples. Genes with different major isoforms in two tissues were retrieved. We compared their coding sequences using the coding sequence annotation from Gencode v12 (*9*). Indeed, some transcripts may differ only on their UTRs, which should not impact the protein produced ultimately (Table S19).

Supplementary figures

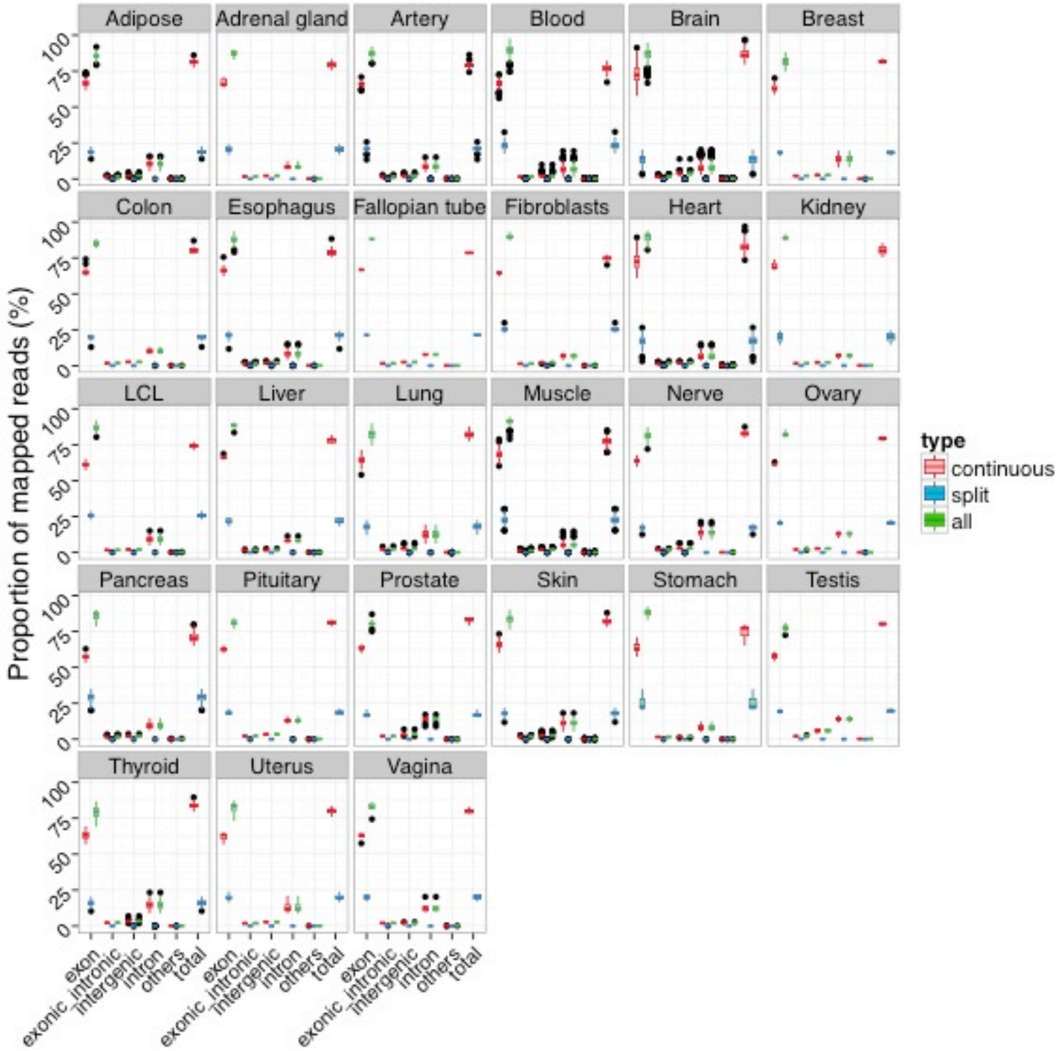


Figure S1. Proportion of mapped reads across the genome. Proportion of different types of mapped reads (split, continuous or both) in different genomic domains (exonic, intronic, exonic-intronic, intergenic and other categories) for each RNA-seq sample. Primary alignments of each read were classified into exonic or intronic when they were fully included in exon or introns respectively, exonic-intronic when they overlapped both, intergenic when they were fully included in intergenic regions and other when they were at the boundary between exons and intergenic regions. Then, we calculated the proportion of reads that fell in each category, considering all reads, those reads that had been split-mapped (different portions of the read mapped in different genomic locations), or those that were contiguously mapped (not split reads).

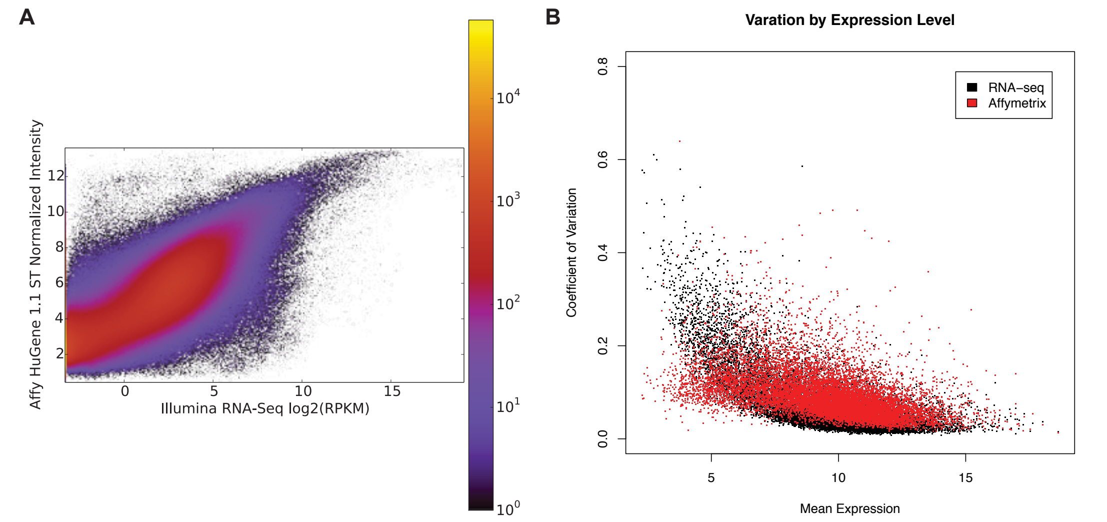


Figure S2. Comparison of Microarray and RNA-seq. A. Density plot showing the correlation between gene expression measured using microarrays (Affymetrix) and RNA-seq (Illumina). **B.** Coefficient of variation as a function of mean gene expression (log2 RPKM).

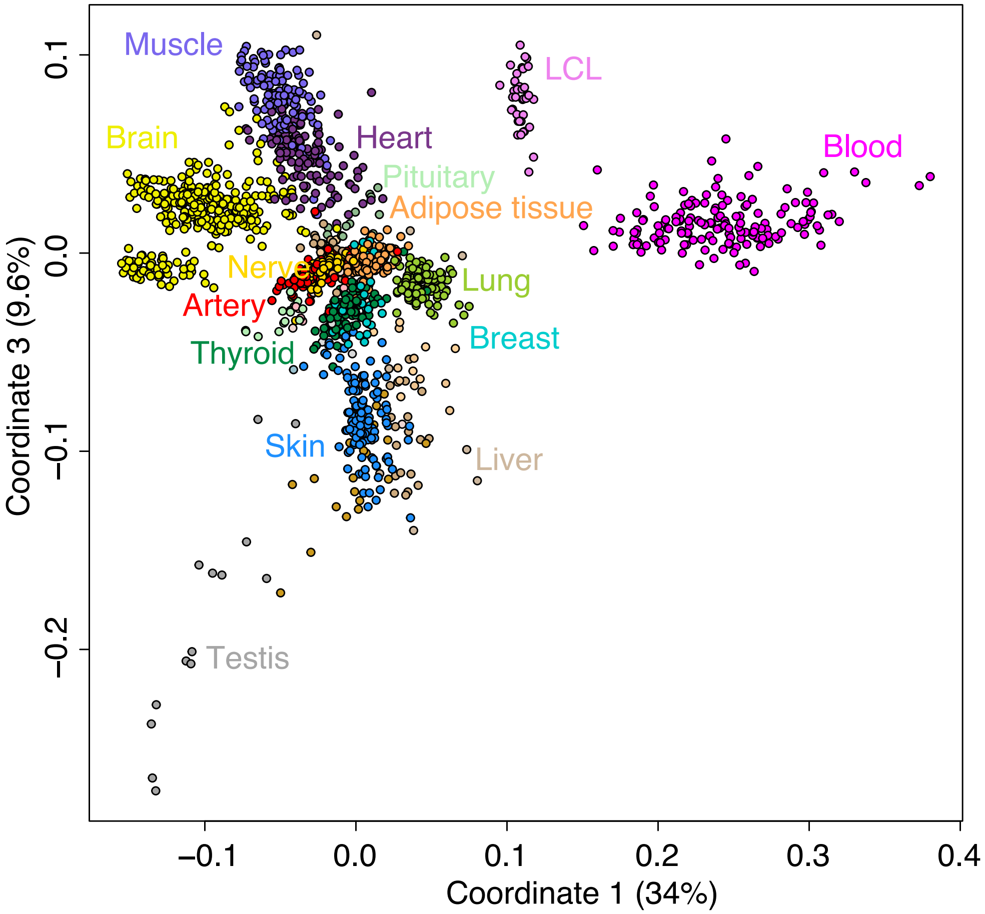


Figure S3. Multi dimensional scaling (MDS) of GTEx samples based on gene expression. First and third principal components (PC) of MDS shown here. The first PC separates well solid and non-solid tissues (LCL and Blood). The third PC separates brain sub-tissues and also clusters muscle and heart separating them from testis, in contrast to what is observed in PC2.

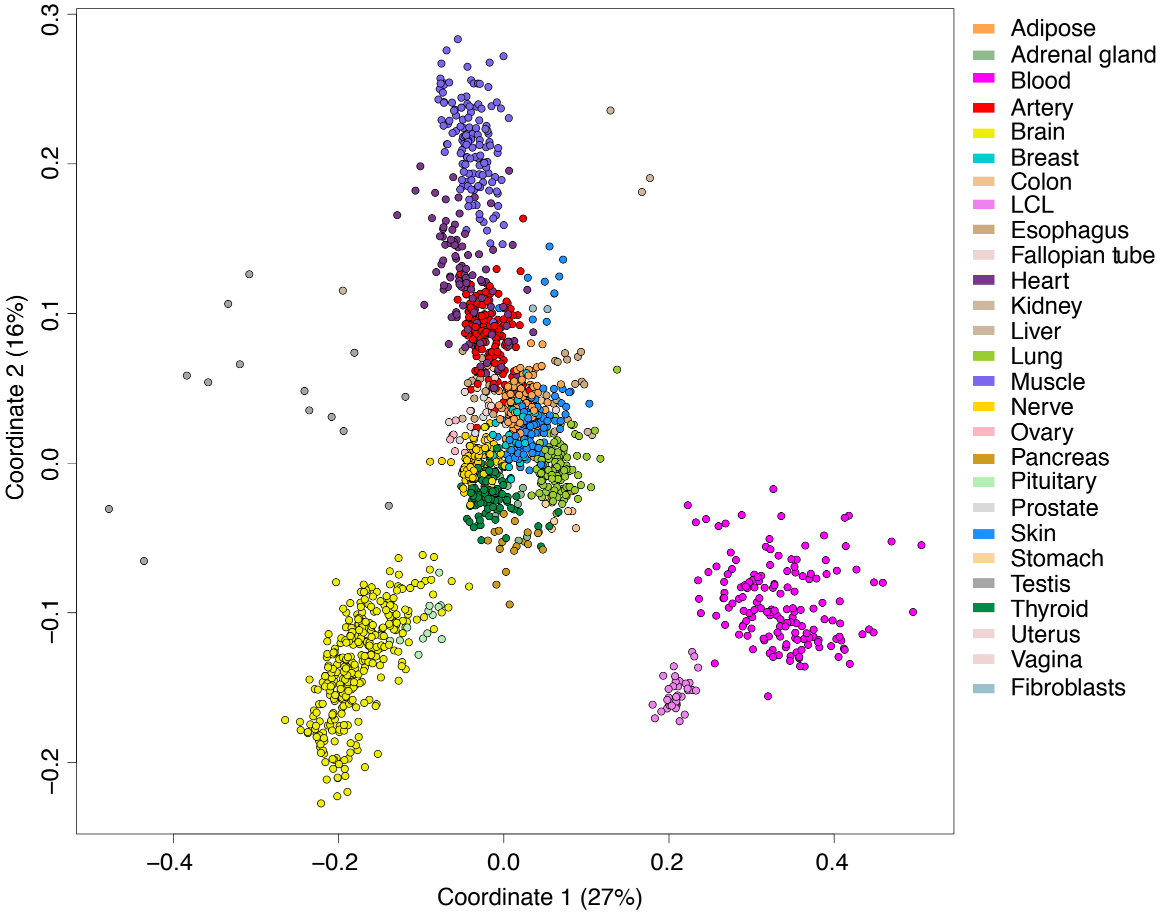


Figure S4. Multi dimensional scaling (MDS) of GTEx samples based on lncRNA expression. Multi-dimensional scaling of all samples based on expression levels of lncRNAs (log2 RPKM transformed, distance = 1 – Pearson correlation). As when using all genes, expression of lncRNAs alone also recapitulates tissue type.

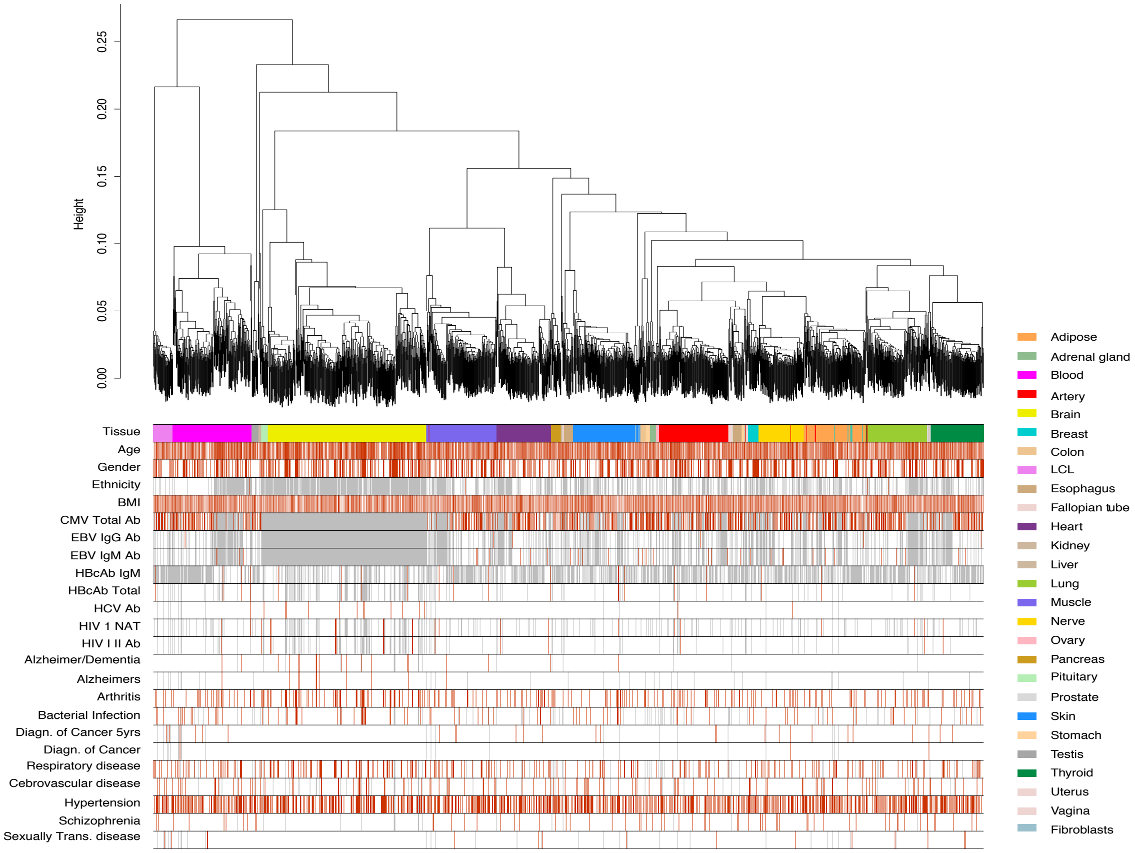


Figure S5. Hierarchical clustering based on gene expression of GTEx samples (annotated with clinical variables). Hierarchical clustering of all the samples based on gene expression of all genes (log2 RPKM transformed, distance= 1 – Pearson correlation). For each sample we show the respective tissue color and the values in some of the clinical traits associated to the samples. Each clinical trait was normalized to [0,1] range, where gray represents unknown values, and the remaining values are represented in a white (lower values) to red (higher values) gradient scale.

Scales and meaning of the traits: Continuous values: Age, BMI (range:18.46-35.87). Gender (1=Male, 2=Female). Ethnicity (0=Not Hispanic or Latino, 1=Hispanic or Latino). Binary Variables (0=Negative=white;1=Positive=red): 'CMV Total Ab' = Test performed to determine the presence of a genus of the family herpesviridae;'EBV IgG Ab' = Test performed to determine the amount of immunoglobulin G present in a sample;'EBV IgM Ab' = Test performed to determine the amount of immunoglobulin M present in a sample;'HBcAb IgM' = Test performed to determine the amount of Hepatitis B virus core antibody and the amount of immunoglobulin M present in a sample;'HBcAb Total' = Test performed to determine the amount of Hepatitis B virus core antibody present in a sample. 'HCV Ab' = Test performed to determine the presence of a small, enveloped, positive sense single strand RNA virus in the family Flaviviridae and protein made by B lymphocytes in response to a foreign substance (antigen);'HIV 1 NAT' = Any of various amplification and detection strategies applied to detection of virus contamination in donor blood, specifically the virus isolated and recognized as the etiologic agent of AIDS. 'HIV I II Ab' = Test performed using HIV antibodies produced by B-cells, human immunodeficiency virus (HIV) antibodies that react with HIV antigens. Other reported conditions: 'Alzheimers','Arthritis','Hypertension','Schizophrenia','Alzheimer/Dementia','Bacterial Infection','Diagn. of Cancer 5yrs' = Cancer diagnosis within the preceding 5 years,'Respiratory disease','Diagn. of Cancer','Cebrovascular disease','Sexually Trans. disease'.

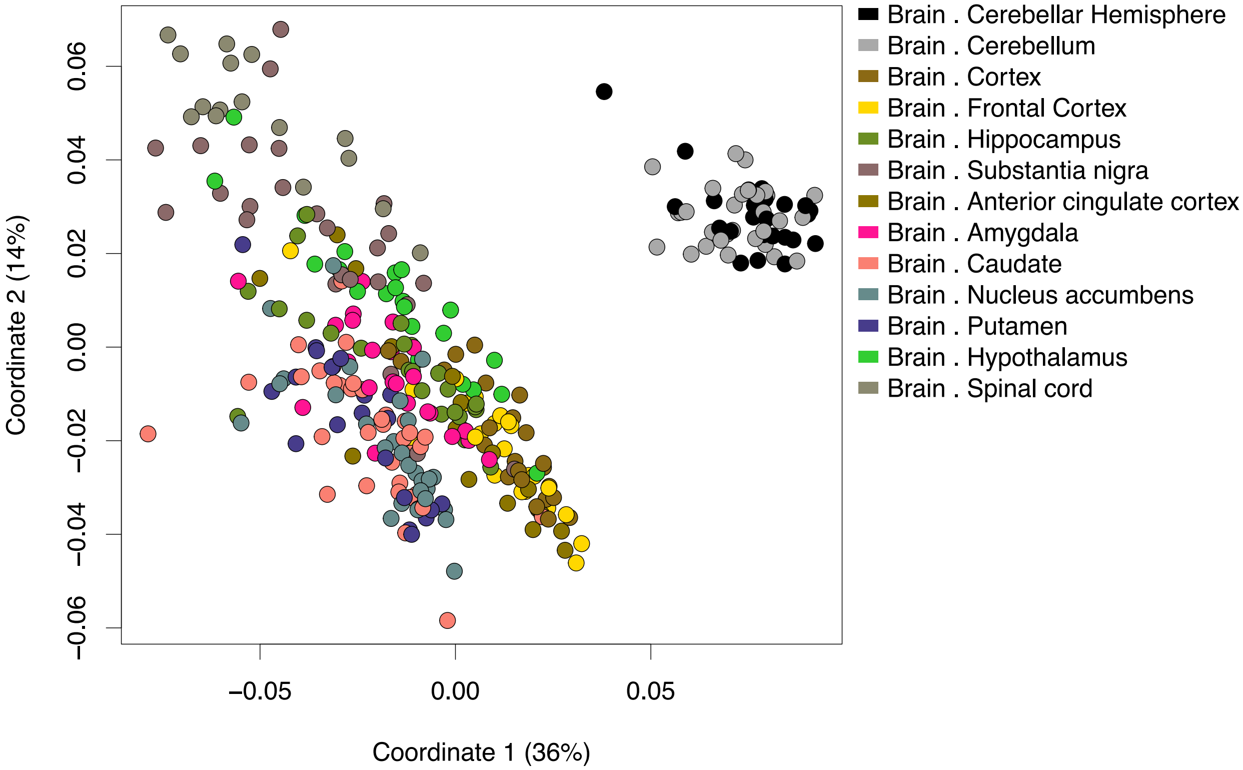


Fig. S6. Multi dimensional scaling (MDS) of GTEx brain samples based on gene expression. Multidimensional Scaling for brain samples based on gene expression of all genes (log2 transformed RPKM, distance = 1 – Pearson correlation). Cerebellum and cerebellar samples are clearly separated from the rest of the brain sub-tissues.

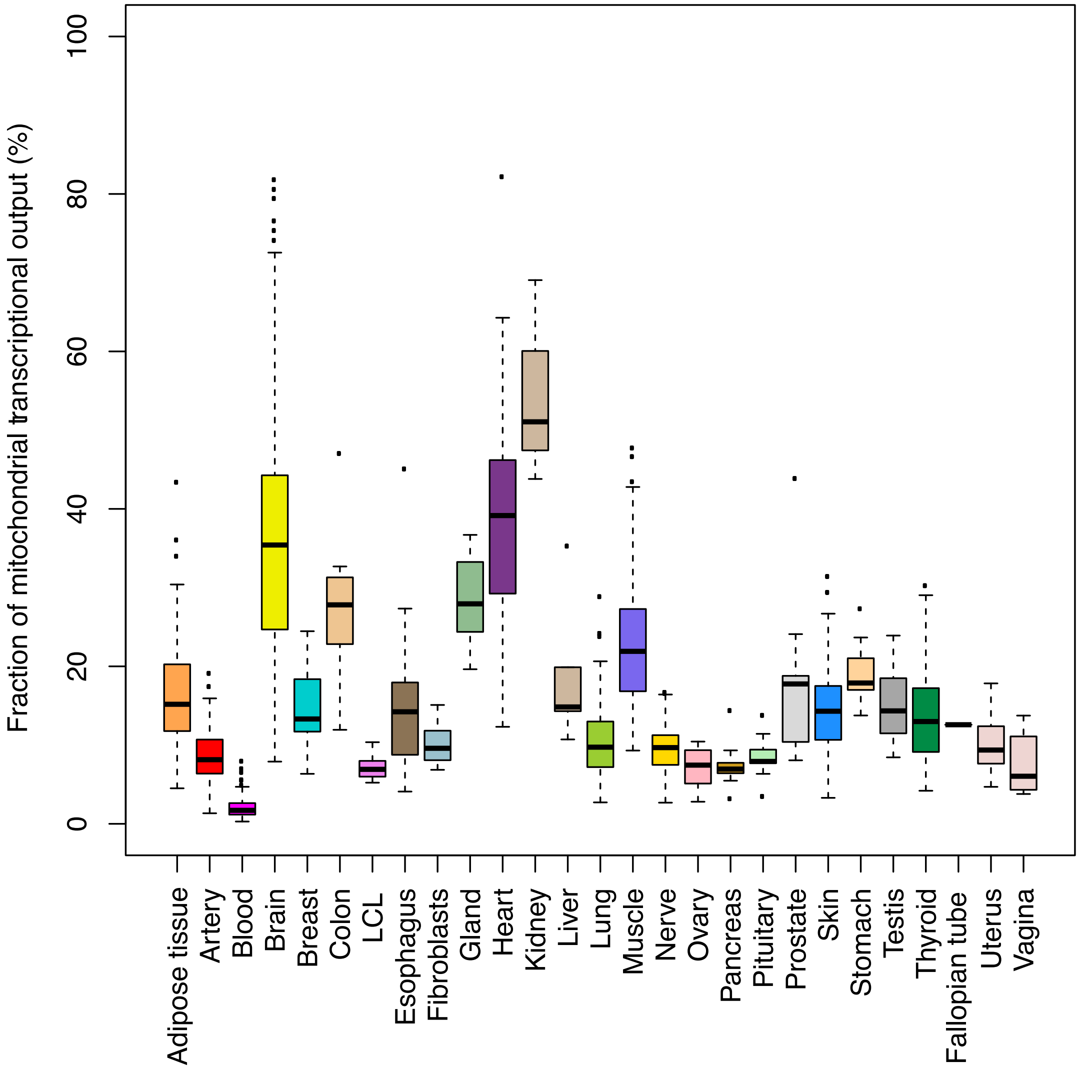
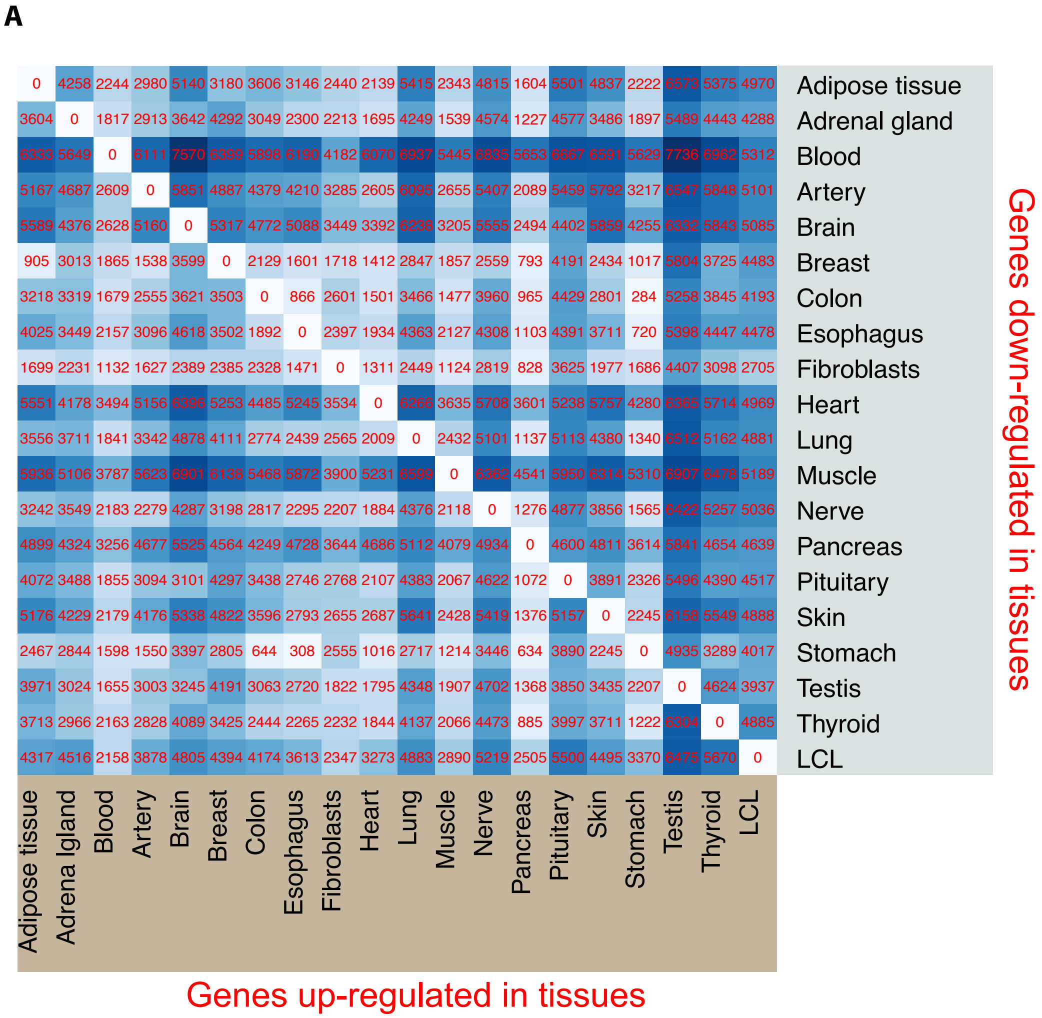
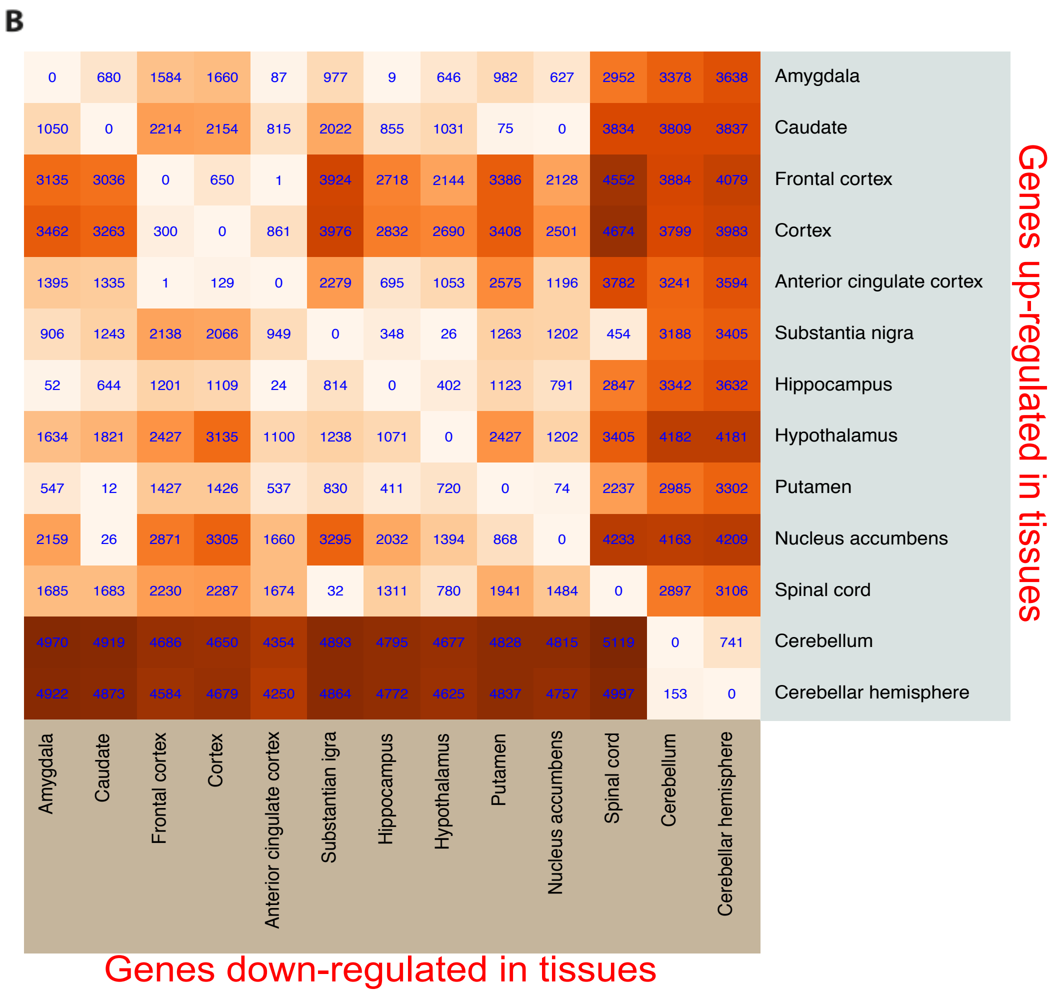


Figure S7. Fraction of the total transcriptional output originating from mitochondrial genes. The fraction is measured in each sample by the proportion of the sum over all RPKM values of genes encoded by the mitochondrial genome. Kidney has the by far highest mitochondrial activity, but also brain, heart and skeletal muscle show elevated levels of mitochondrial genes.



	Adipose tissue	Adrena lgland	Blood	Artery	Brain	Breast	Colon	Esophagus	Fibroblasts	Heart	Lung	Muscle	Nerve	Pancreas	Pituitary	Skin	Stomach	Testis	Thyroid	LCL
Adipose tissue	0	4258	2244	2980	5140	3180	3606	3146	2440	2139	5415	2343	4815	1604	5501	4837	2222	6573	5375	4970
Adrenal gland	3604	0	1817	2913	3642	4292	3049	2300	2213	1695	4249	1539	4574	1227	4577	3486	1897	5489	4443	4288
Blood	6333	5649	0	6111	7570	6399	5898	6190	4182	6070	6937	5445	6835	5653	6667	6591	5629	7736	6962	5312
Artery	5167	4687	2609	0	5851	4887	4379	4210	3285	2605	6095	2655	5407	2089	5459	5792	3217	6547	5848	5101
Brain	5589	4376	2628	5160	0	5317	4772	5088	3449	3392	6238	3205	5555	2494	4402	5859	4255	6332	5843	5085
Breast	905	3013	1865	1538	3599	0	2129	1601	1718	1412	2847	1857	2559	793	4191	2434	1017	5804	3725	4483
Colon	3218	3319	1679	2555	3621	3503	0	866	2601	1501	3466	1477	3960	965	4429	2801	284	5258	3845	4193
Esophagus	4025	3449	2157	3096	4618	3502	1892	0	2397	1934	4363	2127	4308	1103	4391	3711	720	5398	4447	4478
Fibroblasts	1699	2231	1132	1627	2389	2385	2328	1471	0	1311	2449	1124	2819	828	3625	1977	1686	4407	3098	2705
Heart	5551	4178	3494	5156	6396	5253	4485	5245	3534	0	6266	3635	5708	3601	5238	5757	4280	6365	5714	4969
Lung	3556	3711	1841	3342	4878	4111	2774	2439	2565	2009	0	2432	5101	1137	5113	4380	1340	6512	5162	4881
Muscle	5936	5106	3787	5623	6901	6138	5468	5872	3900	5231	6599	0	6362	4541	5950	6314	5310	6907	6478	5189
Nerve	3242	3549	2183	2279	4287	3198	2817	2295	2207	1884	4376	2118	0	1276	4877	3856	1565	6422	5257	5036
Pancreas	4899	4324	3256	4677	5525	4564	4249	4728	3644	4686	5112	4079	4934	0	4600	4811	3614	5841	4654	4639
Pituitary	4072	3488	1855	3094	3101	4297	3438	2746	2768	2107	4383	2067	4622	1072	0	3891	2326	5496	4390	4517
Skin	5176	4229	2179	4176	5338	4822	3596	2793	2655	2687	5641	2428	5419	1376	5157	0	2245	6158	5549	4888
Stomach	2467	2844	1598	1550	3397	2805	644	308	2555	1016	2717	1214	3446	634	3890	2245	0	4935	3289	4017
Testis	3971	3024	1655	3003	3245	4191	3063	2720	1822	1795	4348	1907	4702	1368	3850	3435	2207	0	4624	3937
Thyroid	3713	2966	2163	2828	4089	3425	2444	2265	2232	1844	4137	2066	4473	885	3997	3711	1222	6304	0	4885
LCL	4317	4516	2158	3878	4805	4394	4174	3613	2347	3273	4883	2890	5219	2505	5500	4495	3370	6475	5670	0



	Amygdala	Caudate	Frontal cortex	Cortex	Anterior cingulate cortex	Substantian igra	Hippocampus	Hypothalamus	Putamen	Nucleus accumbens	Spinal cord	Cerebellum	Cerebellar hemisphere
Amygdala	0	680	1584	1660	87	977	9	646	982	627	2952	3378	3638
Caudate	1050	0	2214	2154	815	2022	855	1031	75	0	3834	3809	3837
Frontal cortex	3135	3036	0	650	1	3924	2718	2144	3386	2128	4552	3884	4079
Cortex	3462	3263	300	0	861	3976	2832	2690	3408	2501	4674	3799	3983
Anterior cingulate cortex	1395	1335	1	129	0	2279	695	1053	2575	1196	3782	3241	3594
Substantia nigra	906	1243	2138	2066	949	0	348	26	1263	1202	454	3188	3405
Hippocampus	52	644	1201	1109	24	814	0	402	1123	791	2847	3342	3632
Hypothalamus	1634	1821	2427	3135	1100	1238	1071	0	2427	1202	3405	4182	4181
Putamen	547	12	1427	1426	537	830	411	720	0	74	2237	2985	3302
Nucleus accumbens	2159	26	2871	3305	1660	3295	2032	1394	868	0	4233	4163	4209
Spinal cord	1685	1683	2230	2287	1674	32	1311	780	1941	1484	0	2897	3106
Cerebellum	4970	4919	4686	4650	4354	4893	4795	4677	4828	4815	5119	0	741
Cerebellar hemisphere	4922	4873	4584	4679	4250	4864	4772	4625	4837	4757	4997	153	0

Figure S8. Pairwise differential expression of protein coding genes between tissues. Genes are called differentially expressed if they are called by two different methods NOISeq (q > 0.95, corresponding to FDR < 0.05) and by DESeq2 (FDR < 0.05). **A.** Pairwise differential gene expression for the 20 main tissues with 10 or more samples. Testis is the sample with the largest number of up-regulated genes and blood with largest down-regulated genes. The average number of differentially expressed genes across all pairs is 3918. **B.** Pairwise differential gene expression for brain sub-tissues. Cerebellum and Cerebellar hemisphere have the most differentiated behavior. The average number of differentially expressed genes across all pairs is 2322.

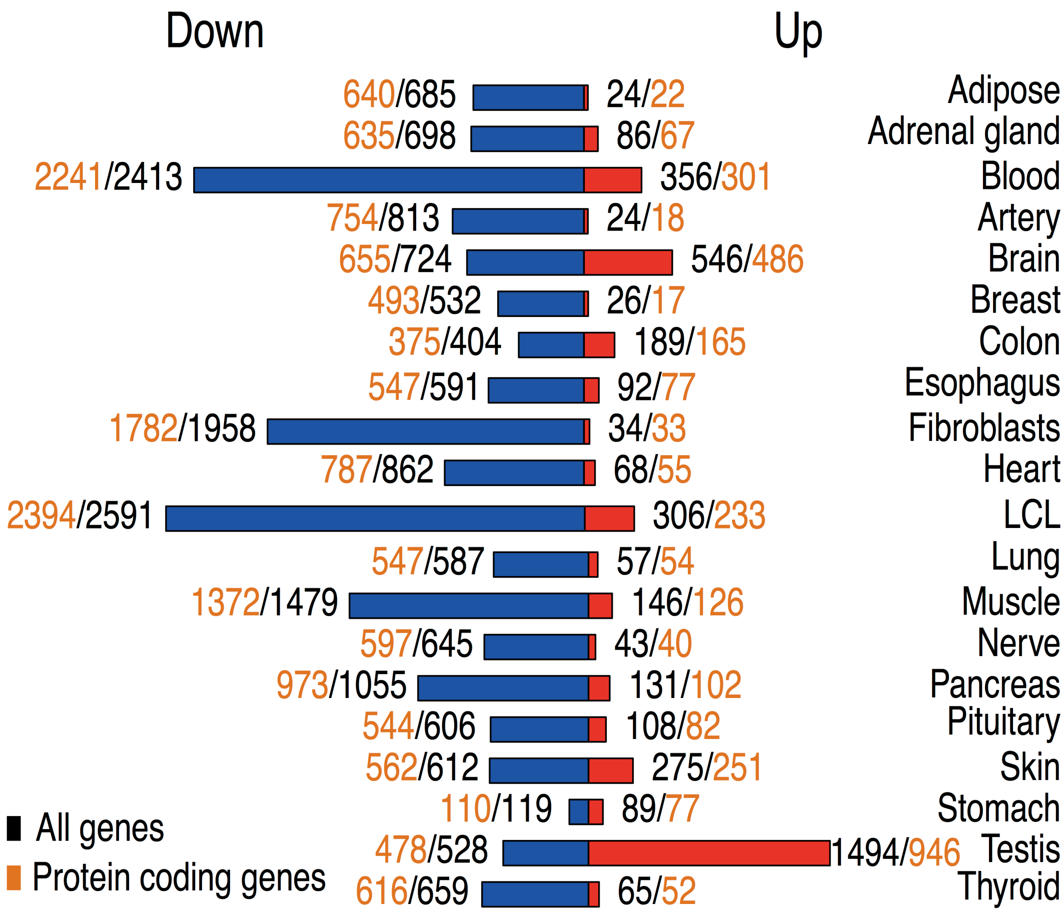


Figure S9. Genes with tissue preferential expression. Number of down and up-regulated genes when comparing expression in the samples of a given tissue to the samples of the remaining tissues. Only tissues with ten or more samples were considered. The analysis was performed with NOISeq and genes were considered tissue specific if q>=0.99 (FDR=1%) and log2 fold change was greater than 4.

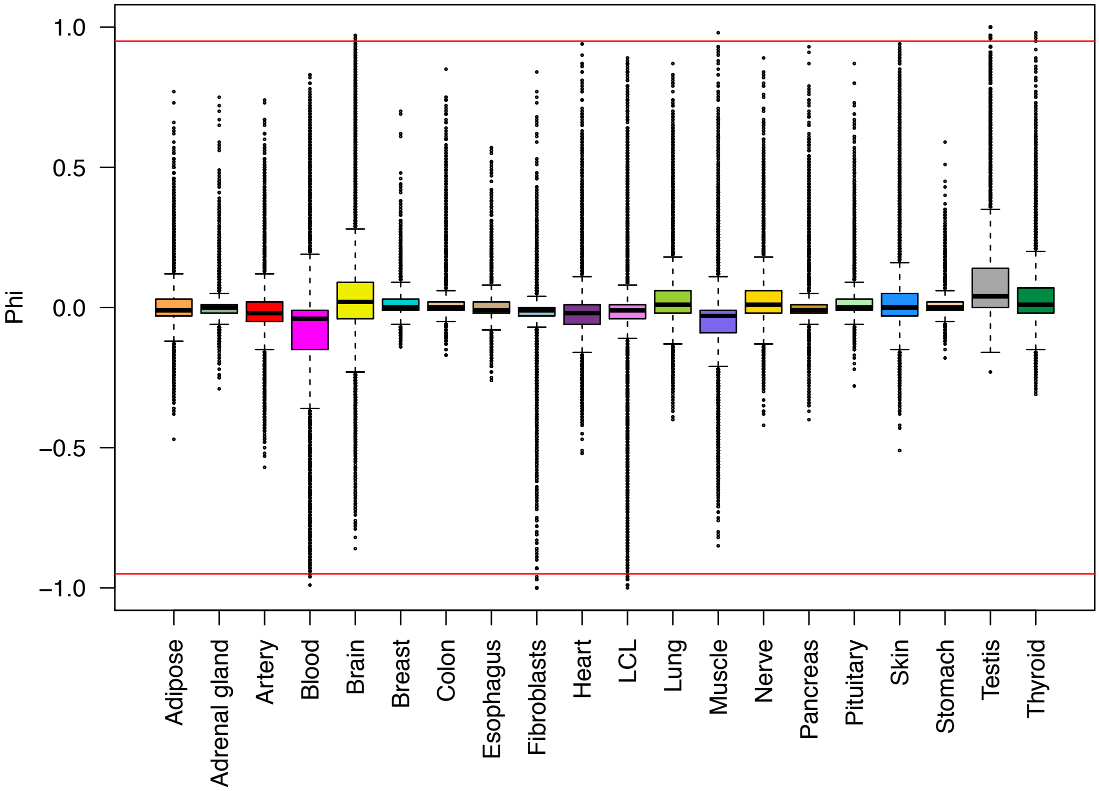


Figure S10. Genes with tissue exclusive expression. The distribution of the phi correlation is given for all genes within each tissue. For each pair (gene, tissue) the phi statistic is computed from a 2x2 contingency table that includes the number of samples from the tissue in which the gene is expressed (RPKM>0.1) and not expressed (RPKM<0.1), and the number of samples from the rest of the tissues in which the gene is expressed and not expressed. Vertical red lines indicated an absolute value of phi equal to 0.95. Values of phi close to 1 indicate that the gene is expressed in (nearly) all samples from the tissue, and (nearly) no samples from the rest of the tissues. Values of phi close to -1 indicate that the gene is not expressed in (nearly) all samples from the tissue and it is expressed in (nearly) all samples form the rest of the tissues. Only a few genes have phi values > 0.95 or < -0.95.

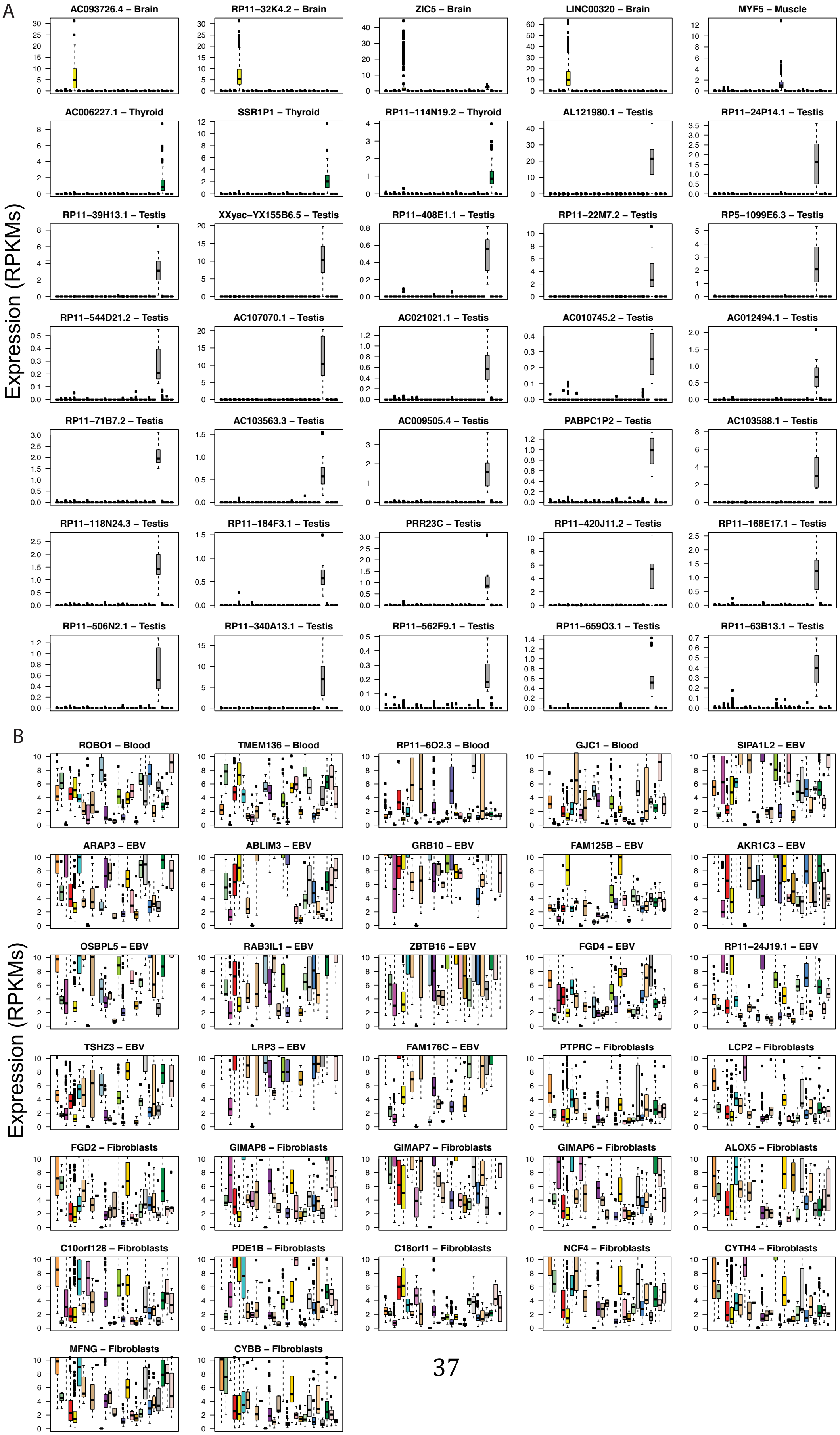

A
Expression (RPKMs)
AC093726.4 – Brain
RP11–32K4.2 – Brain
ZIC5 – Brain
LINC00320 – Brain
MYF5 – Muscle
AC006227.1 – Thyroid
SSR1P1 – Thyroid
RP11–114N19.2 – Thyroid
AL121980.1 – Testis
RP11–24P14.1 – Testis
RP11–39H13.1 – Testis
XXyac–YX155B6.5 – Testis
RP11–408E1.1 – Testis
RP11–22M7.2 – Testis
RP5–1099E6.3 – Testis
RP11–544D21.2 – Testis
AC107070.1 – Testis
AC021021.1 – Testis
AC010745.2 – Testis
AC012494.1 – Testis
RP11–71B7.2 – Testis
AC103563.3 – Testis
AC009505.4 – Testis
PABPC1P2 – Testis
AC103588.1 – Testis
RP11–118N24.3 – Testis
RP11–184F3.1 – Testis
PRR23C – Testis
RP11–420J11.2 – Testis
RP11–168E17.1 – Testis
RP11–506N2.1 – Testis
RP11–340A13.1 – Testis
RP11–562F9.1 – Testis
RP11–659O3.1 – Testis
RP11–63B13.1 – Testis
B
Expression (RPKMs)
ROBO1 – Blood
TMEM136 – Blood
RP11–6O2.3 – Blood
GJC1 – Blood
SIPA1L2 – EBV
ARAP3 – EBV
ABLIM3 – EBV
GRB10 – EBV
FAM125B – EBV
AKR1C3 – EBV
OSBPL5 – EBV
RAB3IL1 – EBV
ZBTB16 – EBV
FGD4 – EBV
RP11–24J19.1 – EBV
TSHZ3 – EBV
LRP3 – EBV
FAM176C – EBV
PTPRC – Fibroblasts
LCP2 – Fibroblasts
FGD2 – Fibroblasts
GIMAP8 – Fibroblasts
GIMAP7 – Fibroblasts
GIMAP6 – Fibroblasts
ALOX5 – Fibroblasts
C10orf128 – Fibroblasts
PDE1B – Fibroblasts
C18orf1 – Fibroblasts
NCF4 – Fibroblasts
CYTH4 – Fibroblasts
MFNG – Fibroblasts
CYBB – Fibroblasts

Figure S11. Expression of genes with tissue exclusive expression. A. Expression patterns of 35 tissue specific genes (phi > 0.95) including all non-testis specific genes and some randomly selected testis specific genes. Gene name and tissue in which the gene is expressed are indicated in the title of each plot. **B.** Expression patterns of all tissue anti-specific genes (phi < -0.95). Gene name and the tissue in which the gene is not expressed are indicated in the title of each plot.

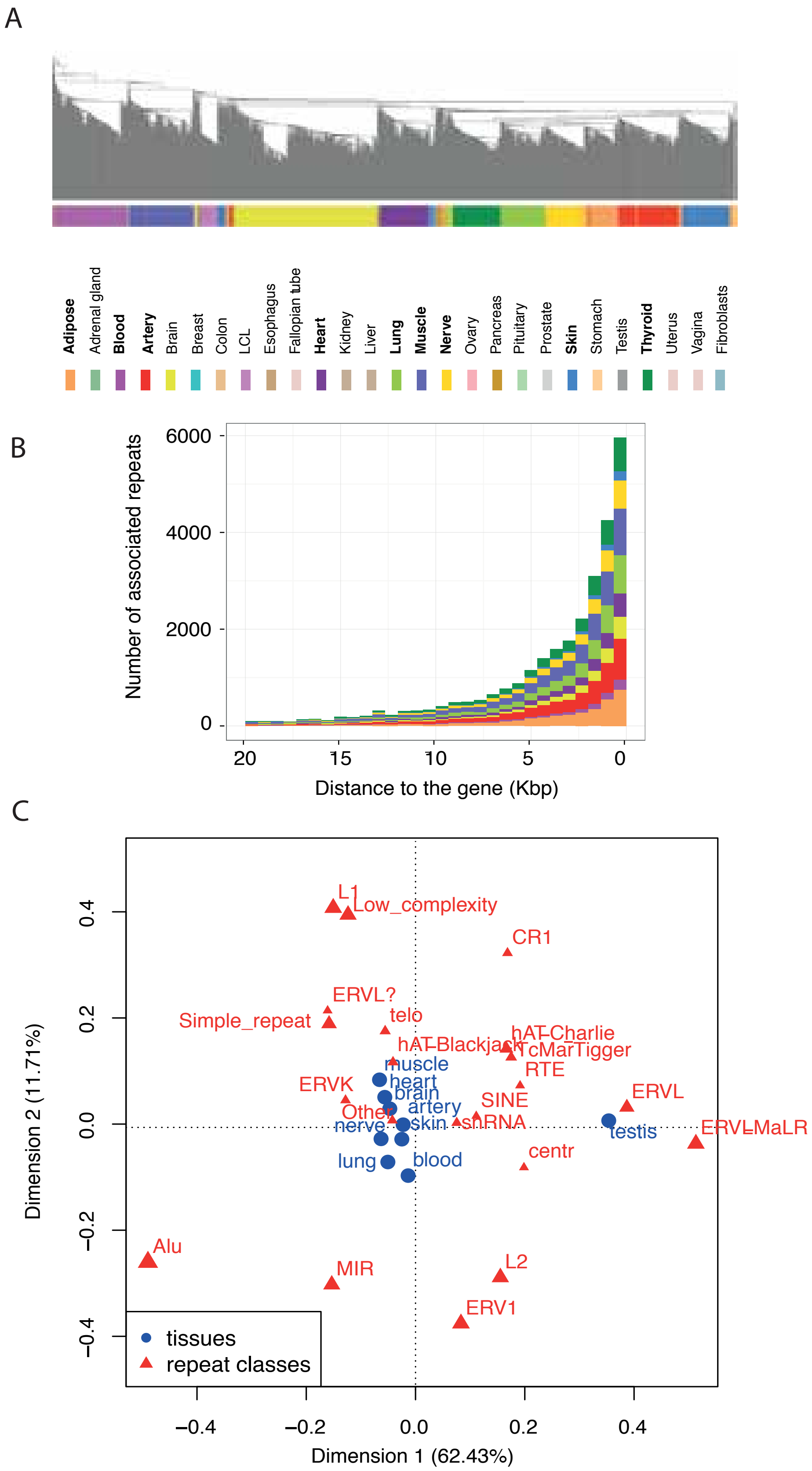

A
Adipose
Adrenal gland
Blood
Artery
Brain
Breast
Colon
LCL
Esophagus
Fallopian tube
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Prostate
Skin
Stomach
Testis
Thyroid
Uterus
Vagina
Fibroblasts
B
Number of associated repeats
6000
4000
2000
0
20
15
10
5
0
Distance to the gene (Kbp)
C
Dimension 2 (11.71%)
0.4
0.2
0.0
−0.2
−0.4
Dimension 1 (62.43%)
−0.4
−0.2
0.0
0.2
0.4
L1
Low_complexity
CR1
ERVL?
Simple_repeat
telo
hAT-Charlie
hAT-Blackjack
TcMarTigger
muscle
RTE
heart
ERVK
brain
ERVL
artery
SINE
Other
nerve
skin
snRNA
testis
ERVL-MaLR
centr
lung
blood
Alu
MIR
L2
ERV1
tissues
repeat classes

Fig. S12. Repeat expression. A. Hierarchical clustering of all samples based on repeat **expression.** We analyzed expression patterns of 62,539 repeats. Distance between samples was defined as distance = 1 – Spearman. We used the average linkage criterion to perform the clustering. **B.** Distance between a gene and the closest upstream associated repeat. Repeats and genes are considered to be associated when they have significant correlated expression across samples within a particular tissue. The nine main tissues used in the analysis are highlighted in bold in the tissue legend. **C.** Correspondence analysis of tissues and repeat classes. The analysis was carried out using transcript expression values across tissues and information on which repeats intersected the promoter region of each of these transcripts. Correspondence analysis plots tissues and repeat classes onto two leading principal axes and the distance between repeat classes and tissues in the chart represents the strength of the correlation between having a repeat class in the promoter and being expressed in that tissue. The plot suggests that ERVL repeats may be driving expression of certain lncRNAs in testis.

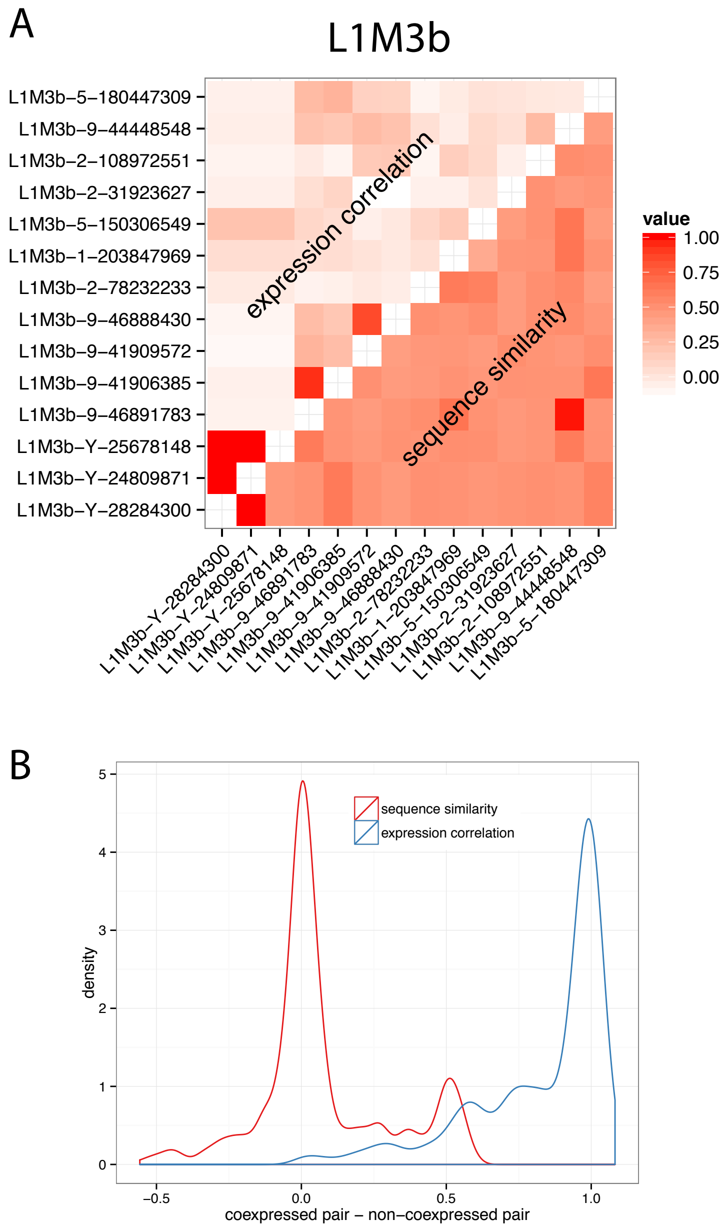


Figure S13. A. Pairwise repeat expression correlation (top left) and pairwise sequence similarity between all intergenic expressed repeat elements from the L1M3b family (bottom right). **B.** Distribution of the difference between sequence similarity of a pair of coexpressed repeats and a pair of non- coexpressed repeats of the same family (red) and distribution of the difference in expression correlation between a pair of coexpressed repeats and a non-coexpressed pair of repeats of the same family (blue). On average, two co-expressed repeats in a family have as similar sequence similarity as two non-coexpressed ones.

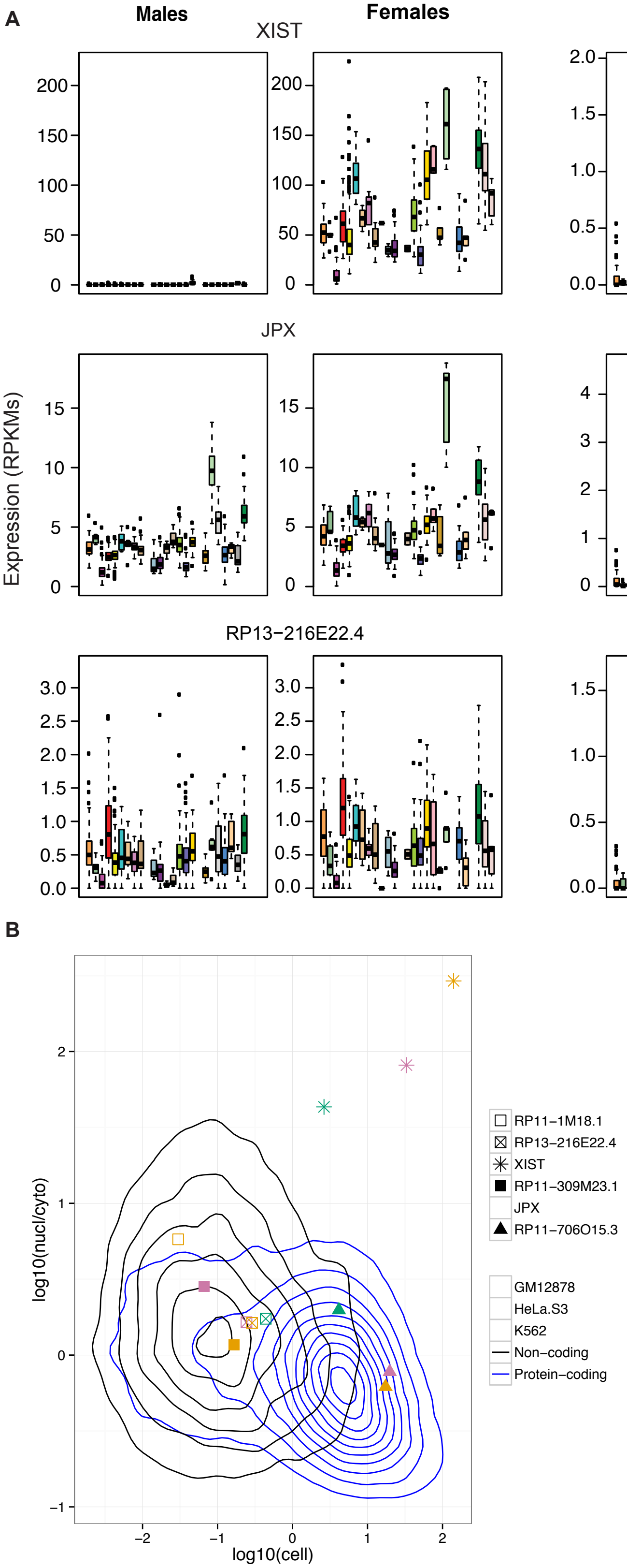

A
Males
Females
XIST
JPX
RP13−216E22.4
RP11−309M23.1
RP11−706O15.3
RP11−1M18.1
Expression (RPKMs)
B
log10(nucl/cyto)
log10(cell)
RP11−1M18.1
RP13−216E22.4
XIST
RP11−309M23.1
JPX
RP11−706O15.3
GM12878
HeLa.S3
K562
Non-coding
Protein-coding

Figure S14 Expression levels of intergenic X-chromosomal lncRNAs A. Expression levels of intergenic X-chromosomal lncRNAs upregulated in females. Intergenic X-chromosomal lncRNA ordered left to right and top to bottom from higher to lower expression differences between males and females. XIST and JPX, known to be involved in X chromosome inactivation, have the first and third largest differences respectively. The second largest is a lincRNA located in the PAR region. **B.** Nuclear vs cytosolic enrichment of these lincRNAs. Using fractionation data from the ENCODE project, we computed the ratio of nuclear vs cytosolic enrichment for each lincRNA as the ratio of the gene RPKM in the nucleus over the gene RPKM in the cytosol. Nuclear/ cytosolic ratios for each gene in those cell lines of female origin are given in the Y-axis. In the X-axis we provide the value of gene expression in whole cell. The kernel distributions correspond to the distributions of protein coding genes (black) and lncRNAs (blue), as in Figure 3 from (*17*). Most of the X chromosome lincRNAs overexpressed in females are enriched in the nucleus, a necessary property if they were involved in the process of X inactivation.

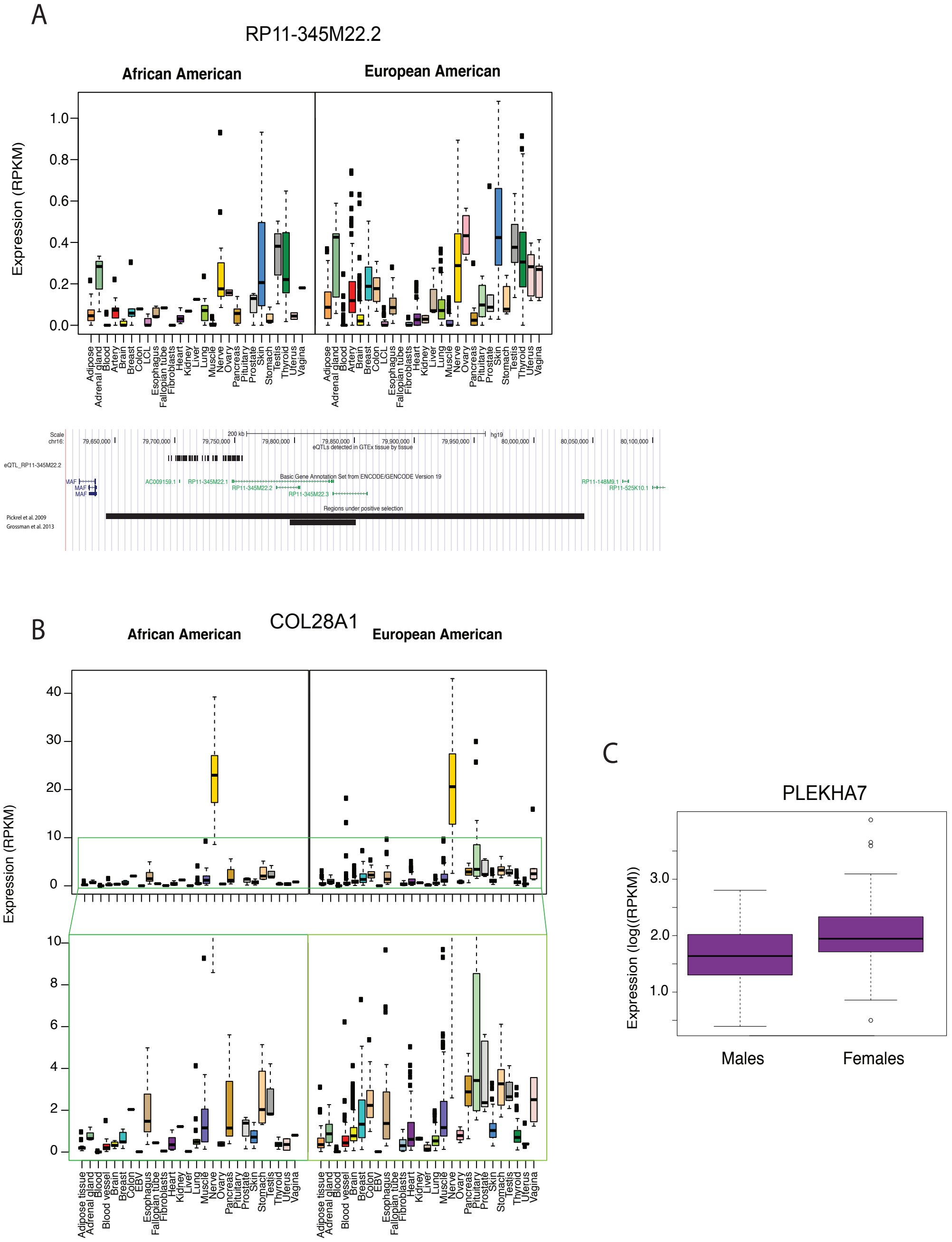
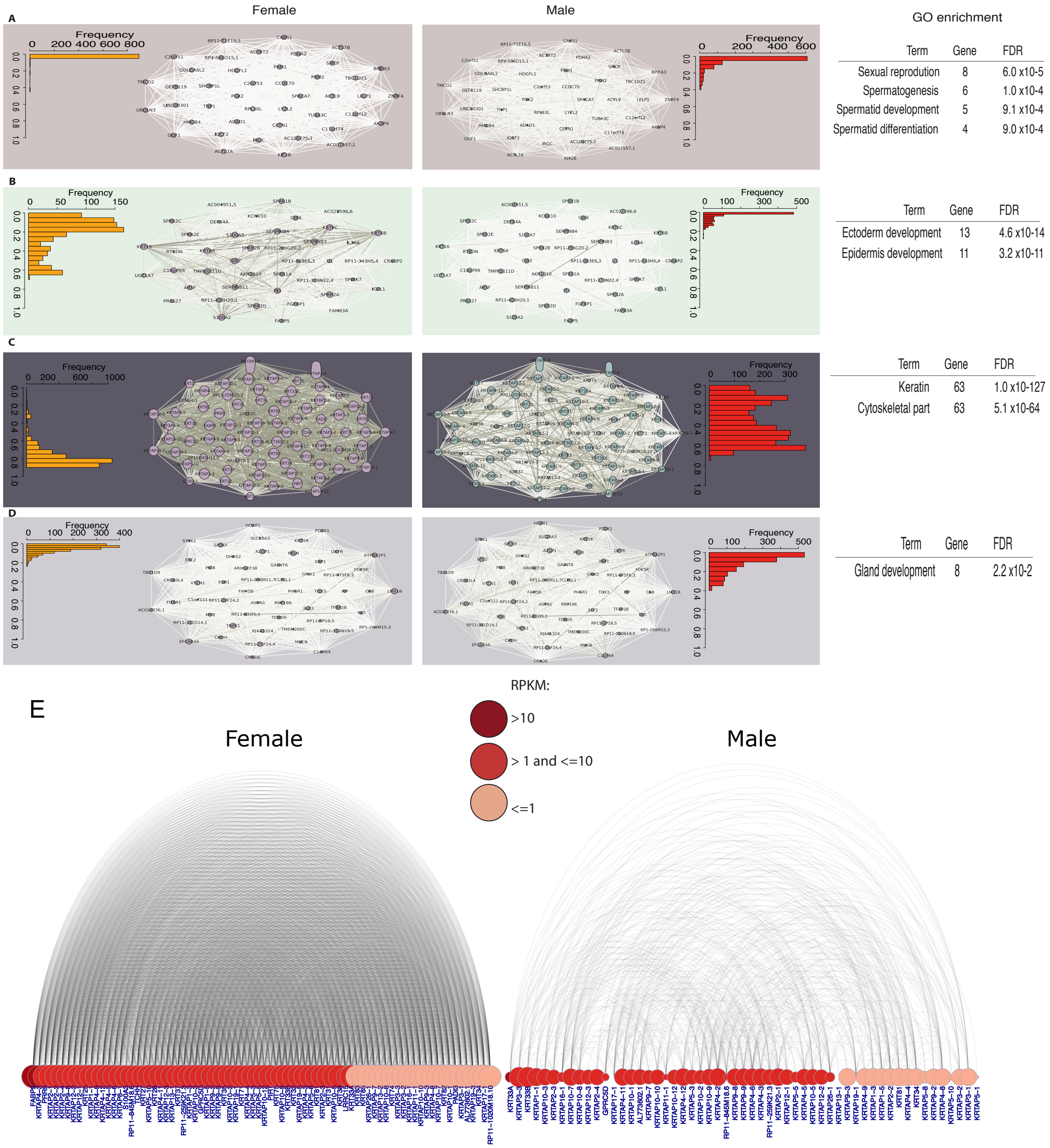


Figure S15. A. Top. Expression of the RP11-345M22.2 lincRNA in African and European Americans across tissues**.** In Europeans – the population under positive selection - this gene is most highly expressed in skin. Bottom. Genomic location of RP11-345M22.2 lincRNA showing that this region has been detected to be under positive selection in two independent studies (*53, 54*). **B.** Top. Expression of the COL28A1 in African and European Americans across tissues. This gene is mostly expressed in nerve. Bottom. Zoom in of the expression of COL28A1 across tissues. COL28A1 lies in a region under positive selection in Europeans and harbors a SNP (rs17168526) that had been linked to resistance to smallpox. **C**. Expression of PLEKHA7 in male and female heart.



Term	Gene	FDR
Sexual reproduction	8	6.0 x10-5
Spermatogenesis	6	1.0 x10-4
Spermatid development	5	9.1 x10-4
Spermatid differentiation	4	9.0 x10-4

Term	Gene	FDR
Ectoderm development	13	4.6 x10-14
Epidermis development	11	3.2 x10-11

Term	Gene	FDR
Keratin	63	1.0 x10-127
Cytoskeletal part	63	5.1 x10-64

Term	Gene	FDR
Gland development	8	2.2 x10-2

Figure S16. Sex differential expression network modules. A-D Co-expression networks were built for males and females independently. Nodes in networks correspond to genes and their size represents intramodular connectivity, i.e., connectivity of nodes to other nodes within the same module. Edges represent the topological overlap between two nodes, which measures interconnectedness between nodes. A higher topological overlap between two nodes means a node is connected to all of the neighbors of the other node. Edges are colored according to their topological overlap with darker colors corresponding to higher topological overlap. The histograms close to each module represent the distribution of topological overlap of the network edges. The same set of genes is represented in male and female modules. On the right of the figure are the GO terms that were found to be enriched among the genes in the modules.

Among male specific modules we found for instance a module related to sexual reproduction and spermatid differentiation depicted in **A**. Among female specific modules, we found one related to epidermis and ectoderm development, depicted in **B**. Differential network expression, capture therefore the important structural differences between male's and female's skin, which are not well captured by differential expression of individual genes. This is further supported by one of the two modules with conserved gene sets, but contrasting network topology. This module is enriched for genes related to keratin, hair follicle morphogenesis and development shown in **C**. Conversely in **D** we have a module with male to female topology change with a loss of interconnectedness between nodes from male to female. **E.** Alternative view of network module depicted in C (81 genes). Arcs represent network edges with correlation > 0.6 and node degrees are proportional to node size. The number of edges between sexes differs with females having a higher number of edges than males.

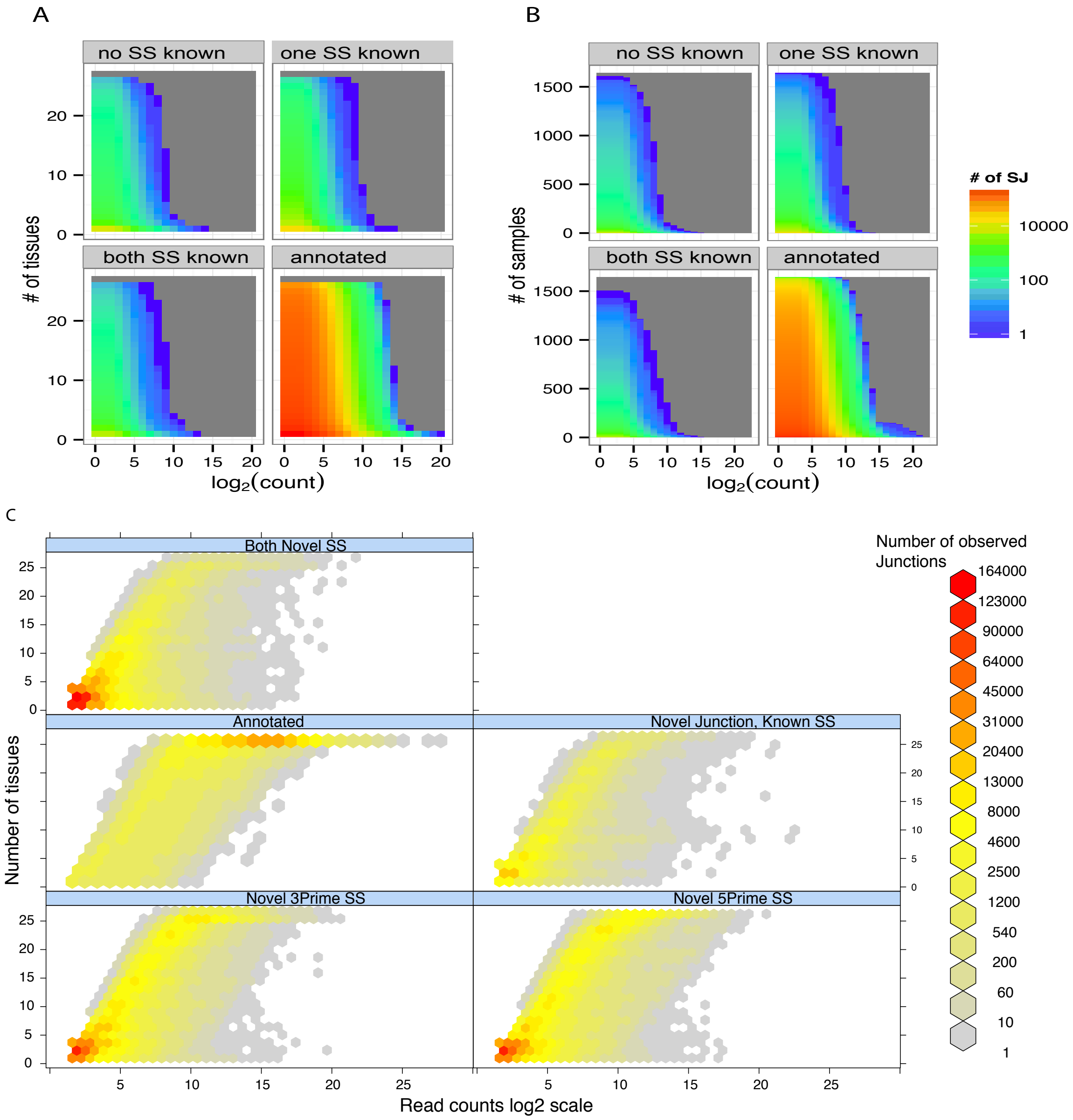


Figure S17. **Splicing across tissues.** Upper panel: High confidence splice junctions. The number of splice junctions (SJ) with log2 (average count)>=x in at least y tissues (**A**) or in at least y samples (**B**). The four categories of SJs are: no SS known (no splice site is annotated); one SS known (one splice site is annotated and the other is not), annotated (both splice sites are annotated and so is intron between them); and, both SS known (both splice sites are annotated but the intron between them is not). Lower panel: Unfiltered junctions. **C.** Distribution and support of detected splice junctions. Number of reads supporting five types of SJs: Both Novel SS (none of splice sites is annotated); 5' or 3' SS known (one splice site is annotated and the other is not), annotated (both splice sites are annotated and so is intron between them); and, Novel junction, known SS (both splice sites are annotated but the intron between them is not). The x-axis shows the number of reads (in log2 scale) that support the junction; the y-axis the number of tissues where the junction is found (pooling all samples of a given tissue). The '*both novel SS*, '*novel junction, both SS known',* and '*3prime/5prime SS known'* tend to be more tissue specific, occurring in fewer tissues and they are less frequent than annotated junctions.

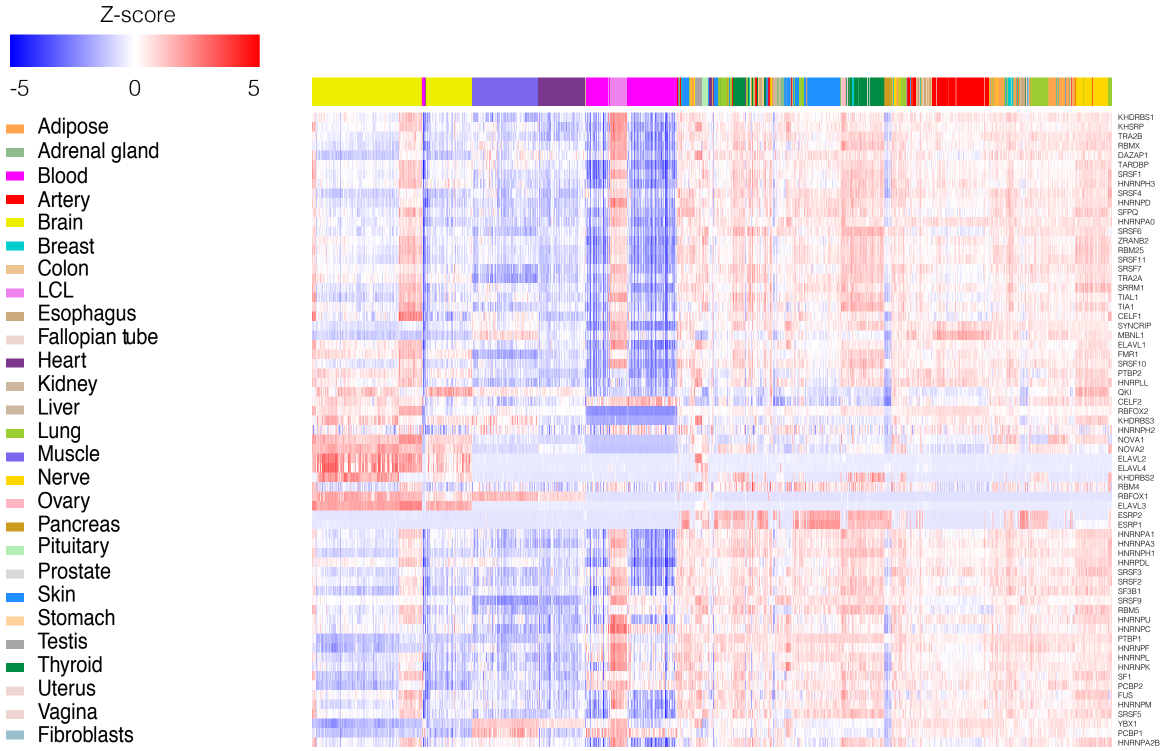


Figure S18 Relation between tissue specific splicing patterns and expression of RNA binding proteins. The heatmap shows normalized gene expression profiles (Z-score normalization by gene) of 67 RNA binding proteins (RBP) in all samples. The order of the samples, however, is determined by the hierarchical clustering solution based on the exon inclusion values for ~50,000 exons expressed in those samples (PSI, see Figure 1(*6*)). Tissues are described by the color row in the top of the matrix. In Some RBP, expression is higher in samples that have been clustered together because they share a specific splicing program. This suggests that some of these RBP may be playing an active role in generating such differential splicing patterns. Twelve out of the 67 analyzed RNA binding proteins (*63*) show tissue preferential gene expression. Seven out of these twelve are preferentially expressed in brain.

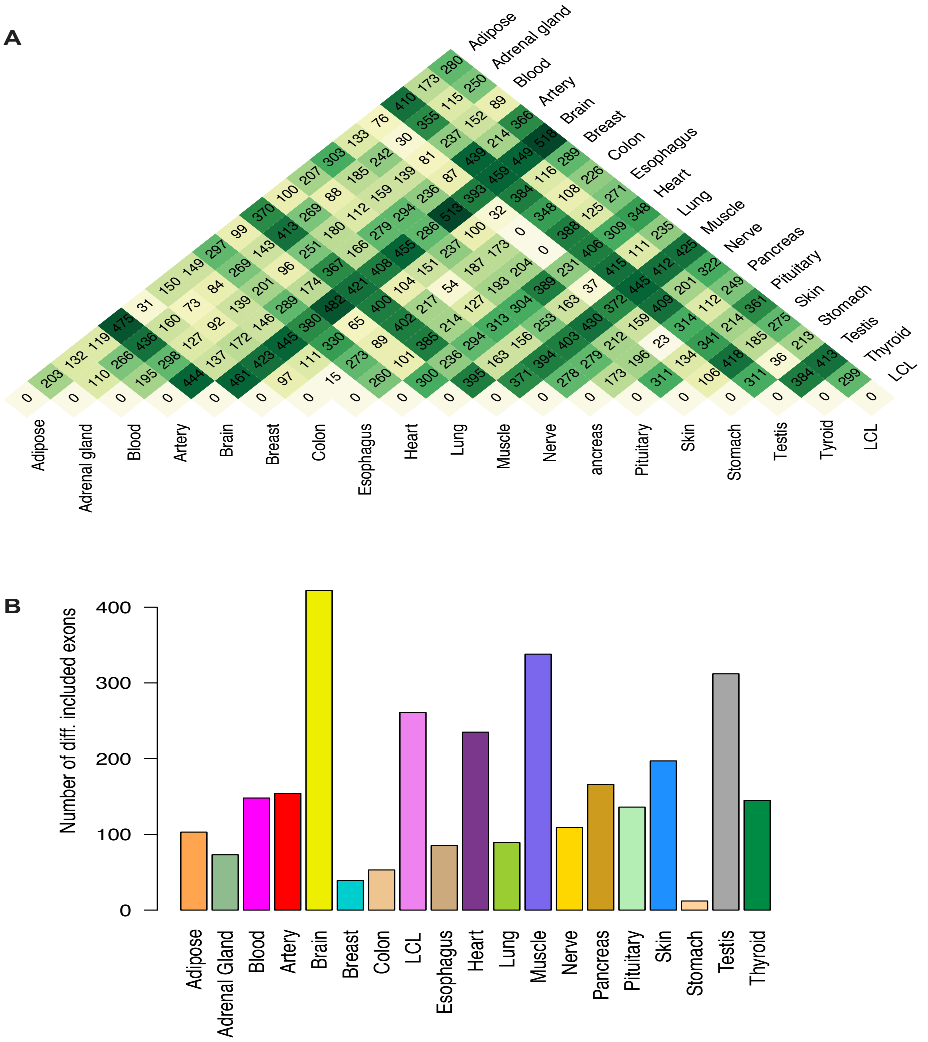


Figure S19. Tissue differential, preferential and exclusive exon inclusion

A. Pairwise differential exon inclusion between tissues. Only tissues with 10 or more samples are considered. Exons are considered to be differentially included if the absolute differences between the median PSI values in the two tissues is larger than 0.1 and the FDR is below 0.01 (adjusted non-parametric Wilcoxon test). **B**. Number of exons with tissue-specific differential inclusion for all tissues with 10 or more samples. A comparison is performed between the samples of a given tissue and the samples from the remaining tissues. Exons are considered to be differentially included if the absolute differences between the median PSI values in the two tissues is >0.1 and the FDR < 0.01 (adjusted non-parametric Wilcoxon test).

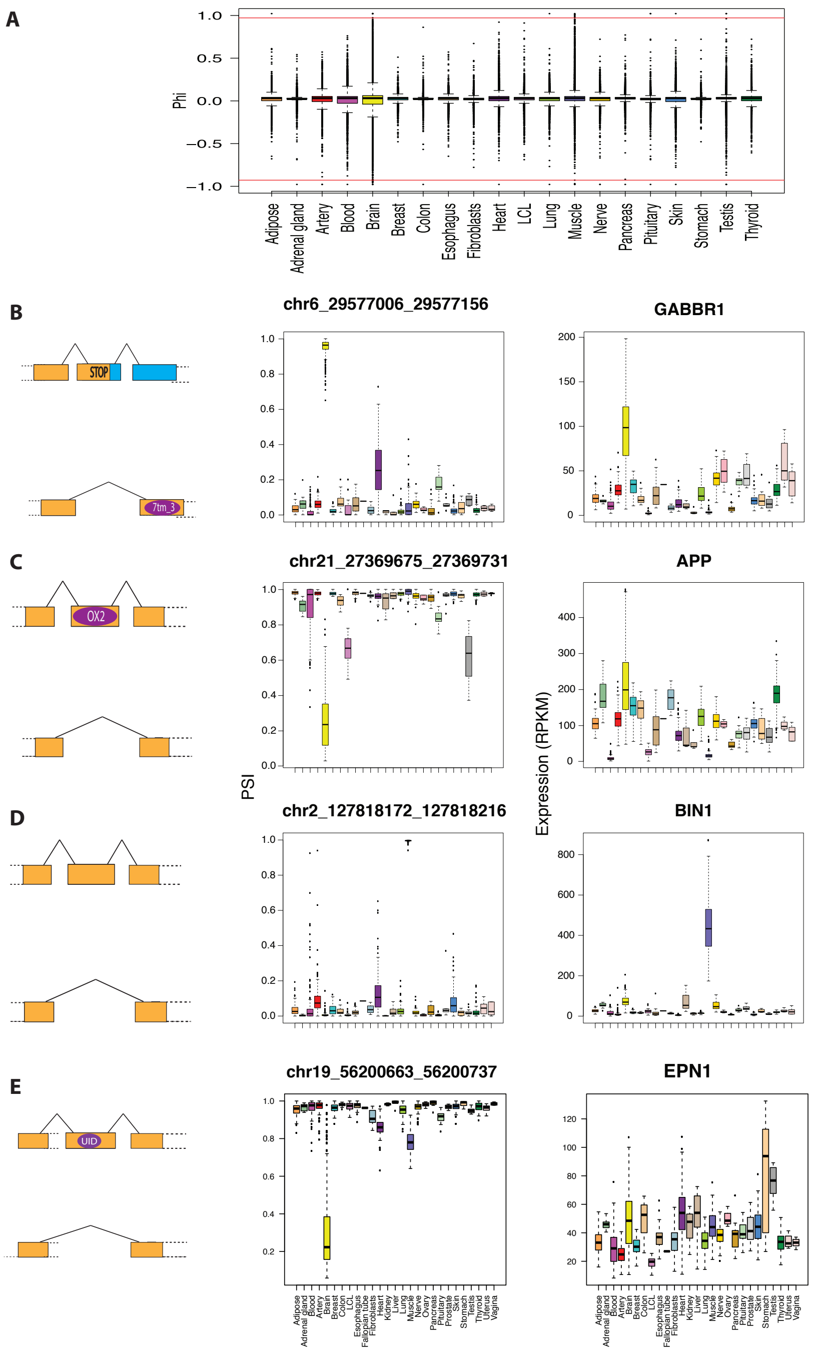


Figure S20. Tissue exclusive inclusion across tissues. A. Phi statistic distribution across tissues. For each pair (exon, tissue) the phi statistic is computed from a 2x2 contingency table with the number of samples from the tissue in which the exon is included (PSI>0.8) and excluded (PSI<0.5), and the number of samples from the rest of the tissues in which the exon is included and excluded. Vertical red lines indicated an absolute value of phi equal to 0.95. Values of phi

close to 1 indicate that the exon is included in (nearly) all samples from the tissue, and excluded in (nearly) all samples from the rest of the tissues. Values of phi close to -1 indicate that the exon is excluded in (nearly) all samples from the tissue and included in (nearly) all samples form the rest of the tissues. There are only a few exons with tissue specific inclusion or exclusion at the 0.95 and -0.95 phi thresholds.**B-E.** Exon–intron structure, exon inclusion levels and expression values for several exons. Exon inclusion may be accompanied by changes in gene expression, but this is not always the case. **B.** An exon from the GABBR1 gene specifically included in brain. This exon contains a STOP codon that shortens the protein and impairs it to have a transmembrane domain that normally anchors it to the membrane. So GABBR1 isoforms expressed in brain will most likely be soluble (*70*). **C**. An exon from the APP genes specifically excluded in brain. This exon contains glycosylation domain so that the proteins expressed in brain will not have it (*71*). **D.** An exon from the BIN1 gene specifically included in muscle. BIN1 is related to a myopathy (*72*). **E.** Brain-specific exon exclusion on exon 7 of the EPN1 gene. Exon 7 is specifically excluded in brain and lower expression levels of EPN1 in brain cannot explain this pattern. 7tm_3 = 7 transmembrane sweet-taste receptor of 3 GCPR**.** OX2 = glicosilation domain. UID = ubiquitin domain.

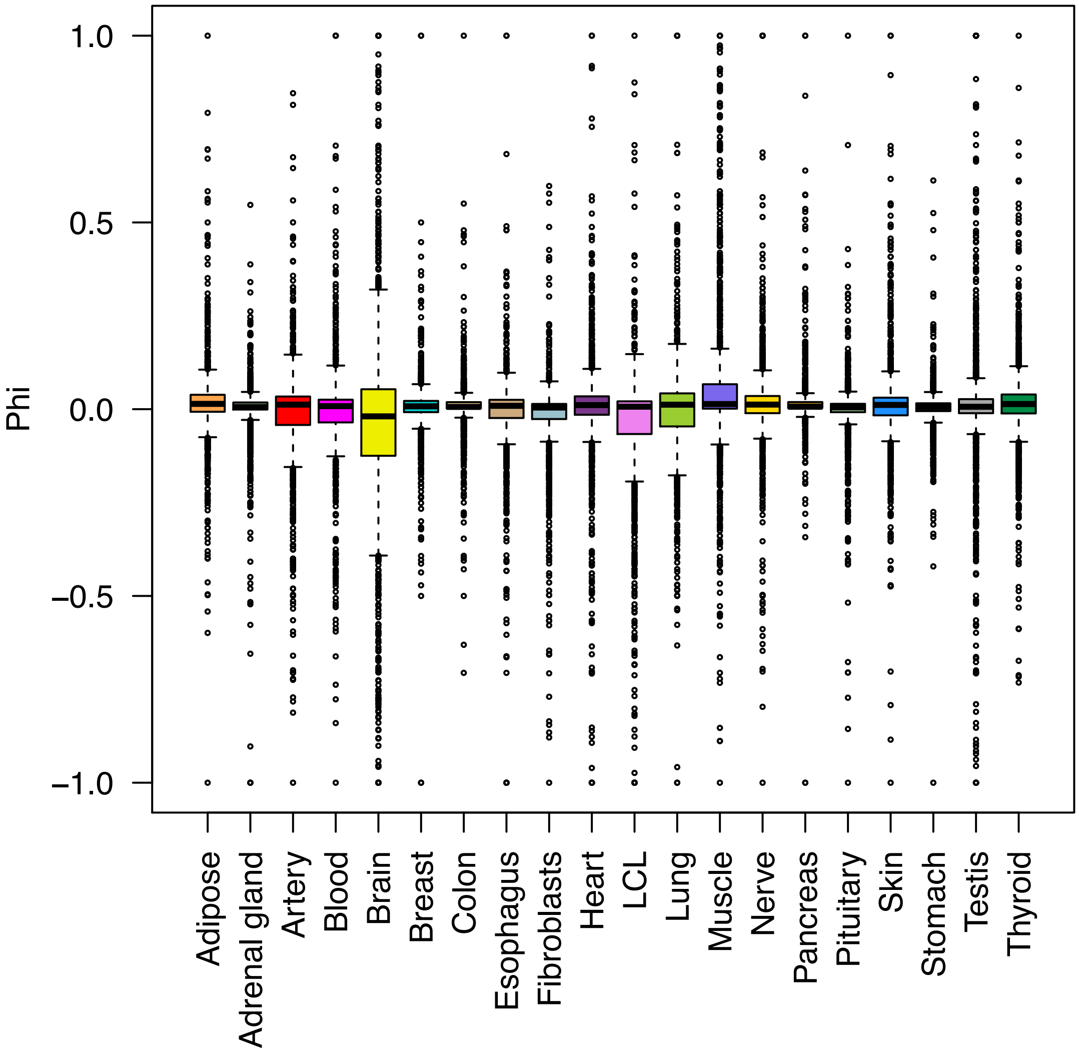


Figure S21. Short exon inclusion exclusivity across tissues. The distribution of phi is given across all short exons (15bp<length<60bp) within each tissue. For each short exon and for each tissue, the phi statistic is computed from a 2x2 contingency table with the number of samples from the tissue in which the exon is included (PSI>0.8) and excluded (PSI<0.5), and the number of samples from the rest of the tissues in which the exon is included (PSI>0.8) and excluded (PSI<0.5). Vertical red lines indicated an absolute value of phi equal to 0.95. Values of phi close to 1 indicate that the exon is included in (nearly) all samples from the tissue, and excluded in (nearly) all samples from the rest of the tissues. Values of phi close to -1 indicate that the exon is excluded in (nearly) all samples from the tissue and included in (nearly) all samples form the rest of the tissues. The observation that microexons (length<15bp) tend to be more included in brain does not hold for short exons shown here.

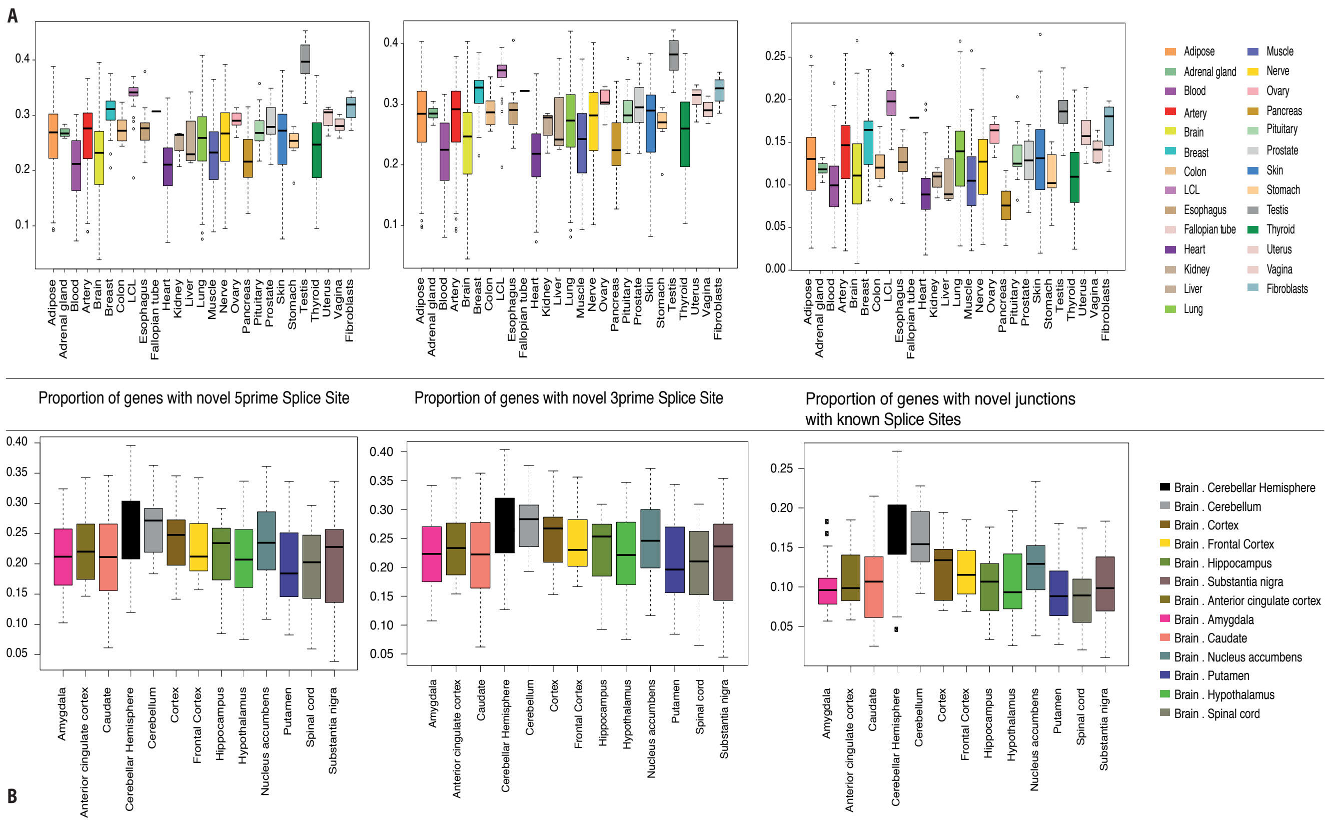


Figure S22. Proportion of genes with novel splicing events. Proportion of genes with a detected: novel 5' splice site usage (if 3' splice site is annotated and the 5' is not), novel 3'splice site usage (if 5' splice site is annotated and the 3' is not) or novel junction with known splice sites (if both splice sites are present in the annotation but the intron is not annotated) over the number of genes with a detected annotated splice junction. **A.** Tissues. **B.** Brain sub-tissues. Testis is the tissue with the largest proportion of expressed genes with unnanotated splicing events, and cerebellar tissues have the largest among brain subtissues.

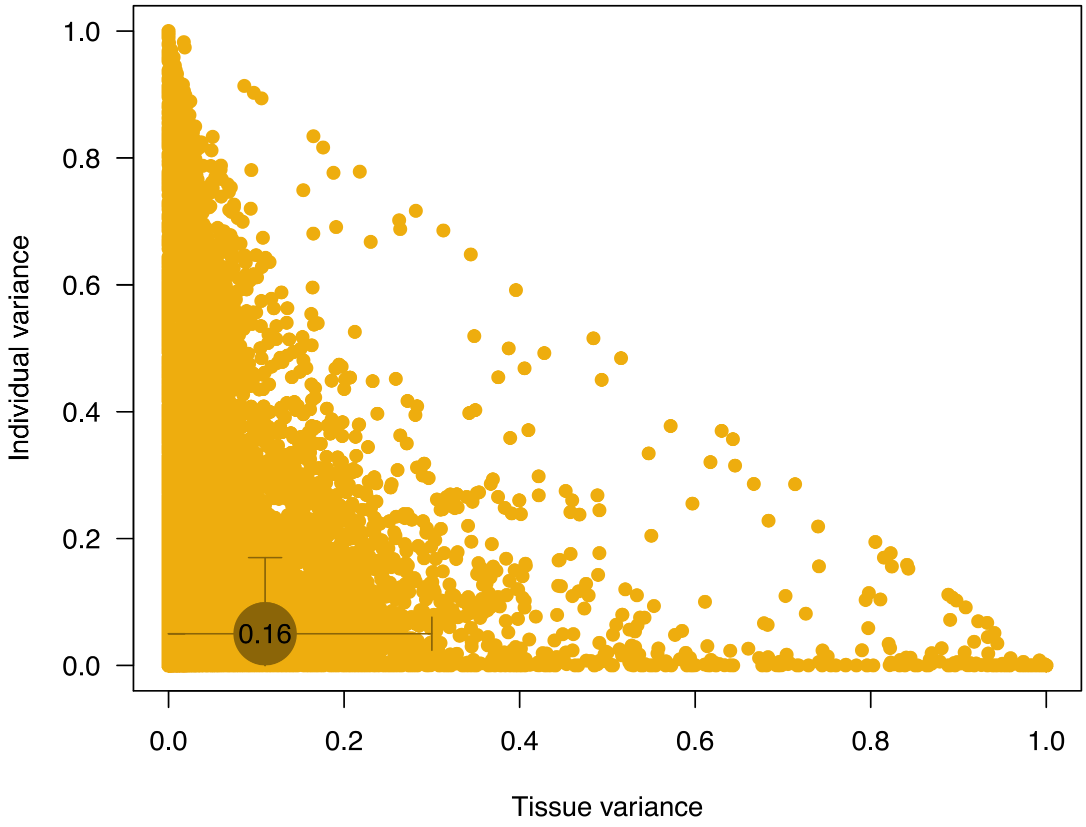


Figure S23. Contribution of tissue and individual to exon inclusion (PSI) variation in protein-coding genes. The circle is centered at the mean of individual and tissue contributions to exon inclusion variation and the segment lines correspond to half standard deviation. The number inside the circle is the sum individual and tissue contributions to exon inclusion variation.

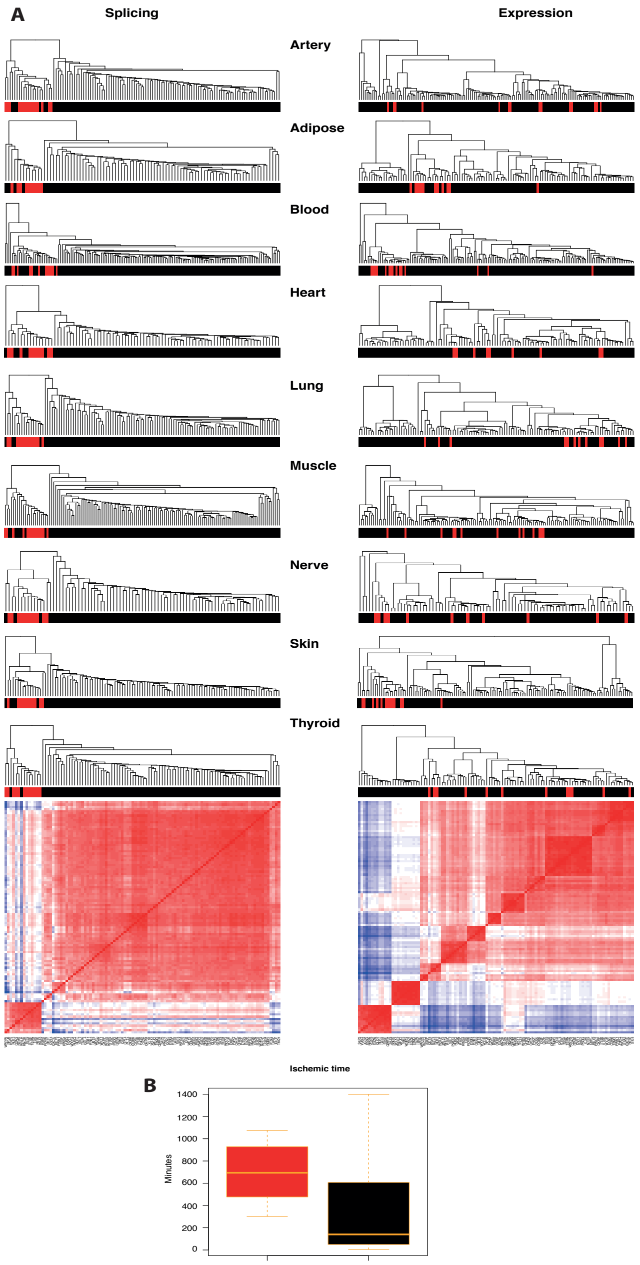


Figure S24. Effect of ischemic time on splicing and expression. **A**. Hierarchical clustering for splicing (exon inclusion levels quantified as PSI) and expression (gene expression levels quantified as RPKM) for the nine main tissues. For thyroid, both dendrogram and heatmap are shown. Clustering based on PSI values consistently identified an outgroup that included from 16 to 26 individuals (mean of 18). Samples from 17 individuals that belong to the outgroup in at least 5 out of the 9 main tissues are highlighted in red. **B.** Distribution of ischemic time for the samples that belong to the 17 individuals in the outgroup (red) and for the remaining samples (black). Differences between the two groups are significant (Wilcoxon test W = 2095.5, p-value < 5-e05).

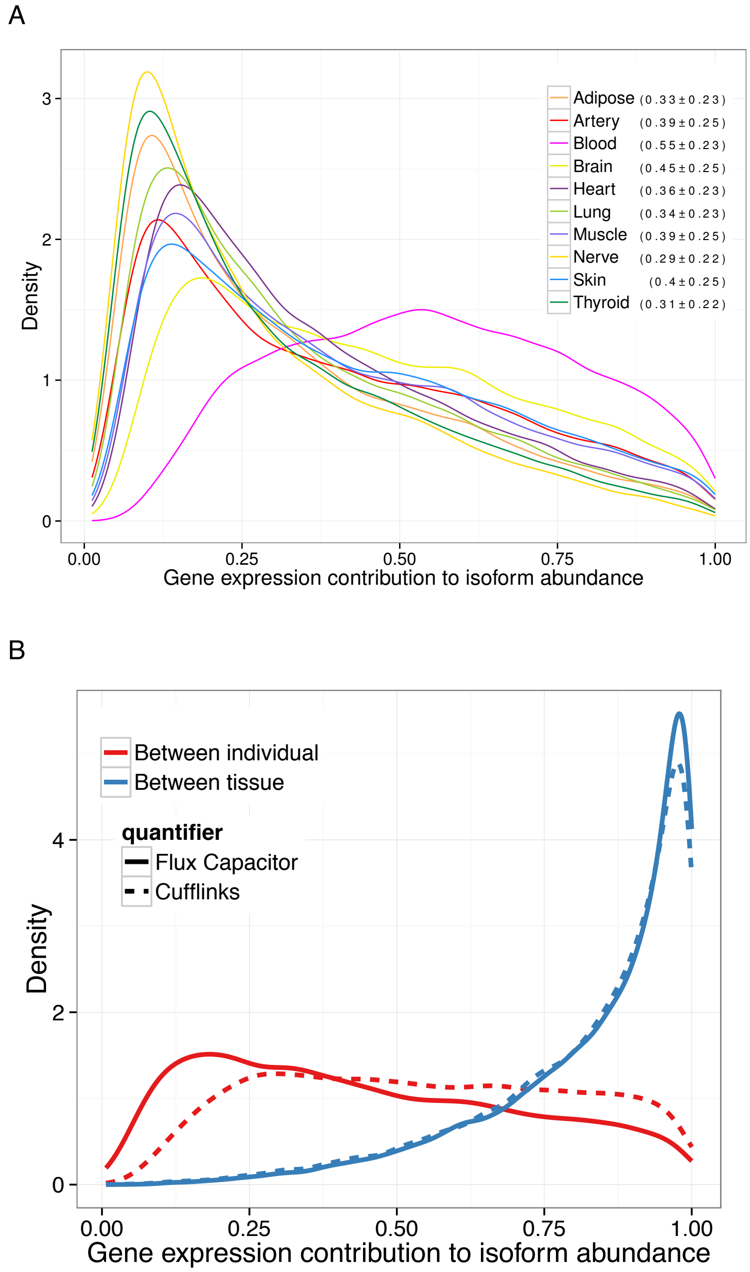


Figure S25. A. Contribution of gene expression to variation in isoform abundance within each tissue. Using a multiplicative model (see text) we estimate the fraction of the variation in isoform abundance across individuals that can be explained by variation in gene expression. The figure shows the distribution of this fraction for each gene in different tissues. With the exception of blood, variation in gene expression contributes less than 50% to the observed variation in isoform abundance across individuals. A fraction of the unexplained variation can be attributed to splicing variability. **B.** Contribution of gene expression to the between-individual and between-tissue variation in isoform abundance when using Flux Capacitor or Cufflinks quantifications. Due to fewer available samples, Cufflinks analysis used 133 samples (19 individuals across 7 tissues) compared to 380 samples (38 individuals across 10 tissues) for Flux Capacitor analysis. The same genes were analyzed.

Supplementary tables

Table S1. Characteristics of the 1,641 RNA-sequenced samples included in the GTEx pilot data analysis freeze. Column's A-C show the official tissue name, tissue site abbreviation and color assigned to each tissue in the GTEx pilot project (as used in several analyses here, and in (*6*)). The 9 tissues prioritized for sequencing, and with largest sample sizes, are indicated in bold and with a # symbol. Column D shows which tissues were combined for some analyses and their corresponding colour used. The Brain regions shown in the box are two regions each sampled in duplicate, the Cerebellum (BRNCHA and B), and Cortex (BRNCTXA and B). The "A" samples were sampled at the collection site with the other tissues, and preserved in PAXgene tissue preservative [PAX]. The "B" samples were re-sampled at the Miami Brain Bank and were frozen [F] (see Section1, and (*6*)). AA = African American, EA = European American, AS = Asian. The EBV-transformed lymphocytes (LCL) were cultured from whole blood. Fibroblast samples (FIBRBLS) were cultured from skin adjacent to the Skin-Sun Exposed (Lower Leg) (SKINS) samples.

Tissue Site Detail	Tissue site Abbreviation	Color	Combined Tissue	Analysis Freeze n	Sample Ischemic Time (min) Mean	Age Mean	Gender M	F	Ethnicity AA	EA	AS
Adipose - Subcutaneous #	**ADPSBQ**	🟧	**Adipose**	94	422	48	33	61	14	79	1
Adipose - Visceral (Omentum)	ADPVSC	🟧		19	401	47	4	15	3	16	0
Adrenal Gland	ADRNLG	🟩		12	173	51	5	7	3	9	0
Artery - Aorta	ARTAORT	🟥	**Artery**	24	262	51	7	17	5	19	0
Artery - Coronary	ARTCRN	🟥		9	312	55	4	5	1	8	0
Artery - Tibial #	**ARTTBL**	🟥		112	487	48	46	66	15	94	2
Brain - Amygdala	BRNAMY	🟨		23	NA	51	9	14	1	22	0
Brain - Anterior cingulate cortex (BA24)	BRNACC	🟨		17	NA	51	7	10	1	16	0
Brain - Caudate (basal ganglia)	BRNCDT	🟨		36	NA	53	14	22	1	33	1
Brain - Cerebellum [PAX]	BRNCHA	🟨		30	869	52	14	16	0	29	1
Brain - Cerebellar Hemisphere [F]	BRNCHB	🟨		24	NA	50	8	16	1	22	1
Brain - Cortex [PAX]	BRNCTXA	🟨		23	837	51	10	14	1	23	0
Brain - Frontal Cortex (BA9) [F]	BRNCTXB	🟨		24	NA	55	11	12	0	23	0
Brain - Hippocampus	BRNHPP	🟨		24	NA	51	9	15	0	24	0
Brain - Hypothalamus	BRNHPT	🟨		23	NA	51	10	13	0	22	1
Brain - Nucleus accumbens (b.g)	BRNNCC	🟨		28	NA	53	13	15	1	25	1
Brain - Putamen (basal ganglia)	BRNPTM	🟨		20	NA	50	8	12	1	18	1
Brain - Spinal cord (cervical c-1)	BRNSPC	🟨		16	NA	53	9	7	0	15	1
Brain - Substantia nigra	BRNSNG	🟨		25	NA	54	11	14	1	23	1
Breast - Mammary Tissue	BREAST	🟦		27	646	50	13	14	5	22	0
Cells -EBV-transformed lymphocytes	LCL	🟪		39	-60	46	13	26	9	30	0
Cells - Transformed fibroblasts	FIBRBLS	🟦		14	545	49	4	10	1	12	0
Colon - Transverse	CLNTRN	🟫		12	237	46	4	8	1	11	0
Esophagus - Mucosa	ESPMCS	🟫	**Esophagus**	18	331	52	6	12	2	16	0
Esophagus - Muscularis	ESPMSL	🟫		20	311	48	5	15	3	17	0
Fallopian Tube	FLLPNT	⬜		1	520	51	1	0	1	0	0
Heart - Atrial Appendage	HRTAA	🟪	**Heart**	25	492	51	6	19	2	23	0
Heart - Left Ventricle #	**HRTLV**	🟪		83	381	48	28	55	9	72	2
Kidney - Cortex	KDNCTX	🟫		3	583	56	0	3	1	2	0
Liver	LIVER	🟫		5	365	43	2	3	1	4	0
Lung #	**LUNG**	🟩		119	447	49	43	76	14	104	1
Muscle – Skeletal #	**MSCLSK**	🟪		138	486	49	51	87	18	117	2
Nerve – Tibial #	**NERVET**	🟨		88	464	49	34	54	13	73	2
Ovary	OVARY	⬜		6	401	44	6	0	2	4	0
Pancreas	PNCREAS	🟫		19	200	49	6	13	7	12	0
Pituitary	PTTARY	🟩		13	841	52	5	8	0	13	0
Prostate	PRSTTE	⬜		9	231	50	0	9	4	5	0
Skin-Not Sun Expsd (Suprapubic)	SKINNS	🟦	**Skin**	23	557	49	7	16	3	20	0
Skin-Sun Exposed (Lower leg) #	**SKINS**	🟦		96	499	49	36	60	11	83	2
Stomach	STMACH	🟧		12	250	48	6	6	4	8	0
Testis	TESTIS	⬜		14	294	52	0	14	3	11	0
Thyroid #	**THYROID**	🟩		105	429	49	40	65	13	90	2
Uterus	UTERUS	⬜		7	313	49	7	0	2	5	0
Vagina	VAGINA	⬜		6	415	54	6	0	1	5	0
Whole Blood #	**WHLBLD**	🟪		156	238	50	57	99	24	129	2
All				**1641**	**419**	**50**	**618**	**1023**	**203**	**1408**	**24**

Table S2. Data sets used to investigate the post-mortem effect on transcriptome patterns. Data sets used to compare the expression data from GTEx tissues with the same tissues obtained from living donors (e.g. surgical samples). All GTEx data analyzed were generated by RNA sequencing. All external data sets were microarray-based datasets downloaded from GEO or Array express, and derived from normal tissues that matched GTEx tissue sampling sites. A total of 798 GTEx samples were compared to 609 surgical samples representing 8 tissue sites.

Tissue	Number of GTEx (Deceased) samples	Number of External (Surgical) samples	Number of External datasets	Number of External microarray types
Adipose	93	45	4	1
Blood	69 Pre/74 Post	169	3	3
Breast	21	13	3	1
Heart (ventricle)	84	24	4	3
Lung	119	70	4	3
Muscle (skeletal)	136	139	4	2
Skin	95	123	8	4
Thyroid	107	26	5	2
Total	**798**	**609**	**35**	**19**

Table S3. Tissue classification of samples from living donors, using GTEx tissues (A). Results of the support vector machine (SVM) classifier trained on a subset of the GTEx data set (model). Metagene class weights were derived from the training (model) data set and were used to predict the tissue types of the remaining (test) GTEx data set (Section 1 and 2). Green cells along the diagonal highlight correctly predicted tissues. Orange cells highlight misclassified samples. Only a single breast sample was misclassified as adipose, and pathology notes on the sample indicated it had high adipose content. **(B)** Results of the test classification of external, surgically collected samples using the GTEx data as model. Despite the number of different studies represented (each with different processing and array-based assay conditions), the GTEx model provides a highly accurate classification of the tissues. Blood is the only GTEx tissue for which a subset of samples was collected ante-mortem. The observation that the expression patterns of those samples is a better classifier of the external bloods obtained from living donors demonstrates that we do observe an ischemic time effect on the data. Overall, however, we observe that strong tissue-specific expression profiles, representative of living tissues, are maintained in the postmortem GTEx samples. LV = left ventricle, Sk = skeletal.

A

	GTEx Tissues (Model)								
GTEx Tissues (test)	**Adipose**	**Blood (Ante)**	**Blood (Post)**	**Breast**	**Heart (LV)**	**Lung**	**Muscle (Sk)**	**Skin**	**Thyroid**
Adipose	93	0	0	0	0	0	0	0	0
Blood-Ante	0	69	0	0	0	0	0	0	0
Blood-Post	0	0	74	0	0	0	0	0	0
Breast	1	0	0	20	0	0	0	0	0
Heart (LV)	0	0	0	0	84	0	0	0	0
Lung	0	0	0	0	0	119	0	0	0
Muscle (Sk)	0	0	0	0	0	0	136	0	0
Skin	0	0	0	0	0	0	0	95	0
Thyroid	0	0	0	0	0	0	0	0	107

B

	GTEx Tissues (Model)								
External Tissues (test)	**Adipose**	**Blood (Ante)**	**Blood (Post)**	**Breast**	**Heart (LV)**	**Lung**	**Muscle (Sk)**	**Skin**	**Thyroid**
Adipose	43	0	0	0	0	0	1	0	0
Blood	0	166	3	0	0	0	0	0	0
Breast	4	0	0	9	0	0	0	0	0
Heart (LV)	0	0	0	0	24	0	0	0	0
Lung	0	0	1	0	0	69	0	0	0
Muscle (Sk)	0	0	0	0	0	0	137	2	0
Skin	1	0	0	0	0	0	0	122	0
Thyroid	0	0	0	0	0	0	0	0	26

Table S4 Top Expressed genes. The hundred most expressed genes in each tissue and their cumulative contribution to the global amount of gene expression in that tissue. *Table is an excel file.*

Table S5. Tissue preferentially expressed genes. We used NOISeq to call a gene tissue preferentially expressed by comparing the samples from a given tissue to those samples from the rest of the tissues. The table shows all instances where the mean expression of the gene in the tested tissue was significantly higher (FDR<0.01 and a log2 fold change >= 4) than in the samples from the rest of the tissues. *Table is an excel file.*

Table S6. **Tissue exclusive expressed genes.** Genes with high tissue specificity (phi>=0.95) or tissue anti-specificity (phi < -0.95). For each pair (gene, tissue) the phi statistic is computed from a 2x2 contingency table that includes the number of samples from the tissue in which the gene is expressed (RPKM>0.1) and not expressed (RPKM<0.1), and the number of samples from the rest of the tissues in which the gene is expressed and not expressed. Values of phi >=0.95 indicate that the gene is expressed in (nearly) all samples from the tissue (it is exclusive of the tissue), and in (nearly) no samples from the rest of the tissues. Genes with phi < -0.95 indicate that it is (nearly) not expressed in samples from the tissue and it is expressed in (nearly) all samples form the rest of the tissues. *Table is an excel file.*

Table S7. Number of tissue exclusive expressed genes with different expression thresholds. Number of genes with high tissue specificity (phi>=0.95) or tissue anti-specificity (phi < -0.95) using different thresholds to consider a gene expressed (threshold expressed) or not expressed (threshold not expressed). For each pair (gene, tissue) the phi statistic is computed from a 2x2 contingency table that includes the number of samples from the tissue in which the gene is expressed (RPKM> threshold expressed) and not expressed (RPKM< threshold not expressed), and the number of samples from the rest of the tissues in which the gene is expressed and not expressed. For all thresholds, most tissue exclusive genes are from Testis. *Table is an excel file.*

Table S8. Repeat expression and correlation with nearby genes. Columns 3-5, number of significantly correlated (Pearson p-val<0.05) repeats and closest gene (average distance 2.8 Kb). Percentage in parenthesis represent the number of repeats correlated with a gene nearby with respect to the number of repeats experessed in that tissue. Column 6, the number of repeat families with significant instances of co-expression. Columns 7-9, intersection of the two analysis above: the number of repeats whose expression is associated with the gene nearby and co-expressed with the other instances of its family. Columns 10-12, "Upstream candidates" refer to repeats with family co-expression, located more than 3kb upstream of the associated gene. Skin has similar numbers of expressed repeats than the other tissues but has half as much repeat-gene association compared to the other tissues (only 1.1% of the repeats that are expressed are co-expressed with a gene).

		Expression correlation between genes and repeats		**Repeat family co-expression**	**Expression correlation between genes and repeats and between repeats and elements of their repeat family**			**Upstream candidates**		
Tissue	**# of repeat expressed**	**# of genes**	**# of repeats (%)**	**# of repeat families**	**# of genes**	**# of repeats (%)**	**# of repeat families**	**# of genes**	**# of repeats (%)**	**# of repeat families**
Adipose	122830	1300	3854 (3.1%)	213	591	1021 (0.83%)	111	27	49 (0.04%)	27
Artery	118196	1513	4712 (4%)	305	783	1488 (1.3%)	187	34	63 (0.053%)	36
Blood	60313	345	890 (1.5%)	98	91	119 (0.2%)	23	4	6 (0.0099%)	4
Brain	98754	938	2451 (2.5%)	271	303	470 (0.48%)	118	21	31 (0.031%)	20
Heart	103411	1074	3354 (3.2%)	252	590	1149 (1.1%)	128	20	33 (0.032%)	23
Lung	134514	1443	4281 (3.2%)	240	729	1305 (0.97%)	125	24	42 (0.031%)	20
Muscle	81594	1272	4533 (5.6%)	169	389	569 (0.7%)	75	17	25 (0.031%)	11
Nerve	145566	1246	3483 (2.4%)	237	521	894 (0.61%)	136	42	70 (0.048%)	31
Skin	113784	470	1226 (1.1%)	235	214	347 (0.3%)	81	5	5 (0.0044%)	5
Thyroid	143520	1425	4058 (2.8%)	233	737	1347 (0.94%)	121	36	73 (0.051%)	32
All	209541	3046	10958 (5.2%)	592	1731	3966 (1.9%)	379	136	276 (0.13%)	97

Table S9. Contribution of individual and tissue to variation in gene expression of protein coding and lncRNAs. Relative contribution of individual, tissue and residual to the variance in gene expression. *Table is an excel file.*

Table S10. Genes differentially expressed between males and females (FDR <0.05) *Table is an excel file.*

Table S11. Genes differentially expressed between African American and individuals of European ancestry (FDR<0.05). *Table is an excel file.*

Table S12. Genes differentially expressed across age. Genes that changed (FDR<0.05) expression with age across all GTEx tissues based on LMM analysis. *Table is an excel file.*

Table S13. **Genes differentially expressed between males and females within each tissue (FDR<0.05)**. *Table is an excel file*.

Table S14. **Genes differentially expressed between African American and individuals of European ancestry within each tissue (FDR<0.05)**. *Table is an excel file*.

Table S15. Tissue preferential Exon Inclusion. Number of exons differentially included between a given tissue and the remaining tissues. An exon is considered to be differentially included if FDR < 0.01 and ΔPSI>0.1 *Table is an excel file.*

Table S16. Tissue exclusive Exon Inclusion. Number of exons that have high tissue exclusivity (phi>=0.95) or tissue anti-exclusivity (phi < -0.95) *Table is an excel file.*

Table S17. Individual and Tissue contribution to variation of splicing in protein coding genes. *Table is an excel file*

Table S18. GO enrichment analysis for genes with high contribution of individual to splicing variation. Analysis was carried out with the 139 genes that could be mapped in the DAVID database. Twenty-nine clusters were detected. The figure shows the first 2 most enriched clusters. The first cluster has 9-fold enrichment for functions related to translation and the ribosomes. *Table is an excel file.*

Table S19. Number of genes changing major isoform between tissues. Top right, the isoform switch involves changes in the coding sequence. Bottom left, the isoform switch involves changes in the 3'UTR, 5'UTR or the coding sequence.

	Adipose tissue	**Blood**	**Artery**	**Brain**	**Heart**	**Lung**	**Muscle**	**Nerve**	**Skin**	**Thyroid**
Adipose tissue		47	3	63	21	14	22	10	21	21
Blood	53		54	113	46	40	47	80	45	68
Artery	3	57		62	24	16	30	11	67	50
Brain	71	121	74		59	57	85	54	67	50
Heart	27	84	29	64		25	20	36	34	33
Lung	14	44	18	65	31		52	23	22	13
Muscle	30	53	37	95	22	62		42	51	58
Nerve	11	87	14	58	39	24	47		30	25
Skin	26	53	45	76	45	25	63	38		28
Thyroid	25	76	38	54	40	14	70	29	37	

Table S20. GENCODE gene biotypes. Gene classification based on GENCODE 12 biotypes.

Gene Group	Number	gencode gene types
protein_coding	20110	protein_coding
pseudogene	12648	pseudogene
lncRNA	11790	3prime_overlapping_ncrna, antisense, lincRNA, non_coding, processed_transcript, sense_intronic, sense_overlapping
smallRNA	6963	Mt_rRNA, Mt_tRNA, miRNA, misc_RNA, rRNA, snRNA, snoRNA
smallRNA_pseudogene	1838	Mt_tRNA_pseudogene, miRNA_pseudogene, misc_RNA_pseudogene, rRNA_pseudogene, scRNA_pseudogene, snRNA_pseudogene, snoRNA_pseudogene, tRNA_pseudogene
IGorTR	364	IG_C_gene, IG_D_gene, IG_J_gene, IG_V_gene, TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene

References and Notes

1. FANTOM Consortium and the RIKEN PMI and CLST (DGT), A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. Jørgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabløs, A. S. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. Prendergast, O. J. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, Y. Hayashizaki, A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014). Medline doi:10.1038/nature13182

2. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). Medline

3. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. 't Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis; Geuvadis Consortium, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013). Medline doi:10.1038/nature12531

4. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector; Multiple Tissue Human Expression Resource (MuTHER) Consortium, Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012). Medline doi:10.1038/ng.2394

5. T. J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, H. F. Moore; GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). Medline doi:10.1038/ng.2653

6. The GTEx Consortium,The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **248**, 648–660 (2015).

7. Materials and methods are available in the supplementary materials on *Science* Online.

8. D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M. D. Nazaire, C. Williams, M. Reich, W. Winckler, G. Getz, RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012). Medline doi:10.1093/bioinformatics/bts196

9. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). Medline doi:10.1101/gr.135350.111

10. M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, A. Brazma, A global map of human gene expression. *Nat. Biotechnol.* **28**, 322–324 (2010). Medline doi:10.1038/nbt0410-322

11. A. C. Birdsill, D. G. Walker, L. Lue, L. I. Sue, T. G. Beach, Postmortem interval effect on RNA and gene expression in human brain tissue. *Cell Tissue Bank.* **12**, 311–318 (2011). Medline doi:10.1007/s10561-010-9210-8

12. J. P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 4164–4169 (2004). Medline doi:10.1073/pnas.0308531101

13. P. Carninci, Y. Shibata, N. Hayatsu, Y. Sugahara, K. Shibata, M. Itoh, H. Konno, Y. Okazaki, M. Muramatsu, Y. Hayashizaki, Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10**, 1617–1630 (2000). Medline doi:10.1101/gr.145100

14. R. D. Kelly, A. Mahmud, M. McKenzie, I. A. Trounce, J. C. St John, Mitochondrial DNA copy number is regulated in a tissue specific manner by DNA methylation of the nuclear-encoded DNA polymerase gamma A. *Nucleic Acids Res.* **40**, 10124–10138 (2012). Medline doi:10.1093/nar/gks770

15. D. Kelley, J. Rinn, Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012). Medline doi:10.1186/gb-2012-13-11-r107

16. L. Carrel, H. F. Willard, X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005). Medline doi:10.1038/nature03479

17. S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Röder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y.

Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigó, T. R. Gingeras, Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012). Medline doi:10.1038/nature11233

18. V. Regitz-Zagrosek, U. Seeland, Sex and gender differences in myocardial hypertrophy and heart failure. *Wien. Med. Wochenschr.* **161**, 109 (2011).

19. M. Yan, L. C. Wang, S. G. Hymowitz, S. Schilbach, J. Lee, A. Goddard, A. M. de Vos, W. Q. Gao, V. M. Dixit, Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290**, 523–527 (2000). Medline doi:10.1126/science.290.5491.523

20. G. Yeo, D. Holste, G. Kreiman, C. B. Burge, Variation in alternative splicing across human tissues. *Genome Biol.* **5**, R74 (2004). Medline doi:10.1186/gb-2004-5-10-r74

21. J. K. Pickrell, A. A. Pai, Y. Gilad, J. K. Pritchard, Noisy splicing drives mRNA isoform diversity in human cells. *PLOS Genet.* **6**, e1001236 (2010). Medline doi:10.1371/journal.pgen.1001236

22. I. Ezkurdia, A. del Pozo, A. Frankish, J. M. Rodriguez, J. Harrow, K. Ashman, A. Valencia, M. L. Tress, Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol. Biol. Evol.* **29**, 2265–2283 (2012). Medline doi:10.1093/molbev/mss100

23. M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol.* **14**, R70 (2013). Medline doi:10.1186/gb-2013-14-7-r70

24. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). Medline doi:10.1093/bioinformatics/btp120

25. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008). Medline doi:10.1038/nmeth.1226

26. F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y. H. Zhou, A. Abdellaoui, S. Batista, C. Butler, G. Chen, T. H. Chen, D. D'Ambrosio, P. Gallins, M. J. Ha, J. J. Hottenga, S. Huang, M. Kattenberg, J. Kochar, C. M. Middeldorp, A. Qu, A. Shabalin, J. Tischfield, L. Todd, J. Y. Tzeng, G. van Grootheest, J. M. Vink, Q. Wang, W. Wang, W. Wang, G. Willemsen, J. H. Smit, E. J. de Geus, Z. Yin, B. W. Penninx, D. I. Boomsma, Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014). Medline

27. S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, E. T. Dermitzakis, Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010). Medline doi:10.1038/nature08903

28. M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, F. Meng, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005). Medline doi:10.1093/nar/gni179

29. R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003). Medline doi:10.1093/biostatistics/4.2.249

30. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003). Medline doi:10.1093/bioinformatics/19.2.185

31. J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad, RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008). Medline doi:10.1101/gr.079558.108

32. N. Raghavachari, J. Barb, Y. Yang, P. Liu, K. Woodhouse, D. Levy, C. J. O'Donnell, P. J. Munson, G. J. Kato, A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med. Genomics* **5**, 28 (2012). Medline doi:10.1186/1755-8794-5-28

33. P. Tamayo, D. Scanfeld, B. L. Ebert, M. A. Gillette, C. W. Roberts, J. P. Mesirov, Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 5959–5964 (2007). Medline doi:10.1073/pnas.0701068104

34. R. Gaujoux, C. Seoighe, A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010). Medline doi:10.1186/1471-2105-11-367

35. T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. G. Knowles, J. Lagarde, L. Veeravalli, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. B. Brown, L. Lipovich, J. M. Gonzalez, M. Thomas, C. A. Davis, R. Shiekhattar, T. R. Gingeras, T. J. Hubbard, C. Notredame, J. Harrow, R. Guigó, The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012). Medline doi:10.1101/gr.132159.111

36. W. N. R. Venables, B. D., *Modern Applied Statistics with S.* (Springer, LOCATION, ed. 4, 2002).

37. D. Ramsköld, E. T. Wang, C. B. Burge, R. Sandberg, An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Comput. Biol.* **5**, e1000598 (2009). Medline doi:10.1371/journal.pcbi.1000598

38. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289 (1995).

39. K. S. Pollard, S. Dudoit, M. J. van der Laan, *Applications to Genomics, in Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, V. C. R. Gentleman, W. Huber, R. Irizarry, S. Dudoit, Eds., Statistics for Biology and Health Series (Springer, LOCATION, 2005), pp. 251–272.

40. A. Dabney, J. Storey, (with assistance from Gregory R. Warnes), qvalue: Q-value estimation for false discovery rate control; www.bioconductor.org/packages/release/bioc/html/qvalue.html.

41. S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21**, 2213–2223 (2011). Medline doi:10.1101/gr.124321.111

42. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). Medline doi:10.1186/s13059-014-0550-8

43. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010). Medline doi:10.1186/gb-2010-11-3-r25

44. W. Revelle, psych: Procedures for Personality and Psychological Research. http://CRAN.R-project.org/package=psych.

45. G. Bourque, Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr. Opin. Genet. Dev.* **19**, 607–612 (2009). Medline doi:10.1016/j.gde.2009.10.013

46. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

47. O. Nenadic, M. Greenacre, Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. *J. Stat. Softw.* **20**, 163–170 (2007).

48. D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear mixed-effects models using Eigen and S4. http://CRAN.R-project.org/package=lme4.

49. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010). Medline doi:10.1186/gb-2010-11-10-r106

50. J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R Core Team, nlme: Linear and nonlinear mixed effects. *R package version 1.2.* (2014).

51. G. Dennis Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, R. A. Lempicki, DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol.* **4**, 3 (2003). Medline doi:10.1186/gb-2003-4-5-p3

52. D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42**, D764–D770 (2014). Medline doi:10.1093/nar/gkt1168

53. S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. Bamezai, A. V. Hill, F. O. Vannberg, J. L. Rinn, E. S. Lander, S. F. Schaffner, P. C. Sabeti; 1000 Genomes Project, Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013). Medline doi:10.1016/j.cell.2013.01.035

54. J. K. Pickrell, G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, G. S. Barsh, R. M. Myers, M. W. Feldman, J. K. Pritchard, Signals of recent positive

selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009). Medline doi:10.1101/gr.087577.108

55. P. Langfelder, S. Horvath, WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). Medline doi:10.1186/1471-2105-9-559

56. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008). Medline doi:10.1093/bioinformatics/btm563

57. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, e17 (2005). Medline doi:10.2202/1544-6115.1128

58. P. Langfelder, R. Luo, M. C. Oldham, S. Horvath, Is my network module preserved and reproducible? *PLOS Comput. Biol.* **7**, e1001057 (2011). Medline doi:10.1371/journal.pcbi.1001057

59. D. D. Pervouchine, E. E. Khrameeva, M. Y. Pichugina, O. V. Nikolaienko, M. S. Gelfand, P. M. Rubtsov, A. A. Mironov, Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA* **18**, 1–15 (2012). Medline doi:10.1261/rna.029249.111

60. O. Denas, R. Sandstrom, Y. Cheng, K. Beal, J. Herrero, R. Hardison, J. Taylor, Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *bioRxiv* 10.1101/010926 (2014).

61. E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, C. B. Burge, Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008). Medline doi:10.1038/nature07509

62. N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, B. J. Blencowe, The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012). Medline doi:10.1126/science.1230612

63. M. Giulietti, F. Piva, M. D’Antonio, P. D’Onorio De Meo, D. Paoletti, T. Castrignanò, A. M. D’Erchia, E. Picardi, F. Zambelli, G. Principato, G. Pavesi, G. Pesole, SpliceAid-F: A database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* **41**, D125–D131 (2013). Medline doi:10.1093/nar/gks997

64. R Core Team, www.R-project.org/ (2014).

65. I. M. Shapiro, A. W. Cheng, N. C. Flytzanis, M. Balsamo, J. S. Condeelis, M. H. Oktay, C. B. Burge, F. B. Gertler, An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLOS Genet.* **7**, e1002218 (2011). Medline doi:10.1371/journal.pgen.1002218

66. S. Falcon, R. Gentleman, Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258 (2007). Medline doi:10.1093/bioinformatics/btl567

67. M. Gonzàlez-Porta, M. Calvo, M. Sammeth, R. Guigó, Estimation of alternative splicing variability in human populations. *Genome Res.* **22**, 528–538 (2012). Medline doi:10.1101/gr.121947.111

68. J. Monlong, M. Calvo, P. G. Ferreira, R. Guigó, Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* **5**, 4698 (2014). Medline doi:10.1038/ncomms5698

69. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010). Medline doi:10.1038/nbt.1621

70. K. Kaupmann, V. Schuler, J. Mosbacher, S. Bischoff, H. Bittiger, J. Heid, W. Froestl, S. Leonhard, T. Pfaff, A. Karschin, B. Bettler, Human gamma-aminobutyric acid type B receptors are differentially expressed and regulate inwardly rectifying K+ channels. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14991–14996 (1998). Medline doi:10.1073/pnas.95.25.14991

71. S. Kitazume, Y. Tachida, M. Kato, Y. Yamaguchi, T. Honda, Y. Hashimoto, Y. Wada, T. Saito, N. Iwata, T. Saido, N. Taniguchi, Brain endothelial cells produce amyloid beta from amyloid precursor protein 770 and preferentially secrete the O-glycosylated form. *J. Biol. Chem.* **285**, 40097–40103 (2010). Medline doi:10.1074/jbc.M110.144626

72. R. J. Wechsler-Reya, K. J. Elliott, G. C. Prendergast, A role for the putative tumor suppressor Bin1 in muscle cell differentiation. *Mol. Cell. Biol.* **18**, 566–575 (1998). Medline