**Possible titles** (Please vote/make new suggestions!)

**1. The impact of biological sex on alternative splicing across tissues**

**2. Sex-biased gene expression and alternative splicing and their interaction across tissues**

Guy Karlebach,1 Diogo Troggian Veiga,1 Anne Deslattes Mays,1 Anil Kesarwani,1 Daniel Danis,1 Georgios Kararigas,2 Joshy George,1 Guru Ananda,1 Chris Bizon,3 Rebecca Boyles,4 Chris Ball,4 Julie A McMurry,5 Melissa A Haendel,5 Jeremy Yang,6 Tudor Oprea,6 Olga Anczukow,1 Jacques Banchereau,1 Peter N Robinson1,7

1. The Jackson Laboratory for Genomic Medicine, Farmington CT 06032, USA

2. Charité – Universitätsmedizin Berlin and DZHK (German Centre for Cardiovascular Research), partner site Berlin, Germany

3. Renaissance Computing Institute (RENCI), 100 Europa Drive, Suite 540, Chapel Hill, NC 27517, United States.

4. RTI International, Durham, NC 27709, USA

5. Linus Pauling institute, Oregon State University, Corvallis OR, USA

6. Translational Informatics Division, Department of Internal Medicine, The University of New Mexico Health Science Center, Albuquerque, NM 87131, USA.

7. Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA

**Abstract**

**Over 95% of human genes undergo alternative splicing in a developmental, tissue-specific or signal transduction-dependent manner.**[1](https://paperpile.com/c/Dj72Qo/Mym2) **A number of factors including binding of cis-acting sequences by RNA-binding proteins (RBPs) and the influence of transcriptional elongation speed on the exon inclusion are known to affect alternative splicing (AS), but the combinatorial mechanisms leading to the distribution of spliced isoforms remain largely unstudied. The extent to which sex-biased gene expression and splicing lead to biological sex differences has yet to be systematically investigated. Here, we investigated 9,011 samples from 532 individuals across 53 tissues from the Genotype-Tissue Expression (GTEx) resource, and identified 4,135 genes with significantly sex-biased expression in at least one tissue, as well as 5,925 significantly sex-biased AS events. We used hierarchical Bayesian modeling to characterize the roles of RBPs and gene expression and showed that 87% of sex-biased AS events correlated with the expression of at least one RBP, and 67% of AS events correlated with gene expression. In 40% of AS events, our model indicated the existence of additional factors correlated with sex differences that were not captured by RBPs or gene expression. About 30% of sex-biased exon skipping events correlated with gene expression were associated with nonsense-mediated decay (NMD) isoforms. The remaining isoforms were significantly more highly expressed. For those exon skipping events most associated with gene expression, inclusion-related isoforms were more likely to be associated with RBP levels than skipping-related isoforms. Overall, our results demonstrate widespread sex differences in alternative splicing associated with RBP levels and gene expression in a characteristic fashion. We provide an integrative method for the interpretation of the RBP and gene expression levels as putative factors that mediate sex-biased alternative splicing.**

AS, a process by which splice sites are used differentially to create protein diversity, plays important roles in development,[2](https://paperpile.com/c/Dj72Qo/PcQx) disease,[3](https://paperpile.com/c/Dj72Qo/mhuQ) and aging.[4](https://paperpile.com/c/Dj72Qo/Ky2A) Although some AS is constitutive, meaning that multiple isoforms are produced in the same proportions in all or most cell types, AS is more often regulated by developmental or differential cues or in response to external stimuli.[1](https://paperpile.com/c/Dj72Qo/Mym2) Several mechanisms have been demonstrated to underlie AS, although their combinatorial interactions remain poorly understood. Intronic and exonic cis-acting regulatory sequences that are recognized by RBPs may promote or suppress AS events such as exon inclusion.[5](https://paperpile.com/c/Dj72Qo/3V13) Several chromatin-level mechanisms play a role in AS. Nucleosome density is higher within exons than introns, suggesting the existence of RNA polymerase II (RNA Pol II) mediated cross-talk between chromatin structure and exon-intron architecture.[6](https://paperpile.com/c/Dj72Qo/0KEo) Alternative exons with suboptimal splicing signals may require more time to be recognized by the splicing machinery, and faster transcriptional elongation by RNA Pol II may result in increased exon skipping. Additionally, certain histone modifications that can be enriched over exons may promote binding of proteins such as HP1α and HP1γ that in turn influence transcriptional speed.[7](https://paperpile.com/c/Dj72Qo/LfY1) Here, we perform a systematic survey of sex-biased AS across multiple tissues using an integrated suite of frequentist and Bayesian approaches to characterize the RBP and gene expression and their interplay in sex-biased AS events (Fig. 1a).

Although sex-biased gene expression is common,[8](https://paperpile.com/c/Dj72Qo/zNFm) and widespread differences in splicing have been identified in the human brain,[9](https://paperpile.com/c/Dj72Qo/K0Ju) no analysis of sex-biased AS has been performed to date over a comprehensive dataset that spans multiple tissue types. The Genotype-Tissue Expression (GTEx) project comprises samples from 53 non-diseased tissue sites across nearly 1000 individuals which have been assayed by whole genome or exome sequencing, and RNA-Seq.[10–15](https://paperpile.com/c/Dj72Qo/d8KX+Ehmb+F9BY+zP2T+a1iT+NoUj) In this study, we explored the GTEx resource to investigate gene expression and AS in male and female subjects. We analyzed gene expression and AS in 8,944 samples from 532 individuals across 53 tissues (Supplemental Tables S1,S2) with the goal of characterizing sex-specific patterns of gene expression, AS, and abundance of RBPs in tissues.

We first sought statistically significant differences between males and females in gene expression using the voom function from the limma package.[16](https://paperpile.com/c/Dj72Qo/3WnC) A total of 4,135 genes with significantly sex-biased expression were detected at a false-discovery rate (FDR) cutoff of 0.05 and a fold change cutoff of 1.5. Y chromosomal genes were excluded from the analysis. The tissue that had the largest number of genes showing significantly sex-biased differential expression (DE) was breast, with 2,579 DE genes, followed by thyroid with 254 DE genes, skin with 224 and adipose-subcutaneous with 201 DE genes. Kidney cortex had 2 DE genes, and bladder had none. Left ventricle (LV) and brain cortex had 61 and 24 DE genes, respectively (Supplemental Figure S1). We created a heatplot based on the mean fold change of genes between male and female samples; this showed that there was a consistent shift in expression patterns in some related tissues, such as 11 structures in the brain (top left), as well as three arterial tissues, two esophageal tissues, and sun-exposed and non-sun-exposed skin (Fig. 1B).

|  |  |
| --- | --- |
|  |  |

Figure 1. **(a)** Flowchart depicting the analysis of GTEx RNA-seq data. Analysis of GTEx gene expression and alternative splicing profiles identified significantly sex-biased genes and AS events. Data were used as input for a hierarchical Bayesian model to characterize the influence of RNA binding proteins and gene expression on sex-biased ASEs. **(b)** Heatplot representing similarity in the fold-changes between male and female samples, with the values in the heatmap being the correlation between the vectors of fold changes of the tissues.

Gene Ontology (GO) analysis of genes showing sex-specific expression in one or more tissues revealed 101 enrichments with a posterior probability greater than 0.5 with 79 unique GO terms in 29 tissues (Supplementary Table S3). A number of the GO terms could reflect known sex differences, such as the enrichment of extracellular matrix in breast tissue, which is differentially expressed between the sexes,[17](https://paperpile.com/c/Dj72Qo/UqaT) or translation initiation factor activity, which was differential in five tissues in our examination; several translation initiation factors have been shown to be differentially expressed between male and female muscle tissue.[18](https://paperpile.com/c/Dj72Qo/f8iQ)

We then investigated AS events. We chose to investigate individual AS events rather than transcript (isoform) abundance, because despite improvements in algorithms, accurate quantification of the expression of individual transcripts is challenging with short-read RNA-seq technology, especially for short or low-abundance transcripts and genes with complex structures.[19–22](https://paperpile.com/c/Dj72Qo/umkd+r9X5+Abpt+FlJE) We focused on five classes of discrete AS events (ASE) comprising exon skipping or inclusion, alternative 5’ or 3’ splicing, mutually exclusive exons, and intron retention (Fig 2a) and defined sex-specific AS based on a statistical model with sex, ASE, and sex:event interaction as covariates. We called AS event sex-biased if the interaction term was significant following multiple testing correction (Methods, Supplemental Figure S2).

For the AS analysis, the 53 tissue types were consolidated into 46 groups with very similar patterns according to YARN (Methods; Supplemental Table S2). Statistical analysis revealed from 0 to 2,579 differentially expressed genes per tissue; from 0 to 2,724 genes per tissue were found to harbor one or more significant AS events. We found a different number of AS events for the different tissues, with the highest number occurring for the LV (3,619 skipped exon [SE] events), breast (1,541 SE events), and brain cortex **(**369 SE events), whereas other tissues contained far fewer AS events; for example, kidney cortex and bladder were not found to have AS events of any type. 2225 genes were affected by AS in the left ventricle (mean of 1.63 AS events per gene), with 934 genes in breast (1.65 AS events/gene) and 321 in brain cortex (1.15 AS events/gene). The total number of AS events over all tissues was 5,925 (Supplemental Figure S3).

It has been reported that differentially expressed sex-biased genes are more likely to be linked to escape from X chromosome inactivation as compared to inactivated or variable genes.[23](https://paperpile.com/c/Dj72Qo/hAfn) We confirmed this result with our data (Figure 2b, Fisher’s exact test, *P* = 5.47 × 10−50). We hypothesized that AS events might be more commonly observed in X chromosomal genes that escape inactivation. Indeed, we found that escaped genes were enriched compare to inactive and variable genes (*P* = 2.46 × 10−20; Figure 2b).

We performed GO analysis over genes harboring one or more ASEs. A total of 54 distinct GO terms were significantly enriched in one or more tissues. For instance, α-actinin is a cytoskeletal actin-binding protein that plays structural and regulatory roles in cytoskeleton organization and muscle contraction;[24](https://paperpile.com/c/Dj72Qo/Ptux) inclusion isoforms were significantly less in males and overall gene expression was significantly higher in females (Fig. 2C; Supplemental Table S4).

|  |  |
| --- | --- |
| A) | **b)** |
| **Fig 2c** |  |

**Figure 2. (a)** The five categories of ASE that were investigated in this work.  **(b)** Percentages of differentially expressed or alternatively spliced genes among X-chromosomal genes escaped from X-chromosomal inactivation, inactivated, or variable. **(c)** Box plot showing fold change of exon inclusion (top) and gene expression (bottom) for genes annotated to the Gene Ontology (GO) term *alpha-actinin binding* (GO:0051393).

|  |  |  |  |
| --- | --- | --- | --- |
| a)   |  | | --- | |  | |  | | b) |
| c) | d) |
| e) |  |

**Figure 3**. **Hierarchical Bayesian Modeling**. **(a)** Structure of the HBM used in this work. The figure illustrates the prediction of skipping probabilities in a sample for m skipped exon events. **(b)** An example Bayesian high density interval plot for the effect of the RBP ZNF638 on a skipped exon event (chrX:24208217-24208370) in the *ZFX* gene in the left ventricle. The 3 curves correspond the 3 different Markov chains, and the three short vertical bars that intersect the x-axis correspond to the 95% HDIs of these chains. **(c)** 3-dimensional plot illustrating the relationship between gene expression and inclusion and exclusion counts for an exon skipping event of CDKN2A in breast. Females show both higher gene expression as well as higher skip counts than males, and the 95% HDI of the expression coefficient for this event spans between 0.7 and 0.95 (Supplementary Fig. S3). **(d)** Bar plot of average RBP effects in five tissues. A positive average coefficient indicates a positive correlation between the gene expression of an RBP and the probability of skipping of the genes identified as targets by the HBM. A negative average coefficient indicates a negative correlation. **(e)** Histograms of RBP coefficients for sex-biased AS events (blue) and non-sex-biased events (green).

We hypothesized that differential levels of RBPs with roles in AS could be responsible for some of the AS events. Additionally, we observed a significant overlap between differentially expressed genes and sex-biased alternatively spliced genes (836 observed vs. 462 expected by chance, hypergeometric test p-value 1.55×10-93). Furthermore, because of recent reports that transcription elongation can affect gene expression,[25](https://paperpile.com/c/Dj72Qo/WM76) and the observation that faster transcriptional elongation speed of RNA Pol II can lead to increased exon skipping,[7](https://paperpile.com/c/Dj72Qo/LfY1) we additionally posited that overall gene expression levels could be correlated with exon skipping events. In order to test these hypotheses, we developed a hierarchical Bayesian model and applied it to the ten tissues with the largest number of sex-biased skipped exon (SE) events (Fig 3a). For each tissue, up to the most significant 100 sex-biased SE events and an identical number of SE events with the largest combined variance within the two types of read counts were chosen. If fewer than 100 events had been called significant, then only these significant events and an equal number of non sex-biased events were chosen. The hierarchical model was designed on the basis of a number of assumptions explained in detail in the Methods; the outcome of the splicing process, viz., counts of exon inclusion and exclusion, was modeled as a result of a weighted linear combination of the RNA concentrations of each of the 87 RBPs (Supplemental Tables S4) as well as of the overall expression level of the gene harboring the SE event. An additional sex term was included to model influences not captured by RBP or gene expression. The structure of the model and the priors placed on the individual distributions (nodes) of the model reflect our expectations about the data. The modeling process runs a Monte-Carlo Markov Chain (MCMC) which in effect estimated the posterior probability of the model given the data; our interpretation of the results of modeling is based on the highest posterior density interval (HDI), where we take a parameter to be meaningful if the 95% HDI for a given coefficient of the model does not include zero, and we take the mode of the HDI to be the estimate of the effect size. We examined measures of the convergence of the model such as the autocorrelation to assess model quality (Fig 3b).

A total of 1112 ESEs were modeled in the ten tissues that had shown the highest number of significant ESEs in the above analysis. 87% of sex-biased ESEs correlated with the expression of at least one RBP, and 67% of ESEs correlated with gene expression. In 40% of ESEs, our model indicated the existence of additional factors correlated with sex differences that were not captured by RBP mRNA levels or gene expression (in 0.5% of events, a sex effect was predicted in the absence of RBP or expression effects). Figure 3C shows the results of modeling an SE event in the *CDKN2A* gene, showing both differential expression and differential alternative splicing. Overall, females (shown as orange points) display higher expression of the *CDK2NA* gene (z-axis). As the expression level goes up, there is a slight tendency to more inclusion reads, but a marked tendency to more reads with exon skipping (Fig. 3c). The observed coefficients of RBPs were predicted to be entirely positive or entirely negative for some tissues, but often was positive for some targets and negative for others, suggesting the possibility of context-dependent effects of RBPs (Fig. 3d).

Our model estimated the effects of RBPs and gene expression on SE events in equal numbers of genes previously flagged as having sex-biased SE events or not. The distribution of mean absolute value of the estimated coefficients was significantly lower in the non-sex-biased events (TODO t test?). We interpret this as indicative of the fact that the strongest RBP and gene expression driven differential effects within any given tissue are related to sex rather than to other unidentified factors operating within a given tissue cohort in the GTEx data (Fig 3e).

We then investigated the relationship between SE events, RBP levels, and gene expression in more detail. We plotted the sum (over the 87 RBPs) of the estimated coefficients for RBPs affecting exon inclusion against the mean coefficient for gene expression affecting exon inclusion. For left ventricle, 61 out of 100 sex-biased events in heart-LV, no effect of gene expression was produced (flat line at y=0.0), and for the remaining genes, a correlation with R2=0.35 (p-value 7.98⋅10-5) was detected (Fig 4a). This correlation suggests that in the absence of RBP regulation increased expression correlates with a decreased probability of skipping, and vice versa. A similar correlation was found mammary tissue, with R2=0.64 (p-value 3.6⋅10-12) (Fig 4b) See Supplementary Figures XYZ for other tissues.

To validate this observation with a larger set of events, we performed linear regression of inclusion counts against RBP levels for each of the 87 investigated RBPs. We then plotted the log fold change expression vs. log fold change inclusion for all sex-biased events in breast mammary tissue. As in Fig 4a**,** events can be divided into inclusion increasing with expression and skipping increasing with expression. Points are colored blue instead of black if there was at least one RBP that was associated with the SE event at FDR ≤ 0.05 in the linear regression. The portion of the graph in the upper right quadrant shows SE events in which more skipping in females is correlated with higher gene expression in females. All of these events were positively correlated with at least one RBP (blue points). The lower right quadrant shows SE events in which more inclusion in females is associated with higher expression in females. Many of these events are not associated with any RBP (black points). Our results suggest that one group of SE events shows a positive correlation between gene expression and exon skipping; in all cases, there was at least one significantly associated RBP. Another group of genes shows a positive correlation between gene expression and exon inclusion; intriguingly, all of the genes that failed to demonstrate a significant correlation with one or more RBP were in this group (Fig 4c). We observed similar findings in the left ventricle (Supplemental Fig. S4).

NMD is a translation-coupled mechanism that eliminates mRNAs containing premature translation-termination codons (PTCs). NMD can thus serve as a quality control mechanism to prevent the accumulation of abnormal truncated proteins that could be deleterious to the cell.[26](https://paperpile.com/c/Dj72Qo/8jPO) The NMD additionally regulates the abundance of a large number of naturally occuring cellular mRNAs by degrading PTC-containing splice variant transcripts.[27,28](https://paperpile.com/c/Dj72Qo/eHuG+A1Di) We therefore divided all isoforms of the genes harboring the SE events into isoforms that are predicted to trigger NMD because of the presence of a premature truncation codon (PTC), and isoforms that do not contain a PTC (which we refer to as non-NMD in the following). We first tested whether inclusion counts of SE events associated with at least one NMD isoform differed from those of events not associated with any NMD isoform. As expected, there was a highly significant lower expression in the NMD-associated events (Fig 4d). The remaining non-NMD associated isoforms showed a significant, nearly two-fold increase in the number of inclusion read counts, suggesting that NMD is not the only factor responsible for the observed correlation between gene expression and AS. We additionally tested this by using the results of a study that used knockdowns and rescues of the three NMD factors.[29](https://paperpile.com/c/Dj72Qo/IyxK) The transcripts that we classified as NMD-related both within out Bayesian analysis (Fig 4e) and our larger frequentist analysis showed to classify transcripts as NMD or non-NMD and obtained comparable results. Finally, we reasoned that exons associated with NMD isoforms might be less likely to code for protein domains if their primary function is the induction of NMD. Indeed, there was a significant depletion of domain annotations in NMD associated skipped exons and their flanking exons (Fig 4f).

|  |  |  |  |
| --- | --- | --- | --- |
| a)   |  |  | | --- | --- | |  |  | | b) |
| c) |  |
| **d)** | **e)** |
| **f)** |  |

**Figure 4 Gene expression and SE events. (a)**. Predicted effects of gene expression vs. RBP on exon inclusion in 100 sex-biased SE events. The Y axis shows the mean of the posterior of the coefficient that gene expression affects exon inclusion. Negative values favor skipping and positive values favor inclusion. The X axis shows the sum of the absolute values of the posterior of the coefficients of the 87 RBPs. The higher the value, the more the predicted effect on exon skipping. In the left frame it can be seen that for 61 out of 100 sex-biased events in left ventricle, no effect of gene expression was predicted (flat line at y=0.0). For the remaining genes there was a correlation with R2=0.35 (p-value 7.98⋅10-5). **(b)** similar correlation was found mammary tissue, with R2=0.64 (p-value 3.6⋅10-12) **(c)** log fold change expression vs. log fold change inclusion for all sex-biased events in breast mammary tissue. As in **panel a ,** events can be divided into inclusion increasing with expression and skipping increasing with expression. Blue dots correspond to events whose skip counts were significantly correlated with an RBP using linear regression. The red lines correspond to second order polynomial regression.  **(d)**. Boxplots of log2(mean inclusion counts+0.5) for events with at least one NMD isoform and events that do not have NMD isoforms, calculated for all expression regulated sex-biased events in the HBM in all the tissues that were modeled. **(e)** Boxplots of log2(mean inclusion counts+0.5) for events with at least one NMD isoform and events that do not have NMD isoforms, calculated using inclusion counts from Colombo et al. for all expression regulated sex-biased events in the HBM in all the tissues that were modeled (See supplementary figure S7 for the same plots with all our skipped exon events). **(f)** Depletion of domain annotations in NMD associated skipped exons and their flanking exons. Maser was used for downloading domain annotations associated with each event. The y-axis gives the log2-number of events. p-values were obtained using the hypergeometric density and BH-corrected for multiple testing.

Here, we have analyzed 8,944 samples from 532 individuals across 53 tissues from the GTEx resource to provide the first comprehensive map of sex-biased gene expression and AS. Our hierarchical Bayesian modeling approach uncovered widespread correlations between the levels of RBPs and overall gene expression and SE events. While the role of regulation by RBP recognition of cis-acting binding sites on AS is well known,[5,30,31](https://paperpile.com/c/Dj72Qo/CqG1+3V13+aIGR) the observation that SE events showing higher degrees of correlation with gene expression displayed lower correlation with RBPs was surprising. Our study showed in the two tissues with the highest counts of called SE events (breast and left ventricle) that only about 30% of SE events correlated with gene expression could be ascribed to NMD (these cases were associated with significantly lower gene expression, as expected). Speculatively, the remaining 70% of SE events not associated with NMD isoforms could be related to one or more mechanism coupling epigenetic marks, transcription and splicing,[25,32–39](https://paperpile.com/c/Dj72Qo/WM76+PTU7+YZGn+kSQC+fW5Y+0p3L+Jrft+577B+wpQs) One possible interpretation of our results on the correlation of gene expression with some ASEs is therefore that differential transcription factor activation of promoters has led to both differential gene expression and differential ES. Our investigation has shown pervasive correlations of RBP levels and gene expression on sex-biased AS events, but our hierarchical Bayesian modelling can be extended to examine large datasets to characterize molecular mechanisms underlying other systems.

**ONLINE METHODS**

**GTEx samples**.

FASTQ files as well as transcript per million (TPM) and read counts of 56,202 genes together with the corresponding GTEx sample attributes and phenotypes were downloaded from the most current release, GTEx Analysis V7 (dbGaP Accession phs000424.v7.p2) (<https://www.gtexportal.org/home/datasets>). An approval for use of the the raw GTEx RNA-seq FASTQ files was granted by to Database of Genotypes and Phenotypes (dbGaP).

**Alignment of RNA-seq data**

Analysis was performed on the Institute for Systems Biology Cancer Genomic Cloud (ISB-CGC), an NCI Data Commons Pilot program. Our objective was to construct a matrix of counts for each of a variety of splicing types as discovered and cataloged by the rMATS[40](https://paperpile.com/c/Dj72Qo/Uliq) program (version 3.2.5) for each of the samples available from the GTEx archive. A prerequisite to using the rMATS program is that all reads to be assessed in the matrix must be of the same length. Using the rMATS 3.2.5 version, FASTQ files from the GTEx project were trimmed using the included python script trimFastq.py, Files were trimmed to 48 base pairs, aligned to the Genome Reference Consortium Homo sapiens assembly version hg38 (GRCh38.p7) using hisat2,[41](https://paperpile.com/c/Dj72Qo/hHX5) and duplicates were removed using Picard (http://broadinstitute.github.io/picard/). In order to create a matrix of counts with rows containing unique junction identifiers for each of the splicing types and columns containing the unique GTEx sample identifiers, some modifications were made to the standard process of running the rMATS program. For each file, rMATS identifies specific alternative splicing events capturing skipped exons (SE), retention introns (RI), alternative 3' and 5’ splice sites (A3SS amd A5SS), and mutually exclusive exons (MXE). For each of these 5 different splicing types, rMATS creates two files, with one that contains read counts that span the splicing junctions only, and a second file that counts not only the reads spanning the defined splicing junctions but also the reads that are on target. Custom bash scripts were written to merge the data from individual samples into a single matrix for each of the sample types and for each of the AS types.

**Differential gene expression in male vs female tissues.**

For analysis of differential gene expression and alternative splicing, genes on the Y chromosome were excluded from the analysis. For screening low-expressed events, we require that a number of samples equal to half the size of the smallest study group (always females in this case) will have a count per million (cpm) of at least 1, and in addition that male and female samples will contribute equally to satisfying this bound, i.e. that each will contain at least one fourth of the minimal number of samples with cpm ≥1. Each of the 53 tissue categories was analyzed individually voom[16](https://paperpile.com/c/Dj72Qo/3WnC) to assess gene expression differential between male and female samples.

**Normalization of counts data for alternative splicing analysis**

We used the *Yet Another RNA Normalization* software pipeline (YARN)[42](https://paperpile.com/c/Dj72Qo/p6Pz) to look for samples that are likely to be mis-annotated. We applied the function checkMisAnnotation using chromosome Y genes as control genes, and removed the individual GTEX-11ILO from the dataset (similar to ref. [42](https://paperpile.com/c/Dj72Qo/p6Pz)). We followed the YARN preprocessing procedure for identifying GTEx tissues that can be combined in the differential splicing analysis, using the function checkTissuesToMerge. This function creates MDS plots that reveal similarities and differences between samples in a set of tissues. As the input for the MDS, we concatenated the matrices of the skip and inclusion isoform counts row-wise such that each data point in the MDS is count values of skip and inclusion isoforms in one sample, where a sample corresponds to an individual and tissue. In order to obtain a consistent criterion for merging, we created an MDS for each pair of samples in the following regions: brain, artery, esophagus, skin and fibroblasts, colon and adipose. We calculated the normalized distance () between samples of the same anatomical region defined as the intersample distance divided by the mean distance of all samples for each pairwise-MDS and AS event type. We then merged tissues where for all AS event types (Supplementary figure S1 ). In skin regions the merging was identical to that in the original YARN publication, and in brain regions our procedure further refined merges performed in YARN[42](https://paperpile.com/c/Dj72Qo/p6Pz) into finer subsets. For esophagus, we found that Esophagus - Gastroesophageal Junction and Esophagus - Muscularis can be merged. Similar to YARN[42](https://paperpile.com/c/Dj72Qo/p6Pz) , other regions were not found to be mergeable with respect to the 5 types of AS events investigated in our study.

**Characterization of alternative splicing events (ASEs)**

We used rMATS [40](https://paperpile.com/c/Dj72Qo/Uliq) to identify and count reads that correspond to 5 types of ASEs: (1) skipped exon - the skipping of a single exon in an isoform of the transcript. (2) mutually exclusive exons - two consecutive exons out of which only one is present in each isoform of the transcript. (3) retained intron - the retention of an intron in an isoform of the transcript. (4) alternative 5’ splice site - a different exon at a 5’ position in an isoform of the transcript. (5) alternative 3’ splice site - similar to the previous category, but at a 3’ position (See **Fig. 1b**). rMATs identifies these events from a GTF file of known transcripts using release 25 from gencode annotation for genes for GRCh38.p7. rMATs then counts the number of reads that agree with each of the two alternatives that the event describes. For example, for a skipped exon rMATs will count the reads that fall within splice junctions that connect the skipped exon to its neighboring exons, and the reads that fall within a splice junctions that connects the neighboring exons to each other. A matrix of event counts was generated for all samples according to tissue types; one matrix was generated for each of the 5 categories of ASE and was used for the downstream analysis.

**Statistical approach to differential splicing between males & females**

In order to be able to fit a linear model to the data we use the voom function from the R package limma[16,43](https://paperpile.com/c/Dj72Qo/dzOL+3WnC) to transform the counts into continuous data, appropriate for linear modeling. For each ASE, we combine skip and inclusion event counts as individual samples in a multifactorial linear model, where skip and inclusion counts from the same individual/sample are treated as replicate arrays in order to account for correlation. Limma uses generalized least squares to fit the model, which does not assume that the errors of different samples are independent. The multifactorial model has 3 predictors: sex (male or female), event (skip or inclusion) and a sex:event interaction term:



Events that have a significant sex:event interaction term (FDR<=0.05) and in addition a fold change of at least 1.5 for that term, are considered differentially spliced. The sex predictor accounts for the case where male or females have a higher level of both isoforms but the proportions in both sexes are the same. The event predictor accounts for the case where one event has more reads mapped to it, but not as a result of alternative splicing that is differential between the sexes. For example, if due to the fragmentation process of RNA-Seq more reads are mapped to the inclusion event, there will be a bias in both sexes towards this event. For normalization, we used the edgeR function calcNormFactors.[44](https://paperpile.com/c/Dj72Qo/4ml1) We screened lowly expressed events using the same criterion as for differential expression, where screening is applied to each of the following 4 groups: male inclusion counts, male skip counts, female inclusion counts and female skip counts.

**Definition of set of “interesting” RNA-binding proteins (RBPs)**

We retrieved 87 RNA-binding proteins with a defined position-specific scoring matrix (PSSM) from RBPMap[45](https://paperpile.com/c/Dj72Qo/aU9K) and defined this set as RNA-binding proteins (RBPs). The PSSM of RBPs were visualized in Supplemental Table 4 with the R package “seqLogo”.[46](https://paperpile.com/c/Dj72Qo/iGZp)

**Clustering Analysis**

The log-fold-changes obtained from voom’s differential expression analysis were obtained for each tissues, where genes that were screened out from the DE analysis of a tissue were assigned a log-fold-change of 0. Genes with a logFC of 0 in all tissues were removed, and then the pearson correlation between each pair of tissues was calculated based on the logFC vectors. The vectors of correlations were clustered using hierarchical clustering (Fig. 1b).

**Gene Ontology analysis**

Sets of significantly differentially expressed or spliced genes were obtained for each tissue and analyzed with the model-based gene set analysis procedure in the Ontologizer.[47,48](https://paperpile.com/c/Dj72Qo/LMgp+6W5x) The population set was defined to be the set of all annotated human genes using the Human GO Annotation EBI (<http://geneontology.org/page/download-go-annotations>) that contains 19,712 gene product annotation for the association of genes to GO.

**Hierarchical Bayesian Modeling**

Hierarchical Bayesian modeling (HBM) is a technique for multiparameter modeling in which one assumes a statistical distribution for individual parameters and whose interdependencies are reflected in the structure of the hierarchy. The HBM can use a Markov Chain Monte-Carlo technique to estimate the posterior probability of each parameter. To apply HBM, one must design the structure of the hierarchy and define the probability distribution of each node. The HBM procedure can then use MCMC to estimate the posterior probability distribution at each node. If the 95% high density interval (also called the credibility interval) for a coefficient does not contain zero, then we assume that the corresponding parameter is relevant for the model.

Our assumptions in developing our model were: (i) RNA binding proteins can affect the probability of exon inclusion either negatively or positively; (ii) The effect of each RBP will tend to be consistent across all genes in a given tissue; (iii) the overall expression of a gene can affect the probability of inclusion of an exon of the gene either negatively or positively; (iv) additional sex-related effects may exist that are not captured by RBPs or gene expression. (v) The effect of RBPs, gene expression, and sex is additive (vi) there is a prior assumption of no effect of any of the above mentioned factors (in the absence of evidence against the prior).

We chose up to 100 statistically significant sex-biased ASEs for each tissue together with an equal amount of non-sex-biased ASEs that were chosen as the ASEs showing the highest variance (summed over exclusion and inclusion counts) among all samples regardless of sex.

The observed number of skip counts at event i in sample j (Sij) is modeled as:



where Pij is the probability of skipping at event i in sample j and Iij is the number of inclusion counts at event i in sample j.

The probability Pij of skipping at event i in sample j is:



where is the logistic function, the parameter is an intercept for the ith event, is the effect of sex, is the effect of the expression level of gene g(i) to which the ith event belongs and is a vector of RBP effects on the ith event. Sex takes a value from {0,1}, and the expression level of the gene and the RBPs are normalized to mean 0 and standard deviation 1 over all the samples.

The prior of , and are N(0,), the priors for the effects of the kth RBP on dimorphic events, , are N(, and the prior of is N(0,) , , =1. This reflects our prior knowledge as described in assumption (ii) above. A similar hyperparameter is defined for the common effect of an RBP on all the non-dimorphic events.

We ran the scripts using the R-Stan interface of rstan.[49](https://paperpile.com/c/Dj72Qo/6zf5) The number of chain was set to 3, the number of iterations to 5,000, the number of warmup iterations to 3,000, the thinning parameter was set to 1 and all parameter values were initialized to 0. Each chain was run on a different processor in order to improve performance. The script is called “**AS\_network\_most\_varying\_events\_more\_sigs.R**” in the scripts directory in the github.

**Statistical approach to correlation of gene expression and RBP levels with differential splicing between males & females**

We obtained the log fold-change values computed by limma for differentially expressed genes and significant alternative splicing events in those genes, and plotted them against each other. A second order polynomial regression line was fit separately to points corresponding to positive AS fold change (more skipping in females) and negative AS fold change (more skippind in males). A linear regression was performed between the skip counts of each AS event (dependent variable) and the TPM of each RBP in supplementary table S5. The p-values obtained for each RBP were BH-corrected for multiple testing, and a significant correlation was associated with an FDR0.05

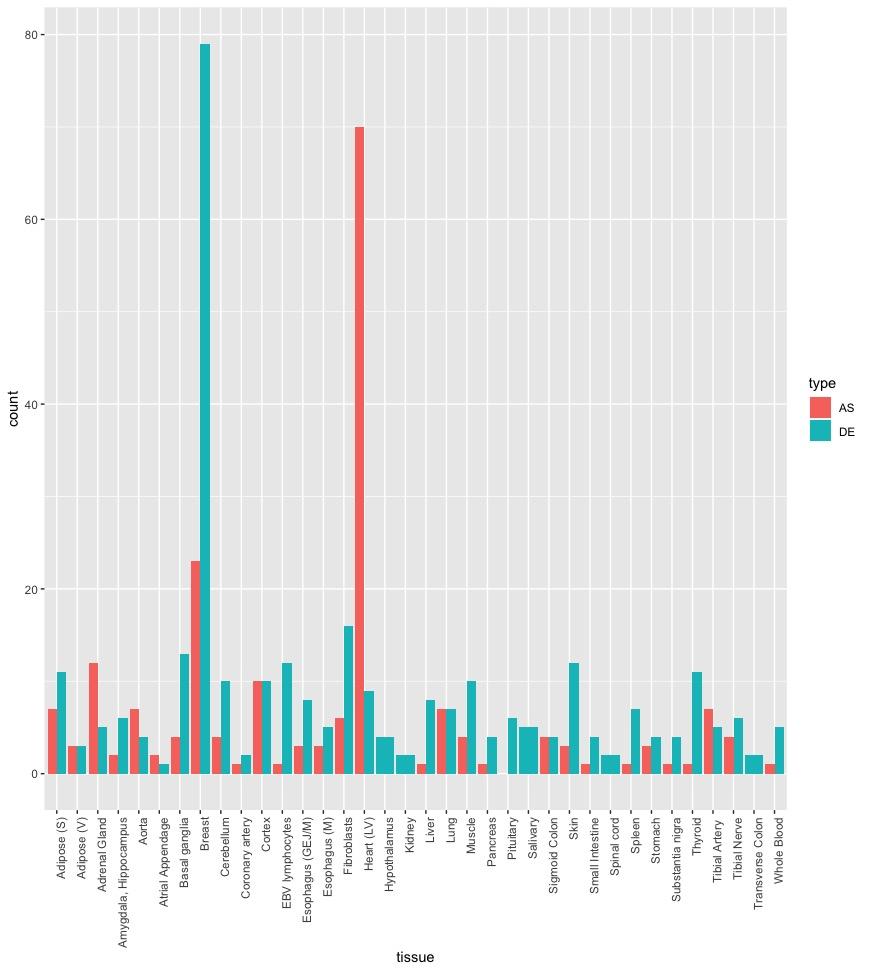
Figure N: **Effect of expression in the hierarchical Bayesian model for sexually dimorphic events and for the most varying events.** The box plot displays mode of the posterior for each event in each one of the tissues. Positive values correspond to increasing skipping, and negative values to increasing inclusion.

|  |  |
| --- | --- |
|  |  |

Figure N: **Effect of sex in the hierarchical Bayesian model for sexually dimorphic events and for the most varying events.** The box plot displays mode of the posterior for each event in each one of the tissues. Positive values correspond to increasing skipping, and negative values to increasing inclusion.

|  |  |
| --- | --- |
|  |  |

Supplementary Figure N: The number of chromosome X genes that were differentially expressed or contained at least one sex-biased AS events. For differential expression, the counts of tissues that were summed in order to match the tissue sets of alternative splicing.

****

**References**

1. [Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–463 (2010).](http://paperpile.com/b/Dj72Qo/Mym2)

2. [Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18,** 437–451 (2017).](http://paperpile.com/b/Dj72Qo/PcQx)

3. [Poulos, M. G., Batra, R., Charizanis, K. & Swanson, M. S. Developments in RNA splicing and disease. *Cold Spring Harb. Perspect. Biol.* **3,** a000778 (2011).](http://paperpile.com/b/Dj72Qo/mhuQ)

4. [Wang, K. *et al.* Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Sci. Rep.* **8,** 10929 (2018).](http://paperpile.com/b/Dj72Qo/Ky2A)

5. [Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84,** 291–323 (2015).](http://paperpile.com/b/Dj72Qo/3V13)

6. [Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16,** 990–995 (2009).](http://paperpile.com/b/Dj72Qo/0KEo)

7. [Zhou, H.-L., Luo, G., Wise, J. A. & Lou, H. Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* **42,** 701–713 (2014).](http://paperpile.com/b/Dj72Qo/LfY1)

8. [Rinn, J. L. & Snyder, M. Sexual dimorphism in mammalian gene expression. *Trends Genet.* **21,** 298–305 (2005).](http://paperpile.com/b/Dj72Qo/zNFm)

9. [Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4,** 2771 (2013).](http://paperpile.com/b/Dj72Qo/K0Ju)

10. [GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).](http://paperpile.com/b/Dj72Qo/d8KX)

11. [Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).](http://paperpile.com/b/Dj72Qo/Ehmb)

12. [GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).](http://paperpile.com/b/Dj72Qo/F9BY)

13. [Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550,** 239–243 (2017).](http://paperpile.com/b/Dj72Qo/zP2T)

14. [Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49,** 692–699 (2017).](http://paperpile.com/b/Dj72Qo/a1iT)

15. [Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27,** 1843–1858 (2017).](http://paperpile.com/b/Dj72Qo/NoUj)

16. [Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).](http://paperpile.com/b/Dj72Qo/3WnC)

17. [Li, Y. *et al.* Cell sex affects extracellular matrix protein expression and proliferation of smooth muscle progenitor cells derived from human pluripotent stem cells. *Stem Cell Res. Ther.* **8,** 156 (2017).](http://paperpile.com/b/Dj72Qo/UqaT)

18. [Welle, S., Tawil, R. & Thornton, C. A. Sex-related differences in gene expression in human skeletal muscle. *PLoS One* **3,** e1385 (2008).](http://paperpile.com/b/Dj72Qo/f8iQ)

19. [Zhang, C., Zhang, B., Lin, L.-L. & Zhao, S. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* **18,** 583 (2017).](http://paperpile.com/b/Dj72Qo/umkd)

20. [Kanitz, A. *et al.* Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol.* **16,** 150 (2015).](http://paperpile.com/b/Dj72Qo/r9X5)

21. [Kratz, A. & Carninci, P. The devil in the details of RNA-seq. *Nat. Biotechnol.* **32,** 882–884 (2014).](http://paperpile.com/b/Dj72Qo/Abpt)

22. [SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32,** 903–914 (2014).](http://paperpile.com/b/Dj72Qo/FlJE)

23. [Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550,** 244–248 (2017).](http://paperpile.com/b/Dj72Qo/hAfn)

24. [Sjöblom, B., Salmazo, A. & Djinović-Carugo, K. Alpha-actinin structure and regulation. *Cell. Mol. Life Sci.* **65,** 2688–2701 (2008).](http://paperpile.com/b/Dj72Qo/Ptux)

25. [Cohen, E., Zafrir, Z. & Tuller, T. A code for transcription elongation speed. *RNA Biol.* **15,** 81–94 (2018).](http://paperpile.com/b/Dj72Qo/WM76)

26. [Brogna, S. & Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* **16,** 107–113 (2009).](http://paperpile.com/b/Dj72Qo/8jPO)

27. [Saltzman, A. L. *et al.* Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. *Mol. Cell. Biol.* **28,** 4320–4330 (2008).](http://paperpile.com/b/Dj72Qo/eHuG)

28. [Baek, D. & Green, P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 12813–12818 (2005).](http://paperpile.com/b/Dj72Qo/A1Di)

29. [Colombo, M., Karousis, E. D., Bourquin, J., Bruggmann, R. & Mühlemann, O. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. *RNA* **23,** 189–201 (2017).](http://paperpile.com/b/Dj72Qo/IyxK)

30. [Fu, X.-D. & Ares, M., Jr. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15,** 689–701 (2014).](http://paperpile.com/b/Dj72Qo/CqG1)

31. [Fredericks, A. M., Cygan, K. J., Brown, B. A. & Fairbrother, W. G. RNA-Binding Proteins: Splicing Factors and Disease. *Biomolecules* **5,** 893–909 (2015).](http://paperpile.com/b/Dj72Qo/aIGR)

32. [Cramer, P. *et al.* Coupling of transcription with alternative splicing: RNA pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Mol. Cell* **4,** 251–258 (1999).](http://paperpile.com/b/Dj72Qo/PTU7)

33. [Kornblihtt, A. R. Promoter usage and alternative splicing. *Curr. Opin. Cell Biol.* **17,** 262–268 (2005).](http://paperpile.com/b/Dj72Qo/YZGn)

34. [de Matos, P. *et al.* Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.* **38,** D249–54 (2010).](http://paperpile.com/b/Dj72Qo/kSQC)

35. [Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144,** 16–26 (2011).](http://paperpile.com/b/Dj72Qo/fW5Y)

36. [Kim, Y.-E., Park, C., Kim, K. E. & Kim, K. K. Histone and RNA-binding protein interaction creates crosstalk network for regulation of alternative splicing. *Biochem. Biophys. Res. Commun.* **499,** 30–36 (2018).](http://paperpile.com/b/Dj72Qo/0p3L)

37. [Li, T., Liu, Q., Garza, N., Kornblau, S. & Jin, V. X. Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing. *Genome Med.* **10,** 30 (2018).](http://paperpile.com/b/Dj72Qo/Jrft)

38. [Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19,** 1732–1741 (2009).](http://paperpile.com/b/Dj72Qo/577B)

39. [Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat. Genet.* **41,** 376–381 (2009).](http://paperpile.com/b/Dj72Qo/wpQs)

40. [Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E5593–601 (2014).](http://paperpile.com/b/Dj72Qo/Uliq)

41. [Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).](http://paperpile.com/b/Dj72Qo/hHX5)

42. [Paulson, J. N. *et al.* Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics* **18,** 437 (2017).](http://paperpile.com/b/Dj72Qo/p6Pz)

43. [Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).](http://paperpile.com/b/Dj72Qo/dzOL)

44. [Lun, A. T. L. & Smyth, G. K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16,** 258 (2015).](http://paperpile.com/b/Dj72Qo/4ml1)

45. [Paz, I., Kosti, I., Ares, M., Jr, Cline, M. & Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **42,** W361–7 (2014).](http://paperpile.com/b/Dj72Qo/aU9K)

46. [Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).](http://paperpile.com/b/Dj72Qo/iGZp)

47. [Bauer, S., Gagneur, J. & Robinson, P. N. GOing Bayesian: model-based gene set analysis of genome-scale data. *Nucleic Acids Res.* **38,** 3523–3532 (2010).](http://paperpile.com/b/Dj72Qo/LMgp)

48. [Bauer, S., Grossmann, S., Vingron, M. & Robinson, P. N. Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24,** 1650 (2008).](http://paperpile.com/b/Dj72Qo/6W5x)

49. [Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles* **76,** 1–32 (2017).](http://paperpile.com/b/Dj72Qo/6zf5)