Profiling genetically driven alternative splicing across the Indonesian archipelago

Authors

Neke Ibeh, Pradiptajati Kusuma,
Chelzie Crenna Darusallam, Safarina G. Malik,
Herawati Sudoyo, Davis J. McCarthy,
Irene Gallego Romero

Correspondence

irene.gallego@svi.edu.au

We have carried out a global splicing analysis across 115 Indonesian samples spanning the genetic ancestry cline that characterizes the region. We identify over 6,000 sQTLs, some of which are driven by Papuan-like genetic ancestry and only show partial overlap with sQTL datasets derived from European populations.

ARTICLE

Profiling genetically driven alternative splicing across the Indonesian archipelago

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

Summary

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 characterized across human populations, and it is unclear how genetic and environmental variation contribute to AS. 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 an event-based statistical model, we detected over 1,500 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 reveal 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 toward 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,[4,7–10] although the degree to which alternative isoforms are functional remains unclear.[11–13] Nevertheless, isoform-level transcriptome analyses have revealed that splicing can play significant roles in the cellular response to environmental cues, including immune pressures.[14–20] In humans, this natural variation in AS has been highlighted as a phenomenon influencing complex traits and disease prevalence.[21]

AS is tightly regulated through an intricate protein-RNA interaction network comprised of *cis* regulatory elements and *trans*-acting factors.[22] Genetic polymorphisms can alter these splicing regulatory elements and splice site usage thereby influencing gene expression and protein products.[23] Splicing quantitative trait locus (sQTL) analysis has become a leading method for explicating 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,[24] Alzheimer disease,[25] and schizophrenia.[26] The GTEx Consortium has recently characterized sQTLs in over 50 human healthy tissues,[9] 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 characterize 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 seven different global populations, population differences in genetic ancestry explained 25% of differences between individuals in genes expressing at least two transcripts,[27] but expression differences, rather than splicing ones, accounted for the bulk of this effect. However, LCLs are known to trend toward a homogeneous pattern of gene expression that blunts the effect of inter-individual diversity,[28] 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.[29]

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 Park et al.[30]). 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[20] 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.[31]

Here, we focus on 115 previously described whole-blood samples drawn from three traditional Indonesian populations[32,33] that span the genetic ancestry cline that makes up the region.[34] We explore the prevalence and functional significance of AS variation across this region of the world, working toward 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 identify 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 AS 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 AS in human populations.

Subjects, material, and methods

Dataset description and ethics statement
We used a previously published set of 115 (all male) matched DNA and RNA data from three distinct Indonesian populations[32] (available at the European Genome-Phenome Archive: EGAS00001003671). The samples span the main west-east axis of genetic diversity in the Indonesian archipelago[34,35]: (1) the inhabitants of Mentawai (MTW, $n = 48$), a small barrier island in western Indonesia, are of West Island Southeast Asian-like genetic ancestry; (2) the Korowai (KOR, $n = 19$), from western New Guinea Island, are of Papuan-like genetic ancestry; while (3) 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.[33] 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).

RNA-seq processing
For all 115 samples, raw RNA-sequencing (RNA-seq) reads were assessed with FastQC (v.0.11.9)[36] and pre-processed with Trimmomatic v0.36,[37] removing leading and trailing bases with a Phred score below 20 prior to any further analysis. RNA-seq reads were aligned to GRCh38 Ensembl release 110 using default STAR (v.2.7) settings[38] with the exception of -alignEndsType EndToEnd to remove soft-clipping of the reads and -sjdbOverhang 100 for optimal splice junction overhang length.

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 cases of retained introns [RIs]) 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 AS events, namely, skipped exons (SEs), RIs, mutually exclusive exons (MXEs), alternative 5′ splice sites (A5SSs), and alternative 3′ splice sites (A3SSs). 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 AS between the three groups, we corrected PSI values for batch effects and other technical confounders using fractional regression.[29,41] 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[32] estimates of the proportion of CD8T, CD4T, natural killer (NK), B cells, monocytes, and granulocytes in each sample. Using these corrected PSI values, differentially spliced events were identified by computing Δ PSI 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 $|\Delta$ PSI $| \geq 0.1$ and FDR < 0.05.

In parallel, we conducted intron-based quantification of splicing variation (necessary for sQTL mapping) using LeafCutter,[42] 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 500 kb. 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[42] 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[43] *sashimi_plot* tool. Read densities for all exons were quantified using RPKM units.[44] 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 a 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*.[45] Prior to isoform switch testing, we filtered out completely unused isoforms, single-isoform genes, and 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[46] *removeBatchEffect* function to the isoform abundance matrix. Statistical identification of isoform switches was conducted via *DEXseq* using default settings.[47] Isoforms were considered differentially switched and retained for further analysis if the 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 Korowai versus Mentawai is one that is used more in Mentawai when compared to Korowai. 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,[48] IUPred2A,[49] SignalP,[50] and Pfam,[51] 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.[45] 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
Aligned reads (GRCh38 Ensembl release 110) were used as input to LeafCutter[42] 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[52] to test for an association between variants and intron ratios within a *cis*-region of $\pm$ 1 Mb of the intron cluster using 10,000 adaptive permutations. After observing that the vast majority (89%) of our detected associations occurred within 250 kb of the intron clusters, we restricted our analyses to this 250 kb window. We controlled latent sources of variation using covariates identified with the probabilistic estimation of expression residuals (PEER) method.[53] 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 five genotyping principal components (PCs) 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[9] (*n* = 670) and BLUEPRINT[8] (*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 ± 1 Mb of each gene's transcription start site (TSS). For consistency, GTEx and BLUEPRINT sQTL sets were restricted to variants within 250 kb of the TSS, and coordinates for the BLUEPRINT data were converted from hg19 to hg38 using the R package *liftOver* v1.26.0.[54] 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*.[55] 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,[56] fitting a linear model $V = \alpha \times PAP + \beta$ for each variant. V is the genotype vector and PAP is a covariate that represents the number of alleles assigned to Papuan-like genetic ancestry. This analysis was conducted using the 73 samples (30 Mentawai, 29 Sumba, and 14 Korowai) with available 30× depth whole-genome sequencing data.[33] 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),[57] 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[58] (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) 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.[59] Specifically, we used DeepCLIP's 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.[60,61] 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).[59]

Colocalization analysis with genome-wide association study hematological traits

Colocalization analyses were performed between sQTLs and 13 hematological traits using coloc v5.2.3.[62] We tested for colocalization between our detected permutation-based sQTLS and genome-wide association study (GWAS) loci from a *trans*-ethnic study that included East Asian populations.[63] We assumed a prior probability that a SNP is associated with (1) the GWAS trait (p1, default = 1×10^{-4}), (2) AS (p2, default = 1×10^{-4}), and (3) both the GWAS trait and AS (p12, default = 1×10^{-6}). We identified robust colocalization with the default threshold of CCV > 0.8 and a ratio of CCV/DCV > 5, where CCV and DCV stand for common causal variant and distinct causal variant, respectively. 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 red blood cell density. All GWAS summary statistics were obtained from the GWAS catalog. The linkage disequilibrium r^2 values were calculated from our phased genotype data (115 samples).

Results

Characterizing differential alternative splicing events between Indonesian populations

To characterize the splicing landscape across the Indonesian archipelago, we carried out AS analyses using paired-end RNA-seq data generated from 115 previously described male whole-blood samples.[32] 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[33]), and the Korowai of New Guinea Island (of Papuan-like genetic ancestry; Figure 1A).

Across the three populations, we observed 38,611 AS events in the Korowai, while the Mentawai and Sumba groups had 39,728 and 39,769 AS events, respectively (Figure S1A). 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 AS events were seen in all three island groups, 1,337 events (3.3%) were shared between only two of the 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, Figures S1B–S1F), although this is likely attributable to our increased power in these two groups relative to the Korowai. To gain further insight into the detected AS genes across all populations, as well as the genes that contain the 1,002 population-specific AS patterns (434 events in Mentawai, 447 events in Sumba, and 121 events in Korowai), we tested for their enrichment against GO and KEGG pathways (Tables S1 and 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 AS analysis. We identified 419 significant ($|\Delta \text{PSI}| \geq 0.1$ and FDR < 0.05) differential AS events between Mentawai and Sumba, 515 events between Korowai and Mentawai, and 738 events between Korowai and Sumba (Figures 1C and 1D).

These differences are in line with previously reported gene expression and methylation patterns across these groups, with the Korowai vs. Sumba comparison yielding the most differentially expressed genes and differentially methylated probes, closely followed by Korowai vs. Mentawai.[32] 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. Given the ancestry proportions across these groups, we would expect to observe the highest number of differential AS events between the Korowai and Mentawai populations instead of the Korowai and Sumba populations. We hypothesize that there are complex gene-by-environment interactions at play, where differences in the local environments at each sampling site might be contributing to these patterns, including the shared subsistence agriculture lifestyle of the inhabitants of both Mentawai and Sumba.[32]

We found that differentially alternatively spliced (DAS) genes between Korowai and Sumba are related to mRNA processing (e.g., *SLTM*, *BARD1* [MIM: 601593], *DYRK1A* [MIM: 600855], *CDK11A* [MIM: 116951], *QKI* [MIM: 609590], and *RNPS1* [MIM: 606447]), splicing regulation (e.g., *RBM23*, *TRA2B* [MIM: 602719], *TMBIM6* [MIM: 600748], *MBNL1* [MIM: 606516], *HNRNPH1* [MIM: 601035], *RNPS1*, and *PTBP2* [MIM: 608449]), and cell cycle signaling (e.g., *MRE11* [MIM: 600814], *BRIP1* [MIM: 605882], *DONSON* [MIM: 611428], *ZWILCH* [MIM: 609984], and *CDC14B* [MIM: 603505]), 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* [MIM: 601671], *ATF7IP* [MIM: 613644], *MTA1* [MIM: 603526], *HDAC9* [MIM: 606543], and *SF1* [MIM: 601516]) and coactivator (e.g., *MED24* [MIM: 607000], *KMT2C* [MIM: 606833], *SMARCA2* [MIM: 600014], *MED12* [MIM: 300188], and *ACTN1* [MIM: 102575]) 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.

Characterization of Indonesian *cis*-sQTLs in whole blood

To understand the genetic regulation of AS 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 (subjects, material, and methods). Splice ratios from a total of 49,122 splice clusters (i.e., overlapping introns that share a splice donor or acceptor site) were included in the sQTL analysis. We detected a total of 6,077 *cis*-sQTLs affecting 1,977 genes
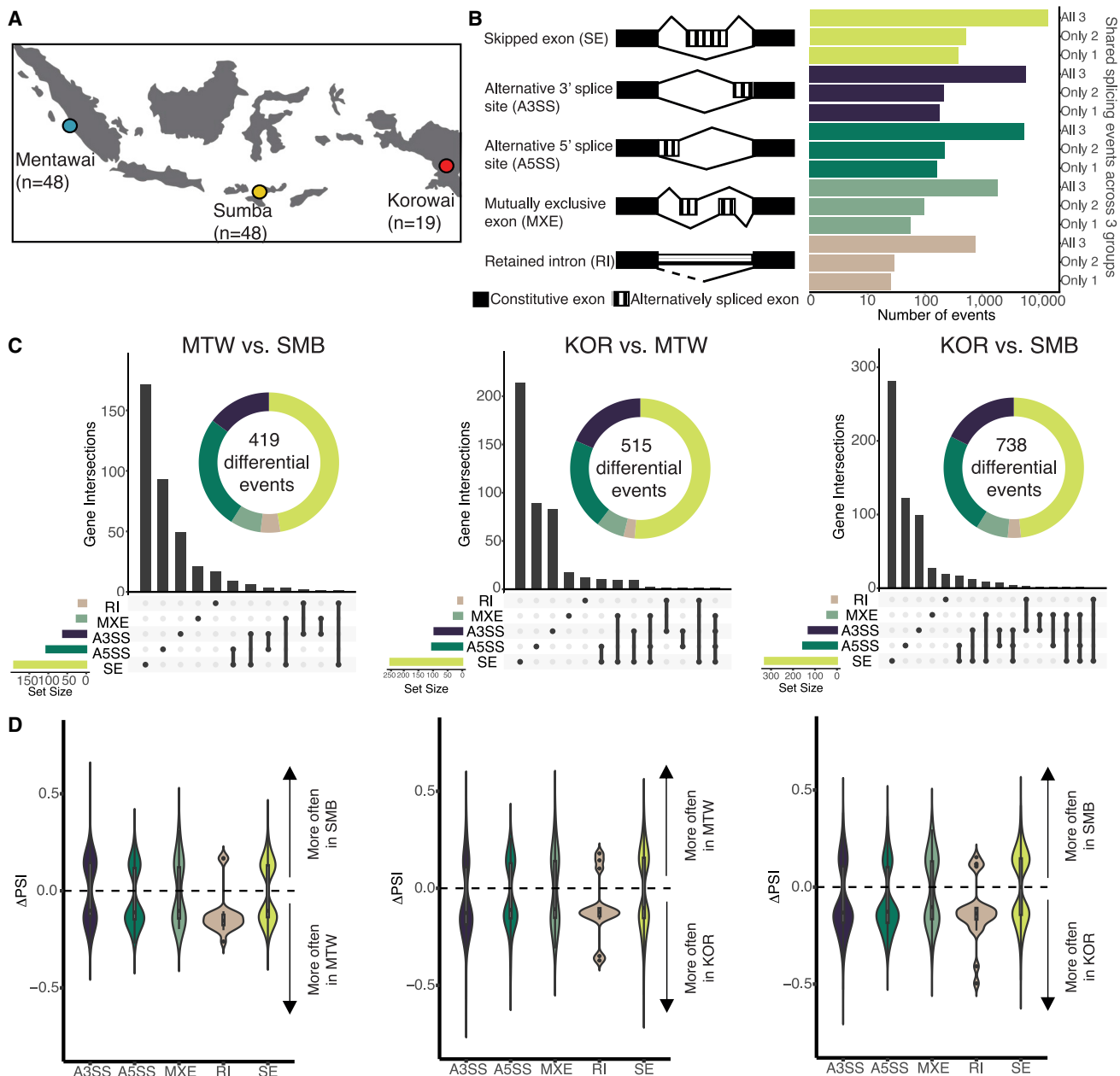
Figure 1. Characterization of differential AS across Indonesian populations

(A) Geographical location of the three sampling sites ($n = 115$).

(B) Schematic diagram of the five AS events profiled with SUPPA2. Horizontal bar plots correspond to the event categories to the left and illustrate the sharing of AS events across the three island groups (i.e., the splicing event either occurs in all three groups, 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 ($|\Delta \text{PSI}| \geq 0.1$, FDR < 0.05) are shown.

(D) For each population pair, the distributions of Δ PSI 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.

at an FDR level of 0.01, comprising 3,658 unique SNPs (Tables S4 and S5). Positional enrichment of significant sQTL SNPs confirmed that sQTLs are located within or near the introns that they regulate, with 56% of SNPs occurring within 5 kb of the LeafCutter-defined splicing cluster (Figure 2A). Additionally, 67% of sQTLs are located

in non-coding genomic regions (Figure S2). This observation is consistent with previous research suggesting that most variants that fall in close proximity to splice junctions influence splicing regulatory functions.[30,64]

GO enrichment analysis of sGenes (genes with a significant sQTL) revealed enrichment for numerous
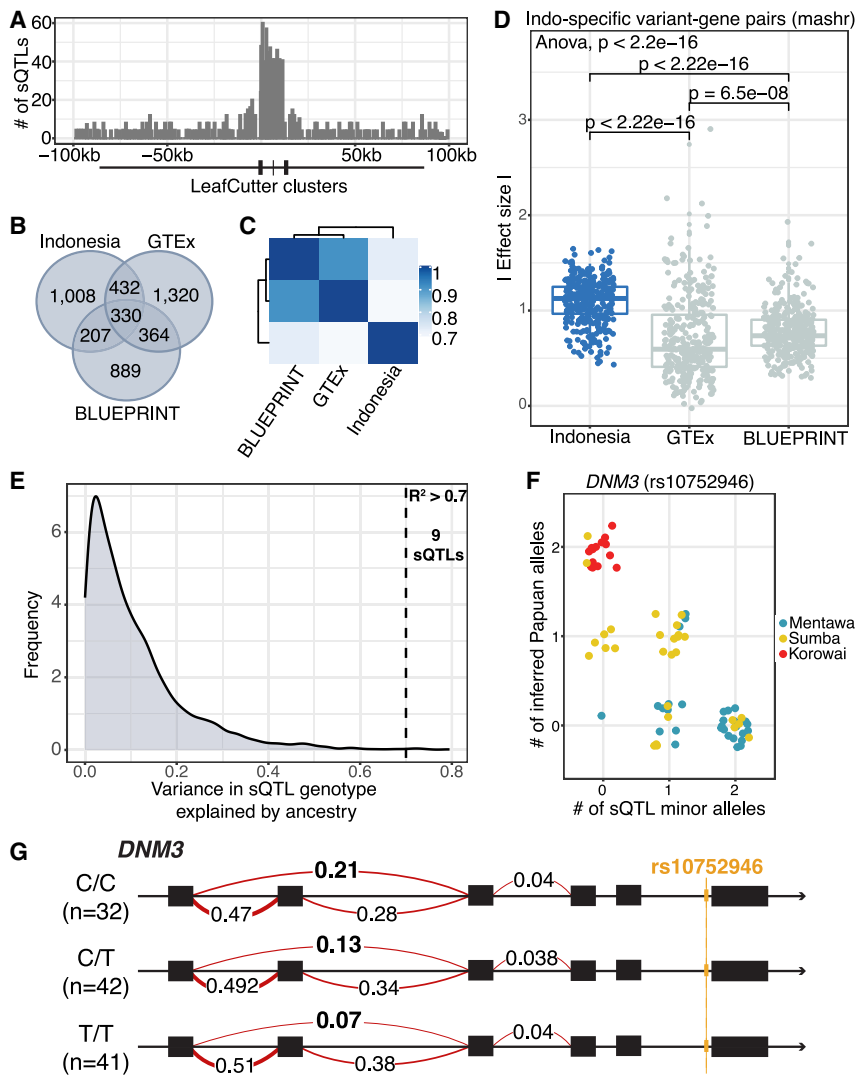
Figure 2. Indonesian *cis*-sQTL characteristics

(A) Meta-cluster representation of the position of sQTLs across LeafCutter-defined intron clusters.

(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 nine sQTLs (across 11 sGenes) 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.

(G) Genotype plot of the *DNM3* Papuan-driven sQTL event. For this AS event, the junction with the most significant association to the SNP (rs10752946) is depicted with bold numbers across all three genotypes, and these numbers represent the proportion of reads spanning the junction (i.e., normalized count values). The thickness of each intron-spanning curve is proportional to the displayed normalized counts.

S4B). *TRIM58* is additionally implicated in malignancies such as lung cancer, liver cancer, and pancreatic ductal adenocarcinoma[65–68] and is involved in innate immunity and cell proliferation.[65,69] *ERMAP* is a B7 family immune regulator that has been found to promote the phagocytosis of tumor cells[70–72] and encodes the protein responsible for the Scianna blood group system.[73]

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[9] ($n = 670$) and BLUEPRINT[8] ($n = 197$). We found that sQTL effects are largely shared between the two European datasets, but they are not shared to the same extent with the Indonesian data. Out of 4,550 unique sGenes (among 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 1,008 (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 generalizability 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

immune-related pathways, including immune-response-regulating signaling (e.g., *THEMIS2* [MIM: 617856], *ERMAP* [MIM: 609017], *GBP2* [MIM: 600412], *VAV3* [MIM: 605541], *FCRL3* [MIM: 606510], *AIM2* [MIM: 604578], and *FCGR3A* [MIM: 146740]), immune-response-activating signaling (e.g., *DENND1B* [MIM: 613292], *MAPKAPK2* [MIM: 602006], *CR1* [MIM: 120620], *NLRC4* [MIM: 606831], *NAGK* [MIM: 606828], *IFIH1* [MIM: 606951], and *PRKCD* [MIM: 176977]), phagocytosis (e.g., *VAV3*, *NCF2* [MIM: 608515], *C4BPA* [MIM: 120830], *DYSF* [MIM: 603009], *MARCO* [MIM: 604870], and *CD302* [MIM: 612246]), regulation of innate immune response (e.g., *GBP2*, *ADAR* [MIM: 146920], *FCRL3*, *IFI16* [MIM: 147586], *AIM2*, and *CFH* [MIM: 134370]), and interleukin-1 beta production (e.g., *IFI16*, *AIM2*, *NLRC4*, *CASP8* [MIM: 601763], *GHRL* [MIM: 605353], and *CX3CR1* [MIM: 601470]), highlighting the impact of splicing on immune-related genes (Figure S3; Table S6). Indeed, some of the strongest sQTL signals are present in immune-related genes such as *TRIM58* (MIM: 620527) and *ERMAP* (Figures S4A and

multivariate adaptive shrinkage model, mashr. Using an LFSR 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 naive 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). These Indonesia-specific sQTLs identified by mashr had higher minor allele frequencies (MAFs) across the Indonesian samples compared to the European samples ($p = 1.52e$-08, $p = 2.96e$-09, Figure S5), while the shared sQTLs had equally high MAFs between both populations ($p = 0.077$, $p = 0.081$). To determine whether this Indonesia-specific signature is driven by Indonesia-specific variants, we calculated the proportion of Indonesian sQTLs that did not occur in the European sQTL databases because the MAF of the variant was too low in the European samples. GTEx and BLUEPRINT include variants with MAFs ≥ 0.01, while we tested those with an MAF above 0.05 to account for diminished power as a consequence of our sample size. Across all Indonesia-specific sQTLs, we found that 87/312 (28%) were not tested in the European datasets due to a European MAF that is below 0.01. Notably, for these 87 SNPs, the mean MAF difference between our Indonesian samples and individuals of non-Finnish European ancestry (gnomAD[74]) is 0.314. In line with previous eQTL analyses carried out on this set of samples,[33] we also 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 (e.g., *IRF1* [MIM: 147575] and *SIRPB1* [MIM: 603889]) and oncogene pathways (e.g., *RAP1A* [MIM: 179520], *RAP1B* [MIM: 179530], *FES* [MIM: 190030], and *DNM3* [MIM: 611445]). 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 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.[29,31]

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 LA in modern Indonesians. Specifically, using LA haplotype information for each significant sQTL, we investigated whether there was a correlation between the inferred ancestral source of the genotype and gene splicing (subjects, material, and methods). We found nine sQTLs (across 11 sGenes) 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[75] and has been implicated in various functions related to the development and colony formation of megakaryocytes.[76–78] Human *DNM3* contains 21 exons with at least five of its exons being alternatively spliced to generate several isoforms.[79] The Papuan-driven sQTL event that we identified for *DNM3* is a skipping event of exon 17 ($\beta = 0.55$, q value = 3.5×10^{-4}) associated with the SNP rs10752946, which lies within intron 20 of *DNM3* (Figures 2F, 2G, and 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 AS event that excludes exon 17, resulting in a loss of four amino acids that 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.[63] 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* (MIM: 601437), which encodes the neonatal Fc receptor (FcRn) colocalized with red blood cell density, hemoglobin concentration, and hematocrit, while *MAEA* (MIM: 606801), a macrophage erythroblast attacher, colocalized with mean corpuscular hemoglobin, mean corpuscular volume, and red blood cell density. We also found that the LA-driven sQTL at *DNM3* significantly colocalizes with hemoglobin concentration (H4 = 0.837, H4/H3 = 8.445), a trait that refers to the amount of hemoglobin protein present in a specific volume of blood (Figure 3B).
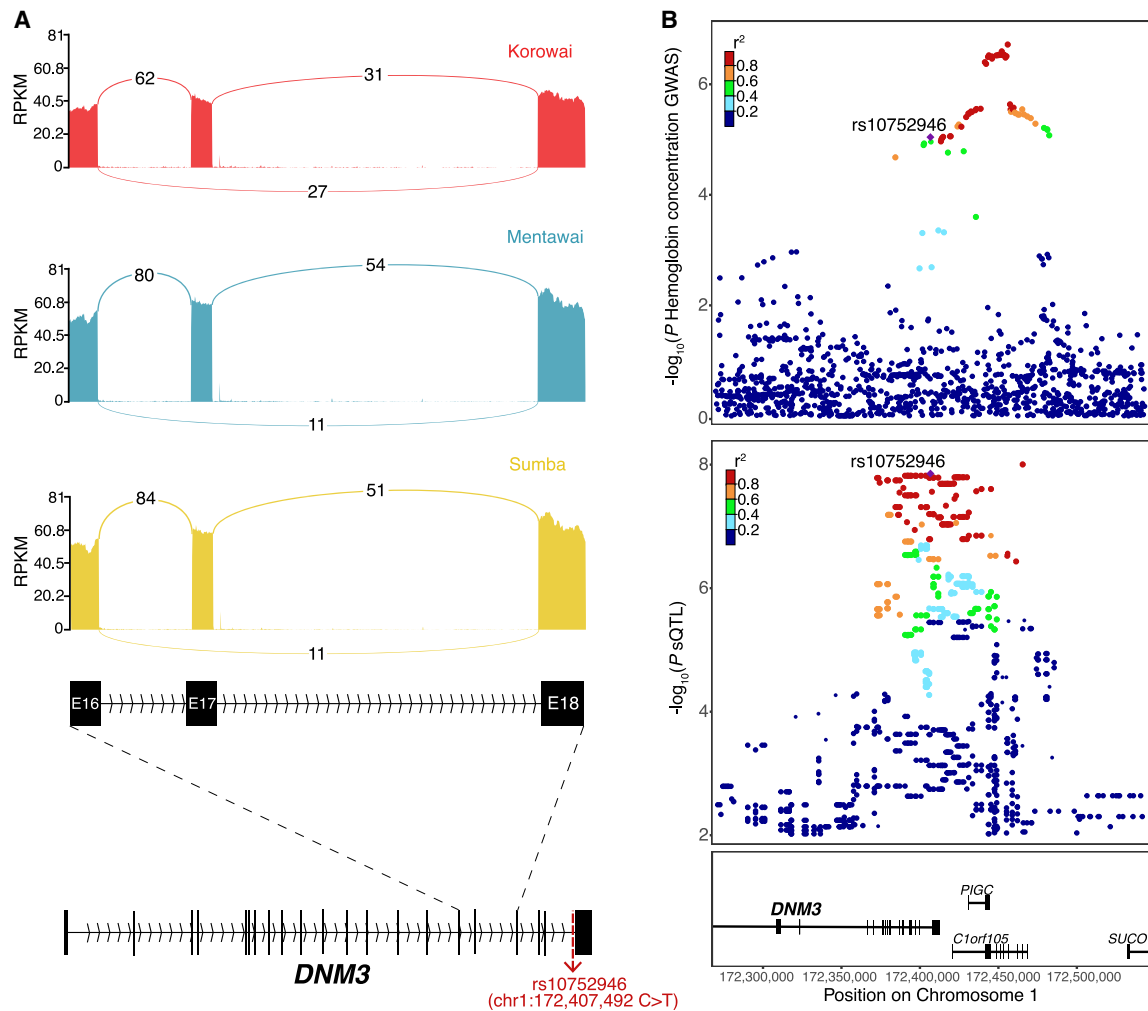
Figure 3. Splicing regulation in *DNM3*
(A) Sashimi plot illustrating the skipped exon event (exon 17) in *DNM3*. Isoform expression was summarized using the MISO framework.[43] The numbers over each curved line represent the junction spanning reads. The canonical isoform for *DNM3* is provided. The labels E16, E17, and E18 represent exons 16, 17, and 18, respectively.
(B) Statistically significant GWAS colocalization between the LAI-driven sQTL in *DNM3* (rs10752946) and the GWAS hemoglobin concentration trait[63] measured in East Asian populations.

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 in ClinVar. We considered two sets of SNPs: (1) all lead sQTL SNPs within the gene body and (2) all sQTL SNPs falling within splice junctions. Across both SNP sets, the analysis yielded 120 total hits with 116 of these having benign or likely benign status, three variants of uncertain significance (VUSs), and one variant with a GWAS disease trait classification (Table S8). The three VUSs occur in *KLRC4* (rs139613925, in-sample alternative allele frequency 0.481 [MIM: 602893]), *HADHB* (rs2303893, in-sample alternative allele frequency 0.366 [MIM: 143450]), and *DENND3* (rs2289001, in-sample alternative allele frequency 0.307 [MIM: 617503]). The protein encoded by *KLRC4* belongs to the natural killer group 2 family of receptors. These receptors are primarily expressed on NK cells and some sub-sets of T cells, where they play important roles in the recognition and regulation of immune responses.[80] The *HADHB* gene product, MTP-beta, plays a central role in the breakdown of long-chain fatty acids within mitochondria.[81] *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.[82,83] The GWAS variant, rs2763979 (in-sample alternative allele frequency 0.468), occurs in *HSPA1B* (MIM: 603012) and has reported associations with cancer (p value = 2×10^{-14}, European population, $n = 475,312$)[84] and systemic lupus erythematosus (SLE [MIM: 152700], p value = 6×10^{-6}, Thai population, $n = 4,088$).[85]

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
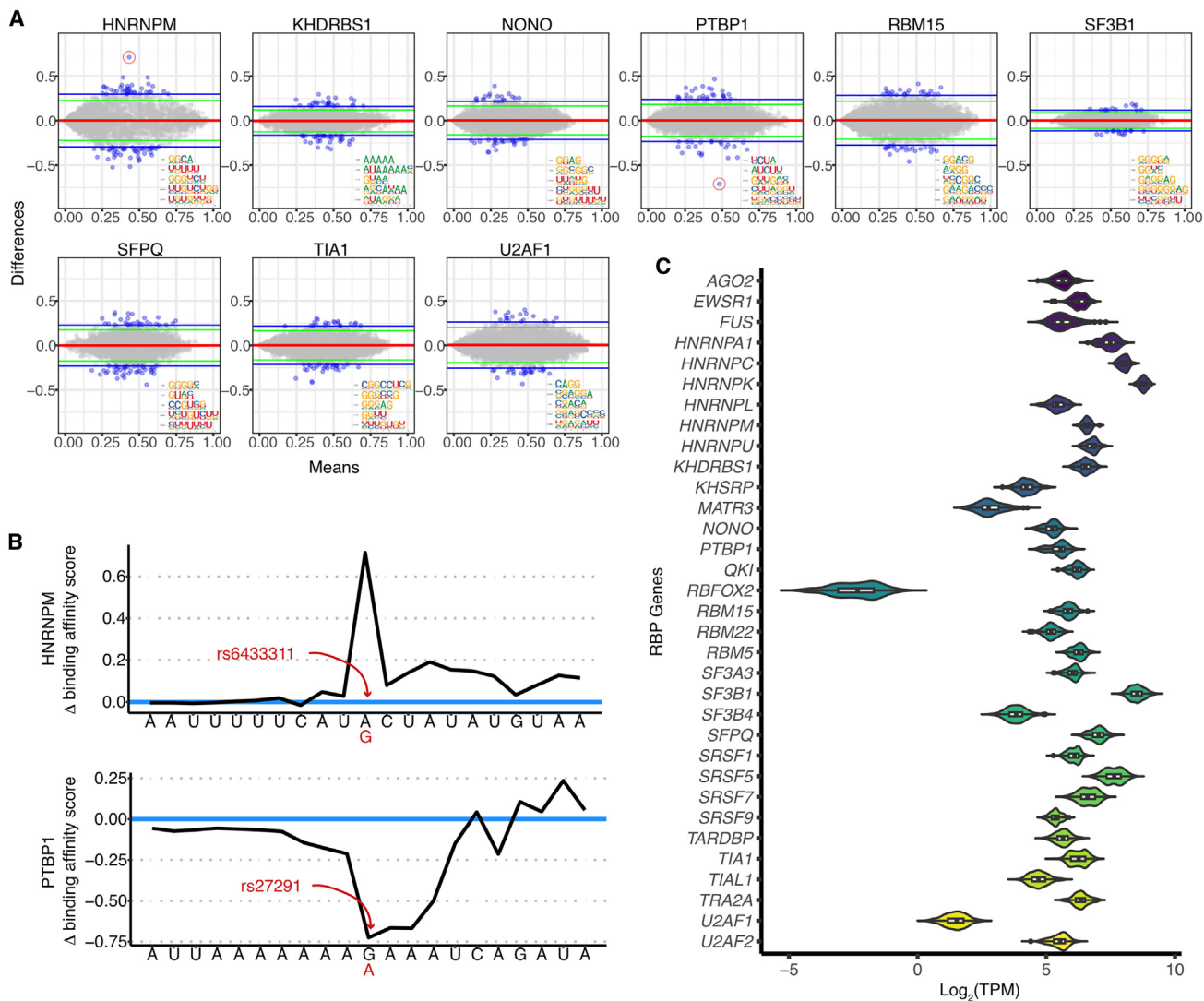
Figure 4. sQTL impact on RBP binding affinity
(A) Mean-difference plots for 9 of the 33 splicing-RBPs. Each point represents a pair of 21-bp-long sequences centered on the lead sQTL SNP (ref/alt alleles). The y axis depicts the difference between the DeepCLIP reference sequence score and the DeepCLIP variant sequence score, while the x axis depicts the average between these two scores. Horizontal green and blue lines represent the 95% and 99% confidence intervals (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 RBPs.

binding as a possible mechanism of action. We employed DeepCLIP's pre-trained models for 33 splicing-associated RBPs with well-documented roles in splicing regulation and spliceosome activity[60] and assessed the impact on their binding abilities across all 3,658 significant lead SNPs within 250 kb of the nearest splice site, adding 10-bp 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 (File S2). These binding affinity scores, ranging from 0 to 1, can be interpreted as the strength of RBP binding.[59] For each RBP, we observe that most sQTL SNPs are not predicted to significantly increase or decrease RNA binding af-

finity, as their difference scores (reference score − alternative score) tend to fall within the range (99% CI) of likely values (Figures 4A and S6). However, globally, we find that 59.6% of all sQTLs fall outside of the 99% CI for at least one RBP, suggesting that these sQTls might impact the binding affinity of at least one of these 33 splicing RBPs (File S2). For RBPs with very uniform motifs, such as EWSR1 (Ewing sarcoma protein),[86] 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 (Figure S6).

Conversely, we also observe RBPs with only one or two outliers, suggesting a biologically significant impact of
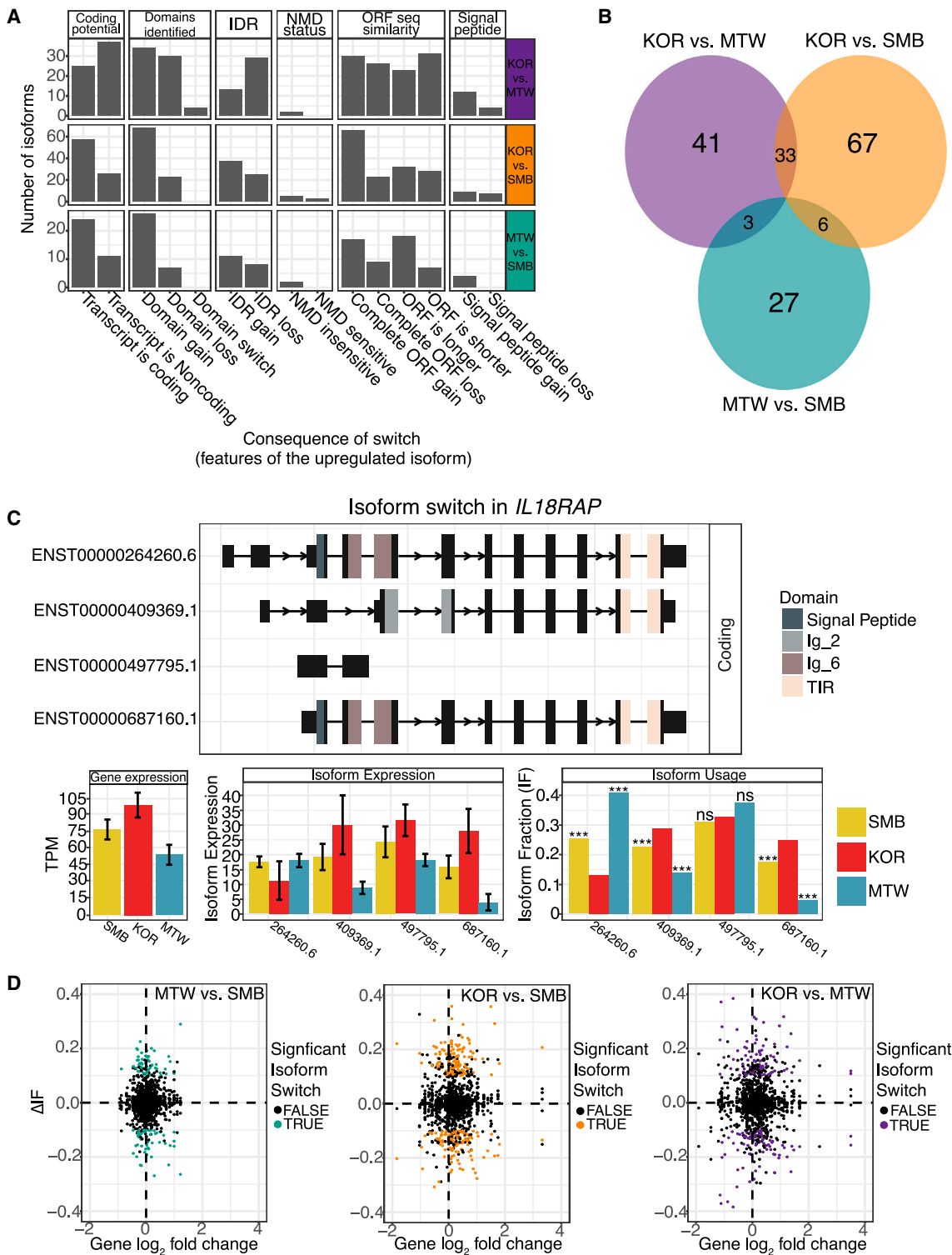
Figure 5. Global isoform switching analysis and functional consequences

(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. KOR, Korowai; MTW, Mentawai; and SMB, Sumba.

(B) Venn diagram of the number of genes with overlapping isoform switching events and functional consequences (FDR < 0.05) across all pairwise comparisons.

(legend continued on next page)

the sQTL SNP on RBP activity (Figures 4A and 4B). 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 *DYNC1I2* (MIM: 603331); 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, Figure S7A) 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.[87–89] HNRNPM is a member of the heterogeneous nuclear ribonucleoprotein family involved in RNA processing, including pre-mRNA splicing and mRNA transport,[90] and has been found to promote exon skipping across various experimental conditions.[89,91–93] *DYNC1I2*, as a component of the dynactin complex,[94] 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 *LNPEP* (MIM: 151300); 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 AS, mRNA stability, mRNA localization, and translation initiation.[95–98] Containing four highly conserved RNA binding domains that recognize short pyrimidine-rich sequences,[99] this splicing factor regulates AS by inducing exon skipping.[100–102] 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 (Figure S7B).

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

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 (Δ IF) >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 (Tables S9 and 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 NMD, and the similarity of the ORF sequences between transcripts. For every detected isoform switch in a given gene, the annotations for the most-used isoform (Δ IF > 0.1) and the least-used isoform (Δ IF < 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%, Files S3–S6). 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 (Korowai vs. Sumba and Mentawai vs. Sumba), 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 vs. 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.[103–105]

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

(C) The isoform switch in *IL18RAP*. Exons with annotated domains are colored, 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 (Δ IF > 0.1 and FDR < 0.05) isoform switches are colored accordingly.

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, such as *IL18RAP* (MIM: 604509), which exhibited the most significant isoform switching event across all comparisons (q value $= 5 \times 10^{-27}$; 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.[106] As such, *IL18RAP* is a crucial component of the IL-18 receptor complex, and the dysregulation of *IL18RAP* expression has been implicated in various diseases, including autoimmune disorders, inflammatory conditions, and cancer.[107] In our data, we observe preferential usage of the longer *IL18RAP* isoform (ENST00000264260.6) in the Mentawai group (isoform switch q value $= 1.8 \times 10^{-8}$), while the Sumba and Korowai 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.[108] 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 *BST1* (MIM: 600387; Figure S8A). *BST1* (bone marrow stromal cell antigen 1) is an immune gene that facilitates pre-B cell growth and regulates leukocyte diapedesis in inflammation.[109] The isoform switch in *BST1* reveals increased usage of the ENST00000265016.9 isoform within the Sumba group (isoform switch q value $= 1.3 \times 10^{-4}$) while Korowai samples exhibit a higher isoform fraction of the ENST00000514989.1 isoform. The ENST00000514989.1 isoform is noncoding, 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 an NAD+ glycohydrolase.[110]

Most differentially expressed genes are not differentially spliced

Our isoform switching analyses illustrate how critical AS is for the diversification of gene expression at both the mRNA and protein levels (Figures 5A and 5D). 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[32] differentially expressed genes in the same groups. We found that differentially spliced genes were very rarely differentially expressed, with only 13 (Fisher's exact test $p = 9.7e-04$), 65 (Fisher's exact test $p = 1.1e-07$), and 39 (Fisher's exact test $p = 2.1e-06$) shared genes across both tests for the Mentawai vs. Sumba, Korowai vs. Sumba, and Korowai vs. Mentawai comparisons, respectively (Figure S8B). 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.[21,111–114]

Discussion

It is well established that AS influences various biological functions.[2,4,7–10,21] 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 region's genetic ancestry cline, we characterize the global AS landscape across these populations and identify genetic variants that are associated with splicing (sQTLs), thereby providing a comprehensive map of genetically regulated AS events in human whole blood.

Our detection of AS 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,[29,115–120] 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.[121,122] 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.[123] Thus, for future isoform-specific AS studies, the application of long-read technologies should be prioritized.

We conducted a global differential AS analysis between these groups and detected over 1,500 significant events across all comparisons with Mentawai and Sumba exhibiting the lowest number of differential events (i.e., the highest level of AS 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 nine sQTLs driven by LA 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 AS variation between populations reflects a complex interplay between genetic and environmental factors.

As a means of characterizing the phenotypic consequence of AS 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 colocalization 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 a notable observation given the role of high altitude corridors in the settlement of New Guinea Island.[124]

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[89,91–93] 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 *LNPEP*. 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),[100–102] the predicted reduction in PTBP1 activity directly correlates with a positive effect

size for this sQTL and a higher exon inclusion rate for individuals harboring the rs27291 A allele. While these results are consistent with previous findings that sQTLs disrupt the binding affinity of core and auxiliary splice proteins,[21,125] 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 AS 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 (1) contain an additional protein domain, (2) contain a signal peptide, and (3) gain a complete ORF. Our analysis identified several significant isoform-switching events in genes critical for the immune response, including *IL18RAP* and *BST1*. In *IL18RAP*, we observed preferential usage of different isoforms across populations, with potential implications for protein translocation and structural stability. Similarly, in *BST1*, isoform switching resulted in changes in coding potential and the presence of functional domains, highlighting the importance of AS in immune regulation and disease processes.

As previous studies have reported,[126–128] 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 DAS genes, indicating that the genetic control of splicing and transcription are independent, as previously observed.[16,20,21,29,129–131] 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.[32] Together, this study provides a comprehensive catalog of genetically regulated AS events in whole-blood Indonesian samples and adds to

our knowledge of genomic and regional drivers of gene regulatory variation across the Indonesian archipelago.

References

1. Sharp, P.A. (1994). Split genes and RNA splicing. Cell *77*, 805–815. https://doi.org/10.1016/0092-8674(94)90130-9.
2. Wright, C.J., Smith, C.W.J., and Jiggins, C.D. (2022). Alternative splicing as a source of phenotypic diversity. Nat. Rev. Genet. *23*, 697–710. https://doi.org/10.1038/s41576-022-00514-4.
3. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415. https://doi.org/10.1038/ng.259.
4. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476. https://doi.org/10.1038/nature07509.
5. Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nat. Rev. Genet. *17*, 19–32. https://www.nature.com/articles/nrg.2015.3.
6. Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. Nat. Rev. Genet. *8*, 749–761. https://www.nature.com/articles/nrg2164.
7. Ergun, A., Doran, G., Costello, J.C., Paik, H.H., Collins, J.J., Mathis, D., Benoist, C.; and ImmGen Consortium (2013). Differential splicing across immune system lineages. Proc. Natl. Acad. Sci. USA *110*, 14.324–14.329. https://doi.org/10.1073/pnas.1311839110.
8. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell *167*, 1398–1414.e24. https://doi.org/10.1016/j.cell.2016.10.026.
9. GTEx Consortium (2020). The GTEx consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
10. Zhang, Y., Yang, H.T., Kadash-Edmondson, K., Pan, Y., Pan, Z., Davidson, B.L., and Xing, Y. (2020). Regional variation of splicing QTLs in human brain. Am. J. Hum. Genet. *107*, 196–210. https://doi.org/10.1016/j.ajhg.2020.06.002.
11. Pickrell, J.K., Pai, A.A., Gilad, Y., and Pritchard, J.K. (2010). Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. PLoS Genet. *6*, e1001236. https://doi.org/10.1371/journal.pgen.1001236.
12. Saudemont, B., Popa, A., Parmley, J.L., Rocher, V., Blugeon, C., Necsulea, A., Meyer, E., and Duret, L. (2017). The fitness cost of mis-splicing is the main determinant of alternative splicing patterns. Genome Biol. *18*, 208. https://doi.org/10.1186/s13059-017-1344-6.
13. Pozo, F., Martinez-Gomez, L., Walsh, T.A., Rodriguez, J.M., Di Domenico, T., Abascal, F., Vazquez, J., and Tress, M.L. (2021). Assessing the functional relevance of splice isoforms. NAR Genom. Bioinform. *3*, lqab044. https://doi.org/10.1093/nargab/lqab044.
14. Carpenter, S., Ricci, E.P., Mercier, B.C., Moore, M.J., and Fitzgerald, K.A. (2014). Post-transcriptional regulation of gene expression in innate immunity. Nat. Rev. Immunol. *14*, 361–376. https://doi.org/10.1038/nri3682.
15. Richards, A.L., Watza, D., Findley, A., Alazizi, A., Wen, X., Pai, A.A., Pique-Regi, R., and Luca, F. (2017). Environmental perturbations lead to extensive directional shifts in RNA processing. PLoS Genet. *13*, e1006995. https://doi.org/10.1371/journal.pgen.1006995.
16. Ye, C.J., Chen, J., Villani, A.-C., Gate, R.E., Subramaniam, M., Bhangale, T., Lee, M.N., Raj, T., Raychowdhury, R., Li, W., et al. (2018). Genetic analysis of isoform usage in the human anti-viral response reveals influenza-specific regulation of ERAP2 transcripts under balancing selection. Genome Res. *28*, 1812–1825. https://doi.org/10.1101/gr.240390.118.
17. Pai, A.A., Baharian, G., Pagé Sabourin, A., Brinkworth, J.F., Nédélec, Y., Foley, J.W., Grenier, J.-C., Siddle, K.J., Dumaine, A., Yotova, V., et al. (2016). Widespread shortening of 3' untranslated regions and increased exon inclusion are

evolutionarily conserved features of innate immune responses to infection. PLoS Genet. *12*, e1006338. https://doi.org/10.1371/journal.pgen.1006338.

18. Kalam, H., Fontana, M.F., and Kumar, D. (2017). Alternate splicing of transcripts shape macrophage response to mycobacterium tuberculosis infection. PLoS Pathog. *13*, e1006236. https://doi.org/10.1371/journal.ppat.1006236.

19. Martinez, N.M., and Lynch, K.W. (2013). Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. Immunol. Rev. *253*, 216–236. https://doi.org/10.1111/imr.12047.

20. Rotival, M., Quach, H., and Quintana-Murci, L. (2019). Defining the genetic and evolutionary architecture of alternative splicing in response to infection. Nat. Commun. *10*, 1671. https://doi.org/10.1038/s41467-019-09689-7.

21. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604. https://doi.org/10.1126/science.aad9417.

22. Wang, Z., and Burge, C.B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. RNA *14*, 802–813. https://doi.org/10.1261/rna.876308.

23. Faustino, N.A., and Cooper, T.A. (2003). Pre-mRNA splicing and human disease. Genes Dev. *17*, 419–437. https://doi.org/10.1101/gad.1048803.

24. Caswell, J.L., Camarda, R., Zhou, A.Y., Huntsman, S., Hu, D., Brenner, S.E., Zaitlen, N., Goga, A., and Ziv, E. (2015). Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors. Hum. Mol. Genet. *24*, 7421–7431. https://doi.org/10.1093/hmg/ddv432.

25. Raj, T., Li, Y.I., Wong, G., Humphrey, J., Wang, M., Ramdhani, S., Wang, Y.-C., Ng, B., Gupta, I., Haroutunian, V., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in alzheimer's disease susceptibility. Nat. Genet. *50*, 1584–1592. https://doi.org/10.1038/s41588-018-0238-1.

26. Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat. Commun. *8*, 14519. https://doi.org/10.1038/ncomms14519.

27. Martin, A.R., Costa, H.A., Lappalainen, T., Henn, B.M., Kidd, J.M., Yee, M.-C., Grubert, F., Cann, H.M., Snyder, M., Montgomery, S.B., and Bustamante, C.D. (2014). Transcriptome Sequencing from Diverse Human Populations Reveals Differentiated Regulatory Architecture. PLoS Genet. *10*, e1004549. https://doi.org/10.1371/journal.pgen.1004549.

28. Çalışkan, M., Cusanovich, D.A., Ober, C., and Gilad, Y. (2011). The effects of EBV transformation on gene expression levels and methylation profiles. Hum. Mol. Genet. *20*, 1643–1652. https://doi.org/10.1093/hmg/ddr041.

29. García-Pérez, R., Ramirez, J.M., Ripoll-Cladellas, A., Chazarra-Gil, R., Oliveros, W., Soldatkina, O., Bosio, M., Rognon, P.J., Capella-Gutierrez, S., Calvo, M., et al. (2023). The landscape of expression and alternative splicing variation across human traits. Cell Genom. *3*, 100244. https://doi.org/10.1016/j.xgen.2022.100244.

30. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The expanding landscape of alternative splicing variation in human populations. Am. J. Hum. Genet. *102*, 11–26. https://doi.org/10.1016/j.ajhg.2017.11.002.

31. Kelly, D.E., Ramdas, S., Ma, R., Rawlings-Goss, R.A., Grant, G.R., Ranciaro, A., Hirbo, J.B., Beggs, W., Yeager, M., Chanock, S., et al. (2023). The genetic and evolutionary basis of gene expression variation in East Africans. Genome Biol. *24*, 35. https://doi.org/10.1186/s13059-023-02874-4.

32. Natri, H.M., Bobowik, K.S., Kusuma, P., Crenna Darusallam, C., Jacobs, G.S., Hudjashov, G., Lansing, J.S., Sudoyo, H., Banovich, N.E., Cox, M.P., and Gallego Romero, I. (2020). Genome-wide DNA methylation and gene expression patterns reflect genetic ancestry and environmental differences across the indonesian archipelago. PLoS Genet. *16*, e1008749. https://doi.org/10.1371/journal.pgen.1008749.

33. Natri, H.M., Hudjashov, G., Jacobs, G., Kusuma, P., Saag, L., Darusallam, C.C., Metspalu, M., Sudoyo, H., Cox, M.P., Gallego Romero, I., and Banovich, N.E. (2022). Genetic architecture of gene regulation in indonesian populations identifies QTLs associated with global and local ancestries. Am. J. Hum. Genet. *109*, 50–65. https://doi.org/10.1016/j.ajhg.2021.11.017.

34. Hudjashov, G., Karafet, T.M., Lawson, D.J., Downey, S., Savina, O., Sudoyo, H., Lansing, J.S., Hammer, M.F., and Cox, M.P. (2017). Complex patterns of admixture across the indonesian archipelago. Mol. Biol. Evol. *34*, 2439–2452. https://academic.oup.com/mbe/article/34/10/2439/3952785.

35. Jacobs, G.S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C.C., Lawson, D.J., Mondal, M., Pagani, L., Ricaut, F.-X., Stoneking, M., et al. (2019). Multiple deeply divergent denisovan ancestries in papuans. Cell *177*, 1010–1021.e32. https://doi.org/10.1016/j.cell.2019.02.035.

36. Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data, p. 2017.

37. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

38. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

39. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., and Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. *19*, 40. https://doi.org/10.1186/s13059-018-1417-1.

40. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

41. Closa, A., Reixachs-Solé, M., Fuentes-Fayos, A.C., Hayer, K.E., Melero, J.L., Adriaanse, F.R.S., Bos, R.S., Torres-Diz, M., Hunger, S.P., Roberts, K.G., et al. (2022). A convergent malignant phenotype in B-cell acute lymphoblastic leukemia involving the splicing factor SRRM1. NAR Cancer *4*, zcac041. https://doi.org/10.1093/narcan/zcac041.

42. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158. https://doi.org/10.1038/s41588-017-0004-9.

43. Katz, Y., Wang, E.T., Silterra, J., Schwartz, S., Wong, B., Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P., Airoldi, E.M., and Burge, C.B. (2015). Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics *31*, 2400–2402. https://doi.org/10.1093/bioinformatics/btv034.

44. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods 5, 621–628. https://doi.org/10.1038/nmeth.1226.

45. Vitting-Seerup, K., and Sandelin, A. (2019). IsoformSwitch AnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics 35, 4469–4471. https://doi.org/10.1093/bioinformatics/btz247.

46. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47. https://doi.org/10.1093/nar/gkv007.

47. Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017. https://doi.org/10.1101/gr.133744.111.

48. Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., and Gao, G. (2017). CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45, W12–W16. https://doi.org/10.1093/nar/gkx428.

49. Mészáros, B., Erdos, G., and Dosztányi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 46, W329–W337. https://doi.org/10.1093/nar/gky384.

50. Almagro Armenteros, J.J., Tsirigos, K.D., Sønderby, C.K., Petersen, T.N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat. Biotechnol. 37, 420–423. https://doi.org/10.1038/s41587-019-0036-z.

51. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419. https://doi.org/10.1093/nar/gkaa913.

52. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A complete tool set for molecular QTL discovery and analysis. Nat. Commun. 8, 15452. https://doi.org/10.1038/ncomms15452.

53. Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat. Protoc. 7, 500–507. https://www.nature.com/articles/nprot.2011.457.

54. Maintainer, B.P. (2020). liftOver: Changing genomic coordinate systems with rtracklayer:: liftOver. R package version 1, 10–18.129.

55. Urbut, S.M., Wang, G., Carbonetto, P., and Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. 51, 187–195. https://doi.org/10.1038/s41588-018-0268-8.

56. Gay, N.R., Gloudemans, M., Antonio, M.L., Abell, N.S., Balliu, B., Park, Y., Martin, A.R., Musharoff, S., Rao, A.S., Aguet, F., et al. (2020). Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. Genome Biol. 21, 233. https://doi.org/10.1186/s13059-020-02113-0.

57. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

58. Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6, e21800. https://doi.org/10.1371/journal.pone.0021800.

59. Grønning, A.G.B., Doktor, T.K., Larsen, S.J., Petersen, U.S.S., Holm, L.L., Bruun, G.H., Hansen, M.B., Hartung, A.-M., Baumbach, J., and Andresen, B.S. (2020). DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. Nucleic Acids Res. 48, 7099–7118.

60. Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 46, D794–D801. https://doi.org/10.1093/nar/gkx1081.

61. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74. https://doi.org/10.1038/nature11247.

62. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 10, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

63. Chen, M.-H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell 182, 1198–1213.e14. https://doi.org/10.1016/j.cell.2020.06.045.

64. Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. the human splicing code reveals new insights into the genetic determinants of disease. Science 347, 1254806. https://doi.org/10.1126/science.1254806.

65. Hatakeyama, S. (2011). TRIM proteins and cancer. Nat. Rev. Cancer 11, 792–804. https://doi.org/10.1038/nrc3139.

66. Tao, R., Li, J., Xin, J., Wu, J., Guo, J., Zhang, L., Jiang, L., Zhang, W., Yang, Z., and Li, L. (2011). Methylation profile of single hepatocytes derived from hepatitis B virus-related hepatocellular carcinoma. PLoS One 6, e19862. https://doi.org/10.1371/journal.pone.0019862.

67. Kajiura, K., Masuda, K., Naruto, T., Kohmoto, T., Watabnabe, M., Tsuboi, M., Takizawa, H., Kondo, K., Tangoku, A., and Imoto, I. (2017). Frequent silencing of the candidate tumor suppressor TRIM58 by promoter methylation in early-stage lung adenocarcinoma. Oncotarget 8, 2890–2905. https://doi.org/10.18632/oncotarget.13761.

68. Xu, R., Xu, Q., Huang, G., Yin, X., Zhu, J., Peng, Y., and Song, J. (2019). Combined analysis of the aberrant epigenetic alteration of pancreatic ductal adenocarcinoma. BioMed Res. Int. 2019, 9379864. https://doi.org/10.1155/2019/9379864.

69. Gan, Y., Cao, C., Li, A., Song, H., Kuang, G., Ma, B., Zhang, Q., and Zhang, Q. (2021). 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. https://doi.org/10.3389/fmolb.2021.655126.

70. Su, M., Lin, Y., Cui, C., Tian, X., and Lai, L. (2021). ERMAP is a B7 family-related molecule that negatively regulates T cell and macrophage responses. Cell. Mol. Immunol. 18, 1920–1933. https://doi.org/10.1038/s41423-020-0494-8.

71. Xu, H., Foltz, L., Sha, Y., Madlansacay, M.R., Cain, C., Lindemann, G., Vargas, J., Nagy, D., Harriman, B., Mahoney, W., and Schueler, P.A. (2001). Cloning and characterization of human erythroid membrane-associated protein, human ERMAP. Genomics 76, 2–4. https://doi.org/10.1006/geno.2001.6600.

72. Li, J., Liu, X.-G., Ge, R.-L., Yin, Y.-P., Liu, Y.-d., Lu, W.-P., Huang, M., He, X.-Y., Wang, J., Cai, G., et al. (2023). The ligation between ERMAP, galectin-9 and dectin-2 promotes kupffer cell phagocytosis and antitumor immunity. Nat. Immunol. 24, 1813–1824. https://doi.org/10.1038/s41590-023-01634-7.

73. Wagner, F.F., Poole, J., and Flegel, W.A. (2003). Scianna antigens including rd are expressed by ERMAP. Blood 101, 752–757. https://doi.org/10.1182/blood-2002-07-2064.

74. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

75. Orth, J.D., and McNiven, M.A. (2003). Dynamin at the actin-membrane interface. Curr. Opin. Cell Biol. 15, 31–39. https://doi.org/10.1016/s0955-0674(02)00010-8.

76. Meisinger, C., Prokisch, H., Gieger, C., Soranzo, N., Mehta, D., Rosskopf, D., Lichtner, P., Klopp, N., Stephens, J., Watkins, N.A., et al. (2009). A genome-wide association study identifies three loci associated with mean platelet volume. Am. J. Hum. Genet. 84, 66–71. https://doi.org/10.1016/j.ajhg.2008.11.015.

77. Cantor, A.B. (2012). Dynamin 3 and platelet size variation. Blood 120, 4666–4667. https://ashpublications.org/blood/article/120/24/4666/30976/Dynamin-3-and-platelet-size-variation.

78. Reems, J.-A., Wang, W., Tsubata, K., Abdurrahman, N., Sundell, B., Tijssen, M.R., van der Schoot, E., Di Summa, F., Patel-Hett, S., Italiano, J., Jr., and Gilligan, D.M. (2008). Dynamin 3 participates in the growth and development of megakaryocytes. Exp. Hematol. 36, 1714–1727. https://doi.org/10.1016/j.exphem.2008.08.010.

79. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al. (2023). Ensembl 2023. Nucleic Acids Res. 51, D933–D941. https://doi.org/10.1093/nar/gkac958.

80. Kelley, J., Walter, L., and Trowsdale, J. (2005). Comparative genomics of natural killer cell receptor gene clusters. PLoS Genet. 1, 129–139. https://doi.org/10.1371/journal.pgen.0010027.

81. Spiekerkoetter, U. (2010). Mitochondrial fatty acid oxidation disorders: clinical presentation of long-chain fatty acid oxidation defects before and after newborn screening. J. Inherit. Metab. Dis. 33, 527–532. https://doi.org/10.1007/s10545-010-9090-x.

82. Wang, X., Luo, X., Wang, Z., Wang, Y., Zhao, J., and Bian, L. (2023). Identification of cancer stemness and M2 macrophage-associated biomarkers in lung adenocarcinoma. Heliyon 9, e19114. https://doi.org/10.1016/j.heliyon.2023.e19114.

83. Wojnacki, J., and Galli, T. (2018). A new actin-binding domain glues autophagy together. J. Biol. Chem. 293, 4575–4576. https://doi.org/10.1074/jbc.H118.002041.

84. Rashkin, S.R., Graff, R.E., Kachuri, L., Thai, K.K., Alexeeff, S.E., Blatchins, M.A., Cavazos, T.B., Corley, D.A., Emami, N.C., Hoffman, J.D., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat. Commun. 11, 4423. https://doi.org/10.1038/s41467-020-18246-6.

85. Tangtanatakul, P., Thumarat, C., Satproedprai, N., Kunhapan, P., Chaiyasung, T., Klinchanhom, S., Wang, Y.-F., Wei, W., Wongshinsri, J., Chiewchengchol, D., et al. (2020). 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, 185. https://doi.org/10.1186/s13075-020-02276-y.

86. Paronetto, M.P. (2013). Ewing sarcoma protein: a key player in human cancer. Int. J. Cell Biol. 2013, 642853. https://doi.org/10.1155/2013/642853.

87. Huelga, S.C., Vu, A.Q., Arnold, J.D., Liang, T.Y., Liu, P.P., Yan, B.Y., Donohue, J.P., Shiue, L., Hoon, S., Brenner, S., et al. (2012). Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. Cell Rep. 1, 167–178. https://doi.org/10.1016/j.celrep.2012.02.001.

88. Datar, K.V., Dreyfuss, G., and Swanson, M.S. (1993). The human hnRNP M proteins: identification of a methionine/arginine-rich repeat motif in ribonucleoproteins. Nucleic Acids Res. 21, 439–446. https://doi.org/10.1093/nar/21.3.439.

89. Xu, Y., Gao, X.D., Lee, J.-H., Huang, H., Tan, H., Ahn, J., Reinke, L.M., Peter, M.E., Feng, Y., Gius, D., et al. (2014). Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. Genes Dev. 28, 1191–1203. https://doi.org/10.1101/gad.241968.114.

90. Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.-F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv. Exp. Med. Biol. 623, 123–147. https://doi.org/10.1007/978-0-387-77374-2_8.

91. West, K.O., Scott, H.M., Torres-Odio, S., West, A.P., Patrick, K.L., and Watson, R.O. (2019). The splicing factor hnRNP M is a critical regulator of innate immune gene expression in macrophages. Cell Rep. 29, 1594–1609.e5. https://doi.org/10.1016/j.celrep.2019.09.078.

92. Harvey, S.E., Xu, Y., Lin, X., Gao, X.D., Qiu, Y., Ahn, J., Xiao, X., and Cheng, C. (2018). Coregulation of alternative splicing by hnRNPM and ESRP1 during EMT. RNA 24, 1326–1338. https://doi.org/10.1261/rna.066712.118.

93. Ho, J.S., Di Tullio, F., Schwarz, M., Low, D., Incarnato, D., Gay, F., Tabaglio, T., Zhang, J., Wollmann, H., Chen, L., et al. (2021). HNRNPM controls circRNA biogenesis and splicing fidelity to sustain cancer cell fitness. Elife 10, e59654. https://doi.org/10.7554/eLife.59654.

94. Pfister, K.K., Shah, P.R., Hummerich, H., Russ, A., Cotton, J., Annuar, A.A., King, S.M., and Fisher, E.M.C. (2006). Genetic analysis of the cytoplasmic dynein subunit families. PLoS Genet. 2, e1. https://doi.org/10.1371/journal.pgen.0020001.

95. Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). 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, 996–1006. https://doi.org/10.1016/j.molcel.2009.12.003.

96. Keppetipola, N., Sharma, S., Li, Q., and Black, D.L. (2012). Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. Crit. Rev. Biochem. Mol. Biol. 47, 360–378. https://doi.org/10.3109/10409238.2012.691456.

97. Kafasla, P., Mickleburgh, I., Llorian, M., Coelho, M., Gooding, C., Cherny, D., Joshi, A., Kotik-Kogan, O., Curry, S., Eperon, I.C., et al. (2012). Defining the roles and interactions of PTB. Biochem. Soc. Trans. *40*, 815–820. https://doi.org/10.1042/BST20120044.

98. Sawicka, K., Bushell, M., Spriggs, K.A., and Willis, A.E. (2008). Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. Biochem. Soc. Trans. *36*, 641–647. https://doi.org/10.1042/BST0360641.

99. Han, A., Stoilov, P., Linares, A.J., Zhou, Y., Fu, X.-D., and Black, D.L. (2014). De novo prediction of PTBP1 binding and splicing targets reveals unexpected features of its RNA recognition and function. PLoS Comput. Biol. *10*, e1003442. https://doi.org/10.1371/journal.pcbi.1003442.

100. Georgilis, A., Klotz, S., Hanley, C.J., Herranz, N., Weirich, B., Morancho, B., Leote, A.C., D'Artista, L., Gallage, S., Seehawer, M., et al. (2018). PTBP1-mediated alternative splicing regulates the inflammatory secretome and the pro-tumorigenic effects of senescent cells. Cancer Cell *34*, 85–102.e9. https://doi.org/10.1016/j.ccell.2018.06.007.

101. Shen, L., Lei, S., Zhang, B., Li, S., Huang, L., Czachor, A., Breitzig, M., Gao, Y., Huang, M., Mo, X., et al. (2020). Skipping of exon 10 in axl pre-mRNA regulated by PTBP1 mediates invasion and metastasis process of liver cancer cells. Theranostics *10*, 5719–5735. https://doi.org/10.7150/thno.42010.

102. Li, S., Shen, L., Huang, L., Lei, S., Cai, X., Breitzig, M., Zhang, B., Yang, A., Ji, W., Huang, M., et al. (2019). PTBP1 enhances exon11a skipping in mena pre-mRNA to promote migration and invasion in lung carcinoma cells. Biochim. Biophys. Acta. Gene Regul. Mech. *1862*, 858–869. https://doi.org/10.1016/j.bbagrm.2019.04.006.

103. Fuxreiter, M., Simon, I., and Bondos, S. (2011). Dynamic protein–DNA recognition: beyond what can be seen. Trends Biochem. Sci. *36*, 415–423.

104. Guo, X., Bulyk, M.L., and Hartemink, A.J. (2012). Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. Pac. Symp. Biocomput., 104–115. https://www.ncbi.nlm.nih.gov/pubmed/22174267.

105. Habchi, J., Tompa, P., Longhi, S., and Uversky, V.N. (2014). Introducing protein intrinsic disorder. Chem. Rev. *114*, 6561–6588. https://doi.org/10.1021/cr400514h.

106. Torigoe, K., Ushio, S., Okura, T., Kobayashi, S., Taniai, M., Kunikata, T., Murakami, T., Sanou, O., Kojima, H., Fujii, M., et al. (1997). Purification and characterization of the human interleukin-18 receptor. J. Biol. Chem. *272*, 25737–25742. https://doi.org/10.1074/jbc.272.41.25737.

107. Hedl, M., Zheng, S., and Abraham, C. (2014). The IL18RAP region disease polymorphism decreases IL-18RAP/IL-18R1/IL-1R1 expression and signaling through innate receptor-initiated pathways. J. Immunol. *192*, 5924–5932. https://doi.org/10.4049/jimmunol.1302727.

108. Tsutsumi, N., Kimura, T., Arita, K., Ariyoshi, M., Ohnishi, H., Yamamoto, T., Zuo, X., Maenaka, K., Park, E.Y., Kondo, N., et al. (2014). The structural basis for receptor recognition of human interleukin-18. Nat. Commun. *5*, 5340. https://doi.org/10.1038/ncomms6340.

109. Ferrero, E., Lo Buono, N., Morone, S., Parrotta, R., Mancini, C., Brusco, A., Giacomino, A., Augeri, S., Rosal-Vela, A., García-Rodríguez, S., et al. (2017). Human canonical CD157/Bst1 is an alternatively spliced isoform masking a previously unidentified primate-specific exon included in a novel transcript. Sci. Rep. *7*, 15923. https://doi.org/10.1038/s41598-017-16184-w.

110. Malavasi, F., Deaglio, S., Funaro, A., Ferrero, E., Horenstein, A.L., Ortolan, E., Vaisitti, T., and Aydin, S. (2008). Evolution and function of the ADP ribosyl cyclase/CD38 gene family in physiology and pathology. Physiol. Rev. *88*, 841–886. https://doi.org/10.1152/physrev.00035.2007.

111. Jacobs, A., and Elmer, K.R. (2021). Alternative splicing and gene expression play contrasting roles in the parallel phenotypic evolution of a salmonid fish. Mol. Ecol. *30*, 4955–4969. https://doi.org/10.1111/mec.15817.

112. Kalsotra, A., and Cooper, T.A. (2011). Functional consequences of developmentally regulated alternative splicing. Nat. Rev. Genet. *12*, 715–729. https://www.nature.com/articles/nrg3052.

113. Huntley, M.A., Srinivasan, K., Friedman, B.A., Wang, T.-M., Yee, A.X., Wang, Y., Kaminker, J.S., Sheng, M., Hansen, D.V., and Hanson, J.E. (2020). Genome-wide analysis of differential gene expression and splicing in excitatory neurons and interneuron subtypes. J. Neurosci. *40*, 958–973. https://doi.org/10.1523/JNEUROSCI.1615-19.2019.

114. Grantham, M.E., and Brisson, J.A. (2018). Extensive differential splicing underlies phenotypically plastic aphid morphs. Mol. Biol. Evol. *35*, 1934–1946. https://doi.org/10.1093/molbev/msy095.

115. Sammeth, M., Foissac, S., and Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. PLoS Comput. Biol. *4*, e1000147. https://doi.org/10.1371/journal.pcbi.1000147.

116. Kim, D., Shivakumar, M., Han, S., Sinclair, M.S., Lee, Y.-J., Zheng, Y., Olopade, O.I., Kim, D., and Lee, Y. (2018). Population-dependent intron retention and DNA methylation in breast cancer. Mol. Cancer Res. *16*, 461–469. https://doi.org/10.1158/1541-7786.MCR-17-0227.

117. Aktas Samur, A., Fulciniti, M., Avet-Loiseau, H., Lopez, M.A., Derebail, S., Corre, J., Minvielle, S., Magrangeas, F., Moreau, P., Anderson, K.C., et al. (2022). In-depth analysis of alternative splicing landscape in multiple myeloma and potential role of dysregulated splicing factors. Blood Cancer J. *12*, 171. https://doi.org/10.1038/s41408-022-00759-6.

118. Bradley, R.K., Merkin, J., Lambert, N.J., and Burge, C.B. (2012). Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. PLoS Biol. *10*, e1001229. https://doi.org/10.1371/journal.pbio.1001229.

119. Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. *11*, 345–355. https://doi.org/10.1038/nrg2776.

120. Jiang, W., and Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. Comput. Struct. Biotechnol. J. *19*, 183–195. https://doi.org/10.1016/j.csbj.2020.12.009.

121. Florea, L., Song, L., and Salzberg, S.L. (2013). Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues. F1000Res. *2*, 188. https://doi.org/10.12688/f1000research.2-188.v2.

122. David, J.K., Maden, S.K., Wood, M.A., Thompson, R.F., and Nellore, A. (2022). Retained introns in long RNA-seq reads are not reliably detected in sample-matched short reads. Genome Biol. *23*, 240. https://doi.org/10.1186/s13059-022-02789-6.

123. Kovaka, S., Ou, S., Jenike, K.M., and Schatz, M.C. (2023). Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. Nat. Methods *20*, 12–16. https://doi.org/10.1038/s41592-022-01716-8.

124. Brucato, N., André, M., Tsang, R., Saag, L., Kariwiga, J., Sesuki, K., Beni, T., Pomat, W., Muke, J., Meyer, V., et al. (2021). Papua New Guinean Genomes Reveal the Complex Settlement of North Sahul. Mol. Biol. Evol. *38*, 5107–5121. https://doi.org/10.1093/molbev/msab238.

125. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., and Guigó, R. (2021). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nat. Commun. *12*, 727. https://doi.org/10.1038/s41467-020-20578-2.

126. Chen, L., Chen, K., Hong, Y., Xing, L., Zhang, J., Zhang, K., and Zhang, Z. (2022). The landscape of isoform switches in sepsis: a multicenter cohort study. Sci. Rep. *12*, 10276. https://doi.org/10.1038/s41598-022-14231-9.

127. Zhang, Y., Weh, K.M., Howard, C.L., Riethoven, J.-J., Clarke, J.L., Lagisetty, K.H., Lin, J., Reddy, R.M., Chang, A.C., Beer, D.G., and Kresty, L.A. (2022). Characterizing isoform switching events in esophageal adenocarcinoma. Mol. Ther. Nucleic Acids *29*, 749–768. https://doi.org/10.1016/j.omtn.2022.08.018.

128. Marques-Coelho, D., Iohan, L.d.C.C., Melo de Farias, A.R., Flaig, A., Brainbank Neuro–CEB Neuropathology Network, Lambert, J.-C., and Costa, M.R. (2021). Differential transcript usage unravels gene expression alterations in alzheimer's disease human brains. NPJ Aging Mech. Dis. *7*, 2. https://doi.org/10.1038/s41514-020-00052-5.

129. Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D.F., Paul, D.S., and Gaffney, D.J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. Elife *8*, e41673. https://doi.org/10.7554/eLife.41673.

130. Lalonde, E., Ha, K.C.H., Wang, Z., Bemmo, A., Kleinman, C.L., Kwan, T., Pastinen, T., and Majewski, J. (2011). RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Res. *21*, 545–554. https://doi.org/10.1101/gr.111211.110.

131. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. *24*, 14–24. https://doi.org/10.1101/gr.155192.113.