Profiling genetically driven alternative splicing across the Indonesian Archipelago

Neke Ibeh[1,2,3,4], Pradiptajati Kusuma[5], Chelzie Crenna Darusallam[5], Safarina Malik[5], Herawati Sudoyo[5], Davis J. McCarthy[2,3,6], Irene Gallego Romero[1,2,4,7,*],

1 School of BioSciences, University of Melbourne, Parkville, VIC 3010, Australia

2 Melbourne Integrative Genomics, University of Melbourne, Parkville, VIC 3010, Australia

3 Bioinformatics and Cellular Genomics, St Vincents Institute of Medical Research, Fitzroy, Victoria 3065, Australia

4 Human Genomics and Evolution, St Vincent's Institute of Medical Research, Fitzroy, Victoria, 3065, Australia

5 Genome Diversity and Disease Laboratory, Mochtar Riady Institute of Nanotechnology, Tangerang 15811, Indonesia

6 School of Mathematics and Statistics, Faculty of Science, University of Melbourne, Parkville, VIC 3010, Australia

7 Centre for Genomics, Evolution and Medicine, Institute of Genomics, University of Tartu, Tartu 51010, Estonia

* irene.gallego@svi.edu.au

Abstract

One of the regulatory mechanisms influencing the functional capacity of genes is alternative splicing (AS). Previous studies exploring the splicing landscape of human tissues have shown that AS has contributed to human biology, especially in disease progression and the immune response. Nonetheless, this phenomenon remains poorly characterised across human populations, and it is unclear how genetic and environmental variation contribute to alternative splicing. Here, we examine a set of 115 Indonesian samples from three traditional island populations spanning the genetic ancestry cline that characterizes Island Southeast Asia. We conduct a global AS analysis between islands to ascertain the degree of functionally significant AS events and their consequences. Using a hierarchical event-based statistical model, we detected over 1,000 significant differential AS events across all comparisons. Additionally, we identify over 6,000 genetic variants associated with changes in splicing (splicing quantitative trait loci; sQTLs), some of which are driven by Papuan-like genetic ancestry, and only show partial overlap with other publicly available sQTL datasets derived from other populations. Computational predictions of RNA binding activity revealed that a fraction of these sQTLs directly modulate the binding propensity of proteins involved in the splicing regulation of immune genes. Overall, these results contribute towards elucidating the role of genetic variation in shaping gene regulation in one of the most diverse regions in the world.

Introduction

Pre-mRNA splicing is a critical and highly regulated process through which multiple mRNA isoforms are produced from a single gene through the excision of introns and ligation of exons [1]. While constitutive splicing yields identically spliced mRNA isoforms, the process of alternative splicing (AS) produces isoforms that differ from each other based

on their unique combinations of exons. AS is one of the regulatory mechanisms influencing the functional capacity of genes, and the resulting alternatively spliced isoforms contribute significantly to the protein diversity and functional complexity observed in eukaryotic organisms [2]. In humans, approximately 95% of multi-exon genes undergo AS [3, 4], and aberrant AS has been implicated in over 15% of human hereditary diseases and cancers [5, 6]. AS has been found to be highly specific, with isoform expression regularly restricted to certain tissues and cell-types [7–11], although the degree to which alternative isoforms are functional remains unclear [12–14]. Nevertheless, isoform-level transcriptome analyses have revealed that splicing can play significant roles in the cellular response to environmental cues, including immune pressures [15–21]. In humans, this natural variation in alternative splicing has been highlighted as a phenomenon influencing complex traits and disease prevalence [22].

Alternative splicing is tightly regulated through an intricate protein-RNA interaction network comprised of *cis* regulatory elements and *trans*-acting factors [23]. Genetic polymorphisms can alter these splicing regulatory elements and splice site usage, thereby influencing gene expression and protein products [24]. Splicing quantitative trait loci (sQTL) analysis has become a leading method for elucidating these genotype-splicing associations. sQTL analyses across multiple tissues have shed light on the contributions of AS to a number of traits, including breast cancer [25], Alzheimer's disease [26], and schizophrenia [27]. The GTEx Consortium has recently characterised sQTLs in over 50 human healthy tissues [10], providing an overview of baseline healthy variation. Such studies are vital for advancing our understanding of gene regulation and disease mechanisms. They provide a bridge between genetic variation and complex traits, contributing to the development of innovative diagnostic tools and therapeutic interventions. The unprecedented scale and granularity of these analyses is, however, frustrated by the fact that the majority of participants involved in these studies are of either European or undocumented ancestry. This lack of diversity limits our collective understanding of variation in mRNA regulation and how these regulatory mechanisms might, in turn, contribute to human phenotypic diversity.

Multiple studies have sought to characterise the extent of variation across human populations in both splicing and the genetic mechanisms that regulate it, with existing data suggesting that both genetic ancestry and environmental variation make substantial contributions to these traits. In lymphoblastoid cell lines (LCLs) from 7 different global populations, population differences in genetic ancestry explained 25% of differences between individuals in genes expressing at least two transcripts [28], but expression differences, rather than splicing ones, accounted for the bulk of this effect. However, LCLs are known to trend towards a homogeneous pattern of gene expression that blunts the effect of inter-individual diversity [29], and thus this number may provide a lower-bound estimate of the true amount of variation. Indeed, the majority of AS events across GTEx tissues have been recently attributed to differences in ancestry between individuals, rather than to other demographic drivers such as age, sex or BMI [30].

The number of studies to map sQTLs is much more limited, and even more so when considering populations other than urban, genetically European cohorts; additionally most studies have used LCLs instead of primary tissue (reviewed in [31]). However, there are notable exceptions. A study of the response to infection in monocytes from individuals of European and African descent living in Belgium showed that genetic ancestry consistently contributed to differences in splicing, and highlighted the role of archaic introgression from Neanderthals [21] in particular. Recently, a separate study examining both expression and splicing in whole blood in East African populations has shown that overall, the genetic architecture of splicing is shared between individuals of African and European ancestry, but nonetheless identified a substantial number of sQTLs in individuals from Tanzania and Ethiopia that have not been reported in European cohorts [32].

Here, we focus on 115 previously described whole blood samples drawn from three traditional Indonesian populations [33, 34] that span the genetic ancestry cline that makes up the region [35]. We explore the prevalence and functional significance of alternative splicing variation across this region of the world, working towards deepening our understanding

of the regulatory mechanisms that shape the landscape of immune gene regulation across the Indonesian archipelago. To elucidate the genetic variants influencing these splicing events, we identified 6,405 *cis*-sQTLs (4,199 unique SNPs) affecting 2,423 genes, and investigate possible biological mechanisms driving these observations. We discuss the implications of these gene regulatory patterns on a global scale, comparing our findings with those derived from European samples. Finally, we characterize the landscape of alternative splicing events across our groups of interest and investigate the functional consequences of the differential isoform usage between them. Thus, our work contributes to the growing catalog of putative regulatory elements that shape and influence alternative splicing in human populations.

Materials and Methods

RNA-seq processing and read alignment

We used a published set of 115 (all male) matched DNA and RNA data from three distinct Indonesian populations [33], available at (EGAS00001003671). The samples span the main west-east axis of genetic diversity in the Indonesian archipelago [35, 36]: (i) the inhabitants of Mentawai (MTW, n = 48), a small barrier island in western Indonesia, are of West Island Southeast Asian-like genetic ancestry; (ii) the Korowai (KOR, n = 19), from western New Guinea Island, are of Papuan-like genetic ancestry; while (iii) the inhabitants of Sumba (SMB, n = 48), a small island east of the Wallace line in central Indonesia, carry an approximately 80/20 admixture of either ancestry. Previous work has shown that the Korowai individuals in the data carry approximately 2% of archaic Denisovan introgression [34]. Both data collection and subsequent analyses were approved by the institutional review board at the Eijkman Institute (EIREC #90 and EIREC #126) and by the University of Melbourne's Human Ethics Sub-Committee (approval 1851639.1). All individuals gave written informed consent for participation in the study. Permission to conduct research in Indonesia was granted by the Indonesian Institute of Sciences and by the Ministry for Research, Technology and Higher Education (RISTEK).

Reads were assessed with FastQC (v.0.11.9) [37] and pre-processed with Trimmomatic v0.36 [38], removing leading and trailing bases with a Phred score below 20 prior to any further analysis.

Differential alternative splicing analysis

For event-based quantification of local splicing variation, we employed SUPPA2 [39]. SUPPA2 generates the percent spliced-in (PSI) values for each splicing event across all samples simultaneously using pseudoalignment-based transcript quantification. First, standard local splicing variations were computed using the SUPPA2 *generateEvents* command, applied to the reference annotation (GRCh38 Ensembl release 110). Then, the sample-wise PSI values for each event were calculated using transcript abundances (TPMs) obtained from Salmon v1.9.0 [40]. These PSI values denote the relative abundance of transcripts containing an exon (or intron, in RI cases) over the relative abundance of transcripts for the gene of interest containing the exon/intron. PSI scores were computed for five well-defined classes of alternative splicing events, namely, skipped exons (SE), retained introns (RI), mutually exclusive exons (MXE), alternative 5' splice sites (A5SS), and alternative 3' splice sites (A3SS). To ensure high confidence event calling, events were restricted to protein-coding and lincRNA genes, the percentage of exons with missing PSI values had to be below 5% per sample, and average TPM within a population had to be > 1).

To identify differential alternative splicing between the three groups, we corrected PSI values for batch effects and other technical confounders using fractional regression [41, 42]. Each splicing event was fit with logit-transformed PSI values (*glm* function in the R stats package (R v4.3.3), setting family = quasibinomial('logit')). We selected the following covariates as the most relevant to incorporate into the model: RIN, sequencing batch, age of donor, and blood cell type

proportions from previously computed [33] estimates of the proportion of CD8T, CD4T, NK, B cells, monocytes, and granulocytes in each sample. Using these corrected PSI values, differentially spliced events were identified by computing ΔPSI values between each pair of sample groups . Only splicing events that were detected in all groups were retained for differential testing (i.e., no NAs), as recommended in the SUPPA2 manual. Differential events were deemed statistically significant if they had a $|\Delta \text{PSI}| \geq 0.1$ and FDR < 0.05.

In parallel, we conducted intron-based quantification of splicing variation (necessary for sQTL mapping) using LeafCutter [43], which identifies intron excision events across samples. Intron clustering (representing alternative intron excision events) was performed using default settings of 50 reads per cluster and a maximum intron length of 500kb. For each intron cluster, the proportion of reads supporting a specific intron excision event was calculated. Intron excision ratios were then standardized across individuals and quantile-normalized. LeafViz [43] was used for the annotation and visualization of intron clusters and splicing events.

Comparative visualization of RNA-seq read alignments

We summarized exon- and junction-spanning RNA-seq densities for all sample groups using MISO's [44] *sashimi_plot* tool. Read densities for all exons were quantified using RPKM units [45]. Junction reads were plotted as arcs spanning the exons that the junction borders. Isoform structure was obtained from the GFF annotation (GRCh38 Ensembl release 110) of each splicing event. For ease of visualization, intron lengths were scaled down by factor of 30, and exon lengths were scaled down by a factor of 4.

Isoform switching analyses

We identified instances of isoform switching by testing each individual isoform for differential usage across the three populations. Changes in isoform usage for a gene are represented by the difference in isoform fraction values (dIF), where IF = isoform expression/gene expression. Pseudoaligned transcript abundance and genomic coordinates were aggregated with the R package *IsoformSwitchAnalyzeR* [46]. Prior to isoform switch testing, we filtered out completely unused isoforms, single-isoform genes, as well as genes with an average TPM level (in each population) < 1. Confounding effects (RIN, sequencing batch, age, and blood cell type proportions, as above) were accounted for by applying the limma [47] *removeBatchEffect* function to the isoform abundance matrix. Statistical identification of isoform switches was conducted via *DEXseq* using default settings [48]. Isoforms were considered differentially switched and retained for further analysis if the difference in isoform fraction (dIF) > 0.1 and Benjamini-Hochberg-adjusted FDR < 0.05. We consider the first population in the switch comparison as the ground state, and the second population as the changed stated. For example, an up-regulated isoform in KOR versus MTW is one that is used more in MTW when compared to KOR. We then translated the coding sequences of the switching isoforms into amino acids, and predicted their coding capabilities, protein structure, peptide signaling, and presence of protein domain families using CPC2 [49], IUPred2A [50], SignalP [51], and Pfam [52], respectively.

The functional consequences of each significant isoform switch were evaluated by analyzing the annotation differences between the isoform(s) used more (switching up, dIF > 0) and the isoform(s) used less (switching down, dIF < 0). In other words, if a gene has at least one isoform with a significant change in usage between the three populations, the functional annotation of this isoform is compared to that of the isoform(s) with the compensatory change in usage. For genes with multiple significant switching events, we compared the functional annotation of all pairwise combinations of the isoforms involved. The following functional properties were compared: isoform coding potential, open reading frame (ORF) sequence similarity, the presence or absence of protein domain families, the presence or absence of signal

peptides, the presence of intrinsically disordered regions (IDRs; regions that lack a fixed or ordered structure), and nonsense-mediated decay (NMD) sensitivity. For sequence similarity comparisons, we used the default minimum length difference cut-off of 10 amino acids [46]. Changes in protein domain or IDR length are only reported if the shorter protein domain (or IDR) is $< 50\%$ of the length of the longer region.

cis-sQTL mapping

RNA-seq reads were aligned to GRCh38 Ensembl release 110 using the default settings of STAR (v.2.7) [53] with the exception of -alignEndsType EndToEnd to remove soft-clipping of the reads and -sdjbOverhang 100 for optimal splice junction overhang length. Mapped reads were used as input to LeafCutter [43] to obtain standardized and normalized intron excision ratios (the number of reads defining an excised intron over the total number of intron cluster reads), which were then used as phenotypes for sQTL mapping. We used QTLTools [54] to test for an association between variants and intron ratios within a *cis*-region of ± 1Mb of the intron cluster, using 10,000 adaptive permutations. After observing that the vast majority of our detected associations occurred within 250kb of the intron clusters, we restricted our analyses to this 250kb window. We controlled latent sources of variation using covariates identified with the Probabilistic Estimation of Expression Residuals (PEER) method [55]. The number of PEER factors was determined as a function of sample size, with the first 29 PEER factors (25% of our sample size) selected. The top 5 genotyping principal components were included to account for population structure, since they account for approximately 9% of the genotype variance observed, with diminishing returns for all subsequent PCs. Nominal p-values for each variant-phenotype pair were obtained by testing the alternative hypothesis that the slope of the linear regression model between genotype and excision ratios deviates from 0.

Identifying shared *cis*-sQTLs and differences in effect sizes

We compared our Indonesian sQTLs to two European whole blood sQTL studies: GTEx v8 [10] (n = 670) and BLUEPRINT [9] (n = 197) with freely available complete summary statistics. For both European studies, the sQTL mapping approaches that were employed are in line with those used for our Indonesian data here, making comparisons possible without extensive raw data reprocessing. Specifically, both studies applied a linear modeling approach, testing for associations with variants within ± 1Mb of each gene's transcription start site. For consistency, GTEX and BLUEPRINT sQTL sets were restricted to 250kb, and coordinates for the BLUEPRINT data were converted from hg19 to hg38 using the R package *liftOver* v1.26.0 [56]. Shared and population-specific sQTLs across Indonesian and European populations were assessed using the multivariate adaptive shrinkage (mash) approach implemented in the R package *mashr* [57]. The input for mash consisted of sQTL effect sizes and their standard errors (obtained from the QTLTools output). The correlation structure among the null tests was estimated using a large, random subset of all tests (40%). For each intron cluster within each gene, the SNP with the smallest p-value across all tested SNPs was retained in order to produce a confident set of sQTLs. The data-driven covariance matrix was constructed using this strong set, and posterior mean effect sizes were calculated by applying the mash model that was built using the random set. sQTLs were considered shared across populations if the magnitude of effect sizes was within a factor of 0.5 between groups, and the sign of the effect was the same. A local false sign rate (LFSR) < 0.05 was used as a threshold for significance.

Variance in sQTL genotype explained by local genetic ancestry

Using all significant permutation-based sQTLs, we quantified the variance in sQTL genotype explained by modern local genetic ancestry (LA). To do this, we adapted a previously described approach [58], fitting a linear model $V = \alpha \times \text{PAP}$

$+\ \beta$ for each variant. V is the genotype vector and PAP is a covariate which represents the number of alleles assigned to Papuan-like genetic ancestry. This analysis was conducted using the 73 samples (30 Mentawai, 29 Sumba, 14 Korowai) with available 30x depth whole-genome sequencing data [34]. Variants were categorized as highly correlated with LA if they had an absolute $R^2 > 0.7$.

Gene ontology overrepresentation analysis

We tested for overrepresentation of GO and KEGG terms using the R package *clusterProfiler* (v.4.0.5) [59], setting the gene universe as all tested genes. We used an FDR threshold of 0.05 to identify significantly enriched terms. For GO term representations across sGenes, we employed REVIGO [60] (parameters: allowed similarity = 0.9, database =*H.sapiens*, semantic metric = SimRel) to remove highly redundant GO terms from *clusterProfiler* output and visualize semantic similarity-based GO term representations.

Estimation of sQTL variant pathogenicity

We intersected our sQTL SNPs with clinically annotated variants in ClinVar (`https://www.ncbi.nlm.nih.gov/clinvar`) in order to assess their possible pathogenicity. We considered two sets of variants for analysis—lead SNPs located within the gene body, and all sQTL SNPs that were located directly within a splice junction.

Evaluating the impact of sequence variations on the binding affinity of splicing RNA binding proteins

We used DeepCLIP to quantify the effects of SNPs on protein-RNA binding [61]. Specifically, we used DeepCLIPs pre-trained models for 33 splicing-associated RNA binding proteins (RBPs). These RBPs were chosen because they have been previously characterized in eCLIP (enhanced Cross-Linking and ImmunoPrecipitation) studies from HepG2 and K562 cell lines and have well-documented roles in splicing regulation and spliceosome activity [62, 63]. Using the pre-trained models, we ran prediction in paired sequence mode using all significant sQTL SNPs, with 10 bp flanking sequences on both sides. The output from DeepCLIP is a set of binding profiles and overall scores for the reference and variant sequences. The binding profile scores range from 0 to 1, and indicate whether regions of a sequence contain potential binding sites (1) or if they are most likely to be random genetic background (0) [61].

Colocalization analysis with GWAS hematological traits

Colocalization analyses were performed between sQTLs and 13 hematological traits using coloc v5.2.3 [64]. We tested for colocalization between our detected permutation-based sQTLS and GWAS loci from a trans-ethnic study that included East Asian populations [65]. We assumed a prior probability that a SNP is associated with i) the GWAS trait (p1, default $= 1 \times 10^{-4}$), ii) alternative splicing (p2, default $= 1 \times 10^{-4}$), and iii) both the GWAS trait and alternative splicing (p12, default $= 1 \times 10^{-6}$). We identified robust colocalization with the default threshold of CCV > 0.8 and a ratio CCV/DCV > 5. The 13 hematological traits measured in the East Asian populations were: basophil count, eosinophil count, hematocrit, hemoglobin concentration, lymphocyte count, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration, mean corpuscular volume, monocyte count, myeloid white blood cell count, neutrophil count, platelet count, and RBC density. All GWAS summary statistics were obtained from the GWAS catalog. Since we did not have access to linkage disequilibrium values for the Indonesian samples and our sample size is too small to build a robust LD matrix [66], the colocalization analyses performed here did not allow for the identification of multiple causal variants.

Results

Characterizing differential alternative splicing events between Indonesian populations

To characterize the splicing landscape across the Indonesian archipelago, we carried out alternative splicing analyses using paired-end RNA-seq data generated from 115 previously described male whole blood samples [33]. Samples were obtained from three Indonesian populations: the people of Mentawai (of genetically West Island Southeast Asian-like ancestry), the people of Sumba (approximately 80/20% admixed between West Island Southeast Asian-like and Papuan-like genetic ancestries), and the Korowai of New Guinea Island (of Papuan-like genetic ancestry) (Figure 1A).

Across the three populations, we observed 38,611 alternative splicing events in the Korowai, while the Mentawai and Sumba groups had 39,728 and 39,769 alternative splicing events, respectively. This slight difference in total detected events can be attributed to the smaller sample size of the Korowai group (n = 19). Our results show that the splicing events across these three populations were generally shared. Specifically, 38,144 (94.2%) of all alternative splicing events were seen in all three island groups, 1,337 events (3.3%) were shared between only two of three groups, and 1,002 events (2.5%) were found in just one of the three groups (Figure 1B). We additionally observe that the Mentawai and Sumba groups make up the majority of the pairwise AS event sharing (74.1% of all events shared by any two island groups, Supplementary Figure 1B-F), although this is likely attributable to our increased power in these two groups relative to the Korowai. To gain further insight into the genes that contain these 1,002 population-specific alternative splicing patterns (434 events in MTW, 447 events in SMB, and 121 events in KOR), we tested for their enrichment against GO and KEGG pathways (Table S2). Overlapping enriched GO categories and KEGG pathways for the Korowai population are related to nervous system function, transmembrane transportation, and substance dependence. In Mentawai, significantly enriched GO categories are generally related to regulation of synaptic signaling, however, no enriched KEGG pathways were detected. For Sumba, enriched GO categories and KEGG pathways include renal system processes, hormone transport, and calcium signaling.

To better understand the splicing variation across these populations, we conducted a differential alternative splicing analysis. We identified 419 significant ($|\Delta PSI| \geq 0.1$ and FDR < 0.05) differential alternative splicing events between Mentawai and Sumba, 515 events between Korowai and Mentawai, and 738 events between Korowai and Sumba (Figure 1C and D). These differences are in line with previously reported gene expression patterns across these groups, with the Korowai vs. Sumba comparison yielding the most differentially expressed genes, closely followed by Korowai vs. Mentawai [33]. These findings suggest that the Korowai population might be driving a lot of the observed inter-island variability, with the Sumba and Mentawai populations exhibiting higher levels of shared homogeneity, although we cannot confidently disambiguate whether these differences are caused by genetic or environmental factors. Differentially alternatively spliced (DAS) genes between Korowai and Sumba are related to mRNA processing (e.g., *SLTM*, *BARD1*, *DYRK1A*, *CDK11A*, *QKI*, *RNPS1*), splicing regulation (e.g., *RBM23*, *TRA2B*, *TMBIM6*, *MBNL1*, *HNRNPH1*, *RNPS1*, *PTBP2*), and cell cycle signaling (e.g., *MRE11*, *BRIP1*, *DONSON*, *ZWILCH*, *CDC14B*), indicating potential population-based differences in post-transcriptional regulatory patterns (Table S3). While DAS genes between Korowai and Mentawai were not enriched for either GO or KEGG terms, we found that DAS genes between Mentawai and Sumba were enriched for transcription corepressor (e.g., *DPF2*, *ATF7IP*, *MTA1*, *HDAC9*, *SF1*) and coactivator (e.g., *MED24*, *KMT2C*, *SMARCA2*, *MED12*, *ACTN1*) activity (Table S3). Overall, these findings indicate that the DAS genes between these populations are predominantly impacting aspects of cell division and regulation of gene expression.
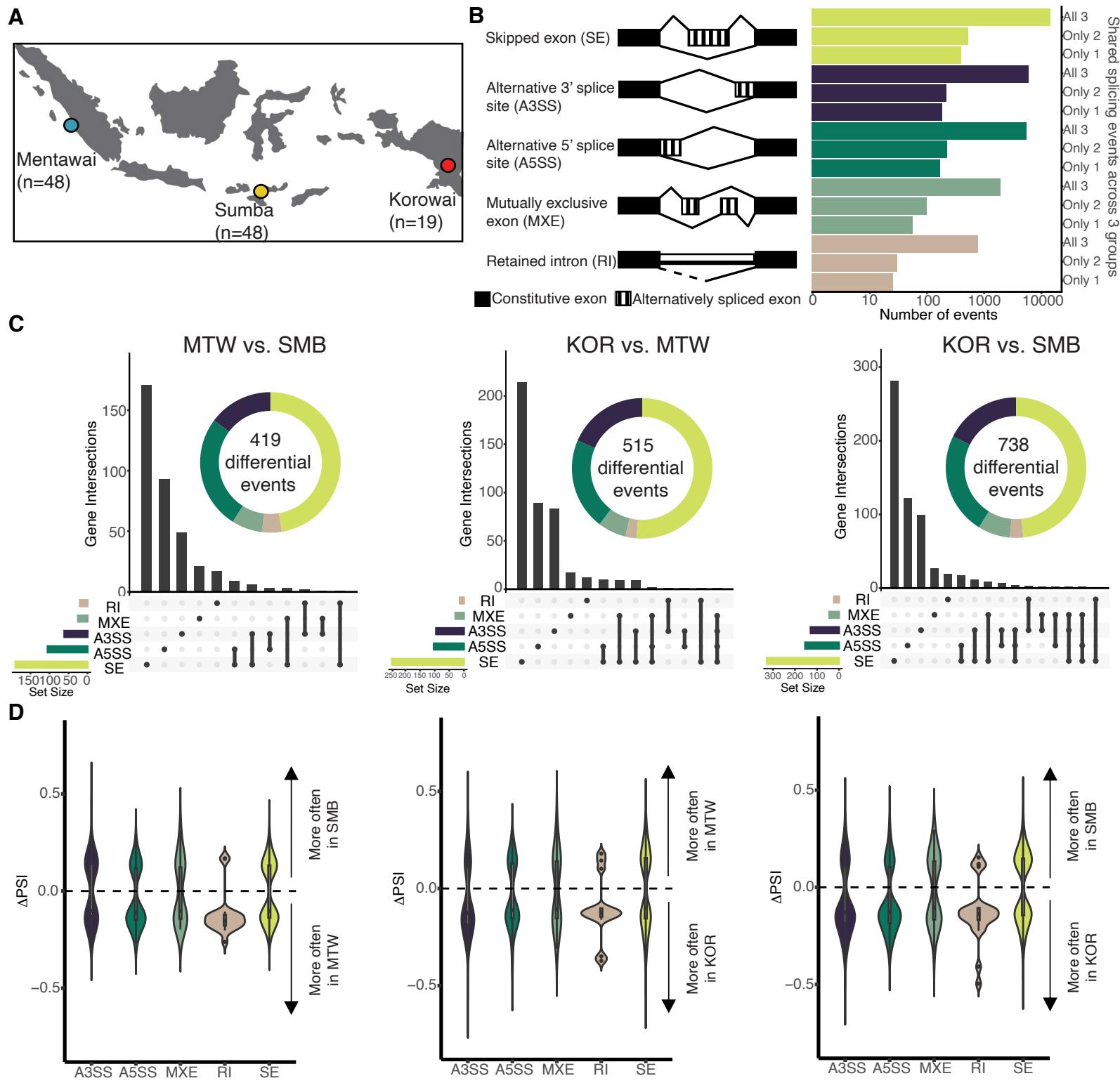
Figure 1. Characterization of differential alternative splicing across Indonesian populations. A) Geographical location of the three sampling sites (n = 115). B) Schematic diagram of the five alternative splicing events profiled with SUPPA2. Horizontal bar plots correspond to the event categories to the left, and illustrate the sharing of alternative splicing events across the three island groups (i.e., the splicing event either occurs in all three groups, or in only two of the three groups, or in only one group). The x-axis is log_{10} transformed. C) UpSet plots summarize the DAS gene intersections for the three pairwise comparisons. Only significant events ($|\Delta PSI| \geq 0.1$, FDR < 0.05) are shown. D) For each population pair, the distributions of ΔPSI values are plotted for each splicing category (A3SS, A5SS, MXE, RI, and SE). The direction of the arrows indicate whether the alternatively spliced exon is excluded more or included more in each sample (i.e. exon exclusion levels). In the case of an MXE event, if the strand is positive, then the inclusion form includes the first exon and excludes the second; if the strand is negative, then the inclusion form includes the second exon and skips the first.

Characterisation of Indonesian *cis*-sQTLs in whole blood

To understand the genetic regulation of alternative splicing across our Indonesian populations, we performed a *cis*-sQTL mapping analysis using all 115 samples. sQTLs were mapped using linear regression to quantify the association between the intron excision ratios and SNP genotypes (Materials and Methods). Splice ratios from a total of 49,122 splice clusters (i.e., overlapping introns that share a spice donor or acceptor site) were included in the sQTL analysis. We detected a total of 6,077 *cis*-sQTLs affecting 1,977 genes at an FDR level of 0.01, comprising 3,658 unique SNPs (Table S4 and Table S5). Positional enrichment of significant sQTL SNPs showed clustering closest to the detected splice junctions, with 75.2% of SNPs occurring within 50kb of the nearest junction (Figure 2A). This observation is consistent with previous research suggesting that most variants that fall in close proximity to splice junctions influence splicing regulatory functions [67,68]

GO enrichment analysis of sGenes (genes with a significant sQTL) revealed enrichment for numerous immune-related pathways, including: immune response-regulating signaling (e.g., *THEMIS2*, *ERMAP*, *GBP2*, *VAV3*, *FCRL3*, *AIM2*, *FCGR3A*), immune response-activating signaling (e.g., *DENND1B*, *MAPKAPK2*, *CR1*, *NLRC4*, *NAGK*, *IFIH1*, *PRKCD*), phagocytosis (e.g., *VAV3*, *NCF2*, *C4BPA*, *DYSF*, *MARCO*, *CD302*), regulation of innate immune response (e.g., *GBP2*, *ADAR*, *FCRL3*, *IFI16*, *AIM2*, *CFH*), and interleukin-1 beta production (e.g., *IFI16*, *AIM2*, *NLRC4*, *CASP8*, *GHRL*, *CX3CR1*), highlighting the impact of splicing on immune-related genes (Supplementary Figure 3, Table S6). Indeed, some of the strongest sQTL signals are present in immune-related genes such as *TRIM58*, which is additionally implicated in malignancies such as lung cancer, liver cancer, and pancreatic ductal adenocarcinoma [69–72], and is involved in innate immunity and cell proliferation [69,73], and *ERMAP*, which is a B7 family immune regulator that has been found to promote the phagocytosis of tumor cells [74–76] and encodes the protein responsible for the Scianna blood group system [77] (Supplementary Figure 4).

Given that the majority of sQTL studies to date have profiled European populations, we wanted to compare the overlap between our detected sQTLs and previously published datasets. We focused our comparison on two large studies of whole blood, namely: GTEx v8 [10] (n = 670) and BLUEPRINT [9] (n = 197). We found that sQTL effects are largely shared between the two European data sets, but they are not shared to the same extent with the Indonesian data. Out of 4,550 unique sGenes (amongst the three datasets), 330 (7.3%) are shared across all datasets, 364 are detected in both European cohorts but not in the Indonesian data (8%), and 1008 (22.2%) are solely in the Indonesian data; similar values were private to both GTEx and BLUEPRINT, suggesting either incomplete power to detect sQTLs across studies, winner's curse, or a genuine poor generalizabilty of sQTLs across studies (Figure 2B). To better estimate shared and population-specific effects and ascertain the degree of sharing between our detected Indonesian sQTLs and the European sQTLs, we applied a multivariate adaptive shrinkage model, mashr. Using a local false sign rate cut-off of 0.05, an sQTL is considered shared if the effect sizes share the same sign and fall within a factor of 0.5 of each other. As with the more naïve approach above, sQTL effects are largely shared across European datasets (Figure 2C). Additionally, we detected 312 Indonesia-specific (LFSR < 0.05) SNP-sGene pairs (Figure 2D). In line with previous eQTL analyses carried out on this set of samples [34], we observe significantly larger effect sizes of Indonesian-specific sQTLs within Indonesia than we do in the European datasets (ANOVA, $p < 2.2e-16$; Figure 2D), again suggestive of incomplete power across all considered datasets, and highlighting the importance of establishing larger cohorts for QTL mapping in general.

While there was no evidence of enrichment of GO or KEGG terms across these Indonesia-specific genes, many genes with sQTLs are implicated in signal regulation (i.e. *IRF1*, *SIRPB1*) and oncogene pathways (i.e. *RAP1A*, *RAP1B*, *FES*, and *DNM3*). In spite of our limited sample size, these findings suggest a distinctive and distinguishable (relative to previously characterized European data) Indonesian sQTL signal, as we would expect sharing between GTEx and BLUEPRINT to also be negatively impacted by the relatively low sample size in BLUEPRINT and their distinct processing pipelines. By profiling the differences and similarities across these groups, we can more accurately characterize the gene
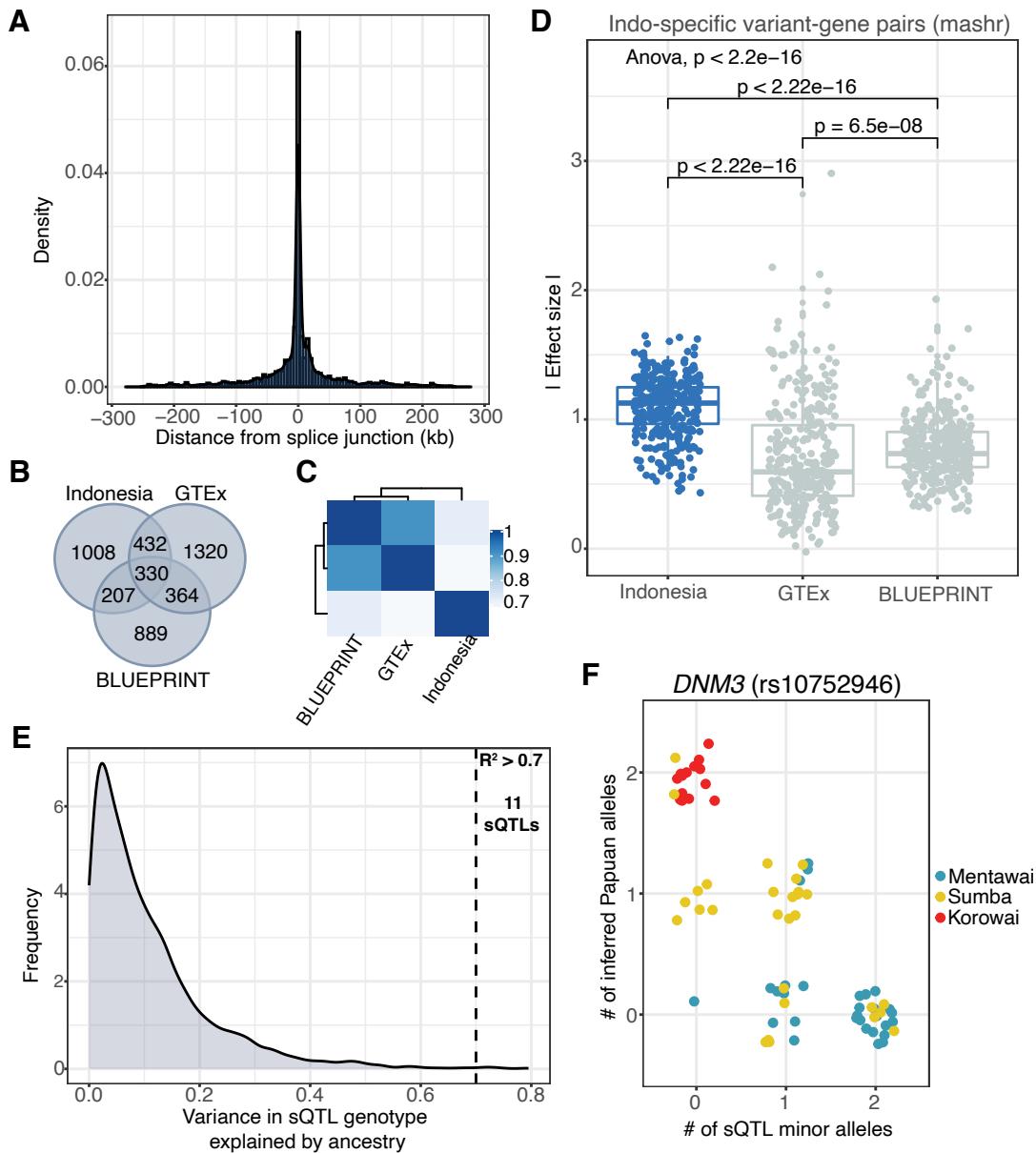
Figure 2. Indonesian *cis*-sQTL characteristics. **A**) Relative distance of sQTLs to the nearest splice junction. **B**) Shared and unique sGenes across the Indonesian and European datasets. **C**) Pairwise sharing of sQTLs between the datasets evaluated by mashR. **D**) Distribution of effect sizes for Indonesia-specific sQTLs. **E**) Linear regression between the number of QTL minor alleles and number of inferred Papuan alleles reveals 11 sQTLs largely driven by Papuan-like genetic ancestry ($R^2 > 0.7$). **F**) An sQTL that is strongly correlated with Papuan-like genetic ancestry. Plot shows the relationship between the number of inferred Papuan alleles and sQTL minor alleles.

regulatory mechanisms underlying diversity in genetic traits, highlighting the value of trans-ethnic, trans-environment QTL mapping to maximize discovery of gene regulatory variation across humans [30, 32].

After quantifying sQTL-level differences between the Indonesian and European cohorts, we wanted to assess the extent to which variance in sQTL genotype is driven by local genetic ancestry in modern Indonesians. Specifically, using local genetic ancestry haplotype information for each significant sQTL, we investigated whether there was a correlation between the inferred ancestral source of the genotype and gene splicing (Materials and Methods). We found 11 sQTLs that show strong evidence ($R^2 > 0.7$) of being driven by Papuan-like genetic ancestry (Figure 2E, Table S7). Although small and likely underpowered due to the very limited representation of Papuan-like genetic ancestry in our dataset, this set included several immune genes, one of which was *DNM3* (Figure 2F). Otherwise known as Dynamin-3, *DNM3* is a microtubule-associated gene that plays a fundamental role in membrane trafficking and the regulation of vesicle formation in cells [78], and has been implicated in various functions related to the development and colony formation of megakaryocytes [79–81]. Human *DNM3* contains 21 exons, with at least 5 of its exons being alternatively spliced to generate several isoforms [82]. The Papuan-driven sQTL event that we identified for *DNM3* is a skipping event of exon 17 ($\beta = 0.55$, q-value $= 3.5 \times 10^{-4}$), associated with the SNP rs10752946, which lies within intron 20 of *DNM3* (Figure 2F, Figure 3A). The canonical isoform for this gene, ENST00000627582.3, includes exon 17 and encodes a protein made up of 863 amino acids (GRCh38 Ensembl release 110). Conversely, isoform ENST00000367731.5 is the product of the alternative splicing event that excludes exon 17, resulting in a loss of 4 amino acids which do not appear to have a significant impact on protein structure predictions made with AlphaFold. Our results indicate that the Korowai samples have a higher rate of exon 17 exclusion/skipping (22.5%) than the Mentawai (7.6%) and Sumba (7.5%) groups; in other words, a higher proportion of expressed *DNM3* isoforms in the Korowai group exclude exon 17.

Indonesian whole blood *cis*-sQTLs colocalize with hematological GWAS traits

To shed light on the potential interplay between gene splicing regulation and complex traits, we sought to explore whether the Indonesian sQTLs colocalized with any hematological GWAS traits. Using coloc, a Bayesian test for genetic colocalization, we performed colocalization analyses between all significant sQTLs and 13 blood-related traits using summary statistics from a trans-ethnic study which included 151,807 East Asian participants [65]. Overall, we identified 45 unique sGenes (3.4 % of all significant sGenes) that colocalized with 12 different traits, from a total of 68 significant GWAS trait-sGene pairs (File S1). Seven of these genes colocalized with three traits simultaneously. Of these, *FCGRT*, which encodes the neonatal Fc receptor (FcRn) colocalized with red blood cell density, hemoglobin concentration, and hematocrit, while *MAEA*, a macrophage erythroblast attacher, colocalised with mean corpuscular hemoglobin, mean corpuscular volume, and red blood cell density. We also found that the local genetic ancestry-driven sQTL at *DNM3* significantly colocalizes with hemoglobin concentration (H4 = 0.837, H4/H3 = 8.445), a trait which refers to the amount of hemoglobin protein present in a specific volume of blood (Figure 3B).

Assessing the pathogenicity of *cis*-sQTL variants

Given the growing awareness of the role of splicing in genetic disease and the ancestral diversity biases in existing sQTL datasets, we intersected sQTL SNPs with clinically annotated variants on ClinVar. We considered two sets of SNPs: i) all lead sQTL SNPs within the gene body and ii) all sQTL SNPs falling within splice junctions. Across both SNP sets, the analysis yielded 120 total hits, with 116 of these having benign/likely benign status, 3 variants of uncertain significance (VUS), and 1 variant with a GWAS disease trait classification (Table S8). The 3 VUS occur in *KLRC4* (rs139613925, in-sample alternative allele frequency 0.481), *HADHB* (rs2303893, in-sample alternative allele frequency
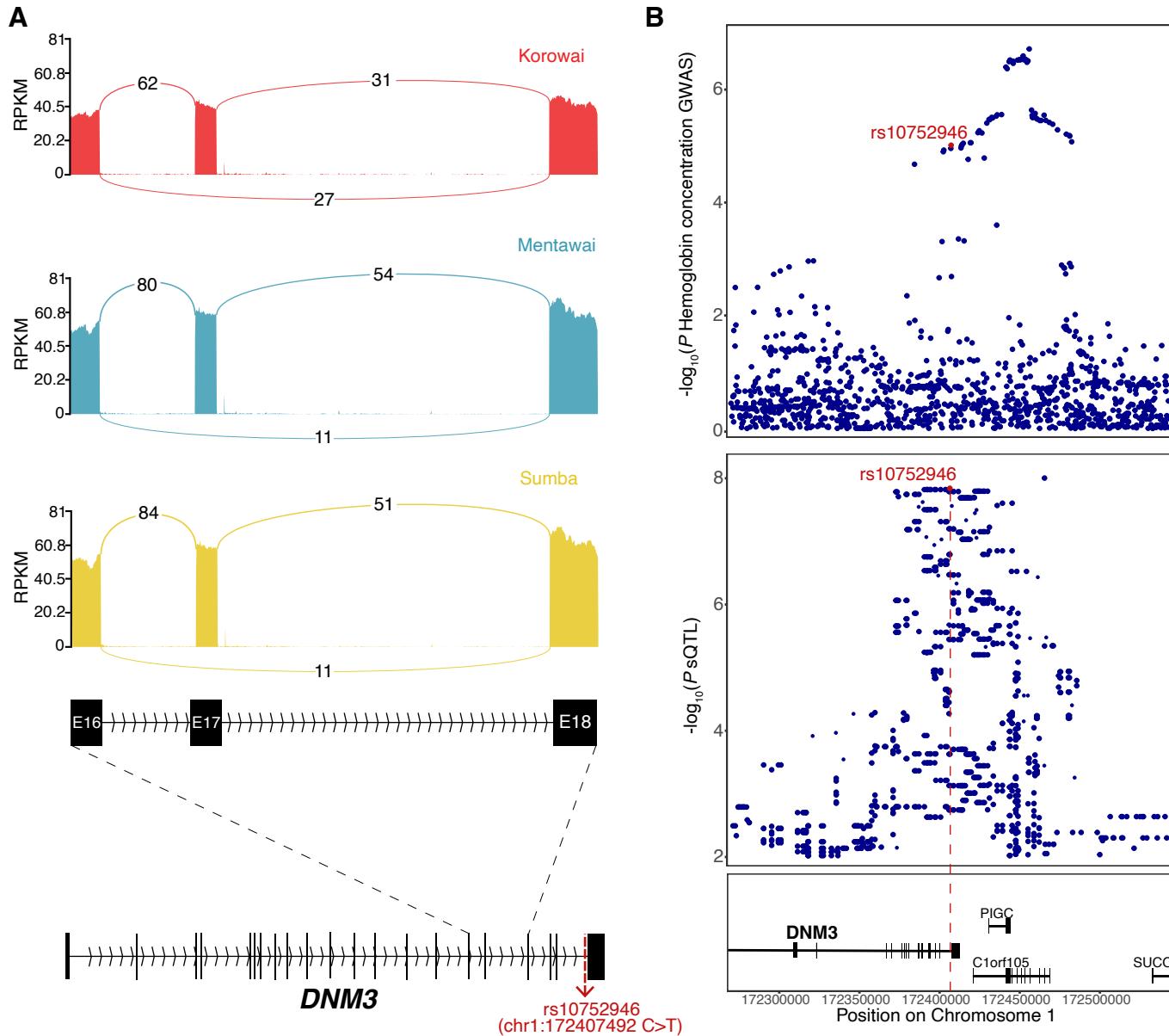
Figure 3. Splicing regulation in the *DNM3* gene. A) Sashimi plot illustrating the skipped exon event (exon 17) in the *DNM3* gene. Isoform expression was summarized using the MISO framework [44]. The numbers over each curved line represent the junction spanning reads. The canonical isoform for *DNM3* is provided. The labels E16, E17, and E18 represent exons 16, 17, and 18, respectively. **B**) Statistically significant GWAS colocalization between the LAI-driven sQTL in the *DNM3* gene (rs10752946) and the GWAS hemoglobin concentration trait [65] measured in East Asian populations.

0.366), and *DENND3* (rs2289001, in-sample alternative allele frequency 0.307). The protein encoded by *KLRC4* belongs to the Natural Killer Group 2 (NKG2) family of receptors. These receptors are primarily expressed on natural killer (NK) cells and some subsets of T cells, where they play important roles in the recognition and regulation of immune responses [83]. The *HADHB* gene product, MTP-beta, plays a central role in the breakdown of long-chain fatty acids within mitochondria [84]. *DENND3* plays a role in regulating intracellular vesicle trafficking, and has been implicated in cancer, where dysregulation of vesicle trafficking pathways can contribute to tumor progression and metastasis [85, 86]. The GWAS variant, rs2763979 (in-sample alternative allele frequency 0.468), occurs in *HSPA1B* and has reported associations with cancer (*p*-value = 2×10^{-14}, European population, n = 475,312) [87] and systemic lupus erythematosus (*p*-value = 6×10^{-6}, Thai population, n = 4,088) [88].

cis-sQTL SNPs influence splicing-RBP binding affinity

We used a deep-learning convolutional neural network, DeepCLIP, to predict the effect of our sQTLs on RBP-RNA binding as a possible mechanism of action. We employed DeepCLIPs pre-trained models for 33 splicing-associated RNA binding proteins (RBPs) with well-documented roles in splicing regulation and spliceosome activity [62], and assessed the impact on their binding abilities across all 3,658 significant lead SNPs within 250kb of the nearest splice site, adding 10bp flanking sequences on both sides. For each reference-alternative allele pair and each RBP, we obtained binding profiles indicating the likelihood of RBP binding (for the respective protein) along the RNA sequence. These binding affinity scores, ranging from 0 to 1, can be interpreted as the strength of RBP binding [61]. Globally, we observe that most sQTL SNPs do not significantly increase or decrease RNA binding affinity, and their difference scores (reference score - alternative score) tend to fall within the range (99% CI) of likely values (Figure 4A, Supplementary Figure 5). However, for RBPs with very uniform motifs, such as EWSR1 (Ewing sarcoma protein) [89], the introduction of sQTL SNPs has a marked effect on the sequence binding scores, yielding large differences in predicted binding affinity between the reference and the alternative allele (Supplementary Figure 5).

Conversely, we also observe RBPs with only one or two outliers, suggesting a biologically significant impact of the sQTL SNP on RBP activity (Figure 4A and B). Specifically, for the HNRNPM protein, which is moderately expressed within our dataset (Figure 4C), we predict a 0.71 increase in binding affinity at rs6433311 (A>G, Figure 4B), which is an sQTL for the *DYNC1I2* gene; in our data, the G allele at rs6433311 is associated with an effect size of -1.174 relative to the A allele (indicating exon exclusion, Supplementary Figure 6A), and segregating at a frequency of 0.739. The predicted increase in HNRNPM binding as a consequence of this A>G intronic variant is consistent with previous findings that HNRNPM recognizes GU-rich cis-elements, primarily in introns [90–92]. HNRNPM is a member of the heterogeneous nuclear ribonucleoprotein (hnRNP) family involved in RNA processing, including pre-mRNA splicing and mRNA transport [93], and has been found to promote exon skipping across various experimental conditions [92, 94–96]. The *DYNC1I2* gene, as a component of the dynactin complex [97], may indirectly influence RNA transport within the cell, potentially interacting with HNRNPM-associated mRNA complexes during transport. In addition, we predict a 0.73 decrease in the binding affinity of *PTBP1* at rs27291 (G>A, Figure 4B), which is an sQTL for the *LNPEP* gene; in this case, the A allele at rs27291 in our data is associated with an effect size of 0.672 (indicating exon inclusion) and segregating at an allele frequency of 0.434. PTBP1, also known as Polypyrimidine Tract Binding Protein 1, plays crucial roles in alternative splicing, mRNA stability, mRNA localization, and translation initiation [98–101]. Containing four highly conserved RNA binding domains that recognize short pyrimidine-rich sequences [102], this splicing factor regulates alternative splicing by inducing exon skipping [103–105]. Thus, a decrease in the binding activity of this protein may result in higher exon inclusion, which is in line with the positive effect size that we observe at this splicing junction (Supplementary Figure 6B).
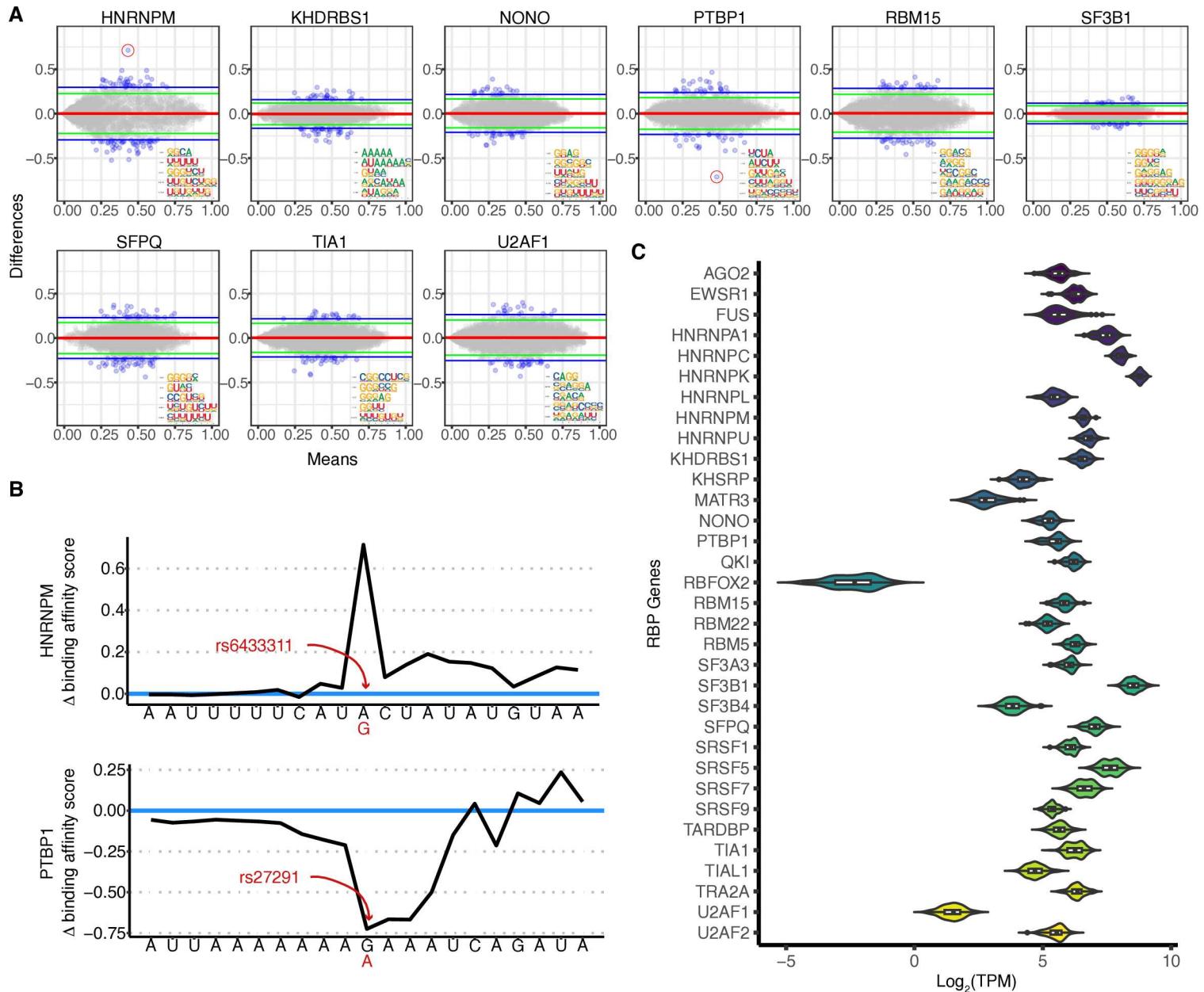
Figure 4. sQTL impact on RBP binding affinity. A) Mean-difference plots for 9 of the 33 splicing-RBPs. Each point represents a pair of 21bp-long sequences centered on the lead sQTL SNP (ref/alt alleles). The y-axis depicts the difference between the DeepCLIP reference sequence score and the DeepCLIP variant sequence score, while the x-axis depicts the average between these two scores. Horizontal green and blue lines represent the 95% and 99% CIs, respectively. Points that are outside of the 99% CI are colored blue. CNN filters predicted by DeepCLIP, ranked by score, are provided for each protein. **B)** DeepCLIP binding profiles for HNRNPM and PTBP1. The variant-reference sequence pairs in question are circled in panel A. **C)** Normalized expression (TPM) of the 33 chosen splicing-associated RNA binding proteins.

Genome-wide isoform switching analysis reveals significant changes in functional protein domains as a consequence of fluctuations in isoform usage

In order to better assess the functional implications of our detected splicing events, we conducted an isoform switching analysis across each of the three pairwise group comparisons (Mentawai vs. Sumba, Korowai vs. Sumba, and Korowai vs. Mentawai). Isoform switching is defined as the change in relative abundance of different isoforms from the same gene between conditions. Isoforms were considered differentially switched and retained for further analysis if the difference in isoform fraction (ΔIF) > 0.1 and FDR < 0.05. In total, we identified 165 significant isoform switches across 110 genes between Korowai and Mentawai, 244 switches across 158 genes between Korowai and Sumba, and 100 switches across 66 genes between Mentawai and Sumba(Table S9 and Table S10). Altogether, these events impacted 239 unique genes and 389 unique isoforms. Then, to ascertain the potential biological impact of these isoform switches, we tested for changes in the functional properties of the transcripts at the sequence level. These properties included the coding potential of the transcripts, the functional domains and signal peptides (that are either present or absent), the presence of intrinsically disordered regions, or IDRs (regions that lack a fixed or ordered structure), the sensitivity of the transcript to nonsense-mediated decay (NMD), and the similarity of the open reading frame (ORF) sequences between transcripts. For every detected isoform switch in a given gene, the annotations for the most-used isoform (ΔIF > 0.1) and the least-used isoform (ΔIF < 0.1) were compared. Differences in the annotations between the up-regulated isoform and the down-regulated isoform were categorized as isoform switches with predicted functional consequences (Figure 5A).

Across all switching events between the three Indonesian populations, the most frequent functional consequences of isoform switching were changes in ORF sequence similarity (36.7%), protein domains identified (23%), and coding potential (21.4%, File S2-File S5). When focusing on switches between Korowai and either Mentawai or Sumba, isoforms expressed at higher levels in the Mentawai or Sumba groups were more likely to contain an additional domain, with 14.8% of all switching events resulting in a protein domain gain in the up-regulated isoform, while 7.1% resulted in a domain loss in the up-regulated isoform (Figure 5A). Furthermore, for up-regulated isoforms in the Sumba population (KOR vs. SMB and MTW vs. SMB), a greater number of switches resulted in the increased usage of a coding transcript (15%) as opposed to a non-coding transcript (6.8%). Interestingly, when compared to the Korowai population, isoforms up-regulated in the Sumba population are more likely to have an IDR gain (9.2%) rather than an IDR loss (5%) and are also more likely to be coding, whereas the inverse is true for isoforms unregulated in the Mentawai population versus the Korowai population (Figure 5A). IDRs act as a flexible point of contact for protein-protein interactions and are often enriched in protein products that play a critical role in signaling, regulation, and molecular recognition [106–108].

We also found that the number of these functionally significant switches, and the extent of overlap between pairwise comparisons, varied greatly. For the Mentawai vs. Sumba comparison, we detected statistically significant isoform switching events with a functional consequence across 36 genes, 106 genes for Korowai vs. Sumba, and 77 genes for the Korowai vs. Mentawai comparison (58.3%, 68.8%, and 73.2% of all isoform switches in each comparison; Figure 5B). The overlap in functionally significant switching genes was greatest between the pairs involving the Korowai group, with the Mentawai vs. Sumba comparison sharing very few switching genes with either of the other two pairs. Moreover, there were no functionally significant switching genes that were shared across the three sets.

Many of the significant isoform switching events occurred in genes that play critical roles in the immune response, one of which is the *IL18RAP* gene, the most significant isoform switching event across all comparisons (q-value = 5×10^{-27}, Table S9). The IL18RAP protein (Interleukin 18 Receptor Accessory Protein) serves as a receptor accessory protein for interleukin-18 (IL-18), a cytokine involved in immune responses and inflammation [109]. As such, *IL18RAP* is a crucial component of the IL-18 receptor complex, and the dysregulation of *IL18RAP* expression has been implicated in
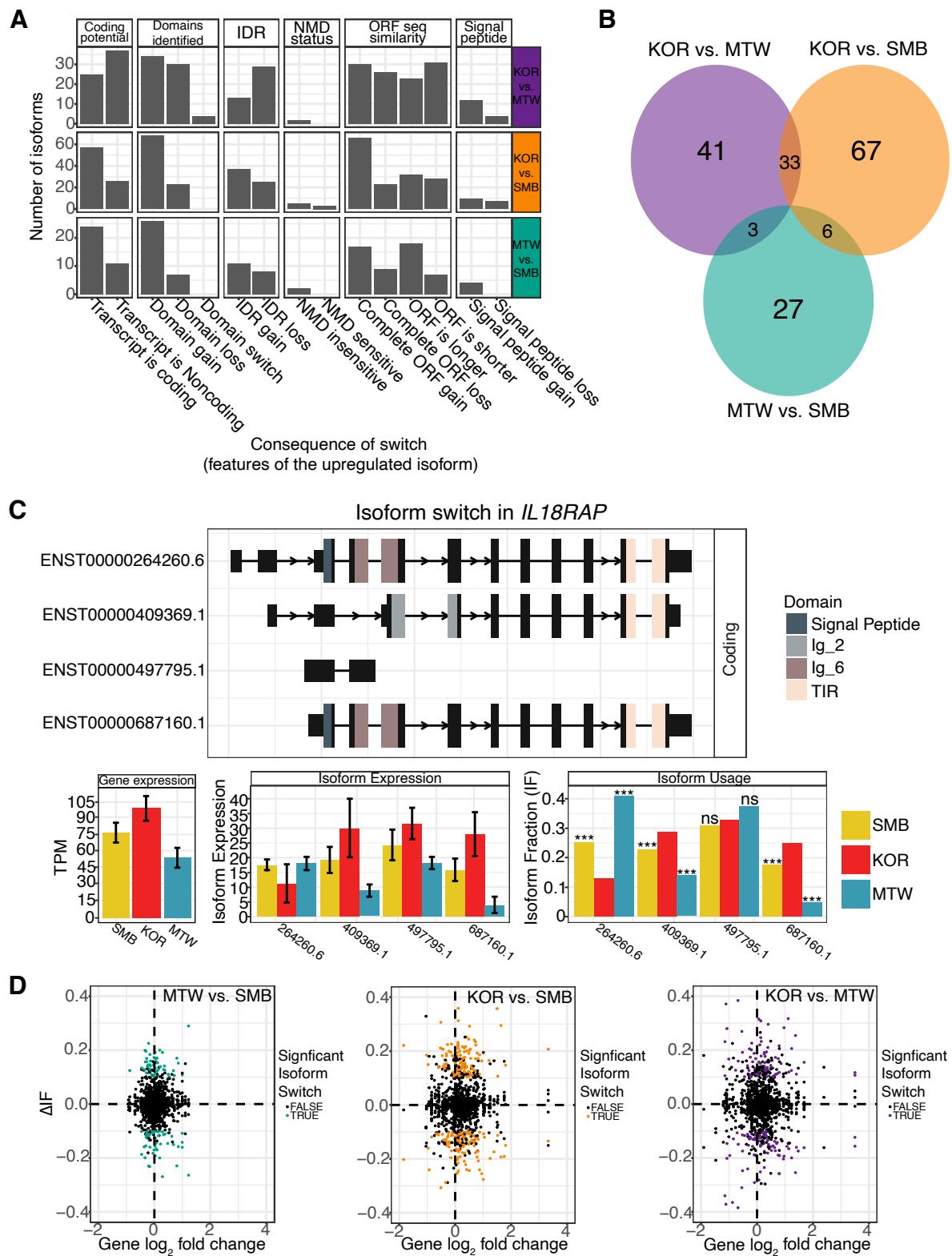
Figure 5. Global isoform switching analysis and functional consequences. **A**) For each pairwise group comparison, the consequences of all significant isoform switching events are described. Specifically, the annotation differences between the up-regulated isoform (dIF > 0) and the down-regulated isoform (dIF < 0) are summarized. The first population in the switch comparison is the ground state, and the second population is the changed stated. Thus, an up-regulated isoform in the KOR vs. MTW comparison, for example, is one that is used more in MTW relative to KOR. **B**) Venn diagram of the number of genes with overlapping isoform switching events and functional consequences (FDR < 0.05) across all pairwise comparisons. **C**) The isoform switch in *IL18RAP*. Exons with annotated domains are coloured, and those without are black boxes. Normalized *IL18RAP* expression (TPM), isoform expression, and isoform usage are plotted. Error-bars indicate 95% confidence intervals. Level of significance (with KOR as the reference) is denoted by the asterisks. **D**) Changes in isoform switches against changes in gene expression (x-axis) for each pairwise comparison. Significant (ΔIF > 0.1 and FDR < 0.05) isoform switches are coloured accordingly.

various diseases, including autoimmune disorders, inflammatory conditions, and cancer [110]. In our data, we observe preferential usage of the longer *IL18RAP* isoform (ENST00000264260.6) in the MTW group (isoform switch q-value $= 1.8 \times 10^{-8}$), while the SMB and KOR groups exhibit preferential usage of the shorter isoforms ENST00000409369.1 and ENST00000687160.1 (Figure 5C). Notably, the ENST00000264260.6 isoform contains a signal peptide at the third exon, while isoforms ENST00000409369.1 and ENST00000497795.1 do not. Signal peptides play a key role in directing proteins to their appropriate cellular locations. Thus, higher usage of the *IL18RAP* signal peptide-containing isoforms (ENST00000264260.6 and ENST00000687160.1) might be indicative of an up-regulation of IL18RAP protein translocation. Furthermore, isoforms ENST00000264260.6, ENST00000409369.1, and ENST00000687160.1 contain immunoglobulin domainsIg_2 (PF13895), and Ig_6 (PF18452), which play essential roles in antigen recognition, cell adhesion, receptor-ligand interactions, and structural stability [111]. Higher isoform usage of these isoforms might therefore influence structural stability and protein diversity.

In addition, we looked for evidence of significant isoform switching across our 11 detected Papuan ancestry-driven sGenes, such as *DNM3*. We observed statistically significant isoform switching that also resulted in functional domain changes in the *BST1* gene (Supplementary Figure 7A). *BST1* (bone marrow stromal cell antigen 1) is an immune gene that facilitates pre-B-cell growth and regulates leukocyte diapedesis in inflammation [112]. The isoform switch in the *BST1* gene reveals increased usage of the ENST00000265016.9 isoform within the SMB group (isoform switch q-value $= 1.3 \times 10^{-4}$), while KOR samples exhibit a higher isoform fraction of the ENST00000514989.1 isoform. The ENST00000514989.1 isoform is non-coding, which means that a switch in isoform usage results in a loss of coding potential. Interestingly, the ENST00000265016.9 isoform contains a rib hydrolase domain at exons 1-8, which is essential for *BST1* enzymatic activity as a NAD+ glycohydrolase [113].

Most differentially expressed genes are not differentially spliced

Our isoform switching analyses illustrate how critical alternative splicing is for the diversification of gene expression at both the mRNA and protein levels (Figures 5A and D). To assess the extent to which splicing and transcriptional regulation are linked across each pairwise comparison, we intersected the differentially spliced genes with previously reported [33] differentially expressed genes in the same groups. We found that differentially spliced genes were very rarely differentially expressed, with only 13 (Fishers exact test $p = 9.7e\text{-}04$), 65 (Fishers exact test $p = 1.1e\text{-}07$), and 39 (Fishers exact test $p = 2.1e\text{-}06$) shared genes across both tests for the Mentawai vs. Sumba, Korowai vs. Sumba, and Korowai vs. Mentawai comparisons, respectively (Supplementary Figure 7B). In other words, for most genes, we observe changes in the mRNA splice variants produced without a reciprocal change in total expression. These results support previous findings of independent regulation of gene splicing and expression [22, 114–117].

Discussion

It is well-established that alternative splicing influences various biological functions [2, 7–11, 22]. Studies of the link between genotypic and splicing diversity in humans, however, have not yet sampled the breadth of genetic and environmental diversity that characterizes our species, making it challenging to assess the degree to which population-specific forces contribute to splicing variation. In this work, using a set of 115 samples from three traditional island populations spanning the regions genetic ancestry cline, we characterize the global alternative splicing landscape across these populations and identify genetic variants that are associated with splicing (sQTLs), thereby providing a comprehensive map of genetically regulated alternative splicing events in human whole blood.

Our detection of alternative splicing revealed that the most frequent splicing event types were skipped exon events (50.1%) and the least frequent were retained intron events (3.5%). While these proportions are expected in higher eukaryotes [41, 118–123], it is important to note that short-read algorithms for detecting AS events are less likely to misassemble skipped exon events and can therefore identify them with high accuracy, but they suffer from poor precision with respect to RI detection [124, 125]. Accurately disambiguating real splicing events from background noise (such as partial transcript processing) is greatly facilitated by long-read sequencing technologies, which enable much more precise discovery and reconstruction of full-length isoforms, significantly improving the characterization of functional transcripts [126]. Thus, for future isoform-specific alternative splicing studies, the application of long-read technologies should be prioritized.

We conducted a global differential alternative splicing analysis between these groups and detected over 1500 significant events across all comparisons, with Mentawai and Sumba exhibiting the lowest number of differential events (i.e., the highest level of alternative splicing similarity). Given that these two populations share very similar proportions of West Island Southeast Asian-like genetic ancestry, while the Korowai are of Papuan-like genetic ancestry, our results suggest that ancestral differences might play a role in driving some of the variation in splicing across these populations. Interestingly, our *cis*-sQTL mapping analysis revealed 11 sQTLs driven by local genetic ancestry differences between groups. Furthermore, we detected population-specific patterns of splicing regulation at a global scale, identifying 312 Indonesia-specific SNP-sGene pairs with no evidence of shared effects with European sQTLs. Although our analysis is limited by our sample size, these findings suggest that alternative splicing variation between populations reflects a complex interplay between genetic and environmental factors.

As a means of characterizing the phenotypic consequence of alternative splicing variation across these Indonesian samples, we investigated whether GWAS signals for hematological traits measured in East Asian participants colocalized with Indonesian sQTLs. Of the 13 GWAS traits examined, 45 unique sGenes showed strong evidence of colocalisation with 12 distinct traits. Although colocalization analysis alone is not a sufficient means by which causal genes can be defined, the Papuan-driven splicing gene *DNM3* significantly colocalized with hemoglobin concentration, indicating its potential role in regulating this trait, and is notable observation given the role of high altitude corridors in the settlement of New Guinea Island [127].

To further bridge the gap between our identified sQTLs and their functional consequences, we employed a deep-learning model, DeepCLIP, to predict changes in RBP binding affinity. We found that the DeepCLIP predictions of the impact of sQTL SNPs on RBP binding dynamics correlated with the direction of effects that were observed at these splice junctions. In particular, we predicted a potential increase in HNRNPM binding affinity at rs6433311 (A>G, *DYNC1I2*), while our sQTL mapping analysis showed that the G allele at rs6433311 is associated with an effect size of -1.174 relative to the A allele. Leafcutter-predicted skipping junctions (exon skipping events) yield negative effect sizes, while inclusion junctions (exon inclusion events) yield positive effect sizes. Therefore, the predicted increase in binding affinity of a splicing factor that promotes exon exclusion [92, 94–96] should be associated with a negative effect size at this SNP, which is indeed what we observe. For another splicing factor, PTBP1, DeepCLIP analysis highlighted a potential binding disruption at the rs27291 G>A mutation in the *LNPEP* gene. In our data, the A allele of this sQTL SNP is associated with an effect size of 0.672, which is indicative of increased exon inclusion. Given the role PTBP1 plays in suppressing exon inclusion (thereby promoting exon exclusion) [103–105], the predicted reduction in PTBP1 activity directly correlates with a positive effect size for this sQTL and a higher exon inclusion rate for individuals harbouring the rs27291 A allele. While these results are consistent with previous findings that sQTLs disrupt the binding affinity of core and auxiliary splice proteins [22, 128], it is important to note that we have not conducted any experimental validation of the predicted protein binding activity, thereby limiting our ability to evaluate the accuracy of the deep-learning predictions. Nonetheless, our results highlight

the molecular mechanisms through which genetic variants impact splicing, and demonstrate the value of applying deep learning methods to refine our collective understanding of the consequences of human genetic variation.

We also explored evidence of functionally significant isoform switching events across the three populations, identifying frequent changes in protein domains (domain gain/loss), NMD status, and ORF sequence similarity. We observed the greatest number of isoform switching genes in comparisons involving the Korowai population, in line with the population-level alternative splicing differences we identified with SUPPA2. Notably, we report that when an isoform switch occurs between the Korowai group and the Sumba or Mentawai groups, the up-regulated isoforms in Sumba or Mentawai are more likely to: i) contain an additional protein domain, ii) contain a signal peptide, and iii) gain a complete open reading frame. Our analysis identified several significant isoform switching events in genes critical for the immune response, including the *IL18RAP* and *BST1* genes. In the *IL18RAP* gene, we observed preferential usage of different isoforms across populations, with potential implications for protein translocation and structural stability. Similarly, in the *BST1* gene, isoform switching resulted in changes in coding potential and the presence of functional domains, highlighting the importance of alternative splicing in immune regulation and disease processes.

As previous studies have reported [129–131], we find that a large subset of genes exhibit isoform switches without significant changes in their global expression levels. Such genes would typically be overlooked in RNA-seq studies that strictly focus on differential gene expression, thereby limiting critical biological insight. Irrespective of the global expression levels of a gene, changes in the relative expression of its isoforms influence protein abundance, which in turn modulates cellular processes. Furthermore, we detect minimal overlap between differentially expressed genes and differentially alternatively spliced genes, indicating that the genetic control of splicing and transcription are independent, as previously observed [17, 21, 22, 30, 132–134]. Indeed, across our Indonesian dataset, splicing and expression affect different pathways, as DE and DAS genes were enriched for distinct biological processes and molecular functions. DAS genes were mostly involved in processes related to post-transcriptional regulation, DNA replication, and cell-cycle signaling. In contrast, DE genes were enriched in regulatory pathways related to the adaptive immune response and nervous system function [33]. Together, this study provides a comprehensive catalog of genetically regulated alternative splicing events in whole blood Indonesian samples, and adds to our knowledge of genomic and regional drivers of gene regulatory variation across the Indonesian archipelago.

Acknowledgements

Data and code availability

All data used in this project was previously published and is deposited at the European Genome Archive under accession EGAS00001003671. Genome-wide significant and nominal test statistics for all sQTL tests are provided as supplementary tables and are publicly available on Figshare: `https://melbourne.figshare.com/projects/Profiling_genetically_`

driven_alternative_splicing_across_the_Indonesian_Archipelago/204111. Analysis code is available at `https://gitlab.svi.edu.au/igr-lab/indo_splicing`

Supplementary Figures
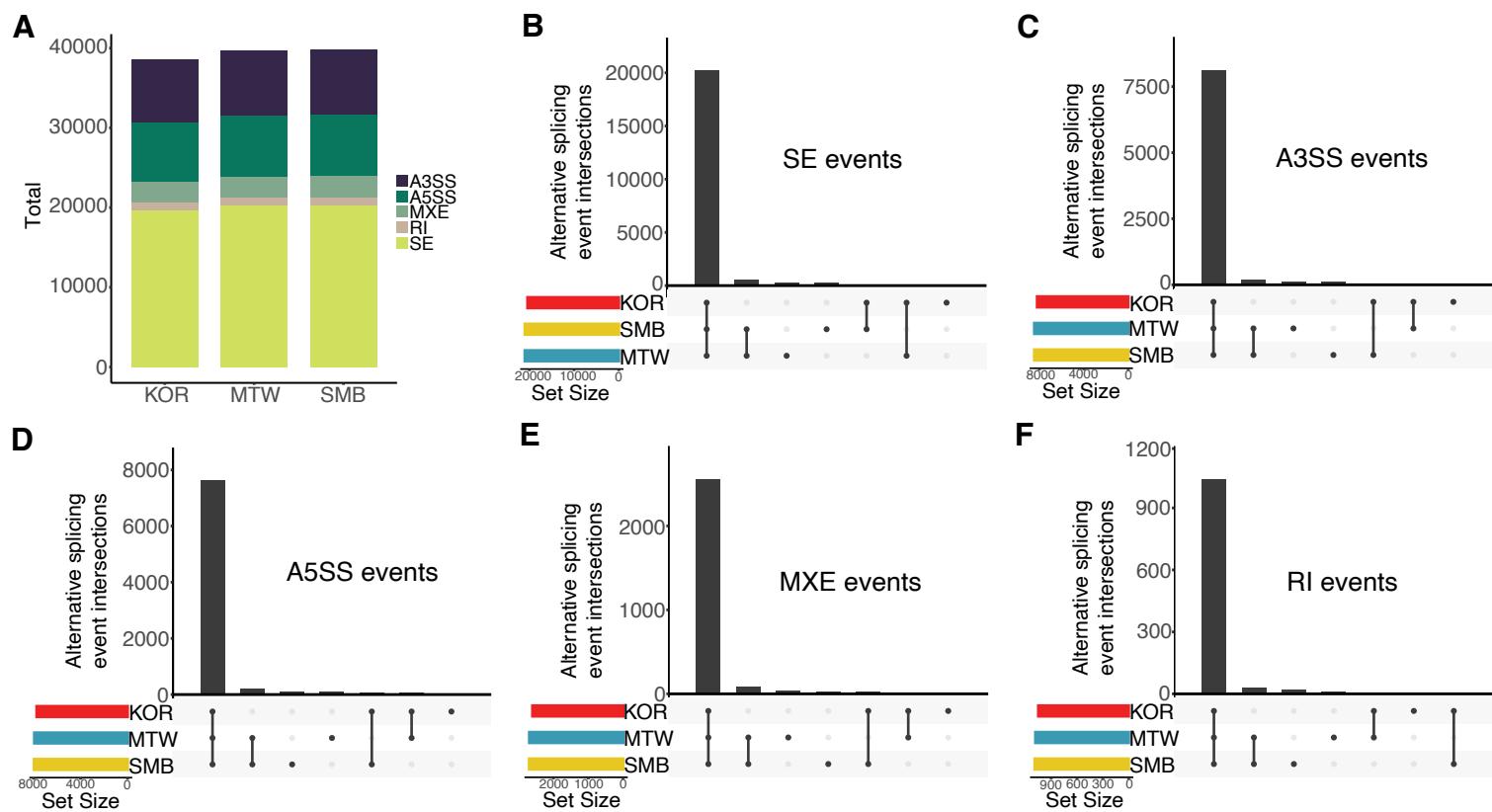
Supplementary Figure 1. AS events quantified with SUPPA2. A) Proportion of each AS event type for Korowai, Mentawai, and Sumba. **B)-F)** AS event intersections between sample groups.
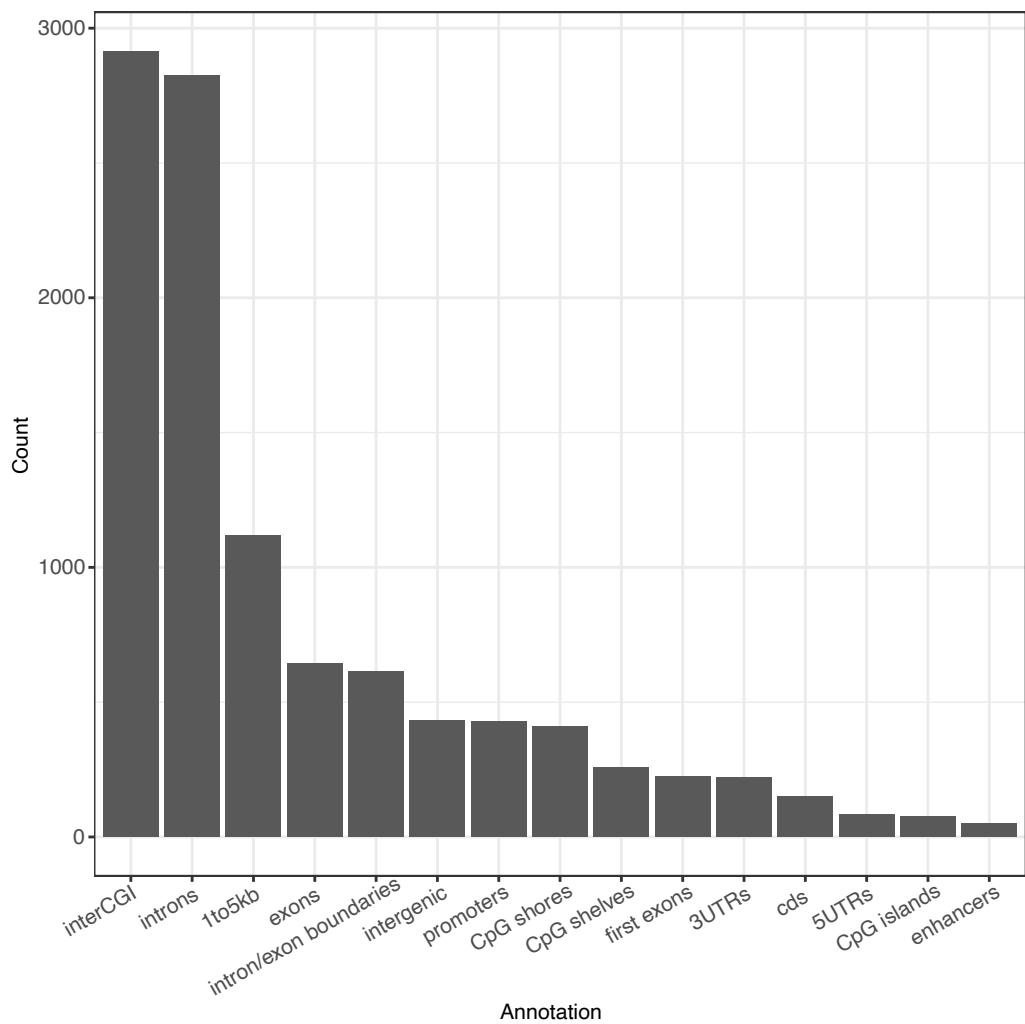
Supplementary Figure 2. Genomic locations of sQTLs. Significant sQTLs and their prevalance in annotated regions.
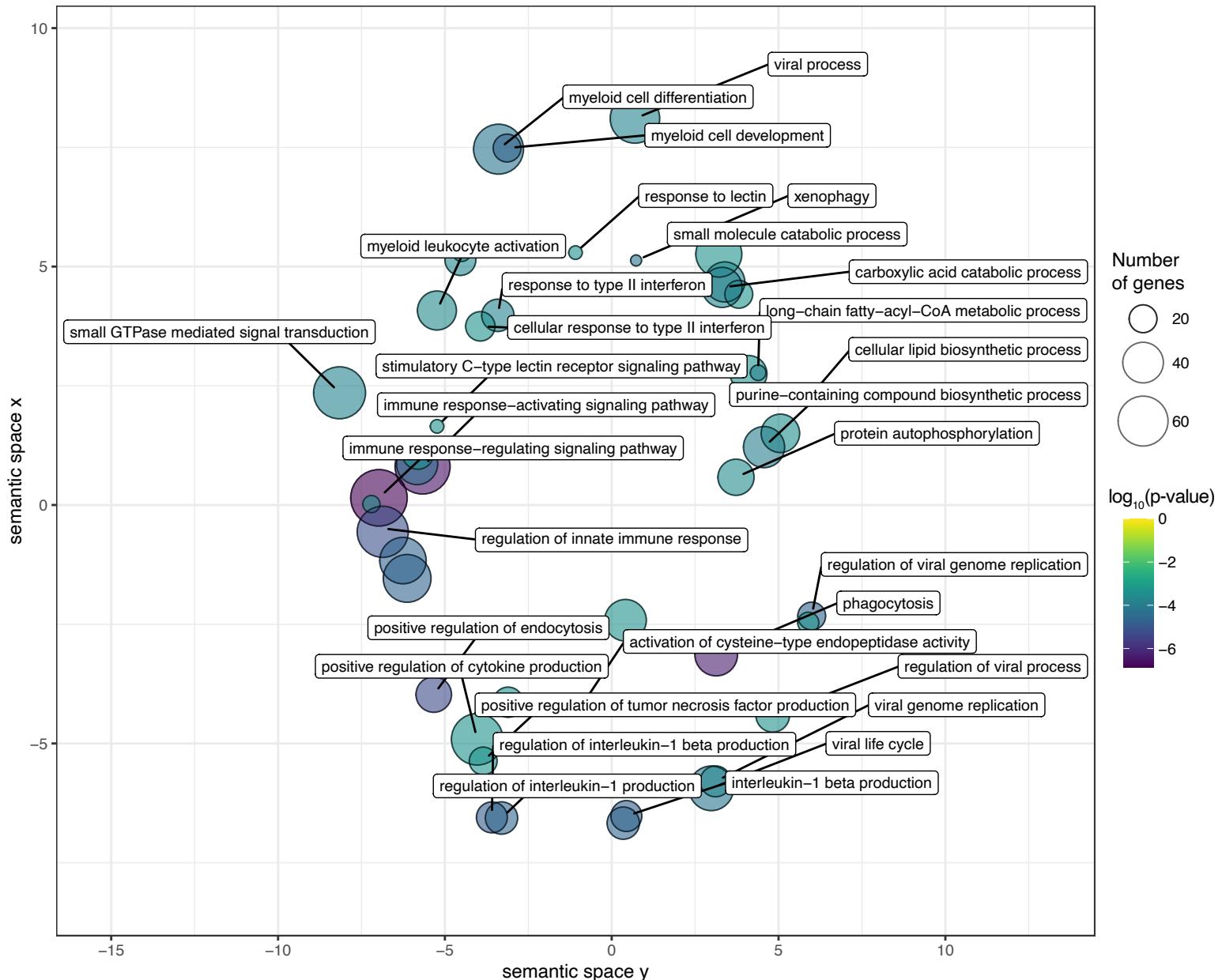
Supplementary Figure 3. GO term clustering of differentially alternatively spliced genes using semantic similarity measures in Revigo.
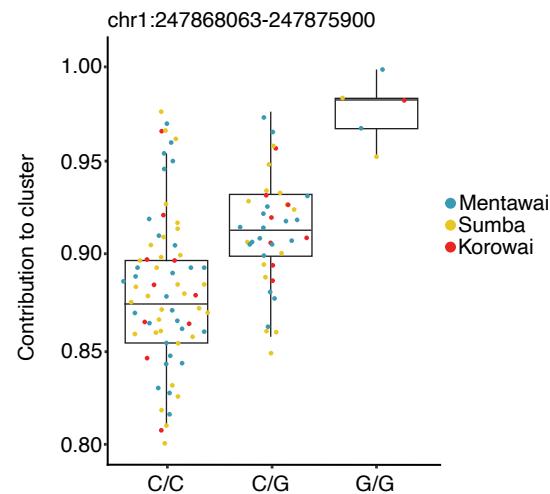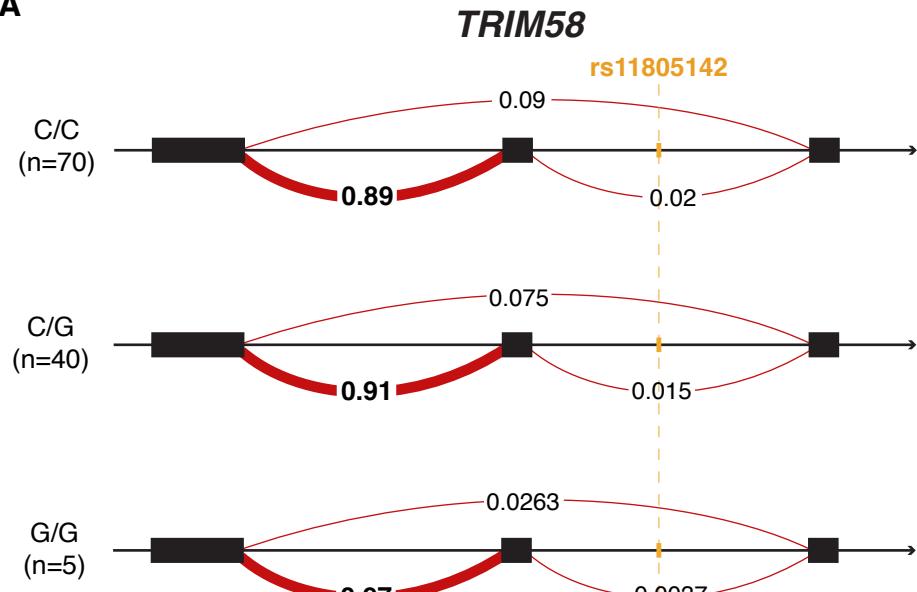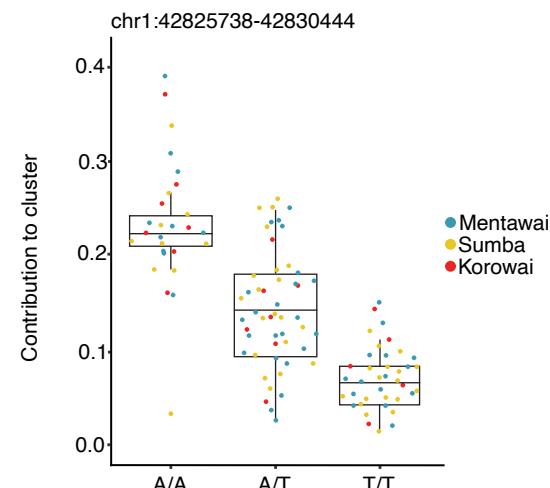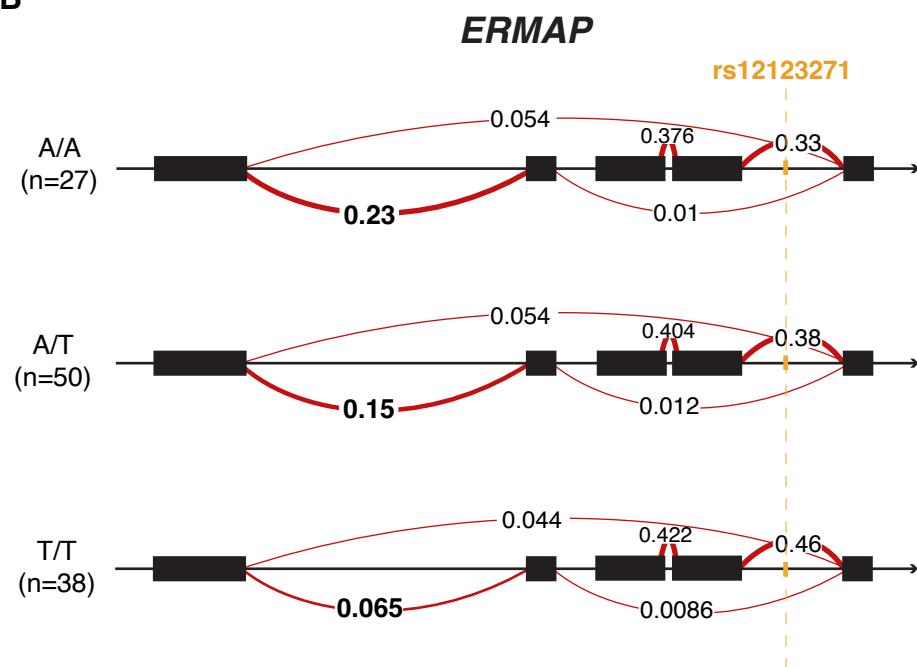
Supplementary Figure 4. sQTL examples. A) Genotype plot of the *TRIM58* sQTL event. For this alternative splicing event, the junction with the most significant association to the SNP is highlighted with bold numbers across all three genotypes, and these numbers represent the proportion of reads spanning the junction. On the right, boxplots depict the distributions for each genotype. **B)** Genotype plot of the *ERMAP* sQTL event.
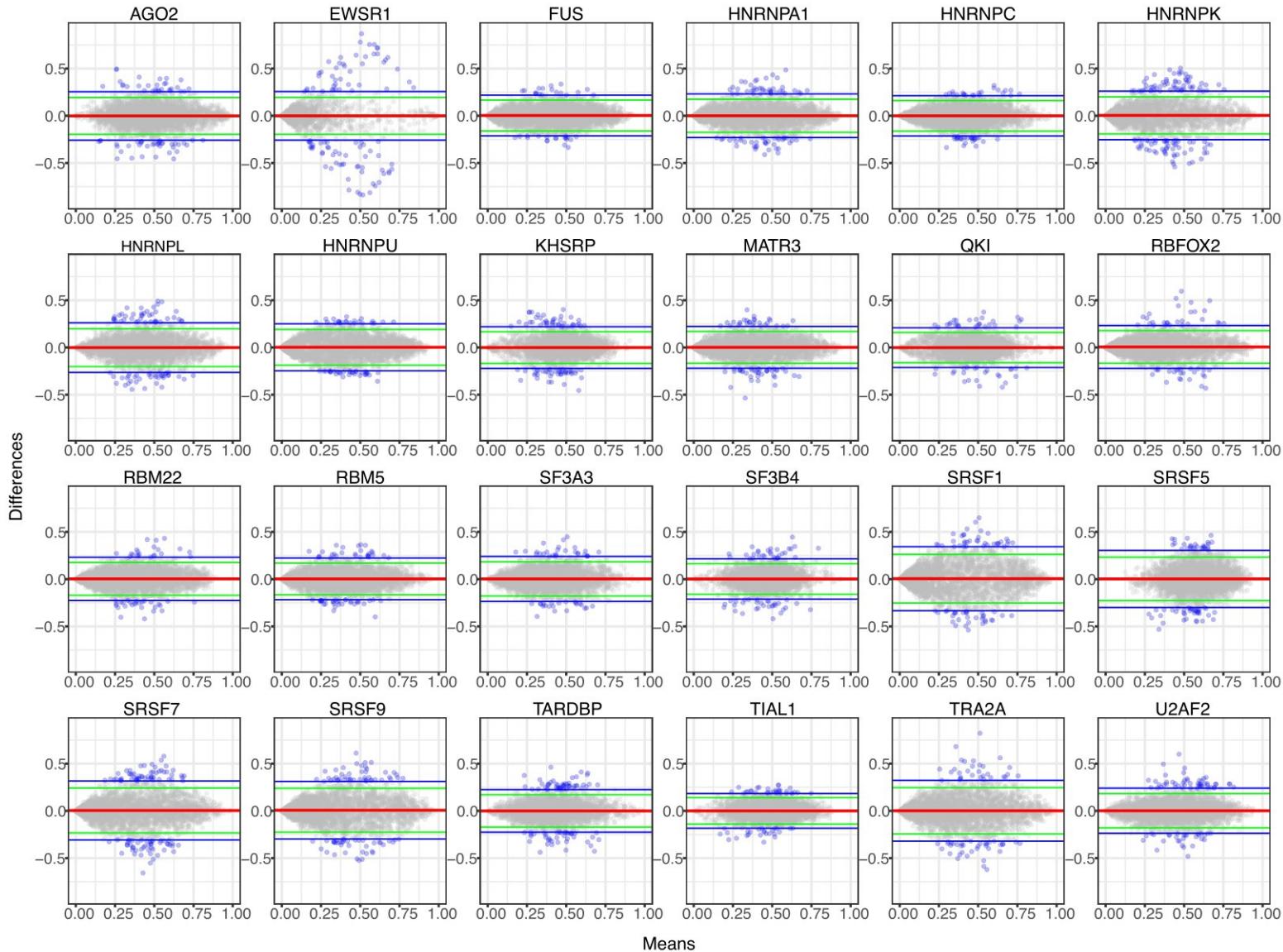
Supplementary Figure 5. sQTL impact on RBP binding affinity. Mean-difference plots for 24 of the 33 splicing-RBPs. Each point represents an sQTL variant sequence (i.e., the sQTL with a 10bp flanking sequence) and reference sequence pair. The y-axis represents the difference between the DeepCLIP reference sequence score and the variant sequence score, while the x-axis depicts the average between these two scores. Horizontal green and blue lines represent the 95% and 99% CIs, respectively. Points that are outside of the 99% CI are colored blue.
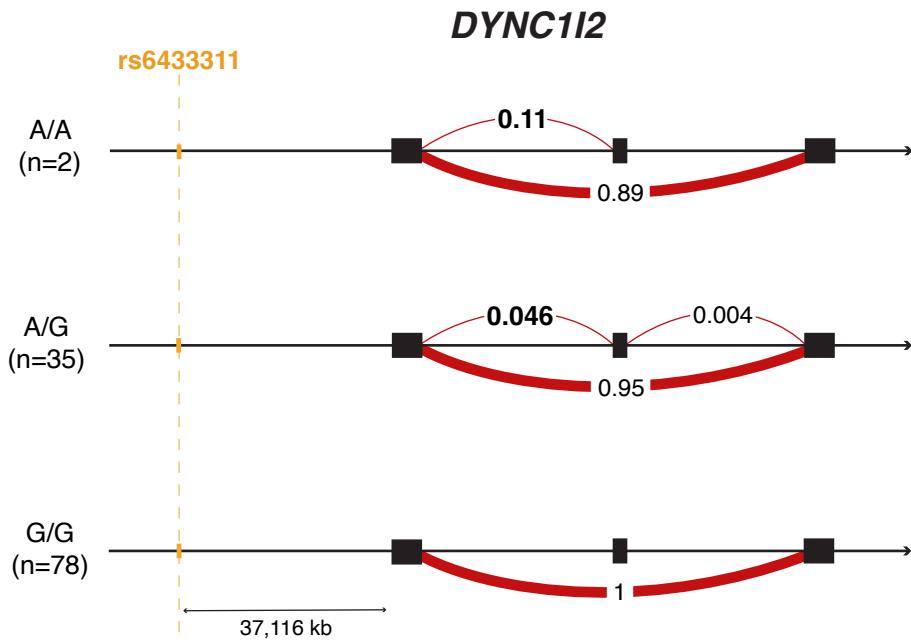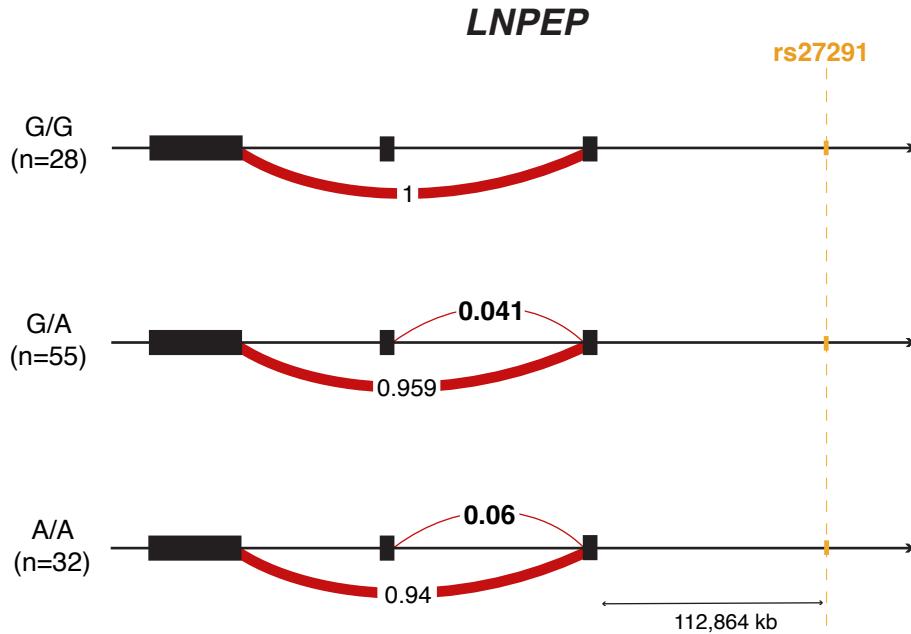
Supplementary Figure 6. sQTL genotype plots for *DYNC1I2* and *LNPEP*. A) Genotype plot of the *DYNC1I2* sQTL event. For this alternative splicing event, the junction with the most significant association to the SNP is highlighted with bold numbers across all three genotypes, and these numbers represent the proportion of reads spanning the junction. **B**) Genotype plot of the *LNPEP* sQTL event.
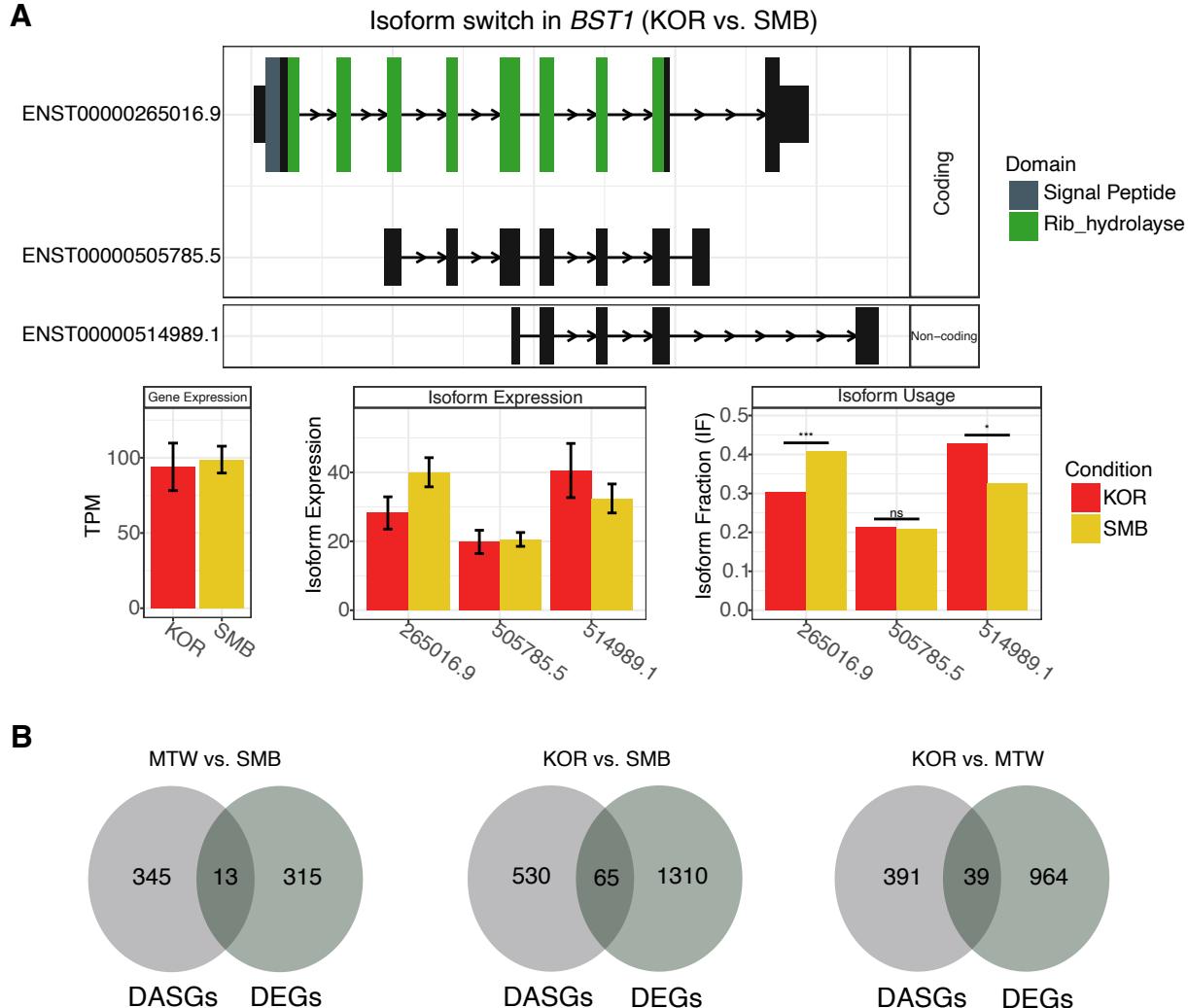
Supplementary Figure 7. Comparing isoform expression and gene expression. A) Isoform switching analysis for *BST1* (between the Korowai population and the Sumba population). Exons with annotated domains are coloured, and those without are black boxes. Gene expression (TPM), as well as isoform expression and isoform usage are plotted for all detected isoforms. Error-bars indicate 95% confidence intervals. Level of significance for differences in isoform fraction (IF) is denoted by the asterisks. **B)** The overlap between differentially expressed genes (DEGs) and differentially alternatively spliced genes (DASGs).

Supplementary Tables

Table S1

Gene Ontology and KEGG overrepresentation analysis for all AS genes.

Table S2

Gene Ontology and KEGG overrepresentation analysis for population-specific AS genes.

Table S3

Gene Ontology and KEGG overrepresentation analysis for DAS genes.

Table S4

Permutation-significant sQTLs. Available on Figshare: `https://melbourne.figshare.com/projects/Profiling_genetically_driven_alternative_splicing_across_the_Indonesian_Archipelago/204111`

Table S5

Nominal sQTL statistics. Available on Figshare: `https://melbourne.figshare.com/projects/Profiling_genetically_driven_alternative_splicing_across_the_Indonesian_Archipelago/204111`

Table S6

Gene Ontology and KEGG overrepresentation analysis for sGenes.

Table S7

sQTL-LAI correlation results for Papuan-like genetic ancestry.

Table S8

ClinVar annotation of lead sQTL SNPs.

Table S9

Summary of isoform switch analyses.

Table S10

Exon coordinates (GRCh38 Ensembl release 110) for all isoforms involved in switching events.

Supplementary Files

File S1

sQTL-GWAS colocalization results.

File S2

CPC2 isoform coding potential prediction.

File S3

IUPred2A prediction of Intrinsically Disordered Regions (IDRs).

File S4

SignalP prediction of signal peptides.

File S5

PFAM domain prediction.

References

1. P A Sharp. Split genes and RNA splicing. *Cell*, 77(6):805–815, June 1994.

2. Charlotte J. Wright, Christopher W. J. Smith, and Chris D. Jiggins. Alternative splicing as a source of phenotypic diversity. *Nature Reviews Genetics*, 23(11):697–710, November 2022. Number: 11 Publisher: Nature Publishing Group.

3. Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, December 2008.

4. Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008. Number: 7221 Publisher: Nature Publishing Group.

5. Marina M Scotti and Maurice S Swanson. RNA mis-splicing in disease. *Nat. Rev. Genet.*, 17(1):19–32, November 2015.

6. Guey-Shin Wang and Thomas A. Cooper. Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews Genetics*, 8(10):749–761, October 2007. Number: 10 Publisher: Nature Publishing Group.

7. Ayla Ergun, Graeme Doran, James C Costello, Henry H Paik, James J Collins, Diane Mathis, Christophe Benoist, and ImmGen Consortium. Differential splicing across immune system lineages. *Proc. Natl. Acad. Sci. U. S. A.*, 110(35):14324–14329, August 2013.

8. Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, November 2008.

9. Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, Avik Datta, David Richardson, Frances Burden, Daniel Mead, Alice L Mann, Jose Maria Fernandez, Sophia Rowlston, Steven P Wilder, Samantha Farrow, Xiaojian Shao, John J Lambourne, Adriana Redensek, Cornelis A Albers, Vyacheslav Amstislavskiy, Sofie Ashford, Kim Berentsen, Lorenzo Bomba, Guillaume Bourque, David Bujold, Stephan Busche, Maxime Caron, Shu-Huang Chen, Warren Cheung, Oliver Delaneau, Emmanouil T Dermitzakis, Heather Elding, Irina Colgiu, Frederik O Bagger, Paul Flicek, Ehsan Habibi, Valentina Iotchkova, Eva Janssen-Megens, Bowon Kim, Hans Lehrach, Ernesto Lowy, Amit Mandoli, Filomena Matarese, Matthew T Maurano, John A Morris, Vera Pancaldi, Farzin Pourfarzad, Karola Rehnstrom, Augusto Rendon, Thomas Risch, Nilofar Sharifi, Marie-Michelle Simon, Marc Sultan, Alfonso Valencia, Klaudia Walter, Shuang-Yin Wang, Mattia Frontini, Stylianos E Antonarakis, Laura Clarke, Marie-Laure Yaspo, Stephan Beck, Roderic Guigo, Daniel Rico, Joost H A Martens, Willem H Ouwehand, Taco W Kuijpers, Dirk S Paul, Hendrik G Stunnenberg, Oliver Stegle, Kate Downes, Tomi Pastinen, and Nicole Soranzo. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414.e24, November 2016.

10. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, September 2020.

11. Yida Zhang, Harry Taegyun Yang, Kathryn Kadash-Edmondson, Yang Pan, Zhicheng Pan, Beverly L Davidson, and Yi Xing. Regional variation of splicing QTLs in human brain. *Am. J. Hum. Genet.*, 107(2):196–210, August 2020.

12. Joseph K. Pickrell, Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genetics*, 6(12):e1001236, December 2010.

13. Baptiste Saudemont, Alexandra Popa, Joanna L. Parmley, Vincent Rocher, Corinne Blugeon, Anamaria Necsulea, Eric Meyer, and Laurent Duret. The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. *Genome Biology*, 18(1):208, October 2017.

14. Benjamin Fair, Carlos Buen Abad Najar, Junxing Zhao, Stephanie Lozano, Austin Reilly, Gabriela Mossian, Jonathan P Staley, Jingxin Wang, and Yang I Li. Global impact of aberrant splicing on human gene expression levels. *bioRxiv*, October 2023.

15. Susan Carpenter, Emiliano P Ricci, Blandine C Mercier, Melissa J Moore, and Katherine A Fitzgerald. Post-transcriptional regulation of gene expression in innate immunity. *Nat. Rev. Immunol.*, 14(6):361–376, June 2014.

16. Allison L Richards, Donovan Watza, Anthony Findley, Adnan Alazizi, Xiaoquan Wen, Athma A Pai, Roger Pique-Regi, and Francesca Luca. Environmental perturbations lead to extensive directional shifts in RNA processing. *PLoS Genet.*, 13(10):e1006995, October 2017.

17. Chun Jimmie Ye, Jenny Chen, Alexandra-Chloé Villani, Rachel E Gate, Meena Subramaniam, Tushar Bhangale, Mark N Lee, Towfique Raj, Raktima Raychowdhury, Weibo Li, Noga Rogel, Sean Simmons, Selina H Imboywa, Portia I Chipendo, Cristin McCabe, Michelle H Lee, Irene Y Frohlich, Barbara E Stranger, Philip L De Jager, Aviv Regev, Tim Behrens, and Nir Hacohen. Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. *Genome Res.*, 28(12):1812–1825, December 2018.

18. Athma A Pai, Golshid Baharian, Ariane Pagé Sabourin, Jessica F Brinkworth, Yohann Nédélec, Joseph W Foley, Jean-Christophe Grenier, Katherine J Siddle, Anne Dumaine, Vania Yotova, Zachary P Johnson, Robert E Lanford, Christopher B Burge, and Luis B Barreiro. Widespread shortening of 3' untranslated regions and increased exon inclusion are evolutionarily conserved features of innate immune responses to infection. *PLoS Genet.*, 12(9):e1006338, September 2016.

19. Haroon Kalam, Mary F Fontana, and Dhiraj Kumar. Alternate splicing of transcripts shape macrophage response to mycobacterium tuberculosis infection. *PLoS Pathog.*, 13(3):e1006236, March 2017.

20. Nicole M Martinez and Kristen W Lynch. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol. Rev.*, 253(1):216–236, May 2013.

21. Maxime Rotival, Hélène Quach, and Lluis Quintana-Murci. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nature Communications*, 10(1):1671, April 2019.

22. Yang I Li, Bryce van de Geijn, Anil Raj, David A Knowles, Allegra A Petti, David Golan, Yoav Gilad, and Jonathan K Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 2016.

23. Zefeng Wang and Christopher B Burge. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, 14(5):802–813, May 2008.

24. Nuno André Faustino and Thomas A Cooper. Pre-mRNA splicing and human disease. *Genes Dev.*, 17(4):419–437, February 2003.

25. Jennifer L Caswell, Roman Camarda, Alicia Y Zhou, Scott Huntsman, Donglei Hu, Steven E Brenner, Noah Zaitlen, Andrei Goga, and Elad Ziv. Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors. *Hum. Mol. Genet.*, 24(25):7421–7431, December 2015.

26. Towfique Raj, Yang I Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, Bernard Ng, Ishaan Gupta, Vahram Haroutunian, Eric E Schadt, Tracy Young-Pearse, Sara Mostafavi, Bin Zhang, Pamela Sklar, David A Bennett, and Philip L De Jager. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.*, 50(11):1584–1592, November 2018.

27. Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing QTLs in the human brain and their enrichment among Schizophrenia-associated loci. *Nat. Commun.*, 8:14519, February 2017.

28. Alicia R. Martin, Helio A. Costa, Tuuli Lappalainen, Brenna M. Henn, Jeffrey M. Kidd, Muh-Ching Yee, Fabian Grubert, Howard M. Cann, Michael Snyder, Stephen B. Montgomery, and Carlos D. Bustamante. Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. *PLoS Genetics*, 10(8):e1004549, August 2014.

29. Minal Çalışkan, Darren A. Cusanovich, Carole Ober, and Yoav Gilad. The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics*, 20(8):1643–1652, April 2011.

30. Raquel García-Pérez, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, Paul Joris Rognon, Salvador Capella-Gutierrez, Miquel Calvo, Ferran Reverter, Roderic Guigó, François Aguet, Pedro G. Ferreira, Kristin G. Ardlie, and Marta Melé. The landscape of expression and alternative splicing variation across human traits. *Cell Genomics*, 3(1), January 2023. Publisher: Elsevier.

31. Eddie Park, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *The American Journal of Human Genetics*, 102(1):11–26, January 2018. Publisher: Elsevier.

32. Derek E. Kelly, Shweta Ramdas, Rong Ma, Renata A. Rawlings-Goss, Gregory R. Grant, Alessia Ranciaro, Jibril B. Hirbo, William Beggs, Meredith Yeager, Stephen Chanock, Thomas B. Nyambo, Sabah A. Omar, Dawit Woldemeskel, Gurja Belay, Hongzhe Li, Christopher D. Brown, and Sarah A. Tishkoff. The genetic and evolutionary basis of gene expression variation in East Africans. *Genome Biology*, 24(1):35, February 2023.

33. Heini M Natri, Katalina S Bobowik, Pradiptajati Kusuma, Chelzie Crenna Darusallam, Guy S Jacobs, Georgi Hudjashov, J Stephen Lansing, Herawati Sudoyo, Nicholas E Banovich, Murray P Cox, and Irene Gallego Romero. Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the Indonesian archipelago. *PLoS Genet.*, 16(5):e1008749, May 2020.

34. Heini M Natri, Georgi Hudjashov, Guy Jacobs, Pradiptajati Kusuma, Lauri Saag, Chelzie Crenna Darusallam, Mait Metspalu, Herawati Sudoyo, Murray P Cox, Irene Gallego Romero, and Nicholas E Banovich. Genetic

architecture of gene regulation in Indonesian populations identifies QTLs associated with global and local ancestries. *Am. J. Hum. Genet.*, 109(1):50–65, January 2022.

35. Georgi Hudjashov, Tatiana M Karafet, Daniel J Lawson, Sean Downey, Olga Savina, Herawati Sudoyo, J Stephen Lansing, Michael F Hammer, and Murray P Cox. Complex patterns of admixture across the Indonesian archipelago. *Mol. Biol. Evol.*, 34(10):2439–2452, July 2017.

36. Guy S Jacobs, Georgi Hudjashov, Lauri Saag, Pradiptajati Kusuma, Chelzie C Darusallam, Daniel J Lawson, Mayukh Mondal, Luca Pagani, François-Xavier Ricaut, Mark Stoneking, Mait Metspalu, Herawati Sudoyo, J Stephen Lansing, and Murray P Cox. Multiple deeply divergent Denisovan ancestries in Papuans. *Cell*, 177(4):1010–1021.e32, May 2019.

37. Simon Andrews and Others. FastQC: a quality control tool for high throughput sequence data. 2010, 2017.

38. Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, August 2014.

39. Juan L Trincado, Juan C Entizne, Gerald Hysenaj, Babita Singh, Miha Skalic, David J Elliott, and Eduardo Eyras. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, 19(1):40, March 2018.

40. Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, 14(4):417–419, April 2017.

41. Raquel García-Pérez, Jose Miguel Ramirez, Aida Ripoll-Cladellas, Ruben Chazarra-Gil, Winona Oliveros, Oleksandra Soldatkina, Mattia Bosio, Paul Joris Rognon, Salvador Capella-Gutierrez, Miquel Calvo, Ferran Reverter, Roderic Guigó, François Aguet, Pedro G Ferreira, Kristin G Ardlie, and Marta Melé. The landscape of expression and alternative splicing variation across human traits. *Cell Genom*, 3(1):100244, January 2023.

42. Adria Closa, Marina Reixachs-Solé, Antonio C Fuentes-Fayos, Katharina E Hayer, Juan L Melero, Fabienne R S Adriaanse, Romy S Bos, Manuel Torres-Diz, Stephen P Hunger, Kathryn G Roberts, Charles G Mullighan, Ronald W Stam, Andrei Thomas-Tikhonenko, Justo P Castaño, Raúl M Luque, and Eduardo Eyras. A convergent malignant phenotype in b-cell acute lymphoblastic leukemia involving the splicing factor SRRM1. *NAR Cancer*, 4(4):zcac041, December 2022.

43. Yang I Li, David A Knowles, Jack Humphrey, Alvaro N Barbeira, Scott P Dickinson, Hae Kyung Im, and Jonathan K Pritchard. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, 50(1):151–158, January 2018.

44. Yarden Katz, Eric T Wang, Jacob Silterra, Schraga Schwartz, Bang Wong, Helga Thorvaldsdóttir, James T Robinson, Jill P Mesirov, Edoardo M Airoldi, and Christopher B Burge. Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics*, 31(14):2400–2402, July 2015.

45. Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, July 2008.

46. Kristoffer Vitting-Seerup and Albin Sandelin. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21):4469–4471, November 2019.

47. Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, 43(7):e47, April 2015.

48. Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res.*, 22(10):2008–2017, October 2012.

49. Yu-Jian Kang, De-Chang Yang, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei, and Ge Gao. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, 45(W1):W12–W16, July 2017.

50. Bálint Mészáros, Gábor Erdos, and Zsuzsanna Dosztányi. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, 46(W1):W329–W337, July 2018.

51. José Juan Almagro Armenteros, Konstantinos D Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, 37(4):420–423, April 2019.

52. Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Res.*, 49(D1):D412–D419, January 2021.

53. Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

54. Olivier Delaneau, Halit Ongen, Andrew A Brown, Alexandre Fort, Nikolaos I Panousis, and Emmanouil T Dermitzakis. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.*, 8:15452, May 2017.

55. Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.*, 7(3):500–507, February 2012.

56. Bioconductor Package Maintainer. liftover: Changing genomic coordinate systems with rtracklayer:: liftover. *R package version*, 1(0):10–18129, 2020.

57. Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.*, 51(1):187–195, January 2019.

58. Nicole R Gay, Michael Gloudemans, Margaret L Antonio, Nathan S Abell, Brunilda Balliu, Yoson Park, Alicia R Martin, Shaila Musharoff, Abhiram S Rao, François Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Farhad Hormozdiari, GTEx Consortium, Kristin G Ardlie, Christopher D Brown, Hae Kyung Im, Tuuli Lappalainen, Xiaoquan Wen, and Stephen B Montgomery. Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.*, 21(1):233, September 2020.

59. Tianzhi Wu, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, Lang Zhou, Wenli Tang, Li Zhan, Xiaocong Fu, Shanshan Liu, Xiaochen Bo, and Guangchuang Yu. clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)*, 2(3):100141, August 2021.

60. Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, 6(7):e21800, July 2011.

61. Alexander Gulliver Bjørnholt Grønning, Thomas Koed Doktor, Simon Jonas Larsen, Ulrika Simone Spangsberg Petersen, Lise Lolle Holm, Gitte Hoffmann Bruun, Michael Birkerod Hansen, Anne-Mette Hartung, Jan Baumbach, and Brage Storstein Andresen. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic Acids Res.*, 48(13):7099–7118, June 2020.

62. Carrie A Davis, Benjamin C Hitz, Cricket A Sloan, Esther T Chan, Jean M Davidson, Idan Gabdank, Jason A Hilton, Kriti Jain, Ulugbek K Baymuradov, Aditi K Narayanan, Kathrina C Onate, Keenan Graham, Stuart R Miyasato, Timothy R Dreszer, J Seth Strattan, Otto Jolanki, Forrest Y Tanaka, and J Michael Cherry. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, 46(D1):D794–D801, January 2018.

63. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

64. Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, 10(5):e1004383, May 2014.

65. Ming-Huei Chen, Laura M Raffield, Abdou Mousas, Saori Sakaue, Jennifer E Huffman, Arden Moscati, Bhavi Trivedi, Tao Jiang, Parsa Akbari, Dragana Vuckovic, Erik L Bao, Xue Zhong, Regina Manansala, Véronique Laplante, Minhui Chen, Ken Sin Lo, Huijun Qian, Caleb A Lareau, Mélissa Beaudoin, Karen A Hunt, Masato Akiyama, Traci M Bartz, Yoav Ben-Shlomo, Andrew Beswick, Jette Bork-Jensen, Erwin P Bottinger, Jennifer A Brody, Frank J A van Rooij, Kumaraswamynaidu Chitrala, Kelly Cho, Hélène Choquet, Adolfo Correa, John Danesh, Emanuele Di Angelantonio, Niki Dimou, Jingzhong Ding, Paul Elliott, Tõnu Esko, Michele K Evans, James S Floyd, Linda Broer, Niels Grarup, Michael H Guo, Andreas Greinacher, Jeff Haessler, Torben Hansen, Joanna M M Howson, Qin Qin Huang, Wei Huang, Eric Jorgenson, Tim Kacprowski, Mika Kähönen, Yoichiro Kamatani, Masahiro Kanai, Savita Karthikeyan, Fotis Koskeridis, Leslie A Lange, Terho Lehtimäki, Markus M Lerch, Allan Linneberg, Yongmei Liu, Leo-Pekka Lyytikäinen, Ani Manichaikul, Hilary C Martin, Koichi Matsuda, Karen L Mohlke, Nina Mononen, Yoshinori Murakami, Girish N Nadkarni, Matthias Nauck, Kjell Nikus, Willem H Ouwehand, Nathan Pankratz, Oluf Pedersen, Michael Preuss, Bruce M Psaty, Olli T Raitakari, David J Roberts, Stephen S Rich, Benjamin A T Rodriguez, Jonathan D Rosen, Jerome I Rotter, Petra Schubert, Cassandra N Spracklen, Praveen Surendran, Hua Tang, Jean-Claude Tardif, Richard C Trembath, Mohsen Ghanbari, Uwe Völker, Henry Völzke, Nicholas A Watkins, Alan B Zonderman, VA Million Veteran Program, Peter W F Wilson, Yun Li, Adam S Butterworth, Jean-François Gauchat, Charleston W K Chiang, Bingshan Li, Ruth J F Loos, William J Astle, Evangelos Evangelou, David A van Heel, Vijay G Sankaran, Yukinori Okada, Nicole Soranzo, Andrew D Johnson, Alexander P Reiner, Paul L Auer, and Guillaume Lettre. Trans-ethnic and Ancestry-Specific Blood-Cell genetics in 746,667 individuals from 5 global populations. *Cell*, 182(5):1198–1213.e14, September 2020.

66. Christian Benner, Aki S. Havulinna, Marjo-Riitta Jrvelin, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *The American Journal of Human Genetics*, 101(4):539–551, October 2017.

67. Eddie Park, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. The expanding landscape of alternative splicing variation in human populations. *Am. J. Hum. Genet.*, 102(1):11–26, January 2018.

68. Hui Y Xiong, Babak Alipanahi, Leo J Lee, Hannes Bretschneider, Daniele Merico, Ryan K C Yuen, Yimin Hua, Serge Gueroussov, Hamed S Najafabadi, Timothy R Hughes, Quaid Morris, Yoseph Barash, Adrian R Krainer, Nebojsa Jojic, Stephen W Scherer, Benjamin J Blencowe, and Brendan J Frey. RNA splicing. the human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218):1254806, January 2015.

69. Shigetsugu Hatakeyama. TRIM proteins and cancer. *Nat. Rev. Cancer*, 11(11):792–804, October 2011.

70. Ran Tao, Jun Li, Jiaojiao Xin, Jian Wu, Jing Guo, Liyuan Zhang, Longyan Jiang, Wu Zhang, Zhe Yang, and Lanjuan Li. Methylation profile of single hepatocytes derived from hepatitis B virus-related hepatocellular carcinoma. *PLoS One*, 6(5):e19862, May 2011.

71. Koichiro Kajiura, Kiyoshi Masuda, Takuya Naruto, Tomohiro Kohmoto, Miki Watabnabe, Mitsuhiro Tsuboi, Hiromitsu Takizawa, Kazuya Kondo, Akira Tangoku, and Issei Imoto. Frequent silencing of the candidate tumor suppressor TRIM58 by promoter methylation in early-stage lung adenocarcinoma. *Oncotarget*, 8(2):2890–2905, January 2017.

72. Rui Xu, Qiuyan Xu, Guanglei Huang, Xinhai Yin, Jianguo Zhu, Yikun Peng, and Jukun Song. Combined analysis of the aberrant epigenetic alteration of pancreatic ductal adenocarcinoma. *Biomed Res. Int.*, 2019:9379864, December 2019.

73. Ying Gan, Congcong Cao, Aolin Li, Haifeng Song, Guanyu Kuang, Binglei Ma, Quan Zhang, and Qian Zhang. Silencing of the TRIM58 gene by aberrant promoter methylation is associated with a poor patient outcome and promotes cell proliferation and migration in clear cell renal cell carcinoma. *Front Mol Biosci*, 8:655126, March 2021.

74. Min Su, Yujun Lin, Cheng Cui, Xiaohong Tian, and Laijun Lai. ERMAP is a B7 family-related molecule that negatively regulates T cell and macrophage responses. *Cell. Mol. Immunol.*, 18(8):1920–1933, August 2021.

75. H Xu, L Foltz, Y Sha, M R Madlansacay, C Cain, G Lindemann, J Vargas, D Nagy, B Harriman, W Mahoney, and P A Schueler. Cloning and characterization of human erythroid membrane-associated protein, human ERMAP. *Genomics*, 76(1-3):2–4, August 2001.

76. Jie Li, Xiao-Gang Liu, Rui-Liang Ge, Yu-Peng Yin, Yong-da Liu, Wan-Peng Lu, Mei Huang, Xue-Ying He, Jinghan Wang, Guoxiang Cai, Shu-Han Sun, and Ji-Hang Yuan. The ligation between ERMAP, galectin-9 and dectin-2 promotes kupffer cell phagocytosis and antitumor immunity. *Nat. Immunol.*, 24(11):1813–1824, November 2023.

77. Franz F Wagner, Joyce Poole, and Willy A Flegel. Scianna antigens including rd are expressed by ERMAP. *Blood*, 101(2):752–757, January 2003.

78. James D Orth and Mark A McNiven. Dynamin at the actin-membrane interface. *Curr. Opin. Cell Biol.*, 15(1):31–39, February 2003.

79. Christa Meisinger, Holger Prokisch, Christian Gieger, Nicole Soranzo, Divya Mehta, Dieter Rosskopf, Peter Lichtner, Norman Klopp, Jonathan Stephens, Nicholas A Watkins, Panos Deloukas, Andreas Greinacher, Wolfgang Koenig, Matthias Nauck, Christian Rimmbach, Henry Völzke, Annette Peters, Thomas Illig, Willem H Ouwehand,

Thomas Meitinger, H-Erich Wichmann, and Angela Döring. A genome-wide association study identifies three loci associated with mean platelet volume. *Am. J. Hum. Genet.*, 84(1):66–71, January 2009.

80. Alan B Cantor. Dynamin 3 and platelet size variation. *Blood*, 120(24):4666–4667, December 2012.

81. Jo-Anna Reems, Wenjing Wang, Ken Tsubata, Najla Abdurrahman, Birgitta Sundell, Marloes R Tijssen, Ellen van der Schoot, Franca Di Summa, Sunita Patel-Hett, Joseph Italiano, Jr, and Diana M Gilligan. Dynamin 3 participates in the growth and development of megakaryocytes. *Exp. Hematol.*, 36(12):1714–1727, December 2008.

82. Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Alexandra Bignell, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Mehrnaz Charkhchi, Alexander Cockburn, Luca Da Rin Fiorretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Cristi Guijarro, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Diego Marques-Coelho, José Carlos Marugán, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, José G Pérez-Silva, Ahamed Imran Abdul Salam, Nuno Saraiva-Agostinho, Helen Schuilenburg, Dan Sheppard, Swati Sinha, Botond Sipos, William Stark, Emily Steed, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Michal Szpak, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Elizabeth Wass, Natalie Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, John Tate, David Thybert, Stephen J Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Magali Ruffier, Fiona Cunningham, Sarah Dyer, Robert D Finn, Kevin L Howe, Peter W Harrison, Andrew D Yates, and Paul Flicek. Ensembl 2023. *Nucleic Acids Res.*, 51(D1):D933–D941, January 2023.

83. James Kelley, Lutz Walter, and John Trowsdale. Comparative genomics of natural killer cell receptor gene clusters. *PLoS Genet.*, 1(2):129–139, August 2005.

84. Ute Spiekerkoetter. Mitochondrial fatty acid oxidation disorders: clinical presentation of long-chain fatty acid oxidation defects before and after newborn screening. *J. Inherit. Metab. Dis.*, 33(5):527–532, October 2010.

85. Xiaofang Wang, Xuan Luo, Zhiyuan Wang, Yanghao Wang, Juan Zhao, and Li Bian. Identification of cancer stemness and M2 macrophage-associated biomarkers in lung adenocarcinoma. *Heliyon*, 9(9):e19114, September 2023.

86. José Wojnacki and Thierry Galli. A new actin-binding domain glues autophagy together. *J. Biol. Chem.*, 293(12):4575–4576, March 2018.

87. Sara R Rashkin, Rebecca E Graff, Linda Kachuri, Khanh K Thai, Stacey E Alexeeff, Maruta A Blatchins, Taylor B Cavazos, Douglas A Corley, Nima C Emami, Joshua D Hoffman, Eric Jorgenson, Lawrence H Kushi, Travis J Meyers, Stephen K Van Den Eeden, Elad Ziv, Laurel A Habel, Thomas J Hoffmann, Lori C Sakoda, and John S Witte. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.*, 11(1):4423, September 2020.

88. Pattarin Tangtanatakul, Chisanu Thumarat, Nusara Satproedprai, Punna Kunhapan, Tassamonwan Chaiyasung, Siriwan Klinchanhom, Yong-Fei Wang, Wei Wei, Jeerapat Wongshinsri, Direkrit Chiewchengchol, Pongsawat Rodsaward, Pintip Ngamjanyaporn, Thanitta Suangtamai, Surakameth Mahasirimongkol, Prapaporn Pisitkun, and Nattiya Hirankarn. Meta-analysis of genome-wide association study identifies FBN2 as a novel locus associated with systemic lupus erythematosus in thai population. *Arthritis Res. Ther.*, 22(1):185, August 2020.

89. Maria Paola Paronetto. Ewing sarcoma protein: a key player in human cancer. *Int. J. Cell Biol.*, 2013:642853, September 2013.

90. Stephanie C Huelga, Anthony Q Vu, Justin D Arnold, Tiffany Y Liang, Patrick P Liu, Bernice Y Yan, John Paul Donohue, Lily Shiue, Shawn Hoon, Sydney Brenner, Manuel Ares, Jr, and Gene W Yeo. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.*, 1(2):167–178, February 2012.

91. K V Datar, G Dreyfuss, and M S Swanson. The human hnRNP M proteins: identification of a methionine/arginine-rich repeat motif in ribonucleoproteins. *Nucleic Acids Res.*, 21(3):439–446, February 1993.

92. Yilin Xu, Xin D Gao, Jae-Hyung Lee, Huilin Huang, Haiyan Tan, Jaegyoon Ahn, Lauren M Reinke, Marcus E Peter, Yue Feng, David Gius, Kalliopi P Siziopikou, Junmin Peng, Xinshu Xiao, and Chonghui Cheng. Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. *Genes Dev.*, 28(11):1191–1203, June 2014.

93. Rebeca Martinez-Contreras, Philippe Cloutier, Lulzim Shkreta, Jean-François Fisette, Timothée Revil, and Benoit Chabot. hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.*, 623:123–147, 2007.

94. Kelsi O West, Haley M Scott, Sylvia Torres-Odio, A Phillip West, Kristin L Patrick, and Robert O Watson. The splicing factor hnRNP M is a critical regulator of innate immune gene expression in macrophages. *Cell Rep.*, 29(6):1594–1609.e5, November 2019.

95. Samuel E Harvey, Yilin Xu, Xiaodan Lin, Xin D Gao, Yushan Qiu, Jaegyoon Ahn, Xinshu Xiao, and Chonghui Cheng. Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT. *RNA*, 24(10):1326–1338, October 2018.

96. Jessica Sy Ho, Federico Di Tullio, Megan Schwarz, Diana Low, Danny Incarnato, Florence Gay, Tommaso Tabaglio, Jingxian Zhang, Heike Wollmann, Leilei Chen, Omer An, Tim Hon Man Chan, Alexander Hall Hickman, Simin Zheng, Vladimir Roudko, Sujun Chen, Alcida Karz, Musaddeque Ahmed, Housheng Hansen He, Benjamin D Greenbaum, Salvatore Oliviero, Michela Serresi, Gaetano Gargiulo, Karen M Mann, Eva Hernando, David Mulholland, Ivan Marazzi, Dave Keng Boon Wee, and Ernesto Guccione. HNRNPM controls circRNA biogenesis and splicing fidelity to sustain cancer cell fitness. *Elife*, 10, June 2021.

97. K Kevin Pfister, Paresh R Shah, Holger Hummerich, Andreas Russ, James Cotton, Azlina Ahmad Annuar, Stephen M King, and Elizabeth M C Fisher. Genetic analysis of the cytoplasmic dynein subunit families. *PLoS Genet.*, 2(1):e1, January 2006.

98. Yuanchao Xue, Yu Zhou, Tongbin Wu, Tuo Zhu, Xiong Ji, Young-Soo Kwon, Chao Zhang, Gene Yeo, Douglas L Black, Hui Sun, Xiang-Dong Fu, and Yi Zhang. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, 36(6):996–1006, December 2009.

99. Niroshika Keppetipola, Shalini Sharma, Qin Li, and Douglas L Black. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit. Rev. Biochem. Mol. Biol.*, 47(4):360–378, June 2012.

100. Panagiota Kafasla, Ian Mickleburgh, Miriam Llorian, Miguel Coelho, Clare Gooding, Dmitry Cherny, Amar Joshi, Olga Kotik-Kogan, Stephen Curry, Ian C Eperon, Richard J Jackson, and Christopher W J Smith. Defining the roles and interactions of PTB. *Biochem. Soc. Trans.*, 40(4):815–820, August 2012.

101. Kirsty Sawicka, Martin Bushell, Keith A Spriggs, and Anne E Willis. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.*, 36(Pt 4):641–647, August 2008.

102. Areum Han, Peter Stoilov, Anthony J Linares, Yu Zhou, Xiang-Dong Fu, and Douglas L Black. De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. *PLoS Comput. Biol.*, 10(1):e1003442, January 2014.

103. Athena Georgilis, Sabrina Klotz, Christopher J Hanley, Nicolas Herranz, Benedikt Weirich, Beatriz Morancho, Ana Carolina Leote, Luana D'Artista, Suchira Gallage, Marco Seehawer, Thomas Carroll, Gopuraja Dharmalingam, Keng Boon Wee, Marco Mellone, Joaquim Pombo, Danijela Heide, Ernesto Guccione, Joaquín Arribas, Nuno L Barbosa-Morais, Mathias Heikenwalder, Gareth J Thomas, Lars Zender, and Jesús Gil. PTBP1-Mediated alternative splicing regulates the inflammatory secretome and the pro-tumorigenic effects of senescent cells. *Cancer Cell*, 34(1):85–102.e9, July 2018.

104. Lianghua Shen, Sijia Lei, Bin Zhang, Shuaiguang Li, Luyuan Huang, Alexander Czachor, Mason Breitzig, Yuman Gao, Meiyan Huang, Xuemei Mo, Qing Zheng, Hanxiao Sun, and Feng Wang. Skipping of exon 10 in axl pre-mRNA regulated by PTBP1 mediates invasion and metastasis process of liver cancer cells. *Theranostics*, 10(13):5719–5735, April 2020.

105. Shuaiguang Li, Lianghua Shen, Luyuan Huang, Sijia Lei, Xingdong Cai, Mason Breitzig, Bin Zhang, Annan Yang, Wenzuo Ji, Meiyan Huang, Qing Zheng, Hanxiao Sun, and Feng Wang. PTBP1 enhances exon11a skipping in mena pre-mRNA to promote migration and invasion in lung carcinoma cells. *Biochim. Biophys. Acta Gene Regul. Mech.*, 1862(8):858–869, August 2019.

106. Monika Fuxreiter, Istvan Simon, and Sarah Bondos. Dynamic protein–DNA recognition: beyond what can be seen. *Trends Biochem. Sci.*, 36(8):415–423, August 2011.

107. Xin Guo, Martha L Bulyk, and Alexander J Hartemink. Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. *Pac. Symp. Biocomput.*, pages 104–115, 2012.

108. Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. Introducing protein intrinsic disorder. *Chem. Rev.*, 114(13):6561–6588, July 2014.

109. K Torigoe, S Ushio, T Okura, S Kobayashi, M Taniai, T Kunikata, T Murakami, O Sanou, H Kojima, M Fujii, T Ohta, M Ikeda, H Ikegami, and M Kurimoto. Purification and characterization of the human interleukin-18 receptor. *J. Biol. Chem.*, 272(41):25737–25742, October 1997.

110. Matija Hedl, Shasha Zheng, and Clara Abraham. The IL18RAP region disease polymorphism decreases IL-18RAP/IL-18R1/IL-1R1 expression and signaling through innate receptor-initiated pathways. *J. Immunol.*, 192(12):5924–5932, June 2014.

111. Naotaka Tsutsumi, Takeshi Kimura, Kyohei Arita, Mariko Ariyoshi, Hidenori Ohnishi, Takahiro Yamamoto, Xiaobing Zuo, Katsumi Maenaka, Enoch Y Park, Naomi Kondo, Masahiro Shirakawa, Hidehito Tochio, and Zenichiro Kato. The structural basis for receptor recognition of human interleukin-18. *Nat. Commun.*, 5:5340, December 2014.

112. Enza Ferrero, Nicola Lo Buono, Simona Morone, Rossella Parrotta, Cecilia Mancini, Alfredo Brusco, Alice Giacomino, Stefania Augeri, Antonio Rosal-Vela, Sonia García-Rodríguez, Mercedes Zubiaur, Jaime Sancho, Alessandra Fiorio Pla, and Ada Funaro. Human canonical CD157/Bst1 is an alternatively spliced isoform masking a previously unidentified primate-specific exon included in a novel transcript. *Sci. Rep.*, 7(1):15923, November 2017.

113. Fabio Malavasi, Silvia Deaglio, Ada Funaro, Enza Ferrero, Alberto L Horenstein, Erika Ortolan, Tiziana Vaisitti, and Semra Aydin. Evolution and function of the ADP ribosyl cyclase/CD38 gene family in physiology and pathology. *Physiol. Rev.*, 88(3):841–886, July 2008.

114. Arne Jacobs and Kathryn R Elmer. Alternative splicing and gene expression play contrasting roles in the parallel phenotypic evolution of a salmonid fish. *Mol. Ecol.*, 30(20):4955–4969, October 2021.

115. Auinash Kalsotra and Thomas A Cooper. Functional consequences of developmentally regulated alternative splicing. *Nat. Rev. Genet.*, 12(10):715–729, September 2011.

116. Melanie A Huntley, Karpagam Srinivasan, Brad A Friedman, Tzu-Ming Wang, Ada X Yee, Yuanyuan Wang, Josh S Kaminker, Morgan Sheng, David V Hansen, and Jesse E Hanson. Genome-Wide analysis of differential gene expression and splicing in excitatory neurons and interneuron subtypes. *J. Neurosci.*, 40(5):958–973, January 2020.

117. Mary E Grantham and Jennifer A Brisson. Extensive differential splicing underlies phenotypically plastic aphid morphs. *Mol. Biol. Evol.*, 35(8):1934–1946, August 2018.

118. Michael Sammeth, Sylvain Foissac, and Roderic Guigó. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4(8):e1000147, August 2008.

119. Dongwook Kim, Manu Shivakumar, Seonggyun Han, Michael S Sinclair, Young-Ji Lee, Yonglan Zheng, Olufunmilayo I Olopade, Dokyoon Kim, and Younghee Lee. Population-dependent intron retention and DNA methylation in breast cancer. *Mol. Cancer Res.*, 16(3):461–469, March 2018.

120. Anil Aktas Samur, Mariateresa Fulciniti, Herve Avet-Loiseau, Michael A Lopez, Sanika Derebail, Jill Corre, Stephane Minvielle, Florence Magrangeas, Philippe Moreau, Kenneth C Anderson, Giovanni Parmigiani, Mehmet K Samur, and Nikhil C Munshi. In-depth analysis of alternative splicing landscape in multiple myeloma and potential role of dysregulated splicing factors. *Blood Cancer J.*, 12(12):171, December 2022.

121. Robert K Bradley, Jason Merkin, Nicole J Lambert, and Christopher B Burge. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS Biol.*, 10(1):e1001229, January 2012.

122. Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, 11(5):345–355, May 2010.

123. Wei Jiang and Liang Chen. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Comput. Struct. Biotechnol. J.*, 19:183–195, 2021.

124. Liliana Florea, Li Song, and Steven L Salzberg. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. *F1000Res.*, 2:188, September 2013.

125. Julianne K David, Sean K Maden, Mary A Wood, Reid F Thompson, and Abhinav Nellore. Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. unpublished, March 2022.

126. Sam Kovaka, Shujun Ou, Katharine M Jenike, and Michael C Schatz. Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nat. Methods*, 20(1):12–16, January 2023.

127. Nicolas Brucato, Mathilde Andr, Roxanne Tsang, Lauri Saag, Jason Kariwiga, Kylie Sesuki, Teppsy Beni, William Pomat, John Muke, Vincent Meyer, Anne Boland, Jean-Franois Deleuze, Herawati Sudoyo, Mayukh Mondal, Luca Pagani, Irene Gallego Romero, Mait Metspalu, Murray P Cox, Matthew Leavesley, and Franois-Xavier Ricaut. Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. *Molecular Biology and Evolution*, 38(11):5107–5121, November 2021.

128. Diego Garrido-Martín, Beatrice Borsari, Miquel Calvo, Ferran Reverter, and Roderic Guigó. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.*, 12(1):727, February 2021.

129. Lin Chen, Kun Chen, Yucai Hong, Lifeng Xing, Jianjun Zhang, Kai Zhang, and Zhongheng Zhang. The landscape of isoform switches in sepsis: a multicenter cohort study. *Sci. Rep.*, 12(1):10276, June 2022.

130. Yun Zhang, Katherine M Weh, Connor L Howard, Jean-Jack Riethoven, Jennifer L Clarke, Kiran H Lagisetty, Jules Lin, Rishindra M Reddy, Andrew C Chang, David G Beer, and Laura A Kresty. Characterizing isoform switching events in esophageal adenocarcinoma. *Mol. Ther. Nucleic Acids*, 29:749–768, September 2022.

131. Diego Marques-Coelho, Lukas da Cruz Carvalho Iohan, Ana Raquel Melo de Farias, Amandine Flaig, Brainbank Neuro–CEB Neuropathology Network, Jean-Charles Lambert, and Marcos Romualdo Costa. Differential transcript usage unravels gene expression alterations in alzheimer's disease human brains. *NPJ Aging Mech Dis*, 7(1):2, January 2021.

132. Kaur Alasoo, Julia Rodrigues, John Danesh, Daniel F Freitag, Dirk S Paul, and Daniel J Gaffney. Genetic effects on promoter usage are highly context-specific and contribute to complex traits. *Elife*, 8, January 2019.

133. Emilie Lalonde, Kevin C H Ha, Zibo Wang, Amandine Bemmo, Claudia L Kleinman, Tony Kwan, Tomi Pastinen, and Jacek Majewski. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.*, 21(4):545–554, April 2011.

134. Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, Rui Mei, Alexander E Urban, Stephen B Montgomery, Douglas F Levinson, and Daphne Koller. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, 24(1):14–24, January 2014.