Article

Transcriptome variation in human tissues revealed by long-read sequencing

https://doi.org/10.1038/s41586-022-05035-y

Received: 25 February 2021

Accepted: 28 June 2022

Published online: 3 August 2022

Check for updates

Dafni A. Glinos[1,2,15]✉, Garrett Garborcauskas[3,15]✉, Paul Hoffman[1], Nava Ehsan[4], Lihua Jiang[5], Alper Gokden[1], Xiaoguang Dai[6], François Aguet[7], Kathleen L. Brown[1,8], Kiran Garimella[7], Tera Bowers[7], Maura Costello[7], Kristin Ardlie[7], Ruiqi Jian[5], Nathan R. Tucker[9,10], Patrick T. Ellinor[10], Eoghan D. Harrington[6], Hua Tang[5], Michael Snyder[5], Sissel Juul[6], Pejman Mohammadi[4,11], Daniel G. MacArthur[3,12,13], Tuuli Lappalainen[1,2,14,16]✉ & Beryl B. Cummings[3,16]✉

Regulation of transcript structure generates transcript diversity and plays an important role in human disease[1–7]. The advent of long-read sequencing technologies offers the opportunity to study the role of genetic variation in transcript structure[8–16]. In this Article, we present a large human long-read RNA-seq dataset using the Oxford Nanopore Technologies platform from 88 samples from Genotype-Tissue Expression (GTEx) tissues and cell lines, complementing the GTEx resource. We identified just over 70,000 novel transcripts for annotated genes, and validated the protein expression of 10% of novel transcripts. We developed a new computational package, LORALS, to analyse the genetic effects of rare and common variants on the transcriptome by allele-specific analysis of long reads. We characterized allele-specific expression and transcript structure events, providing new insights into the specific transcript alterations caused by common and rare genetic variants and highlighting the resolution gained from long-read data. We were able to perturb the transcript structure upon knockdown of PTBP1, an RNA binding protein that mediates splicing, thereby finding genetic regulatory effects that are modified by the cellular environment. Finally, we used this dataset to enhance variant interpretation and study rare variants leading to aberrant splicing patterns.

Variation in transcript structure by RNA splicing and differences in the 5′ and 3′ untranslated regions (UTRs) is a key feature of gene regulation[1]. Disruption of transcript structure has a major role in human disease, with genetic variants associated with changes in splicing enriched in genome-wide associations for common diseases[2–4] and implicated in many severe Mendelian diseases[5–7]. Common genetic variants affecting transcript structure can be mapped by transcript ratio quantitative trait locus (trQTL) and and splicing quantitative trait locus (sQTL) analyses that have further shown that genetic variants affecting gene expression levels and splicing tend to be distinct[17–19]. An orthogonal method to analyse genetic regulatory effects, allele-specific expression (ASE) analysis, has proven to be a highly sensitive method for studying rare genetic variants in *cis*[20–22]. However, the application of these approaches to short-read data relies on proxies for the full transcript structure and quantification, which are often inaccurate[23–27]. Furthermore, most metrics have focused on alternative splicing, leaving the role of UTRs obscure despite their critical role in disease being demonstrated, with recent progress[28–30]. Long-read RNA-sequencing technologies[8,9] have now reached a mature stage, having already been used to study transcript structures[10,11] and novel transcripts[12–14], as well as for early allele-specific analyses[15,16]. Allele-specific transcript structure (ASTS) analysis, enabled by long-read transcriptome data, could therefore provide important information on how rare and common variants affect transcript structure and disease risk.

Overview of dataset

Altogether, complementary DNA from 90 samples from 56 donors and 4 K562 cell line samples were sequenced on the MinION and GridION Oxford Nanopore Technologies (ONT) platforms. Fibroblast cell lines were used to test the platform and to assess the direct-cDNA versus PCR-cDNA RNA-seq protocols (Extended Data Fig. 1a–c). As the primary purpose of this study was to study allelic events, which require high coverage, we prioritized depth and sequenced the remaining samples

[1]New York Genome Center, New York, NY, USA. [2]Department of Systems Biology, Columbia University, New York, NY, USA. [3]Medical and Population Genetics Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA, USA. [5]Department of Genetics, Stanford University, Stanford, CA, USA. [6]Oxford Nanopore Technology, New York, NY, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Department of Biomedical Informatics, Columbia University, New York, NY, USA. [9]Masonic Medical Research Institute, Utica, NY, USA. [10]Cardiovascular Disease Initiative, The Broad Institute of Harvard and MIT, Cambridge, MA, USA. [11]Scripps Research Translational Institute, La Jolla, CA, USA. [12]Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, New South Wales, Australia. [13]Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. [14]Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden. [15]These authors contributed equally: Dafni A. Glinos, Garrett Garborcauskas. [16]These authors jointly supervised this work: Tuuli Lappalainen, Beryl B. Cummings. ✉e-mail: dafni.glinos@gmail.com; garrett.garbo@gmail.com; tlappalainen@nygenome.org; berylbcummings@gmail.com

using the PCR-cDNA protocol, which does not require high RNA input and is preferred given the precious GTEx tissue samples. To evaluate the messenger RNA isolation protocol, we used the K562 cell lines (Supplementary Methods). The 90 GTEx samples included the following. (1) Assessment of replicability by five samples sequenced in duplicate and five samples in triplicate. Replicability was high (Spearman ρ = 0.86–0.95; Extended Data Fig. 1d), leading us to merge the samples to increase depth. (2) The main dataset for analysis of transcriptome variation across tissues, consisting of 1–5 donors from 14 tissues. (3) Analysis of the effects of transcript perturbation by comparison of five GTEx fibroblast cell lines with and without PTBP1 RNA binding protein (RBP) knockdown. Data were produced across two research centres (Supplementary Methods; Suppl. Table 1). All the GTEx samples had Illumina TruSeq short-read RNA-seq data and 85 samples (51 donors) had whole-genome sequencing (WGS) data made available by the GTEx Consortium[4].

Principal component analysis (PCA) and hierarchical clustering of samples based on transcript expression correlation showed tissue clustering (Fig. 1a,b and Extended Data Fig. 1e), similar to the GTEx consortium analysis of short-read RNA-seq data[4]. Gene and transcript quantifications from long-read data were highly concordant with those from Illumina RNA-seq (median R^2 = 0.75 for genes and R^2 = 0.57 for transcripts; Fig. 1c and Extended Data Fig. 2a). Genes and transcripts with low correlation were enriched for lower expression in ONT data, higher complexity genes and transcripts with multiple exons (Extended Data Fig. 2b–d). We manually checked the read coverage of some of the genes that displayed low correlation, such as *PRELID1*, which is better captured by ONT, and *ARSB*, which displays 3′ bias (Fig. 1d). Overall, longer transcripts displayed higher 3′ bias, as assessed using only the mitochondrial transcripts[14] (Methods; Fig. 1e), and tissue-specific patterns were observed, such as shorter *MT-ND5* in brain tissue and greater variability in cultured cell lines (Extended Data Fig. 3).

Discovery of novel transcripts

We used FLAIR[31] to quantify transcripts and identify novel ones, defined as transcripts with intron chains not matching with any transcript in GENCODE (v.26) (Methods). We found 93,718 transcripts across 21,067 genes (Supplementary Table 2), of which 77% were novel (Extended Data Fig. 4a–c). In most cases we quantified one, often already annotated, transcript for a gene, whereas more novel transcripts were discovered in genes with a high number of annotated transcripts (Fig. 2a). Of the novel transcripts, 47,678 shared at least one splice junction with annotated transcripts and 21,620 had intron retention. In fact, 87% of all intron retention events were novel (Fig. 2b), which suggests the presence of pre-mRNA despite carrying out a poly(A) enrichment step. On the other hand, only 37% of exon skipping events were novel, suggesting they are better represented in the existing annotations (Fig. 2b). We compared our findings with the 33,984 transcripts defined by Workman et al.[14] based on GM12878 cell lines using ONT direct and cDNA RNA-sequencing, which matched 13.1% of our transcripts, 3,604 of which were novel. Similarly, we compared our findings to the CHESS project, which identified 116,156 novel transcripts using short-read RNA-seq from multiple tissues, matching 32.6% of our transcripts, 10,630 of which were novel (Extended Data Fig. 4d and Supplementary Table 3). Despite differences in the tissue samples, sequencing method and parameters used to identify novel transcripts, these provide further evidence to support the identified transcripts.

We validated our novel transcripts by using proteome mass spectrometry data of 32 GTEx samples[32]. For most tissues a similar number of samples using long-read RNA-seq and proteomics were assayed, apart from brain tissue, for which in addition the subregions between the two assays did not match (Suppl. Table 4). We limited this analysis to 33,251 transcripts (63% of which were novel) expressed at ≥5 transcripts per million (TPM) in a sample per tissue and tested for matches in the predicted amino-acid chain. Across tissues, 2,575 novel transcripts were validated and increasing the RNA abundance threshold did not affect this number (Extended Data Fig. 5a–c and Supplementary Table 3). When compared with annotated alternative transcription events, higher validation was observed for novel alternative 5′ UTR and skipped exons, and both annotated and novel intron retention events showed low validation rates that were not different from each other (Fig. 2c and Extended Data Fig. 5d). This depletion could be partially explained by nonsense-mediated decay or other post-transcriptional events depleting the protein products rather than the poor quality of the transcript annotations. Alternative 3′ and 5′ splicing showed higher validation in annotated transcripts, suggesting that these types of events annotated by long reads might be due to technical limitations. For 608 genes we validated more than one transcript (1,304 total), with 823 transcripts being novel, often detecting tissue-specific protein transcript validation (Supplementary Table 5 and Extended Data Fig. 5e).

Novel transcripts resulted in clearer clustering of samples by tissue based on transcript expression correlations and PCA (Extended Data Fig. 6a,b), indicating that novel transcripts capture tissue-specific expression patterns. We therefore examined the gene and transcript expression across nine tissues with at least five samples. Highly expressed (>1 TPM) novel transcripts were tissue specific, with 31.5% expressed in a single tissue (Fig. 2d and Extended Data Fig. 6c). This may explain their absence in existing annotations and highlights the potential for characterizing tissue-specific gene expression and regulation with long-read transcript analysis. We found thousands of transcripts exclusively expressed in a single tissue or having different transcript ratios across all nine tissues (Extended Data Fig. 7). The tissues with the highest ratio of tissue-specific transcripts were the cerebellar hemisphere, liver and fibroblasts (8% of all differentially expressed transcripts), in agreement with previous observations of high transcript diversity[33,34].

Allele-specific analysis

Allele-specific analysis captures *cis*-regulatory genetic effects on expression and transcript structure[17]. The expression of a gene or a transcript is quantified for each haplotype of a sample, separated on the basis of the allele at a heterozygous site. Sixty-four of the long-read RNA-seq samples also had phased whole genome sequencing information from GTEx[4], which allowed us to carry out allelic analysis. To address local alignment biases caused by sequencing errors adjacent to the variant sites of interest, we developed an alignment pipeline in which two haplotype-specific references are created for each donor (Extended Data Fig. 8). To perform ASE and ASTS analysis, in which we test the relative usage of a transcript in relation to the other transcripts of the same gene (Fig. 3a), we developed a new software package, LORALS (long-read allelic analysis). In addition to adopting mappability and genotyping error filters previously developed for short-read data[35], we introduced flags addressing the higher error rate of long-read data (Methods; Extended Data Fig. 9). We performed power calculations using simulated data to test how read counts, number of transcripts and effect size affect ASTS detection power (Methods; Extended Data Fig. 10a).

Having established and optimized our pipeline, we performed the analysis using the FLAIR-aligned transcripts. Per sample, an average of 8.9% of genes analysed for ASE and 7.7% of genes analysed for ASTS had a statistically significant event, with the discovery being proportional to the library size. To maximize power for generalizable insights, we analysed all ASE (3,437 significant out of 36,077 across 6,680 unique genes) and ASTS events (331 significant out of 3,858 across 1,207 unique genes) combined across samples (Extended Data Fig. 10b). For 77% of genes analysed for ASTS we quantified and tested the counts of 2 transcripts per gene, whereas the remaining ranged between 3 and 14 (Extended Data Fig. 10c). Per tissue, 71% of the genes were tested for ASTS in a single donor (Extended Data Fig. 10d). Within the remaining 29%, there

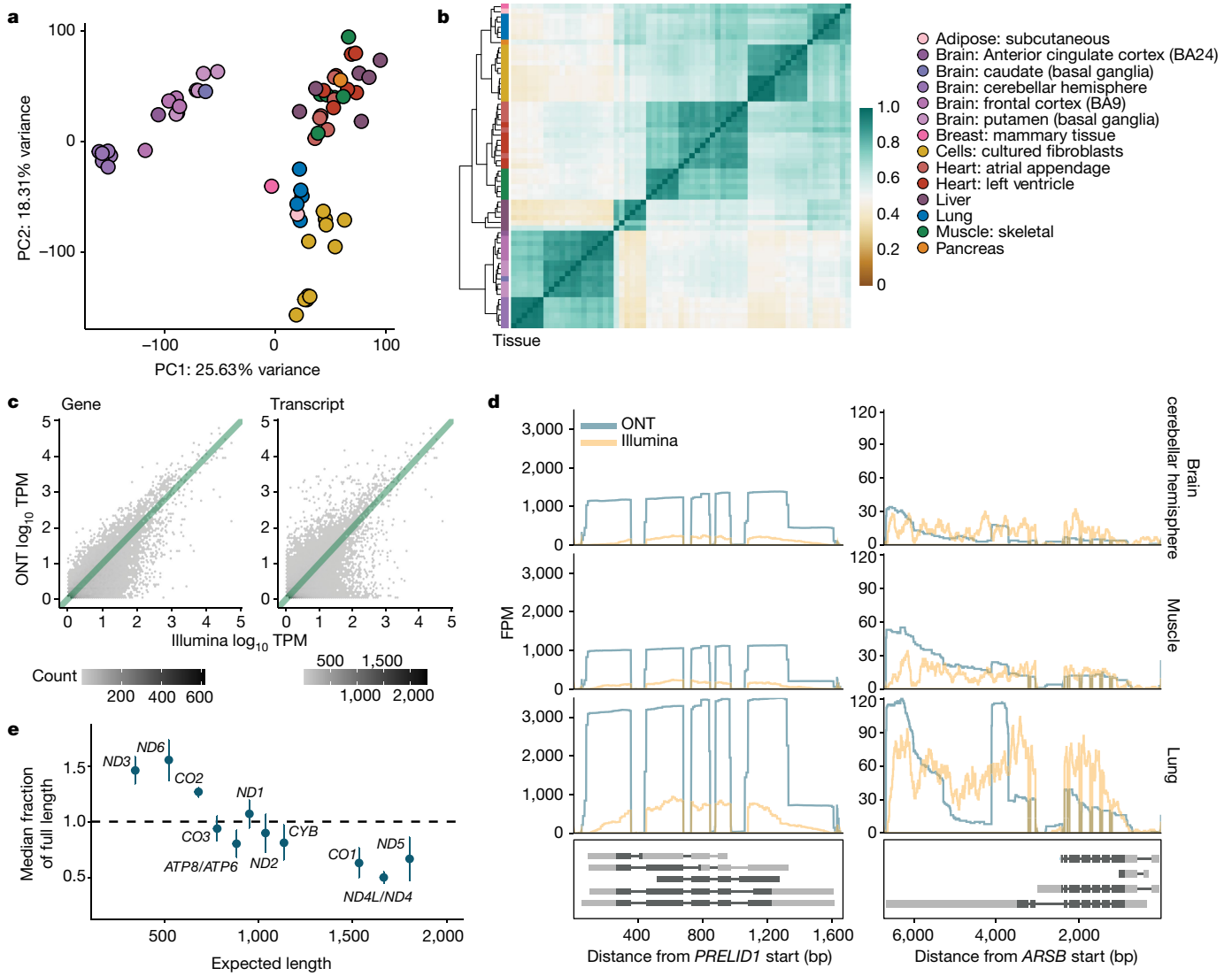


Fig. 1 | Overview and quality control of the dataset. a, PCA of samples with replicates merged, without K562 cell lines and without PTBP1 knockdown samples, based on GENCODE transcript expression (>3 TPM in more than five samples). **b**, Hierarchical clustering of samples based on correlation of transcript expression (as in **a**), using Euclidean distance. **c**, Example of gene and transcript expression correlation between Illumina and ONT in the muscle tissue of GTEX-1LVA9. **d**, Two examples of genes displaying low correlation between ONT and Illumina. *PRELID1* was better captured by ONT than Illumina, whereas *ARSB* had 3′ bias when assayed by ONT. They are shown across three different tissues and all protein-coding transcripts are plotted below. FPM, fragments per million. **e**, Relationship between the expected transcript read length and the fraction of observed nanopore poly(A) RNA reads over the expected full length. Labels are for mitochondrial genes without the MT prefix. The transcript median was calculated per sample, and the median across all samples is plotted ($n = 90$). Error bars represent standard deviation.

were 47 genes that consistently displayed ASTS across donors within a tissue (Supplementary Table 6). Most of these had over two highly expressed transcripts (Binomial test $P = 3.2 \times 10^{-4}$), suggesting that they can withstand variability.

Comparing the long-read ASE events to the ones reported for short-read GTEx v.8 data[35], we observed moderate concordance when looking at the *P* values in short-read data using the long-read significant ASE events ($\pi1 = 0.23$) and vice versa ($\pi1 = 0.41$) (Extended Data Fig. 11a). Of the 341 events that were significant in both datasets, 83% had the same direction of effect, with the opposite direction mostly observed in fibroblast cell lines that were passaged since Illumina sequencing was carried out (Extended Data Fig. 11b–e). Differences were explained by low read depth and some variants being filtered out in one of the datasets (Extended Data Fig. 11f); for example, 445 variants with significant ASE in long-read data were filtered in short-read data owing to the mapping bias flag. Next, we sought to establish that ASE and ASTS recapitulate genetic regulatory effects of expression quantitative trait locus (eQTL) and sQTL mapped by GTEx[4]. Individuals who are heterozygous for a QTL lead variant are expected to show increased allelic imbalance compared with those who are homozygous, and such significant enrichments were observed in the data (Fig. 3b).

Classification of alternative transcript structure (AltTS) changes enables better understanding of the nature of the ASTS events, and thus genetic variants affecting transcript structure. When considering each AltTS event alone, the most common was exon skipping, followed by alternative 3′ splice sites and 3′ UTR events that were enriched for significant ASTS (Extended Data Fig. 12a). To support this, we found that variants located in the 3′ end were more likely to lead to significant ASTS events, compared to 5′ end variants (chi-squared $P = 2.46 \times 10^{-4}$; Extended Data Fig. 12b). We then examined the combination of two types of AltTS events per gene (Fig. 3c). We observed that certain AltTS events co-occurred more commonly in genes with significant ASTS, compared to all events. For example, the combination of mutually exclusive exons with exon skipping (binomial test $P = 2.05 \times 10^{-8}$). On the other hand,

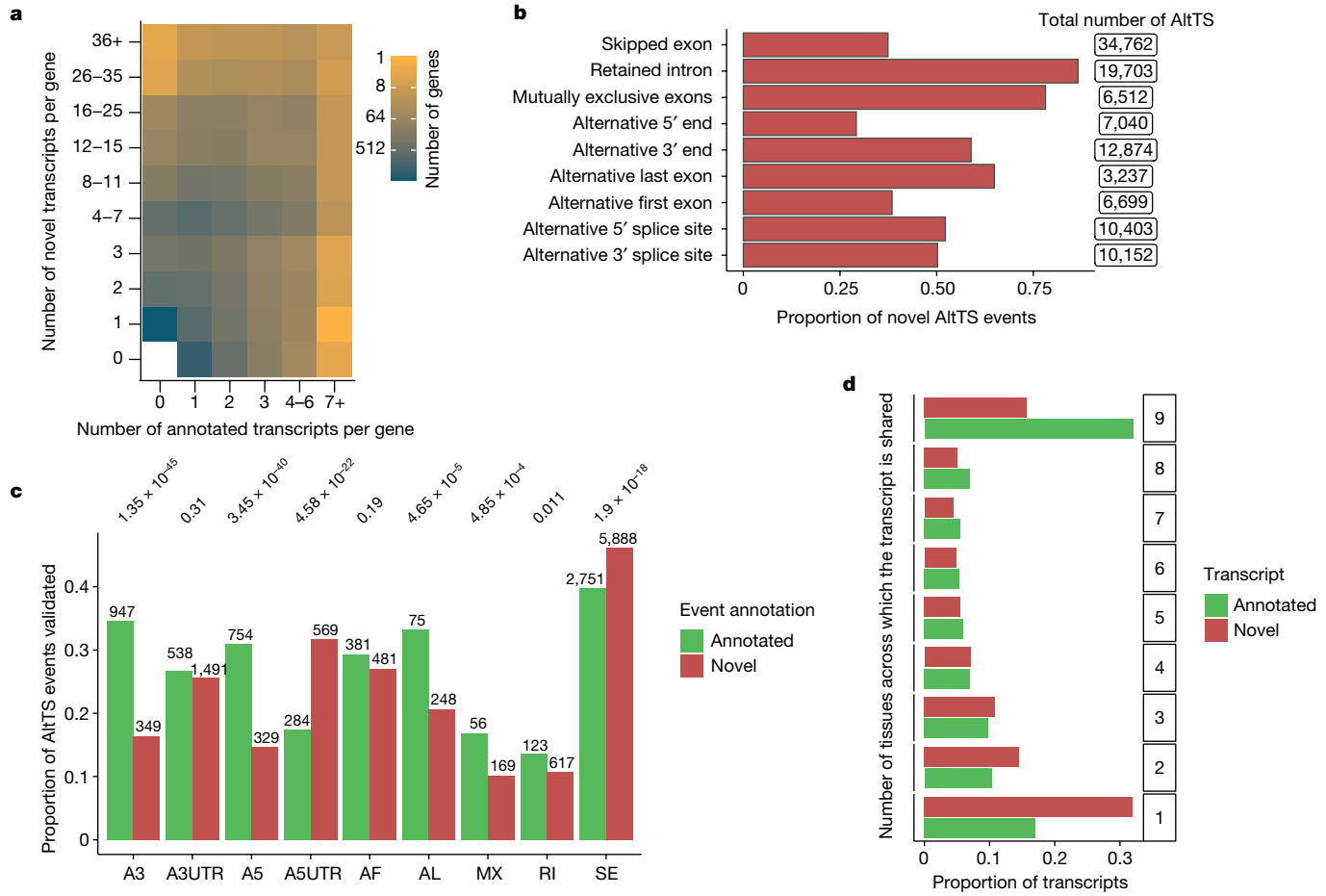


Fig. 2 | Discovery of novel transcripts and comparison between tissues. **a**, Number of annotated and novel transcripts per gene quantified in our dataset. **b**, Proportion of novel AltTS events across all quantified transcripts compared to GENCODE v.26. **c**, Proportion of the AltTS events validated at the protein level by mass spectrometry per novel or annotated. Enrichment was calculated using a two-sided proportionality test. **d**, Number of transcripts expressed at >1 TPM in at least two samples and classified based on how many tissues express the transcript. A3, alternative 3′ splice site; A5, alternative 5′ splice site; AF, alternative first exon; AL, alternative last exon; A3UTR, alternative 3′ end; A5UTR, alternative 5′ end; MX, mutually exclusive exons; RI, retained intron; SE, skipped exon.

there were combinations that were depleted from significant ASTS events, notably the combination of alternative 3′ UTR with any other event. This highlights the distinct effect of alternative UTR regions within the significant ASTS genes, missed in most sQTL mapping approaches.

To better understand the relationship between genetic effects on expression and transcript structure, we compared the ASE and ASTS events. We found that 222 of the 880 significant ASE genes displayed significant nominal *P* values in ASTS ($\pi1 = 0.15$). This proportion was larger when looking at significant ASTS, for which we found that 176 of the 330 genes displayed significant nominal *P* values in ASE ($\pi1 = 0.46$; Fig. 3d). This indicates that changes in transcript structure are often accompanied by changes in transcript levels, but less often the other way around. When repeating this analysis stratified by AltTS events, we observed that an exception to this were ASTS events caused by alternative 3′ ends, for which an equal proportion of events were ASE and ASTS (Fig. 3d).

On the basis of these observations, we examined sQTL-significant genes in ASE, for which we observed a difference between heterozygous and homozygous individuals (Fisher's exact test $P = 1.81 \times 10^{-5}$). When looking at eQTLs, we also observed that more heterozygous donors had significant ASTS compared to homozygous (Fisher's exact test $P = 1.56 \times 10^{-4}$; Fig. 3b), indicating that genetically induced expression differences manifest in ASTS. To test the origin of this, we stratified the events by AltTS events. We observed that the sQTLs were mostly manifesting in differences in exon skipping (34.2%; Fig. 3e), as expected, whereas eQTLs were manifesting not only in total expression differences but also in transcript structure changes of the 5′ end of a gene (33.3%; Fig. 3e). Differences in the 5′ end of a gene are therefore driving the capture of eQTLs in ASTS data, which would be normally missed by sQTL mapping.

This breakdown of events allows us to revisit existing sQTLs and find examples in which ASTS data enable a better understanding of the exact molecular events associated with the genetic variant, potentially contributing to diseases and traits (Methods; Supplementary Table 7). *DUSP13*, for example, is a gene specifically expressed in muscle, and has three sQTL intron excision phenotypes colocalizing with a single locus associated with body fat percentage[4]. Multiple transcripts arise from this gene, but in both donors displaying ASTS we observed that the transcript ENST00000372700 lacking four middle exons was more highly expressed from the risk allele (Extended Data Fig. 12c). As further validation, GTEx short-read transcript ratios recapitulated this pattern (Extended Data Fig. 12d). We were therefore able to pinpoint to the exact events leading to differences in transcript expression from the two alleles and potentially predisposing to high body fat percentage.

To test how ASTS captures changes in the effects of *cis*-regulatory variants due to perturbation of the splicing machinery of the cell, we knocked down PTBP1 RBP in five GTEx fibroblast cell lines. PTBP1 mediates exon skipping in pre-mRNAs and is involved in the 3′-end processing of mRNA. We therefore expected to see a disturbance of transcript expression as well as ASTS patterns for some genes upon small interfering RNA (siRNA)

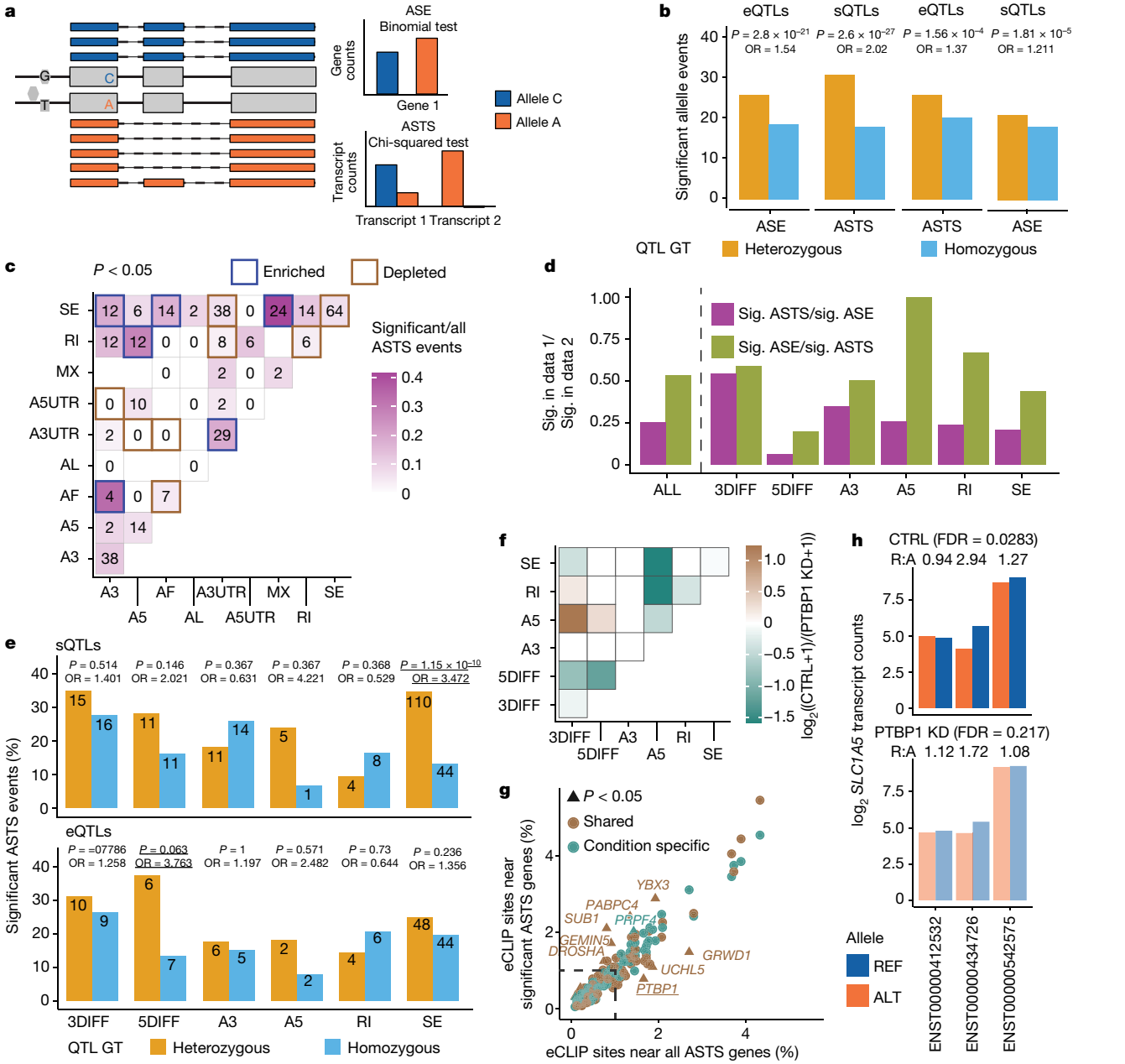


Fig. 3 | Allelic analysis of long-read data. a, Illustration of allele-specific analysis framework, data input and testing performed. **b**, Percentage of significant allele-specific expression and transcript structure events for samples that are heterozygous or homozygous for a lead eQTL or sQTL variant for that gene. *P* values from two-sided Fisher's exact test. GT, genotype; OR, odds ratio. **c**, Co-occurrence of alternative transcript structure events within the transcripts used for ASTS analysis that are observed at least once per each event (or a single time for the diagonal) in a given gene. *P* values from two-sided binomial test. A3, alternative 3′ splice site; A5, alternative 5′ splice site; AF, alternative first exon; AL, alternative last exon; A3UTR, alternative 3′ end; A5UTR, alternative 5′ end; MX, mutually exclusive exons; RI, retained intron; SE, skipped exon. **d**, Sharing of ASE and ASTS events for all events, and stratified by AltTS event. ALL, all events; Sig., significant. **e**, Percentage of significant ASTS for samples that are heterozygous or homozygous for a lead eQTL or sQTL variant for that gene, respectively, by type of event based on whether at least 50% of the differences in transcript can be assigned to that AltTS event. *P* values from two-sided Fisher's exact test. Underlined are the events with the lowest *P* values for eQTLs and sQTLs. **f**, Changes in ASTS by PTPB1 knockdown, with the heatmap showing the co-occurrence of alternative transcript structure events that are observed at least once per each event (or a single time for the diagonal) in a given gene. Colour corresponds to the $\log_2$ ratio of the number of events found in the control (CTRL) over PTBP1 knockdown (KD) samples. **g**, Percentage of eCLIP sites near genes tested for ASTS, annotated using a 10 kb window. Genes stratified into shared or condition specific based on the overlap between control and PTBP1 knockdown. Marked are sets of peaks with $P < 0.05$ using a two-sided binomial test. **h**, Example of a gene, *SLC1A5*, where transcript read counts display significant ASTS only in the PTBP1 knockdown sample. REF, reference allele; ALT, alternative allele; R:A, ratio of reads containing reference over alternative allele.

knockdown. Indeed, we found 3,061 differentially expressed genes, 70% of which were validated with short-read data, and 4,220 differentially expressed transcripts (Extended Data Fig. 13a,b). Exon exclusion and longer alternative 3′ UTR events were enriched in transcripts significantly upregulated in PTBP1 knockdown samples (Extended Data Fig. 13c).

We then compared allelic events in the knockdown and control samples (Methods and Extended Data Fig. 14a). We observed different transcript processing events between the two conditions, indicating that heterozygous genetic variants driving the ASTS in control samples lose their effect in the absence of PTBP1 (Extended Data Fig. 14b).

To increase our power, we re-sequenced the same samples on the PromethION platform, resulting in a minimum of 22 million reads per sample. We re-identified allelic imbalance for 87% of the ASE events and 58% of the ASTS events (Extended Data Fig. 14c). We observed an enrichment of condition-specific events in ASTS compared with ASE (Fisher's exact test $P = 2.89 \times 10^{-10}$; Extended Data Fig. 14d), consistent with the fact that PTBP1 affects splicing and not gene expression at the allelic level. The control samples were enriched for ASTS with 3′ end differences combined with alternative 5′ splice sites, whereas alternative 5′ splice sites combined with exon skipping or intron retention were enriched in knockdown-specific ASTS (Fig. 3f).

We hypothesized that condition-specific ASTS events upon RBP knockdown might reflect different regulation modes to those that are shared. We expect those to be driven by heterozygous variants within RBP sites detectable in eCLIP peaks[36] (Supplementary Table 8). We focused on the genes with at least one heterozygous variant falling in an eCLIP site (82% of ASTS genes), and tested whether specific RBPs were differentially enriched near significant ASTS genes that were specific to a condition or shared. PTBP1 sites harbouring heterozygous variants were depleted from ASTS events shared between the two conditions ($P = 0.0087$; Fig. 3g), in agreement with the expectation that these events are driven by PTBP1 independent processes. We discovered 35 condition-specific ASTS events with PTBP1 eCLIP peaks, equally distributed between the control and the knockdown. For example, in *SLC1A5*, a donor has a heterozygous site within a PTB1 eCLIP site and ASTS that is attenuated upon PTPB1 knockdown (Fig. 3h and Extended Data Fig. 14e). These analyses are consistent with a model in which changes in the cellular environment altering splicing regulation can affect the molecular function of genetic variants.

Rare variant interpretation

Finally, we evaluated the potential to better interpret rare variants with novel transcript annotations and ASTS data from long reads. We complemented the GENCODE v.26 annotation with an additional 73,599 transcripts, and re-annotated genetic variants from GTEx WGS data using VEP[37] (Methods). The most severe consequence for a variant changed for 0.75% of all variants (Extended Data Fig. 15a), 16,435 of which were coding (3.27% of coding variants). We used combined annotation-dependent depletion (CADD) scores as a proxy for the pathogenicity of a variant and as further support for validity of the re-classifications. We observed that variants reassigned to a more severe consequence had on average a higher CADD score than those that retained the same annotation (Fig. 4a and Supplementary Table 9). An exception to this were variants previously annotated as non-coding transcript exons and reassigned as coding but assigned a lower CADD score, suggesting that some of the novel transcripts we identify might not be coding. The higher CADD scores for variants reassigned as pathogenic provides independent evidence that our novel transcripts detect real biology and functional variants that may have been missed before. We therefore re-annotated ClinVar variants, resulting in the reassignment of 9,582 variants (1.23%). We observed that variants with uncertain benign or pathogenic clinical significance and no assertion criteria were reassigned at the highest rate (4% and 3.1%), whereas pathogenic variants with higher reviewer support were reassigned at the lowest rates (Extended Data Fig. 15b). This provides an explanation for the conflicting reports of these variants and a potential pathogenic mechanism.

Long-read allelic data provide the opportunity to observe rare variants disrupting transcriptional regulation. GTEx has previously defined individuals that are extreme ASE, expression and splicing outliers, and shown that they are enriched for having rare genetic variants in the gene's vicinity[22,38]. Although our sample size is insufficient for analogous analysis of ASTS outliers, we tested the presence of rare (minor allele frequency (MAF) < 0.01) heterozygous variants within a 10 kb

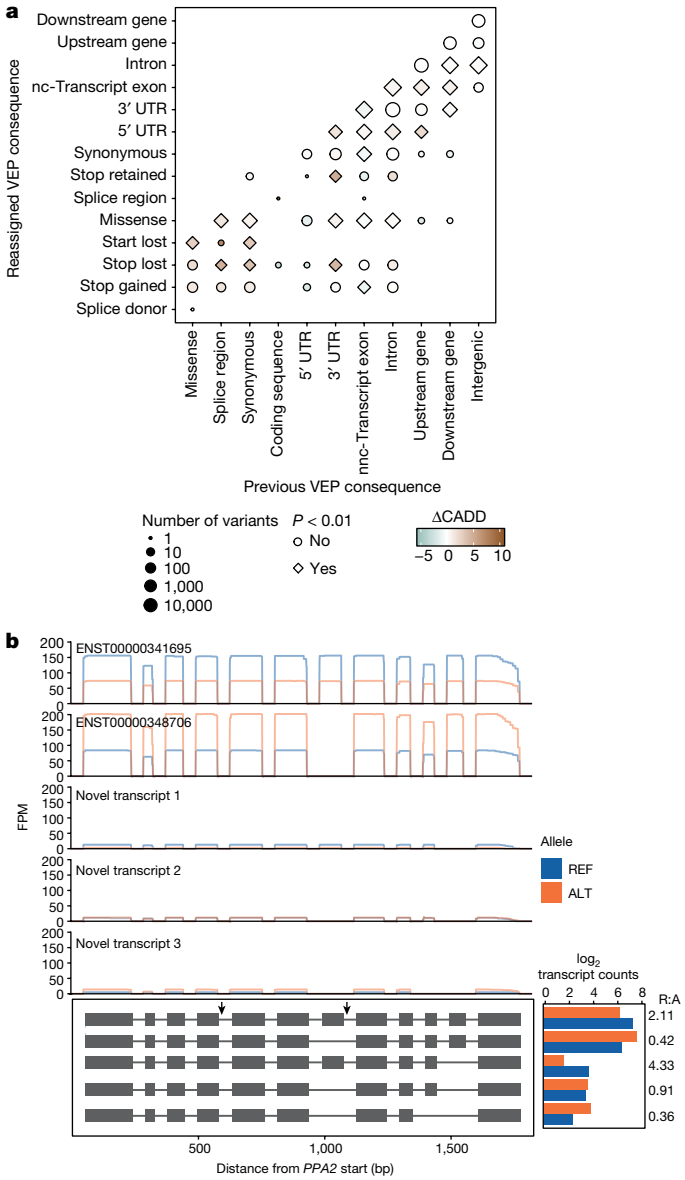


Fig. 4 | Variant interpretation through novel transcripts and ASTS analysis. **a**, Difference in the mean CADD score of variants that were reassigned to a more severe consequence when the GENCODE gene annotations were complemented with the novel FLAIR transcripts, compared to variants that retained their annotation (down sampled to a similar size). *P* values from two-sided *t*-test. VEP, variant effect predictor. **b**, *PPA2* is an example of a gene with a rare heterozygous variant in a sample that is a GTEx splicing outlier and has significant ASTS, with read pile-ups, and grey arrows indicating the rare variants.

window of each ASTS gene. Across all samples, missense variants were enriched for being in significant ASTS genes compared to all genes measured for ASTS (Extended Data Fig. 15c,d). This indicates that ASTS can capture rare variant effects on transcript structure. In addition, we observed that significant ASTS genes were enriched within splicing outliers (Extended Data Fig. 15e). Finally, we searched for specific examples in which a rare variant is probably causing ASTS in our data (Supplementary Table 10). Out of eleven genes for which an individual has a rare heterozygous variant, is a splicing outlier as defined by GTEx and has significant ASTS, we highlight two examples: *PPA2* has two intron variants chr4:105409456:G:A and chr4:105449015:G:A (MAF = 5.97×10^{-4} and 9.55×10^{-3}, respectively), with the alternative allele having higher expression levels of transcript ENST00000348706 and lower expression of ENST00000341695 (Fig. 4b) and *NDUFS4* (Extended Data Fig. 15e,f).

Discussion

In this study, we present a large dataset of long-read RNA-seq, using material derived from cell lines and human tissues collected by the GTEx project. We identified 71,735 novel transcripts, which is high compared to other long-read studies[12–14] probably because of our large sample size and tissue diversity, consistent with the high number of tissue-specific novel transcripts discovered. Supported by a high validation rate of the novel transcripts in high-throughput mass spectrometry proteome data[32], our data make an important contribution to human transcript annotations. Expanding long-read studies to further tissues and cell types, coupled with more extensive validation efforts, will enable a better understanding of the regulatory mechanisms of the different types of transcript changes[12], the functionally distinct protein isoforms that different transcripts can give rise to[39] and the improved variant annotation, as demonstrated by our analysis.

Long reads provide the ability to map allelic effects over transcripts, instead of just expression[40], thus providing the opportunity to analyse *cis* effects of genetic variants on transcripts. We developed LORALS, a toolkit for allelic analysis specific to long reads, considering various biases inherent to the technology. It is tuneable and applicable to any long-read data, improving on previous work in this field[14,15]. We observed that the majority of ASTS events coincided with ASE, indicating that genetic effects on transcript usage rarely happen by reciprocally flipped transcript expression, but are typically accompanied by a change in the total expression levels, which could happen, for example, by altered stability of specific transcripts[41]. However, the widespread co-occurrence of ASTS with ASE and eQTLs manifesting as ASTS are seemingly at odds with multiple QTL mapping studies that have established that expression and splicing are affected by distinct regulatory variants and processes[3,4,17]. The ability to distinguish the exact alternative transcript structure events in ASTS data allowed us to discover allele-specific 5′ differences as the cause of eQTLs manifesting in transcript structure changes, whereas expression and splicing are indeed highly independent. Given that promoter differences greatly affect gene expression levels and that most sQTL mapping methods do not capture the variation in UTRs, this explains both the low overlap between causal variants of sQTLs and eQTLs and the overlap of ASTS with ASE and eQTLs.

These results reinforce the emerging understanding[29] of the importance of analysing the transcriptome not at the level of genes or imprecisely defined splicing, but rather with a detailed characterization of specific transcripts, their changes and combinations. These insights are readily captured by long reads. Given the role of genetic variants affecting transcript structure in disease risk[2–4,42–44], we anticipate that a high-resolution characterization of the transcriptome with long-read data will be an important approach for the discovery of regulatory mechanisms of disease-associated variants.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-022-05035-y.

1. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.* **102**, 11–26 (2018).
2. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
3. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
4. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
5. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
6. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
7. Gonorazky, H. D. et al. Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease. *Am. J. Hum. Genet.* **104**, 466–483 (2019).
8. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
9. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
10. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D. & Ragoussis, J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Sci Rep.* **6**, 31602 (2016).
11. Weirather, J. L. et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6**, 100 (2017).
12. Anvar, S. Y. et al. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**, 46 (2018).
13. Tardaguila, M. et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.* **28**, 396–411 (2018).
14. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
15. Tilgner, H., Grubert, F., Sharon, D. & Snyder, M. P. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. USA* **111**, 9869–9874 (2014).
16. Tilgner, H. et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**, 736–742 (2015).
17. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506 (2013).
18. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).
19. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
20. Rivas, M. A. et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
21. Smith, D. et al. A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. *PLoS Genet.* **13**, e1006659 (2017).
22. Mohammadi, P. et al. Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **366**, 351–356 (2019).
23. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
24. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
25. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Erratum: near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 888 (2016).
26. Teng, M. et al. A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).
27. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
28. Pai, A. A. et al.Widespread shortening of 3′untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection *PLoS Genet.* **12**, e1006338 (2016).
29. Alasoo, K. et al. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *eLife* **8**, e41673 (2019).
30. Mittleman, B. E. et al. Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**, e57492 (2020).
31. Tang, A. D. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.* **11**, 1438 (2020).
32. Jiang, L. et al. A quantitative proteome map of the human body. *Cell* **183**, 269–283.e19 (2020).
33. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5**, R74 (2004).
34. 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).
35. Castel, S. E. et al. A vast resource of allelic expression data spanning human tissues. *Genome Biol.* **21**, 234 (2020).
36. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
37. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
38. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, eaaz5900 (2020).
39. Yang, X. et al. Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* **164**, 805–817 (2016).
40. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
41. Sibley, C. R. et al. Recursive splicing in long vertebrate genes. *Nature* **521**, 371–375 (2015).
42. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
43. Gandal, M. J. et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018).
44. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

Methods

Fibroblasts cell culture and PTBP1 siRNA transfection

Fibroblast cell lines derived from skin samples from the lower leg and biobanked as part of the GTEx Consortium were cultured in DMEM media supplemented with 10% FBS and 1% penicillin/streptomycin (Corning). Transfections were performed 24 h after initial seeding of 500,000 cells in 10 cm dishes. Transfection mixtures were prepared with 6 μg per dish of siRNA pools (Dharmacon SO-2720501G, SO-2703775G), Lipofectamine 2000 (Thermo Fisher) and Opti-MEM reduced serum media (Corning) according to proprietary guidelines. Mixtures were added to cell cultures containing reduced volumes of 5 ml DMEM media for 6 h before increasing volumes to 10 ml with fresh media. Cells were harvested 96 h after transfection.

SDS–PAGE and western blotting

Protein was extracted by boiling 75,000 cells at 95 °C for 5 min in 100 μl 2× Laemmli Sample Buffer (Bio-Rad) and 2-mercaptoethanol (5%) as a reducing agent. SDS–PAGE was run on 10% Mini-PROTEAN TGX gels (Bio-Rad) in Tris/glycine/10%SDS buffer. Proteins were transferred onto nitrocellulose membranes. Then 5% non-fat milk was used for blocking. Primary antibodies from mouse for PTBP1 (used in 1:4,000 dilution; Thermo Fisher Scientific) and rabbit for GAPDH (used in 1:10,000 dilution; Cell Signaling Technology) were incubated overnight at 4 °C. Secondary antibodies (LI-COR IRDye; donkey anti-mouse IgG polyclonal antibody (800CW; size 100 μg) and donkey anti-rabbit IgG polyclonal antibody (680RD; size 100 μg) were incubated for 1 h at room temperature. Membranes were imaged on the Li-cor Odyssey CLx system.

Generation of long-read RNA-seq data

Generally following the manufacturer's instructions, the protocol detailed in the Supplementary Methods was used.

Sequencing and base-calling

Libraries were prepared with 300 ng of input total RNA using the Illumina TruSeq kit and sequenced on the NextSeq 550 platform. Sequencing of mRNA samples was performed on the GridION X5 and MinION platform (Oxford Nanopore Technologies) for 48 h. To basecall the raw data we used ONT's Guppy tool (v.3.2.4).

Genome and transcriptome alignments

We used minimap2 v.2.11 (ref. [45]) to align the reads to the GRCh38 human genome reference using -ax splice -uf -k14 –secondary=no parameters. We also aligned to the GENCODE v.26 transcriptome using -ax map-ont parameters. We used NanoPlot[46] to calculate alignment statistics. We obtained a median of 6,343,016 raw reads per sample, of which on average 80% (s.d. 16%) aligned to the genome (Extended Data Fig. 1a). The median read length was 709 base pairs (bp) and 789 bp for raw and aligned reads, respectively (Extended Data Fig. 1b). We observed a higher median read length in samples sequenced using the direct-cDNA ONT protocol when compared to the PCR-cDNA protocol (*t*-test $P = 0.022$), at the expense of lower read depth (*t*-test $P = 6.45 \times 10^{-3}$) (Extended Data Fig. 1c).

We used the method outlined in Workman et al.[28] to calculate 3′ bias in our data, which only focuses on reads assigned to transcripts encoded in the mitochondrial genome. The reasoning for using mitochondrial transcripts was that they are abundant across all tissues, are single exon and have variable lengths. We limited our analysis to reads that aligned within a 50 nucleotide window of the 3′ end of the gene. We calculated the median proportion of full-length reads per sample, and across all samples, along with standard deviations.

All read pile-up plots were made using wiggleplotr[47].

Transcript detection and characterization

We defined transcripts using FLAIR v.1.4 (ref. [31]). Four heart left ventricle samples from patients with cardiovascular disease were included for the novel transcript calling (phs001539.v1.p1). We used the samples that had been aligned to the genome and applied FLAIR-correct to correct misaligned splice sites using GENCODE v.26 annotations. We merged all samples and ran FLAIR-collapse per chromosome to generate a first-pass transcript set by grouping reads on their splice junction chains and only keeping transcripts supported by at least ten reads. We only kept reads with transcription start sites that fell within promoter regions defined by taking a window 10 bp upstream and 50 bp downstream of the gene start site based on the GENCODE v.26 build and that spanned at least 80% of the transcript with at least 25 nucleotide coverage into the first and last exon. Reads that passed these filters were then re-aligned to the first-pass transcript set, retaining alignments with mapping quality score (MAPQ) > 10.

We further filtered our transcript discovery set using TransDecoder software (https://github.com/TransDecoder/TransDecoder/) to remove transcripts with no open reading frames (ORFs). We integrated Pfam and Blast databases in this search, using the default parameters, to select the ORFs with the most functional coding potential. We removed transcripts for which all ORF were marked as being partial 3′ and 5′. We further limited our discovery to transcripts encoding at least 100 amino acid long transcripts. This step decreased the number of novel transcripts from 159,882 to 93,718.

Transcripts were compared to GENCODE v.26, Workman et al. flair-called transcripts[28] and CHESS transcripts[48] using gffcompare[49]. Transcripts with exact intron chain-match were marked as annotated, whereas all others were marked as novel.

Transcript quantification

We used flair quantify[31] to quantify transcripts from all samples for which reads had been aligned with (1) GENCODE v.26 and (2) the newly identified transcripts. Reads were normalized using TPM normalization and were filtered for transcripts expressed at least five TPM in at least three samples before clustering analysis. Similarly, for the comparison between ONT and Illumina, reads were normalized using TPM normalization, filtered for protein-coding genes and limited to those with expression higher than one TPM in both Illumina and ONT. Lowly correlated genes were defined by residual analysis of the Spearman correlations (Extended Data Fig. 3).

Alternative transcript structure events definition

We used SUPPA (v.2.3)[50] to define alternative 3′ splicing (A3), 5′ splicing (A5), first exon (AF), last exon (AL), intron retention (RI), exon skipping (SE) and mutually exclusive exons (MX). We supplemented these annotations with alternative UTR regions, which for the purposes of this study were assumed to be the last exons. We used a window size of ten nucleotides around splice sites to allow for error.

Protein validation of highly expressed transcripts

For the tissues assayed in the GTEx proteomics database[32] (heart, brain, liver, lung, muscle, pancreas and breast), we identified the transcripts expressed at higher than 5 TPM per sample. We used the output peptide fasta file from TransDecoder analysis to get the amino-acid sequence for each of the maintained transcripts. In total, 33,251 transcripts were maintained. To optimize our search space, we grouped together brain samples from different regions and heart samples from different regions.

Raw files from the GTEx proteomics study[32] were first converted to mzXML files and submitted to the Trans-Proteomic Pipeline (http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP) for database search. The Comet search engine was used for the database[51]. The mass tolerance of precursor ions was set to 10 ppm and fragment ions was set to 1.0 atomic mass unit (amu). Up to two missed cleavages were allowed for trypsin digestion. Methionine oxidation was set to variable modification. Cysteine carbamidomethylation and peptide N-terminal and lysine tandem mass tag modifications were set to be

static modifications. After searches, peptides were filtered and scored by the PeptideProphet algorithm and proteins were scored afterwards using ProteinProphet[52]. Protein probability greater than 0.99 and group_sibling_id of 'a', which marks the protein containing the largest number of total peptides, were used for the confident identification of the transcript.

Differential transcript expression and transcript usage

We used the nine tissues with at least five samples (brain cerebellar hemisphere, frontal cortex and putamen, cultured fibroblasts, atrial appendage and left ventricle from the heart, liver, lung and muscle). Differential expression was performed with DEseq2 (ref. [53]) pairwise using the Wald method and across all samples using the likelihood ratio test. We used replicates using the function collapseReplicates. We used a cut-off for statistical significance at false detection rate (FDR) = 0.05. Differential transcript usage was performed with DRIMSeq[54]. Only the replicate with the highest read coverage was maintained in the analysis. All analysis was done in a pairwise manner, with a cut-off for statistical significance at FDR = 0.05.

Differential gene and transcript expression analysis between the control and PTBP1 knockdown samples were performed in the same way as above. For differential gene expression we used quantifications made based on the GENCODE gene annotation, as each gene's differential gene expression status was validated using the Illumina RNA-seq protocol on the same samples. For transcript differential expression we used the FLAIR transcripts.

Allele-specific analysis

Alignment strategy. We used the bcftools package to filter for only heterozygous variants per donor. We complemented the WGS and short-read RNA-seq phasing by long-read RNA-seq read phasing with HAPCUT2 (ref. [55]), run using all available RNA-seq libraries per subject. The haplotype phasing had been informed by the short-read RNA-seq data and we further switched the phase of a median of 0.05% of the heterozygous variants using the long-read data.

We generated a reference genome per haplotype of each donor and re-aligned the reads to each of the two references using the same parameters as described above. For each read, we retrieved the two MAPQ scores, and if different kept the one with the highest score; the ties were randomly chosen between the two references. This approach led to a difference in alignment for on average 4.8% (Extended Data Figs. 8c,9c) of the reads containing a heterozygous variant. We examined the first position of each aligned read to better understand the source of the high reference bias observed. Most reads (98.4%) aligned to the exact same location, which suggests that the reference bias was mediated by local misalignment within the read, probably stemming from insertions/deletions adjacent to the variant of interest (Extended Data Figs. 8d,9d). A small proportion of reads (1.2%) did not align when using the personalized reference genomes. Therefore, our mapping approach allows most long reads to reliably be assigned to a haplotype.

Data acquisition. Single-nucleotide polymorphism level allele-specific data was generated using software developed specifically for long-read data (LORALS). We flagged multi-mappability sites, so sites that were part of the blacklist regions from ENCODE and monoallelic sites as determined by GTEx. Regions with multi-mapping reads were constructed using the alignability track from UCSC Genome Browser using a threshold of 0.1, meaning that a 100*k*-mer aligning to that site aligns to at least five other locations in the genome with up to two mismatches. Monoallelic sites were defined across all their tissue for each sample, by testing whether there are no more reads supporting two alleles than would be expected from sequencing noise alone, indicating potential genotyping errors (FDR < 1%).

We introduced two ONT-specific flags, namely, the ratio of reference and alternative allele containing reads to the total read number for a site, which we set to greater than 80%, and the number of reads containing indels within a 10 bp window of the heterozygous variant. This filter was determined by counting the number of base pairs that were matched within the window and requiring at least eight of them to not be INDELs. If, at the site, the proportion of indel containing reads was greater than 80%, then it was flagged. In addition, the reads that contained over eight INDELs within the window were filtered out. Finally, only variants that were covered by at least 20 reads were kept.

After filtering the flagged sites, we maintained a median of 77% (s.d. 4%) of the sites per sample, with the most stringent filter being the ratio of indel containing reads, which removed 22% of the sites (s.d. 4%). For the variant sites that passed these filters we checked to which transcript each read was assigned. We then created haplotype tables per gene across all of its transcripts. These tables were filtered for genes that had at least two transcripts, for which each transcript had at least 10 reads and the total expression of a gene was greater than 36 reads. In the case when multiple variants associated with a gene, the one with the highest total coverage was selected for the analysis.

We compared the reference ratios per gene and transcript across the samples for which we had either data from more than one tissue or which were sequenced in duplicates or triplicates. We observed a higher Spearman correlation for samples from the same tissue (median $R^2 = 0.72$ for ASE and $R^2 = 0.96$ for ASTS) compared to samples from different tissues (median $R^2 = 0.65$ for ASE and $R^2 = 0.83$ for ASTS). We therefore merged the duplicate samples to increase our read depth.

Statistical analysis, simulations and power analysis. Allele-specific analysis was based on the framework outlined in refs. [40,56]. For a given gene and biallelic variant, we define allelic expressions e_0, and e_1 as the sum of all transcripts produced from a gene located on the same chromosome copy as each allele. We define log aFC as the expression originating from the alternative allele versus the reference allele (equation (1)) and the reference ratio as the proportion of the reads originating from the reference allele over the total number of reads (equation (2)):

$$\log \text{aFC} = \log 2 \frac{e_1}{e_0} \tag{1}$$

$$\text{Reference ratio} = \frac{e_0}{e_0 + e_1} \tag{2}$$

To test for statistically significant allele-specific analysis, a binomial test was used to determine whether it was significantly different from the expected value of 0.5. Binomial test *P* values were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure (FDR < 5%).

When testing for allele-specific transcript structure, we performed power analysis to estimate the fraction of the cases in which the distribution of transcript expression produced from the gene on the haplotypes were significantly different. Let $e^{t_i}_{h_j}$ be the allele-specific dosage for the transcript (t_i) from haplotype h_j. We denote $p^{t_i}_{h_j}$ as the allelic expression fraction of the transcript t_i, where $\sum_i^t = 1 \times p^{t_i}_{h_j} = 1$. The dependence of the two distributions $e^{(t)}_{h_1}$ and $e^{(t)}_{h_2}$ is determined by the chi-squared test (x^2). *P* values are corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure (FDR < 5%).

The read counts, number of transcripts for each gene and the log aFC (ref. [16]) are the factors that affect the power of the statistical test. Regarding the aFC factor, the maximum power happens at log aFC equal to zero, indicating equal expression in both haplotypes. Thus, for our analysis we assume that log aFC is zero and statistical power is estimated to determine the dependency of ASTS analysis on the total coverage and transcript counts to detect the effect of a given size. The effect size is given by Cohen's w, defined as[57]

$$w = \sqrt{\frac{x^2}{N}} \quad (3)$$

This is applied on a $2 \times m$ (where m is the number of transcripts) contingency table from $p_{h_1}^{(t)}$ and $p_{h_2}^{(t)}$, where N is the total count table, which in this case is 2. To give an idea of how the change in transcript ratios affects the magnitude of the effect size, the $p_{h_1}^{(t)}$ and $p_{h_2}^{(t)}$ pairs are presented in Supplemetnary Table 12 for $w = 0.3$ (interpreted as medium effect size).

Data simulations. To perform the power estimation, the simulated allelic expressions $e_{h_1}^{(t)}$ and $e_{h_2}^{(t)}$ were produced from a multinomial distribution of two normalized random vectors $p_{h_1}^{(t)}$ and $p_{h2}^{(t)}$ that specify the effect size of interest. The significant difference of $e_{h1}^{(t)}$ and $e_{h2}^{(t)}$ was determined by the chi-squared test (nominal $P < 0.01$). Power estimation based on simulated transcript count data for a set of read counts and the number of transcripts for the effect sizes of 0.1 (small effect), 0.3 (medium effect) and 0.5 (large effect) was calculated. The effect size is rounded for one digit. In order to detect ASTS with an effect size of 0.5 with 60% power, assuming aFC = 0, the total read coverage of 36 was required. For an effect size of 0.3, at least 100 reads were needed. For the detection of smaller effect sizes, we were underpowered, for even up to 500 reads. These simulations informed our power to detect events of different effect sizes (Extended Data Fig. 10a).

Comparison across datasets

When comparing the significant results derived from different methods or datasets, we used the π1 statistic, setting lambda between 0 and 0.8 in increments of 0.001 (http://github.com/jdstorey/qvalue). For all π1 calculations we only used genes that could be captured in both datasets. Comparison to GTEx ASE events obtained by Illumina were done using the single-nucleotide variant level read counts, annotated using the GENCODE annotations, for continuity. The datasets were merged and the variant with the highest read count across both methods was selected per gene across samples.

Colocalization analysis

We mined all colocalization results between GTEx sQTLs and 5,586 GWAS traits[4] and filtered for loci with regional colocalization probability (rcp) > 0.5 and removed the human leukocyte antigen (HLA) region. We then mapped each sQTL to its corresponding gene and overlapped that gene set with the significant ASTS genes per tissue. For the overlapping genes we verified that the lead sQTL used for colocalization was a heterozygous variant in the donor for which we had ASTS data. This strict filtering resulted in five genes *SRP14*, *DUSP13*, CD36, *IFITM2* and *ELP5*.

Combinatorial allele-specific analysis in control and PTBP1 knockdown samples

For each donor the control and the knockdown samples were processed together, and the most highly covered variants using both samples were selected per gene. Specific allelic events per condition were defined using an FDR threshold of 0.05.

We downloaded all eCLIP (bed narrowPeak) RNA protein binding data in GRCh38 (ref. [58]). All peaks were overlapped with the heterozygous variants per donor using bedtools intersect[59]. Finally, the maintained peaks were annotated to the nearest gene using a 10 kb window around each gene.

Annotation of variant consequences

Annotation of protein-coding regions was generated by running Ensembl VEP (v.104) with the --most_severe flag on the GTEx v.8 release. We did two rounds of annotation, the first one using non-small RNA genes from the GENCODE v.26 GTF file and the second one by supplementing this annotation with newly identified FLAIR transcripts for these genes. We predicted the productivity of each transcript using flair predictProductivity.py (v.1.4)[31] using only the longest ORF for each transcript. The frame of each transcript was corrected using genomeTools (v.1.6.1)[60].

Transcripts were first annotated based on the gene biotypes. Transcripts originating from protein-coding genes were classified as 'protein-coding' if both a start and a stop codon were found, 'nonsense-mediated-decay' if a premature termination codon was found, 'processed transcript' if there was no start codon and 'non-stop decay' if there was a start but no stop codon. Novel transcripts without a conclusive coding sequence frame found had their biotype revised. Novel transcripts marked as protein-coding, processed transcript, sense intronic, antisense or long intergenic non-coding RNA (lincRNA) with intron retention, had their biotype changed to 'retained intron'. Similarly, protein-coding and processed transcripts that came from the opposite strand were re-annotated as 'antisense', those that overlapped an intron as 'sense-overlapping' and those that were intergenic as 'lincRNAs'. If none of these conditions were filled, protein-coding transcripts had their annotation changed to processed transcript. The gene coordinates were extended if one of the transcripts was found to be outside them. This led to 73,599 transcripts being added.

CADD scores for all annotated variants were obtained using the v.1.5 release[61]. We compared the CADD scores between the reassigned and the non-reassigned variants (down sampled to match the size of the total number of reassigned variants per consequence group). We then used a *t*-test to compare the means of the two groups.

Rare variant analysis

We extracted all heterozygous variants within a 10 kb window around each gene assessed for ASTS in a donor-specific manner. Variants were filtered for MAF < 0.01 and the worst consequence was maintained per variant. We found a median of four rare variants per gene. We observed that 50% of genes across all samples had at least one rare variant. We calculated enrichment using a binomial test, setting all variants as background.

Reporting summary

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

Data availability

Raw long-read data generated as part of this manuscript are available in the GTEx v.9 release under dbGAP accession number phs000424.v9 and on AnVIL at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V9_hg38. The GTF file of flair transcripts, along with the transcript-level overall and allelic expression quantifications from GENCODE v.26 and flair transcripts, are available on the GTEx portal (https://gtexportal.org/home/datasets). The GTEx WGS, Illumina short-read, the allelic analysis, eQTLs and sQTLs and enloc colocalization files are all part of the GTEx v.8 release phs000424.v8. In addition, we used the transcript and gene counts available from https://gtexportal.org/home/datasets. The GRCh38 human genome reference and GENCODE v.26 processed for GTEx were used in this analysis (https://console.cloud.google.com/storage/browser/gtex-resources). The CHESS and Workman transcript datasets were downloaded from GitHub (https://github.com/chess-genome/chess and https://github.com/nanopore-wgs-consortium/NA12878). ENCODE eCLIP data was downloaded from https://www.encodeproject.org/.

Code availability

All original code used in the manuscript is released as part of a software package, https://github.com/LappalainenLab/lorals. General scripts are available at https://github.com/LappalainenLab/lorals_paper_code (https://doi.org/10.5281/zenodo.6529254).

45. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
46. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
47. Alasoo, K. Wiggleplotr: Make read coverage plots from bigwig files. *Bioconductor* https://doi.org/10.18129/B9.bioc.wiggleplotr (2017).
48. Pertea, M. et al. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.* **19**, 208 (2018).
49. Pertea, G. & Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Research* **9**, 304 (2020).
50. Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
51. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
52. Deutsch, E. W. et al. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteomics Clin. Appl.* **9**, 745–754 (2015).
53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
54. Nowicka, M. & Robinson, M. D. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research* **5**, 1356 (2016).
55. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
56. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).
57. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. (Academic Press, 2013).
58. Van Nostrand, E. L. et al. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
59. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
60. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10**, 645–656 (2013).
61. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

Acknowledgements We thank M. Micorescu and K. Potamousis from the Oxford Nanopore Technologies commercial team for their help in generating the data. D.A.G. was funded by NIH grant nos. R01GM124486 and U24DK112331. T.L. was funded by NIH grant nos. R01GM124486, R01GM122924, R01AG057422 and UM1HG008901. P.H. was funded by NIH grant no. R01GM124486. A.G. was funded by Roy and Diana Vagelos Pilot Grant. Funding for long-read sequencing of GTEx samples at the Broad was provided by a Broad Ignite grant. N.E. and P.M. were supported by NIGMS award no. R01GM140287. N.R.T. was funded by NIH grant no. K01-HL140187.

Author contributions D.A.G., T.L. and B.C. conceived and designed the project. D.A.G. performed most of the data analysis. G.G., A.G. and X.D. carried out the library preparation and sequencing. P.H. packaged the code. L.J., R.J., H.T. and M.S. provided and analysed the data for the proteomic validation. P.H. and K.B. assisted in allelic expression analysis. K.G. carried out the base-calling. N.E. and P.M. performed power analyses and advised on analysis methods. A.G. carried out the PTBP1 knockdown. N.R.T. and P.E. provided the CVD samples. T.B. and M.C. aided in the data generation. F.A., K.A., E.D.H., S.J., D.G.M., B.C. and T.L. provided feedback on the study design and data analysis. N.E., F.A., N.R.T., E.D.H., S.J., P.M. and D.G.M. provided feedback on the manuscript. E.D.H., S.J., D.G.M. and T.L. supervised the work. D.A.G., T.L. and B.C. wrote the manuscript with contributions from other authors. All authors read and approved the manuscript.

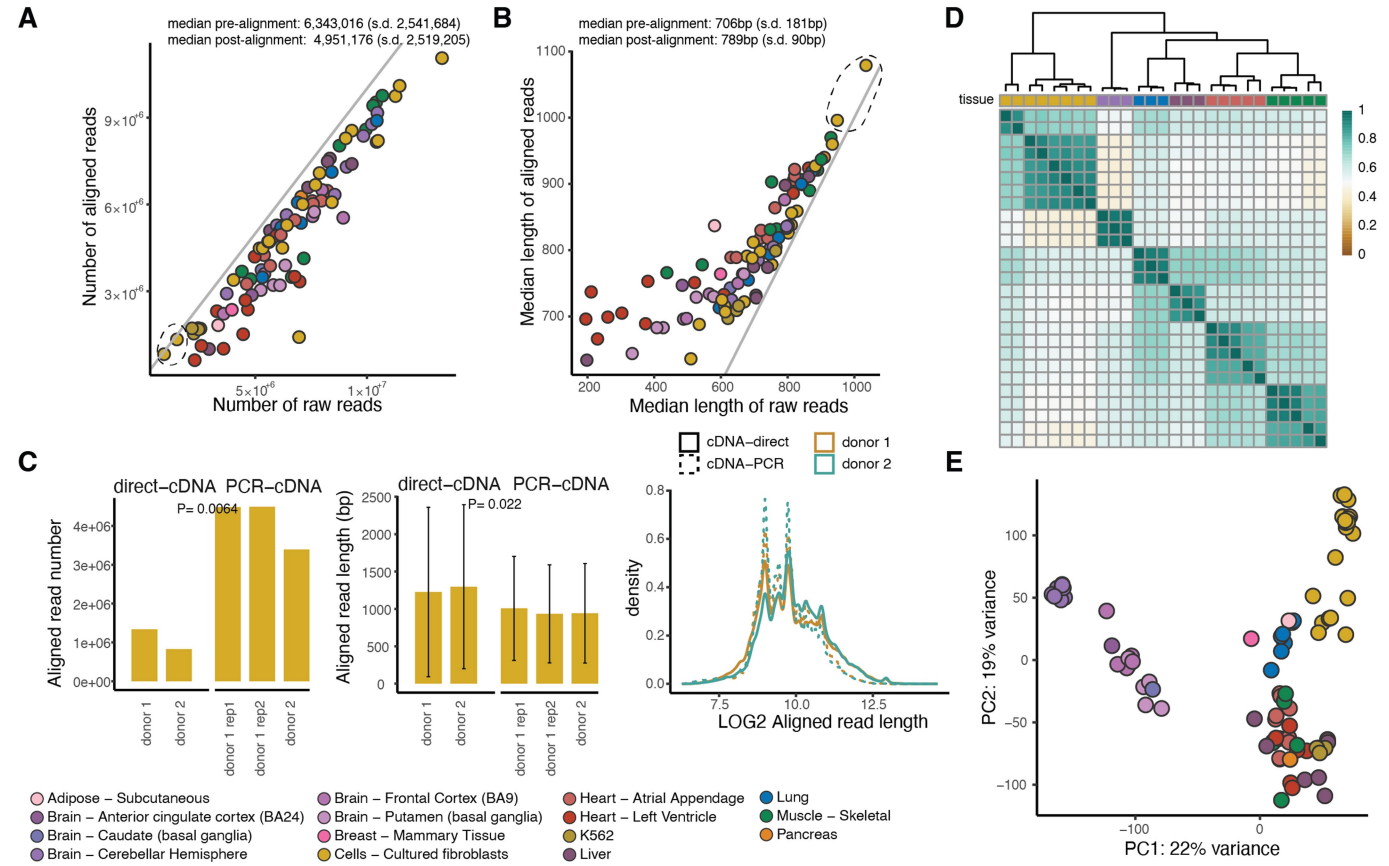
Competing interests D.A.G. is currently a fellow at Vertex Pharmaceuticals. X.D., E.D.H. and S.J. are employees of Oxford Nanopore Technologies and are shareholders and/or share option holders. F.A. has been an employee of Illumina, Inc., since 8 November 2021. P.T.E. has received sponsored research support from Bayer AG and IBM Health, and he has served on advisory boards or consulted for Bayer AG, MyoKardia and Novartis; none of these activities are related to the work presented here. D.G.M. is a founder with equity in Goldfinch Bio, a paid advisor to GSK, Insitro, Third Rock Ventures and Foresite Labs, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer and Sanofi-Genzyme; none of these activities are related to the work presented here. B.C. is currently employed at Third Rock Ventures. The other authors declare no competing interests.

Additional information

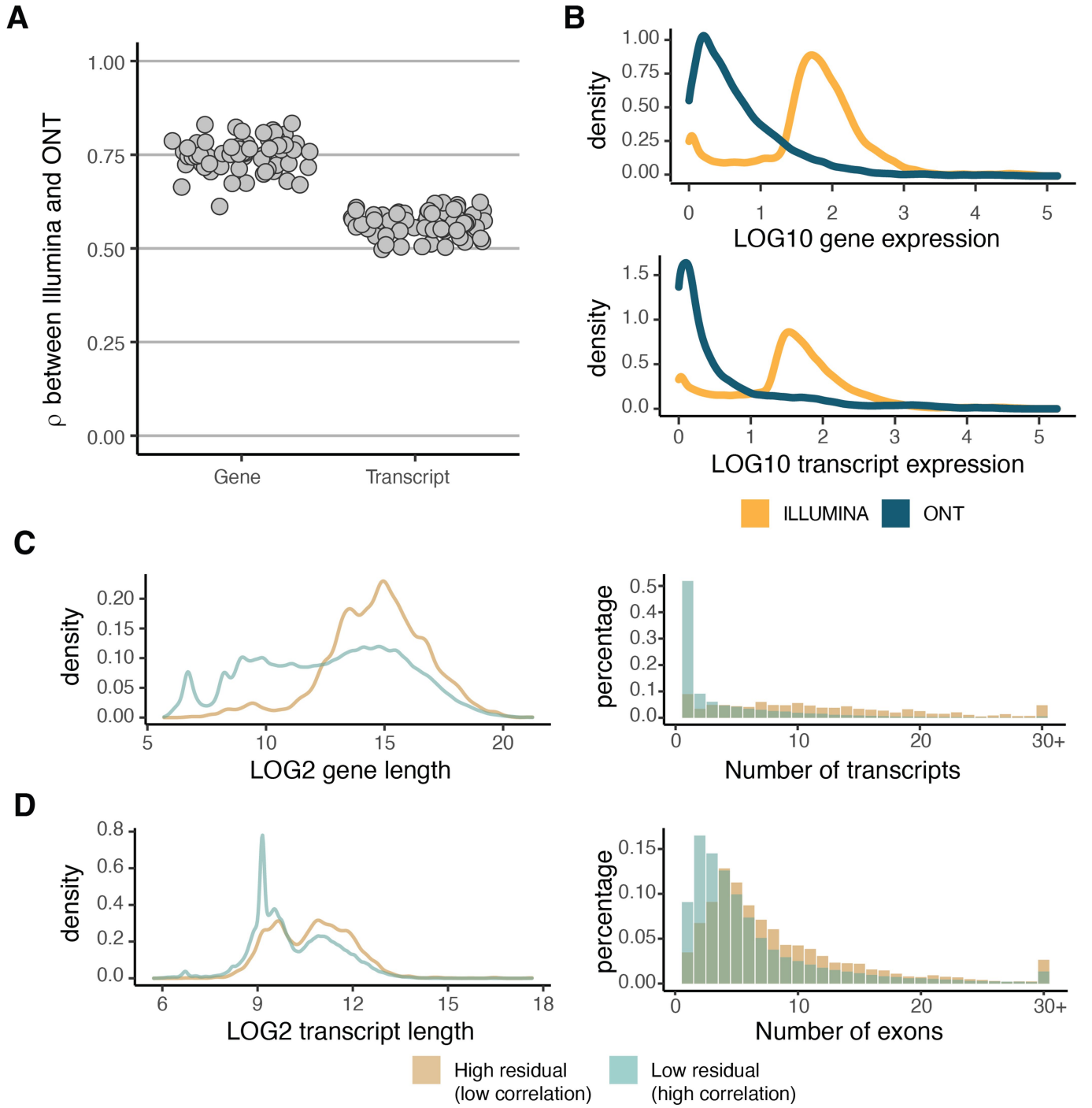
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-022-05035-y.

Correspondence and requests for materials should be addressed to Dafni A. Glinos, Garrett Garborcauskas, Tuuli Lappalainen or Beryl B. Cummings.

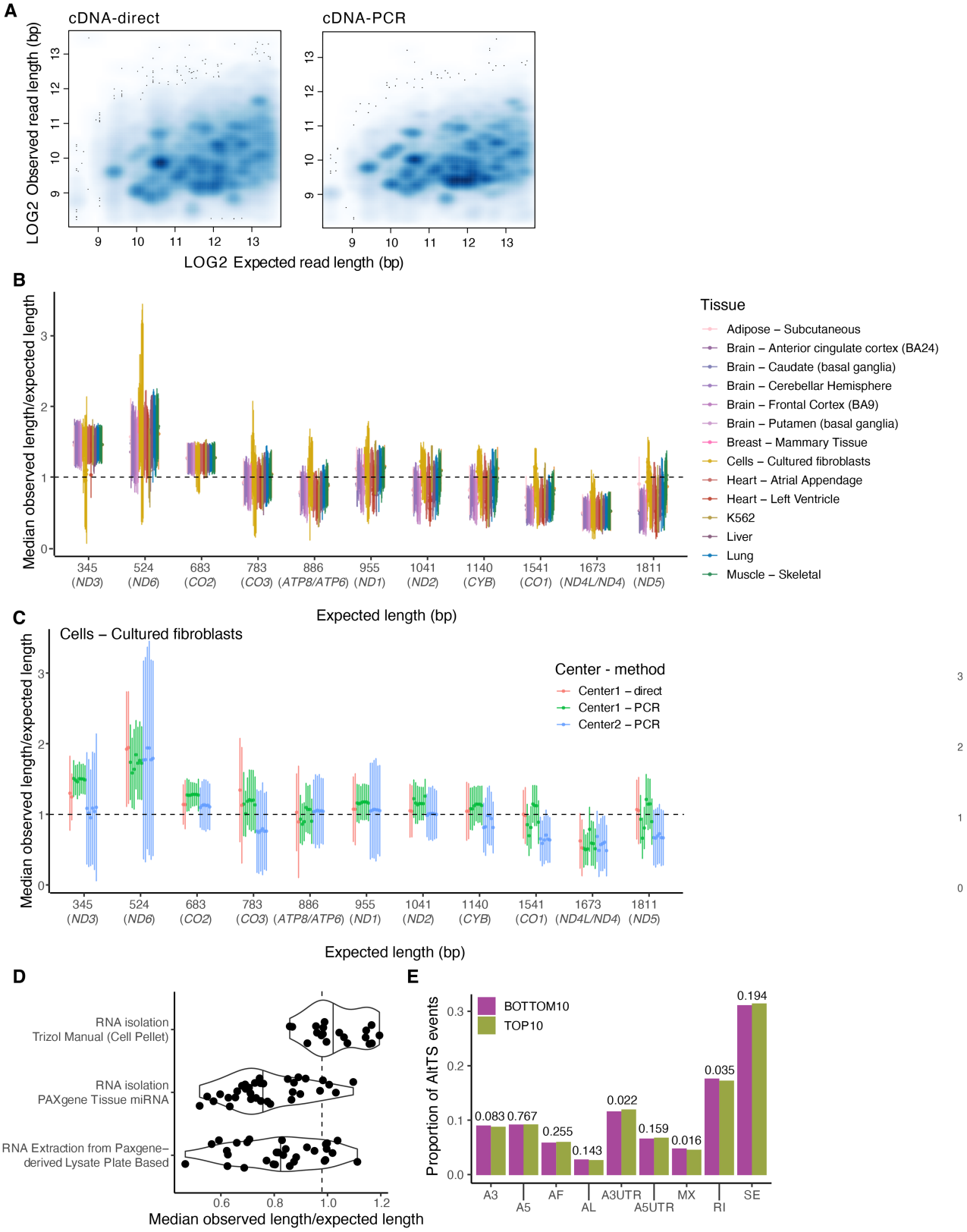
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



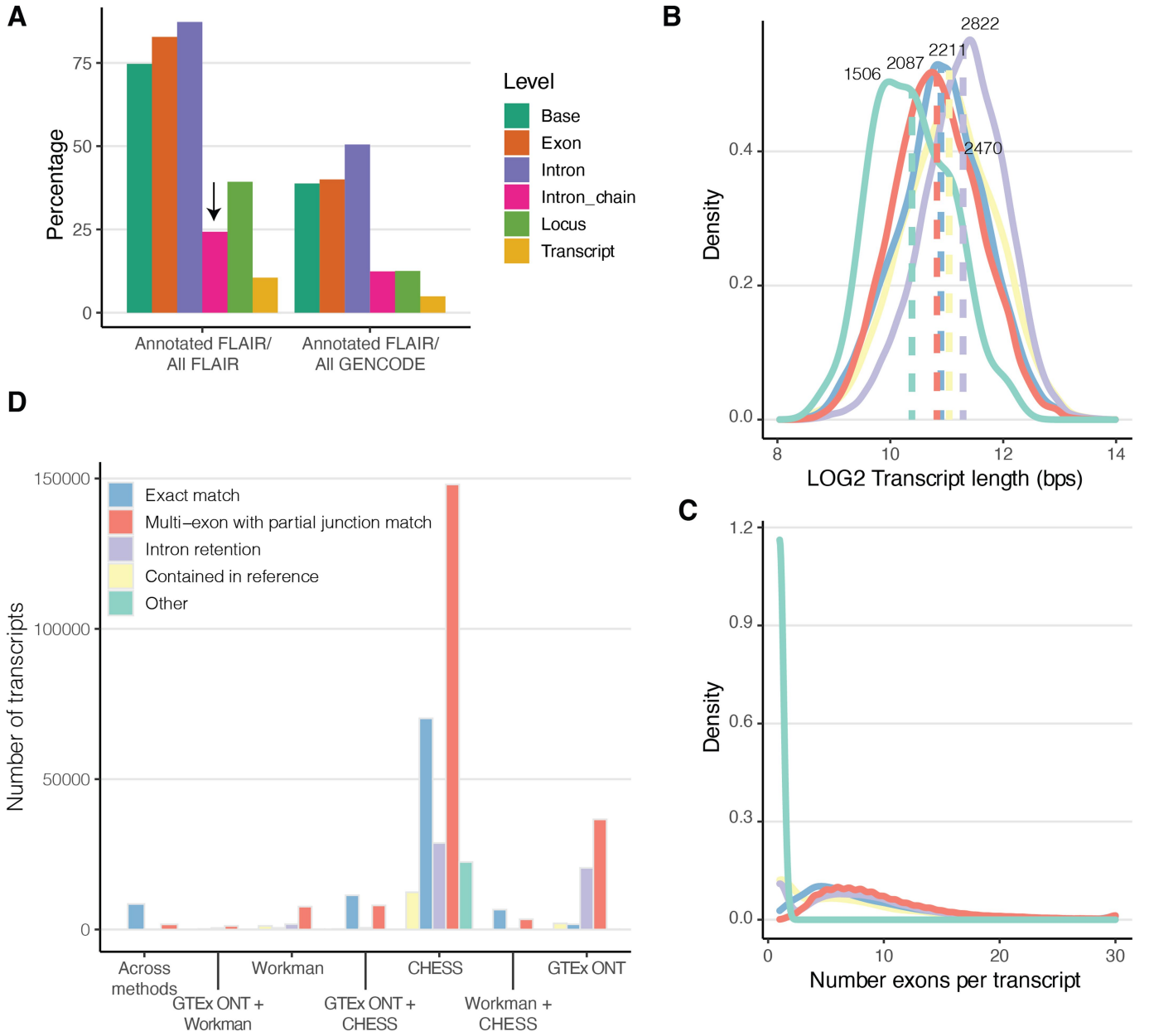
Extended Data Fig. 1 | Quality control of the dataset. A) Number and **B**) median length of raw and aligned reads per sample. The diagonal lines correspond to intercept = 0. With the dashed black circle, we highlight the two samples sequenced using the direct-cDNA technology. **C**) Read number and read length in two fibroblast cell line samples (one of which was sequenced in replicate) that were sequenced using both the direct-cDNA and the PCR-cDNA protocol for 48 h. *P* values were calculated using a two-sided t-test. Error bars: standard deviation from the mean. **D**) Hierarchical clustering using Euclidean distance for replicate samples aligned to GENCODE for transcripts with expression above 3 TPM in at least 5 samples. **E**) Principal component analysis using all 88 samples aligned to GENCODE (v26) for transcripts with expression above 3 TPM in at least 5 samples.



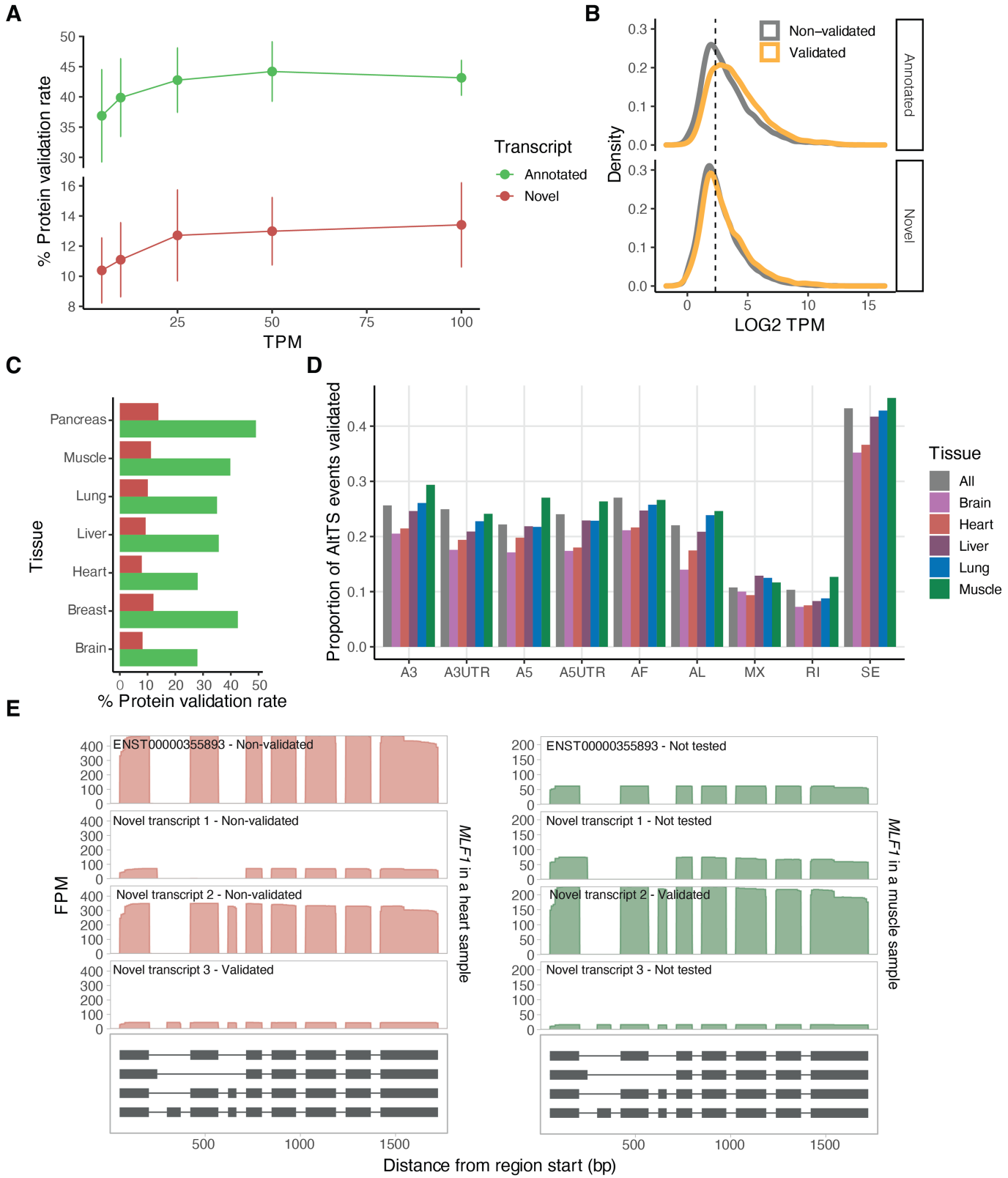
Extended Data Fig. 2 | Comparison between ONT and Illumina gene expression. A) Correlation between the transcriptome of each sample quantified by ONT and by Illumina sequencing technologies. **B**) Normalized gene and transcript expression for high residual (|residual| > 1) genes and transcripts retrieved from the Spearman correlation analysis. **C**) Characteristics of genes and **D**) transcripts with high or low residuals with respect to gene/transcript length, number of transcripts per gene and number of exons per transcript.



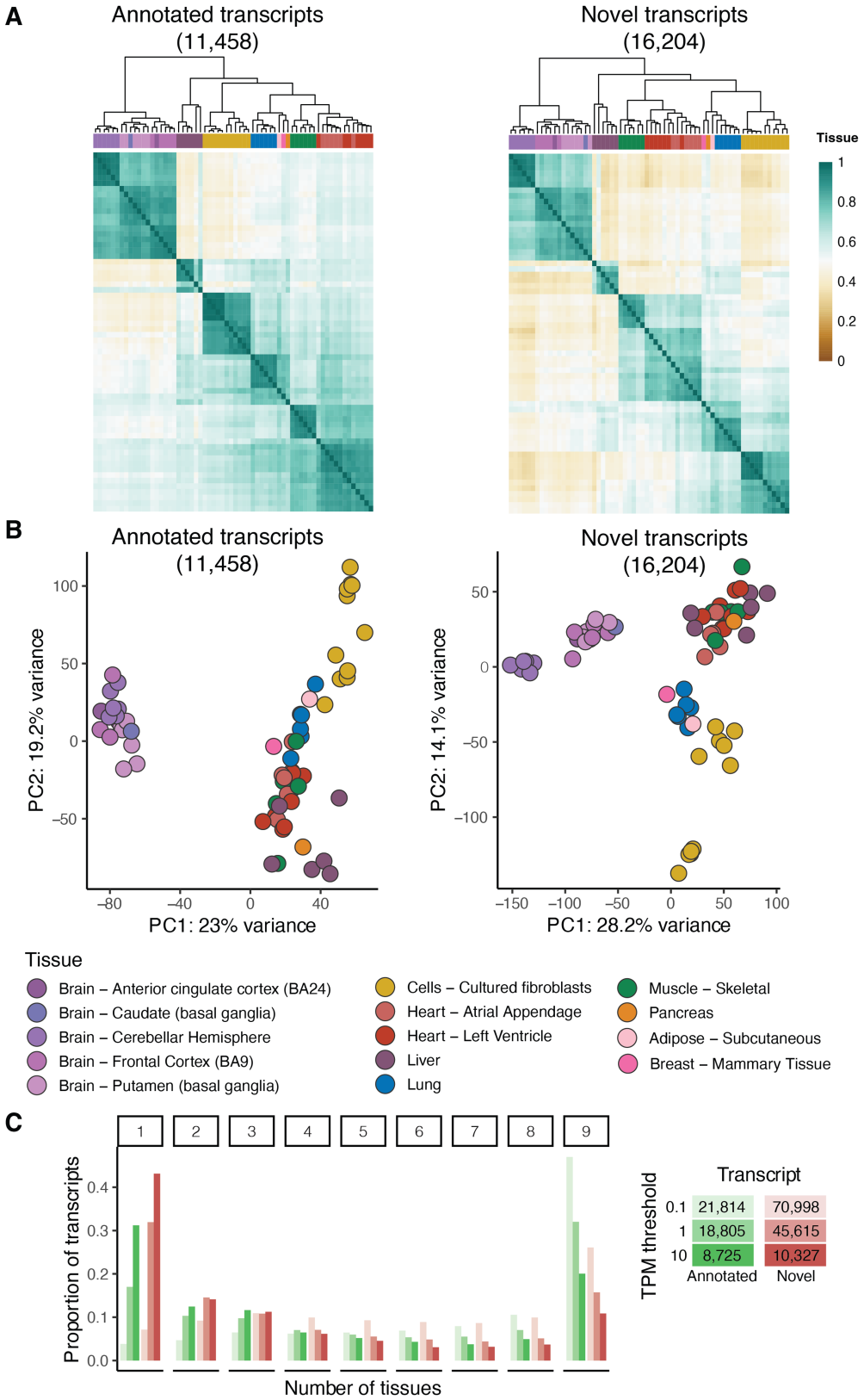
Extended Data Fig. 3 | Three prime bias analysis using mitochondrial reads. **A**) Observed versus expected read length for one sample sequenced using both direct (Spearman $R^2 = 0.3$) and PCR cDNA (Spearman $R^2 = 0.26$) protocols. The discrete clusters below the diagonal represent incorrect assignments to GENCODE isoforms (potential novel transcripts), and the diffuse shading represents fragmented RNA. Relationship between the expected transcript read length and the fraction of observed nanopore poly(A) RNA reads over the expected full length by sample for **B**) all samples and **C**) only fibroblasts. Labels are for mitochondrial genes without the MT prefix. The median was calculated per sample and error bars represent standard deviation. **D**) Median fraction of full-length per method by which RNA was isolated. **E**) Comparison of alternative transcription structure events found in highly expressed transcripts in the top and bottom 10% of samples ranked by 3′ bias. We observed no difference between the two deciles when using a two-sided proportionality test.



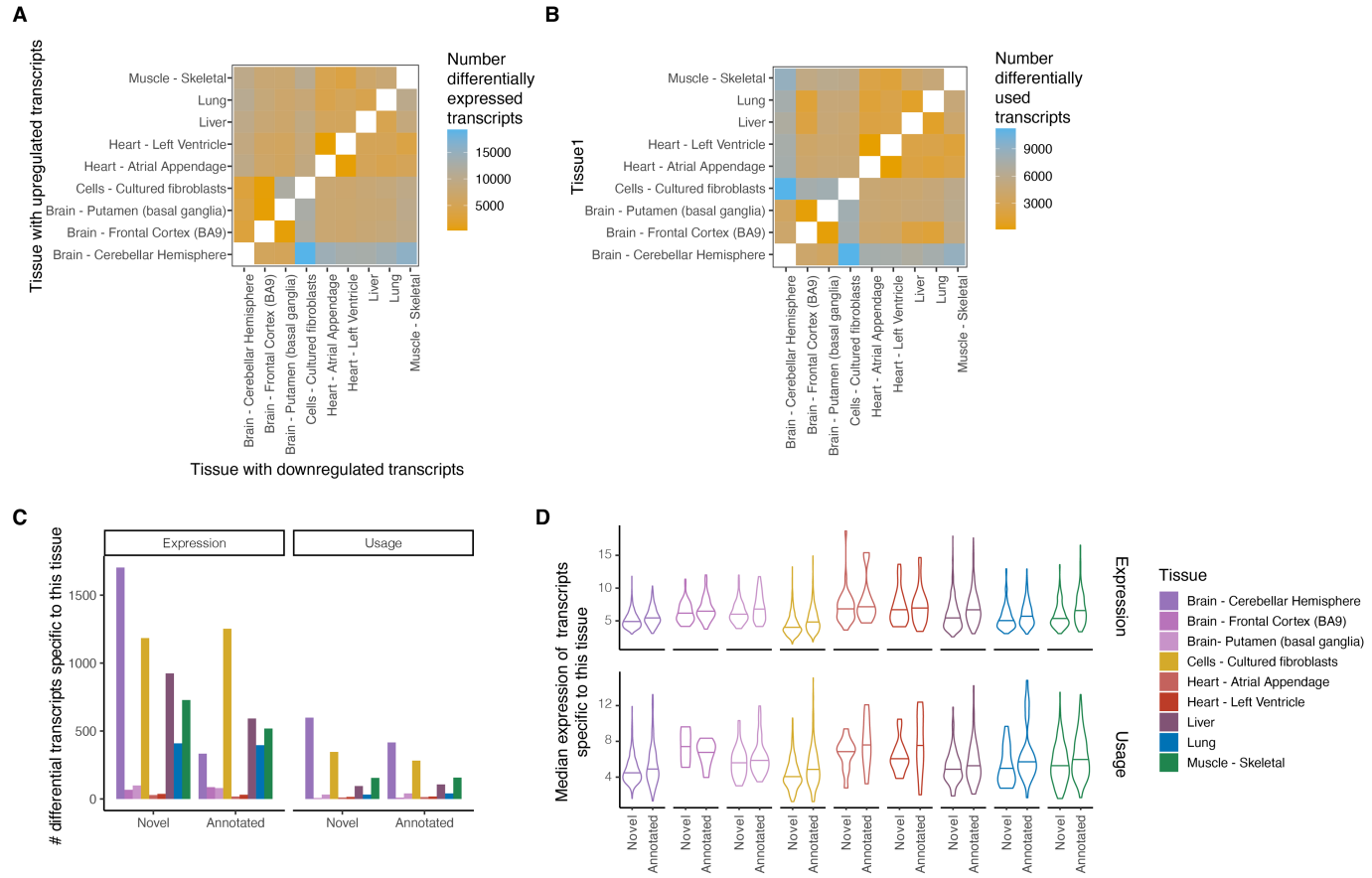
Extended Data Fig. 4 | FLAIR transcript characterisation. A) FLAIR transcripts comparison to GENCODE with respect to different genomic levels. The difference between intron chain and transcript is that the former only looks at matching the intron boundaries, therefore allowing variation in the UTR regions. **B**) Transcript length and **C**) number of exons per transcript classified by comparison to GENCODE. **D**) Number of overlapping transcripts between the ones identified in this paper and the ones released by a) CHESS and b) Workman.



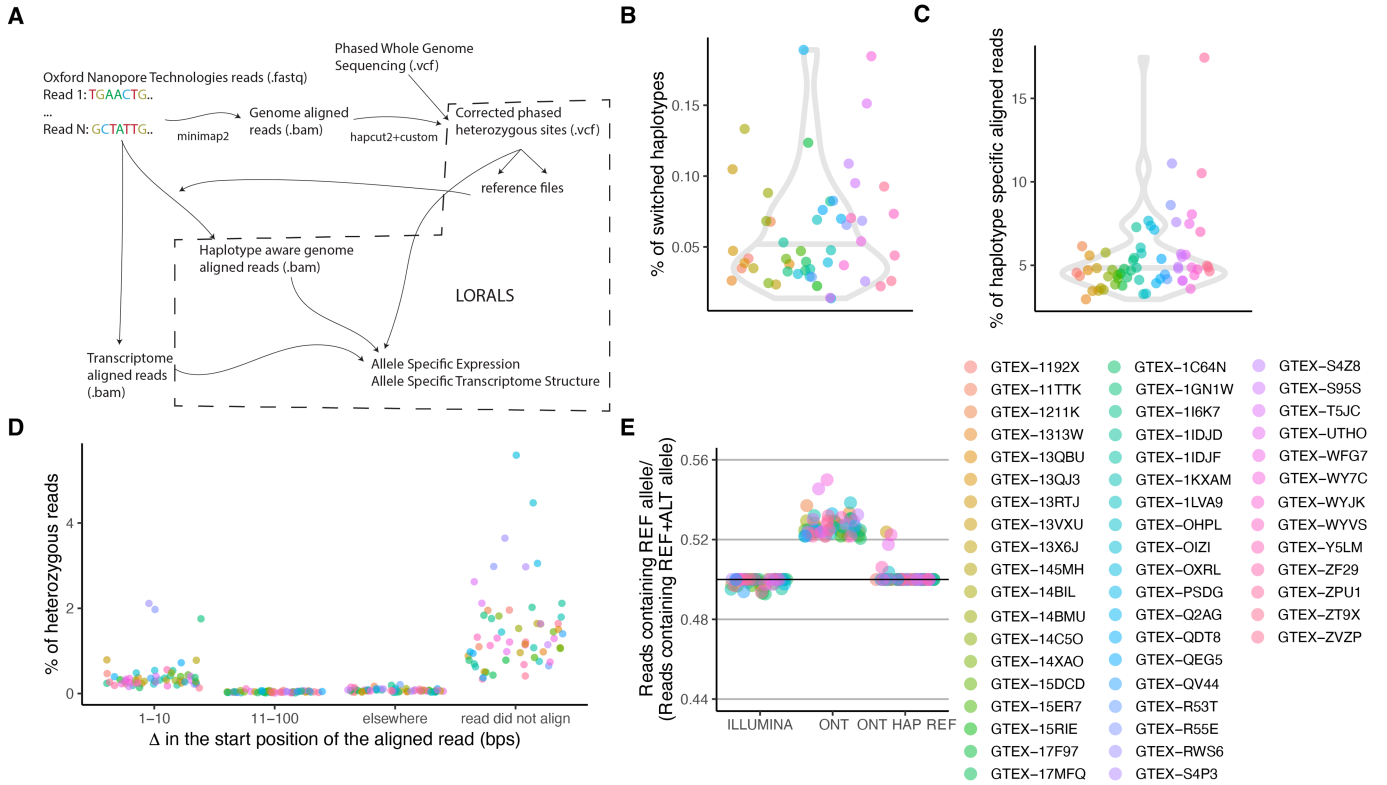
Extended Data Fig. 5 | Protein validation analysis of transcripts from matched tissues. A) Percentage of validated transcripts at the protein level using mass spectrometry for different TPM thresholds. Each point represents the mean across n = 7 assayed tissues and error bars represent the standard deviation. **B**) Mean expression per tissue with over one sample (lung, liver, heart, muscle and brain) of annotated and novel transcripts stratified by their validation status. The vertical line denotes the 5 TPM threshold used. **C**) Percentage of validated transcripts at the protein level using mass spectrometry per primary tissue. **D**) Proportion of the AltTS events validated per tissue. **E**) *MLF1* is an example of a gene with multiple highly-expressed transcripts across both muscle and heart tissues with a different transcript validated in each tissue. A3: alternative 3′ splice site; A5: alternative 5′ splice site; AF: alternative first exon; AL: alternative last exon; A3UTR: alternative 3′ end; A5UTR: alternative 5′ end; MX: mutually exclusive exons; RI: retained intron; SE: skipped exon.



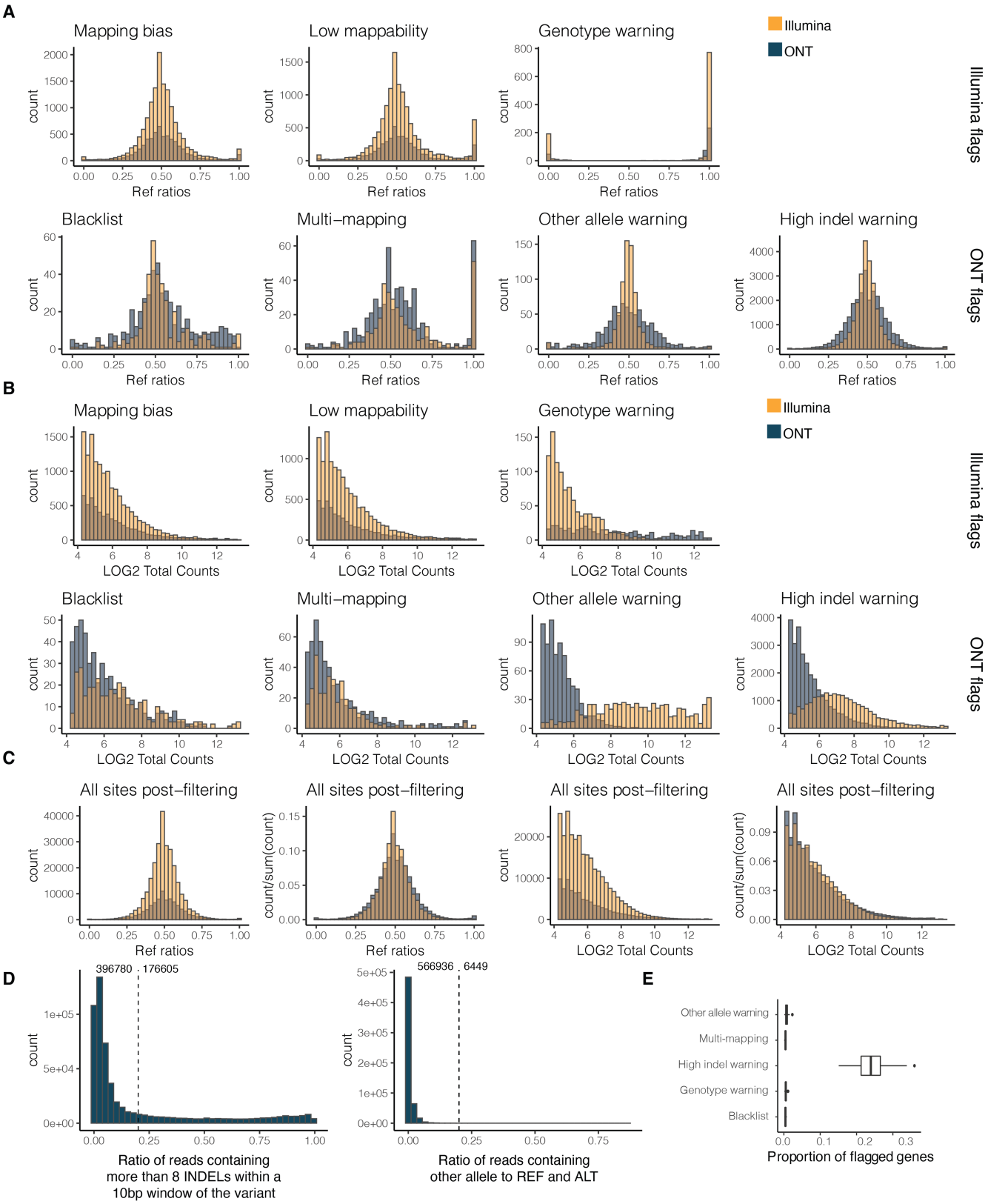
Extended Data Fig. 6 | Transcript expression overview of novel and annotated transcripts. A) Hierarchical clustering using euclidean distance and **B**) principal component analysis for selected samples aligned to GENCODE for transcripts with expression above five TPM in at least three samples separated based on whether they are novel or not. **C**) Proportion of transcripts expressed at different TPM thresholds and classified based on how many tissues express the transcript in at least two samples. The total number of transcripts per group and threshold is included in the legend.



Extended Data Fig. 7 | Differential transcript expression and usage between tissues. Heatmap of number of differentially **A**) upregulated transcripts and **B**) used transcripts (FDR ≤ 0.05) in pairwise comparison of tissues with at least five samples. In differential expression analysis we identify up- or down-regulated transcripts per pairwise comparison (asymmetrical heatmap) while in the differential transcript usage analysis there is no direction of effect (symmetrical heatmap). **C**) Number of differentially expressed or used transcripts that were specific to that tissue. **D**) Median gene expression across all transcripts that are specific to a tissue.

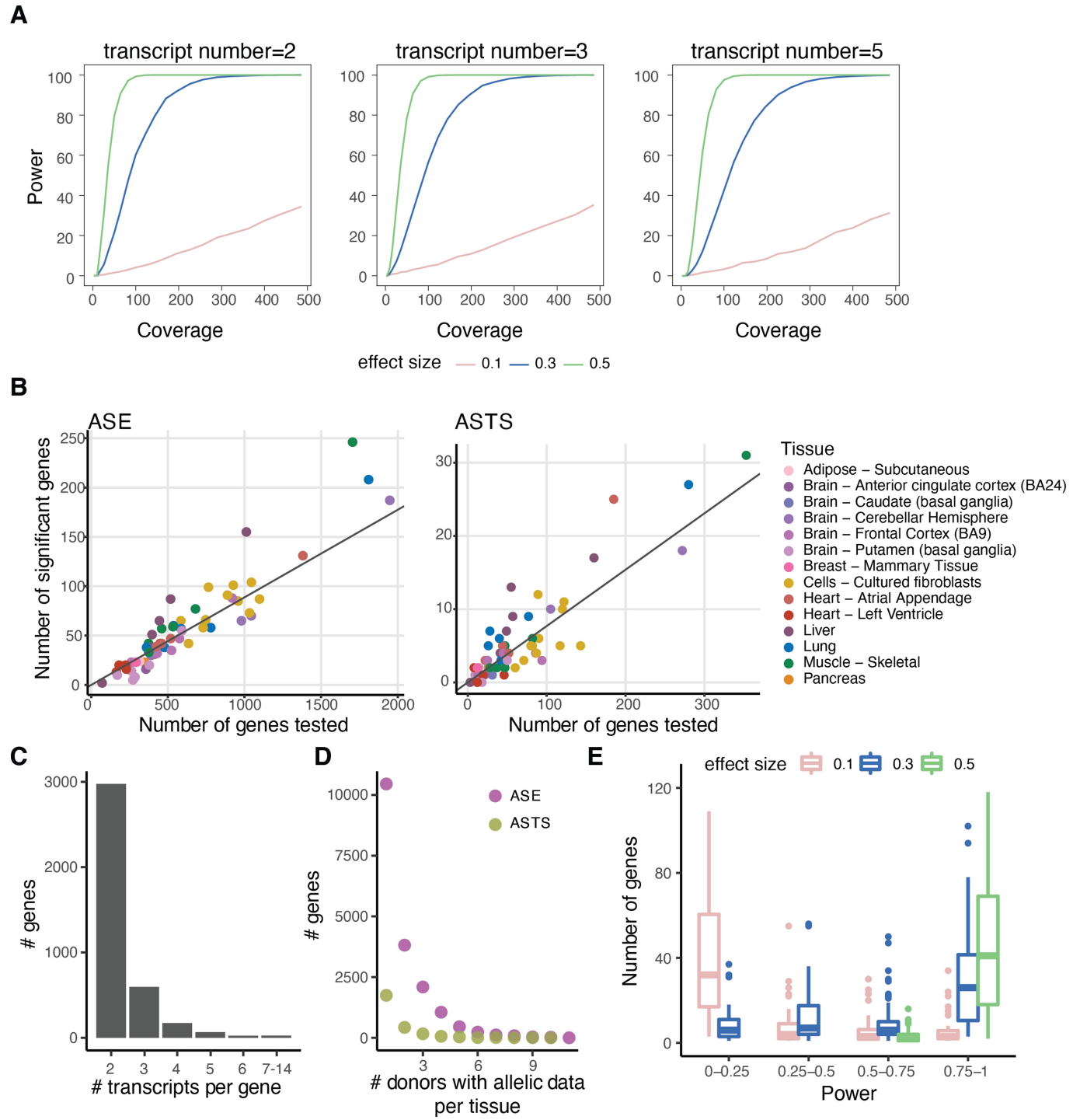


Extended Data Fig. 8 | LORALS pipeline development and aligning statistics. A) Pipeline for allele-specific analysis. Raw long-reads are first aligned to the genome using minimap2. This alignment is used to correct the phase of some of the heterozygous variants on the whole genome sequencing vcf. This new file is then used to generate personalized genome reference files against which the raw reads are again aligned using minimap2. The raw reads are also aligned to the transcriptome using minimap2. The VCF file along with the genome aligned reads and the transcriptome aligned reads are then fed into LORALS for allelic analysis. **B**) Percentage of switched haplotypes per donor informed by the long-read data. For this all samples from the same donor were merged to harmonize the files. **C**) Percentage of haplotype specific reads calculated as reads having a higher mapping score when using a personalized genome reference. **D**) Delta calculated as the difference in the start position of the aligned read between the genome aligned and the personalized genome aligned reads. Not shown are the reads that had Delta = 0. **E**) Reference ratio for the samples present in this study sequenced using Illumina technology and ONT technology aligned with two different approaches.

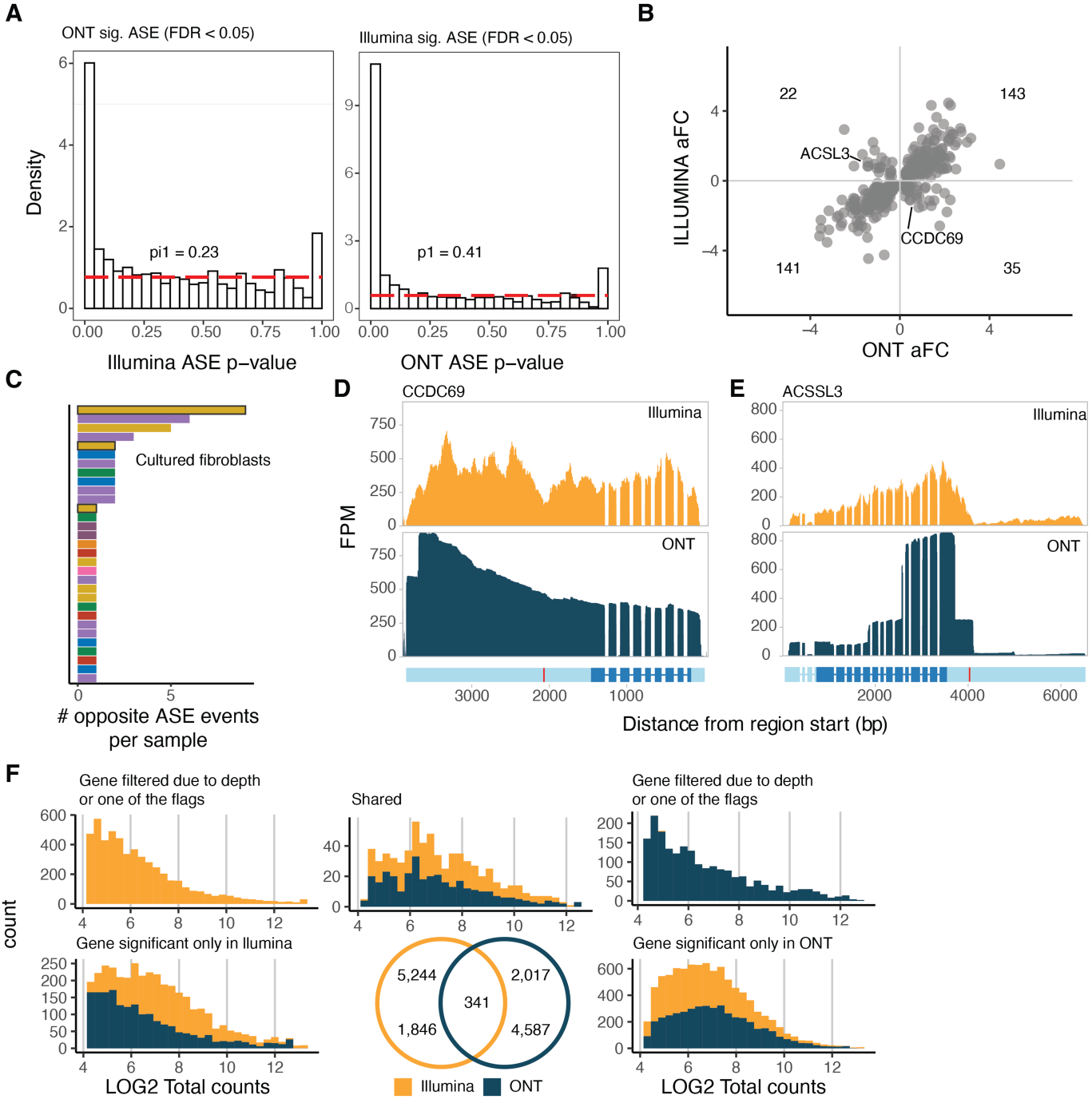


Extended Data Fig. 9 | See next page for caption.

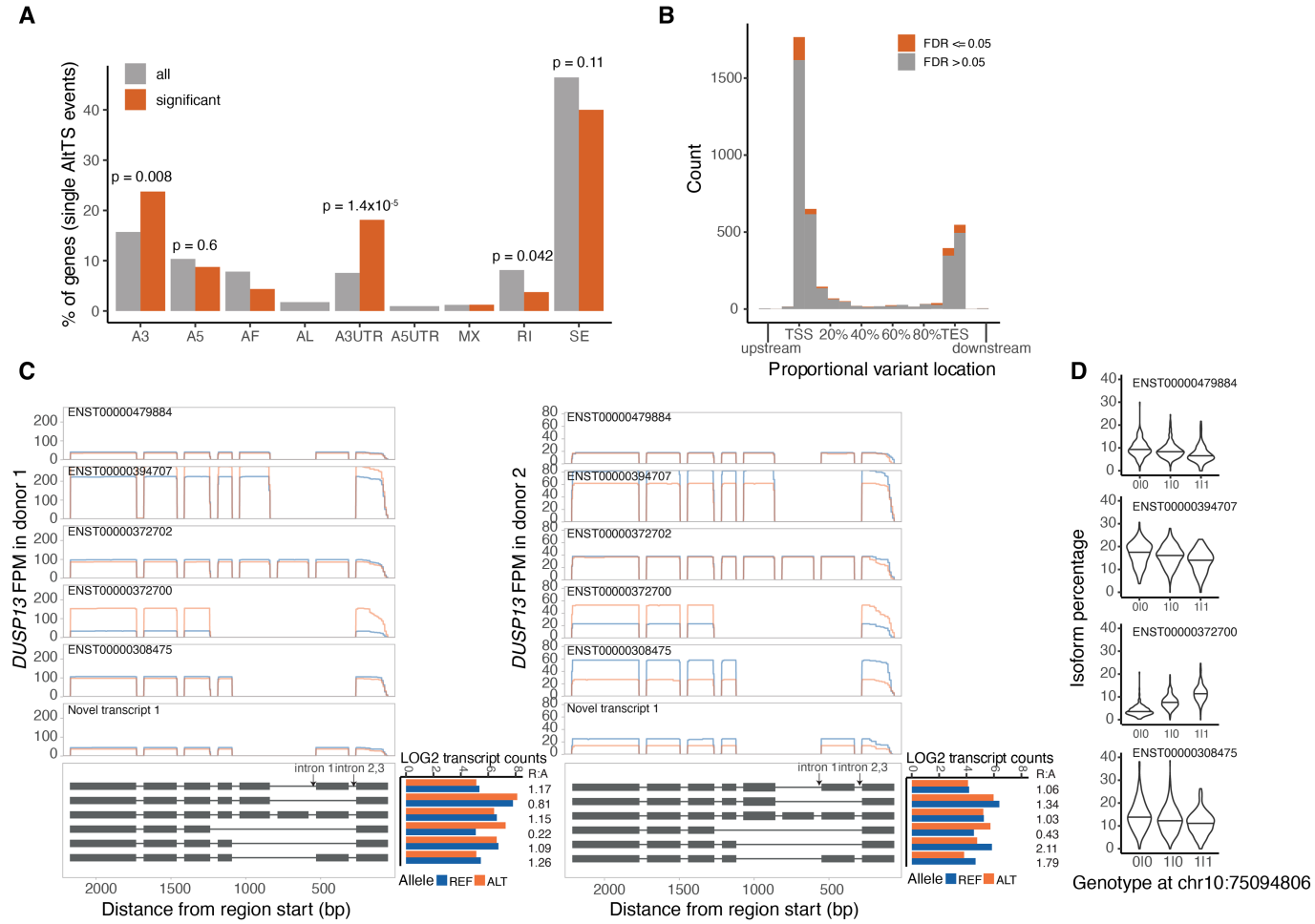
Extended Data Fig. 9 | LORALS pipeline allele specific analysis filter setting. A) Reference ratio and **B**) normalized reads counts across different Illumina and ONT flags for both of these sequencing technologies. Mapping bias: mapping bias in simulations; Low mappability: low-mappability regions (75-mer mappability < 1 based on 75mer alignments with up to two mismatches based on the pipeline for ENCODE tracks and available on the GTEx portal); Genotype warning: no more reads supporting two alleles than would be expected from sequencing noise alone, indicating potential genotyping errors (FDR < 1%); Blacklist: ENCODE blacklist. Multi-mapping: regions with multi-mapping reads constructed using the alignability track from UCSC using a threshold of 0.1 (so that a 100kmer aligning to that site aligns to at least 5 other locations in the genome with up to 2 mismatches); Other allele warning: regions where the proportion of reference or alternative allele containing reads is lower than 0.8; High indel warning: sites where the proportion of non-indel containing is lower than 0.8. **C**) Reference ratios and normalised reads counts of all kept sites across Illumina and ONT sequencing technologies. **D**) Distribution of the high indel warning ratios and the other allele ratios across all samples. **E**) Proportion of genes with at least 20 overlapping reads flagged per filter. The proportion was calculated across all genes for each sample (n = 59). The center corresponds to the median, the lower and upper hinges correspond to the 25th and 75th percentiles and the whiskers extend from the hinge to the smallest/largest value no further than 1.5 * inter-quartile range from the hinge.



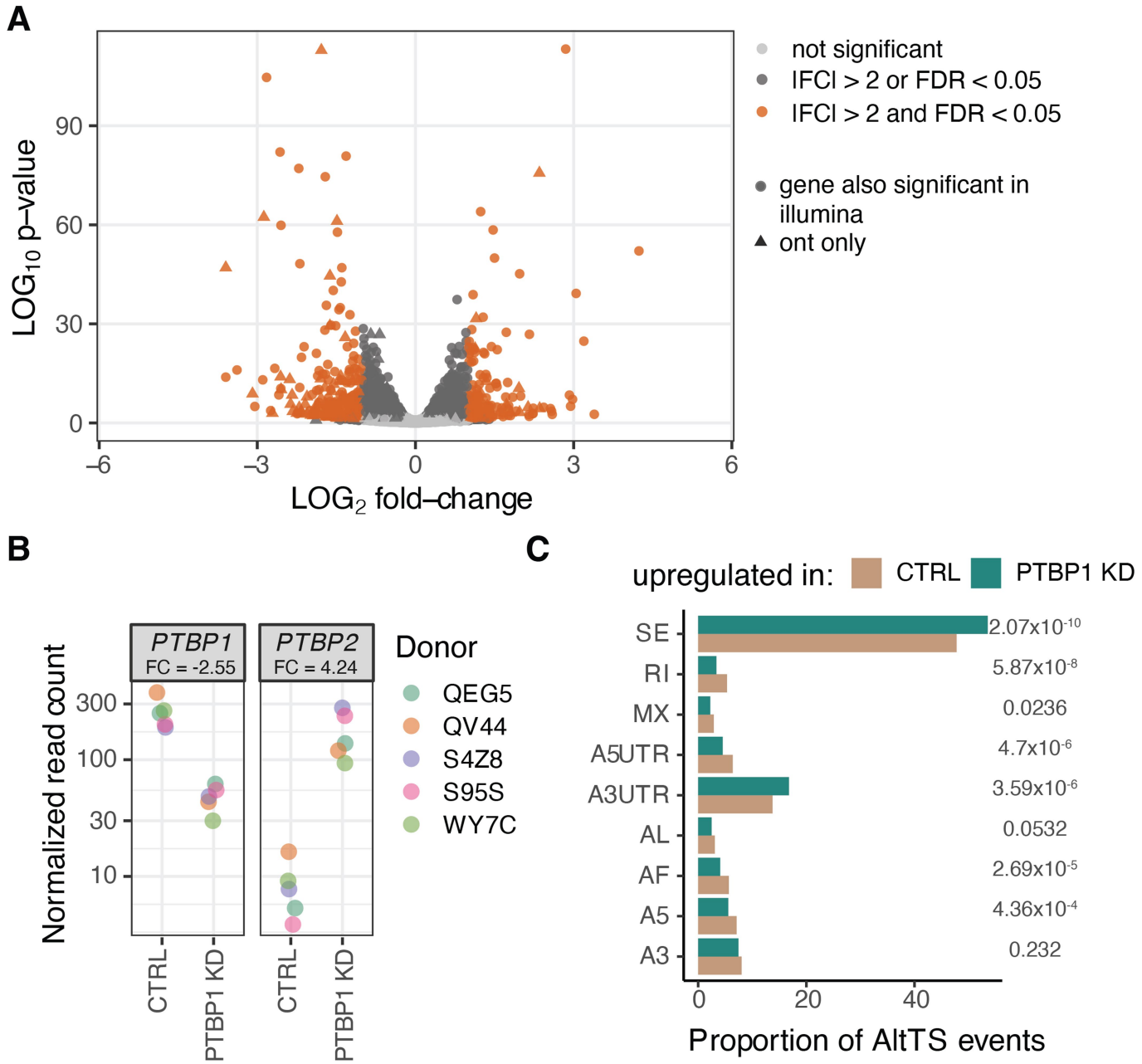
Extended Data Fig. 10 | Allele specific analysis on all GTEx samples. **A**) Estimated power for different number of transcripts (2, 3 or 5) with respect to the coverage (x-axis) for effect sizes 0.1 (low), 0.3 (medium) and 0.5 (high), derived from simulated count data. **B**) Number of genes tested for allele-specific expression (ASE) and allele-specific transcript structure (ASTS) and number of significant genes. The diagonal indicates the median percentage of significant genes (9% and 9%, respectively). **C**) Number of transcripts per gene tested for ASTS. **D**) Number of genes with allelic data across donors per tissue for ASE and ASTS. **E**) Total number of genes calculated per sample (n = 59) at different levels of power. Outliers are hidden for ease of viewing. The center corresponds to the median, the lower and upper hinges correspond to the 25th and 75th percentiles and the whiskers extend from the hinge to the smallest/largest value no further than 1.5 * inter-quartile range from the hinge.



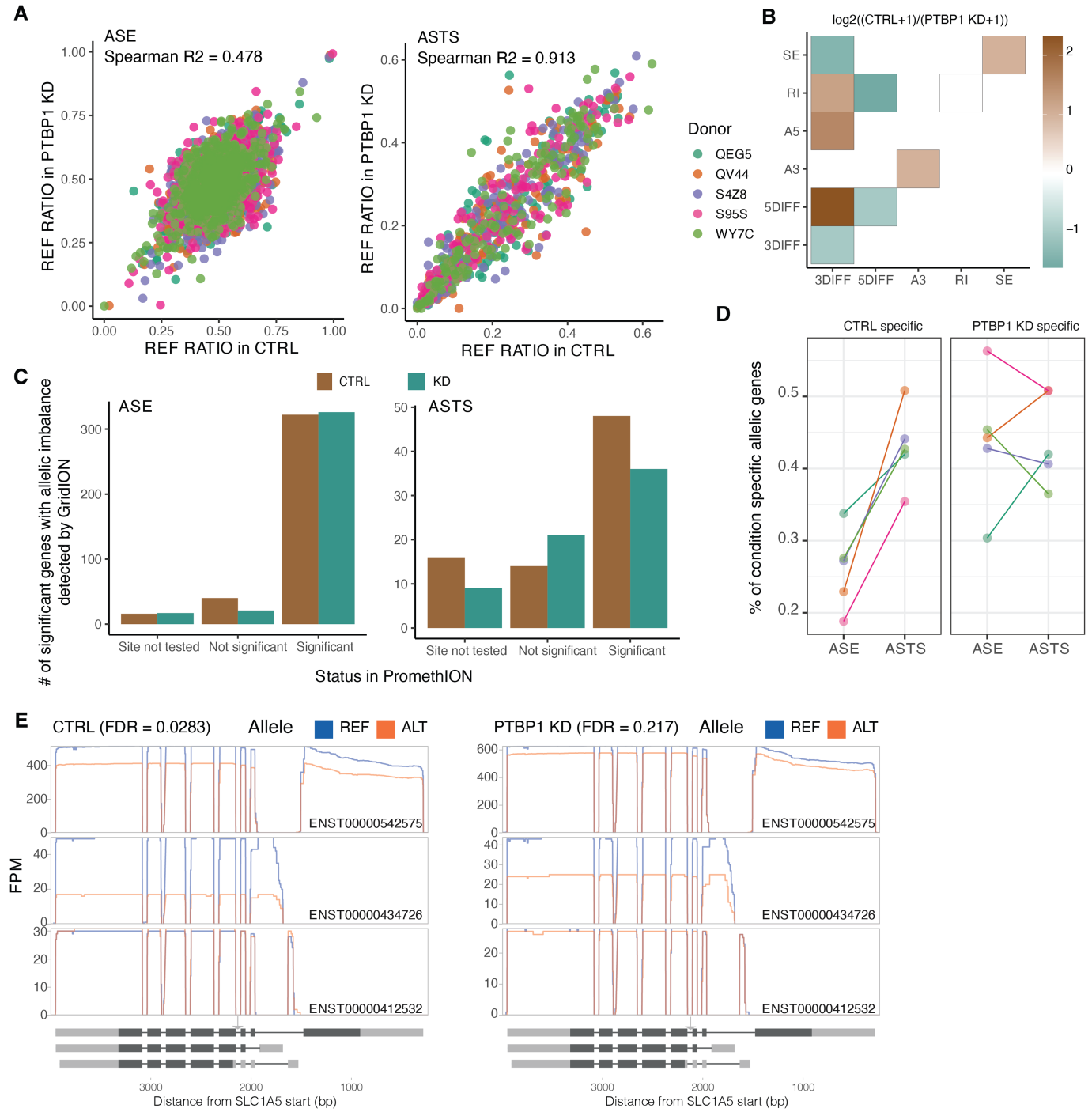
Extended Data Fig. 11 | Comparison of allele specific expression between ONT and Illumina. A) Proportion of significant ASE genes discovered using Illumina or ONT and replicated in the other method. π1 calculations are carried out up to *P* value = 0.5. **B**) Log allelic fold-change of Illumina and ONT of the shared ASE genes. **C**) Number of ASE events with opposite directions between ONT and Illumina per sample. Highlighted are the five fibroblast cell lines that were further cultured prior to sequencing, where 12/57 events were observed. RNA-seq read pile-ups for **D**) *CCDC69* and **E**) *ACSSL3* which have ASE in the opposite direction between the two methods. In red is shown the variant used to parse the reads between the two haplotypes. *CCDC69* differences can be attributed to a depression in the Illumina read pile-ups while *ACSSL3* can be attributed to the variant being in the 3′UTR, which is not well captured by Illumina reads. **F**) Venn diagram of the significant ASE genes discovered by Illumina and ONT. The LOG2 of total counts for each method is shown for each group of the Venn diagram.



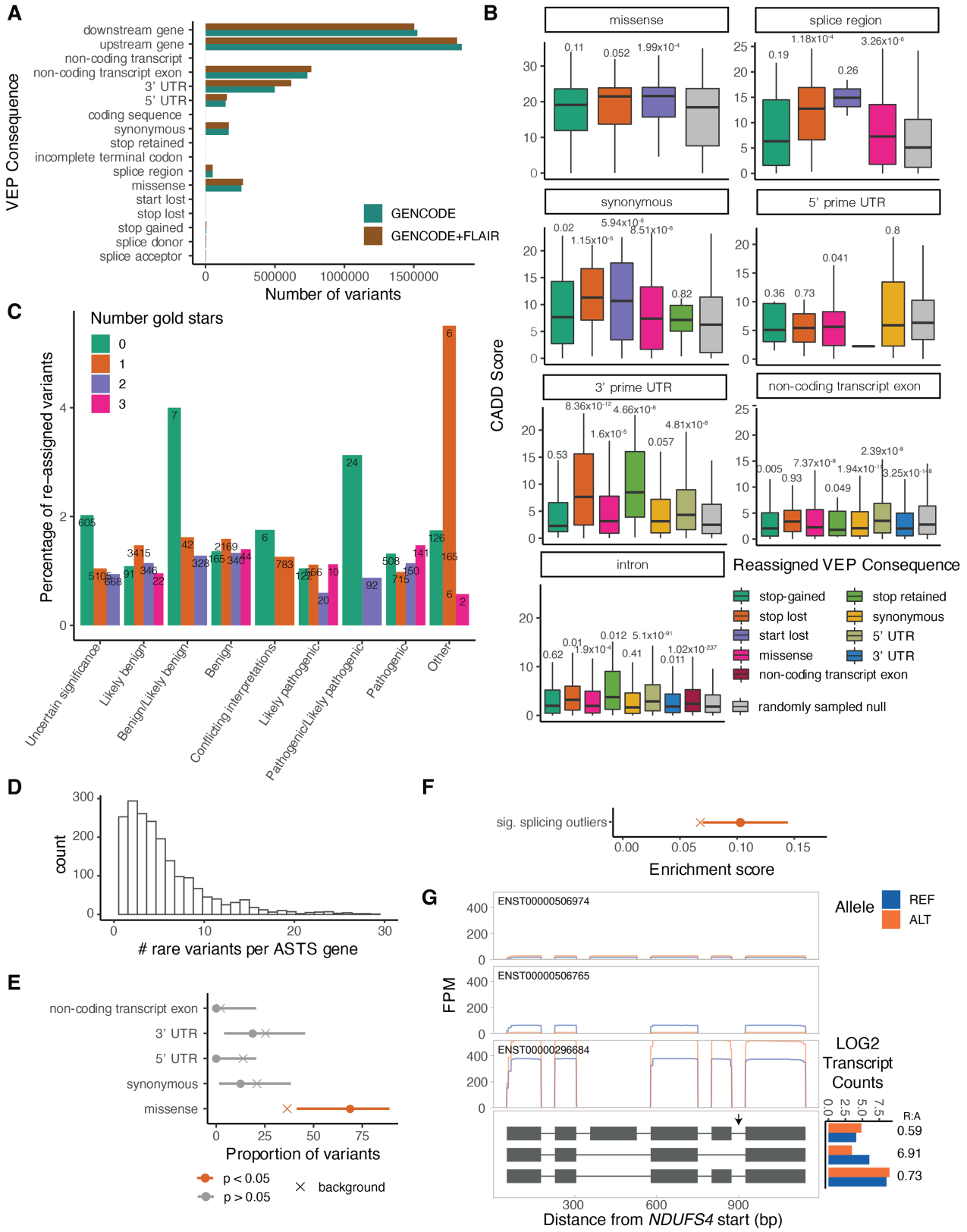
Extended Data Fig. 12 | Alternative transcript structure event annotation of allele specific events. A) Percentage of genes displaying a single alternative transcript structure event. *P* values were calculated using a two-sided binomial test. A3: alternative 3′ splice site; A5: alternative 5′ splice site; AF: alternative first exon; AL: alternative last exon; A3UTR: alternative 3′ end; A5UTR: alternative 5′ end; MX: mutually exclusive exons; RI: retained intron; SE: skipped exon. **B**) Average relative location of the heterozygous variant used for ASTS event, by grouping all the transcripts of an ASTS event together. **C**) Read pile-ups per transcript for the two donors displaying ASTS in *DUSP13* gene. In the lower panel the transcript structure is shown, without details of the coding sequence. **D**) Transcript percentage for four of the five *DUSP13* annotated transcripts with high read coverage in the GTEx v8.



Extended Data Fig. 13 | Differential expression analysis between PTBP1-KD and control samples. A) Volcano plot from differential gene expression between control and samples with PTBP1 knockdown using ONT data. *P* values were calculated using the Wald test in DESeq2. **B**) Gene expression profile in PTBP1 and PTBP2 genes (PTBP2 under normal circumstances has its expression restricted to the brain). **C**) Proportion of different alternative transcript structure events in transcripts upregulated in the control or the PTBP1 knockdown samples. *P* values were calculated using a two-sided proportionality test.



Extended Data Fig. 14 | Allele specific analysis of PTBP1-KD and control samples. A) Correlation between the control and the PTBP1 knockdown samples in the reference ratio of gene expression and transcript structure. **B**) Changes in ASTS by PTPB1 knockdown, as assayed by gridION, with the heatmap showing the co-occurrence of alternative transcript structure events that are observed at least once per each event (or a single time for the diagonal) in a given gene. Color corresponds to the $\log_2$ ratio of the number of events found in the control over PTBP1 knockdown (KD) samples. **C**) Number of significant ASE and ASTS genes found by gridION categorized based on their status in the PromethION data from the same samples. **D**) Proportion of genes displaying allele-specific patterns specifically in either control or PTBP1 knockdown samples. **E**) *SLC1A5* gene transcript read pile-ups which display significant ASTS only in the control sample only. The arrow indicates the location of the PTBP1 eCLIP site which contains a heterozygous variant in that donor.



Extended Data Fig. 15 | See next page for caption.

Extended Data Fig. 15 | Variant interpretation through novel transcripts and allele-specific transcript structure analysis. A) Number of variants per variant effect predictor (VEP) category using GENCODE v26 protein-coding genes with or without novel FLAIR transcripts. **B**) CADD score distribution of variants that were reassigned to a more severe consequence when the GENCODE gene annotations were complemented with the novel FLAIR transcripts, compared to variants that retained their annotation (down sampled to a similar size). *P* values from two-sided t-test. The center corresponds to the median, the lower and upper hinges correspond to the 25th and 75th percentiles and the whiskers extend from the hinge to the smallest/largest value no further than 1.5 * inter-quartile range from the hinge. **C**) Percentage of variants per clinical significance category that get reassigned when supplementing the gene annotation with the novel transcripts. The numbers above the bars correspond to the number of re-assigned variants. **D**) Number of rare variants per ASTS gene (10kb window around gene). **E**) Proportion of rare heterozygous variants per annotation in significant ASTS events. As a background all ASTS events were used, and *P* values were calculated using a two-sided binomial testing. **F**) Enrichment of the significant ASTS genes within splicing outliers. As a background all ASTS genes were used and *P* values were calculated using a two-sided binomial test. **G**) *NDUFS4* as an example of a gene with a rare heterozygous variant in a sample that is a GTEx splicing outlier and has significant ASTS, with read pileups and grey arrows indicating the rare variants. Log normalised transcript counts per allele are plotted per transcript, with the REF:ALT ratios marked for each.

Corresponding author(s): DAG, BBC, TL

Last updated by author(s): May 8, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

- Guppy (v3.2.4) for basecalling
- NanoPlot (v1.33) for collection of summary statistics
- minimap2 (v2.11) for aligning reads to genome/transcriptome
- samtools (v1.9) throughout the analysis
- flair (v1.4) for novel transcript calling, quantification and predicting productivity
- TransDecoder for predicting open reading frame (https://github.com/TransDecoder/TransDecoder/)
- genomeTools (v1.6.1) for correcting open reading frame
- bedtools (v2.27.1) for processing eCLIP peaks
- SUPPA (v2.3) for annotating alternative transcript structure events
- Trans-Proteomic Pipeline (TPP) (http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP) for database search
- PeptideProphet for peptide filtering and scoring
- bcftools (v1.9) for processing vcf files
- HAPCUT2 for vcf phasing
- Ensembl VEP (version 104) for variant annotation

Data analysis	• wiggleplotr for read pile-up plots (https://github.com/kauralasoo/wiggleplotr) • qvalue for pi1 analysis (http://github.com/jdstorey/qvalue) • DEseq2 for differential gene expression • DRIMSeq for differential transcript usage • gffcompare for GTF file comparisons • LORALS for allelic analysis (https://lappalainenlab.github.io/lorals/index.html) • Script repository used in this paper (https://github.com/LappalainenLab/lorals_paper_code) (https://doi.org/10.5281/zenodo.6529254)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw long read data generated as part of this manuscript are available in the GTEx v9 release under dbGAP accession number phs000424.v9 and on AnVIL at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_GTEx_V9_hg38. The GTF file of flair transcripts, along with the transcript-level overall and allelic expression quantifications from GENCODE v26 and flair transcripts are available on the GTEx portal (https://gtexportal.org/home/datasets). The GTEx WGS, Illumina short-read, the allelic analysis, eQTLs and sQTLs and enloc colocalization files which are all part of the GTEx v8 release phs000424.v8. Additionally, we used the transcript and gene counts available on (https://gtexportal.org/home/datasets). The GRCh38 human genome reference and GENCODEv26 processed for GTEx were used in this analysis (https://console.cloud.google.com/storage/browser/gtex-resources). The CHESS and Workman transcript datasets were downloaded from GitHub (https://github.com/chess-genome/chess and https://github.com/nanopore-wgs-consortium/NA12878). ENCODE eCLIP data was downloaded from https://www.encodeproject.org/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In total 90 samples were sequenced. For differential analysis a minimum of five samples per group is recommended, which is what we had.
Data exclusions	A single sample was excluded from this study (sample-id GTEX-T5JC-0011-R10A-SM-2TT23), because it had less than 50,000 reads.
Replication	Assessment of replicability was carried out by sequencing five samples in duplicate and five samples in triplicate. All replicates showed high concordance. Additionally, 10 samples were sequenced on both the GridION and PromethION platforms.
Randomization	Randomization is not applicable in this study.
Blinding	Blinding is not applicable in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: Primary antibodies from mouse for PTBP1 (used in 1:4,000 dilution; # 32-4800 Thermo Fisher Scientific) and rabbit for GAPDH (used in 1:10,000 dilution; #2118 Cell Signaling Technology). Secondary antibodies (LI-COR IRDye; donkey anti-mouse IgG polyclonal antibody (800CW; Size=100 µg) and donkey anti-rabbit IgG polyclonal antibody (680RD; Size=100 µg)

Validation: All antibodies were validated by the manufacturer. In addition, we validated them using western blot.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s): Human cultured fibroblast lines from GTEx donors (donors GTEX-OIZI, GTEX-OXRL, GTEX-PSDG, GTEX-R55E, GTEX-RWS6, GTEX-WFG7, GTEX-QEG5, GTEX-QV44, GTEX-S4Z8, GTEX-S95S, GTEX-WY7C); K562 cell line. Cell lines were created as part of the GTEx project from GTEx donors. Available through the biobank located as www.gtexportal.org

Authentication: House generated cell-lines as part of the GTEx project.

Mycoplasma contamination: All cell lines were tested negative for mycoplasma contamination.

Commonly misidentified lines (See ICLAC register): No

Human research participants

Policy information about studies involving human research participants

Population characteristics: Data from 90 samples from 56 donors from the GTEx project and 4 K562 cell line samples. All the GTEx samples had Illumina TruSeq short-read RNA-seq data and 85 samples (51 donors) had whole genome sequencing data made available by the GTEx Consortium. The median age was 54 (+- 13years) with both male and female human donors included.

Recruitment: N/A

Ethics oversight: All human subjects were deceased. IRB and/or ORSP determination for each recruitment and biospecimen collection site is described on pg. 312 and in Table 1, Carithers et al. (2015). DOI: 10.1089/bio.2015.0032

Note that full information on the approval of the study protocol must also be provided in the manuscript.