nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20578-2 OPEN

Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome

Diego Garrido-Martín[1✉], Beatrice Borsari[1], Miquel Calvo[2], Ferran Reverter[2] & Roderic Guigó[1,3✉]

Alternative splicing (AS) is a fundamental step in eukaryotic mRNA biogenesis. Here, we develop an efficient and reproducible pipeline for the discovery of genetic variants that affect AS (splicing QTLs, sQTLs). We use it to analyze the GTEx dataset, generating a comprehensive catalog of sQTLs in the human genome. Downstream analysis of this catalog provides insight into the mechanisms underlying splicing regulation. We report that a core set of sQTLs is shared across multiple tissues. sQTLs often target the global splicing pattern of genes, rather than individual splicing events. Many also affect the expression of the same or other genes, uncovering regulatory loci that act through different mechanisms. sQTLs tend to be located in post-transcriptionally spliced introns, which would function as hotspots for splicing regulation. While many variants affect splicing patterns by altering the sequence of splice sites, many more modify the binding sites of RNA-binding proteins. Genetic variants affecting splicing can have a stronger phenotypic impact than those affecting gene expression.

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003 Catalonia, Spain. [2] Section of Statistics, Faculty of Biology, Universitat de Barcelona (UB), Av. Diagonal 643, Barcelona 08028, Spain. [3] Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain. ✉email: diego.garrido@crg.eu; roderic.guigo@crg.eu

Alternative splicing (AS) is the process through which multiple transcript isoforms are produced from a single gene[1]. It is a key mechanism that increases functional complexity in higher eukaryotes[2]. Often, its alteration leads to pathological conditions[3]. AS is subject to a tight regulation, usually tissue-, cell type-, or condition-specific, that involves a wide range of *cis* and *trans* regulatory elements[4,5]. Since AS is generally coupled with transcription, transcription factors and chromatin structure also play a role in its regulation[6].

In recent years, transcriptome profiling of large cohorts of genotyped individuals by RNA-seq has allowed the identification of genetic variants affecting AS, i.e. splicing quantitative trait loci or sQTLs[7–12]. sQTL analyses in a variety of experimental settings have helped to gain insight into the mechanisms underlying GWAS associations for a number of traits, such as adipose-related traits[13], Alzheimer's disease[10], schizophrenia[9] or breast cancer[14], among others. sQTLs might actually contribute to complex traits and diseases at a similar or even larger degree than variants affecting gene expression[15].

The vast majority of methods commonly used for sQTL mapping treat splicing as a univariate phenotype. They assess association between genetic variants and the abundance of individual transcripts[7,16], or the splicing of individual exons[9,17] or introns[12,15]. However, this approach ignores the strongly correlated structure of AS measurements (e.g. at constant gene expression level, higher levels of a splicing isoform correspond necessarily to lower levels of other isoforms). In contrast, we propose an approach that takes into account the intrinsically multivariate nature of alternative splicing: variants are tested for association with a vector of AS phenotypes, such as the relative abundances of the transcript isoforms of a gene or the intron excision ratios of an intron cluster obtained by LeafCutter[18].

Based on this approach, we have developed a pipeline for efficient and reproducible sQTL mapping. Here we employ it to leverage the multi-tissue transcriptome data generated by the Genotype-Tissue Expression (GTEx) Consortium, producing a comprehensive catalog of genetic variants affecting splicing in the human genome. Downstream analyses of this catalog uncover a number of relevant features regarding splicing regulation. Thus, consistent with the multivariate nature of splicing, we observe that sQTLs tend to involve multiple splicing events. A substantial fraction of sQTLs also affects gene expression, a reflection of the intimate relationship between splicing and transcription. We find, however, many cases in which the expression of a gene other than the sQTL target is affected by the same variant. In these cases, the pleiotropic effect of the regulatory locus is not mediated by the interplay between the splicing and transcription processes, but it is exerted through different mechanisms, acting upon different genes that otherwise may not appear to be directly interacting. We also find that sQTLs tend to be preferentially located in introns that are post-transcriptionally spliced: these introns would be consequently acting as hotspots for splicing regulation. While many variants affect splicing patterns by directly altering the sequence of splice sites, many more modify the binding of RNA-binding proteins (RBPs) to target sequences within the transcripts. We observe that sQTLs often impact GWAS traits and diseases more than variants affecting only gene expression, confirming earlier reports which suggest that splicing mutations underlie many hereditary diseases[15,19]. For several conditions, GWAS associations are particularly strong for sQTLs altering RBP binding sites.

Results

Identification of *cis* splicing QTLs across GTEx tissues. For sQTL mapping, we developed sQTLseekeR2, a software based on sQTLseekeR[20], which identifies genetic variants associated with changes in the relative abundances of the transcript isoforms of a given gene. sQTLseekeR uses the Hellinger distance to estimate the variability of isoform abundances across observations, and Anderson's method[21,22], a non-parametric analog to multivariate analysis of variance, to assess the significance of the associations (see Methods and Supplementary Note 1). Among other enhancements, sQTLseekeR2 improves the accuracy and speed of the *p*-value calculation, and allows to account for additional covariates before testing for association with the genotype, while maintaining the multivariate statistical test in sQTLseekeR. It also implements a multiple testing correction scheme that empirically characterizes, for each gene, the distribution of *p*-values expected under the null hypothesis of no association (see Methods and Supplementary Note 1). To ensure highly parallel, portable and reproducible sQTL mapping, we embedded sQTLseekeR2 in a Nextflow[23] (plus Docker, https://www.docker.com) computational workflow named sqtlseeker2-nf, available at https://github.com/guigolab/sqtlseeker2-nf.

Here we extensively analyze the sQTLs identified by sqtlseeker2-nf, using the expression and genotype data produced by the GTEx Consortium. For most of the analyses, we employed isoform quantifications obtained from the V7 release (dbGaP accession phs000424.v7.p2), corresponding to 10,361 samples from 53 tissues of 620 deceased donors. 48 tissues with sample size ≥ 70 were selected for sQTL analyses. Under the assumption that most variants with *cis* effects on alternative splicing are likely to be carried on the sequence of the primary transcript or its close vicinity, we tested variants in a *cis* window defined as the gene body plus 5 Kb upstream and downstream the gene boundaries. In addition, to demonstrate that the statistical framework of sQTLseekeR2 is not restricted to the analysis of transcript abundances, but it can leverage other splicing-related multivariate phenotypes, we have also computed the sQTLs based on the intron excision ratios obtained by LeafCutter[18] from the GTEx RNA-seq data (Supplementary Note 2). Finally, we also provide the sQTLs identified by sqtlseeker2-nf in GTEx V8, which we compare to the sQTLs produced by the GTEx Consortium, described in a recent publication[12]. We show that the two sets of sQTLs differ indeed in the nature of the AS events captured. Our approach detects more events affecting the gene termini and intron retention, while the approach by the GTEx Consortium tends to detect more events involving internal exons. The two sQTL sets also differ regarding several biological features at variant and gene level. For instance, our approach seems to have more power to identify sQTLs affecting genes with lower expression levels and shorter introns, as well as involving variants with lower minor allele frequency (MAF). In contrast, the GTEx Consortium's approach has larger power to identify sQTLs affecting genes with higher expression and longer introns (Supplementary Note 3, Supplementary Data 7).

At a 0.05 false discovery rate (FDR), we found in GTEx V7 a total of 210,485 *cis* sQTLs affecting 6,963 genes (i.e. sGenes: 6,685 protein coding genes and 278 long intergenic non-coding RNAs, lincRNAs). On average, per tissue, we identified 1,158 sGenes (Supplementary Table 1). 44% and 34% of all tested protein coding genes and lincRNAs, respectively, were found to be sGenes. In an analogous experimental setting, the GTEx Consortium reported genetic variants affecting gene expression (expression QTLs, eQTLs) for 95% and 71% of all tested protein coding genes and lincRNAs, respectively[24]. To illustrate the nature of the sQTLs identified with sqtlseeker2-nf, in Fig. 1 we show the example of the SNP rs2295682, an sQTL for the gene *RBM23* shared across 46 tissues, with larger effect in brain subregions such as the cortex. The SNP strongly affects the relative abundances of the AS isoforms of the target gene, the dominant isoform depending on the genotype at the sQTL.

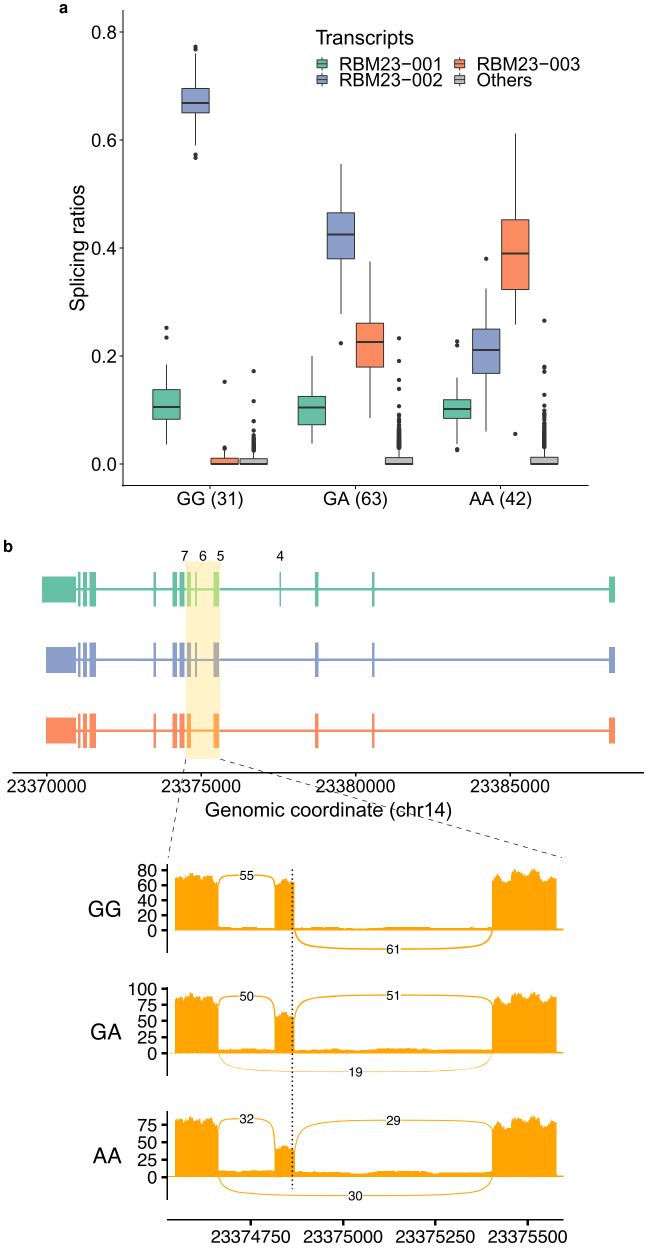


Fig. 1 sQTL example. a Relative abundances of the three most expressed isoforms in the brain cortex from the gene *RBM23* (chr14:23,369,854-23,388,393, reverse strand, *RBM23-001*, *RBM23-002* and *RBM23-003*, all protein coding), for each genotype group at the rs2295682 locus (chr14:23,374,862, G/A in the reverse strand), represented as boxplots. *RBM23* encodes for an RBP that may be itself involved in splicing. The least abundant isoforms are grouped in Others. The number of individuals in each genotype group is shown between parentheses. Individuals that are homozygous for the reference allele (GG) at the SNP locus, express preferentially *RBM23-002* (blue), while they barely express *RBM23-003* (red). In contrast, AA homozygous express preferentially *RBM23-003* (red). Heterozygous individuals exhibit intermediate abundances. *RBM23-001* (green) has similar levels in the three genotype groups. In boxplots, the box represents the first to third quartiles and the median, while the whiskers indicate ± 1.5 × interquartile range (IQR). Source data are provided as a Source Data file. **b** Exonic structure of the isoforms of *RBM23*. Compared to *RBM23-001* (green), *RBM23-002* (blue) lacks exon 6, and *RBM23-003* (red), exons 4 and 6. A sashimi plot corresponding to the highlighted area displays the mean exon inclusion of exon 6 of RBM23 across all brain cortex samples of each genotype group at rs2295682. The plot was obtained using ggsashimi[86]. The dotted vertical line marks the location of the SNP. The number of reads supporting exon skipping increases with the number of copies of the alternative allele A, matching the changes observed in isoform abundances. This allele has been previously associated with increased skipping of exon 6[87].

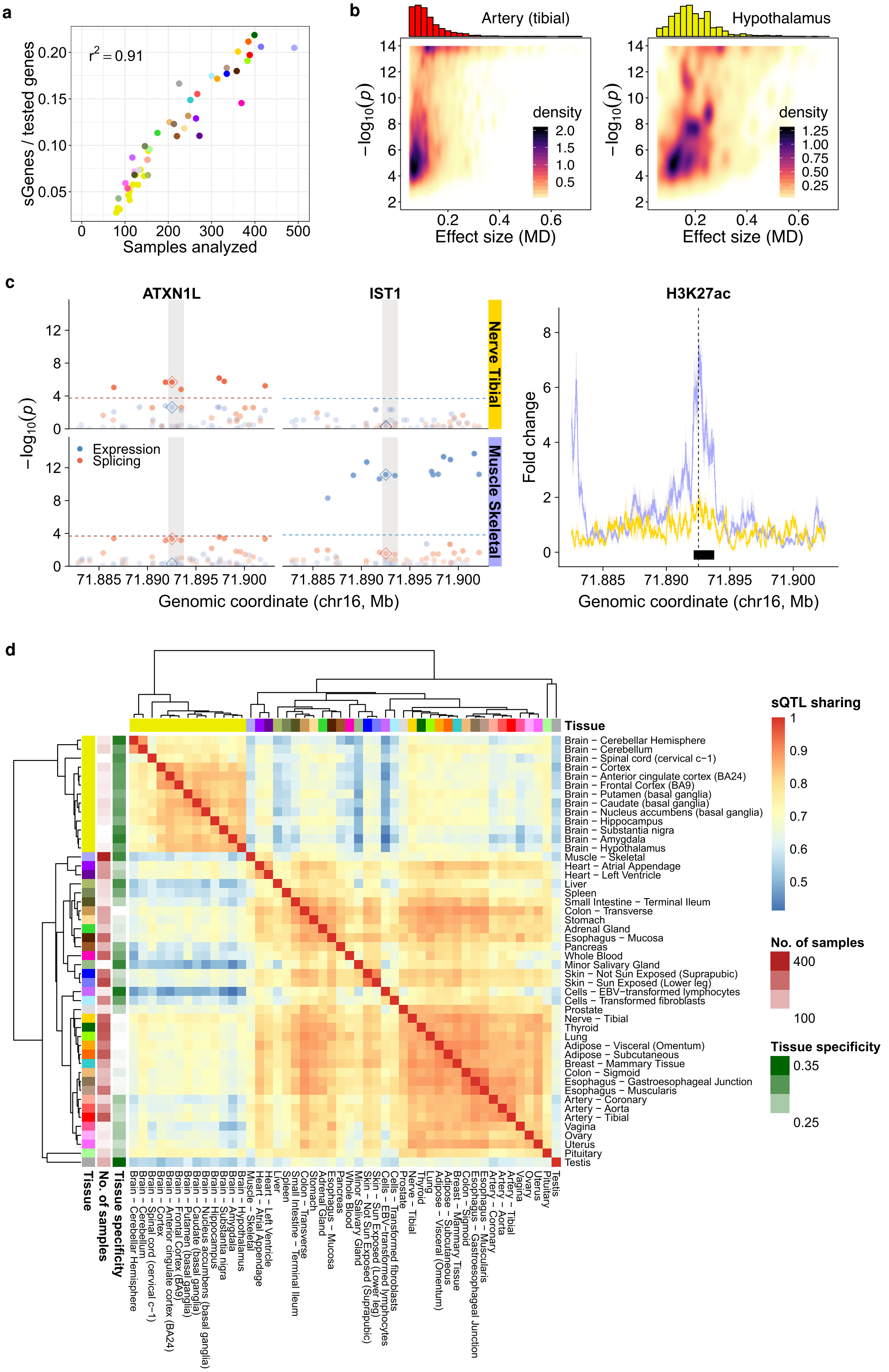
As expected, the number of sGenes over the number of tested genes grows with the tissue sample size ($r^2 = 0.91$). This is explained by the gain of power to detect sQTLs as the number of samples increases (Fig. 2a). No signs of saturation are observed. Some tissues, such as skeletal muscle or whole blood (with less sQTLs than expected), and testis (with more sQTLs than expected), escape the general trend. This was also observed for eQTLs[24]. The cell type heterogeneity of the tissue, estimated using xCell[25], does not seem to have a large impact on sQTL discovery compared to the tissue sample size (the partial correlation between the number of sGenes over the number of tested genes and the estimated cell type heterogeneity, controlling for the tissue sample size, is 0.23, p-value 0.11, see Methods).

sQTL effect sizes, measured as the absolute maximum difference (MD) in adjusted transcript relative expression between genotype groups (see Methods), are generally low to moderate (MD from 0.05 to 0.20). Nevertheless, around 20% of sQTLs account for large effects (MD ≥ 0.20). As one would expect, the median effect size detected across tissues drops substantially with increasing sample sizes (Supplementary Fig. 1), given that larger sample sizes allow the detection of smaller effects. Figure 2b represents sQTL effect sizes (MD values) *vs* p-values, together with the distribution of the former, for two tissues with markedly different sample sizes: tibial artery ($n = 388$) and hypothalamus ($n = 108$).

GO enrichment analysis of sGenes shows a wide variety of biological processes, including cellular transport, immune response, mitochondrial functions and, interestingly, RNA processing (Supplementary Fig. 2a). This might suggest some mechanism of splicing autoregulation, as it has been previously described[26]. In contrast, tested genes without sQTLs are enriched in functions related to signaling and, especially, development (Supplementary Fig. 2b). This resembles the behavior reported for genes without eQTLs[24], as it does the fact that genes without sQTLs are less expressed than sGenes in all tissues (two-sided Wilcoxon Rank-Sum test p-value $< 10^{-16}$).

The sQTLs found here are highly replicated in other studies. We compared them with those obtained in the Blueprint Project[27] for three major human blood cell types (CD14$^+$ monocytes, CD16$^+$ neutrophils, and naive CD4$^+$ T cells, see Methods). The majority of GTEx sQTLs replicate at 0.05 FDR (from $\pi_1 = 0.80$ in brain subregions to $\pi_1 = 0.96$ in whole blood). As expected, whole blood displays the highest sQTL replication rate (Supplementary Fig. 3).

We characterized the types of AS events associated with sQTLs (see Methods, Supplementary Fig. 4a). Note that here we also account for other relevant sources of transcript diversity, such as alternative transcription initiation and termination[28]. sQTLs generally involve multiple events (on average 2.63). Around 34% of sQTLs are related to at least one AS event involving internal exons and/or introns. Among them, exon skipping is the most frequent simple event (7% to 10% of all events). In addition, 58% of sQTLs are associated with events affecting first/last exons and untranslated regions (UTRs). The landscape of AS events associated with sQTLs is very similar across tissues. However, brain subregions present some particularities when compared to non-brain tissues, such as a larger proportion of exon skipping



events and a smaller proportion of complex events involving the 3′ gene terminus (see Supplementary Fig. 4b, c for details).

We found that 52% of the identified sQTLs are also eQTLs for the same gene and tissue, although the top sQTL coincides with the top eQTL only in 3% of the cases. This relatively large overlap, which departs from that reported in some previous studies[15], matches what was observed for sQTLseekeR sQTLs in the GTEx pilot study[29]. This is partially due to our sQTLs being able to involve transcriptional termini, in addition to canonical splicing events. It also indicates a substantial degree of co-regulation of gene expression and splicing, either at the level of transcription (e.g. variants that impact transcription and thus, splicing), or at

Fig. 2 Overall results, heteropleiotropy and sQTL sharing across tissues. a Proportion of sGenes (over tested genes) per tissue (y-axis) with respect to the tissue sample size (x-axis). Tissue color codes are shown in Supplementary Table 1. **b** For two tissues with markedly different sample sizes, such as tibial artery (left panel, $n = 388$ samples) and hypothalamus (right panel, $n = 108$ samples), we display the effect sizes (MD values, x-axis) of significant sQTLs *vs* the $-\log_{10}$ of their association *p*-value (Anderson test) with the target sGene (y-axis). The density of points is shown, together with the sQTL effect size distribution. Note that MD for sQTLs is bounded to [0.05, 1] (see Methods). **c** Example of a heteropleiotropic locus. The SNP rs8046859 (chr16:71,892,531, C/T) is an sQTL for the gene *ATXN1L* (chr16:71,879,894-71,919,171, forward strand) in Nerve Tibial ($n = 361$), but not in Muscle Skeletal ($n = 491$). The SNP is not an eQTL for *ATXN1L* in any of the two tissues. In contrast, the SNP is an eQTL for the gene *IST1* (chr16:71,879,899-71,962,913, forward strand) in Muscle Skeletal, but not in Nerve Tibial. The SNP is not an sQTL for *IST1* in any of the two tissues. In the left panel, the dots represent the $-\log_{10}$ *p*-values of association with the expression (two-sided *t*-test, blue) and splicing (Anderson test, red) of the two genes in the two tissues, for variants in a 20 Kb window centered at rs8046859 (the $-\log_{10}$ *p*-values corresponding to rs8046859 are highlighted by a diamond). The transparency of the dots depends on the $-\log_{10}$ *p*-value. The significance level for each molecular trait, gene and tissue is shown as a colored, horizontal dashed line. When this line is not present, the gene-level *p*-value is above the 0.05 FDR threshold and hence no variant is significantly associated with this molecular trait in this tissue (see Methods). The shaded area represents the position of a H3K27ac ChIP-seq peak (see below). The right panel shows the fold-change signal of the H3K27ac histone mark with respect to the input across ENTEx donors in Nerve Tibial and Muscle Skeletal, in the same genomic region of the left panel. The line and colored area correspond, respectively, to the mean fold-change signal and its standard error (SEM) across four ENTEx donors (i.e. mean ± SEM). The location of the SNP (vertical dashed line) and the overlapping ChIP-seq peak (intersection of the peaks in the four donors, black rectangle) are also displayed. **d** Heatmap of sQTL sharing across GTEx tissues. Sharing estimates (see Methods) range from 0 (low sharing, blue) to 1 (high sharing, red). In addition, hierarchical clustering of the tissues based on sQTL sharing is displayed, together with the tissue sample sizes and tissue specificity estimates. Source data for **a**–**d** are provided as a Source Data file.

the level of transcript stability (e.g. variants that affect splicing, and as a consequence, transcript stability and gene expression).

We focused on a set of 148,618 variants that were tested for association with both the expression and splicing of two genes (i.e. g_1 and g_2) or more, in at least two tissues, and identified 6,552 cases in which the variant is only sQTL for gene g_1, but not for gene g_2, in one tissue, and it is only eQTL for gene g_2, but not for gene g_1, in a different tissue (Supplementary Fig. 5a). These cases uncover regulatory loci in the genome that, either through the same causal variant or through different causal variants in linkage disequilibrium (LD), have different effects on different genes through likely different molecular mechanisms. We term this phenomenon heteropleiotropy. Note that our identification of heteropleiotropic loci should be considered a first approximation, since we lack a specific statistical test to assess heteropleiotropy (see Methods and Discussion). Nevertheless, we found additional biological support for the dual regulatory behavior of these loci. We identified the ChIP-seq peaks corresponding to six histone modifications from the ENTEx Project overlapping the heteropleiotropic variants above (see Methods). We hypothesized that loci with different regulatory effects (i.e. splicing and expression) in different tissues would be differently marked by histone modifications in these tissues. Indeed, we observed histone modification changes in 24% of the heteropleiotropic variants (Supplementary Data 1), compared to 19% of the non-heteropleiotropic variants (two-sided Fisher's exact test *p*-value 0.045, see Methods). Regardless of the underlying causal structure, heteropleiotropic loci would uncover genomic regions that allow the coordinated regulation of different processes and affect different genes which otherwise do not appear to interact directly with each other. While further work is required to establish the relevance and generality of this phenomenon, Fig. 2c and Supplementary Fig. 5b show some potentially interesting examples.

sQTLs are highly shared across tissues. The large number of tissues available in GTEx allowed us to evaluate tissue sharing and specificity of sQTLs. For every pair of tissues, we selected variant-gene pairs tested in both and found significant in at least one, and computed the Pearson correlation (r) between their effect sizes (MD values). Hierarchical clustering based on these correlations grouped tissues with similar sQTL sharing patterns (Fig. 2d). A comparable clustering was obtained when using the more stringent Jaccard index (Supplementary Fig. 6). Brain subregions cluster together and apart from the rest of the tissues, which form a second major cluster. We observe a high degree of sQTL sharing within each of the two groups ($\bar{r} = 0.80$ and 0.78, respectively), but lower between them ($\bar{r} = 0.64$). The same pattern was depicted for eQTLs in GTEx[24]. We further estimated tissue specificity as $s_t = 1-\bar{r}_t$, where $\bar{r}_t$ is the mean correlation between a given tissue t and the others (tissue specificity estimates shown in Fig. 2d). On average, brain sQTLs are more tissue-specific than non-brain sQTLs ($\bar{s}_t = 0.31$ *vs* 0.25, two-sided Wilcoxon Rank-Sum test *p*-value $9.32 \cdot 10^{-5}$). Other tissues with relatively high tissue-specific sQTLs include testis (0.37), skeletal muscle (0.33) or liver (0.32).

sQTLs with large effects are more shared than those with smaller effects (Supplementary Fig. 7a). As with eQTLs[24], the detection of sQTLs with small effects requires larger sample sizes, thus sQTLs in tissues with small sample sizes tend to be more shared, while sQTLs identified in tissues with large sample sizes tend to be more tissue-specific (Supplementary Fig. 7b). To rule out an effect of the sample size in the patterns of sQTL sharing, we downsampled the original dataset to 100, 200 and 300 samples per tissue, and evaluated again sQTL sharing. We found that the patterns of sQTL sharing above are replicated independently of the sample size (Supplementary Fig. 8).

To capture more complex sharing patterns, we further designed a geometric approach that compares changes in the whole splicing phenotype due to sQTLs between tissues (see Methods and Supplementary Fig. 9a). The derived tissue dendrogram (Supplementary Fig. 9b) displayed high similarity with the ones generated by simpler approaches (i.e. based on MD values and Jaccard index), and also with the one obtained using multivariate adaptive shrinkage[30] on LeafCutter sQTLs from GTEx V8[12] (Supplementary Fig. 9c). This strongly supports the robustness of the sQTL sharing patterns observed.

sGenes are also markedly shared: 66% of the genes tested in all tissues are sGenes in at least two tissues. To identify tissue-specific sGenes, we computed τ_s, a variation of the τ index[31] based on sGene significance. We also employed the standard τ to determine the tissue specificity of sGene expression (see Methods). We found 469 genes under strong tissue-specific splicing regulation (highly tissue-specific sGenes), 81 of which did not display tissue-specific expression (Supplementary Data 2). GO enrichment of these genes (universe: all sGenes) identified biological processes related to RNA processing and its regulation (three out of five significant terms at FDR < 0.1: RNA splicing via

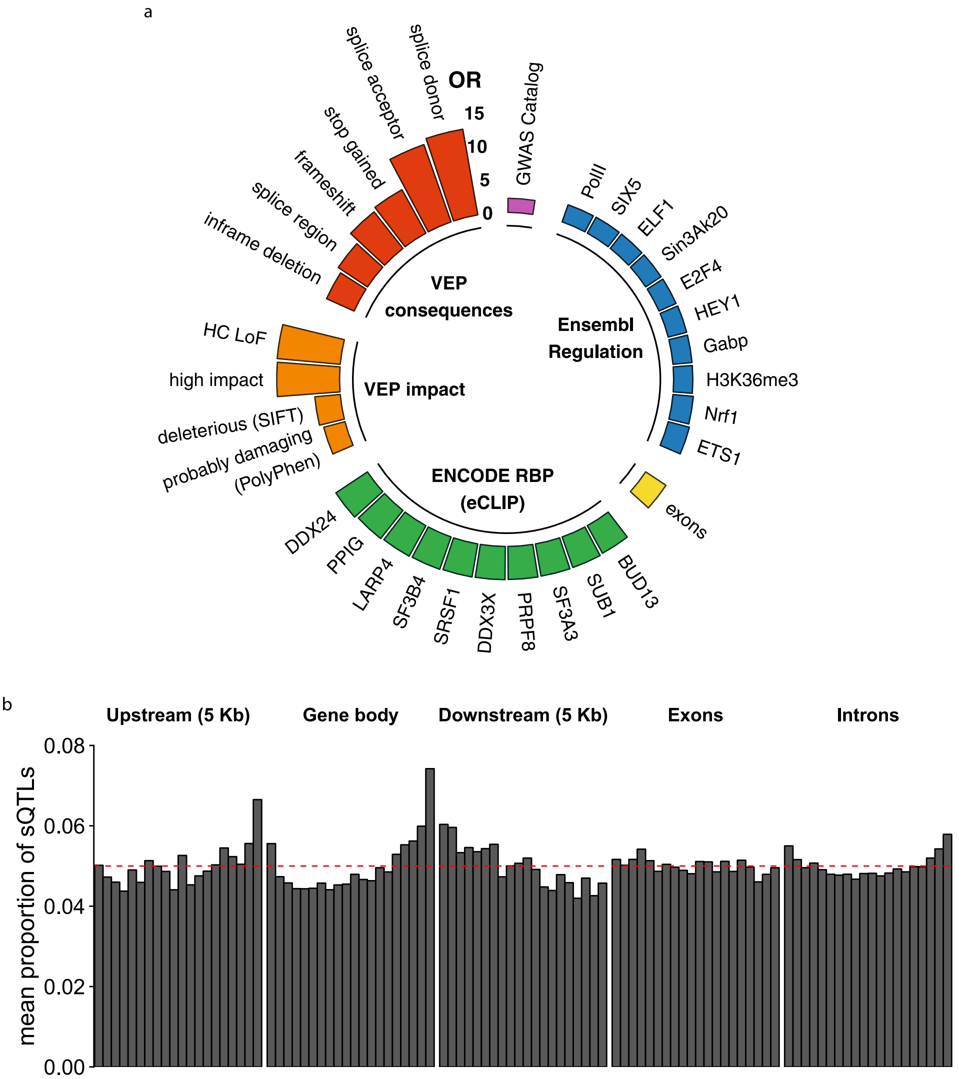


Fig. 3 Functional enrichment and distribution of sQTLs. a Top enrichments of sQTLs in functional annotations. The height of the bars represents the odds ratio (OR) of the observed number of sQTLs to the expected number of variants that are not sQTLs overlapping a given annotation (see Methods): Variant Effect Predictor (VEP) categories (red) and impact (orange), ENCODE RBP eCLIP peaks (green), exons of GENCODE v19 protein coding and lincRNA genes (yellow), Ensembl Regulatory Build elements (blue) and GWAS Catalog hits (purple). All these enrichments are significant at FDR < 0.05 and have OR confidence intervals not overlapping the range [1/1.50, 1.50]. **b** Distribution of the mean proportion of sQTLs along the gene bodies of sGenes, their upstream and downstream regions, introns and exons. The red dashed line represents the expected distribution under a uniform model (see Methods). Source data for both **a** and **b** are provided as a Source Data file.

transesterification reactions, regulation of RNA splicing and regulation of mRNA processing) suggesting again some mechanism of splicing autoregulation[26].

sQTLs are enriched in functional elements of the genome related to splicing and in high-impact variants. To shed light on the mechanisms through which sQTLs may impact splicing, we built a comprehensive functional annotation of the human genome (see Methods). Overall, we observed a high density of functional elements in the proximity of sQTLs (Supplementary Fig. 10). We next evaluated the enrichment of sQTLs in every functional category, with respect to a null distribution of similar variants not associated with splicing (two-sided Fisher's exact test, FDR < 0.05). The top enrichments are summarized in Fig. 3a (the complete list, together with the statistical significance associated with each enrichment, is shown in Supplementary Fig. 11).

As one would expect from *bona fide* variants affecting splicing, sQTLs are strongly enriched in splice sites (donors: OR = 12.98, adj. p-value < 10^{-16}; acceptors: OR = 12.23, adj. p-value $1.22 \cdot 10^{-15}$). They also display enrichments in exons, transcription factor binding sites (TFBS, both activator and repressor), RBP binding sites (including several relevant splicing factors and spliceosomal components), and RNA Pol II sites. sQTLs tend to fall in open chromatin regions and show enrichments for several chromatin marks, particularly for H3K36me3 (OR = 2.85, adj. p-value < 10^{-16}). H3K27me3 regions, in contrast, are depleted of sQTLs (OR = 0.63, adj. p-value < 10^{-16}). sQTLs display large enrichments in predicted protein loss-of-function consequences (stop-gained, frameshift, VEP high impact variants, LOFTEE high-confidence loss-of function variants (HC-LoF)) and potentially deleterious variants (according to Polyphen[32] and SIFT[33] scores). In addition, we found an enrichment in variants in high LD ($r^2 \geq 0.80$) with GWAS hits (OR = 2.08, adj. p-value < 10^{-16}). When performing stratified enrichments (see Methods), we found that sQTLs with large effect sizes are more enriched in high impact variants, splice sites and GWAS hits, while sQTLs with small effect sizes show larger enrichments in RBP binding sites, TFBS and open chromatin regions (Supplementary Fig. 12).

In contrast to eQTLs, which tend to cluster around transcription start sites (TSS)[7,24], we found sQTLs preferentially located towards transcription termination sites (TTS) (Fig. 3b), as previously observed[15]. In addition, while exonic sQTLs are uniformly distributed, intronic sQTLs are biased towards splice sites. Overall, sQTLs are closer to splice sites than non-sQTLs (two-sided Wilcoxon Rank-Sum test p-value $< 10^{-16}$, Supplementary Fig. 13).

sQTLs affect splice site strength and RBP binding. Enrichments in functional annotations (Fig. 3a) suggested several mechanisms through which sQTLs may affect splicing. One of them is the modification of splice site strength. Thus, for each variant within the sequence of an annotated splice site, we scored the site considering the reference and the alternative allele, using position weight matrices (PWMs) (see Methods). Overall, when compared to non-sQTL variants, a larger fraction of sQTLs modifies splice site strength (63% *vs* 49%, OR = 1.79, two-sided Fisher's exact test p-value $< 10^{-16}$). The absolute difference in splice site strength is also larger for sQTLs (two-sided Wilcoxon Rank-Sum test p-value 1.98×10^{-7}), and increases with the sQTL effect size (Fig. 4a).

Another mechanism through which sQTLs may affect splicing is the modification of RBP binding sites. To investigate it, we used eCLIP peaks of 113 RBPs available for HepG2 and K562 cell lines from the ENCODE project[34]. We employed a k-mer-based machine learning approach, which has been shown to outperform PWMs to identify transcription factor binding sites[35] and provides a unique framework to assess the impact of genetic variants on the binding[36]. First we trained, for each RBP, a gapped k-mer support vector machine (gkm-SVM)[37] on the sequences of high-confidence eCLIP peaks. 79 RBPs with a mean cross-validation ROC AUC ≥ 0.8 were kept. Then, we estimated the impact of all variants (whether sQTLs or not) overlapping the eCLIP peaks of each of these RBPs via the deltaSVM metric[36], which measures the difference in predictive potential between the variant alleles (see Methods). To ensure the robustness of our results, we further restricted the analysis to RBPs with at least 30 sQTLs among the top 5% variants most predictive of the binding of the RBP at either allele, resulting in a final set of 32 RBPs (see Methods).

At FDR < 0.05, differences in |deltaSVM| between sQTLs and non-sQTLs were found significant for ten RBPs (Fig. 4b, the corresponding gkm-SVM ROC curves and motif logos are shown in Supplementary Figs. 14 and 15, respectively). Notably, for nine of these proteins the |deltaSVM| values are larger for sQTLs than for non-sQTLs, as expected from variants regulating splicing. In addition, three of them (PPIG, SF3B4 and PRPF8) are among the top ten RBPs whose binding sites are more enriched in sQTLs (Fig. 3a). In Fig. 4c, we show examples of the impact of the SNPs rs4959783 and rs9876026, which are sQTLs for the genes *PSMG4* and *TAMM41* (see also Supplementary Fig. 16) and disrupt the binding sites of the RBPs RBFOX2 and PRPF8, respectively.

We further investigated whether allele-specific RBP binding (ASB) was occurring specifically at sQTLs. We obtained a set of ASB variants identified in the ENCODE eCLIP dataset using BEAPR (Binding Estimation of Allele-specific Protein-RNA interaction)[38] and overlapped them with our sQTLs (see Methods). We found that sQTLs were highly enriched in ASB variants, when compared to non-sQTLs, across all RBPs (OR = 2.30, two-sided Fisher's exact test p-value $< 10^{-16}$). When considering individual RBPs, at FDR < 0.05 we found a significant enrichment of sQTLs among ASB variants for 22 of them (Supplementary Fig. 17), including six of the ones identified above with larger |deltaSVM| values for sQTLs. Altogether, these results suggest that sQTLs may affect splicing through allele-specific binding of RBPs.

Overall, the effect sizes (MD) of sQTLs in splice sites are larger than those of sQTLs overlapping RBP eCLIP peaks (two-sided Wilcoxon Rank-Sum test p-value $1.98 \cdot 10^{-7}$, Supplementary Fig. 18), although the proportion of sQTLs in splice sites is much smaller (1.5% *vs* 8.3% out of all sQTLs). Often, both mechanisms may co-occur, as many RBPs bind near splice sites. This is the case of PRPF8, which binds specifically to the sequence of splice donors[39]. Indeed, the SNP rs9876026 (Fig. 4c), which modifies |deltaSVM| and has been identified as an allele-specific binding SNP for PRPF8 by BEAPR, also disrupts a donor splice site.

sQTLs are preferentially located on post-transcriptionally spliced introns. Although splicing generally occurs co-transcriptionally (most introns are spliced prior to transcription termination and polyadenylation), there is a group of transcripts, often alternatively spliced, that tend to be processed more slowly, even post-transcriptionally[40]. We evaluated the role of genetic variants in the regulation of co- and post-transcriptional splicing (here referred to as *cs* and *ps*, respectively). In order to identify *cs* and *ps* introns, we determined the degree of splicing completion of annotated introns in nuclear and cytosolic RNA-seq data available for 13 cell lines from the ENCODE project (see Methods). We focused on a subset of introns consistently classified as either *cs* or *ps* in at least 10 of the analyzed cell lines (14,699 and 6,419 introns, respectively).

We observe a higher variant density in *ps* introns than in *cs* introns (4.38 *vs* 3.34 variants/Kb, differently distributed along the intron, Supplementary Fig. 19a). The proportion of variants that are sQTLs in *ps* introns is larger than in *cs* introns (9.2% compared to 6.6%, OR = 1.47, two-sided Fisher's exact test p-value $< 10^{-16}$). This enrichment is stronger when considering sQTLs that are not eQTLs for the same gene and tissue (OR = 1.67, p-value $< 10^{-16}$). Furthermore, sQTLs in *ps* introns display a substantial enrichment, with respect to sQTLs in *cs* introns, in RBPs and Pol II binding sites, and less markedly, in histone marks such as H3K36me3 and H3K4me3, open chromatin regions and TFBS (Supplementary Fig. 19b). The proportion of sQTLs overlapping splice sites and GWAS hits is not significantly different between the two types of introns.

These results suggest that splicing regulation occurs preferentially at *ps* introns. This is expected, since these introns are retained longer within the primary transcript, offering more opportunities for regulation through the interaction with RBPs and other factors, including chromatin-related ones.

sQTLs help to gain insight into disease and complex traits. To explore the relevance of regulatory variation affecting splicing in disease and complex traits, we assessed the overlap between GTEx sQTLs and the GWAS Catalog (https://www.ebi.ac.uk/gwas), extended to include variants in high LD ($r^2 \geq 0.80$) with the GWAS hits. sQTLs display a substantial enrichment, when compared to non-sQTLs, in variants associated with a wide variety of GWAS traits and diseases (median OR = 3.23). Among the diseases with the largest sQTL enrichment, we find many for which alternative splicing has been previously related to their pathophysiology (Supplementary Data 3). We integrated the enrichment information with estimates of semantic similarity between individual GWAS terms, computed from the Experimental Factor Ontology (EFO)[41]. Then, we applied multidimensional scaling (MDS) to summarize and represent the results (see Methods). This allowed us to identify the major

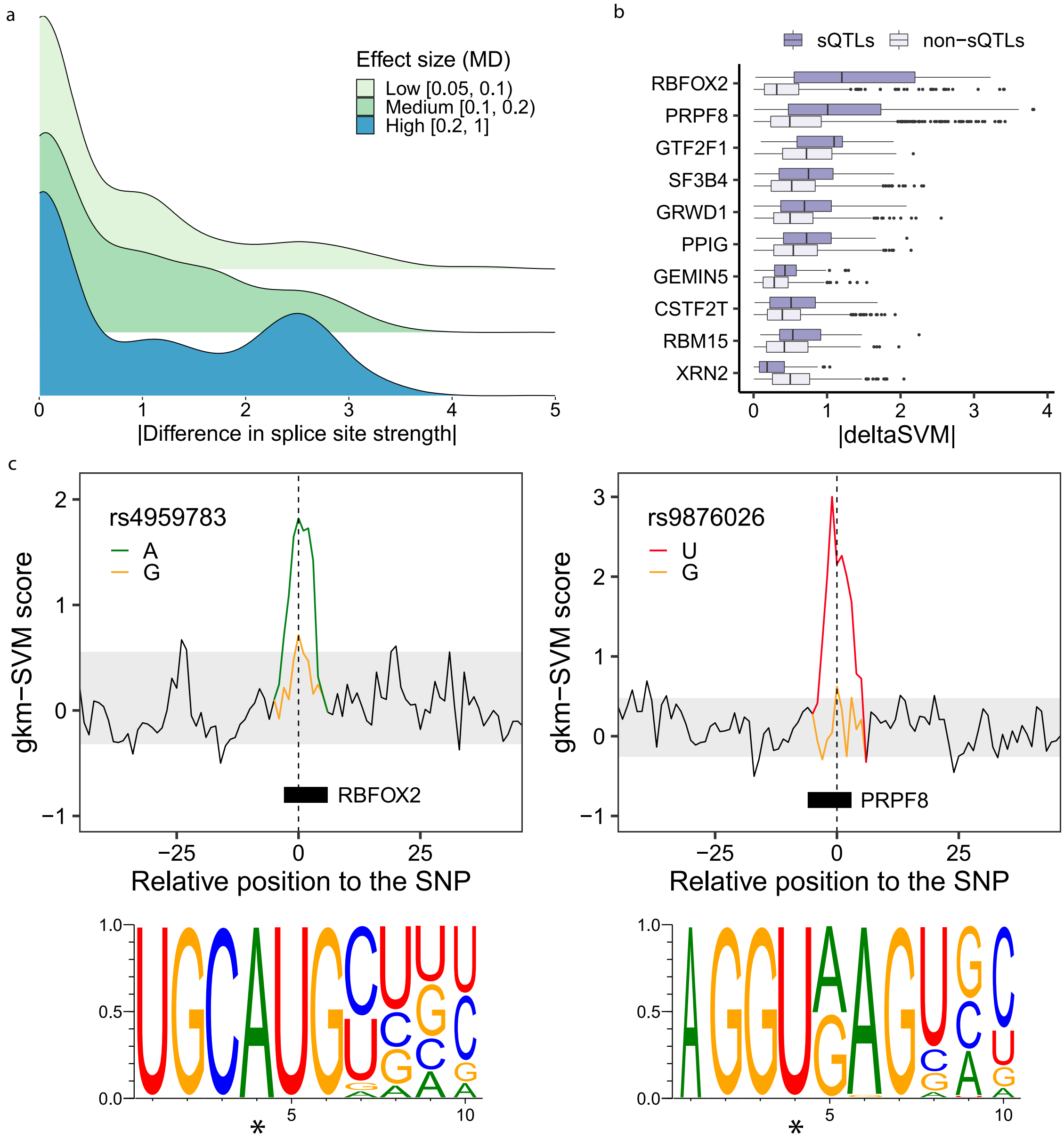
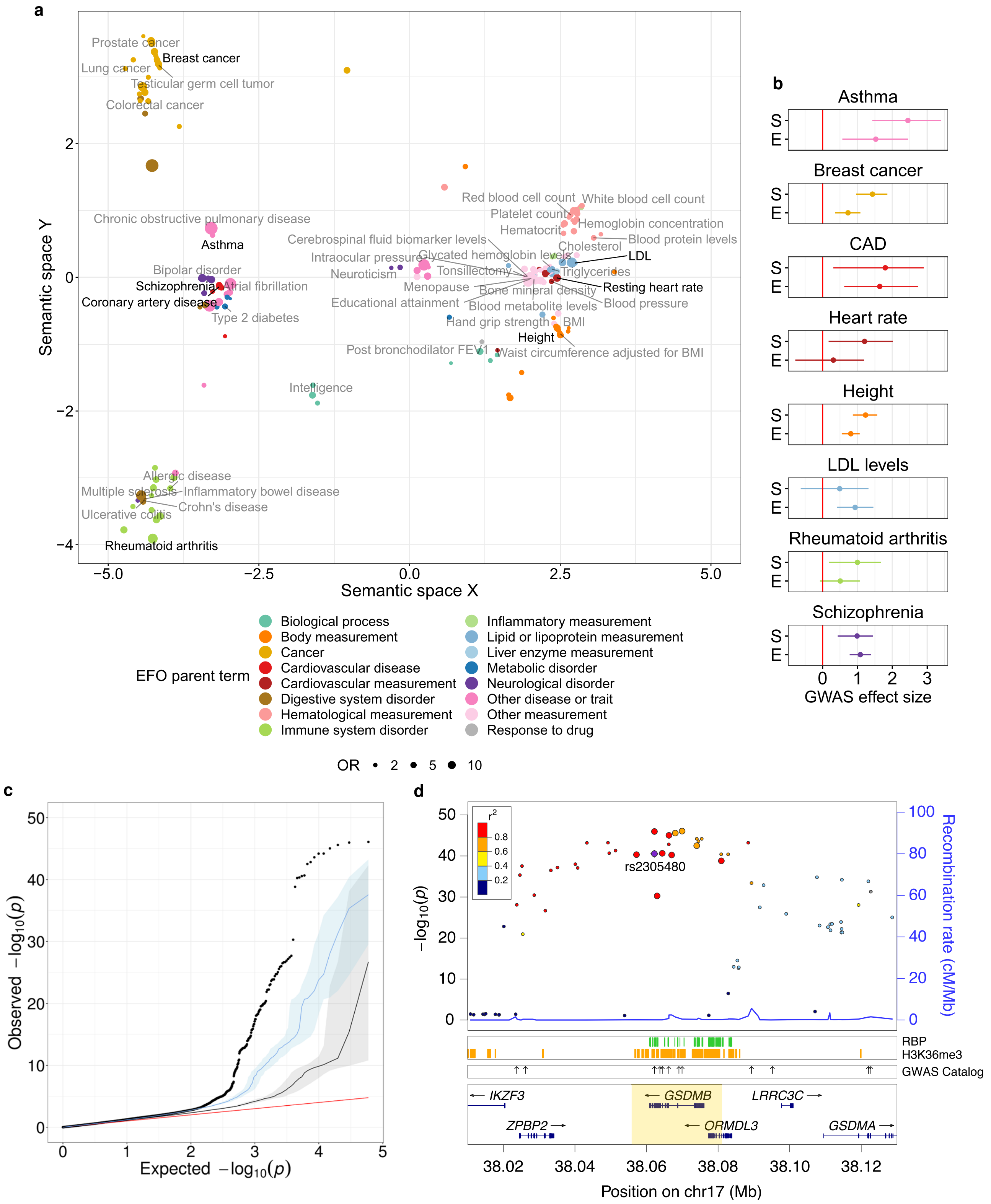


Fig. 4 Impact of sQTLs on splice sites and RBP binding sites. a Distribution of the absolute change in splice site strength for sQTLs with low, moderate and high effect sizes (MD value). **b** Distribution of the absolute deltaSVM value (|deltaSVM|) of sQTLs and non-sQTLs, for RBPs with significantly different mean |deltaSVM| between sQTLs and non-sQTLs (two-sided Wilcoxon Rank-Sum test, FDR < 0.05, total sample size for each test listed as follows: $n_{\mathrm{RBFOX2}} = 509$, $n_{\mathrm{PRPF8}} = 1{,}133$, $n_{\mathrm{GTF2F1}} = 376$, $n_{\mathrm{SF3B4}} = 595$, $n_{\mathrm{GRWD1}} = 736$, $n_{\mathrm{PPIG}} = 691$, $n_{\mathrm{GEMIN5}} = 345$, $n_{\mathrm{CSTF2T}} = 1{,}170$, $n_{\mathrm{RBM15}} = 373$ and $n_{\mathrm{XRN2}} = 450$). Data is shown as boxplots, where the box represents the first to third quartiles and the median, and the whiskers indicate ± 1.5 × interquartile range (IQR). **c** Modification of the binding sites of the RBPs RBFOX2 (left) and PRPF8 (right) by SNPs rs4959783 (chr6:3,260,093, G/A, |deltaSVM| = 2.48) and rs9876026 (chr3:11,849,807, T/G in the reverse strand, |deltaSVM| = 4.77), respectively. The lines represent the gkm-SVM scores of all possible (overlapping) 10-mers in a 100 bp window around the SNP. Those corresponding to the 10-mers overlapping the SNP are colored according to the allele. SNP positions are marked with a dashed line. The gray area includes the 90% middle gkm-SVM scores of 10-mers not overlapping the variant. The relative location of the predicted RBP motifs and the corresponding sequence logos are also displayed. In the logos, the SNP position is marked with an asterisk. Source data for **a**–**c** are provided as a Source Data file.

phenotype groups related to sQTLs. Trait measurements (right-hand side of the MDS plot) and diseases (left-hand side) are the two main groups of enriched GWAS terms observed (Fig. 5a). Within the latter, we identify subgroups corresponding to cancer, autoimmune diseases and other disorders (neurological, cardiovascular, metabolic, etc.).

We also compiled genome-wide GWAS summary statistics for a subset of enriched traits representative of the observed clusters: asthma[42], breast cancer[43], coronary artery disease[44], heart rate[45], height[46], LDL cholesterol levels[47], rheumatoid arthritis[48] and schizophrenia[49]. We further characterized the contribution to the disease phenotype of variants affecting splicing and variants affecting exclusively gene expression using fgwas[50] (see Methods). Overall, both types of variants display effect sizes significantly different from zero (Fig. 5b). Moreover, for some of the traits analyzed, including asthma, breast cancer, heart rate and height, we observe stronger GWAS associations among sQTLs than among variants affecting only gene expression (Fig. 5c and



Supplementary Fig. 20), suggesting that alterations in splicing might play a relevant role in the molecular mechanisms underlying these traits.

In addition, we observe that GWAS variants are especially enriched among sQTLs located in splice sites (OR = 2.66, two-sided Fisher's exact test *p*-value 1.02×10^{-9}) or within RBP binding sites (OR = 1.78, two-sided Fisher's exact test *p*-value < 10^{-16}). In particular, some of the traits and diseases with available summary statistics display stronger GWAS associations for sQTLs in RBP binding sites than for other sQTLs. Notably, this behavior seems to be trait/disease- and RBP-specific (Supplementary Fig. 21).

An interesting example of how sQTL mapping can help to gain insight into the mechanisms underlying GWAS associations is the case of asthma and the gene gasdermin b (*GSDMB*). Asthma displays the largest effect size for sQTL variants (maximum

Fig. 5 sQTLs and GWAS. a Multidimensional scaling-based representation of the semantic dissimilarities between GWAS traits and diseases whose associated variants are enriched among sQTLs with respect to non-sQTLs (two-sided Fisher's exact test FDR < 0.05). Each GWAS term is represented by a dot, whose size corresponds to the enrichment odds ratio (OR), and its color to the Experimental Factor Ontology (EFO) parent category the term belongs to. GWAS terms that lie close to each other are semantically similar. Eight representative traits with available summary statistics are highlighted. To help visualization, only the labels for the non-redundant, confidently and highly enriched terms are displayed (p-value < 10^{-8}, lower bound of the 95% confidence interval (CI) for the OR estimate > 1.5, width of the 95% CI for the OR estimate below the median). **b** Maximum likelihood estimates and 95% CIs (shown as dots and error bars, respectively) for the GWAS association effect size of variants affecting splicing (S), and variants affecting expression, but not splicing (E), for eight traits and diseases. **c** Quantile-quantile plot of p-values for association with asthma in[42] for sQTLs (black dots), eQTLs without effects on splicing (blue line and area), and variants with effects neither on expression nor on splicing (black line and gray area). Lines and colored areas represent, respectively, medians and middle 95% observed $-\log_{10}$ p-values across 10,000 random samplings from the corresponding variant set, with the same size as the sQTL set. The identity line is shown in red. **d** p-values for association with asthma in[42] (left y-axis) of variants in the region chr17:38,010,000-38,130,000, around the *GSDMB* gene (highlighted). The larger dots correspond to variants identified as sQTLs for the *GSDMB* gene in the lung. Linkage disequilibrium patterns (color-coded) and recombination rates are also displayed. The lower panels represent the location of RBP eCLIP peaks, H3K36me3 marked-regions and other GWAS Catalog associations with asthma (shown as arrows). The highlighted variant (rs2305480) is in perfect LD with rs11078928, previously shown to have an impact on *GSDMB* splicing[54]. Source data for **a**–**d** are provided as a Source Data file.

likelihood estimate = 2.32, Fig. 5b), and stronger associations for sQTLs than for variants affecting only gene expression, or variants affecting neither expression nor splicing (Fig. 5c). Indeed, we identified over 850 sQTLs co-localizing with known asthma loci, affecting the splicing patterns of genes related to immunity, including interleukins and immune cell receptors (*IL13*, *TLSP*, *IL1RL1*, *TLR1*), major histocompatibility complex components (*HLA-DQA1*, *HLA-DQB1*) or interferon-activated transcription factors (*IRF1*). However we also found other genes, such as *GSDMB*, with a priori less clear roles in the pathophysiology of the disease.

The *GSDMB* locus (17q21) has been consistently identified as a contributor to genetic susceptibility to asthma[42] and other autoimmune diseases, such as type 1 diabetes[51], ulcerative colitis[52] or rheumatoid arthritis[53]. Although its exact function is unknown, *GSDMB* is highly expressed in human bronchial epithelial cells in asthma[54,55], and it is known that overexpression of the human *GSDMB* transgene in mice induces an asthma phenotype[55]. In addition, the lipid-binding N-terminal domain of GSDMB and other gasdermins causes pyroptotic cell death[56], potentially leading to the release of inflammatory molecules that trigger the asthma pathophysiology.

GSDMB is an sGene in 39 GTEx tissues, including lung (sGene FDR = 1.42×10^{-10}, median MD = 0.22). Indeed, sQTLs for *GSDMB* are among the top associated variants with asthma in[42] (Fig. 5d). Allele C of the splice acceptor variant rs11078928 (chr17:38064469, T/C) has been shown to lead to the skipping of exon 6, which encodes 13 amino acids in the N-terminal domain, disrupting its pyroptotic activity[54]. While the major allele (T) is associated with a higher incidence of asthma, the C allele confers a lower asthma risk[54]. We have identified rs11078928 as an sQTL for *GSDMB*, whose alternative allele C precisely promotes expression of isoforms *GSDMB-001* and *GSDMB-002* (exon 6 skipping) *vs* isoform *GSDMB-003* (exon 6 inclusion) (Supplementary Fig. 22).

Discussion

Using the unprecedented resource generated by the GTEx Consortium, we have obtained and analyzed a comprehensive set of genetic variants in the human genome affecting transcript isoform abundances (splicing QTLs, sQTLs). Unlike most methods for sQTL detection, we use a multivariate approach that monitors global changes in the relative abundances of a gene's transcript isoforms, rather than targeting specific splicing events. Leveraging the correlated structure of isoform abundances is likely to result in increased power for sQTL mapping. Indeed, our approach has demonstrated the ability to detect sQTLs associated with complex splicing events that often escape univariate approaches[20]. In addition, we show that our method is not restricted to the analysis of transcript abundances, but can also accommodate other AS phenotypes, such as LeafCutter's intron excision ratios[18]. A comparison of the resulting sQTLs obtained employing the two types of input data highlights the complementarity between global and local views of alternative splicing, especially regarding the types of splicing events identified[20,29]. We also compared our sQTL set (obtained using RSEM + sQTLseekeR2) with the one generated by the GTEx Consortium (LeafCutter + FastQTL)[12] and observed analogous differences in the nature of the splicing events identified, showing that the splicing phenotype employed for sQTL mapping is a major determinant of the resulting sQTL catalog. Moreover, while the overlap between the two sets is moderate, the sQTLs identified exclusively by either approach differ in a number of biological features, both at variant and target gene level, underlining the complementarity between the two sQTL mapping pipelines.

We have surveyed a large collection of tissues. Our analyses show that sQTLs tend to be highly shared, suggesting that there is a core set of variants that are involved in the regulation of splicing independently of the tissue or cell type. This has also been recently reported by the GTEx Consortium[12]. Among the genes whose splicing is regulated by genetic variants (i.e. sGenes), there is a consistent enrichment of functions related to RNA processing, maybe reflecting splicing autoregulation. Indeed, several positive and negative autoregulation and cross-regulation mechanisms, such as coupling to nonsense-mediated decay, have been proposed for a large number of splicing factors[26].

Overall, we found fewer genes regulated at splicing than at expression level. This is in line with the smaller contribution of splicing, compared to gene expression, to the global variability in transcript abundances across tissues and individuals[57,58]. Although this observation could be also due to the different preprocessing steps and statistical methodologies applied in both analyses, it is consistent with recent reports that use the same approach to map both sQTLs and eQTLs in the GTEx dataset[12]. In addition, many variants seem to be involved simultaneously in the regulation of both processes. This is not surprising, given that there is a substantial interplay between the molecular mechanisms underlying splicing and transcription, and because splicing often takes place co-transcriptionally[6]. Moreover, variants altering splicing can affect RNA stability and, consequently, gene expression[59].

In this regard, we have observed that introns that are spliced post-transcriptionally (*ps*) tend to be more enriched in sQTLs than introns that are spliced co-transcriptionally (*cs*). This is somehow expected, as *ps* introns are retained longer within the primary transcript, offering more opportunities for splicing regulation. Consistent with this, sQTLs in *ps* introns display a larger enrichment, compared to sQTLs in *cs* introns, in RBP binding sites, but also in Pol II binding sites and histone marks. We note that chromatin-related features play a prominent role in co-transcriptional splicing, often through the regulation of transcription[6]. However, not fully spliced but already 3'-end mature transcripts are present in the fraction of RNA attached to chromatin[60,61]. In this context, interactions between chromatin-side features and not fully spliced transcripts can occur post-transcriptionally. Indeed, similar enrichments have been reported for exons that are spliced more slowly[40]. Overall, it seems that post-transcriptionally spliced introns play a larger role in splicing regulation than introns quickly spliced during transcription.

In addition to variants that are sQTLs and eQTLs for the same gene, we have found many variants that are sQTLs for a gene and eQTLs for a different one. In order to rule out indirect regulatory effects (e.g. when the variant directly affects the expression – splicing – of one gene, and the product of this gene directly affects the splicing – expression – of the other gene), we considered each effect (splicing or expression) occurring in different tissues. Since our multivariate approach is not compatible with currently available co-localization methods (see below), we cannot distinguish the cases in which the two effects are indeed caused by the same variant or by two different variants in LD. Regardless of the underlying causal structure, these variants uncover regulatory loci, which we termed heteropleiotropic, that would be involved in the coordinated regulation, through different mechanisms, of different genes which otherwise do not appear to directly interact. Thus, heteropleiotropic loci could reveal regulatory relationships between genes that may not be easily captured by co-expression or splicing networks, highlighting the complexity of the gene regulation program in eukaryotes. While we provide biological evidence supporting this phenomenon, our results should be considered preliminary. In particular, we lack a specific statistical test to assess whether a given variant has a regulatory effect on the expression of a gene only in one tissue and, simultaneously, on the splicing of a different gene only in a second tissue. Further work is thus required to investigate the biological relevance of this phenomenon.

Our study also helps to understand the molecular mechanisms through which genetic variants impact splicing. Two such mechanisms appear to be the most relevant. On the one hand, direct impact on donor and acceptor splice sites. On the other hand, modification of binding sites of a wide variety of transcriptional regulators, especially RBPs, which are major players in RNA processing, transport and stability[5,62]. While the latter seems to occur in a larger number of cases, the former often leads to stronger effects on splicing. However, in many cases both mechanisms are likely to cooperate, given that RBPs often bind near splice sites.

Finally, our work provides new insights into the relationship between genetic variation, splicing and phenotypic traits. Specifically, we found that sQTLs are enriched in variants associated with a number of complex traits and diseases, some of them previously reported[9,10,14,15]. sQTLs display stronger GWAS associations than variants not associated with splicing and, for some traits, even stronger than variants affecting exclusively gene expression. This grants splicing a key role in mediating the impact of genetic variation in human phenotypes[15]. Because gene expression is the main driver of biological function, we hypothesize that genetic variants affecting expression are likely to have a much larger biological impact than those affecting splicing: often, they could be lethal during development. In contrast, genetic variants affecting splicing may have subtler effects, therefore being better tolerated and leading more frequently to observable phenotypes. That genetic variants affecting splicing may underlie most human hereditary diseases has already been pointed out[19]. Especially relevant seems to be the implication of sQTLs in the mechanisms underlying autoimmune diseases, also supported by the overrepresentation of immune functions among sGenes. Actually, sQTLs have been recently proposed as relevant players in human immune response and its evolution[16]. In addition, sQTLs altering RBP binding seem to play a prominent role in disease. Indeed, the relevance of RBPs in human disorders has been often remarked[62].

A more detailed analysis of the relationship between sQTLs and GWAS variants could be achieved by the usage of statistical methods to assess co-localization[63–65], and subsequent fine-mapping of the sQTL candidates[66–68] to assign causal probabilities. However, currently available methods are not directly applicable within our multivariate, non-parametric framework. In addition, recent works have demonstrated the utility of in silico splicing predictors to identify pathogenic variants affecting splicing, especially in the case of Mendelian disorders[69–71]. These methods provide a complementary view to RNA-seq-based approaches that measure splicing changes associated with genetic variants, such as sQTLseekeR2. Indeed, while the former target rare variants in the vicinity of splice sites with strong phenotypic effects, the latter focus on common regulatory variation, not restricted to the splice region nor necessarily pathogenic. Furthermore, the ability of pathogenicity predictors to account for features such as evolutionary conservation or exon importance provides valuable information about the relevance of individual alleles[71], which may help prioritize sQTLs in clinical settings.

In summary, our implementation of an enhanced pipeline for sQTL mapping based on sQTLseekeR2, Nextflow and Docker will help sQTL discovery in multiple datasets, across different platforms, in a highly parallel and reproducible manner. Here we have employed it to identify sQTLs in the GTEx dataset. The extensive catalog of sQTLs generated constitutes a highly valuable resource for the field. As our initial analyses already show, this resource will contribute to the understanding of the mechanisms underlying alternative splicing regulation and its implication in phenotypic traits, including disease risk.

Methods

GTEx data. Transcript expression (transcripts per million, TPM) and variant calls (SNPs and short indels) were obtained from the V7 release of the Genotype-Tissue Expression (GTEx) Project (dbGaP accession phs000424.v7.p2). These correspond to 10,361 samples from 620 deceased donors with both RNA-seq in up to 53 tissues and Whole Genome Sequencing (WGS) data available. Metadata at donor and sample level and variant annotations (Ensembl's Variant Effect Predictor, VEP, v83, http://www.ensembl.org/info/docs/tools/vep, with the Loss-Of-Function Transcript Effect Estimator extension, LOFTEE, https://github.com/konradjk/loftee) were also retrieved. Data from dbGaP was downloaded using IBM Aspera Connect v3.6.1 (the rest of the data downloaded for this work was obtained via GNU Wget v1.14). In GTEx V7, RNA-seq reads are aligned to the human reference genome (build hg19/GRCh37) using STAR[72] v2.4.2a, based on the GENCODE v19 annotation (https://www.gencodegenes.org/human/release_19.html). Transcript-level quantifications are obtained with RSEM[73] v1.2.22. WGS reads are aligned with BWA-MEM (http://bio-bwa.sourceforge.net) after base quality score recalibration and local realignment at known indels using Picard (http://broadinstitute.github.io/picard). Joint variant calling across all samples is performed using GATK's HaplotypeCaller v3.4 (https://software.broadinstitute.org/gatk/documentation/tooldocs). Further details on GTEx data preprocessing and QC pipelines can be found on the GTEx Portal (https://gtexportal.org).

sQTL mapping

Gene, transcript and variant filtering. 48 tissues with sample size $n \geq 70$ were selected for *cis* sQTL mapping. The *cis* window was defined as the gene body plus 5 Kb upstream and downstream the gene boundaries, and provided a good balance between the number of variants analyzed per gene and the computation time (see also [20], [27] and [29]). We considered genes expressed ≥ 1 TPM in at least 80% of the samples (samples with lower gene expression were removed from the analysis of the gene), with at least two isoforms and a minimum isoform expression of 0.1 TPM (transcripts with lower expression in all samples were removed).

These filters correspond to the default parameters of sQTLseekeR2. We analyzed only biallelic SNPs and short indels (autosomal + X) with MAF ≥ 0.01 and at least 10 samples per observed genotype group. In total, 3,588,609 variants and 16,010 genes (15,195 protein-coding, 815 lincRNA) were analyzed.

Covariates included. To evaluate the impact of known technical and biological covariates at sample and donor level on transcript relative abundances, we regressed the first ten principal components (PCs) of the transcript relative expression matrix per tissue onto each available covariate, determining the percentage of variance explained (R^2_{adj}). We selected donor ischemic time, gender and age, as well as sample RIN (RNA integrity number), as the most relevant covariates. We also included the first three genotype PCs (obtained from dbGap), to control for population (i.e. ancestry) effects, and the genotyping platform employed (Illumina HiSeq 2000 or HiSeq X). Selected covariates were regressed out from the relative abundances of each gene's transcript isoforms by sQTLseekeR2 before testing for association with the genotype.

Software. For sQTL mapping we employed sQTLseekeR2 v1.0.0, an enhanced version (see also Supplementary Note 1) of the sQTLseekeR R package[20], which identifies genetic variants that are associated with multivariate changes in the relative abundances of a gene's transcript isoforms (i.e. splicing ratios). sQTLseekeR2 was embedded in sqtlseeker2-nf, a highly parallel, portable and reproducible pipeline for sQTL mapping developed using Nextflow[23], a framework for computational workflows, and Docker container technology. sQTLseekeR2 and sqtlseeker2-nf are available at https://github.com/guigolab.

Details on significance assessment. We performed *cis* sQTL mapping on each tissue. Nominal *p*-values were obtained using the function `sqtl.seeker`. To correct for the fact that multiple genetic variants in LD were tested per gene, an adaptive permutation scheme was applied (implemented in the function `sqtl.seeker.p`). A Benjamini-Hochberg false discovery rate (FDR) threshold of 0.05 was selected to identify sGenes. To retrieve all significant variant-gene pairs, we employed a procedure identical to the one used in ref. [24] for expression QTLs (implemented in the function `sqtls.p`). See Supplementary Note 1 for details. In addition, as our test statistic is sensitive to the heterogeneity of the splicing ratios' variability among genotype groups, a multivariate homoscedasticity test[74] was also performed for each variant-gene pair. Pairs failing this test (FDR across all nominal tests < 0.05) were still assessed for significance and taken into account for multiple testing correction, but they were not reported as significant sQTLs.

Cell type heterogeneity. We employed xCell[25] to estimate the enrichment of 64 reference cell types from the bulk expression profile of each GTEx sample. We applied the `xCellAnalysis` function in the xCell R package to the full gene expression TPM matrix (56,205 genes × 11,688 samples), in order to maximize tissue heterogeneity. We then applied the τ index[31] (see also section *sQTL sharing*) to median xCell enrichments across samples per tissue. The cell type heterogeneity of a tissue was estimated as $1 - \tau$. While these results should be interpreted with caution, as xCell is not a deconvolution method but an enrichment method, they were generally biologically meaningful. For example, the most homogeneous tissues included brain subregions and transformed fibroblasts, and the most heterogeneous, spleen and whole blood. To determine the impact of the cell type heterogeneity of a tissue on sQTL discovery, we computed the partial correlation between the number of sGenes over the number of tested genes and the estimated cell type heterogeneity (i.e. $1 - \tau$), controlling for the tissue sample size.

sQTL effect size. We used the absolute maximum difference (MD) in mean adjusted transcript relative expression between genotype groups as a measure of the size of the effect. MD takes values in the interval [0, 1]. In practice, usual MD values belong to [0.01, 0.4]. As a general rule, we considered MD values < 0.1 as small effect sizes, MD values between 0.1 and 0.2 as moderate effect sizes, and MD values greater than 0.2 as large effect sizes. sQTLs with MD values below 0.05 were not taken into account for further analyses (default in sQTLseekeR2).

GO enrichment of sGenes. For each tissue, we obtained the corresponding set of sGenes, and performed hypergeometric tests to assess Gene Ontology (GO) Biological Process (BP) term over-representation, selecting as gene universe all tested genes. We set a FDR threshold of 0.1 to identify significantly enriched terms. Similarly, we selected genes that were not sGenes in any tissue, and performed a hypergeometric test to assess GO BP term over-representation in this set (FDR < 0.1, universe: all tested genes). Then, we employed REVIGO[75] (http://revigo.irb.hr, with parameters: allowed similarity = 0.9, database = *H. sapiens*, semantic metric = SimRel) to remove highly redundant terms and generate semantic similarity-based GO term representations for sGenes and non-sGenes.

sQTL replication. To assess replication of GTEx sQTLs, we examined the *p*-values for matched variant-gene pairs identified as splicing QTLs by sQTLseekeR for three immune cell types (CD14$^+$ monocytes, CD16$^+$ neutrophils, and naive CD4$^+$ T cells) in the Blueprint Project[27]. Both studies have large differences in RNA sources (tissues in GTEx *vs* cell types in Blueprint), library preparation (unstranded polyA$^+$ *vs* stranded Ribo-Zero), sequencing strategy (e.g. paired-end *vs* single-end in monocytes and neutrophils) and data processing pipelines (e.g. different transcript quantification software). π_1 statistics, that provide an estimate of the proportion of true positives[76], were computed for each pair GTEx tissue/Blueprint cell type. A final replication rate for each GTEx tissue was calculated as the average π_1 value across the three Blueprint cell types.

Alternative splicing events associated with sQTLs. To determine the nature of the splicing events associated with sQTLs we selected, for each sQTL, the two isoforms of the target sGene that changed the most (in opposite directions) across genotypes. Then, we compared the exonic structure of both transcripts using the function `classify.events` of sQTLseekeR, which extends the classification proposed in AStalavista[77]. We considered the same event categories depicted in[20]: exon skipping, alternative 5′ and 3′ splice sites, intron retention, mutually exclusive exons, alternative first and last exons, alternative 5′ and 3′ UTRs, tandem 5′ and 3′ UTRs, complex splicing events (complex combinations of events affecting internal exons) and complex 5′/3′ events (complex combinations of events affecting 5′/3′ termini). Some of these events are not explicitly involving splicing, but alternative transcription initiation and termination sites. Note that each transcript pair, and therefore each sQTL, can be associated with more than one event.

Heteropleiotropy and ENTEx histone modification analysis. Given a genetic variant v and a pair of genes (i.e. g_1 and g_2) and tissues (i.e. t_1 and t_2), we define v as heteropleiotropic with effects in different tissues if (i) v is an sQTL – but not an eQTL – for gene g_1 in tissue t_1, (ii) v is an eQTL – but not an sQTL – for gene g_2 in tissue t_2, (iii) v is neither an sQTL nor an eQTL for gene g_2 in tissue t_1 and (iv) v is neither an sQTL nor an eQTL for gene g_1 in tissue t_2. Note that here we propose a definition rather than a thorough methodological approach to determine whether a locus displays (or not) this behavior. To the best of our knowledge, currently available statistical tests cannot be used to assess this phenomenon. Hence, our identification of heteropleiotropic loci should be considered a first approximation based on indirect evidence. Out of 148,618 variants tested for association with both the expression and splicing of at least two genes in at least two tissues, we identified 6,552 heteropleiotropic cases. In order to evaluate whether changes at epigenetic level were occuring at these positions, we obtained ChIP-seq peaks corresponding to six histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3 and H3K9me3) from the ENTEx data collection of the ENCODE Project[78,79] (https://www.encodeproject.org, accessed 2019-10-04, accession numbers and URLs provided in Supplementary Data 4). ENTEx is a joint effort between the GTEx and ENCODE consortia to deeply profile overlapping tissues from the same four donors (two male, two female) using shared technologies. The two tissues of interest were available for at least three out of four ENTEx donors for 2,855 heteropleiotropic variants. By overlapping these with the ChIP-seq peaks in the corresponding tissues, we identified 699 cases where one or more histone marks present in a tissue were absent in the other (in at least three donors). We compared this number with the one obtained for variants v' affecting both the splicing and expression of the two genes (g_1 and g_2) in the two tissues (t_1 and t_2), using two-sided Fisher's exact test for significance assessment.

sQTL sharing. For every pair of tissues, we selected variant-gene pairs tested in both and found significant in at least one. We computed Pearson correlation (r) between their effect sizes (MD values). Tissue specificity was estimated as $s_t = 1 - \bar{r}_t$, where $\bar{r}_t$ is the mean correlation between a given tissue t and the others. To determine the robustness of the observed sharing patterns with changes in the sample size, we randomly downsampled every original tissue dataset once to 100, 200 and 300 samples, ran our sQTL mapping pipeline again and re-evaluated the sharing patterns. Alternatively, we computed the Jaccard index on the sets of variant-gene pairs identified in every pair of tissues. In this case, tissue specificity estimates corresponded to $1 - \bar{j}_t$, where $\bar{j}_t$ is the mean Jaccard index between a given tissue t and the others.

We further compared these approaches with a third strategy, aimed at evaluating the changes in the whole splicing phenotype due to sQTLs between different tissues, rather than relying on MD values or sQTL presence/absence. This allows more flexibility, likely resulting in an increased ability to capture complex sharing patterns. In short, we focused on variant-gene pairs tested in all tissues and found significant in at least one. For every tissue t_i, variant-gene pair $j \in \{1 \ldots p\}$, and genotype group $k \in \{0, 1, 2\}$, we computed the centroid of the adjusted (square root transformed, covariate corrected) splicing ratios, $\mathbf{c}_{t_i jk}$. Then, we obtained:

$$d(t_1, t_2) = \frac{1}{p} \sum_{j=1}^{p} \sum_{k=0}^{2} \| \mathbf{c}_{t_1 jk} - \mathbf{c}_{t_2 jk} \| \qquad (1)$$

where d measures the distance between any two tissues (t_1 and t_2) in terms of sQTL sharing, as the mean (across variant-gene pairs) of the sum (across genotype groups) of the Euclidean distance between centroids ($\|\mathbf{x}\|$ represents the Euclidean norm of vector $\mathbf{x}$). Small values of d correspond to large sQTL sharing, and vice versa (Supplementary Fig. 9a illustrates the behavior of d for a single variant-gene pair evaluated in 4 tissues). A distance matrix built upon d values was then employed as input for hierarchical clustering.

To compare the tissue clusters obtained using different approaches we computed Baker's Gamma (Γ), a metric of similarity between two dendrograms given by the rank correlation between the stages at which pairs of objects combine in each of the two trees[80]. Γ ranges from -1 to 1, with values close to 1 corresponding to high similarity between the two dendrograms. To assess the significance of this similarity, we performed a permutation test (shuffling the labels of one tree 10,000 times, keeping tree topologies constant). We also employed Baker's Gamma to compare our trees with the one obtained using mashr[30] for LeafCutter sQTLs in GTEx V8[12], available at https://github.com/broadinstitute/gtex-v8.

Of note, we employed pairwise approaches to study sQTL sharing, rather than methods to analyze QTL sharing jointly across tissues (such as mashr, to cite an example), given that the latter, to the best of our knowledge, cannot be applied in our multivariate, non-parametric setting.

In addition, for each sGene tested in all tissues and found significant in at least one, we determined tissue specificity of the sGene expression, using the τ index[31]:

$$\tau = \frac{\sum_{t=1}^{n}(1-\hat{x}_t)}{n-1}; \quad \hat{x}_t = \frac{x_t}{\max\limits_{1 \le t \le n}(x_t)} \tag{2}$$

where x_t is the expression of the gene in tissue t and n the number of tissues. τ takes values between 0 (i.e. genes equally expressed in all tissues) and 1 (i.e. tissue-specific genes). We calculated τ using median gene expression across tissues. In addition, to assess tissue specificity of splicing regulation, we computed a variation of τ, τ_s, where x_t was the $-\log_{10}$(FDR) of the sGene in tissue t. For sGenes in the top 20 percentile of the distribution of τ_s values, and the bottom 20 percentile of the distribution of τ values, we evaluated GO BP term over-representation (hypergeometric test, FDR < 0.1, universe: all sGenes).

Functional enrichment of sQTLs. ChIP-seq peaks (transcription factor binding sites, histone marks) and open-chromatin regions were obtained from the Ensembl Regulation dataset (ftp://ftp.ensembl.org/pub/grch37/release-86/regulation/homo_sapiens/AnnotatedFeatures.gff.gz). eCLIP peaks in HepG2 and/or K562 cell lines for 113 RBPs[34] were obtained from the ENCODE Project[78,79] (https://www.encodeproject.org, see section *sQTL impact on splice site strength and RBP binding sites* for details). Disease and complex-trait associated variants were retrieved from the GWAS Catalog (https://www.ebi.ac.uk/gwas, accessed 2018-09-18), extended to GTEx variants in high linkage disequilibrium ($r^2 > 0.8$) with the GWAS hits. Protein coding and lincRNA exons were derived from the GENCODE v19 annotation. The coordinates of these functional elements were overlapped with all the tested variants (either sQTLs or not) to obtain a functional annotation per variant. The functional consequences of each variant (stop-gained, frameshift, etc.), computed by the Variant Effect Predictor (VEP, http://www.ensembl.org/info/docs/tools/vep), were obtained from dbGap (accession phs000424.v7.p2). Note that the VEP leverages the Ensembl Variation dataset, which contains data from a wide variety of sources (https://www.ensembl.org/info/genome/variation/species/sources_documentation.html). From the VEP result we also identified variants with high impact or in the categories: probably damaging (PolyPhen, http://genetics.bwh.harvard.edu/pph2), deleterious (SIFT, https://sift.bii.a-star.edu.sg), pathogenic (ClinVar, https://www.ncbi.nlm.nih.gov/clinvar) and high-confidence loss-of-function (LOFTEE, https://github.com/konradjk/loftee).

The top ten most significant sQTLs per gene and tissue were compared to a null distribution of 1,000 sets of randomly sampled variants not associated with splicing (FDR > 0.05, i.e. non-sQTLs), of the same size of the sQTL set. The top ten were selected to ensure the coverage of the less common annotations. Non-sQTLs were matched to sQTLs in terms of relative location within the gene and MAF. Specifically, we selected non-sQTLs so that they were located in the same bins (see section *sQTL location*) within the genes for which they were not sQTLs, as sQTLs within the genes for which they were sQTLs, and had MAFs equal to the sQTLs' MAFs ± 0.02. The enrichment was calculated as the odds ratio (OR) of the frequency of a certain annotation among sQTLs to the mean frequency of the same annotation across the 1,000 non-sQTLs sets. To ensure enrichment reliability, we filtered out annotations with a mean frequency across the non-sQTLs sets lower than five. The significance of each enrichment was assessed using a two-sided Fisher's exact test. p-values were corrected for false discovery rate, selecting a threshold of FDR < 0.05. Enrichments in a subset of relevant features, such as high impact/potentially damaging variants, splice sites, GWAS hits, exons, TFBS (all TFs pooled together), RBP binding sites (all RBPs pooled together), Pol II binding sites, HK36me3 and open chromatin regions, were also carried out separately for high effect size (MD ≥ 0.2) and low effect size (MD < 0.1) sQTLs.

sQTL location. We divided every sGene body into 20 bins of equal size and assigned each sQTL to the corresponding bin according to its location. The number of bins (20) was selected in order to provide a good balance between granularity and bin size. We computed the mean proportion of sQTLs (with respect to the total number of sQTLs per gene) on each bin. An identical procedure was applied to exons, introns, downstream and upstream regions. In each case, to ensure a minimum bin size, we filtered out the 20% shortest regions. Under the null hypothesis of no preference in location, a uniform distribution for the mean proportion of sQTLs across bins was expected.

sQTL impact on splice site strength and RBP binding sites. To estimate the impact of genetic variants on splice sites, for each variant (either sQTL or not) within the sequence of an annotated splice site we scored the site considering the reference and the alternative allele, using PWMs built upon human splice sites[81]. High scores corresponded to common/strong splice sites, while low scores corresponded to rare/weak sites, probably leading to less efficient splicing. Then we estimated the change in splice site strength as the absolute value of the difference between alternative and reference scores.

To estimate the impact of genetic variants on RBP binding sites, we obtained eCLIP peaks in HepG2 and K562 cell lines for 113 RBPs[34] from the ENCODE Project[78,79] (https://www.encodeproject.org, accessed 2018-04-16, accession numbers and URLs provided in Supplementary Data 5). For each RBP, we selected the peaks significant at FDR < 0.01 and with a fold-change (FC) with respect to the mock input ≥ 2. We further required a minimum overlap between replicates (50% of the length of the union of a given pair of peaks). This constituted our positive set of RBP-binding sequences. We generated an equally-sized negative set of matched (in terms of GC content, length and repeats) sequences, not overlapping eCLIP peaks from the same RBP. We combined both sets of sequences to build our training set. To achieve feasible memory usage and running times, we limited the size of the training set to 30,000 sequences.

We then trained a gapped k-mer support vector machine (gkm-SVM)[37] with default parameters (word length $l = 10$, informative columns $k = 6$), as recommended for our training set size range[36]. Other choices of l and k barely changed the overall performance (Supplementary Fig. 23). The option `addRC` (add reverse complementary) was set to `FALSE` as we were working with RNA sequences. The classification performance was evaluated using a 5-fold cross-validation. 79 RBPs with a mean cross-validation area under the Receiver Operating Characteristic curve (ROC AUC) ≥ 0.8 were kept. To predict the impact of variants on RBP binding, for all the variants overlapping the eCLIP peaks (FDR < 0.01, FC ≥ 2) of a given RBP, we computed the deltaSVM metric[36]. The gkmSVM assigns a weight to each possible 10-mer, quantifying its contribution to the prediction of RBP binding. Each variant is given a score computed as the sum of the weights of the 10-mers overlapping it (10-mer SVM scores were used as a proxy for weights). deltaSVM computes the difference between the score of the alternative and the reference allele, quantifying their difference in predictive potential. Here we used the minor and the major allele instead of the alternative and the reference allele, respectively.

We focused on the most predictive variants of the binding of each RBP (score of the variant at either allele among the 5% highest scores for this RBP). This was done to target those variants lying on sequences likely to be highly relevant for RBP binding (i.e. potential binding sites). To ensure the robustness of our results, we further required at least 30 sQTLs with deltaSVM values per RBP, resulting in a final set of 32 RBPs. Of these, for 10 RBPs with significantly different |deltaSVM| values between sQTLs and non-sQTLs (two-sided Wilcoxon Rank-Sum test, FDR < 0.05), we obtained the 100 highest-scoring 10-mers, aligned them using mafft v7.407 (high accuracy mode L-INS-i)[82], removed the columns of the alignment with more than 50% of gaps and built sequence logos using WebLogo standalone v3.6.0 (http://weblogo.threeplusone.com).

To evaluate allele-specific RBP binding (ASB), we obtained the ASB variants identified in the same eCLIP dataset using BEAPR (Binding Estimation of Allele-specific Protein-RNA interaction), available from Yang et al.[38]. In short, BEAPR is a method to identify ASB events in protein-RNA interactions from eCLIP data. It accounts for crosslinking-induced sequence propensity and variability between replicates, outperforming commonly used count-based approaches. We only considered ASB variants for which the same alleles had been genotyped in GTEx. We focused on sQTLs, non-sQTLs and ASB variants overlapping eCLIP peaks (FDR < 0.01, FC ≥ 2) for any of the 113 RBPs of interest in HepG2 and/or K562 cell lines. We assessed the significance of the difference in the proportion of sQTLs and non-sQTLs overlapping ASB variants across RBPs using two-sided Fisher's exact test. We also performed this analysis separately for each RBP, using false discovery rate for multiple testing correction (FDR < 0.05).

Co- and post-transcriptional splicing. We obtained RNA-seq data from nuclear and cytoplasmic fractions (2 replicates per fraction) corresponding to 13 cell lines available from the ENCODE project[78,79] (https://www.encodeproject.org, accessed 2018-05-25, accession numbers and URLs provided in Supplementary Data 6). A Nextflow implementation of the Integrative Pipeline for Splicing Analyses (IPSA), developed in-house (https://github.com/guigolab/ipsa-nf), was employed to determine the number of reads supporting splicing completion and splicing incompletion, for each intron annotated in GENCODE v19. We excluded from this analysis introns that overlapped either exons or non-identical introns in terms of chromosome, start and end positions. To assess the significance of the difference in the proportion of reads supporting splicing completion between nuclear and cytoplasmic compartments we employed two-sided Fisher's exact test. False discovery rate was employed for multiple testing correction (FDR < 0.05). Introns with significantly larger proportions of reads supporting splicing completion in the cytoplasm were classified as post-transcriptionally spliced (here referred to as *ps*). Introns that did not pass the FDR threshold were labeled as either unprocessed (i.e.

intron retention events) or co-transcriptionally spliced (here referred to as *cs*), depending on the degree of splicing completion in both cellular compartments. We focused on introns consistently classified as either *ps* or *cs* in at least ten of the analyzed cell lines. We computed variant density (number of variants per Kb of intron) at ten bins of equal size along both types of introns (ten was selected to ensure that enough variants were present in each bin). We also assessed the enrichment in functional elements of sQTLs in *ps* introns with respect to sQTLs in *cs* introns using two-sided Fisher's exact test. False discovery rate was employed for multiple testing correction (FDR < 0.05).

GWAS analyses. We downloaded the GWAS Catalog, including the Experimental Factor Ontology (EFO) annotations for the GWAS terms (https://www.ebi.ac.uk/gwas, accessed 2018-09-18). We used LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert variant coordinates from hg38 to hg19 and PLINK v1.90b6.2 (https://www.cog-genomics.org/plink2) to extend the Catalog to the variants in high linkage disequilibrium ($r^2 \geq 0.8$) with the GWAS hits. The sQTL enrichment was calculated as the odds ratio (OR) of the frequency of GWAS variants among sQTLs to the mean frequency of GWAS variants across 1,000 matched non-sQTL sets (see section *Functional enrichment of sQTLs*). In parallel, we obtained the complete EFO ontology (https://www.ebi.ac.uk/efo) in Open Biomedical Ontologies (OBO) format. For the GWAS terms with an OR > 1, we used the ontologySimilarity R package[83] to compute the pairwise semantic similarity (method = resnik) between the enriched GWAS terms, and built a similarity matrix, *S*. From it, we derived a distance matrix, *D*, as $max(S) - S$, and performed multidimensional scaling (MDS). This is an analogous strategy to the one employed in REVIGO[75] to visualize GO terms.

We further compiled genome-wide GWAS summary statistics for eight traits representative of the clusters observed in the MDS representation: asthma[42], breast cancer[43], coronary artery disease[44], heart rate[45], height[46], LDL cholesterol levels[47], rheumatoid arthritis[48] and schizophrenia[49]. In each case, we employed fgwas[50] v0.3.6 (https://github.com/joepickrell/fgwas, default parameters, except for window size set to 2,500 bp to ensure convergence) to obtain the maximum likelihood estimate and 95% confidence interval for the association effect size, both for (i) sQTLs (variants affecting splicing, independently of their effect on expression), and (ii) variants affecting expression, but not splicing (GTEx V7 eQTLs tested also in our setting and not identified as sQTLs). To display the regional GWAS association results for the *GSDMB* locus we employed LocusZoom standalone v1.4 (https://github.com/statgen/locuszoom-standalone).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All the data employed in this study is publicly available. GTEx data was obtained from dbGaP (https://www.ncbi.nlm.nih.gov/gap), accessions phs000424.v7.p2 and phs000424.v8.p2. ENCODE and ENTEx data was obtained from the ENCODE Portal (www.encodeproject.org, accession numbers and URLs provided in Supplementary Data 4-6). The Ensembl Regulation dataset was obtained from ftp://ftp.ensembl.org/pub/grch37/release-86/regulation/homo_sapiens/AnnotatedFeatures.gff.gz. The GWAS Catalog and the Experimental Factor Ontology (EFO) were obtained from https://www.ebi.ac.uk/gwas and https://www.ebi.ac.uk/efo, respectively. A detailed description of the data can be found in Methods and Supplementary Note 3. The sQTL catalog generated is available at https://doi.org/10.5281/zenodo.4058759[84]. Source data are provided with this paper.

Code availability

Our pipeline for sQTL mapping is publicly available at https://github.com/guigolab/sqtlseeker2-nf (https://doi.org/10.5281/zenodo.4065497)[85]. Detailed information about the software can be found in Methods and Supplementary Note 1.

Received: 11 May 2020; Accepted: 2 December 2020;
Published online: 01 February 2021

References

1. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
2. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–55 (2010).
3. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
4. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.* **10**, 741–754 (2009).
5. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* **15**, 689–701 (2014).
6. Kornblihtt, A. R. et al. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14**, 153–165 (2013).
7. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013).
8. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
9. Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat. Commun.* **8**, 14519 (2017).
10. Raj, T. et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* **50**, 1584–1592 (2018).
11. Tian, J. et al. CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res.* **47**, D909–D916 (2019).
12. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
13. Ma, L., Jia, P. & Zhao, Z. Splicing QTL of human adipose-related traits. *Sci. Rep.* **8**, 318 (2018).
14. Caswell, J. L. et al. Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors. *Human Mol. Genet.* **24**, 7421–31 (2015).
15. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
16. Rotival, M., Quach, H. & Quintana-Murci, L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat. Commun.* **10**, 1671 (2019).
17. Ongen, H. & Dermitzakis, E. T. Alternative splicing QTLs in European and African populations. *Am. J. Human Genet.* **97**, 567–75 (2015).
18. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
19. López-Bigas, N., Audit, B., Ouzounis, C., Parra, G. & Guigó, R. Are splicing mutations the most frequent cause of hereditary disease? *FEBS Lett.* **579**, 1900–1903 (2005).
20. Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* **5**, 4698 (2014).
21. Anderson, M. A new method for non-parametric multivariate analysis of variance. *Australian Ecol.* **26**, 32–46 (2001).
22. Anderson, M. J. & Robinson, J. Generalized discriminant analysis based on distances. *Australian N. Zealand J. Stat.* **45**, 301–318 (2003).
23. Di Tommaso, P. et al. Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
24. The GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
25. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
26. Jangi, M. & Sharp, P. Building robust transcriptomes with master splicing factors. *Cell* **159**, 487–498 (2014).
27. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).
28. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).
29. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–60 (2015).
30. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
31. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
32. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
33. Sim, N.-L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
34. Van Nostrand, E. L. et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
35. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
36. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
37. Ghandi, M. et al. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–7 (2016).

38. Yang, E.-W. et al. Allele-specific binding of RNA-binding proteins reveals functional genetic variants in the RNA. *Nat. Commun.* **10**, 1338 (2019).
39. Wickramasinghe, V. O. et al. Regulation of constitutive and alternative mRNA splicing across the human transcriptome by PRPF8 is determined by 5' splice site strength. *Genome Biol.* **16**, 201 (2015).
40. Tilgner, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22**, 1616–1625 (2012).
41. Malone, J. et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).
42. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature Genet.* **50**, 42–53 (2018).
43. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
44. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
45. den Hoed, M. et al. Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat. Genet.* **45**, 621–631 (2013).
46. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
47. The Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
48. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–81 (2014).
49. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
50. Pickrell, J. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Human Genet.* **94**, 559–573 (2014).
51. Saleh, N. M. et al. Genetic association analyses of atopic illness and proinflammatory cytokine genes with type 1 diabetes. *Diabetes* **27**, 838–43 (2011).
52. McGovern, D. P. B. et al. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332–7 (2010).
53. Eyre, S. et al. High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–40 (2012).
54. Panganiban, R. A. et al. A functional splice variant associated with decreased asthma risk abolishes the ability of gasdermin B to induce epithelial cell pyroptosis. *J. Allergy Clin. Immunol.* **142**, 1469–1478 (2018).
55. Das, S. et al. GSDMB induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. *Proc. Natl Acad. Sci. USA* **113**, 13132–13137 (2016).
56. Ding, J. et al. Pore-forming activity and structural autoinhibition of the gasdermin family. *Nature* **535**, 111–116 (2016).
57. Gonzàlez-Porta, M., Calvo, M., Sammeth, M. & Guigó, R. Estimation of alternative splicing variability in human populations. *Genome Res.* **22**, 528–38 (2012).
58. Melé, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–5 (2015).
59. Sibley, C. R. Regulation of gene expression through production of unstable mRNA isoforms. *Biochem. Soc. Trans.* **42**, 1196–1205 (2014).
60. Bhatt, D. M. et al. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
61. Pandya-Jones, A. et al. Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA* **19**, 811–827 (2013).
62. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
63. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet.* **10**, e1004383 (2014).
64. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–17 (2016).
65. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Human Genet.* **99**, 1245–1260 (2016).
66. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
67. Brown, A. A. et al. Predicting causal variants affecting expression by using whole-genome sequencing and RNA-seq from multiple human tissues. *Nat. Genet.* **49**, 1747–1751 (2017).
68. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Human Genet.* **98**, 1114–1129 (2016).
69. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
70. Cheng, J. et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. *Genome Biol.* **20**, 48 (2019).
71. Jagadeesh, K. A. et al. S-CAP extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat. Genet.* **51**, 755–763 (2019).
72. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
73. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
74. Anderson, M. J. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**, 245–253 (2006).
75. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).
76. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–5 (2003).
77. Sammeth, M., Foissac, S. & Guigó, R. A general definition and nomenclature for alternative splicing events. *PLOS Comput. Biol.* **4**, e1000147 (2008).
78. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
79. Davis, C. A. et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
80. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
81. Parra, G., Blanco, E. & Guigó, R. GeneId in Drosophila. *Genome Res.* **10**, 511–515 (2000).
82. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Mol. Biol. Evol.* **30**, 772–80 (2013).
83. Greene, D., Richardson, S. & Turro, E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics* **33**, 1104–1106 (2017).
84. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. sQTL catalog. Zenodo; https://doi.org/10.5281/zenodo.4058759 (2020).
85. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. sQTL mapping pipeline. Zenodo; https://doi.org/10.5281/zenodo.4065497 (2020).
86. Garrido-Martín, D., Palumbo, E., Guigó, R. & Breschi, A. ggsashimi: Sashimi plot revised for browser- and annotation-independent splicing visualization. *PLOS Comput. Biol.* **14**, e1006360 (2018).
87. Hull, J. et al. Identification of common genetic variation that modulates alternative splicing. *PLOS Genet.* **3**, e99 (2007).

Acknowledgements

We thank the donors and their families for their generous gifts of organ donation for transplantation and tissue donations for the GTEx research study. We thank Manuel Muñoz, Emilio Palumbo, Aliaksei Holik, François Aguet and Kristin Ardlie for useful discussions. We thank Romina Garrido for administrative support. This project was supported by the National Human Genome Research Institute of the National Institutes of Health under grants R01MH101814 and 5U24HG009446, as well as by the BIO2015-70777-P grant from the Spanish Ministry of Economy, Industry and Competitiveness (MEIC), and the PGC2018-094017-B-100 grant from the Agencia Estatal de Investigación/European Regional Development Fund (ERDF). It has been also possible in part thanks to the grant CZF2019-002436 from the Chan Zuckerberg Initiative. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (http://commonfund.nih.gov/GTEx). D.G.-M. is supported by a 'la Caixa'-Severo Ochoa pre-doctoral fellowship (LCF/BQ/SO15/52260001). B.B. is supported by the fellowship 2017FI_B 00722 from the Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement (Generalitat de Catalunya) and the European Social Fund (ESF). We thank the ENCODE Consortium and, in particular, Thomas Gingeras', Gene Yeo's and Bradley Bernstein's laboratories for data production. We also acknowledge support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, 'Centro de Excelencia Severo Ochoa', the CERCA Programme/Generalitat de Catalunya and the European Regional Development Fund (ERDF).

Author contributions

D.G.-M. and R.G. conceived and designed the study. D.G.-M. implemented the software and analyzed the data. B.B. contributed to several analyses, provided analysis tools and helped with the interpretation of the results. M.C. and F.R. contributed ideas and statistical advice, helping with the design of the software. D.G.-M. and R.G. wrote the original draft. All the authors reviewed the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20578-2.

Correspondence and requests for materials should be addressed to D.G.-M. or R.G.

Peer review information *Nature Communications* thanks Alvaro Barbeira, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.