The impact of sex on alternative splicing

Guy Karlebach,[1] Diogo F.T. Veiga,[1] Anne Deslattes Mays,[1] Anil K. Kesarwani,[1] Daniel Danis,[1] Georgios Kararigas,[2,3] Xingmin Aaron Zhang,[1] Joshy George,[1] Guru Ananda,[1] Robin Steinhaus,[2] Peter Hansen,[2] Dominik Seelow,[2] Chris Bizon,[4] Rebecca Boyles,[5] Chris Ball,[5] Julie A McMurry,[6] Melissa A Haendel,[6] Jeremy Yang,[7] Tudor Oprea,[7] Mitali Mukerji,[8] Olga Anczukow,[1] Jacques Banchereau,[1] Peter N Robinson[1,9,*]

1. The Jackson Laboratory for Genomic Medicine, Farmington CT 06032, USA
2. Charité-Universitätsmedizin Berlin, Charitéplatz 1, 10115 Berlin, Germany
3. DZHK (German Centre for Cardiovascular Research), partner site Berlin, Germany
4. Renaissance Computing Institute (RENCI), 100 Europa Drive, Suite 540, Chapel Hill, NC 27517, United States.
5. RTI International, Durham, NC 27709, USA
6. Linus Pauling institute, Oregon State University, Corvallis OR, USA
7. Translational Informatics Division, Department of Internal Medicine, The University of New Mexico Health Science Center, Albuquerque, NM 87131, USA.
8. Genomics and Molecular Medicine, CSIR-Institute of Genomics and Integrative Biology, Mathura Road (CSIR-IGIB), Mathura Road, New Delhi 110 025, India. E-mail: mitali@igib.res.in, 9. Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA
*) correspondence to Peter N. Robinson, The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington CT 06032, USA. peter.robinson@jax.org.

Abstract

Over 95% of human genes undergo alternative splicing (AS) in a developmental, tissue-specific, or signal transduction-dependent manner. A number of factors including binding of cis-acting sequences by RNA-binding proteins (RBPs) are known to affect AS, but the combinatorial mechanisms leading to the distribution of spliced isoforms remain largely unstudied. Here, in 9011 samples from 532 individuals across 53 tissues from the Genotype-Tissue Expression (GTEx) resource, we identified 4,135 genes with sex-biased expression and 5,925 sex-biased AS events. We find that factors including escape from X-chromosomal inactivation, presence of Alu elements, and oestrogen receptor binding sites affect sex-biased AS. We utilize hierarchical Bayesian modelling to characterize the interactions of exon skipping, gene expression, and RBPs, and demonstrate two categories of sex-biased AS that differ with respect to splice site scores, gene expression, RBP levels, and skipping/inclusion ratio.

Introduction

Alternative splicing (AS), a process by which splice sites are used differentially to create protein diversity, plays an important role in development,[1] disease,[2] and aging.[3] Although some splicing isoforms are produced in the same proportions in all or most cell types, AS is often regulated by developmental or differential cues or in response to external stimuli.[4] Several mechanisms have been demonstrated to regulate AS, although their combinatorial interactions remain poorly understood. Binding of RNA-binding proteins (RBPs) to intronic or exonic cis-acting regulatory sequences may promote or suppress local AS events.[5] Additionally, chromatin-level mechanisms also play a role in AS regulation. Nucleosome density is higher within exons than in introns, suggesting the existence of RNA polymerase II (RNA Pol II)-mediated cross-talk between chromatin structure and exon-intron architecture.[6] 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 influence exon skipping.[7] Additionally, specific histone modifications that can be enriched over exons may promote binding of proteins such as HP1α and HP1γ that in turn influence transcriptional speed (fig 1a).[8]

Although sex-biased gene expression is common,[9] and widespread differences in AS have been identified in the human brain,[10] no analysis has been performed to date over a comprehensive dataset that spans multiple tissue types. Here, we perform a systematic survey of sex-biased AS across multiple tissues using a systems biology approach to characterize RBP levels and gene expression and their interplay (Fig. 1b). The Genotype-Tissue Expression (GTEx) project comprises samples from 53 non-diseased tissue sites assayed by whole genome or exome sequencing, and RNA-Seq.[11–16] In this study, we leveraged the GTEx data to investigate gene expression and AS in male and female subjects. We analysed gene expression and AS in 9,011 samples from 532 individuals across 53 tissues (Supplemental Data Tables 1 and 2) with the goal of characterizing sex-specific patterns of gene expression, AS, and abundance of RBPs in multiple tissue types.

Sex-biased differences in gene expression

A total of 4,135 genes with significantly sex-biased expression were detected at a false-discovery rate (FDR) cut-off of 0.05 and a fold change cut-off of 1.5 (Methods; Supplementary File 1). The tissue that had the largest number of genes showing significant differential expression (DE) was breast, followed by thyroid, skin, and adipose-subcutaneous (Supplemental Data Fig. 1). Figure 1b displays a heatplot of a hierarchical clustering of the different tissues, where tissues are clustered by similarity of the mean fold changes of gene expression between male and female sample. The clustering analysis revealed a consistent shift in global expression patterns between males and females in some related tissues, such as 11 brain regions (top left), as well as three arterial tissues, two oesophageal tissues, and sun-exposed and non-sun-exposed skin (Fig. 1b).

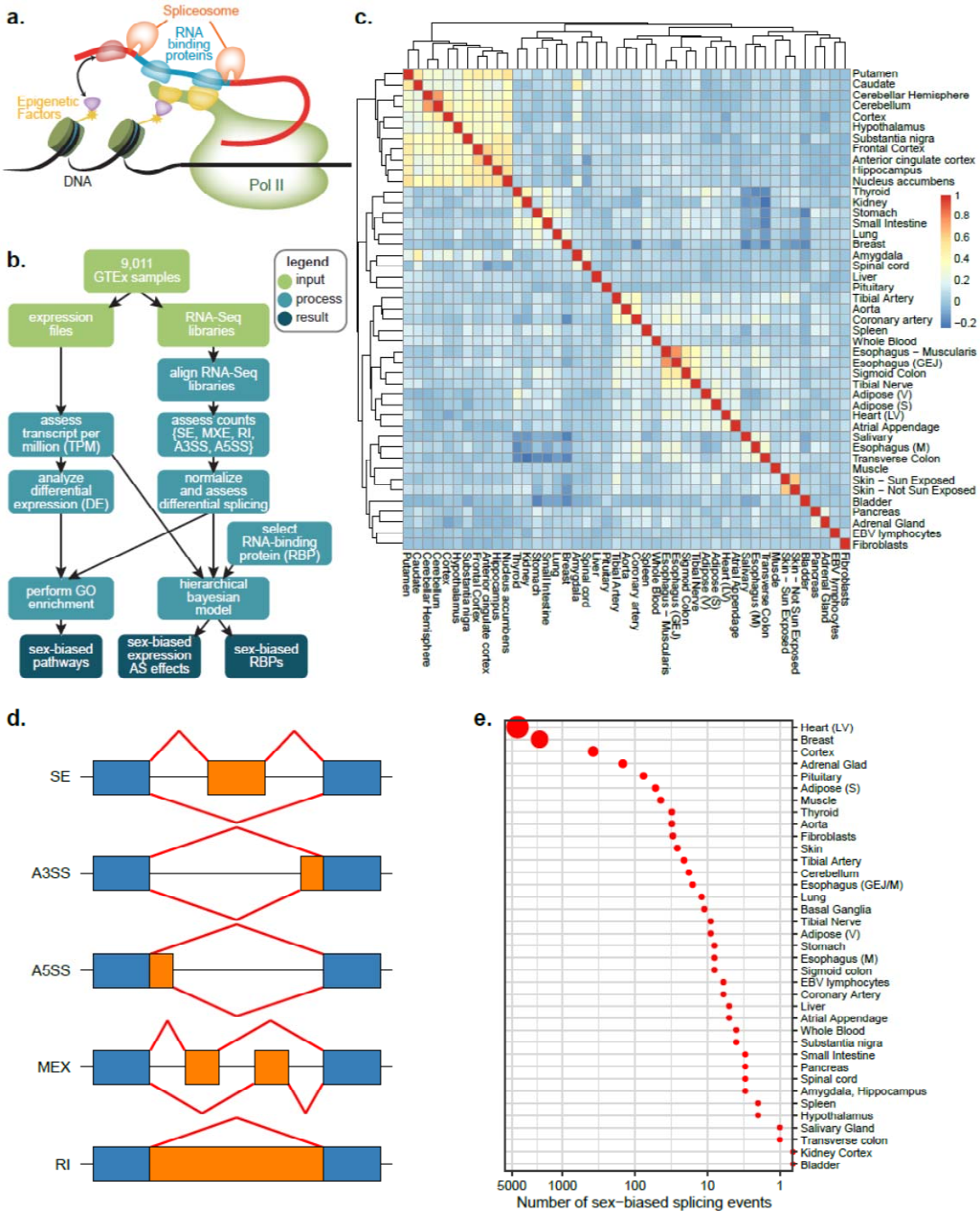

Figure 1. Sex-biased gene expression and alternative splicing. (a) Pre-messenger RNA splicing occurs co-transcriptionally and is influenced by RNA binding proteins and epigenetic factors such as histone modifications that interact with the transcriptional machinery or other proteins to influence splicing and transcription.[17] **(b)** Flowchart depicting the analysis of GTEx RNA-seq data. Analysis of GTEx gene expression and AS profiles identified significantly sex-biased genes and AS events. Data were used as input for a hierarchical Bayesian model to characterize the influence of RBPs and gene expression on sex-biased AS events. **(c)** 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. **(d)** The five categories of AS events that were investigated in this work (SE: skipped exon/exon inclusion; A3SS/A5SS: alternative 3'/5' splice site; MEX: mutually exclusive exons; RI: retained intron). **(e)** Number of sex-biased AS events per tissue type.

Gene Ontology (GO) analysis of genes with significant sex-biased differential expression identified 79 significantly enriched terms including 15 terms enriched in two or more tissues (Supplemental Data Table 3). A number of the GO terms could reflect known sex differences, such as the enrichment of *extracellular matrix* (ECM) in breast tissue, which is known to display sex-biased differential expression of multiple ECM proteins,[18] 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.[19]

Sex-biased differences in alternative splicing

We investigated 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.[20–23] We focused on five classes of discrete AS events (Fig. 1c) and defined sex-biased AS based on a statistical model with sex, AS event, and sex:event interaction as covariates. We called AS events sex-biased if the interaction term was significant following multiple testing correction (Methods, Supplemental Data Figure 2).

For the AS analysis, the 53 tissue types were consolidated into 46 groups by merging samples with highly similar distributions of skipping and inclusion counts (Methods; Supplemental Data Table 2). Statistical analysis revealed between 0 and 2724 genes with at least one significant AS event per tissue. The total number of AS events over all tissues was 5925 (Supplementary File 1). The overall count of sex-biased AS events was strikingly different in different tissue types. We identified four tissues with over 100 AS events, 12 tissues with between 10 and 100 AS events, and 21 tissues with less than 10 AS events (Fig. 1d). In our analysis of the left ventricle, there were over 50 times more statistically significant AS events than significantly differentially expressed genes, suggesting the importance of investigating sex-biased AS and not just sex-biased gene expression as potential contributory mechanisms to the pronounced sex differences in cardiac physiology and heart disease.[24]

We then tested whether sex-biased differential expression and AS occurs independently. The overlap between these two groups contained 836 genes, which is significantly more than expected by chance (p=1.55×10^{-93}, hypergeometric test). It has been reported that differentially expressed sex-biased genes are likely to be linked to escape from X chromosome inactivation.[25] We confirmed this result with our data (p = 5.47 × 10^{-50}, Fisher's exact test). Because of the observed overlap between sex-biased differential expression and AS, 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 amongst genes displaying sex-biased AS (P = 2.46 × 10^{-20}, Fisher's exact test; Figure 2a).

Characterization of exons showing sex-biased alternative splicing events

Alu elements are primate-specific repeats and comprise 11% of the human genome, with over 1 million occurrences in the human genome. Thousands of human genes contain spliced exons derived from Alu elements, and some exert various effects on gene regulation and AS.[26] We therefore hypothesized that some of the sex-biased AS events might be associated with Alu elements. Since Alu elements occur with a particularly high density in introns and intronic Alu elements can influence alternative splicing,[27] we tested for enrichment of Alu elements in the introns flanking exons that displayed sex-biased skipping events, comparing the occurrence in sex-biased skipping events with the occurrence in all skipping events in our dataset. Twelve Alu

subfamilies were found to be enriched using an FDR threshold of 0.05 (Supplemental Data Table 3). Alu elements contain numerous binding sites for transcription factors that may partially mediate their effects on gene regulation, and the AluSp family (which showed the most significant enrichment in our analysis) has been noted to be enriched for a predicted oestrogen receptor (ER) binding site,[28,29] We therefore compared the distribution of predicted ER binding sites in all AS events with those of the sex-biased events, and found that AluSx (FDR 2.32×10^{-6}) and AluSp (FDR 0.039) showed significant enrichment (Supplemental Data Table 4). Speculatively, oestrogen-Alu interactions could contribute to the observed sex bias in these AS events.

The X chromosome also showed the highest normalized number of sex-biased AS events per exon. We defined a sex-biased splicing index (SBSI) as the number of statistically significant AS events per 1000 exons, and calculated the SBSI for each chromosome (excluding the Y chromosome). The X chromosome had by far the highest SBSI with 12.15 per 1000 exons showing sex-biased AS events, with most of the remaining chromosomes having an SBSI of between 5 and 7.5 (Fig. 2b). Most of the AS events were specific to one tissue, but slightly over 12% were found in 2-5 tissues, with only 9 AS events being found in more than 5 tissues (Fig. 2c). 27 genes were found to have >10 AS events (Fig. 2d). 8 of these genes (30%) were X chromosomal. 16 of the 27 genes had various roles in signalling and gene regulation, including 2 genes involved in X inactivation, 5 in chromatin modification, 13 in signal transduction, and 7 in transcriptional regulation (Supplemental Data Table 6). We performed GO analysis on all genes harbouring one or more AS events. A total of 54 distinct GO terms were significantly enriched in one or more tissues, reflecting a wide range of biological processes. The term *translation initiation factor activity* was the most commonly observed term with 12 tissues (Supplemental Data Table 7). The most common type of AS event was exon skipping in all tissues except the pituitary, in which intron retention was the most common. Oesophagus, lung, and basal ganglia showed a relatively high proportion of A3'SS and A5'SS events, and cerebellum displayed the highest frequency of mutually exclusive exons (Fig. 2e).

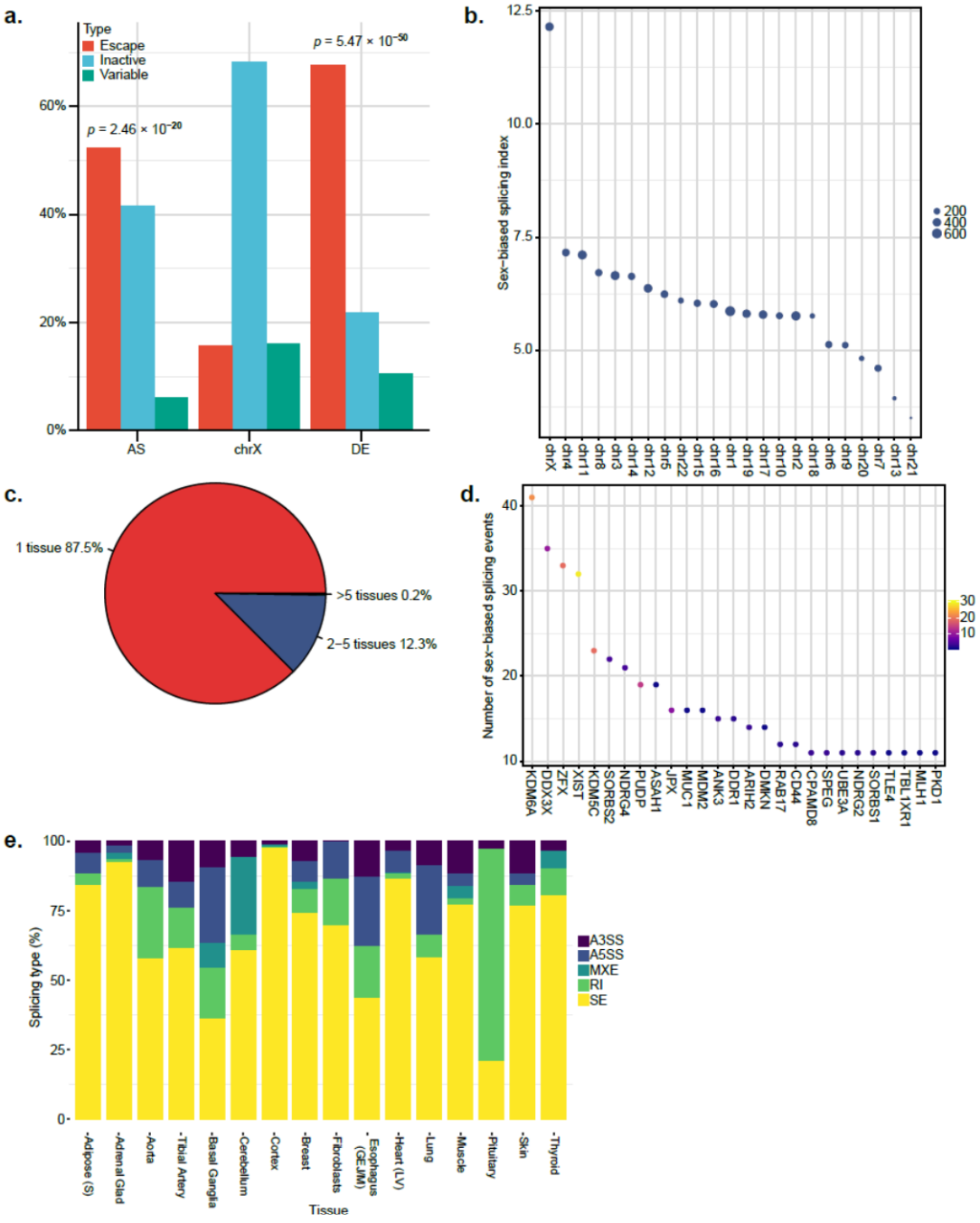

Figure 2. Characteristics of sex-biased alternative splicing. (a) Percentage of X chromosomal genes showing inactivation, escape from inactivation, or variable escape.[25] Percentage is shown for the entire set of 631 investigated X chromosomal genes, as well as the subsets of 248 genes with differential expression (DE) and 166 with at least one AS event. P-values were derived with Fisher's exact test comparing genes with DE or AS to the entire set of 631 genes. **(b)** Sex-biased alternative splicing index for each chromosome. **(c)** Proportions of AS events that were strictly tissue-specific, present in 2-5, or over 5 tissues. **(d)** Genes with over 10 significant sex-biased AS events. The colour coding represents the number of affected tissues. **(e)** Distribution of categories of AS events that were significantly different in males and females (tissues with > 10 AS events).

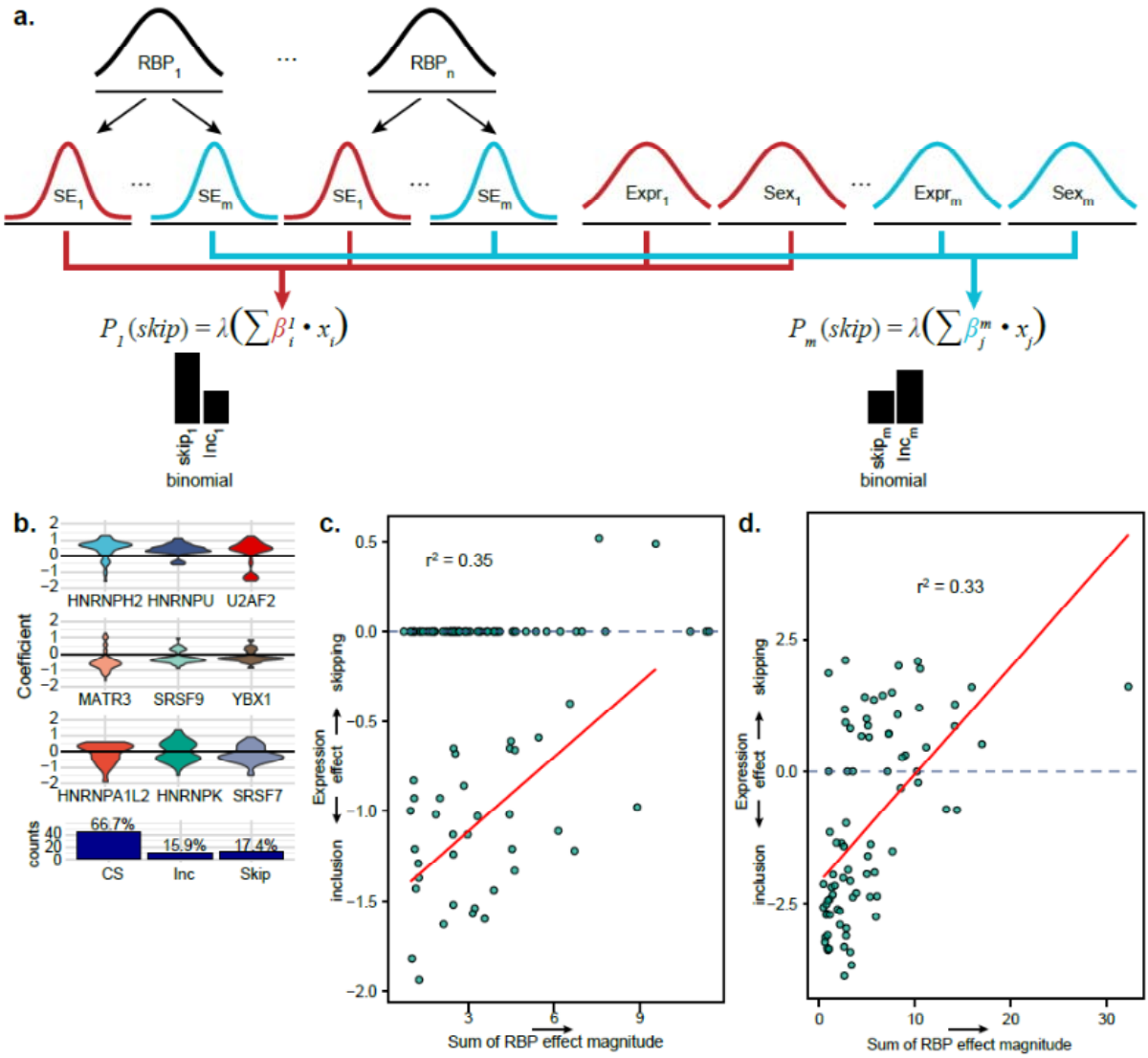

Figure 3. **Hierarchical Bayesian Modelling**. **(a)** Structure of the HBM used in this work. The model posits that RBPs can affect the probability of exon skipping (shown as SE_1,...,SE_m) either negatively or positively but that the effect of each RBP will tend to be consistent across all genes in a given tissue (each one of the RBP_1,...,RBP_n distributions acts as a prior for the effect of the RBP at each exon skipping event). Furthermore, the overall expression of a gene *Exprj* can affect the probability of skipping of an exon of the gene either negatively or positively (the expression node), and additional sex-related effects may exist that are not captured by RBPs or gene expression (Sex_j). All effects are combined in a logistic function λ that models the probability of exon skipping. The MCMC algorithm generates accurate and stable estimates of the model parameters, which are then exploited by our subsequent analysis. **(b)** Average RBP effects for nine exemplary RBPs: 3 that tend to promote skipping (top row), 3 that tend to promote inclusion (middle row) and 3 that are context specific (bottom row). A positive 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 coefficient indicates a negative correlation. The barplot at the bottom of the figure displays the percentage of RBPs from each type among all the RBPs in the study whose absolute sum of effects was in the 0.8 quantile. **(c)** Predicted effects of gene expression vs. RBP levels on exon inclusion in 100 sex-biased SE events in the left ventricle. The Y axis shows the mean of the posterior of the coefficient that determines the effects of gene expression on exon inclusion. Negative values favour skipping and positive values favour 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 R^2=0.35 (p=7.98x10^{-5}). **(d)** A similar correlation was found in mammary tissue, with R^2=0.33 (p=3.6x10^{-12}).

Interactions of exon skipping, gene expression, and RNA binding proteins

We hypothesized that differential levels of RBPs with roles in AS could be responsible for some of the AS events. Furthermore, because of recent reports that transcription elongation rate can affect gene expression,[30] and the observation that faster transcriptional elongation speed can lead to increased exon skipping,[8] 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 events. If available, the top 100 significant sex-biased were chosen for each tissue; otherwise, if fewer than 100 events had been called significant, the corresponding smaller number of events was 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, *i.e.*, counts of exon inclusion and exclusion reads, was modelled as a result of a weighted linear combination of the individual mRNA levels of 87 RBPs (Supplemental File 2) as well as of the overall expression level of the gene harbouring the exon-skipping event. Furthermore, a 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 (Fig 3a). The modelling process makes use of Monte-Carlo Markov Chain (MCMC) sampling in order to estimate the posterior probability of the model given the data; our interpretation of the results of modelling is based on the highest posterior density interval (HDI), whereby 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. If the HDI for some parameter included zero, our interpretation is that the parameter is not relevant for determining exon skipping (Supplemental Data Figure 4). We examined measures of the convergence of the model such as the autocorrelation to assess model quality, and assessed the 95% HDI for all of the parameters in the model.

A total of 556 SE events were modelled in the ten tissues that had shown the highest number of significant SE events in the above analysis. 87% of sex-biased exon-skipping events correlated with the expression of at least one RBP, and 67% of exon-skipping events correlated with gene expression. In 40% of exon-skipping events, our model indicated the existence of additional factors correlated with sex differences that were not captured by RBP mRNA levels or gene expression, and in 0.5% of events, a sex effect was predicted in the absence of RBP or expression effects (Supplemental File 3). We analysed the tendency of RBPs to promote exon skipping or inclusion by examining the distribution of coefficients for all SE events for which the interaction was predicted to be meaningful by the Bayesian model. Roughly 17% of the RBPs were classified as exon skipping-promoting factors because at least 75% of the distribution was positive, and likewise roughly 16% of the RBPs were classified as exon inclusion-promoting factors. The remaining 67% were classified as context dependent (Fig. 3b). Histograms of the coefficients determined by our model for the effects of RBPs, gene expression, and the residual effect of sex are shown in Supplemental Data Figure 3.

We then investigated the relationship between exon-skipping events, RBP levels, and gene expression in more detail. We plotted the sum (over the 87 RBPs) of the absolute 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 AS events showed no effect of gene

expression (flat line at y=0.0), and for the remaining genes, a correlation with R^2=0.35 (p=$7.98 \cdot 10^{-5}$) was detected (Fig 3c). That is, for the AS events in which gene expression tended to promoter exon inclusion (negative values on the X axis), the predicted effect of RBPs was lower. A similar correlation was found in mammary tissue, with R^2=0.33 (p=$3.6 \cdot 10^{-12}$) (Fig 3d). The overall tendency for the influence of RBPs to increase as the influence of gene expression on exon inclusion decreased was similar for both tissues.

Sex-biased alternative splicing and Nonsense-Mediated Decay

Nonsense-mediated decay (NMD) is a translation-coupled mechanism that eliminates mRNAs containing premature 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.[31] NMD additionally regulates the abundance of a large number of naturally occurring cellular mRNAs by degrading PTC-containing AS transcripts.[32,33] In order to investigate a potential role of such physiological NMD in sex-biased AS, we divided all isoforms of the genes harbouring the SE events into isoforms that are predicted to trigger NMD because of the presence of a PTC, and isoforms that do not contain a PTC (which we refer to as non-NMD in the following; Methods). 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.

Of the 5568 significantly sex-biased SE events in our dataset, 1284 were predicted to be associated with at least one NMD isoform (23%), as compared to 39180 total SE events (significant or not), in which 9719 corresponded to NMD isoforms (24.8%). As expected, there was a significantly lower inclusion count in the NMD-associated events (Fig 4a). The remaining non-NMD associated isoforms showed a significant, nearly two-fold increase in the number of inclusion read counts. 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 4b). Fig. 3c shows an exemplary exon-skipping event in the *CDKN2A* gene in which there was both higher gene expression and higher exon-skipping levels in females compared to males (Fig. 4c). The skipped exon contains a PTC and is predicted to induce NMD. This finding is therefore compatible with NMD-induced down-regulation of *CDKN2A* in male breast tissue as compared to female breast tissue.

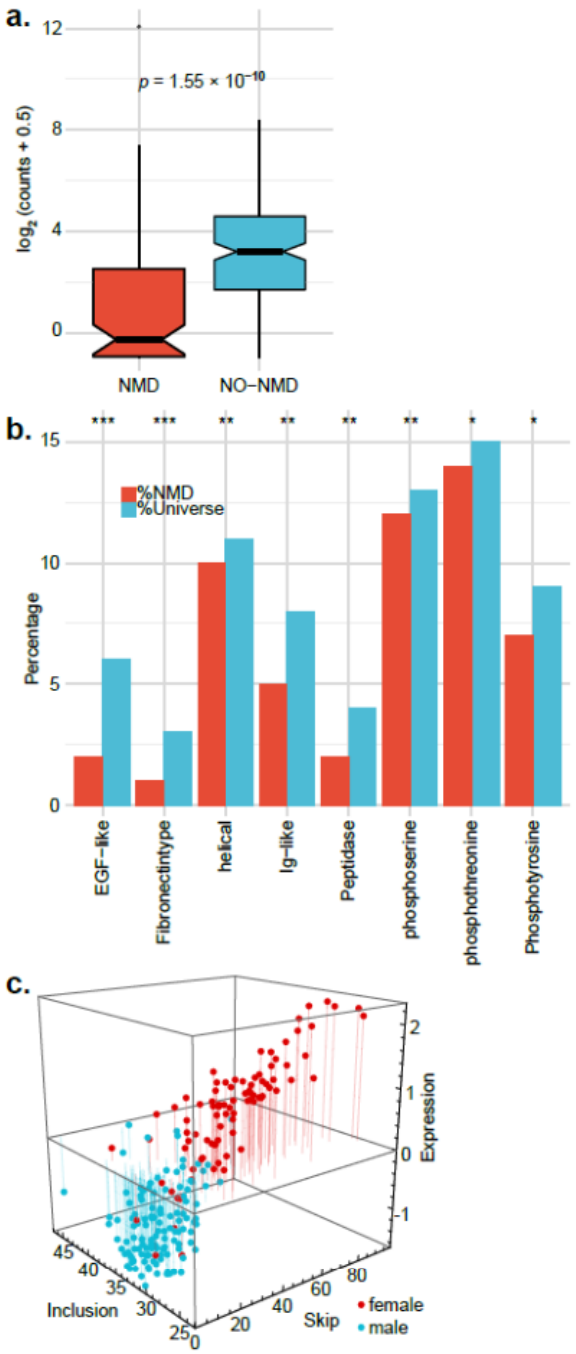

Figure 4 Nonsense-mediated decay in sex-biased alternative splicing. (a) Expression levels $\log_2$(mean inclusion counts+0.5) for SE events where the skipped exon belongs to at least one NMD isoform and where the exon does not belong to any NMD isoform. The plot shows the values for sex-biased events that were correlated with expression in our hierarchical Bayesian model. A similar distribution was observed for all the events in the dataset. **(b)** Depletion of domain annotations in NMD-associated skipped exons and their flanking exons. The y-axis indicates the percentage of ES events with the corresponding domain. The p-values were obtained using the hypergeometric density and Benjamini-Hochberg (BH) corrected for multiple testing. **(c)** 3-dimensional plot illustrating the relationship between gene expression and inclusion and exclusion counts for an exon skipping event in *CDKN2A* in mammary tissue. Females show both higher gene expression as well as higher skip counts than males. The skipped exon is present in isoform 5 of *CDKN2A* (NM_001195132.1) and causes a frameshift that is predicted to induce NMD.[34]

Interplay between splicing and gene expression: Type I and Type II exons

To validate this observation with a larger set of events, we performed linear regression of skipping counts against RBP levels for each of the 87 investigated RBPs. We then plotted the log fold change in expression vs. log fold change in inclusion for all sex-biased events in mammary tissue. Most of the exon-skipping events identified in mammary tissue were found in genes that showed higher expression in females than in males (Fig. 4c). Of these genes, one group showed a higher amount of exon skipping in females (defined as type I), and the other showed a higher amount of inclusion in females (define as type II). Exon-skipping events are shown in blue instead of red if at least one RBP was significantly associated with the exon-skipping event. The type I events in mammary tissue were all positively correlated with at least one RBP (blue points). In contrast, many of the type II AS events were not associated with any RBP (red points). There were 323 type I exons and 827 type II exons. The proportion of NMD-associated exons was not significantly different from the overall proportion (type I: 22.7% ; type II: 20%). The effects of RNA polymerase II (RNAP2) elongation rate on splicing can be studied with RNAP2 mutants that alter the average elongation rates genome-wide. It was previously shown that cassette exons included by slow and excluded by fast elongation display a number of attributes including weaker splice sites.[7] Although gene expression is not necessarily related to RNAP2 elongation speed, a recent study noted a correlation between the two,[30] which motivated us to investigate the sum of donor and acceptor splice scores of the exons associated with sex-biased type I and type II AS events. Indeed, type I exons showed a mean score that was 1.73 less than for type II exons (Fig. 4b; 14.82 vs 16.55; p=9 x 10^{-7}, t-test). We observed similar findings in the left ventricle (Supplemental Data Fig. 5). Sex-biased type I exons were 35 bp shorter than type II exons on average (178.9 vs. 143.8 nt; p=3.95 x 10^{-12}, t-test; Fig. 5c).

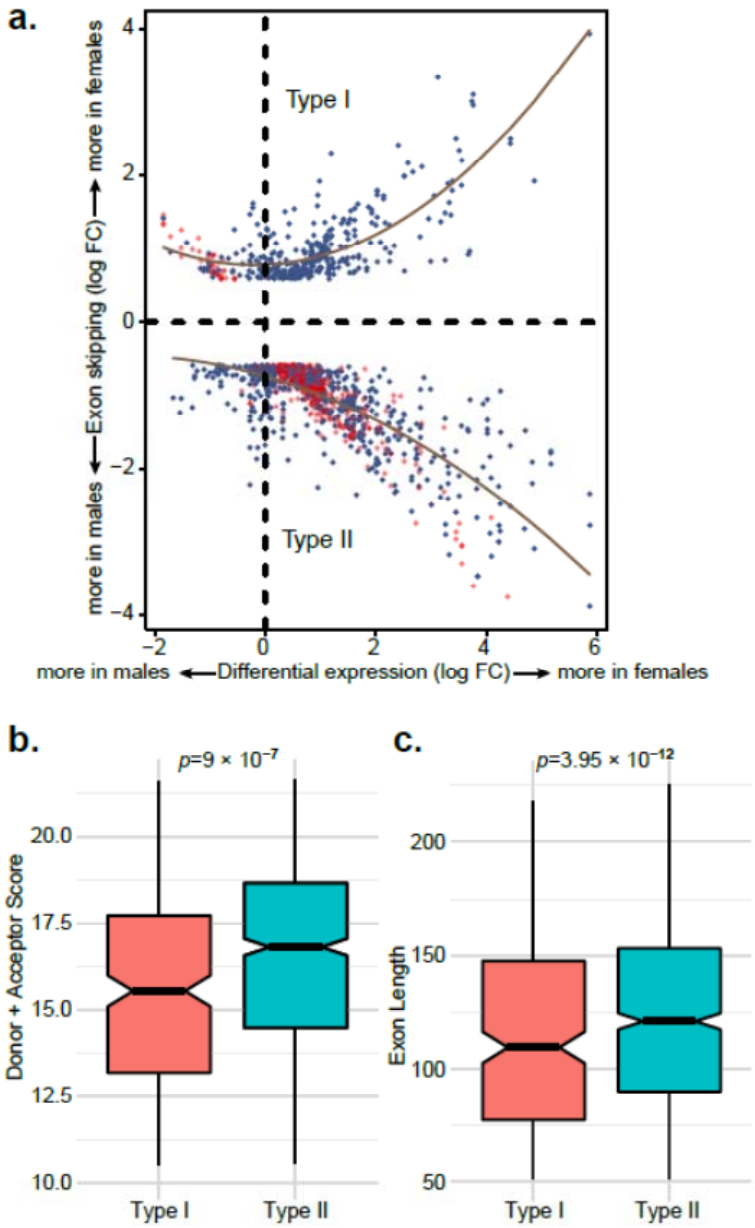

Figure 5 Two types of exons involved in sex-biased exon skipping. (a) log fold change expression vs. log fold change inclusion for all sex-biased events in mammary tissue. A higher fold change of female-to-male expression corresponds to a higher value on the X axis, and a higher fold change of female-to-male skipping-to-inclusion counts on the Y axis. Both fold changes are calculated by limma. Type I events showed a higher frequency of exon skipping in females and type II events showed a higher frequency of inclusion in females. Exon-skipping events are shown in blue instead of red if at least one RBP was significantly associated with the exon-skipping event. The brown lines represent a fit of the points by second-order polynomial regression. **(b)** Comparison of the sum of acceptor and donor splice scores for type I and type II exons (t-test). **(c)** Comparison of the sum of exon lengths for type I and type II exons (t-test).

Discussion.

We have provided the first comprehensive map of sex-biased gene expression and AS. Our computational results suggest previously undescribed associations between sex-biased AS events and escape from X-chromosomal inactivation as well as intronic Alu elements and oestrogen-receptor binding sites located therein. The vast majority of sex-biased AS events identified in our study were unique to one tissue. The density of sex-biased AS events per exon (SBSI) showed a striking degree of variability across the genome, being the highest for X chromosomal exons with a range of densities across the autosomes. We identified 27 genes that displayed sex-biased AS in ten or more tissues. 19 of these genes play roles in controlling gene expression or signalling, suggesting the possibility that the corresponding AS events are important for influencing regulatory networks that underlie sex differences in multiple tissues. We developed a hierarchical Bayesian modelling approach that uncovered widespread correlations between the levels of RBPs and overall gene expression and SE events. Transcription and pre-mRNA processing occur in a coordinated fashion, with capping, splicing and polyadenylation occurring while transcription proceeds.[35] These observations suggest that RNAP2 can regulate pre-mRNA splicing; perhaps the best documented association between RNAP2 biology and splicing is the observation of a coupling between RNAP2 elongation speed and alternative splicing.[7,8,36] One study showed a series of differences between cassette exons that tend to be included by slow and excluded by fast elongation as compared to exons that are excluded by slow and included by fast elongation.[7] In our study, we characterized exons that showed inclusion associated with low expression and exclusion associated with high expression (type I) as well as exons that showed inclusion associated with low expression and exclusion associated with high expression (type II, Fig. 4a). The type I exons showed lower splice scores and were always found to be significantly associated with RBP levels, whereas the type II exons had higher splice scores and often displayed no significant association with RBP levels. Although our findings are reminiscent of the results on RNAP2 elongation speed, gene expression and RNAP2 are not directly coupled, and other explanations such as a potential coupling of RNAP2 promoters to alternative splicing[37] with as yet to be explained effects on gene expression are conceivable. Finally, our results suggest that common approaches to the analysis of alternative splicing that discard differentially expressed genes[38] may be missing a substantial fraction of relevant AS events. To our knowledge, our findings represent the first demonstration of a pervasive relationship between gene expression and sex-biased AS.

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). Approval for use of the raw GTEx RNA-seq FASTQ files was granted by Database of Genotypes and Phenotypes (dbGaP).

Alignment of RNA-seq data
The 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 catalogued by the rMATS[39] 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,[40] and duplicates were removed using the Picard toolkit (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 and A5SS), and mutually exclusive exons (MXE). For each of these 5 different splicing types, rMATS creates two files, one containing the counts of reads that span the splicing junctions only, and a second file that additionally contains the counts of 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. In addition, for each tissue we kept only events for which the number of male and female samples with at least 1 cpm (count per million) for the gene of interest exceeds a threshold X, where X=(¼)min(#male samples, #female samples). For each of the 53 tissues, genes differentially expressed between male and female were individually determined using the voom function from the R package limma.[41,42] The heatplot shown in Fig. 1b was created with R software by calculating the fold change of the mean expression of each gene between male and female samples and displaying the correlation between the vectors of fold changes of the tissues.

Normalization of counts data for alternative splicing analysis
We used the *Yet Another RNA Normalization* software pipeline (YARN)[43] 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. [43]). 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 multidimensional scaling (MDS) plots that reveal similarities and differences between samples in a set of tissues. As input for MDS, we

concatenated the matrices of the skip and inclusion isoform counts row-wise such that each data point in the MDS is the counts 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, oesophagus, 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. 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[43] into finer subsets. For oesophagus, we found that Oesophagus - Gastroesophageal Junction and Oesophagus - Muscularis can be merged. Similar to YARN[43] , 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 [39] to identify and count reads that correspond to each of the 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 neighbouring exons, and the reads that fall within a splice junction that connects the neighbouring 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 voom to transform the counts into continuous data, appropriate for linear modelling. 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:

$$log_2(\#reads + 0.5) = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot event + \beta_3 \cdot event \cdot sex$$

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] For each tissue we kept only

events with ≥X/2 male samples with cpm ≥1 and analogously for the female samples, where X equals the size of the smallest study group (male or female).

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] and defined this set as RNA-binding proteins (RBPs).

Clustering Analysis

The log-fold-changes obtained from voom's differential expression analysis were obtained for each tissue, 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. 1c).

Gene Ontology analysis

Sets of significantly differentially expressed or spliced genes were obtained for each tissue and analysed with the model-based gene set analysis procedure in the Ontologizer.[46,47] The population set was defined to be the set of all annotated human genes using the Human GO Annotation from the EBI release 2018-03-26 (http://geneontology.org/page/download-go-annotations) that contains 19,712 gene product annotations for the association of genes to GO.

Hierarchical Bayesian Modelling

Hierarchical Bayesian modelling (HBM) is a technique for multiparameter modelling in which one assumes a statistical distribution for individual parameters whose interdependencies are reflected in the structure of the hierarchy. The HBM can use a Markov Chain Monte-Carlo (MCMC) 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 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). Additionally, our model does not assume that the effects of an RBP must be the same for sex-biased and non-sex-biased skipping events.

For the analysis, we chose up to 100 statistically significant sex-biased AS events for each tissue. If a tissue had less than 100 significant events, all of them were modelled.

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

$$S_{ij} \sim binomial\left(P_{ij}, S_{ij} + I_{ij}\right)$$

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

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

$$P_{ij} = \lambda\left(\beta_0^i + \beta_1^i \cdot sex_j + \beta_2^i \cdot expression_{j,g(i)} + \vec{\beta}_3^i \cdot \overrightarrow{RBP_j}\right)$$

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 to which the ith event belongs and is a vector of RBP effects on the ith event. is a vector of the normalized RNA levels of each RBP derived from the GTEx tpm matrix. Sex takes a value from {0,1}, and the expression level of the gene and the RBPs are normalized to have mean 0 and standard deviation 1 over all the samples.

The priors of , and are *N*(0,2), the priors for the effects of the kth RBP on skipping events. The vector contains normally distributed, unit-variance parameters that represent the estimated effects of each of the RBPs on skipping at event i. The model contains one such vector for each event. An additional normal distribution with a prior of *N*(0,) serves as prior for the mean of the coefficient of each RBP in each of the events. 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, the R interface of Stan.[48] The number of chains 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 available as Supplemental File 4.

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 skipping in males):

$$\{log_2(AS\ FC) |\ AS\ FC > 0\} = \beta_0 + \beta_1 \cdot log_2(DE\ FC) + \beta_2 \cdot log_2(DE\ FC)^2$$

$$\{log_2(AS\ FC) |\ AS\ FC < 0\} = \beta_0' + \beta_1' \cdot log_2(DE\ FC) + \beta_2' \cdot log_2(DE\ FC)^2$$

A linear regression was performed between the skip counts of each AS event (dependent variable) and the transcript per million (TPM) of each RBP.

$$\forall RBP_j, event_i :\ skip\ count(event_i) = \beta_0'' + \beta_1'' \cdot TPM\left(RBP_j\right)$$

The p-values obtained for each RBP were Benjamini-Hochberg (BH)-corrected for multiple testing, and a significant correlation was associated with an FDR 0.05.

Identification of events included in a nonsense-mediated decay isoform

For each skipped exon event, we find all isoforms that contain this exon using the Ensembl GTF file GRCh38.91. For each isoform, we compare the amino acid length of the inclusion isoform to the sum of lengths of the skip isoform and the skipped exon. If the former is smaller, then the inclusion isoform is identified as an NMD isoform containing a premature truncation codon as a consequence of including the exon. We do the reverse process to identify NMD isoforms in

which the isoform without the exon contains a PTC, e.g., because of a frameshift. For finding the lengths of each isoform, we use the program gffread from the GffCompare package (https://github.com/gpertea/gffcompare). All of the isoforms identified in the set of significantly sex-biased skipped exons fell into the first category.

Analysis of Alu subfamilies and predicted oestrogen receptor binding sites associated with sex-biased exon skipping

Repeat-masker data (http://www.repeatmasker.org) were downloaded from UCSC Table Browser in BED format for the human genome assembly hg38. These sequences were filtered by name to contain only Alu elements of length ≥50 bp on chromosomes 1-22,X,Y. The sequences were then partitioned by Alu class and scanned in both strands with five PWMs of estrogen receptor binding sites (Supplemental Data Table 5).

We defined an ER element as being present when the PWM score at any position in a sequence reached 80% of the maximum score.

To find over representation of Alu elements in the proximity of sex-biased events compared to all the events in our data, we used the hypergeometric enrichment test, where an event was considered proximal to an Alu element if the element is in the region that spans from the start if its upstream intron to the end of its downstream intron. To find the enrichment of oestrogen receptor binding sites within Alu elements in the proximity of sex-biased events compared to their enrichment in Alu elements in the proximity of all AS events in our data that passed our screening criteria in at least one tissue, we used a similar enrichment test, but the universe was restricted to all Alu-proximal events in our data, and enrichment of oestrogen receptor binding sites was tested in sex-biased Alu-proximal events. Categories that had a count of less than 25 for sex-biased events were considered insignificant and not tested. In addition, we only considered binding sites that were in an antisense orientation to the Alu element as these were found to occur in Alu elements more often than expected by chance. Multiple testing correction was performed using the Benjamini-Hochberg procedure.

Domain depletion in NMD-associated events

For each skipped exon event identified by rMATS analysis, we extracted the positions of the exons and its flanking exons, and downloaded domain annotation using MASER.[49] We calculated the probability of finding the observed number of domain-containing skipped exon events or a smaller number in our NMD-associated events using the hypergeometric distribution. We tested domains, modified residues, motifs, active sites, topological domains and trans- and intramembrane domains that occurred repeatedly in our set of sex-biased events (Supplemental Data Table 8).

Calculation of the Sex-Biased Splicing Index

Chromosomes are ranked by the normalized splicing index, which is the number of splicing events per 1000 exons in the chromosome. The normalized splicing index equals the number of exon skipping events divided by the number of exons on the chromosome and multiplied by 10^3.

Calculation of the Splice Score

Donor and acceptor splice scores were calculated using an information-content analysis.[50] Briefly, Information content is defined on the basis of a set of aligned donor or acceptor splice junction recognition sites by counting the frequencies of bases at each position and calculating the bits per base as 2 + log2f(b,l) with a sample-size correction factor. The individual information of a sequence is designated as R_i and is calculated as the dot product between the sequence and the weight matrix. The R_i was calculated for all donor or splice sites in the genome and for the indicated groups.

Funding

Support for this work was provided by the National Institutes of Health, through the Data Commons program (award 1OT3OD025646-01 [Helium]) and by the National Heart, Lung, and Blood Institute, through the Data STAGE program (award 1OT3HL142479-01 [Helium+]). Any opinions expressed in this document are those of the Commons and STAGE communities writ large and do not necessarily reflect the views of NIH, NHLBI, individual Commons and STAGE team members, or affiliated organizations and institutions. Additional support was provided by the Donald A. Roux Family Fund (PNR), The Jackson Laboratory (OA) and the NCI (R00CA178206 to OA).

Supplemental Files

Supplemental File 1 is an Excel file with data on log fold change and FDR for genes showing significant sex-biased differential expression and significantly sex-biased alternative splicing events.
Supplemental File 2 is an Excel file showing the 87 RNA-binding proteins examined in this work.
Supplemental File 3 is an Excel file with a summary of the results of hierarchical Bayesian modelling.
Supplemental File 4 is a text file that contains the STAN script used to perform the Bayesian analysis.

References

1. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18,** 437–451 (2017).
2. Poulos, M. G., Batra, R., Charizanis, K. & Swanson, M. S. Developments in RNA splicing and disease. *Cold Spring Harb. Perspect. Biol.* **3,** a000778 (2011).
3. 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).
4. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457–463 (2010).
5. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.* **84,** 291–323 (2015).
6. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16,** 990–995 (2009).
7. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28,** 2663–2676 (2014).
8. 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).
9. Rinn, J. L. & Snyder, M. Sexual dimorphism in mammalian gene expression. *Trends Genet.* **21,** 298–305 (2005).
10. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4,** 2771 (2013).
11. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

12. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348,** 660–665 (2015).
13. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
14. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550,** 239–243 (2017).
15. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49,** 692–699 (2017).
16. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27,** 1843–1858 (2017).
17. Merkhofer, E. C., Hu, P. & Johnson, T. L. Introduction to cotranscriptional RNA splicing. *Methods Mol. Biol.* **1126,** 83–96 (2014).
18. 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).
19. Welle, S., Tawil, R. & Thornton, C. A. Sex-related differences in gene expression in human skeletal muscle. *PLoS One* **3,** e1385 (2008).
20. 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).
21. 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).
22. Kratz, A. & Carninci, P. The devil in the details of RNA-seq. *Nat. Biotechnol.* **32,** 882–884 (2014).
23. 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).

24. Regitz-Zagrosek, V. & Kararigas, G. Mechanistic Pathways of Sex Differences in Cardiovascular Disease. *Physiol. Rev.* **97,** 1–37 (2017).
25. Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* **550,** 244–248 (2017).
26. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12,** 236 (2011).
27. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300,** 1288–1291 (2003).
28. Polak, P. & Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7,** 133 (2006).
29. Shankar, R., Grover, D., Brahmachari, S. K. & Mukerji, M. Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol. Biol.* **4,** 37 (2004).
30. Cohen, E., Zafrir, Z. & Tuller, T. A code for transcription elongation speed. *RNA Biol.* **15,** 81–94 (2018).
31. Brogna, S. & Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat. Struct. Mol. Biol.* **16,** 107–113 (2009).
32. 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).
33. 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).
34. Lin, Y.-C. *et al.* Human p16gamma, a novel transcriptional variant of p16(INK4A), coexpresses with p16(INK4A) in cancer cells and inhibits cell-cycle progression. *Oncogene* **26,** 7017–7027 (2007).
35. Cramer, P. *et al.* Coordination between transcription and pre-mRNA processing. *FEBS Lett.*

498, 179–182 (2001).

36. de la Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol. Cell* **12,** 525–532 (2003).
37. 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).
38. Climente-González, H., Porta-Pardo, E., Godzik, A. & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell Rep.* **20,** 2215–2226 (2017).
39. 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).
40. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12,** 357–360 (2015).
41. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
42. 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).
43. Paulson, J. N. *et al.* Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. *BMC Bioinformatics* **18,** 437 (2017).
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).
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).
46. 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).
47. 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).

48. Carpenter, B. *et al.* Stan: A Probabilistic Programming Language. *Journal of Statistical Software, Articles* **76,** 1–32 (2017).

49. maser. *Bioconductor* Available at: https://bioconductor.org/packages/release/bioc/html/maser.html. (Accessed: 7th December 2018)

50. Rogan, P. K., Faux, B. M. & Schneider, T. D. Information analysis of human splice site mutations. *Hum. Mutat.* **12,** 153–171 (1998).