Supplementary Materials for

Transcriptomic signatures across human tissues identify functional rare genetic variation

Nicole M. Ferraro*, Benjamin J. Strober*, Jonah Einson, Nathan S. Abell, Francois Aguet, Alvaro N. Barbeira, Margot Brandt, Maja Bucan, Stephane E. Castel, Joe R. Davis, Emily Greenwald, Gaelen T. Hess, Austin T. Hilliard, Rachel L. Kember, Bence Kotis, YoSon Park, Gina Peloso, Shweta Ramdas, Alexandra J. Scott, Craig Smail, Emily K. Tsang, Seyedeh M. Zekavat, Marcello Ziosi, Aradhana, TOPMed Lipids Working Group, Kristin G. Ardlie, Themistocles L. Assimes, Michael C. Bassik, Christopher D. Brown, Adolfo Correa, Ira Hall, Hae Kyung Im, Xin Li, Pradeep Natarajan, GTEx Consortium, Tuuli Lappalainen, Pejman Mohammadi†‡, Stephen B. Montgomery†‡, Alexis Battle†‡

*These authors contributed equally to this work.
†These authors contributed equally to this work.
‡Corresponding author. Email: pejman@scripps.edu (P.M.); smontgom@stanford.edu (S.B.M.); ajbattle@jhu.edu (A.B.)

Published 11 September 2020, *Science* **369**, eaaz5900 (2020)
DOI: 10.1126/science.aaz5900

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at science.sciencemag.org/content/369/6509/eaaz5900/suppl/DC1)

Tables S1 to S3 and S6 to S11 (Excel files)
MDAR Reproducibility Checklist (PDF)

Materials and Methods

GTEx data

All human donors were deceased, and informed consent was obtained via next-of-kin consent for the collection and banking of deidentified tissue samples for scientific research. The research protocol was reviewed by Chesapeake Research Review Inc., Roswell Park Cancer Institute's Office of Research Subject Protection, and the institutional review board of the University of Pennsylvania. We used the RNA-sequencing, allele-specific expression, and whole-genome sequencing (WGS) data from the v8 release of the GTEx project and assessed expression data across the 49 biological tissues with at least 70 samples. Sample size varied across tissues, with average missingness of ~50%. Self-reported ancestry for these individuals spanned three of the major continental populations with the majority (n=714 with WGS) comprising individuals of predominantly European ancestry, 121 individuals with African ancestry, 11 with Asian ancestry, and 12 unknown or other. The generation of these data are described in the supplementary information of (*18*).

Rare variant annotations

We retained all SNVs and indels that passed quality control in the GTEx VCF, variant calling described in (*18*), using the hg38 genome build. Structural variants were called according to (*65*) on the subset of individuals available from V7 with GenomeSTRiP (*66*) GSCNQUAL set to limit the false discovery rate (FDR) for each variant type. Genome STRiP's IntensityRankSumAnnotator was used to evaluate FDR based on available Illumina Human Omni 5M gene expression array data. GSCNQUAL was limited to ≥ 1 for GenomeSTRiP deletions and ≥ 8 for multi-allelic copy number variants, corresponding to an FDR of 10%. The GSCNQUAL cutoff for GenomeSTRiP duplications was set at ≥ 17, the point where the FDR plateaued at 15.1% and did not fluctuate more than ±1% for over 50 steps in increasing GSCNQUAL score. Additionally, the Mobile Element Locator Tool (MELT) version 2.1.4 (*67*) was run using MELT-SPLIT to identify ALU, SVA, and LINE1 insertions into the test genomes. MELT calls that were categorized as "PASS" in the VCF info field, had an ASSESS score ≥ 3, and SR count ≥ 3 were retained. Structural variant (SV) calls were then lifted to the hg38 genome build using liftOver from the Genome Browser (*68*).

We defined rare variants as those with < 1% MAF within GTEx and, for SNVs and indels, also occurring at < 1% frequency in non-Finnish Europeans within gnomAD (*21*). Novel variants were those that occurred in GTEx but were not found in gnomAD. GTEx singletons had an average allele frequency of 0.0030 in gnomAD and doubletons had an average frequency of 0.0096.

Annotation of protein-coding regions and transcription factor binding site motifs was generated by running Ensembl VEP (version 88). Loss of function (LoF) annotation was generated using loftee. Conservation scores (Gerp, PhyloP, PhastCons) were downloaded from UCSC genome browser and CADD scores were extracted from a pre-compiled annotation file (https://cadd.gs.washington.edu/download) using variant scores from the hg38 genome build.

Expression outlier calling

Within each tissue, we log2-transformed the expression values ($\log_2$(TPM + 2)), where TPM is the number of transcripts per million mapped reads, generated by RNA-SeQC (*69*) using the GENCODE v26 gene annotation, available through the GTEx portal. We subsetted to autosomal lincRNA and protein-coding genes and restricted to genes with at least 6 reads and TPM > 0.1 in at least 20% of individuals. We scaled the expression of each gene to mean of 0 and standard deviation of 1 to avoid the deflation of outlier values caused by quantile normalization. As we expected unmeasured technical confounders to impact expression, for each tissue we estimated hidden factors for the transformed expression matrix using PEER (*52*). The number of PEER factors retained was based on sample size and matched the values chosen in the GTEx eQTL analyses (*18*), which were 15 for sample sizes less than or equal to 150, 30 for less than 250, 45 for less than 350, and 60 otherwise. We obtained expression residuals by regressing out PEER factors, the top three genotype principal components, sex, and the genotype of the strongest cis-eQTL per gene in each tissue using the following linear model:

$$Y_g = \mu_g + \sum_{n=1}^{N} \alpha_{g,n} P_n + \sum_{k=1}^{3} \beta_{g,k} G_k + \gamma_g S + \delta_g Q + \varepsilon_g$$

where Y_g is the transformed expression of gene g, μ_g is the mean expression level for the gene, P_n is the nth PEER factor, G_k are the top k genotype principal components, S is the sex covariate, and Q is the genotype of the strongest cis-eQTL for gene g. We then re-scaled the expression residuals ε_g for each gene, to obtain corrected expression Z-scores for each individual per gene per tissue.

For each gene, we calculated an individual's median Z-score across all tissues for which data were available, restricting to individuals with measurements in at least five tissues. To account for situations where widespread extreme expression might occur in an individual due to non-genetic influences, we excluded 39 individuals where the proportion of tested genes that were multi-tissue outliers at a threshold of |median Z-score| > 3 exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals. We then use the median Z-scores per individual across tissues to determine eOutliers and used a threshold of |median Z| > 3 or an equivalent median p-value of 0.0027 for aseOutliers and sOutliers to determine the outlier set of genes. This threshold was chosen to balance the number of outliers identified with increases in nearby rare variant enrichments, though the conclusions are robust to threshold choice (Fig S1D, Fig S7). Controls were defined as any individual with a |median Z-score| of less than 3 (or another threshold as indicated) for the same set of genes as those with any outlier individual. We allowed a gene to have multiple outlier individuals and an individual could be an outlier for multiple genes. Code for generating eOutlier calls was modified from scripts available at https://github.com/joed3/GTExV6PRareVariation.

ASE outlier calling

Allelic expression (ASE) data was produced as described in (*70*). We used the Analysis of Expression VAriation Dosage Outlier Test (ANEVA-DOT; (*17*)) to identify genes in each individual that showed an excessive imbalance of ASE, relative to the population. Briefly, ANEVA-DOT relies on tissue-specific estimates of genetic variation in gene dosage, V^G, derived by Analysis of Expression VAriation (ANEVA) on a reference population ASE data to identify genes in individual test samples that are likely affected by rare variants with unusually large regulatory effects. We calculated reference V^G estimates from GTEx v8 data from 15,201 RNA-seq samples spanning 49 tissues and 838 individuals with WGS data (*17*). Across all analyzed tissues we estimated V^G a total of 2,727,867 times using all available autosomal aeSNPs (variants used to assess allelic expression) with at least 30 reads in 6 individuals. These estimates are publicly available at https://doi.org/10.5281/zenodo.3897759, version 2.31. We used the ANEVA-DOT tool R package (*60*) to calculate a p-value for every gene-individual pair with allelic expression data and a corresponding V^G estimate (Fig S3). The p-value can be interpreted as the result of a binomial test of allelic imbalance, that is overdispersed for each gene individually according to its expected dosage variation in a given tissue in the population. Genes with significant ANEVA-DOT p-values are referred to as aseOutliers in this text. We tested all tissues available for each GTEx v8 individual, using only genes with a minimum coverage of 8 reads spanning an aeSNP and with V^G estimates available (49 tissues, median genes per tissue = 4899, Fig S2). For each gene expressed we considered the aeSNP with the highest coverage in an individual.

For all single-tissue analyses, we removed global outlier genes and individuals from each tissue group independently. Global outlier genes are likely to be ASE outliers at 5% FDR in more than 1% of tested individuals per tissue, as has been previously described in (*17*). These genes are likely to have poor V^G estimates due to the presence of different ASE patterns within the gene or other global biological factors. Outlier individuals were also defined as in (*17*), and were removed from downstream single tissue analysis. These samples contain an unusually high number of outliers (n > Q3+1.5*IQR) at 5% FDR, and are likely to be caused by technical errors. Tissue specific lists of global outlier genes and individuals for outlier threshold of 5% FDR are here:(*59*). In all other analyses unless otherwise specified, we did not apply an FDR control procedure and instead imposed a higher threshold for declaring significance, to be consistent with expression and splicing outliers. For cross-tissue analyses, we calculated median ANEVA-DOT p-values for genes which were expressed in more than 5 tissues, without removing known global outliers first. Therefore, to account for genes with poor V^G, we applied the filtering steps described in (*17*) on the resulting individual-level median p-values. Briefly, we removed individuals with too few genes tested (n < Q1-1.5*IQR), removed individuals with too many outliers (n > Q3+1.5*IQR), and removed genes which appeared as outliers too many times across individuals with a score available (genes that are likely to be called as outliers in more than 1% of cases, Fig S2). To define multi-tissue outliers, we used a threshold of median p-value < 0.0027, equivalent to |median Z| > 3, to determine outlier status.

Split read count quantification and processing

LeafCutter (*35*) provided an annotation-free approach for RNA splicing quantification allowing us to capture split reads overlapping rare exon-exon junctions. Junctions were extracted from WASP-corrected BAM files with a modified version of the “bam2junc.sh” script from LeafCutter that only retained reads that passed WASP filters (*18*). Next in each tissue separately, junction reads were clustered using the “leafcutter_cluster.py” script from LeafCutter, with the options “--maxintronlen 500000” and “mincluratio 0”. LeafCutter assigns exon-exon junctions into mutually exclusive sets, termed clusters. Each exon-exon junction in a cluster had to share a splice site with at least one other exon-exon junction in that cluster, but could not share a splice site with an exon-exon junction from another cluster. A cluster had to contain at least two exon-exon junctions.

Next, in each tissue separately, we applied the following series of custom filters to the LeafCutter results in order to remove exon-exon junctions with low expression while retaining rare exon-exon junctions:

1. Removed exon-exon junctions where no sample has >= 15 split reads
2. Re-defined LeafCutter cluster assignments after removal of exon-exon junctions (according to the above filter) and removed exon-exon junctions that no longer shared a splice site with any other exon-exon junction.
3. Removed all exon-exon junctions belonging to a LeafCutter cluster where more than 10% of the samples had less than 3 reads summed across all exon-exon junctions assigned to that LeafCutter cluster.

Next, we merged LeafCutter cluster assignments across all 49 tissues to make a specific LeafCutter cluster comparable across multiple tissues. For this, we re-defined LeafCutter cluster assignments using the union of all exon-exon junctions that passed the above filters across 49 tissues. Lastly, we mapped our LeafCutter clusters to genes by intersecting splice sites, defining a Leafcutter cluster with splice sites of annotated exons. We limited to genes used in expression outlier calling (described in “Expression outlier calling” section). If an annotated splice site was in a LeafCutter cluster, we considered the LeafCutter cluster mapped to the gene. It was therefore possible for a LeafCutter cluster to map to multiple genes. We filtered LeafCutter clusters, and their corresponding exon-exon junctions, to those that were mapped to at least one gene. Finally, we removed any LeafCutter clusters with more than 20 exon-exon junctions due to computational limitations of SPOT.

SPOT: Overview

sOutliers were identified separately for each LeafCutter cluster in each tissue using Splicing Outlier deTection (SPOT). For a given LeafCutter cluster in a given tissue, we defined a matrix, X (dim NxJ), where each row corresponds to one of N samples, each column corresponding to one of J exon-exon junctions, and each element was the number of raw split read counts corresponding to that row’s sample and that column’s exon-exon junction. We were able to compute a p-value representing how abnormal a given sample’s splicing patterns were for the given LeafCutter cluster as follows:

1. Fitted parameters of Dirichlet-Multinomial distribution based on observed data X in order to capture the distribution of split read counts mapping to this LeafCutter cluster
2. Used fitted Dichlet-Multinomial distribution to compute the Mahalanobis distance for each of the N samples
3. Computed Mahalanobis distance for 1,000,000 samples simulated from the fitted Dirichlet-Multinomial and use these 1,000,000 Mahalanobis distances as an empirical distribution to assess the significance of the N real Mahalanobis distances

SPOT: Dirichlet-Multinomial parameter estimation

We defined a Dirichlet-Multinomial (DM) probability distribution based on data from N samples to capture the probability that a split read would map to each of the J junctions in the Leafcutter cluster:

Let x_{nj} be the raw number of split reads mapped to the j[th] junction in the n[th] sample and $t_n = \sum_{j=1}^{J} x_{nj}$ be the total number of split reads mapped to any junction in this LeafCutter cluster in the nth sample. Then

$x_{n1}, ..., x_{nJ} \mid t_n \sim DM(t_n, \alpha_1 p_1, ..., \alpha_J p_J)$ where $p_j = \frac{exp(c_j)}{\sum_{k=1}^{J} exp(c_k)}$

We used the following non-informative Gamma prior distribution to stabilize optimization:

$\alpha_j \sim Gamma(1 + 1e^{-4}, 1e^{-4})$

We then performed maximum likelihood estimation (via LBFGS as implemented in STAN) to learn the optimal parameter settings of $\alpha_1, ..., \alpha_J$ and $c_1, ..., c_J$ ($\hat{\alpha_1}, ..., \hat{\alpha_J}$ and $\hat{c_1}, ..., \hat{c_J}$) from the N samples. We were able to also deterministically compute the optimal values of each p_j ($\hat{p_j}$) from each $\hat{c_j}$.

SPOT: Mahalanobis distance

The Mahalanobis distance is the multivariate generalization of how many standard deviations a point is from the mean taking into account the covariance structure. After learning the parameters of the Dirichlet-Multinomial distribution for a specific LeafCutter cluster (ie $\hat{\alpha_1}, ..., \hat{\alpha_J}$ and $\hat{c_1}, ..., \hat{c_J}$; see "SPOT: Dirichlet-Multinomial parameter estimation"), we were able to compute the mean vector (μ_n) and covariance matrix (Σ_n) for a specific sample n, according to the Dirichlet-Multinomial. Using μ_n and Σ_n we were able to compute the Mahalanobis distance of sample n (MD_n). The covariance matrix of the Dirichlet-Multinomial (Σ_n) is of rank $J - 1$ because one of the dimensions is always a linear combination of the other $J - 1$ dimensions. As such, we approximated Σ_n^{-1} with the pseudo-inverse of Σ_n when computing the Mahalanobis distance.

SPOT: Empirical distribution to assess significance

For a given LeafCutter cluster, we have already computed the Mahalanobis distance of each of the N samples according to the fitted Dirichlet-Multinomial distribution for that LeafCutter cluster. However, the Mahalanobis distance is biased by the dimensionality of the space (i.e. the number of junctions assigned to the LeafCutter cluster). In order to convert the Mahalanobis distance to a test statistic that was not biased by dimensionality, we simulated an empirical distribution of Mahalanobis distances for each LeafCutter cluster. Specifically, for one LeafCutter cluster we drew 1,000,000 random samples from the fitted Dirichlet-Multinomial distribution assuming each of these random samples has 20,000 reads mapped to the LeafCutter cluster $(t_n = 20000)$. We then computed the Mahalanobis distance of each of these 1,000,000 samples and used the 1,000,000 Mahalanobis distances as an empirical distribution that converted our N Mahalanobis distances (from the real data) into p-values.

SPOT: Gene level correction

To compute a splicing outlier p-value for a gene associated with C LeafCutter clusters, we first computed minimum p-value across all C clusters for the gene. However, the minimum of a list of p-values is not a valid p-value. To address this, we computed the probability of observing a minimum p-value according to a probability density function defining the minimum across C independent uniform random variables between 0 and 1:

$$p(min(pvalue_1, \ ... \ , pvalue_C) <= z) = 1 - (1 - z)^C$$

This approach made the conservative, simplifying assumption that all clusters mapped to a gene were independent of one another.

We excluded individuals (global outliers) where the proportion of tested genes that were multi-tissue outliers (at a threshold of median p-value < .0027) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

SPOT: Robustness to hyperparameter choice

SPOT, under default settings, makes the assumption that each random sample, used in generating an empirical distribution for each LeafCutter cluster, has 20,000 total reads mapped to that cluster (see “SPOT: Empirical distribution to access significance”). To understand if our sOutlier p-values were sensitive to the choice of 20,000 total reads, we re-computed sOutlier calls in Muscle-Skeletal tissue using SPOT with 10,000 total reads and 100,000 total reads (Fig S6). sOutlier p-values generated from SPOT under default settings (20,000 reads) are highly correlated to sOutlier p-values generated from SPOT using 10,000 reads (Spearman’s ρ = .997) and 100,000 reads (Spearman’s ρ = .997). Only .052% and .046% of sample-LeafCutter cluster pairs had a -log10(p-value) change greater than 1 between SPOT under default settings compared to SPOT run with 10,000 and 100,000 reads, respectively. All of the sample-LeafCutter cluster pairs that had a -log10(p-value) change greater than 1 correspond to LeafCutter clusters where more than 95% of the total observed reads mapping to the cluster, summed across samples, map to a single exon-exon junction. These rare instances of

divergence in sOutlier p-values between SPOT under different hyperparameter settings are caused by numerical instability in computing the pseudo-inverse (See “SPOT: Mahalanobis distance”) when distributions are heavily skewed towards a particular junction.

SPOT, under default settings, uses a Gamma prior (on each α_j) when fitting a Dirichlet-Multinomial distribution to each LeafCutter cluster (See “SPOT: Dirichlet-Multinomial parameter estimation”). This prior is intended to stabilize the LBFGS-based optimization routine, while having minimal consequences on parameter estimates. To see if the prior had minimal impact on parameter estimates, we re-computed sOutlier calls in Muscle-Skeletal tissue using a version of SPOT where no prior was used (Fig S6). To encourage SPOT with no prior to converge to a reasonable estimate, we performed Dirichlet-Multinomial parameter estimation 10 times (with 10 random initializations) and selected the Dirichlet-Multinomial parameter estimate whose expected value had the smallest Euclidean norm with expected value of the maximum likelihood estimate of a Multinomial distribution fitted to the same data. sOutlier p-values generated from SPOT using default settings (ie. with the prior) are highly correlated to sOutlier p-values generated from SPOT when no prior is used (Spearman’s ρ = .997) .Only .049% of sample-LeafCutter cluster pairs had a -log10(p-value) change greater than 1 between SPOT under default settings compared to SPOT with no prior. Similar to the above comparison of SPOT using variable number of simulated reads, these rare instances of divergence in sOutlier p-values between SPOT with and without a prior are caused by numerical instability in computing the pseudo-inverse when distributions are heavily skewed towards a particular junction.

Outlier sharing across single tissues

Among all individual-gene outliers across all methods in a discovery tissue, we calculated the percentage of times the same individual-gene pair was detected as an outlier (nominal p-value < 0.0027) in a test tissue, limiting to tissues where both genes are expressed. We then aggregated this calculation across all individuals and genes (Figs 3A, S21). We assess this both when limiting to only the genes tested in both tissues, to answer the biological question of how consistent the outlier status is across tissues that co-express a gene, and when considering a missing datapoint as a non-shared outlier instance, addressing the utility of each method in diagnosing expression outlier status in a tissue of interest using a different tissue as a proxy.

For each of three clinically accessible tissues, whole blood, fibroblasts and lymphoblastoid cells (LCLs), we assessed the proportion of single tissue outliers (|Z| > 3, SPOT p-value < 0.0027 or ANEVA-DOT p-value < 0.0027) that replicate at the same threshold in each of the other 46 tissues, restricting to genes expressed in both tissues (Fig S22). We also restricted to outliers of each type that were seen in more than 1 of the three clinically accessible tissues, at the same thresholds, and assessed the proportion that replicate in each of the other tissues, again filtering each time to genes that are also measured in the replication tissue.

Enrichment calculations

We calculated relative risk enrichments as the proportion of outliers with a given variant type nearby the outlier gene over the proportion of non-outlier individuals with the given variant type nearby the same set of genes. We included 95% confidence intervals estimated via a normal approximation. When assessing rare variant enrichments overall and by category, we used a 10kb +/- window around the gene body. When considering variant categories per outlier, if more than one rare variant was present nearby the outlier gene, we assigned each gene-individual to a single variant category based on the following ordering: duplications (DUP), copy number variations (CNV), deletions (DEL), breakend (BND), inversions (INV), transposable elements (TE), splice, frameshift, stop, transcription start site (TSS), conserved non-coding, coding, or other non-coding, and subsetted to the 527 individuals with structural variant calls. Unless otherwise specified, we used a threshold of median p-value < 0.0027 (chosen to match |median Z-score| > 3) to define multi-tissue outliers, though provide results over a range of thresholds (Fig S1). A categorical model of outlier status was used as opposed to a continuous model because small changes in continuous outlier p-values do not reliably reflect true biological effects due to technical variation from RNA-sequencing, as well as to demonstrate the impact of thresholding choices for downstream applications. Additionally, this allows for matching the genes included in both the outlier and control category, defining an appropriate background distribution for statistical hypothesis testing so that we are not simply identifying differences between genomic regions rather than individual genetic effects on a given gene's expression. When considering variants in different windows upstream from the gene, we constructed exclusive distance ranges from each gene, beginning with the gene body + 10kb window used previously, and then we intersected rare variants with windows 1bp-200kb, 200kb-400kb, 400kb-600kb, 600kb-800kb, and 800kb-1Mb upstream from the set of outlier genes.

Alternative splicing enrichment calculations

We performed several enrichment analyses specific to splicing outliers to better characterize the variants underlying splicing outliers. For all of these analyses, we used sOutlier calls at the LeafCutter cluster level (instead of the gene level) in order to get more accurate enrichments. We excluded individuals identified as global outliers at the gene level (see "SPOT: gene level correction"). We limited enrichment analysis to SNVs. We used a stringent median p-value threshold of 1×10^{-5} in order to isolate the highest confidence instances of outlier splicing, according to SPOT. In Fig S17A, we show the relative risk of rare variants nearby splice sites is robust to a range of median p-value thresholds and becomes more enriched at more stringent p-value thresholds. .

1. **Relative risk of rare variant in window around splice site**. We computed the relative risk of rare variants being located at various windows around splice sites for outlier clusters relative to non-outlier clusters. For example, if the window was [0,2], we mapped a variant to a cluster if that variant were less than or equal to two base pairs away from observed donor and acceptor splice sites ([D-2, D+2] and [A-2, A+2] based on notation in Fig 2C) for that cluster. Relative risk was then calculated as the proportion of outlier (LeafCutter cluster, individual) pairs with a mapped rare variant over the proportion non-outlier (LeafCutter cluster, individual) pairs with a mapped rare variant, while limiting

analysis to LeafCutter clusters with at least one outlier individual. We included 95% confidence intervals estimated via a normal approximation.

2. **Relative risk of rare variant at position relative to splice site.** We first mapped rare variants to clusters if the rare variants were less than or equal to 1000 base pairs from an observed donor or acceptor splice site ([A-1000, A+1000] and [D-1000, D+1000] based on notation in Fig 2C). We then mapped each variant to its nearest splice site in that cluster and calculated its position relative to that splice site. Then, to compute the positional relative risk at position D-1 (for example), we computed the fraction of outlier variants mapped to a donor splice site that were at position D-1 over the fraction of non-outlier variants mapped to a donor splice site that were at position D-1. We added a constant of 1 to all counts in the contingency table to stabilize enrichments. We included 95% confidence intervals estimated via a normal approximation.
3. **Junction Usage for splicing median p-value outliers**. We used the "junction usage" statistic to quantify whether an individual used a splice site more or less than the background population. A positive junction usage value intuitively means the individual uses the splice site more than the background population, while a negative junction value means an individual uses a splice site less than the background population. More concretely to compute the junction usage for an individual i and junction j, we first computed the following ratio in each tissue (in which that individual i is expressed) separately: $\frac{Fraction\ of\ reads\ in\ cluster\ mapping\ to\ junction\ j\ for\ individual\ i}{Fraction\ of\ reads\ in\ cluster\ \ mapping\ to\ junction\ j\ for\ non-outliers\ individuals}$
   We added a constant of 1 to the above contingency table to stabilize enrichments. The "junction usage" statistic is simply the natural logarithm of the median of the above statistic across all tissues in which individual i is expressed.

Enrichment of outlier pairs within a given window

To test if nearby genes were more likely to share outlier status, we counted how many times two consecutive genes within a given genomic distance (defined based on the gene start position) in a given individual were both reported outliers. We considered multi-tissue outliers and analyzed each class of outliers independently. To derive the expected number of such occurrences, for each individual we used sampling without substitution to produce a random set of genes of the same size. Samples were drawn from a list of all genes that had been reported as an outlier at least once across all methods to avoid skewing the statistic by genes never reported as outliers. The expected value for each given window size was derived by averaging over all individuals. To ensure the stability of enrichment estimates at each window size, the sampling process was repeated until Monte Carlo error dropped below 10% of the expected number of outlier co-occurrences. For sOutliers this procedure was repeated once with all outlier genes included and once after removing 80 genes sharing a cluster with another outlier gene, see "Split read count quantification and processing".

We annotated all outliers occurring in a given window with the set of nearby rare variants for each gene in the pair. For each included variant category, defined above, we calculated a relative risk by taking the proportion of outlier pairs within the window for which one or both

genes had a rare variant in that category near the gene over the proportion of control individuals for which the same was true for the same gene, restricting to individuals with genetic data available. We included 95% confidence intervals estimated via a normal approximation, and we defined controls as individuals who were outliers for only one of the genes in the outlier pair.

Visualizing structural variants affecting multiple genes

To visualize rare, outlier-associated SVs as in Fig S15, we use the Integrative Genomics Viewer (IGV) tool (*71*, *72*). Because the SV calls were processed on the GTEx v7 release in hg19, we used RefSeq genes in hg19 coordinates and overlaid the SV start and end position to visualize the impacted genomic regions. To generate the plots in Fig S16, we used SAMtools (*73*) to subset the RNA-sequencing bam files to reads from the region spanning the two genes involved in the fusion for the individual carrying the rare deletion in two tissues, nerve tibial and lung, selected at random from the set of tissues with outlier signal in both genes, and two control individuals without the deletion, one for each tissue. After uploading the subsetted bam files to IGV, we selected “Sashimi plot” from the junctions track pop-up menu to display the reads spanning all junctions in the region of the rare deletion for both the outlier and control individuals. We only display read counts for the fusion transcript, but the line thickness correlates with read count for all other junctions.

Single-tissue rare variant enrichment

We tested for enrichment of rare variants near single-tissue gene expression outliers using the same variant list and relative risk enrichment definition as for cross-tissue outliers and with all individuals with both an expression outlier score and genotype information available. Under this definition of an eOutlier, a gene is only considered in one tissue at a time, i.e. without aggregating the gene's score across all tissues in an individual where it is expressed. Among ASE and splicing outliers, we removed tissue-specific global outlier genes prior to performing enrichment analysis. We converted expression Z-scores to a two-tailed z-test p-value for direct comparison to the other outlier methods. We tested for enrichment of rare variants at multiple, increasingly stringent significance thresholds for each individual tissue to ensure conclusions are not threshold dependent, and then reported the range of enrichment scores across all tissues, separated by outlier type and significance threshold. When assessing single-tissue outlier enrichments split by variant type, we use a more stringent threshold than with multi-tissue outliers of $|Z| > 4$ (p-value < 0.000063) as we do not have repeat measurements per individual-gene in this case.

Correlation tissue-specific expression outlier calling

We subsetted to a set of individuals and tissues with < 75% missingness, leading to 762 individuals and 29 tissues. We imputed missing expression values to improve our estimate of the tissue-by-tissue covariance matrix per gene that would be used in outlier calling. We used K-nearest neighbors in the impute R package (*53*) with k = 200 to impute values for missing tissues per individual on a gene by gene basis. We chose the value of k by comparing reconstruction error across k = [1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 80, 90, 100, 200, 300] on a set of 1000 randomly selected genes with 5% of individuals held-out for

evaluation. We tested several other potential imputation methods and saw similar performance (Fig S26), which included a multivariate normal expectation-maximization (EM), mean imputation (MEAN), soft thresholded iterated SVD imputation (SVD), and penalized matrix decomposition (PMD). For these additional imputation methods, we used the following parameters, determined in the same way as described above: EM - max iterations = 3 and tolerance = 1 x 10^{-6}, SVD - lambda = 20 and rank = 20, and PMD - lambda = 1 and rank = 5.

From the imputed matrix, we estimated the tissue covariance matrix, $\hat{\Sigma}$, for each gene. We calculated the Mahalanobis distance for each individual-gene pair as follows:

$$d^2 = x_g^T \hat{\Sigma}_g^{-1} x_g,\ d^2 \sim \chi_p^2$$

Where x_g is the vector of observed expression values for gene g across tissues, $\hat{\Sigma}_g$ is the estimated covariance matrix for gene g. We assigned a p-value to each gene-individual from the chi-squared distribution with degrees of freedom p equal to the number of tissues available for that individual. We used a two step correction procedure, first correcting via Bonferroni for all genes tested within an individual and then applying Benjamini-Hochberg correction across all tests with p < 0.0027. When assessing nearby rare variant enrichments, we removed genes that had an extreme number of outlier individuals, based on 3*IQR, as compared to the total set of tested genes. For the set of tissue-specific correlation outliers, we subsetted to outliers driven by a single tissue, requiring remaining available tissues for that individual to have a |Z-score| < 2 for the outlier gene.

Tissue-specific enhancer enrichments

We obtained tissue-specific enhancer annotations for 12 tissues from Epigenomics Roadmap (*74*) and mapped to GTEx tissues (Table S2). We subsetted to the tissue-specific correlation outliers that occurred in one of the 12 mapped tissues. To calculate the relative risk of a rare variant, including both SNVs and indels, in a tissue-matched enhancer, we took controls as all individual-gene pairs that were not correlation outliers and randomly assigned them to the same set of tissues as in the outlier group, matched by gene. We used any enhancer region annotated to a given tissue within a 500kb window around the outlier gene to capture the majority of potential enhancers, which can act at longer distances (*19*). We calculated matched enhancer enrichments as the proportion of tissue-specific outliers for which a rare variant fell within a nearby tissue-matched enhancer over the proportion of control individuals for which the same was true. For unmatched enhancer enrichments, we calculated the proportion of tissue-specific outliers with a rare variant falling in any tissue-specific enhancer region across the 12 tissues considered, without regard to the tissue driving the outlier call, within a 500kb window in either direction from the gene over the proportion of controls with a rare variant in any enhancer region within the same window of the same gene set.

Gene Ontology enrichment analysis

The list of genes with no outliers in any tissue was generated by taking the intersection of genes in each tissue that have no eOutlier, aseOutlier or sOutlier individuals, resulting in 261, 11,573

and 9,548 genes respectively. For the set of genes with extreme multi-tissue outliers, we take any genes with an individual having an absolute median Z-score exceeding 7 for eOutliers, or with an aseOutlier or sOutlier p-value < 2.56e-12, resulting in 127, 261 and 389 genes respectively. This threshold was chosen to retain enough outlier examples per outlier type to assess enrichments while focusing on the extremes of the distributions. For enrichments, we assess all Gene Ontology biological process terms (*75,76*) and use as background the complete set of 24,719 lincRNA and protein-coding genes used across outlier analyses. Significance was determined per gene set using FDR-corrected p-values from a Fisher's Exact test using PANTHER (*77*). We include the top ten terms by p-value in Fig S10 per gene set of interest without any filtering or combining of similar terms.

Watershed model overview

Watershed is a hierarchical Bayesian model that predicts regulatory effects of rare variants on a specific outlier signal based on the integration of multiple transcriptomic signals along with genomic annotations describing the rare variants. Watershed models instances of (gene, individual) pairs to predict the regulatory effects of rare variants nearby the gene. The Watershed model for a particular (gene, individual) pair, assuming K outlier signals, consists of three layers (Fig 4a):

1. A set of variables **G** $= G_1, ..., G_P$ representing the P observed genomic annotations aggregated over all rare variants in the individual that are nearby the gene.
2. A set of binary latent variables **Z** = $Z_1, ..., Z_K$ representing the unobserved functional regulatory status of the rare variants on each of the K outlier signals. Let Z^S be the set of all possible values that **Z** can take on. The size of Z^S is 2^K.
3. A set of categorical nodes **E** = $E_1, ..., E_K$ that represents the observed outlier status of the gene for each of the K outlier signals. We allow for missingness in **E**.

A fully connected conditional random field (CRF) (*78*) is defined over variables Z given G, where we let W represent the set edges among Z. Variables E_i are each connected only to the corresponding latent variable Z_i. Specifically, the following conditional distributions together define the full Watershed model:

A. $Z \mid G \sim CRF(\alpha, \beta_1, ..., \beta_k, \theta)$
B. $E_k \mid Z_k \sim \text{Categorical}(\varphi_k) \; \forall \; k \in K$
C. $\varphi_k \sim \text{Dirichlet}(C, ..., C)$
D. $\beta_k \sim \text{Normal}(0, \frac{1}{\lambda})$

where,

- $\beta_k \in R^P \; \forall \; k \in K$ are the parameters defining the contribution of the genomic annotations to the CRF for each outlier signal (k)
- $\alpha \in R^K$ are the parameters defining the intercept of the CRF for each outlier signal (k)
- $\theta \in R^{(K choose 2)}$ are the parameters defining the edge weights between pairs of outlier signals (Notational note: $\theta_{tq} = \theta_{qt}$)

- $\varphi_k \forall\, k \in K$ are the parameters defining the categorical distributions of each outlier signal
- C and λ are hyper-parameters of the model

Explicitly, our CRF probability distribution is defined as:

$$P(Z \mid G, \beta_1, ..., \beta_K, \alpha, \theta) = exp(\sum_{k \in K} \alpha_k Z_k + \sum_{(t,q) \in W} \theta_{tq} Z_t Z_q + \sum_{k \in K} \beta_k G Z_k - A(G, \theta, \beta_1, ..., \beta_K))$$

where $A(G, \theta, \beta_1, ..., \beta_K) = log(\sum_{Z^* \in Z^S} exp(\sum_{k \in K} \alpha_k Z^*_k + \sum_{(t,q) \in W} \theta_{tq} Z^*_t Z^*_q + \sum_{k \in K} \beta_k G Z^*_k))$

Because **Z** is unobserved, the Watershed log-likelihood objective over instances $n = 1, ..., N$:

$$\sum_{n=1}^{N} log \sum_{Z^* \in Z^S} P(E_n, G_n, Z^* \mid \beta_1, ..., \beta_K, \alpha, \theta, \varphi_1, ..., \varphi_K)$$

is non-convex. We therefore optimize model parameters using Expectation-Maximization (EM) as described in the following sections.

Watershed exact inference optimization routine

When the number of outlier signals (K) is small (an approximate rule being 4 or less), Watershed parameters can be optimized using exact inference updates within EM as follows:

In the E-step for instances $n = 1, ..., N$: we compute posterior distributions over the latent variables ($Z^{(n)}$), conditioned on the current model parameters ($\beta_1, ..., \beta_K, \alpha, \theta, \varphi_1, ..., \varphi_K$) and the observed data ($G^{(n)}$ and $E^{(n)}$). For example, the joint posterior probability of $Z^{(n)} = Z$ for the nth instance can be computed as:

$$\omega^{(n)}(Z^{(n)} = Z) = exp(\sum_{k \in K} (\alpha_k Z_k + \beta_k G^{(n)} Z_k + I(E_k^{(n)}) log(P(E_k^{(n)} \mid Z_k)) + \sum_{(t,q) \in W} \theta_{tq} Z_t Z_q - A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \theta, \varphi))$$

$$A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \varphi) = log(\sum_{Z^* \in Z^S} exp(\sum_{k \in K} (\alpha_k Z^*_k + \beta_k G^{(n)} Z^*_k + I(E_k^{(n)}) log(P(E_k^{(n)} \mid Z^*_k)) + \sum_{(t,q) \in W} \theta_{tq} Z^*_t Z^*_q))$$

where,

$I(E_k^{(n)})$ is an indicator function for whether $E_k^{(n)}$ is observed. Given the joint posterior probability distribution, we can marginalize (sum over) specific dimensions (outlier signals) to obtain:

1. Marginal posterior distributions for each dimension i (where Z^W is the set of all possible values that **Z** can take on excluding dimension i):

$$\omega^{(n)}{}_{single}(Z_i) = \sum_{Z^* \in Z^W} \omega^{(n)}(Z^*)$$

2. Pairwise marginal posterior distributions for each pair of dimensions i,j (where Z^W is the set of all possible values that **Z** can take on (excluding dimension i and dimension j)):

$$\omega^{(n)}{}_{pair}(Z_i, Z_j) = \sum_{Z^* \in Z^W} \omega^{(n)}(Z^*)$$

Both the marginal posterior distributions and the pairwise marginal posterior distributions are used in the M-step as follows. We update β, α, and θ by optimizing the conditional random field as follows:

$$argmax_{\beta, \alpha, \theta} \sum_{n=1}^{N} \sum_{Z^* \in Z^S} log(P(Z^* \mid G^{(n)}, \beta, \alpha, \theta))\ \omega^{(n)}(Z^*) - \tfrac{\lambda}{2}\|\beta\|_2 - \tfrac{\lambda}{2}\|\theta\|_2$$

Here λ is an L2 penalty hyper-parameter derived from the Gaussian priors on β and θ. We optimized this objective function by running L-BFGS on the closed-form gradient updates.

In the second part of the M-step, we update $\varphi_k \forall\ k \in K$ as follows:

$$\varphi_k(s, t) = \sum_{n=1}^{N} I(E_k^{(n)} = t)\ \omega^{(n)}{}_{single}(Z_k^{(n)} = s) + C$$

where,
I is an indicator operator, t is the categorical value of expression $E_k^{(n)}$, s is the possible binary values of $Z_k^{(n)}$, and C is the hyperparameter based on the Dirichlet prior on φ.

Once the EM algorithm has converged, we use the marginal posterior distributions for each dimension i in each instance n ($\omega^{(n)}{}_{single}(Z_i = 1)$) as estimates of probability that the nth (gene, individual) pair has a nearby variant that has a functional effect on the gene (with respect to outlier dimension i).

<u>Watershed approximate inference optimization routine</u>

When the number of outlier signals (K) is large (an approximate rule being 5 or more), it becomes computationally intractable to optimize Watershed parameters using exact inference updates, so we use approximate inference updates within EM as follows:

For the E-step, we wish to compute approximate estimates of the following posterior probability distribution:

$$\omega^{(n)}(Z^{(n)} = Z) = exp(\sum_{k \in K} (\alpha_k Z_k + \beta_k G^{(n)} Z_k + I(E_k^{(n)}) log(P(E_k^{(n)} \mid Z_k)) + \sum_{(t,q) \in W} \theta_{tq} Z_t Z_q - A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \theta, \varphi)$$

$$A(G^{(n)}, E^{(n)}, \theta, \beta, \alpha, \varphi) = log(\sum_{Z^* \in Z^S} exp(\sum_{k \in K} (\alpha_k Z^*_k + \beta_k G^{(n)} Z^*_k + I(E_k^{(n)}) log(P(E_k^{(n)} \mid Z^*_k))$$

$$+ \sum_{(t,q)\in W} \theta_{tq} Z^*_t Z^*_q))$$

To approximate this function $\omega^{(n)}(Z^{(n)})$, we use the Mean-Field Approximation (a subclass of Variational Inference) (*73*) and optimize $q^{(n)}(Z^{(n)})$ to minimize the KL-divergence between $q^{(n)}(Z^{(n)})$ and $\omega^{(n)}(Z^{(n)})$

where,

$q^{(n)}(Z^{(n)}) = \prod_{k\in K} q_k{}^{(n)}(Z_k{}^{(n)})$ where $q_k{}^{(n)}(Z_k{}^{(n)}) = (\mu_k{}^{(n)})^{z_k{}^{(n)}}(1-\mu_k{}^{(n)})^{(1-z_k{}^{(n)})}$

To minimize the KL-divergence for a given sample n, we perform coordinate descent on each $\mu_k{}^{(n)}$ while holding all other dimensions (values of $\mu_j{}^{(n)}$) constant. Given that $N(k)$ represents the set of all nodes that share an edge with node k, the variational update for each $\mu_k{}^{(n)}$ is then:

$$\mu_k{}^{(n)(update)} = \frac{exp(a_k + I(E_k{}^{(n)})log(P(E_k{}^{(n)} \mid Z_k{=}1)))}{exp\,(I(E_k{}^{(n)})log(P(E_k{}^{(n)} \mid Z_k{=}0)) + exp(a_k + I(E_k{}^{(n)})log(P(E_k{}^{(n)} \mid Z_k{=}1)))} \quad \text{where}$$

$$a_k = \alpha_k + \beta_k G^{(n)} + \sum_{j\in N(k)} \theta_{kj}\mu_j{}^{(n)}$$

More specifically, for one instance n, we iteratively do the following until convergence:

1. Loop through all K dimensions in a random order, and update each $\mu_k{}^{(n)}$ given the most recent values of $\mu_j{}^{(n)} \forall\, j \in N(k)$. Since coordinate ascent is not guaranteed to reach the global optimum, we used damped updates for each $\mu_k{}^{(n)}\ \forall\ k\ \in K$ in order to decrease the chance of getting stuck at a local optimum:
    a. $\mu_k{}^{(n)(iter\ i+1)} = (1-\eta) * \mu_k{}^{(n)(iter\ i)} + (\eta) * \mu_k{}^{(n)(update)}$
    b. We use a damping value (η) of 0.8.
2. Compute the average difference, across all K dimensions, between the values of $\mu_k{}^{(n)}$ from the current iteration and values of $\mu_k{}^{(n)}$ from the previous iteration. Converge if the average difference is less than 1x 10^{-8}.

Using the same notation as in "Watershed exact inference optimization routine", Mean Field allows us to approximate the following expectations using converged estimates of $\mu_k{}^{(n)}$:

1. $\omega^{(n)}(Z^{(n)}) \approx \prod_{k\in K} (\mu_k{}^{(n)})^{z_k{}^{(n)}}(1-\mu_k{}^{(n)})^{(1-z_k{}^{(n)})}$
2. $\omega^{(n)}{}_{pair}(Z_i{}^{(n)}, Z_j{}^{(n)}) \approx (\mu_i{}^{(n)})^{z_i{}^{(n)}}(1-\mu_i{}^{(n)})^{(1-z_i{}^{(n)})}\ (\mu_j{}^{(n)})^{z_j{}^{(n)}}(1-\mu_j{}^{(n)})^{(1-z_j{}^{(n)})}$
3. $\omega^{(n)}{}_{single}(Z_i{}^{(n)}) \approx (\mu_i{}^{(n)})^{z_i{}^{(n)}}(1-\mu_i{}^{(n)})^{(1-z_i{}^{(n)})}$

We use both the approximate marginal posterior distributions and the approximate pairwise marginal posterior distributions in the M-step. However, when the number of dimensions (K) is large, optimization of the parameters (β, α, and θ) defining the conditional random field becomes intractable. Therefore, we approximated the CRF objective function with the Pseudolikelihood (*80*) of the CRF. Given variational estimates of $\mu_i{}^{(n)}(Z_i{}^{(n)})$ for all values of dimensions (i) and all samples (n), the (log) Pseudolikelihood objective function (including priors on coefficients) is given by:

$$\sum_{n=1}^{N} \sum_{k \in K} (\alpha_k \mu_k^{(n)} + \beta_k G^{(n)} \mu_k^{(n)} + \sum_{j \in N(k)} \theta_{kj} \mu_k^{(n)} \mu_j^{(n)} - A(k, n, \theta, \beta, \alpha)) - \frac{\lambda}{2} \|\beta\|_2 - \frac{\lambda}{2} \|\theta\|_2$$

$$A(k, n, \theta, \beta, \alpha) = log(\sum_{z=0}^{1} exp(\alpha_k z + \beta_k G^{(n)} z + \sum_{j \in N(k)} \theta_{kj} z \mu_j^{(n)}))$$

We computed closed form gradient updates of the above objective function and then optimized it using L-BFGS.

In the second part of the M-step, we update $\varphi_k \forall k \in K$ as follows:

$$\varphi_k(s, t) = \sum_{n=1}^{N} I(E_k^{(n)} = t) \, \omega^{(n)}{}_{single}(Z_k^{(n)} = s) + C$$

Where I is an indicator operator, t is the categorical value of expression $E_k^{(n)}$, s is the possible binary values of $Z_k^{(n)}$, and C is the hyperparameter based on the Dirichlet prior on φ.

Once the EM algorithm has converged, we use marginal posterior distributions for each dimension i, in each instance n ($\omega^{(n)}{}_{single}(Z_i = 1)$) as estimates of probability that the nth (gene, individual) pair has a nearby variant that has a functional effect on the gene (with respect to outlier dimension i).

GAM and RIVER

The genomic annotation model (GAM) is L2-regularized logistic regression using genomic annotations (**G**) as features and the binary outlier status of a specific outlier signal as the response variable. One GAM model was trained for each outlier signal.

The only difference between Watershed and RIVER is that in RIVER θ is fixed to be a vector of zeros. This allows RIVER to be optimized precisely as described in “Watershed exact inference optimization routine” assuming θ is fixed to be zero. It is important to note that RIVER has changed slightly since its initial development (*15*) in the following way: we now use a categorical distribution (φ) with three categories instead of two to model $E \mid Z$. This change in RIVER was made in order to make it directly comparable to Watershed.

Applying Watershed to jointly model ASE, splicing, and expression

We first applied Watershed to the GTEx v8 data using 3 outlier signals: median ASE, splicing, and expression. Recall, Watershed requires a set of genomic annotations (**G**) and a corresponding set of categorical outlier signals (**E**) over (gene, individual) instances. We first limited to a set of (gene, individual) pairs with a rare variant that fell within the gene body or +/- 10kb of each gene and that passed the following set of filters in all 3 outlier signals:

1. The individual was not a global outlier
2. The gene has measured outlier signal for the corresponding individual

3. The gene has at least one individual that is an outlier (median p-value < .01)

This yielded a set of 36,702 (gene, individual) pairs that we used for training and evaluating the Watershed framework.

To generate the genomic annotations (**G**) for each (gene, individual) pair, we limited to SNVs that fell within the gene body or +/- 10kb of each of the gene and then extracted 47 genomic annotations (Table S3) describing each of the SNVs including regulatory element annotations, conservation scores, and derived genomic scores from other models such as CADD. If a (gene, individual) pair had more than one SNV mapped to the gene, the genomic annotations were aggregated across the SNVs with simple transformations to generate gene-level genomic annotations (Table S3). The resulting gene-level genomic annotations were standardized (mean 0 and standard deviation 1) before running Watershed.

We generated the categorical outlier signals (**E**) for each (gene, individual) pair using 3 categories per outlier signal. It is important to note that because of the filters described above there is no missingness in **E**. For aseOutliers and sOutliers, we assigned a gene with median p-value (p) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) < 1$
2. Category 2 if $1 <= -log_{10}(p + 10^{-6}) < 4$
3. Category 3 if $-log_{10}(p + 10^{-6}) >= 4$

For eOutliers, we assigned a gene with median p-value (p) and median Z-score (z) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) > 1$ and z < 0
2. Category 2 if $-log_{10}(p + 10^{-6}) <= 1$
3. Category 3 if $-log_{10}(p + 10^{-6}) > 1$ and z > 0

We note that these thresholds are arbitrary, but were selected to distinguish non-outliers, moderate outliers, and extreme outliers for aseOutliers and sOutliers, and distinguish non-outliers, under-expression outliers, and over-expression outliers for eOutliers.

To train and evaluate Watershed, we identified the 3,411 cases where two or more individuals had the same rare SNV(s) near a particular gene. We held out those instances and trained Watershed on the remaining instances. For training, we set the hyperparameter C equal to 30, motivated by the number of training instances. To select the hyperparameter λ, we trained and evaluated GAM on the training data for each outlier signal independently(assigning a sample an outlier label if outlier p-value < .01) with 5-fold cross validation while running a gridsearch on λ =.1,.01,.001. We selected the λ with the largest median area under the precision recall curve (AUPRC) across the 5 folds. Each precision recall curve aggregated predictions across the three outlier signals. The optimal λ was found to be 0.001. Before running Watershed, we initialized α_k and β_k to be the intercept and slope parameters, respectively, of GAM (when $\lambda = 0.001$) trained on the full training data for outlier signal k. θ was initialized to all zeros. φ_k

was initialized using the MAP updates described in “Watershed exact inference optimization routine”, except we used the GAM (when $\lambda = 0.001$) posterior probabilities to approximate $\omega^{(n)}_{single}(Z_k^{(n)} = s)$.

We evaluated various trained models (Watershed, RIVER, GAM, CADD) using the 3,411 cases where two individuals had the same rare SNV(s) near a particular gene (we will refer to these instances as N2 pairs). Specifically, we estimated the posterior probability of a functional rare variant (according to each of the models) in the first individual from the pair, allowing Watershed to use all data available for that individual. We then used the outlier status of the second individual as a ‘label’ for evaluation. In order to make the fraction of outlier instances comparable between different outlier signals, we defined a (gene, individual) pair to be an outlier for a specific outlier signal if its outlier p-value was ranked amongst the 1% most significant p-values for that outlier signal (across training and N2 pair instances). For an N2 pair, we did this evaluation in both directions: predict on the first individual and evaluate on the second, as well as predict on the second individual and evaluate on the first. Importantly, none of the N2 pairs were used in training any of the models.

Watershed with data generated using various filters

Recall from the previous section (“Applying Watershed to jointly model ASE, splicing, and expression”), Watershed training data was generated through the following approach: we limited to a set of (gene, individual) pairs with a rare variant that fell within the gene body or +/- 10kb of each gene and that passed the following set of filters in all 3 outlier signals:

1. The individual was not a global outlier
2. The gene has measured outlier signal for the corresponding individual
3. The gene has at least one individual that is an outlier (median p-value < 0.01)

These strict thresholds were set in order to reduce the imbalance between outliers and non-outliers in the training data set. We next assessed how sensitive Watershed was to these filters by training Watershed with three different training data sets generated by relaxing the above third filter as follows:

- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05)
- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1)
- At least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01)

We evaluated various trained models (Watershed, RIVER, GAM) using held out pairs of individuals generated under the default filtering in order to make precision-recall curves comparable to those in Fig 4D (Fig S28A-C, Table S4). We found the improvements of Watershed over RIVER decreased when using training data generated under more relaxed thresholds, while the improvements of Watershed and RIVER relative to GAM remained. The increased class imbalance (resulting from the relaxed thresholds) caused the fraction of positive training instances to decrease. This further imbalance resulted in Watershed learning considerably smaller magnitude edge weights, increasing the similarity of the Watershed model with the RIVER model.

We therefore recommend using training data generated through our default filtering approach when running Watershed.

We further accessed sensitivity of our analysis to these filters by training Watershed with training data generated through our default filtering approach, while evaluating Watershed on three different sets of held out pairs of individuals generated by relaxing the above third filter as follows:

- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05)
- All 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1)
- At least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01)

Importantly, improvements of Watershed over both RIVER and GAM were robust to relaxing these thresholds . Specifically, the difference in AUPRC between Watershed and RIVER, when evaluating performance on default held out pairs of individuals, is strictly bounded above zero for splicing, but for other phenotypes there is some overlap. But the difference in AUPRC between Watershed and RIVER is strictly bounded above zero for all phenotypes when evaluating on a larger set of held out pairs of individuals selected with less stringent filters (Table S4, Fig S28).

Applying Watershed to jointly model outlier signals from each tissue (tissue-Watershed)

Next, we trained three independent tissue-Watershed models (one each for ASE, splicing, and expression) where each model considered effects in all tissues, giving 49 phenotypes, corresponding to 49 Z and E variables each. In order for these models to be comparable to the model described in “Applying Watershed to jointly model three outlier types”, we used the same set of (gene, individual) pairs. We therefore used the same extracted and processed genomic annotations (**G**).

We generated the categorical outlier signals (**E**) for each (gene, individual) pair in a particular tissue (for a particular outlier signal) using 3 categories. It is important to note that, unlike the first application of Watershed to three median signals, there is now missingness in **E** as a (gene, individual) pair does not have measured outlier signal across all 49 tissues in GTEx. For ASE and splicing outliers, for a particular tissue, we assigned a gene with p-value (p) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) < 1$
2. Category 2 if $1 <= -log_{10}(p + 10^{-6}) < 4$
3. Category 3 if $-log_{10}(p + 10^{-6}) >= 4$

For expression, outliers, for a particular tissue, we assigned a gene with p-value (p) and Z-score (z) to:

1. Category 1 if $-log_{10}(p + 10^{-6}) > 1$ and z < 0
2. Category 2 if $-log_{10}(p + 10^{-6}) <= 1$
3. Category 3 if $-log_{10}(p + 10^{-6}) > 1$ and z > 0

To train and evaluate tissue- Watershed, we identified the 3,411 cases where two individuals had the same rare SNV(s) near a particular gene. We held out those instances and trained Watershed on the remaining instances. For training, we set the hyperparameter C equal to 10, motivated by the number of training instances with observed outlier calls. We selected $\lambda = 0.001$ based on cross-validation in "applying Watershed to jointly model three outlier types". We initialized α_t and β_t to be the intercept and slope parameters, respectively, of GAM (when $\lambda = 0.001$) trained on the full training data from tissue *t*. θ was initialized to all zeros. φ_t was initialized using the MAP updates described in "Watershed exact inference optimization routine", except we used the GAM (when $\lambda = 0.001$) posteriors to approximate $\omega^{(n)}{}_{single}(Z_k^{(n)} = s)$.

We took a very similar approach as described in "Applying Watershed to jointly model ASE, splicing and expression" to evaluate various trained models (tissue-Watershed, tissue-RIVER, tissue-GAM). In this setting however, both model predictions and outlier labels were in a single tissue as opposed to the median across tissues. As **E** contains missingness in this setting, we required both individuals in the N2 pair to have observed outlier signal for the gene of interest in the corresponding tissue.

Non-parametric bootstrapping of change in area under precision recall curves

We utilize non-parametric bootstrapping to assess the significance of the difference in area under a precision recall curve for two different models (assume the two models are called "model 1" and "model 2", respectively). Assume there are N observations involved in generating the precision-recall curves, meaning there exist N predictions from model 1, N predictions from model 2, and N binary labels. We can then compute the area under the precision recall curve for model 1 and model 2 ($auprc_1$ and $auprc_2$, respectively), as well as the difference between the areas ($\Delta auprc = auprc_1 - auprc_2$). Next, we generate B non-parametric bootstrapped samples of $\Delta auprc$. To generate one non-parametric bootstrapped sample (b) of $\Delta auprc$ we:

1. Randomly sample, with replacement N observations from the original N observations
2. Generate $auprc_1{}^{(b)}$ and $auprc_2{}^{(b)}$ according to the sub-sampled observations.
3. Compute $\Delta auprc^{(b)} = auprc_1{}^{(b)} - auprc_2{}^{(b)}$

We can compute a 95% confidence interval on $\Delta auprc$ using the B bootstrapped samples by first computing the .025 quantile and .975 quantile (across the B bootstrapped samples) of $\Delta auprc^{(b)} - \Delta auprc$ ($\delta_{.025}$ and $\delta_{.975}$, respectively). The 95% confidence interval is then $[\Delta auprc - \delta_{.975},\ \Delta auprc - \delta_{.025}]$.

Rare variant Watershed posterior predictions with trained Watershed model

We used the Watershed model that was previously trained on the 34,837 (gene, individual) pairs described in "Applying Watershed to jointly model ASE, splicing, and expression" to make Watershed posterior predictions on the remainder of rare variants in GTEx. To make genomic annotations comparable, the genomic annotations describing the SNVs we wish to predict on

were standardized according to the mean and standard deviation of the genomic annotations from “Applying Watershed to jointly model ASE, splicing, and expression”. It is important to note that the Watershed model was trained across (gene, individual) pairs and predictions were made across (gene, SNV, individual) triplets.

Note on applying Watershed to new data sets

While we are restricted here to making predictions of variant effect on transcriptomic signals, our framework, including enrichment analysis and Watershed, could be straightforwardly applied to ribosome profiling data and/or mass-spectrometry based protein measurements by researchers using a cohort with WGS or exome sequencing to capture post-translational and structural changes.

Replication in ASMAD Cohort

As previously reported (*40*), 394 family members were genotyped on Illumina Omni 2.5 arrays and 80 individuals were subjected to whole genome sequencing by Complete Genomics. Genotyping was performed at the Center for Applied Genomics and the Children's Hospital of Pennsylvania. Genotype based identity by state metrics validate all familial relationships in the pedigree. All variants with Mendelian inconsistencies or missing in more than 1% of individuals were removed. Haplotypes were phased using SHAPEIT2 with duoHMM (*81*). Imputation was performed using IMPUTE2 (*82*) and the TopMed Anabaptist reference panel of haplotypes. LCL lines from 100 individuals of the pedigree were obtained from the Coriell Institute. These individuals represent the 80 individuals who had been whole genome sequenced, plus an additional 20 closely related individuals.

Total RNA was extracted from LCL cultures using RNAeasy. Paired end RNA sequencing libraries were constructed using the Illumina [TruSeq stranded mRNA library prep kit] (http://www.illumina.com/products/truseq_stranded_mrna_library_prep_kit.html) with 100 independent index barcodes. Paired, 125bp reads were generated on an Illumina HiSeq2500 at the Next Generation Sequencing Core Facility at the University of Pennsylvania. Read level quality was assessed using FastQC (*83*). Reads were trimmed to remove Illumina adapters and low quality sequence using TrimGalore! ('stringency 5, length 50, q 20') (*84*). Reads were aligned to the human genome (hg38) with GENCODE gene annotations (v24) using the STAR aligner (*57*) in 2-Pass mode. Gene level read counts were quantified using Feature Counts. After genotype and RNAseq quality control, 97 samples were included for further analysis.

To control for reference mapping bias and remove reads derived from PCR duplication, reads aligned to the human genome were processed using WASP (*85*). At each heterozygous site, reference and alternate allele read depth was quantified using PySam. Overlapping read pairs were only counted once. Splicing clusters were identified within each sample using Leafcutter (*35*).

aseOutlier calls in the ASMAD cohort were generated as follows. Allele specific read counts were generated with quasar. ASE snps were annotated by overlapping with coding regions of the genome. Then, for all ASE snps which overlapped a gene, the one with the highest read coverage was used to represent that gene's ASE counts. Mono-allelic sites and sites with fewer than 5 reads per allele were discarded. Genes which appeared as frequent outliers in GTEx LCL samples (available at (*59*)) were removed as well. ANEVA-DOT was then run on all available genes per individual, using LCL V^G scores from GTEx as the reference. The results across all 97 available samples were compared, and individuals with more than 61 ASE outliers, after FDR correction (11 in total), were removed from downstream analysis. On average an individual in the ASMAD cohort had 176 ASE outlier genes, before FDR correction.

We next called sOutliers in the ASMAD cohort. As there are relatively few ASMAD RNA-seq samples (n=97), we used Dirichlet-Multinomial parameter estimates for each LeafCutter cluster learned from GTEx Cells EBV-transformed lymphocyte samples and then assessed how extreme each ASMAD sample was according that pre-trained distribution. More specifically, we first filtered ASMAD exon-exon junction counts to exon-exon junctions that passed the filters involved in processing GTEx Cells EBV-transformed lymphocytes (see "Split read count quantification and processing"). Then for each Leafcutter Cluster tested with SPOT in the GTEx Cells EBV-transformed lymphocytes tissue, we:

1. Retrieved Dirichlet-Multinomial parameter estimates for this LeafCutter cluster from when SPOT was trained using GTEx Cells EBV-transformed lymphocytes samples.
2. Generated a junction count matrix for the ASMAD samples. This junction count matrix will be of dimension $N \text{ X } J$ where N is the number of ASMAD samples and J is the number of junctions assigned to this tissue in GTEx Cells EBV-transformed lymphocytes. If a particular junction in this cluster is not expressed in the ASMAD cohort, the column corresponding to this junction in the matrix will be filled in with zeros.
3. Used the GTEx-fitted Dirichlet-Multinomial distribution (from step 1) to compute the Mahalanobis distance of each of the N ASMAD samples.
4. Computed Mahalanobis distance for 1,000,000 samples simulated from the fitted Dirichlet-Multinomial and used these 1,000,000 Mahalanobis distances as an empirical distribution to assess the significance of the N real Mahalanobis distances.

We then converted from ASMAD sOutlier p-values at the LeafCutter cluster level to sOutlier p-values at the gene level using the approach described in "SPOT: Gene level correction". We excluded individuals (global outliers) where the proportion of tested genes that were outliers (at a threshold of p-value < .0027) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

Finally, we called eOutliers in the ASMAD cohort. As there are relatively few ASMAD RNA-seq samples (n=97), we concatenated ASMAD samples and GTEx Cells EBV-transformed lymphocyte samples and called eOutliers across the concatenated samples. More specifically, we first computed the TPM of each sample-gene pair independently for the ASMAD samples

and the GTEx Cells EBV-transformed lymphocyte samples (using transcript lengths specific to each study). Next we concatenated the two TPM matrices into one large TPM matrix of dimension N X G where N is the sum of the number of samples in ASMAD and the number of samples in GTEx Cells EBV-transformed lymphocyte tissue, and G is the number of genes used in the GTEx Cells EBV-transformed lymphocyte eOutlier analysis. We then filtered to genes where at least 10% of the total samples (N) have greater than or equal to 6 raw counts and have greater than .1 TPM. We next log2-transformed the expression values ($\log_2(\text{TPM} + 2)$). We then scaled the expression of each gene to have mean 0 and standard deviation of 1, regressed out the top 30 principal components, and finally standardized each gene, again, to have mean 0 and standard deviation 1. We excluded individuals (global outliers) where the proportion of tested genes that were outliers (at a threshold of |Z-score| > 3) exceeded 1.5 times the interquartile range of the distribution of proportion outlier genes across all individuals.

Massively Parallel Reporter Assay (MPRA) variant selection and sequence design

We selected 1144 individual-gene pairs which were called as multi-tissue eOutliers in GTEx v6p (*15*). We removed any outliers with any of the following: a rare SV within 200kb of the TSS, a rare indel within 10kb of the TSS, any rare coding SNV. Then, we required that all outliers have at least one rare, non-coding SNV within 10kb of the TSS. This procedure yields 194 multi-tissue outlier individual-gene pairs. From this set, we obtained all rare, non-coding SNVs at each outlier gene in its respective individual yielding a set of 284 variants (with a median of 1 and mean of 1.46 per individual-gene pair).

To obtain a set of control variants, for the same 194 genes derived from the individual-gene pairs noted above we found all individuals-gene pairs with |median Z| < 0.5. This yields 14303 individual-gene pairs across the 194 genes. Then, we apply the same filters as in the outlier variant set, yielding 3744 control individual-gene pairs across 193 genes. Finally, we obtain the set of all rare variants found in those control individual-gene pairs and, for each outlier variant, obtain only the closest control variant. These steps yield 271 outlier and 248 control variants.

We designed a set of synthetic DNA fragments by retrieving the genomic sequence corresponding to a 150bp window centered at each variant of interest (and additional flanking constant sequences for cloning). For each variant a reference and alternative sequence was designed, corresponding to each allele, and in cases where multiple variants were in the same window, each possible combination was included (2^n sequences where n is the number of variants). This procedure yielded a set of 1108 unique genomic sequences which were obtained as an oligonucleotide library pool from Agilent Technologies.

MPRA plasmid library construction

We prepared a randomly barcoded library broadly as described in (*55*). To summarize, the oligonucleotide library was randomly barcoded using emulsion PCR as described in (*55*, *86*), with 96 reactions in 100 uL volume each which were pooled prior to purification. The pMPRA1 plasmid was obtained from Addgene (Plasmid #49349) and digested with SfiI to obtain the

plasmid backbone. The purified plasmid backbone and randomly barcoded library were ligated using Gibson assembly (*87*), electroporated into 10-beta *E coli* in 8 parallel reactions, recovered overnight, and pooled. The expanded plasmid library was isolated by Qiagen Midiprep and prepared for sequencing to determine oligo-barcode mappings.

To introduce a minimal reporter-GFP-partial 3'UTR construct, we digested the barcoded oligo plasmid library with AsiSI, cutting between the oligo sequence and barcode. We generated a minp-GFP amplicon by PCR from the pGL4.26-SS-136 vector (Addgene Plasmid #68744) and inserted it into the linearized plasmid library by Gibson assembly. Following SPRI purification, the library was re-digested with AsiSI and Exonuclease V (to remove plasmids lacking a reporter) and electroporated into 10-beta cells in eight parallel reactions. Cultures were recovered for 12 hours, pooled, and purified by Qiagen Gigaprep. The final purified reporter library was used for transfection and direct amplification to assess plasmid-level oligo frequencies.

MPRA cell culture and transfection

GM12878 cells were cultured in RPMI supplemented with 15% FBS and 1% penicillin/streptomycin maintaining a density of 0.2-1.0x10^6 cells/mL at 37C and 5% CO2. For each biological replicate, 5x10^7 cells were collected by centrifugation, washed with PBS, and resuspended in 5 mL RPMI containing 60 ug of plasmid library. Cell suspensions were serially electroporated in 100 uL volumes using a Lonza Nucleofector with program T-25 in 2mm cuvettes. Immediately after the pulse, cells were washed into 50 mL warm RPMI with 15% FBS (without antibiotics) and recovered for 24 hours at 37C. After visually verifying heterogenous GFP expression, cells were collected, washed, and frozen at -80C.

MPRA reporter mRNA isolation and normalization

Total RNA was isolated using 4 mL Trizol per replicate following the standard protocol. DNase digestion was performed on all recovered RNA to prevent DNA contamination in the following reaction: 5 uL Turbo DNase + 75 uL 10X Buffer for 60 minutes at 37C, followed by quenching with 75 uL EDTA and 7.5 uL 10% SDS and then the DNase-digested RNA was SPRI purified. The total eluent was diluted to 1 mL in water and used as input for GFP mRNA isolation via solution hybridization to biotinylated antisense oligonucleotides with the following components added: 1 mL 20X SSC, 2 mL formamide, and 2 uM Biotin-anti-GFP probe. GFP mRNA was isolated on magnetic streptavidin beads following the manufacturer's protocol, washed, and subjected to a second DNase digestion under the same conditions as the first. Following another SPRI purification, mRNA was reverse transcribed using the SuperScript III enzyme and a gene-specific primer according to the manufacturer's instructions, SPRI purified, and quantified using a Qubit fluorometer. Replicates of the input plasmid library were diluted to approximately match the concentrations of the GFP cDNA samples, and all plasmid and cDNA samples were collectively quantified and normalized by qPCR in the following reaction: 5 uL Q5 NEBNext MasterMix, 1 uL Sybr Green (1:1000), 0.5 uM forward and reverse primers, and 1uL cDNA or plasmid-DNA sample. Samples were amplified until saturation using the following

conditions: 95C for 20 seconds followed by 40 cycles of 95C for 20s - 65C for 20s - 72C for 30s. Samples were normalized by dilution to the least concentrated measurement.

MPRA library preparation and sequencing

To prepare libraries to map oligo-barcode pairings by 2x150 sequencing: we designed an amplicon corresponding to the 223bp sequence covering the 150bp genomic sequence, the intervening constant sequence, and the 20bp random barcode. This amplicon was generated by PCR under the following conditions: 95C for 20 s, (95C for 20s, 65C for 30s, 72C for 30c) for 6 cycles, and 72C for 2 min. After SPRI purification, Illumina adapters were attached by PCR, the product purified by SPRI, and assessed by Nanodrop, Qubit, and Bioanalyzer. The final library was sequenced on an Illumina NextSeq instrument using a 2x150 high-output kit.

To prepare libraries to quantify plasmid (DNA) and cDNA (RNA) barcodes and oligos by 1x30 sequencing: we performed PCR targeting the GFP 3' UTR including the random barcode under the same conditions as qPCR quantification increasing reaction volume to 50 uL and decreasing cycles to 12. After SPRI purification, each sample was input into another PCR reaction to attach Illumina adapters and multiplexing indices and purified again. These final libraries were assessed by Nanodrop, Qubit, and Bioanalyzer and pooled according to their Biolanalyzer molarity estimates. The pooled libraries were sequenced on an Illumina NextSeq instrument using 30 cycles of a 1x75 high-output kit.

Quantifying reporter activity across sequences in MPRA results

To assemble oligo-barcode pairings, we merged all paired-end reads using FLASH2 (*56*), requiring a minimum 10bp overlap to retain each pair. Then, extracted the regions of each fragment corresponding to the genomic sequence, the flanking constant sequences, and the random barcode. Sequences corresponding to genomic fragments were mapped using STAR (*57*) against a reference assembled using the designed oligo library sequences. We required that each sequence map uniquely and perfectly to retain the pair, and thus associated the 20bp barcode with the designed sequence. We filtered out any barcode-oligo pairings that were mutually incompatible (i.e. the same barcode point to two oligos), contained errors, or otherwise did not match the expected fragment model.

To count reads per unique barcode sequence, we took raw single-end reads, extracted the 20bp region corresponding to the random barcode, and counted the number of reads per unique sequence using fastx_collapser. We required that all barcodes perfectly match a barcode detected in the barcode-oligo pairing data. Finally, to generate oligo-level read counts, we computed the sum of all barcodes for each oligo (along with other summaries of the counts) within each sample.

Modeling and inference of regulatory signal in MPRA results

We used negative binomial regression with an interaction term, implemented via DESeq2 (*58*), to identify significant allele-independent and allele-dependent regulatory effects. Specifically, we

obtain twelve measurements for each variant position (three each of DNA-ref, DNA-alt, RNA-ref, and RNA-alt) and model the read counts as:

[count ~ 1 + material + allele + material:allele]

We apply the DESeq2 model across the oligo-level counts across three plasmid and three cDNA replicates, and compute Wald tests on each coefficient. A significant material term indicates an allele-independent regulatory effect, a significant allele term indicates a significant (but irrelevant) cloning artifact, and a significant interaction term indicates an allele-dependent effect. Thus, we are mostly interested in tests on the third term, but also the first. We took as our set of “expression hits” and “allelic hits” all variants which had an adjusted p-value <= 0.05 in either the first or third terms, respectively.

Allele Specific CRISPR Assay for functional validation

To perform functional validation, we selected 14 rare stop-gained variants which were good candidates for the CRISPR assay via (1) filtering to rare stop-gained variants with expression and ASE watershed scores > 0.9, (2) filtering to multi-tissue outlier status in both, and (3) keeping 4 remaining candidates which lie in complex trait genes, and the next 10 with the highest individual outlier signal and Watershed score. Variants were tested using the polyclonal editing assay described in (*41*). Briefly, inducible-Cas9 293T cells were transfected with a gRNA and single stranded homologous template specific to each variant. Nine days after transfection, cells were harvested for mRNA and gDNA which were amplified with sequencing adapter primers specific to the variant locus. Samples were run in parallel on the MiSeq and the proportion of alternative allele in both the mRNA and gDNA were calculated using EdiTyper (*88*) . Briefly, EdiTyper is a command line utility designed to process targeted sequencing data from CRISPR genome editing experiments. It applies quality filtering, performs sequence alignment using RecNW (*89*), a modified version of the Needleman-Wunsch algorithm, and classifies reads as containing the alternative or reference allele, discarding reads with indels indicative of non-homologous end joining. EdiTyper code is available at https://github.com/LappalainenLab/edityper.

Effect size was calculated as log2((Alt/Ref in cDNA) / (Alt/Ref in gDNA)), or allelic fold change (aFC) (*54*). Significance of the effect size was calculated with Bonferroni-corrected p values based on z-scores calculated from the distribution of a set of control variants which are not associated with expression of their respective genes in GTEx. Specifically, the non-eQTL negative controls were common synonymous variants in GTEx v8 with an eQTL association p > 0.1 with the gene in which they reside. Eight variants in total passed quality control steps. These results were combined with six previously tested stop-gained and six non-eQTL control variants for which Watershed posteriors were available.

UKBB and MVP GWAS integration

We assessed GWAS summary statistics from the UK Biobank (UKBB) phase 2, as made available by the Neale lab (http://www.nealelab.is/uk-biobank/). We subsetted the variants,

either genotyped or imputed, in UKBB phase 2 to those SNVs that also appeared in any GTEx individuals and had a frequency of < 1% in GTEx, which resulted in 45,415 SNVs, filtered to those not flagged as low confidence due to very low allele counts. Because we are targeting rare variants occurring at frequencies too low to obtain a trait association with genome-wide significance, we focused on the effect size estimates and did not filter by p-value. We defined outlier variants in this context as any rare variant appearing near an eOutlier, sOutlier, or aseOutlier in GTEx and also appearing in UKBB. We defined non-outlier variants as rare GTEx variants appearing in UKBB, but not falling near an outlier of any type, though within 10kb of a gene for which any individual was an outlier. We subsetted to 34 traits tested for colocalization between the UKBB GWAS and GTEx eQTL/sQTL studies as described in (*45*). When filtering to colocalized regions, we included as a colocalization event any gene that had a colocalization posterior probability > 0.5, for both eQTLs and sQTLs. We combine both enloc (*90*) and coloc (*91*) results for eQTL colocalization and enloc results for sQTL colocalization. This resulted in 5,386 gene-trait pairs with significant co-localizations across 34 UKBB traits (Table S9). We transformed the |effect sizes| to percentiles, based on all rare GTEx SNVs that also appear in any UKBB samples tested for the included traits. When showing actual beta values for binary traits, we scaled according to the case-control ratio μ for the given trait, dividing the effect size estimates by $\mu(1-\mu)$.

We filtered the set of GTEx rare variants in UKBB to those in colocalized regions, defined as being in a colocalized gene or within 10kb, and by the maximum Watershed posterior for that variant-gene combination across all data types (ASE, splicing, expression) and all tested individuals. We compared this to a genomic annotation based metric, CADD. We obtain an effect size β for both Watershed posterior and CADD score in predicting variant effect size percentiles in co-localized regions using the following model: $P \sim \beta X + \varepsilon$, where P is a vector of variant effect size percentiles and X is a vector of either Watershed posteriors or CADD scores for the same variant set.

We calculated the proportion of resulting variants that fall in the top 25% of effect sizes within colocalized regions for the associated trait across a range of posterior thresholds. We compared that proportion to the set we would obtain if filtering by a CADD score chosen to return an equal number of variants, prior to intersecting with colocalized regions. Additionally, we took 1000 random samples from the set of rare variants of an equal number to the actual number obtained by filtering at each threshold and assessed the proportion of random variants that fall in the top 25% of effect sizes for each colocalized trait. For replication in the Million Veterans Program (MVP) (*13*) and Jackson Heart Study (JHS) (*14*), we obtained summary statistics for a 250kb region on either side of the variant of interest for four lipid associated traits. We calculated the |effect size| percentile for all rare variants (gnomAD AF < 0.1%) in that region and plot the absolute effect sizes vs the gnomAD allele frequency.

Supplementary Figures

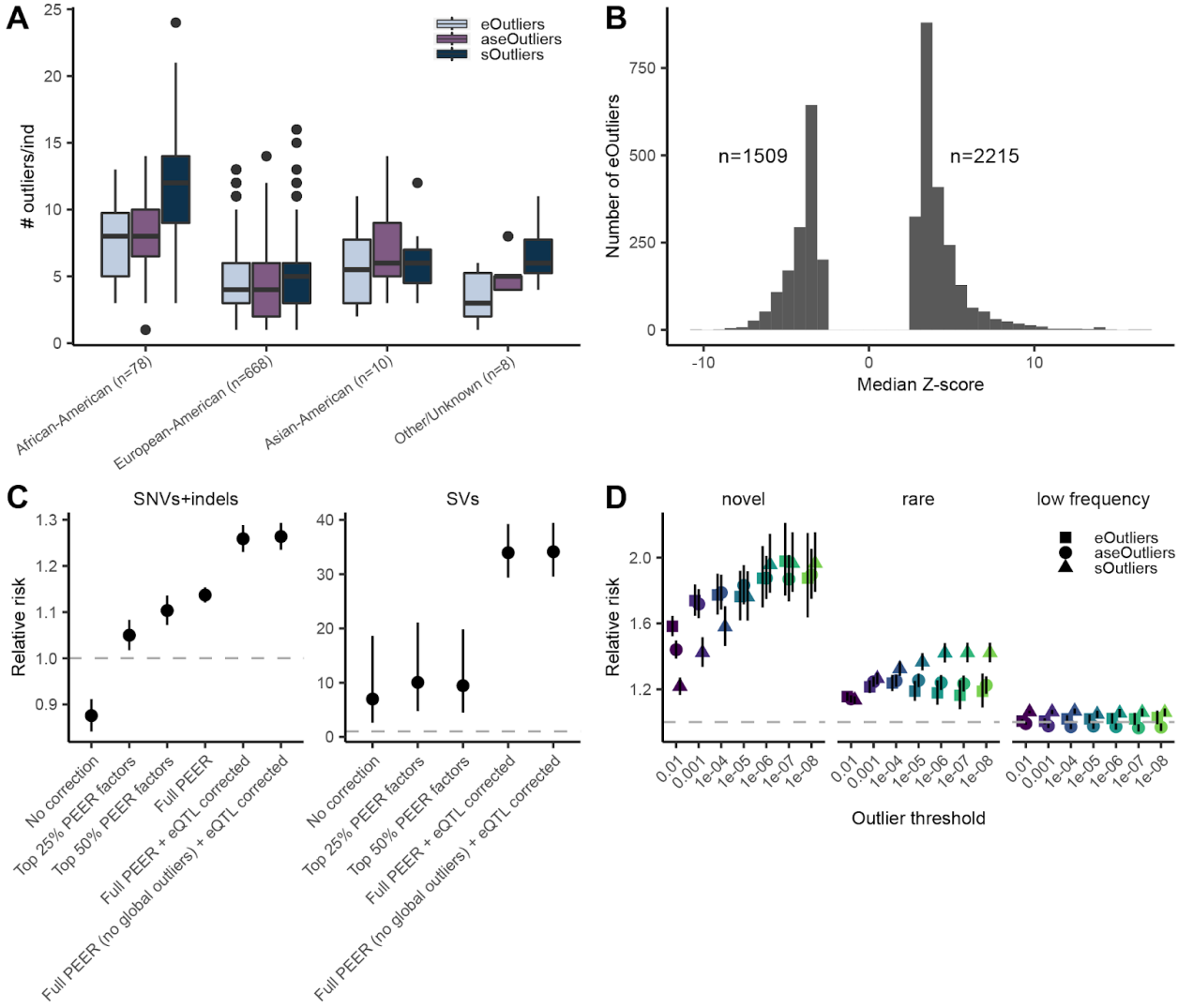

Figure S1. Outlier distribution and effect of expression data correction. (A) Number of outliers per individual across each population defined by self-reported ethnicity, at a threshold of median p-value < 0.0027. **(B)** Number of eOutliers split by direction of the expression effect. **(C)** Effect of different expression data correction procedures on relative risk of an outlier having a nearby rare variant. From left, rare (MAF < 1%) variant enrichments for eOutliers identified from uncorrected data, data corrected for first 25% of PEER factors (based on sample size), first 50% of PEER factors, full PEER factors and known covariates, all PEER factors + strongest cis-eQTL per gene, and all PEER factors learned with global outliers removed plus strongest cis-eQTL per gene. **(D)** Rare SNV and indel enrichments, defined as relative risk, for novel (left), rare (gnomAD AF < 1%), and low frequency (gnomAD AF > 1% and < 5%) within 10kb of outlier genes across a range of outlier thresholds (x-axis).

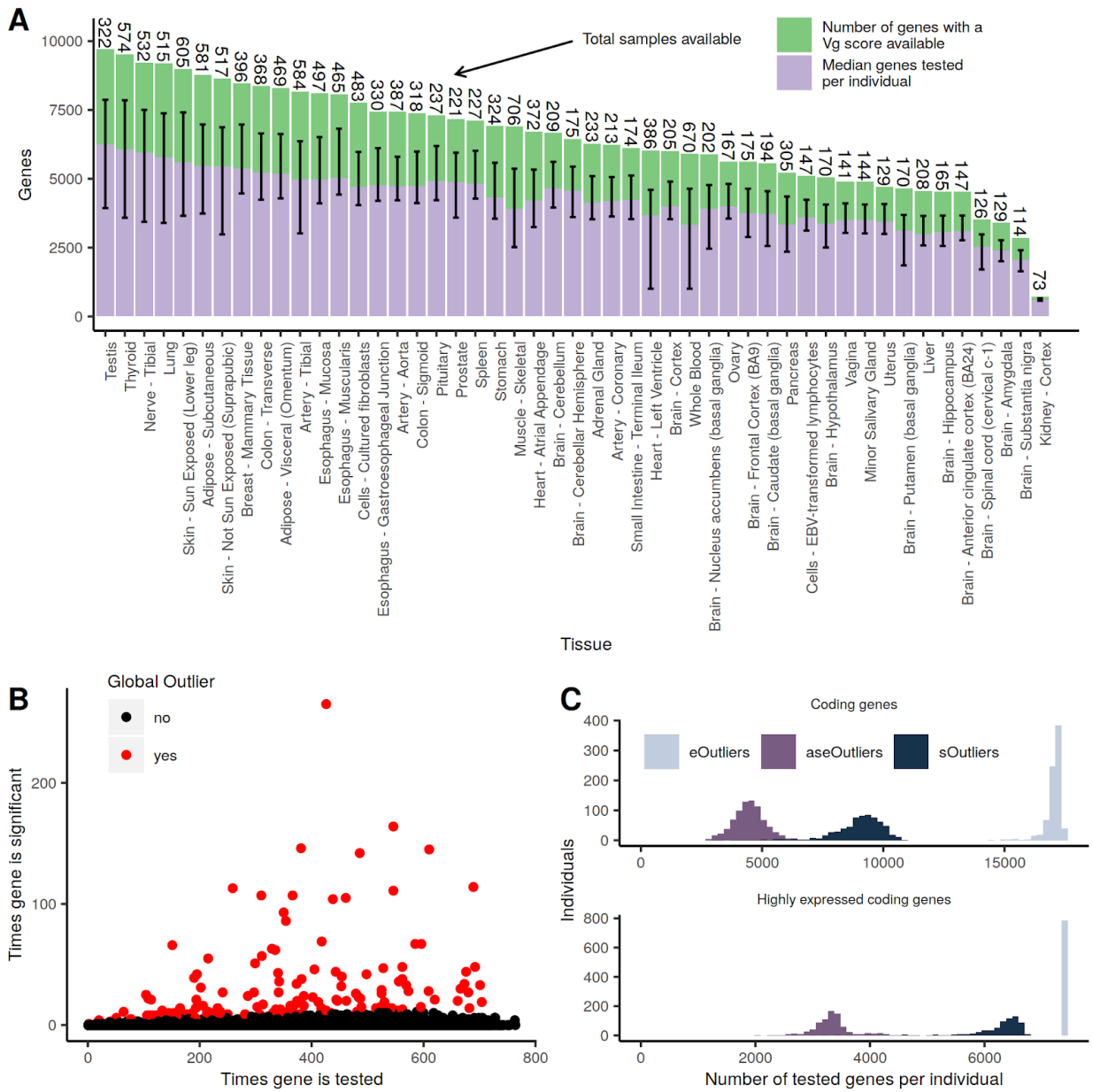

Figure S2. Quality control for ASE processing. (A) Average number of tests per individual tissue sample ± range. The total number of V^G scores available per tissue is shown above in green, with the total samples available per tissue. **(B)** The total number of times a gene was tested by considering its median ANEVA-DOT p-value vs the number of times it was called as an outlier. We call global outliers by drawing a 95% binomial confidence interval around the outlier frequency for each gene, and flagging all genes where the interval contains 1% or greater. Global outlier genes were removed from downstream analysis. **(C)** Distribution of median number of scores available across all three outlier methods, limiting to coding genes above, and coding genes with a median TPM > 10 across all individuals and tissues below.

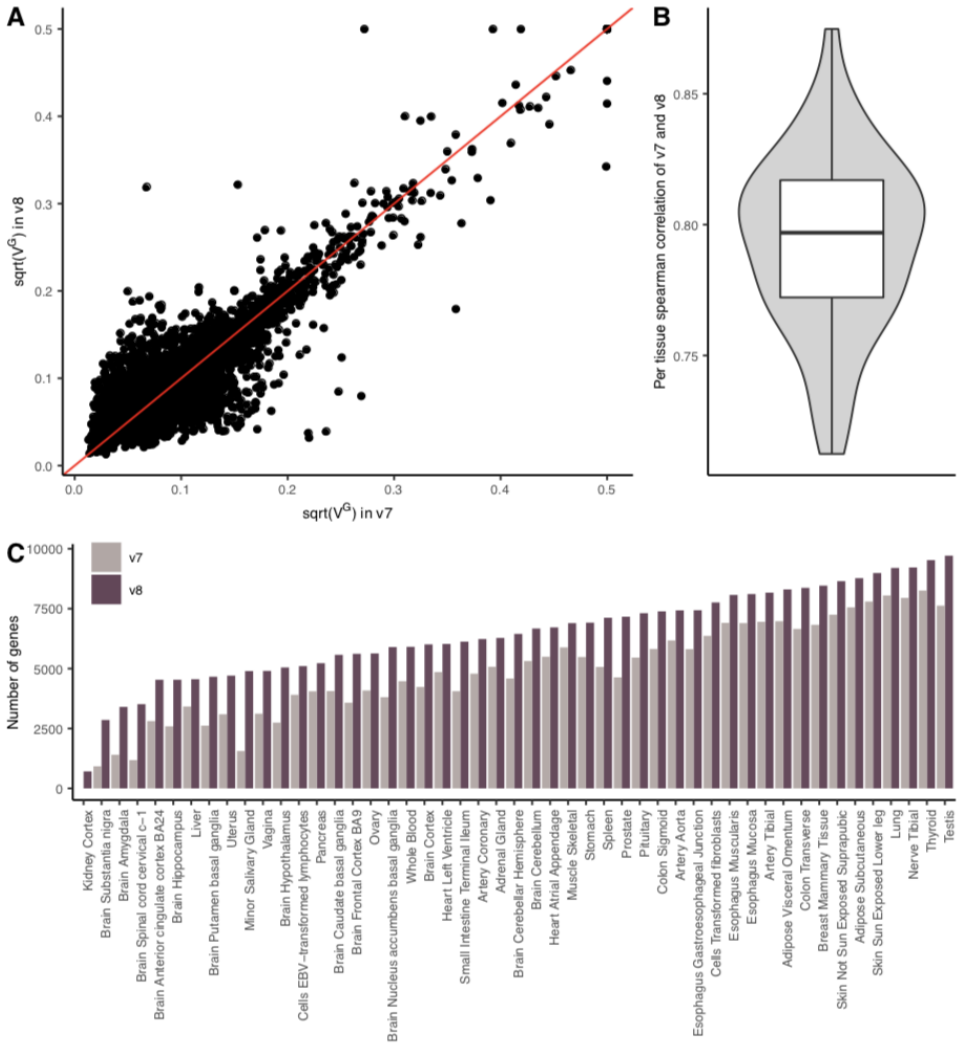

Figure S3. ANEVA estimates of genetic variance in gene expression (V^G). (A) Comparison of V^G estimates for an example tissue (Adipose subcutaneous) derived from GTEx v8 dataset compared to that of v7. The red line represents x=y. **(B)** Distribution of the spearman correlation coefficient between V^G estimates from v7 and v8 across all GTEx tissues. The lower and the upper whiskers indicate 1.5 interquartile range from the first and the third quartile, respectively. **(C)** The number of genes with V^G estimates available across GTEx tissues in each version.

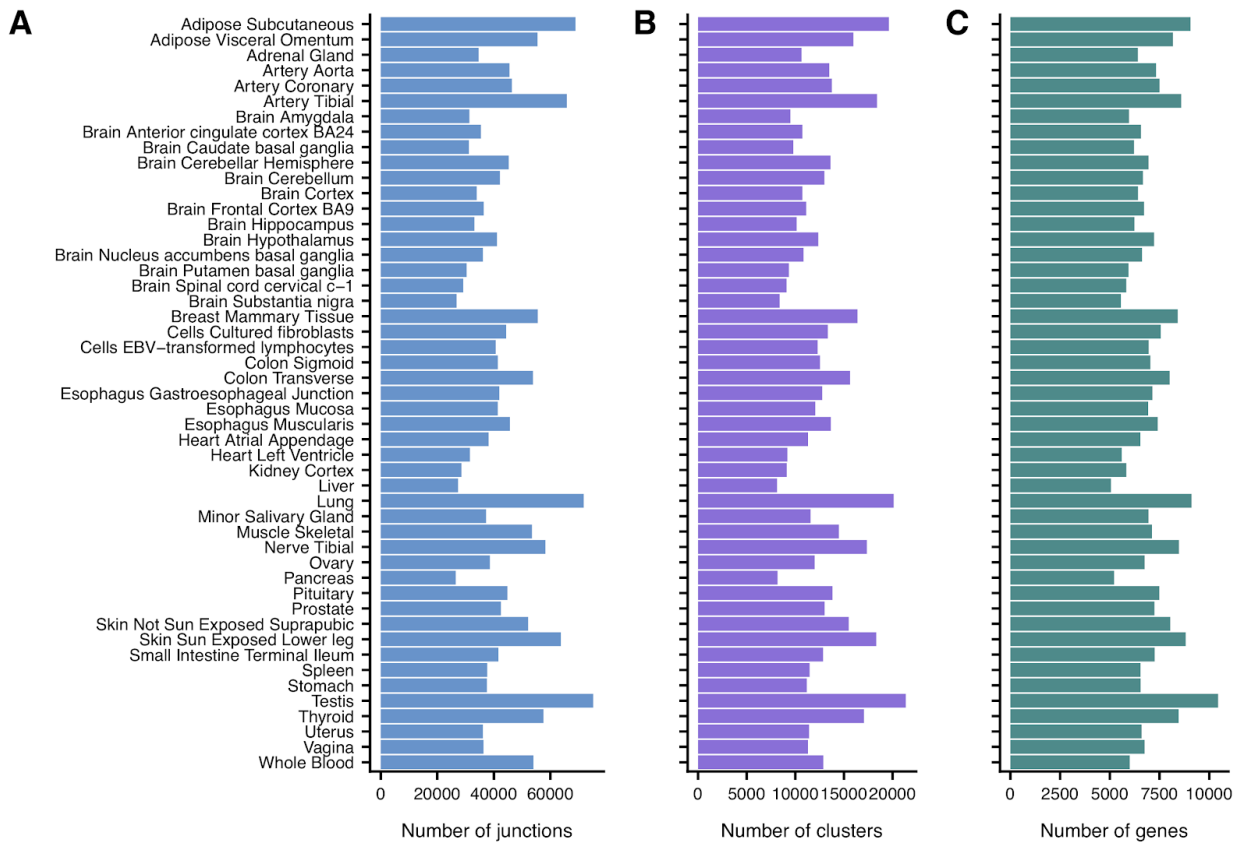

Figure S4. sOutlier split read count processing. The number of unique **(A)** junctions, **(B)** LeafCutter clusters, and **(C)** genes that are found in each tissue (rows) after split read count quantification and processing.

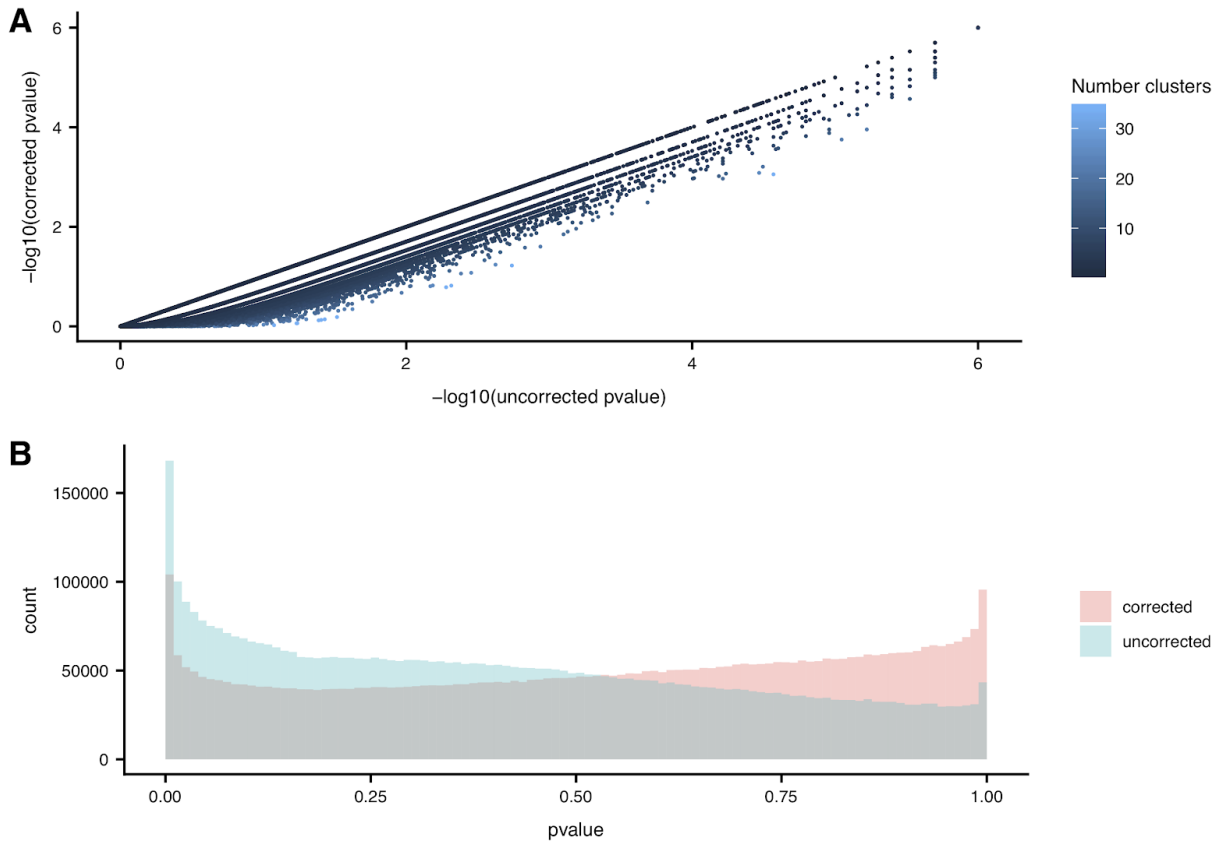

Figure S5. SPOT gene level correction. (A) Scatterplot showing the $-log_{10}$ (sOutlier p-values+ 1x10^{-6}) in Muscle-Skeletal tissue at the gene level before the gene-level correction (x-axis) and after the gene level correction (y-axis) for the number of LeafCutter clusters mapped to each gene (color). **(B)** The distribution of sOutlier p-values in Muscle-Skeletal tissue at the gene level before the gene level correction (teal) and after the gene level correction (salmon) for the number of LeafCutter clusters mapped to each gene.

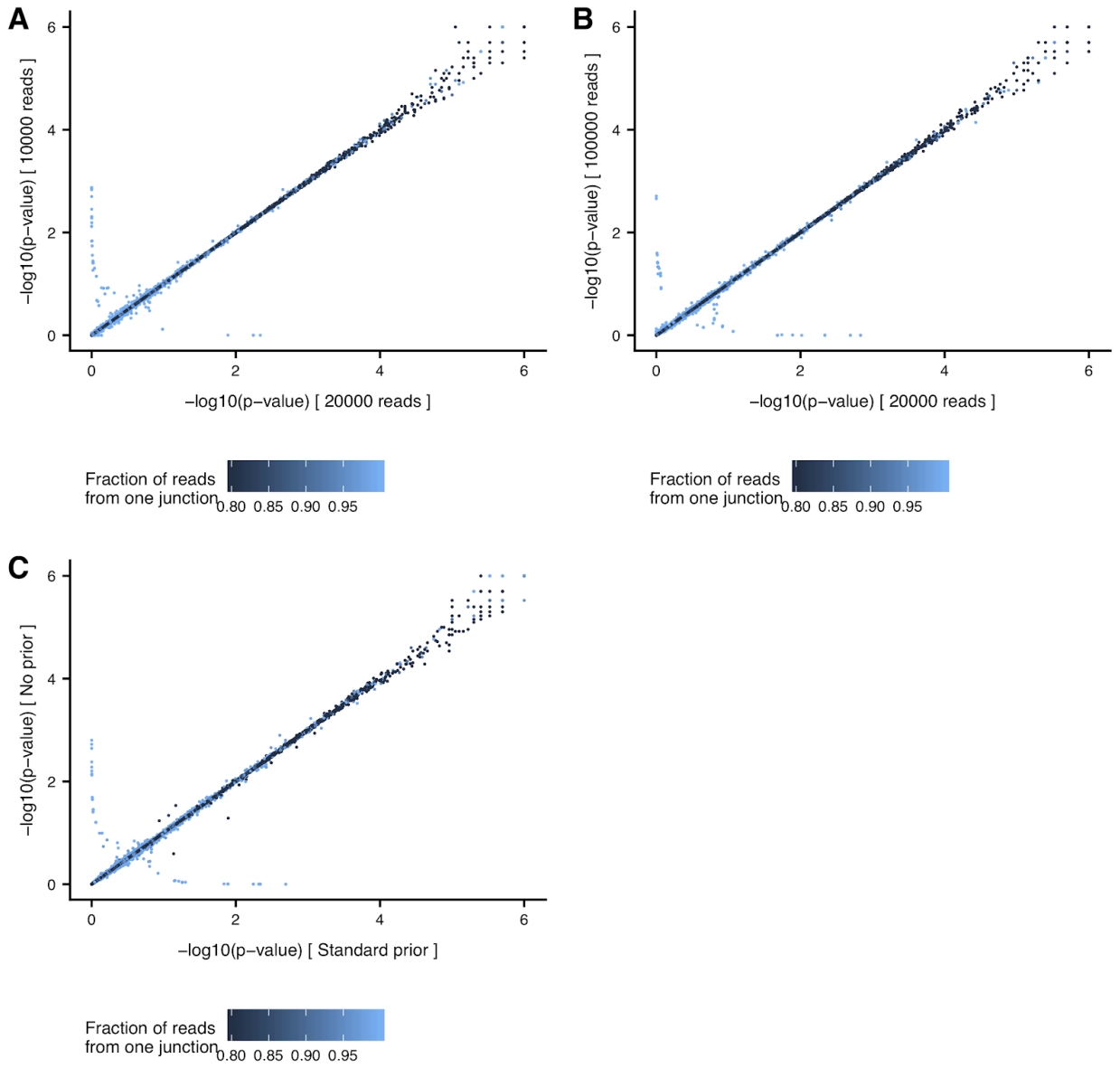

Figure S6. Robustness of SPOT to hyperparameter choice. Scatterplot showing the $-log_{10}$ (sOutlier p-values + 1x10^{-6}) of sample-LeafCutter cluster pairs in Muscle-Skeletal tissue from default implementation of SPOT (x-axis) compared to implementations of SPOT using different hyperparameter settings (y-axis; **A, B, C**) colored by the maximum fraction of reads mapping to a single junction (summed across samples) in the corresponding LeafCutter cluster. Any cluster with a maximum fraction of reads mapping to a single junction that is less than or equal to 80% is colored identically to better highlight differences above 80%. (**A, B**) Comparison of sOutlier p-values from the default implementation of SPOT (x-axis) and an implementation of SPOT where random samples used to generate the empirical distribution have 10,000 (**A)** and 100,000 (**B**) reads mapped to the cluster. (y-axis). (**C**) Comparison of sOutlier p-values from the default implementation of SPOT (x-axis) and an implementation of SPOT where there is no Gamma prior placed on α_j (y-axis).

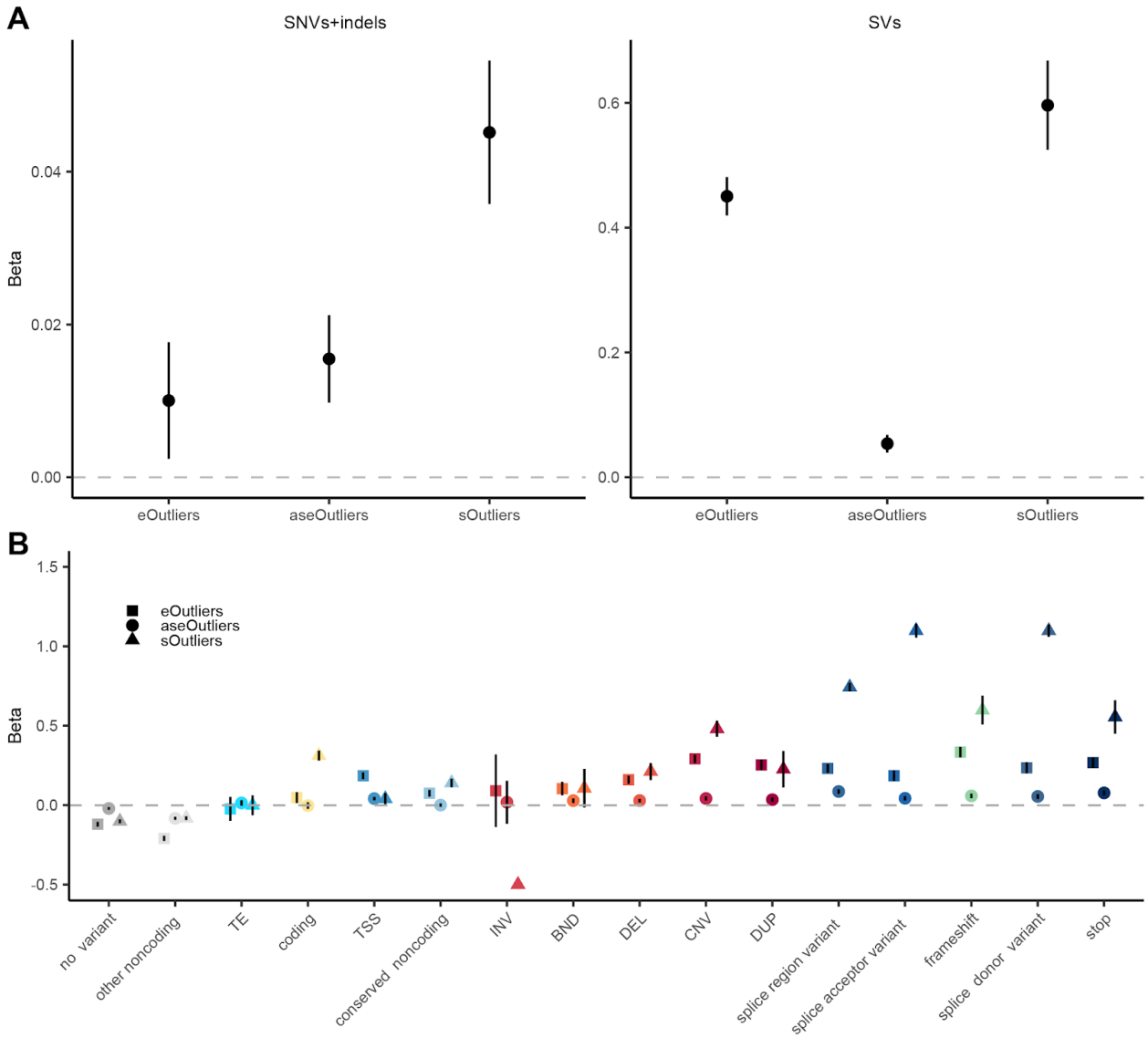

Figure S7. Association of rare variant status and continuous outlier measure. (A) Across each outlier type, the beta coefficient estimate and 95% confidence interval (y-axis) from a linear model of binary rare variant status as the outcome and continuous outlier measure, defined as the -log10(median p-value), as the predictor. Outcome is 1 if the gene has a nearby SNV or indel that is not found in gnomAD, or for SVs if it is a singleton variant within GTEx. **(B)** Beta coefficient estimates from similar models as in **(A)** but considering rare variant status across a range of categories (x-axis).

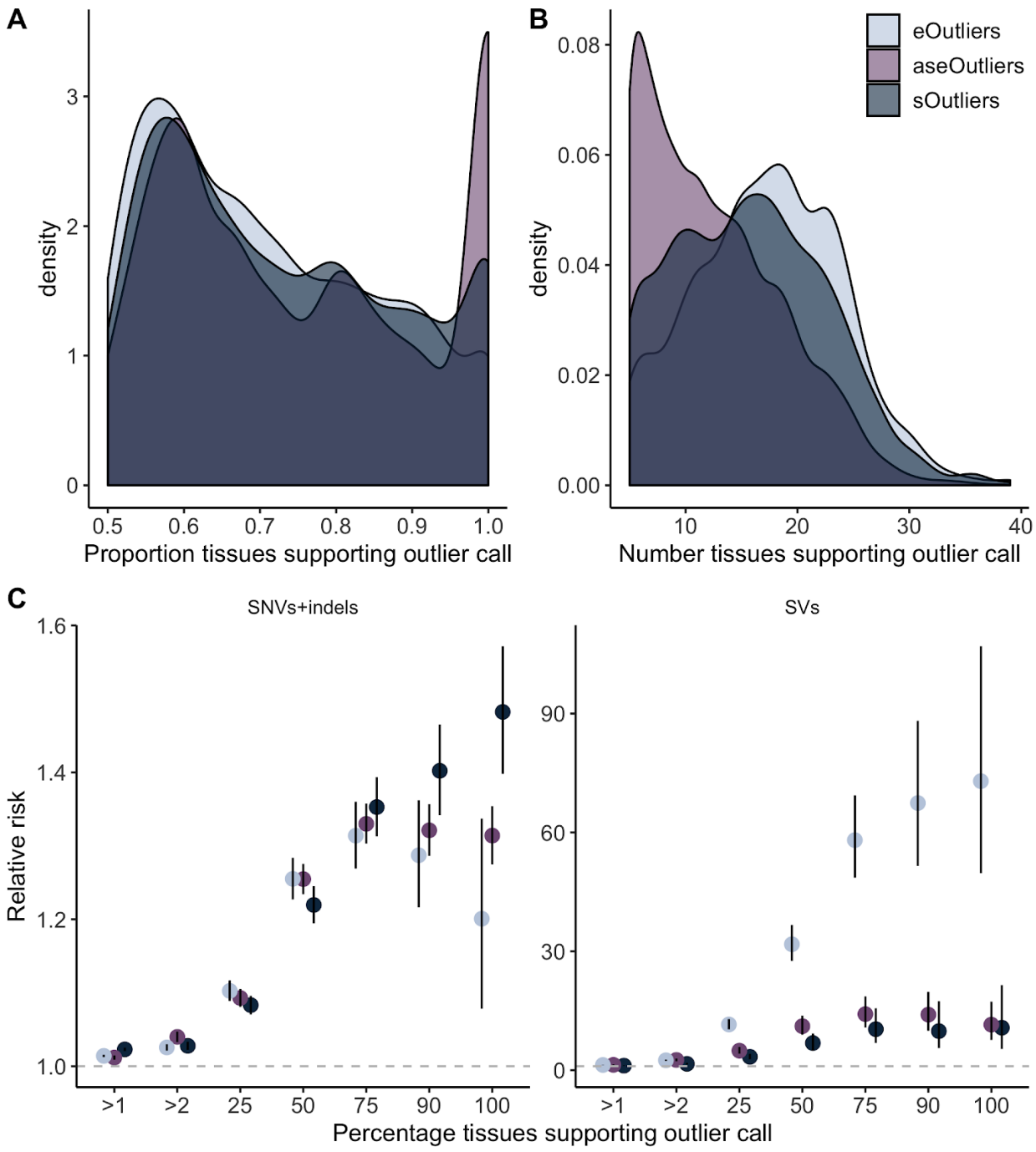

Figure S8. Number of tissues supporting outlier calls. (A) For all multi-tissue outlier calls, the proportion of tested tissues with outlier signal at the same threshold (p-value < 0.0027 or |Z| > 3). **(B)** For all multi-tissue outlier calls, the number of tested tissues with outlier signal at the same threshold (p-value < 0.0027 or |Z| > 3), restricted to individuals with data from at least 5 tissues. **(C)** The impact of the number of tissues supporting the outlier call on the relative risk of outliers having a rare variant (MAF < 1%) within 10kb. For the >1 and >2 bins, this refers to >1

or > 2 tissues, while the remaining bins are percentages of the total number tested. For SVs, sOutlier enrichments stop at the 50% bin due to small numbers at later bins.

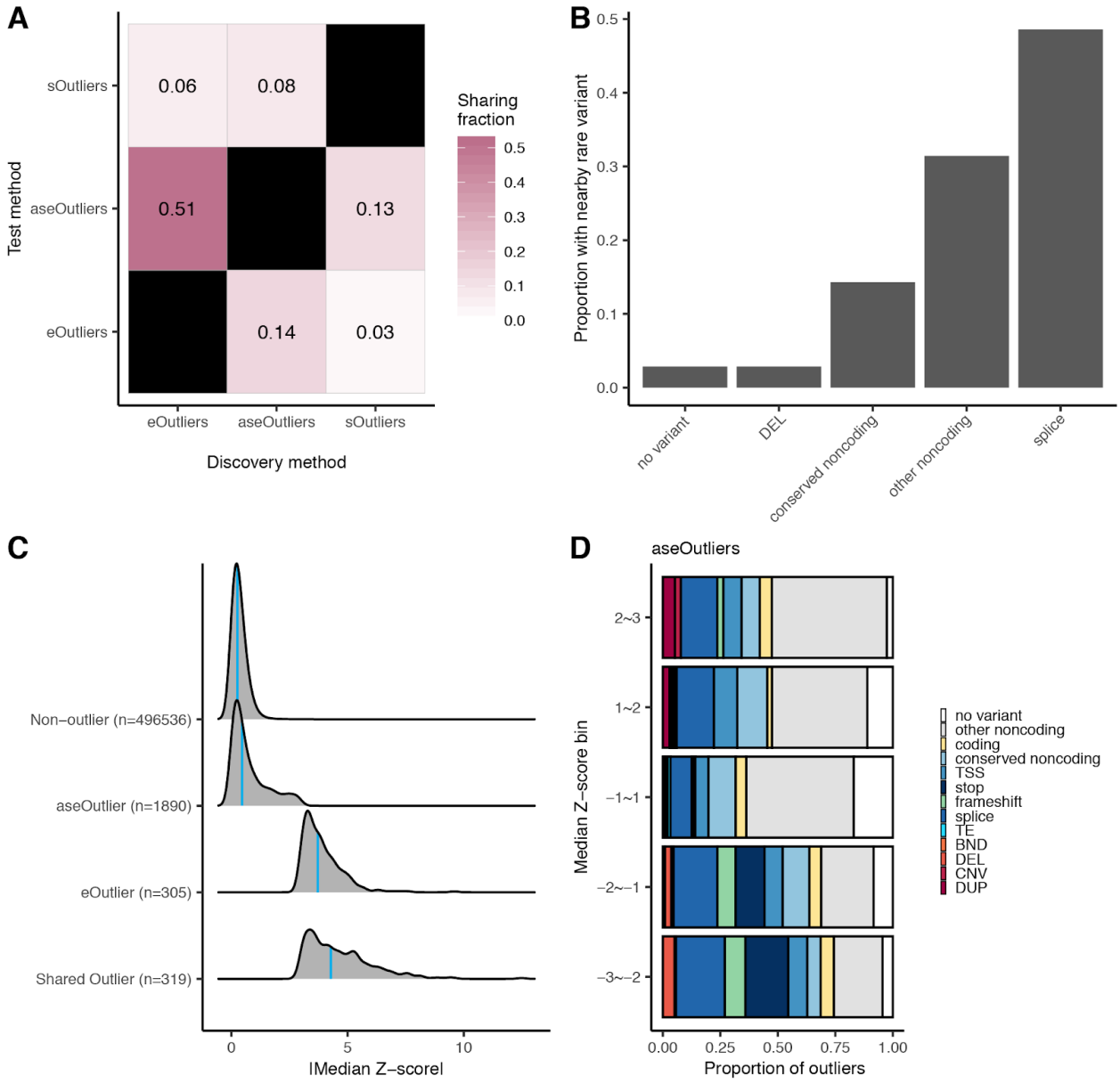

Figure S9. Comparing outliers across methods. (A) Of the set of individuals and genes tested across all data types, the fraction discovered via one method that also meet the outlier thresholds ($p < 0.0027$) in another method. Across all data types, 624 individuals and 8,722 genes, including 2,281,262 unique combinations, were tested by all methods. **(B)** The proportion of outliers shared across all methods assigned to the given rare variant category nearby the outlier gene. Of the 2,209 aseOutliers, 1,385 sOutliers, and 624 eOutliers discovered at this threshold among the shared set, 35 individual-gene pairs are found by all three methods, encompassing 31 unique genes. **(C)** Of the set of eOutliers and aseOutliers within this set, the distribution of |median Z-scores| for outliers in both types, expression alone, ASE alone, or non-outliers for the same set of genes. Blue lines represent the 50th percentile. **(D)** The

proportion of aseOutliers with a nearby rare variant of a given type split by the corresponding median *Z*-score bin for the same individual-gene pair.

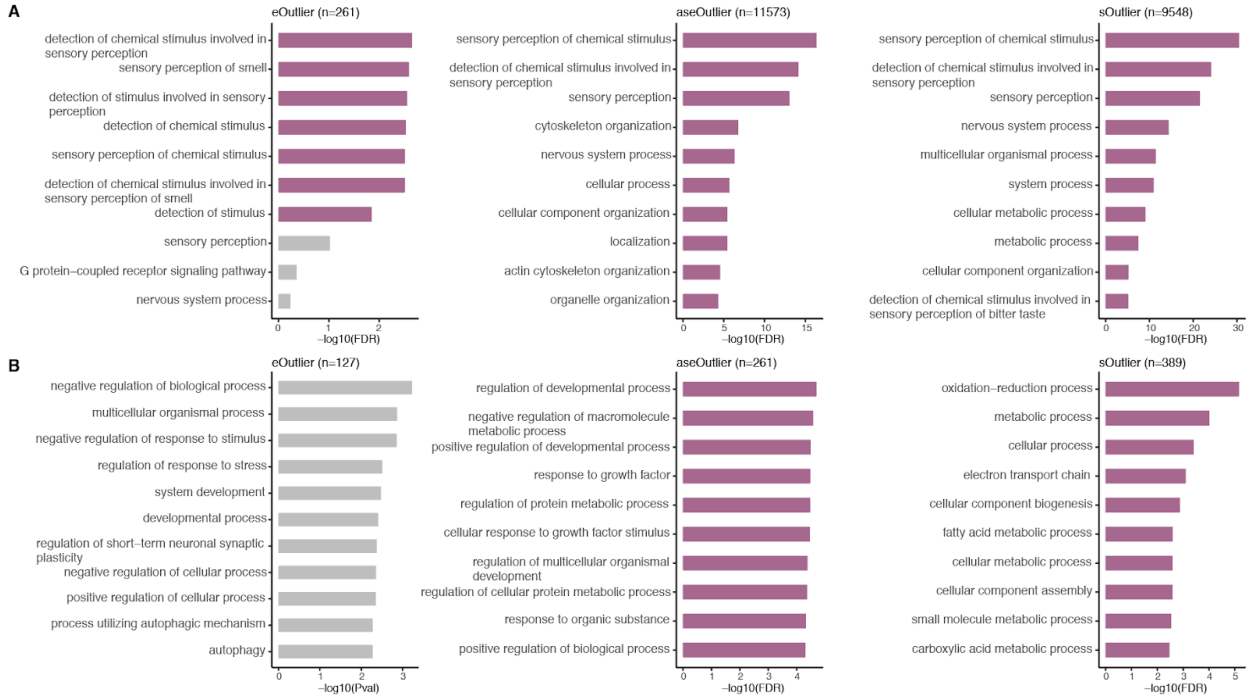

Figure S10. Gene ontology term enrichments for outlier and non-outlier genes. The top ten Gene Ontology (GO) terms enriched, by -log10(FDR-corrected p-value) on the x-axis, in the set of genes with no outliers in any tissue **(A)** and those associated with the most extreme outliers **(B).** Results are included for eOutliers on the left, aseOutliers in the center and sOutliers on the right, with the number of included genes at the top of each plot. Pink bars are significant at an FDR-corrected p-value threshold of 0.05, while the gray bars are not significant. For eOutliers in **(B),** all terms had an FDR corrected p-value of 1, and so nominal p-values are presented instead.

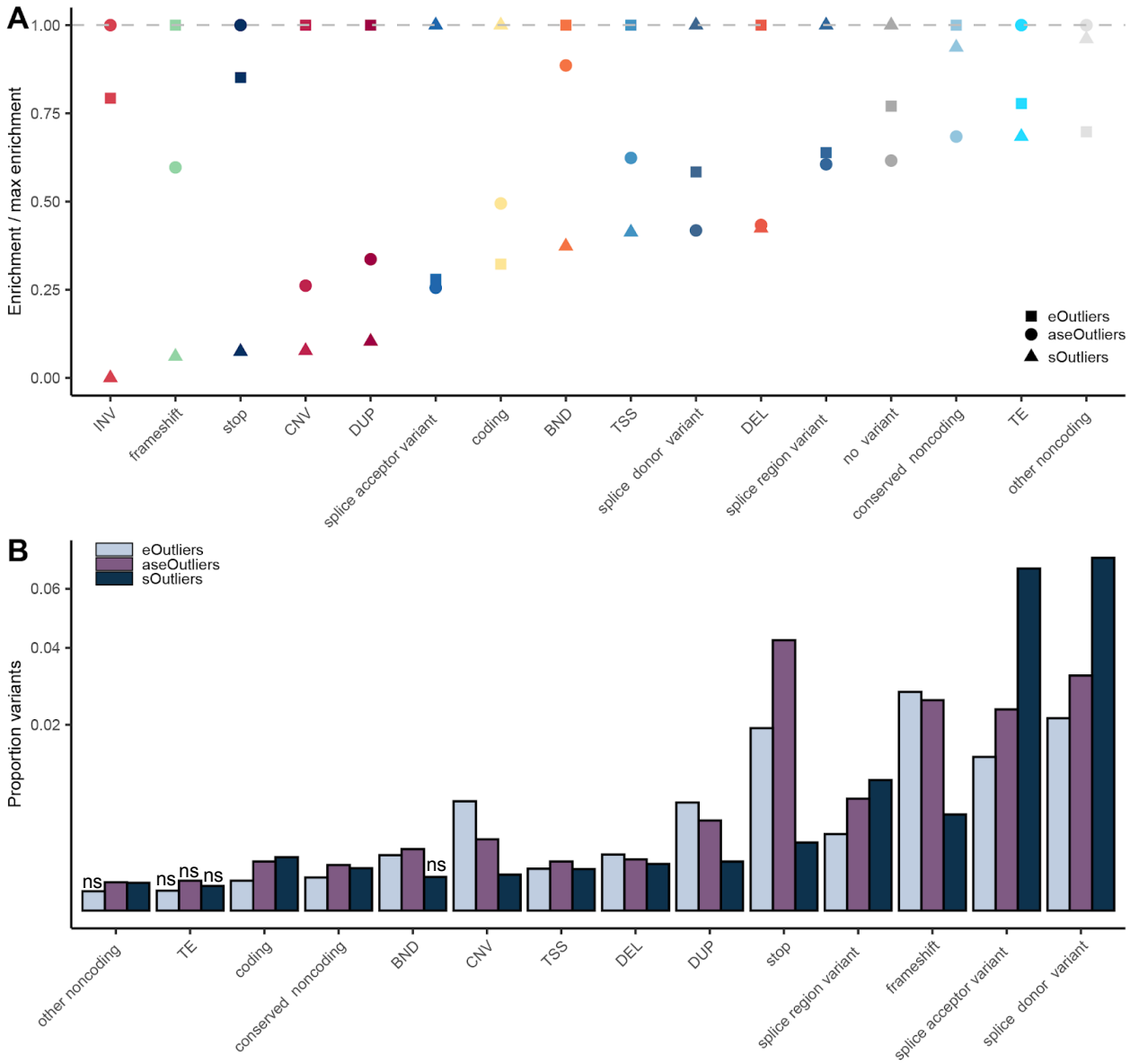

Figure S11. Comparison of variant class enrichments across methods. (A) For each variant category, the relative risk enrichment for each outlier type over the maximum enrichment for that category. **(B)** For each variant category, the proportion of variant occurrences leading to an outlier across all categories, with INV removed due to either very low or zero instances. Those marked ns indicate that in 1000 iterations permuting outlier status, a proportion greater than or equal to the actual proportion was found greater than 5% of the time. TSS = transcription start site, TE = transposable element, INV = inversion, BND = breakend, DEL = deletion, CNV = copy number variation, DUP = duplication.

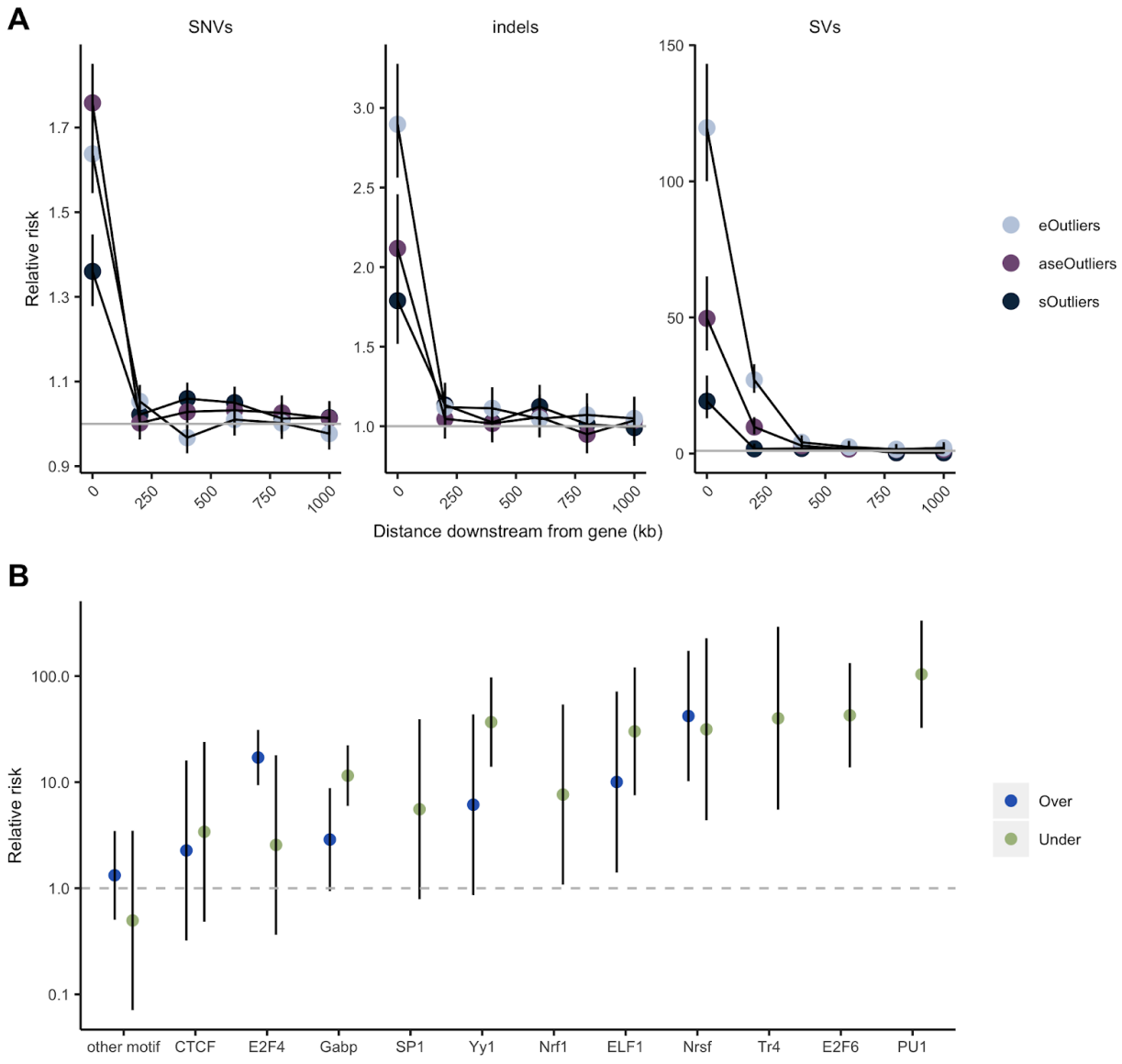

Figure S12. Rare variant enrichments at distances downstream of outlier genes and in promoter motifs. (A) Relative risk of singleton SNVs, indels, and SVs at varying distances downstream of outlier genes (bins exclusive) across data types. **(B)** Relative risk of rare (MAF < 1%) variants interrupting promoter motifs nearby over eOutliers (blue) or under eOutliers (green) relative to controls. For data points not included for one direction, there were not enough instances of rare variants overlapping a given motif near outliers to estimate risk.

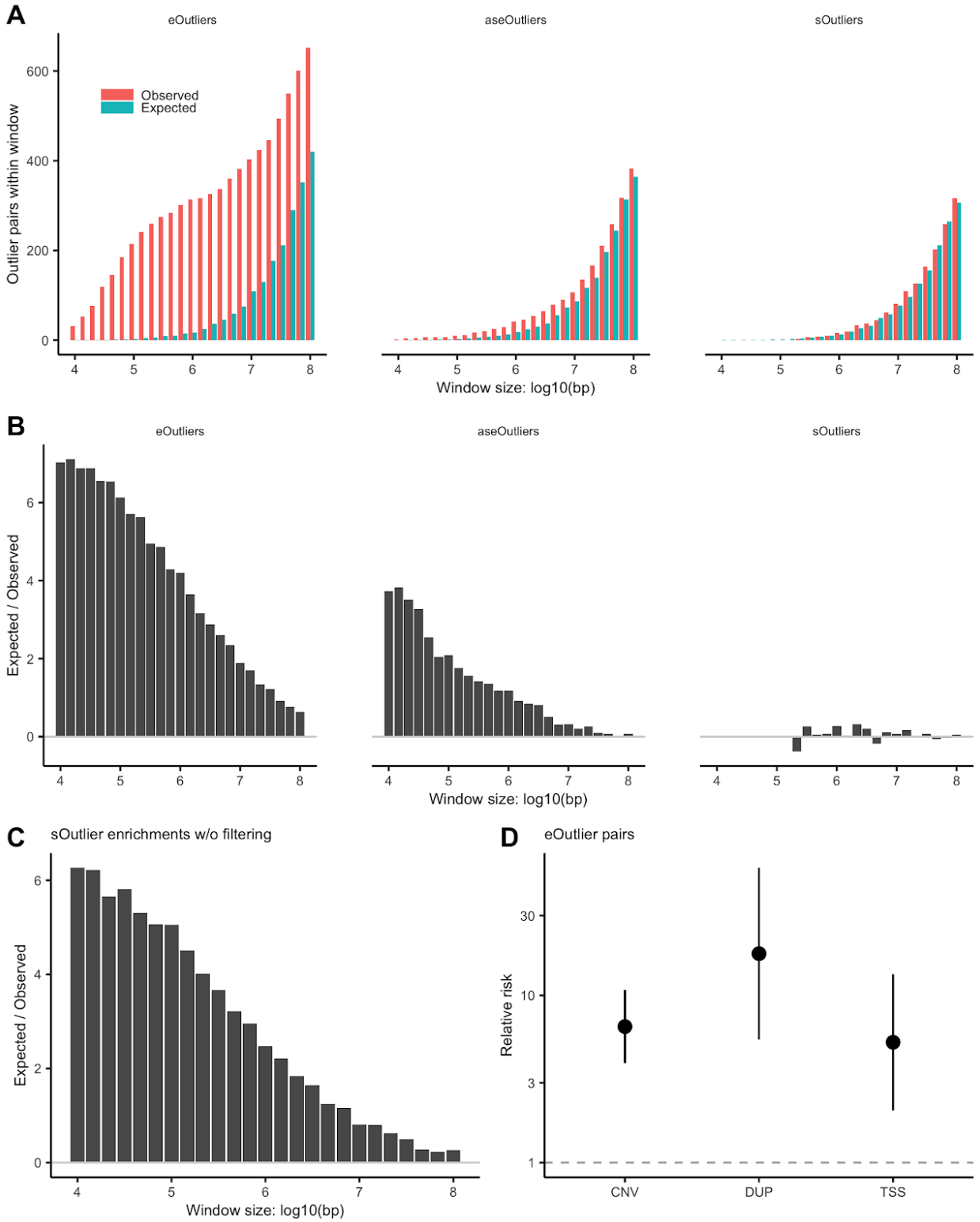

Figure S13. Outliers occurring together within a given window. (A) At varying window sizes, the number of observed vs expected outliers occurring together within that window. Expected numbers were generated from sampling an equal number of outlier genes from randomly

chosen individuals. **(B)** The enrichment, calculated as log2 ratio of the observed number of outliers occurring in the same window over expected, across different window sizes. **(C)** In A and B, we filter out any splicing gene pairs that share a cluster, see Supplemental Methods. Here, we calculate the enrichments for sOutliers including those gene pairs. **(D)** For eOutlier pairs, the relative risk of one or both genes in the pairs found within a 100kb window having a nearby rare CNV, DUP, or TSS variant as compared to individuals who are only outliers for one of the genes in the pair.

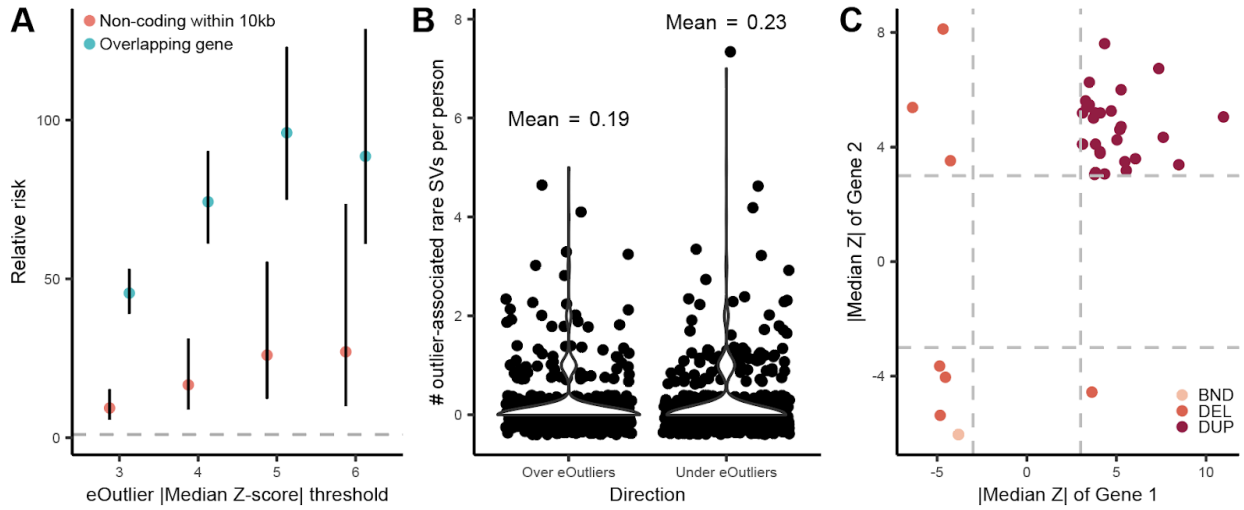

Figure S14. Rare structural variants impacting expression per individual. (A) Relative risk of rare SVs within 10kb of eOutliers split by whether the SV intersects the gene body (teal) or is non-coding but within the 10kb window (pink). **(B)** The number of rare outlier-associated SVs per individual, split by the direction of the eOutlier on the x-axis. **(C)** For a subset of rare SVs associated with a change in the expression of > 1 gene within the same individual, the median Z-score of one gene is plotted on the x-axis vs the other affected gene on the y-axis. The color indicates the type of SV and the grey lines are at median Z = [-3,3]. BND = breakend, DEL = deletion, DUP = duplication

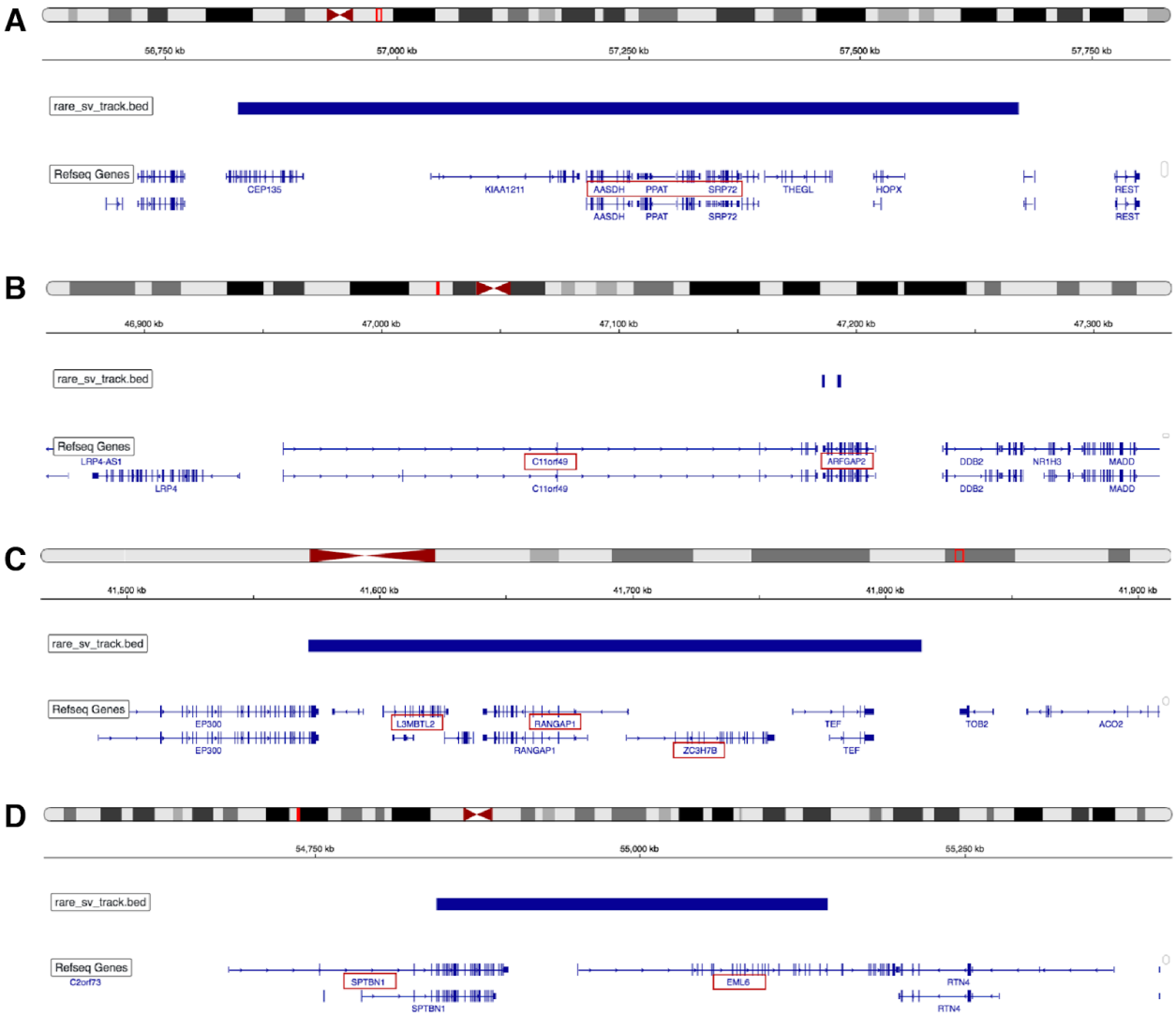

Figure S15. Example rare structural variants with expression effects on multiple genes. **(A)** A rare duplication (dark blue rectangle) with observable increased expression effects on four genes (red squares) PAICS (median Z = 3.83), PPAT (median Z = 3.11), AASDH (median Z = 4.10), and SRP72 (median Z = 5.19). The effect is observed across 14 tissues. Other genes in the region, KIAA1211 and HOPX, show more moderate effects, with median Z = 1.73 in both. **(B)** Three rare breakend mutations (dark blue rectangles) have observable decreased expression effects on two genes (red squares), ARFGAP2 (median Z = -3.81) and C11orf49 (median Z = -6.04). The effect is observed in 18 tissues. **(C)** A rare duplication (dark blue rectangle) with observable increased expression effects on three genes (red squares), L3MBTL2 (median Z = 5.26), RANGAP1 (median Z = 4.71), and ZC3H7B (median Z = 6.00). The effect occurs in 19 tissues. **(D)** A rare deletion (dark blue rectangle) with opposite expression effects on two genes (red squares), SPTBN1 (median Z = -4.67) and EML6 (median Z = 8.12). The effect is observed across 24 tissues, with EML6 expression Z-scores exceeding 15 in 6 tissues.

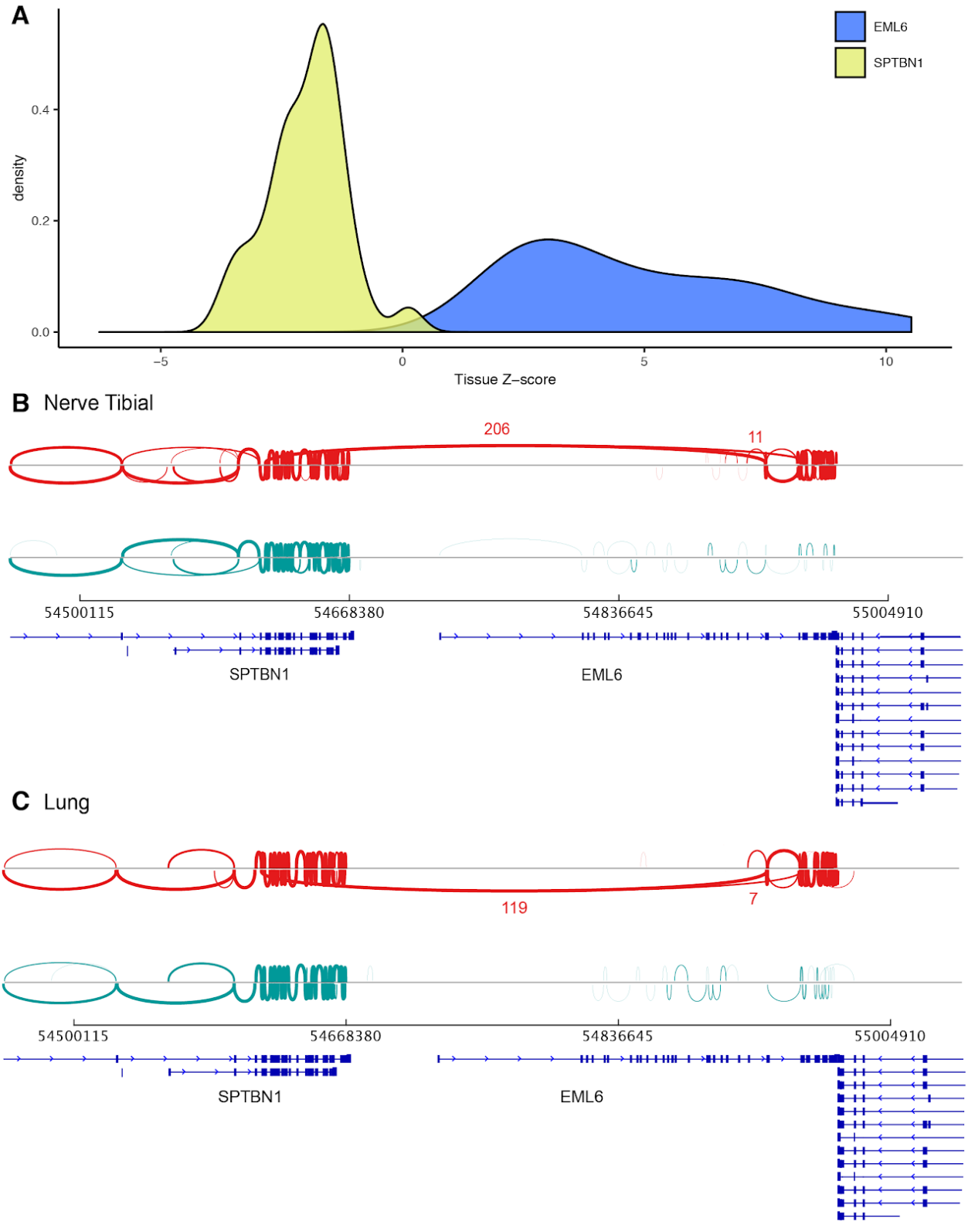

Fig S16. Rare deletion leading to fusion of SPTBN1 and EML6. (A) Distribution of Z-scores across the 31 measured tissues for the individual with the rare, heterozygous deletion (shown in

Fig S15D) for EML6 (blue) and SPTBN1 (yellow). Reads supporting splice junctions found in nerve tibial **(B)** and lung **(C)** for the individual with the rare deletion (red) and two different random non-outlier control individuals (teal), showing the fusion transcript is only found in the individual with the deletion. The width of the lines correspond to the number of reads mapped to each junction.

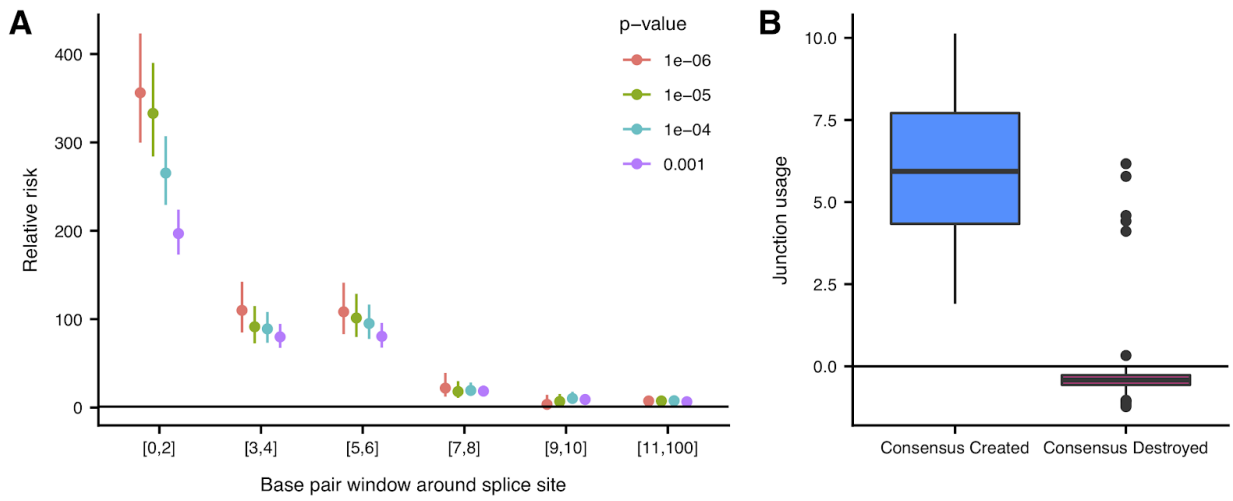

Figure S17. Enrichment of rare variants nearby splice sites in sOutliers. (A) Relative risk (y-axis) of rare variants within various window sizes around splice sites (x-axis) for sOutlier LeafCutter clusters relative to non-outlier clusters at several median LeafCutter cluster p-value thresholds (color). **(B)** Junction usage of a splice site is the natural log of the fraction of reads in a LeafCutter cluster mapping to the splice site of interest in sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) samples relative to the fraction in non-outliers samples aggregated across tissues by taking the median. Junction usage (y-axis) of the closest splice sites to rare variants that lie within the splicing consensus sequence binned by the type of variant (x-axis).

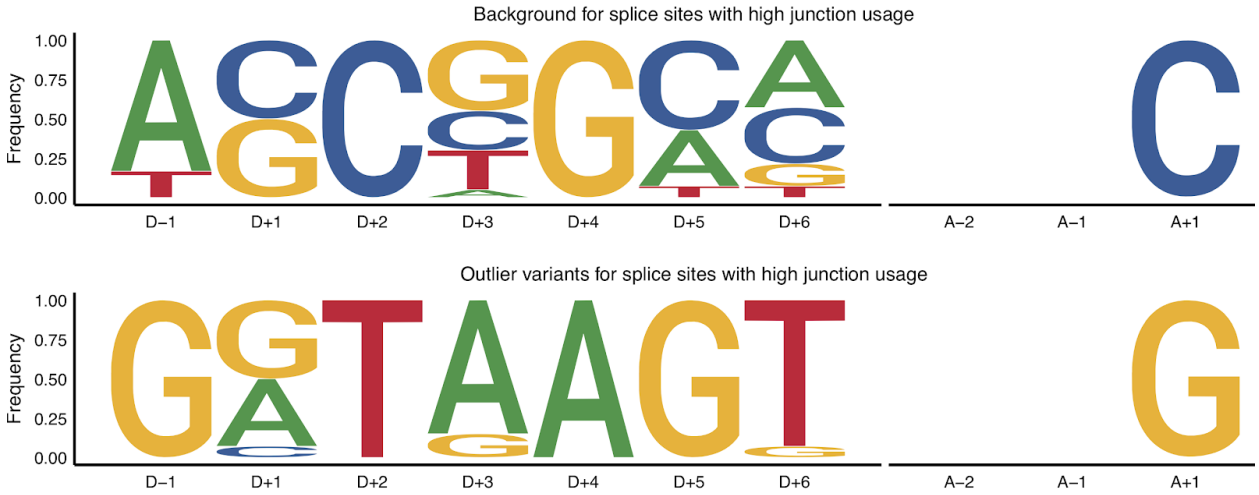

Figure S18. sOutlier variants in consensus sequence of splice sites with high junction usage. Independent position weight matrices showing mutation spectrums of sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) rare variants at positions relative to splice sites with positive junction usage (ie. splice sites used more in outlier individuals than in non-outliers).

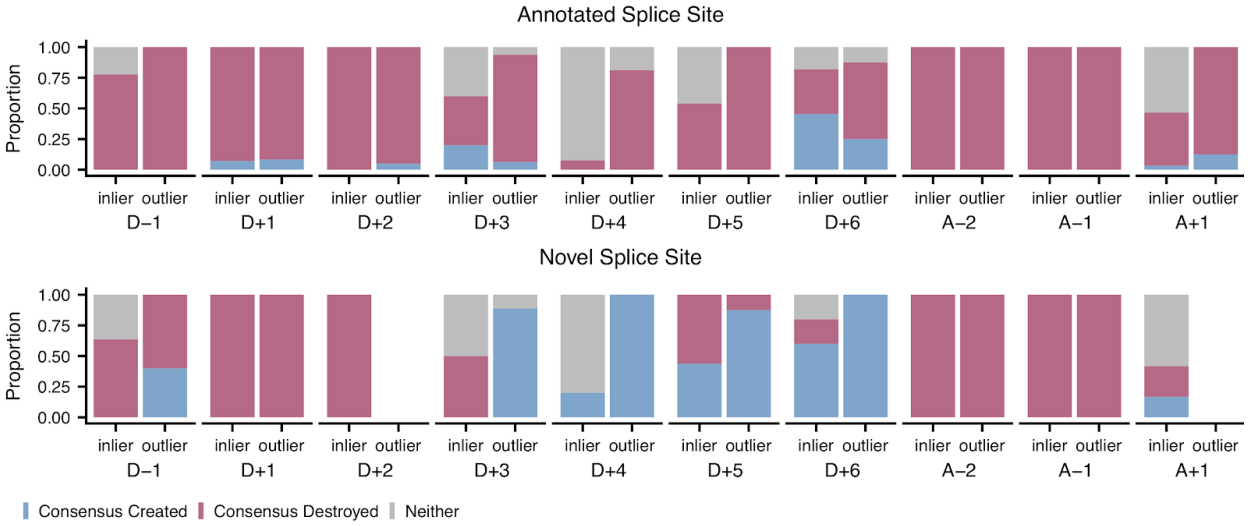

Figure S19. sOutlier variants in consensus sequence of annotated and novel splice sites. Proportion of sOutlier (median LeafCutter cluster p-value < 1 x 10^{-5}) and non-outlier variants, at each position in the splicing consensus sequence, that create the consensus sequence (blue) or destroy the consensus sequence (red) where variants are binned by whether the nearby splice site is annotated or novel (rows).

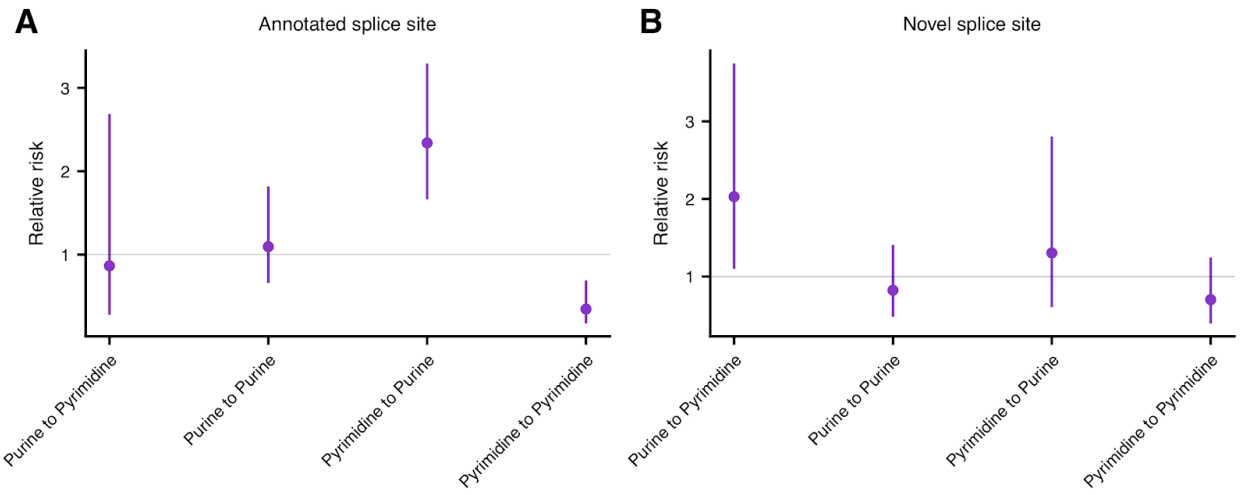

Figure S20. sOutlier variant type enrichments in PPT. Relative risk for sOutliers relative to non-outliers (median LeafCutter cluster p-value < 1 x 10^{-5}) of having a rare variant that is located in PPT (5 to 35 base pairs upstream from an acceptor splice site) having a specific mutation spectrum (x-axis). Relative risk calculation done separately for annotated **(A)** or novel **(B)** splice sites.

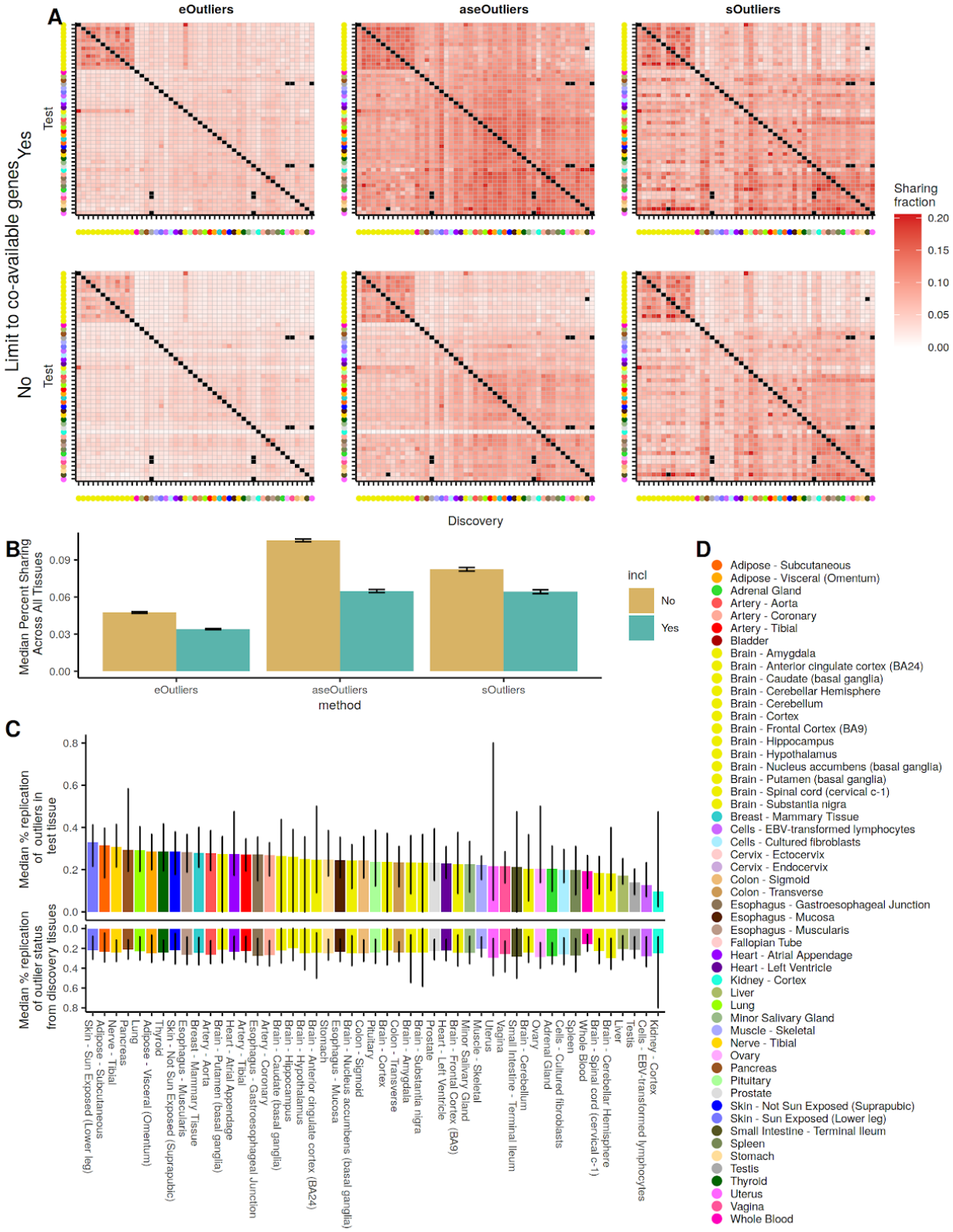
A
eOutliers
aseOutliers
sOutliers
Limit to co-available genes
Yes
No
Test
Sharing fraction
0.20
0.15
0.10
0.05
0.00
Discovery
B
Median Percent Sharing Across All Tissues
0.09
0.06
0.03
0.00
eOutliers
aseOutliers
sOutliers
method
incl
No
Yes
C
Median % replication of outliers in test tissue
Median % replication of outlier status from discovery tissues
0.8
0.6
0.4
0.2
0.0
Skin - Sun Exposed (Lower leg)
Adipose - Subcutaneous
Nerve - Tibial
Pancreas
Lung
Adipose - Visceral (Omentum)
Thyroid
Skin - Not Sun Exposed (Suprapubic)
Esophagus - Muscularis
Breast - Mammary Tissue
Artery - Aorta
Brain - Putamen (basal ganglia)
Heart - Atrial Appendage
Artery - Tibial
Artery - Coronary
Esophagus - Gastroesophageal Junction
Brain - Caudate (basal ganglia)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Anterior cingulate cortex (BA24)
Stomach
Esophagus - Mucosa
Brain - Nucleus accumbens (basal ganglia)
Colon - Sigmoid
Pituitary
Brain - Cortex
Colon - Transverse
Brain - Amygdala
Brain - Substantia nigra
Prostate
Heart - Left Ventricle
Brain - Frontal Cortex (BA9)
Minor Salivary Gland
Muscle - Skeletal
Uterus
Vagina
Small Intestine - Terminal Ileum
Brain - Cerebellum
Ovary
Adrenal Gland
Cells - Cultured fibroblasts
Spleen
Whole Blood
Brain - Spinal cord (cervical c-1)
Brain - Cerebellar Hemisphere
Liver
Testis
Cells - EBV-transformed lymphocytes
Kidney - Cortex
D
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Cultured fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood

Figure S21. Outlier status sharing across tissues detail. (A) Percent sharing heatmaps where for all outlier individual-gene pairs (nominal p value < .0027) in a discovery tissue, we measure the percentage of cases where the same individual-gene pair is also an outlier in a test tissue. In the upper row of heatmaps, we limit the analysis to only the genes tested in both tissues, to answer the biological question of how consistent the outlier status is across tissues that co-express a gene. The lower row of heatmaps considers a missing datapoint as a non-shared outlier status, and addresses the utility of each method in diagnosing expression outlier status in a tissue of interest using a different tissue as a proxy. **(B)** Median percent sharing across all tissue-tissue pairs (± 95% bootstrap confidence interval), with and without considering missing values as “non-shared”. aseOutliers are affected the most by missing values. **(C)** Median replication percentage of aseOutlier status in one discovery tissue across all test tissues (top), and median replication percentage of outlier status in one test tissues across all discovery tissues (bottom). The black bars indicate the observed range of values across all individuals. Here, outlier status is declared when a gene has a Benjamini-Hochberg corrected p-value < .05. While for consistency between the three transcriptome outlier methods we use a high significance threshold on the nominal p-values in all other analyses, the FDR correction is the recommended approach when using ANEVA-DOT p-values in most applications. We observe a considerably higher rate of outlier status sharing, when considering genes passing false discovery rate correction. **(D)** The GTEx tissue color key.

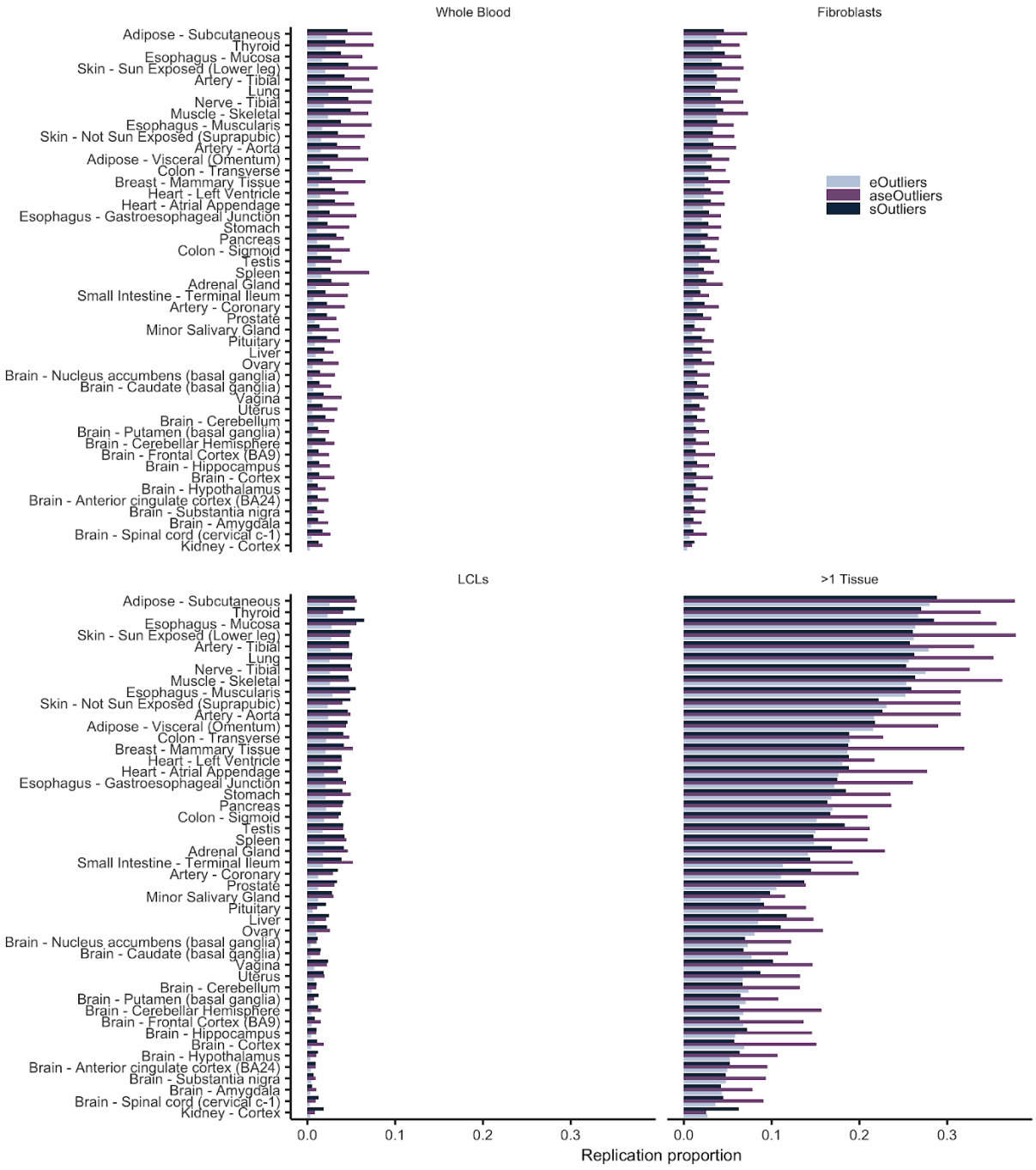

Figure S22. Replication of outliers discovered in clinically accessible tissues. For each of three clinically accessible tissues, the proportion of single tissue outliers (|Z| > 3, SPOT p-value < 0.0027 or ANEVA-DOT p-value < 0.0027) which are also seen in each of the other 46 tissues, restricting each time to genes also measured in the replication tissue. For the bottom right plot, we restrict to outliers seen in more than 1 of the three clinically accessible tissues and assess the replication rate in all other tissues (x-axis).

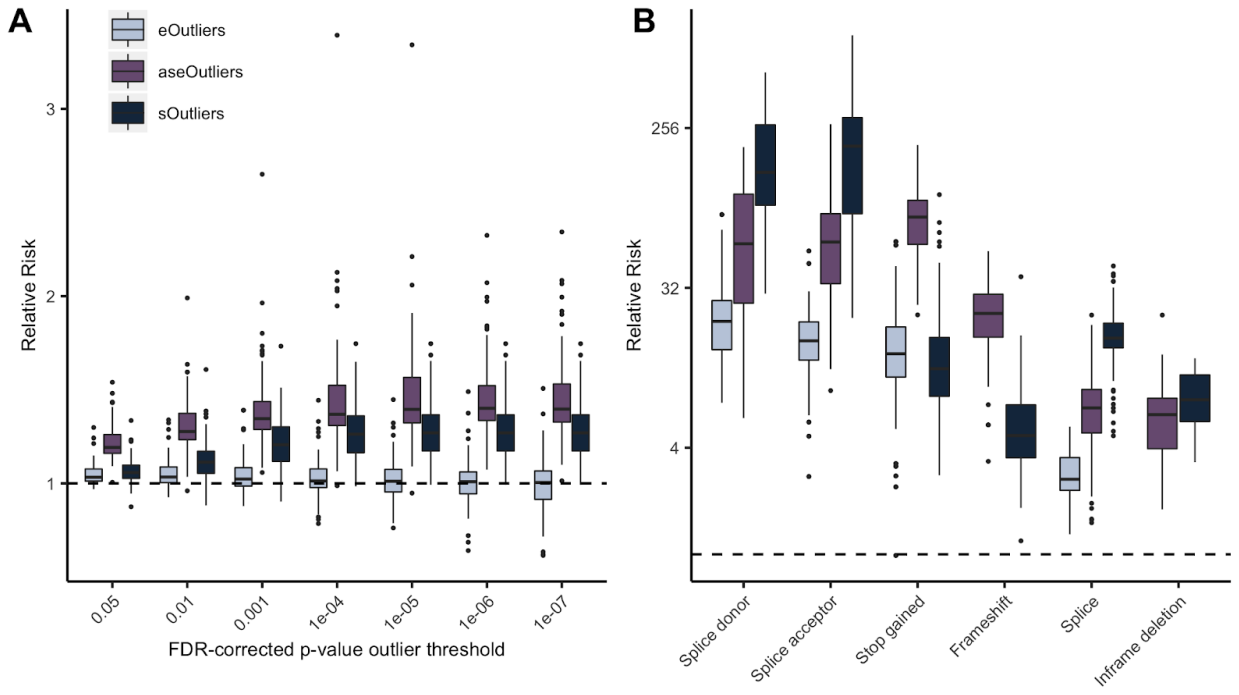

Figure S23. Relative risk of rare variants nearby single tissue outliers with FDR-corrected p-value thresholds. (A) Relative risk point estimate for nearby rare SNVs for outliers across all tissues individually, using various FDR-corrected p-value outlier thresholds. **(B)** Relative risk enrichments for likely gene disrupting rare variants nearby single-tissue outliers using an FDR-corrected p-value threshold of 0.05, with one point per tissue.

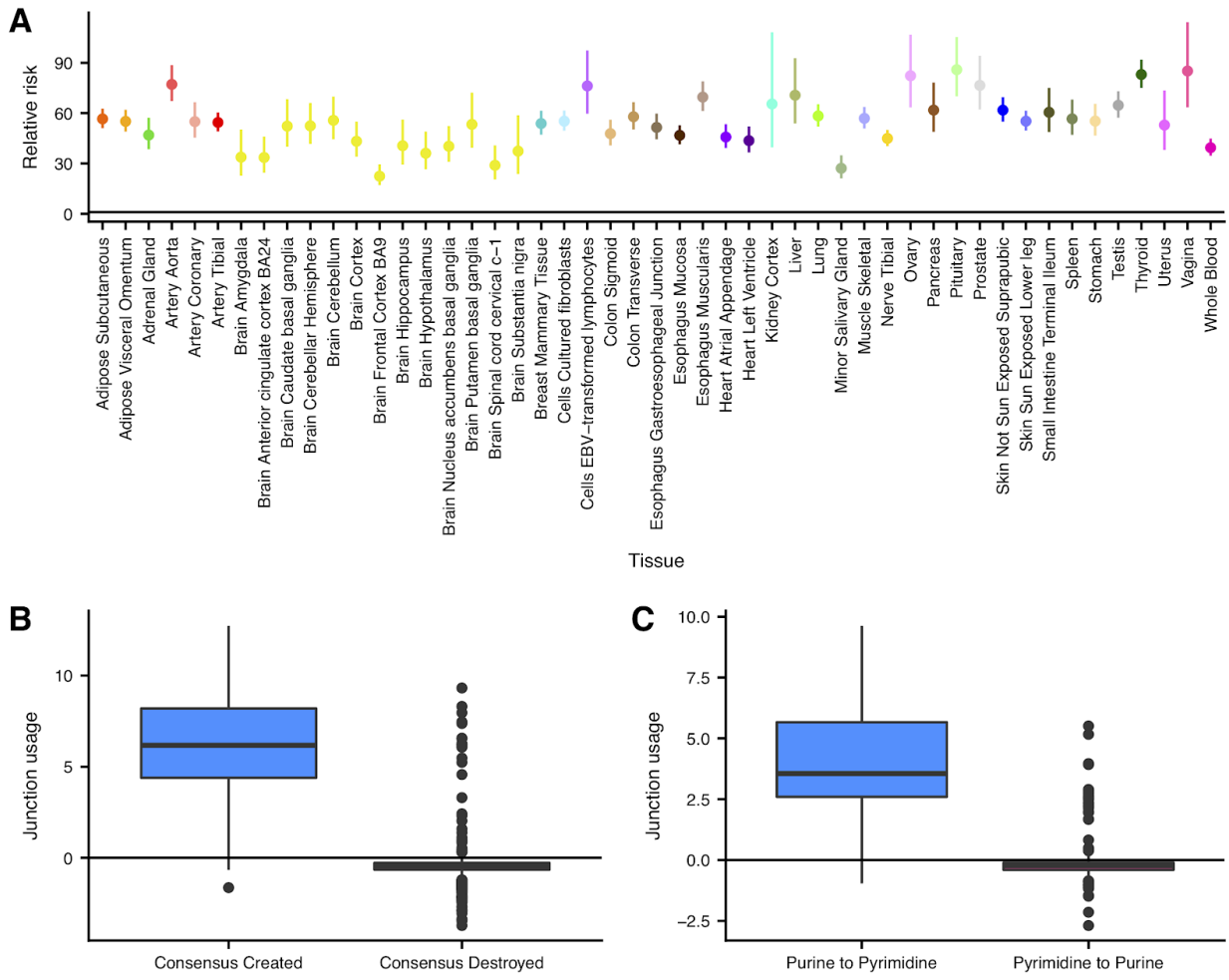

Figure S24. Single tissue sOutlier enrichments. (A) Relative risk, in each tissue independently, of rare variants being located in a 6 base pair window around splice sites for sOutlier LeafCuter clusters (per tissue LeafCutter cluster p-value < 1 x 10^{-5}) relative to non-outlier clusters. **(B, C)** Per tissue junction usage of a splice site is the natural log of the fraction of reads in a LeafCutter cluster mapping to the splice site of interest in sOutlier (per tissue LeafCutter cluster p-value < 1 x 10^{-5}) samples relative to the fraction in non-outliers samples, in a single tissue. **(B)** Per tissue junction usage (y-axis) of the closest splice sites to rare variants that lie within the splicing consensus sequence binned by the type of variant (x-axis). **(C)** Per tissue junction usage (y-axis) of the closest splice sites to rare variants that lie within a PPT ([A-5, A-35]) binned by the type of variant (x-axis).

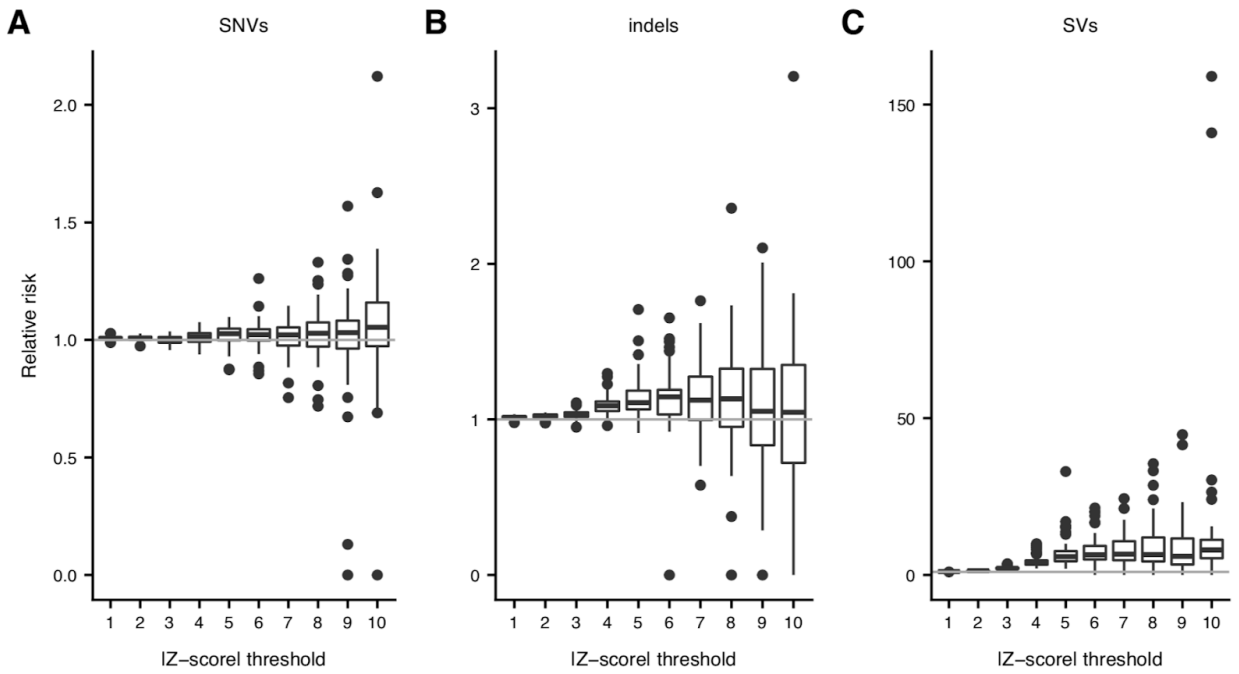

Figure S25. Single tissue eOutlier enrichments across thresholds. Relative risk estimates for nearby rare SNVs **(A)**, indels **(B)** and SVs **(C)** in single-tissue outliers vs controls using |Z-score| thresholds between Z=1 and Z=10, with each point representing a single tissue.

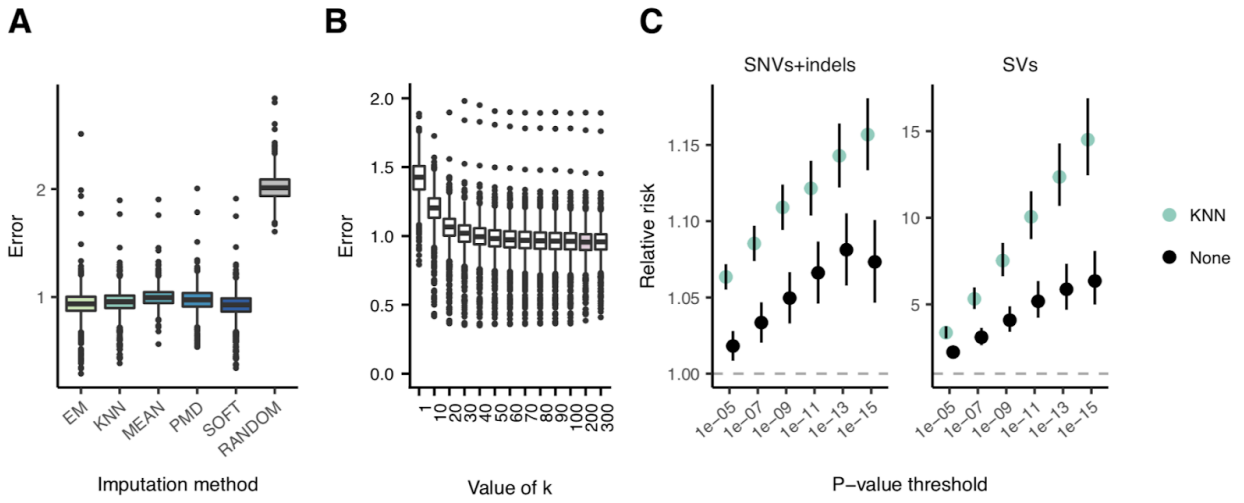

Figure S26. Comparison of imputation methods and correlation outlier enrichments. (A) Reconstruction error across genes when holding out 10% of known expression values for various imputation approaches. **(B)** Reconstruction error across genes for different values of k when performing k-nearest neighbors imputation per gene, with the pink box highlighting the value with the lowest error. **(C)** Relative risk of either a rare SNV or indel or rare SV nearby correlation outliers called using covariance matrices estimated using KNN-imputed expression data across varying thresholds, as compared to an equal number of outliers called by estimating the covariance matrix from complete entries, without imputation. Many more outliers are identified as compared to the median Z-score approach, particularly at the less stringent thresholds.

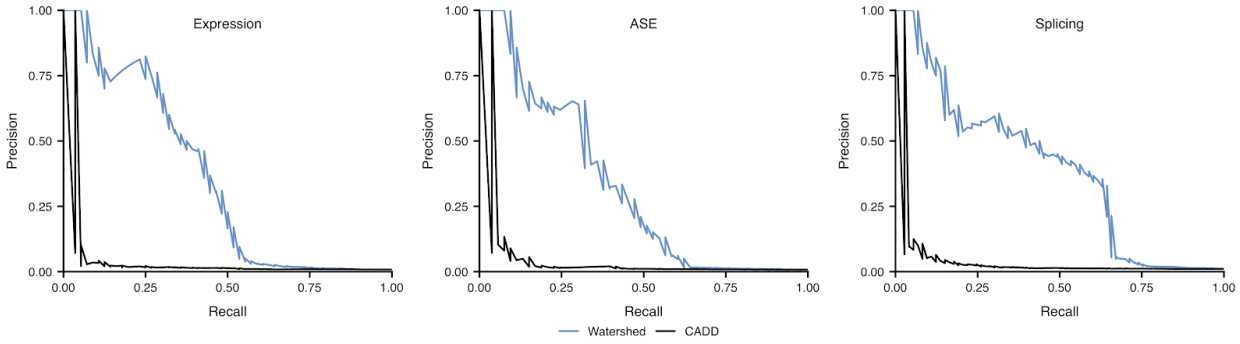

Figure S27. Precision recall curves for Watershed and CADD. Precision-recall curves comparing performance of Watershed and CADD (colors) using held out pairs of individuals for all three median outlier signals.

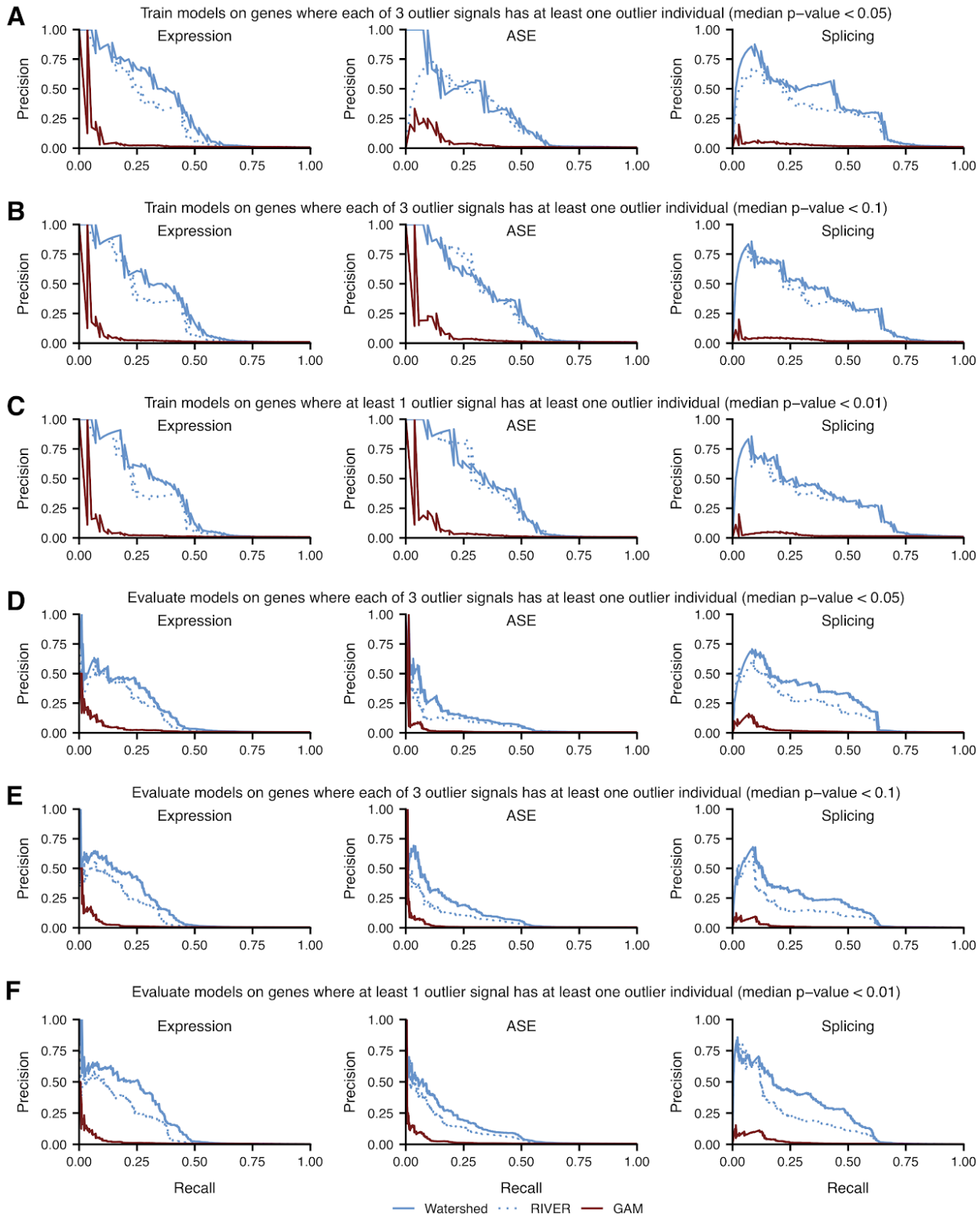

Figure S28. Watershed precision recall curves with different training or evaluation data. Precision-recall curves comparing performance of Watershed, RIVER, and GAM (colors) using held out pairs of individuals for three median outlier signals (columns) when models were

trained with different training data sets (**A, B, C**; see Supplementary methods) or when models were evaluated with different held out pairs of individuals (evaluation data; **D, E, F**; see supplementary methods)**.** Training data for Watershed, RIVER, and GAM filtered to only include genes where **(A)** all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05), **(B)** all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1), **(C)** at least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01). Held out pairs of individuals (evaluation data) used in **A, B, C** were the same held out pairs of individuals used to generate precision-recall curves in Fig 4D. Held out pairs of individuals used to evaluate Watershed, RIVER, and GAM filtered to only include genes where **(D)** all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05), **(E)** all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.1), **(F)** at least 1 outlier signal has at least one individual that is an outlier (median p-value < 0.01). Training data used to train models composing **D, E, F** was the same training data used to generate models underlying precision-recall curves in Fig 4D.

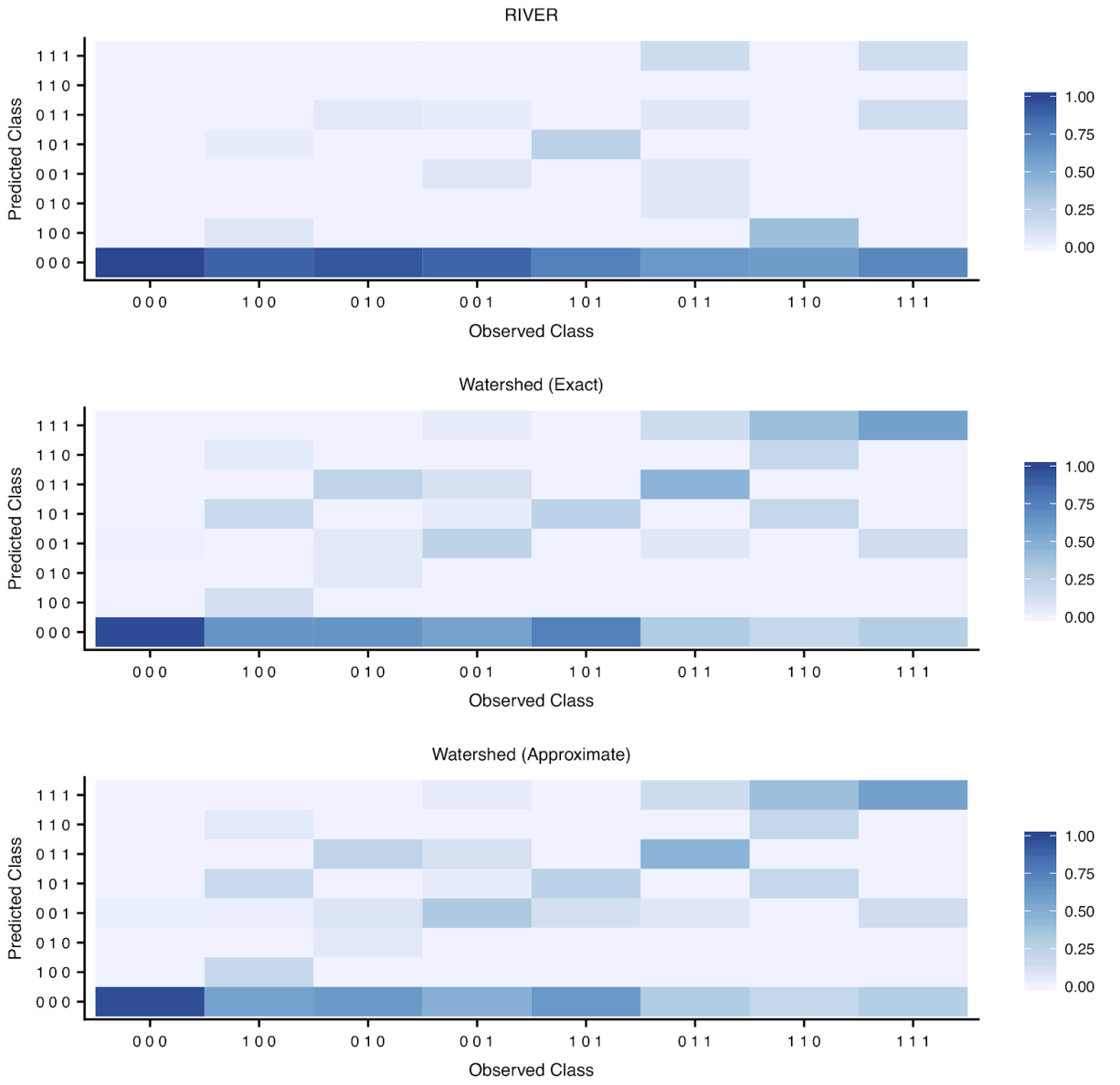

Figure S29. Watershed confusion matrices. Confusion matrices comparing performance of RIVER (top), Watershed with parameters optimized via exact inference (middle), and Watershed with parameters optimized via approximate inference (bottom) in jointly predicting outlier status of all three outlier signals (class) using held out pairs of individuals. The first element of the binary class abbreviations represents median splicing outlier status, the second element of the class abbreviations represents median expression outlier status, and the third element of the class abbreviations represents ASE outlier status. An observed class of “1 0 1” therefor corresponds to a sample that is an outlier for splicing and ASE, but not expression. The predicted class of a sample is the class (out of the 8 classes) that has the largest posterior probability. Columns in each heatmap are normalized to sum to one.

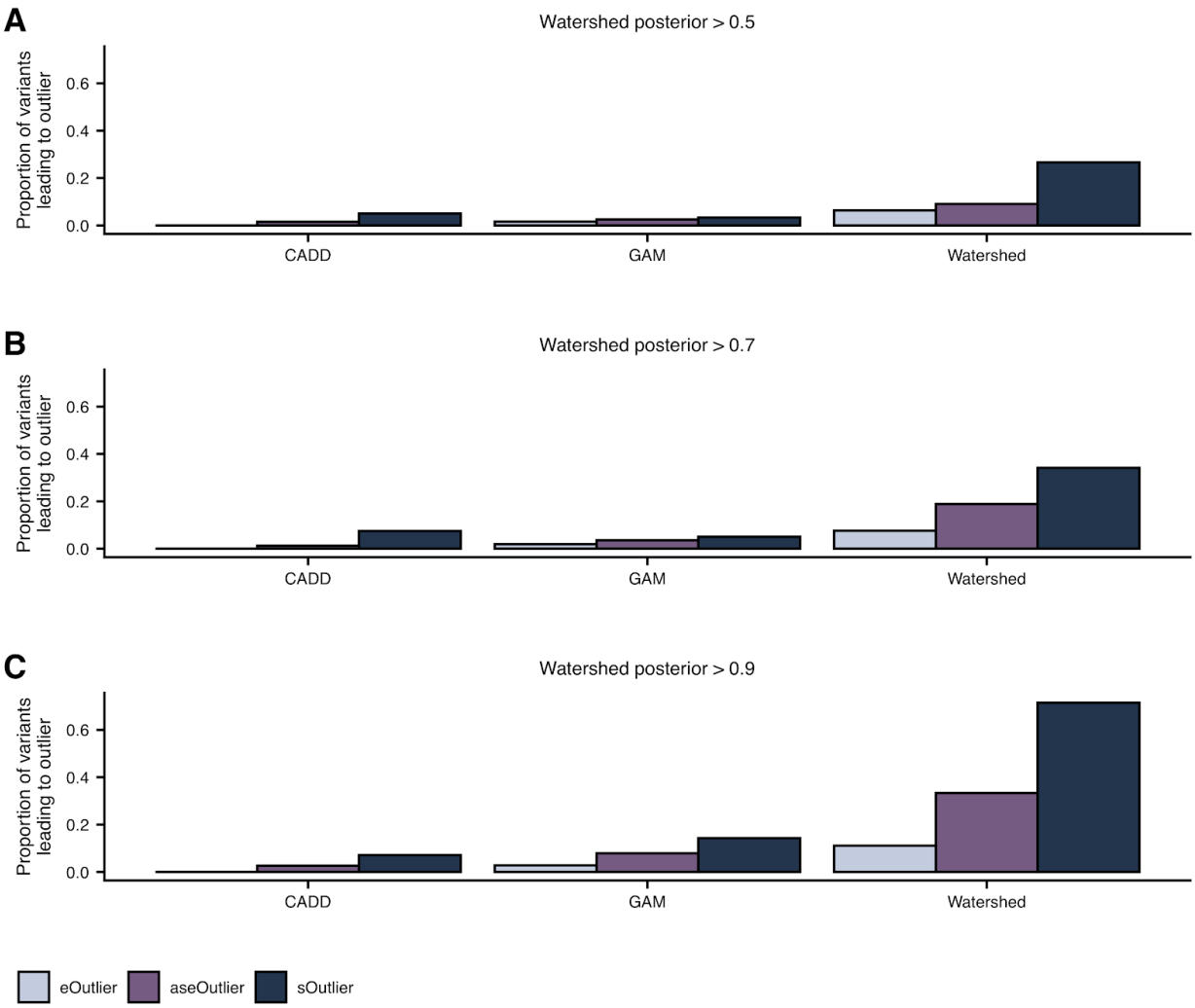

Figure S30. Prioritization of variants that lead to outliers with Watershed. The proportion of rare variants, with Watershed posterior probability greater than 0.5 **(A)**, 0.7 **(B)**, 0.9 **(C)** (right), with GAM probability greater than a threshold set to match the number of Watershed variants for each outlier signal (center), and with CADD score greater than a threshold set to match the number of Watershed variants for each outlier signal (left), that lead to an outlier at a median p-value threshold of 0.0027 across three outlier signals (colors). Watershed, GAM, and CADD models evaluated on held-out pairs of individuals.

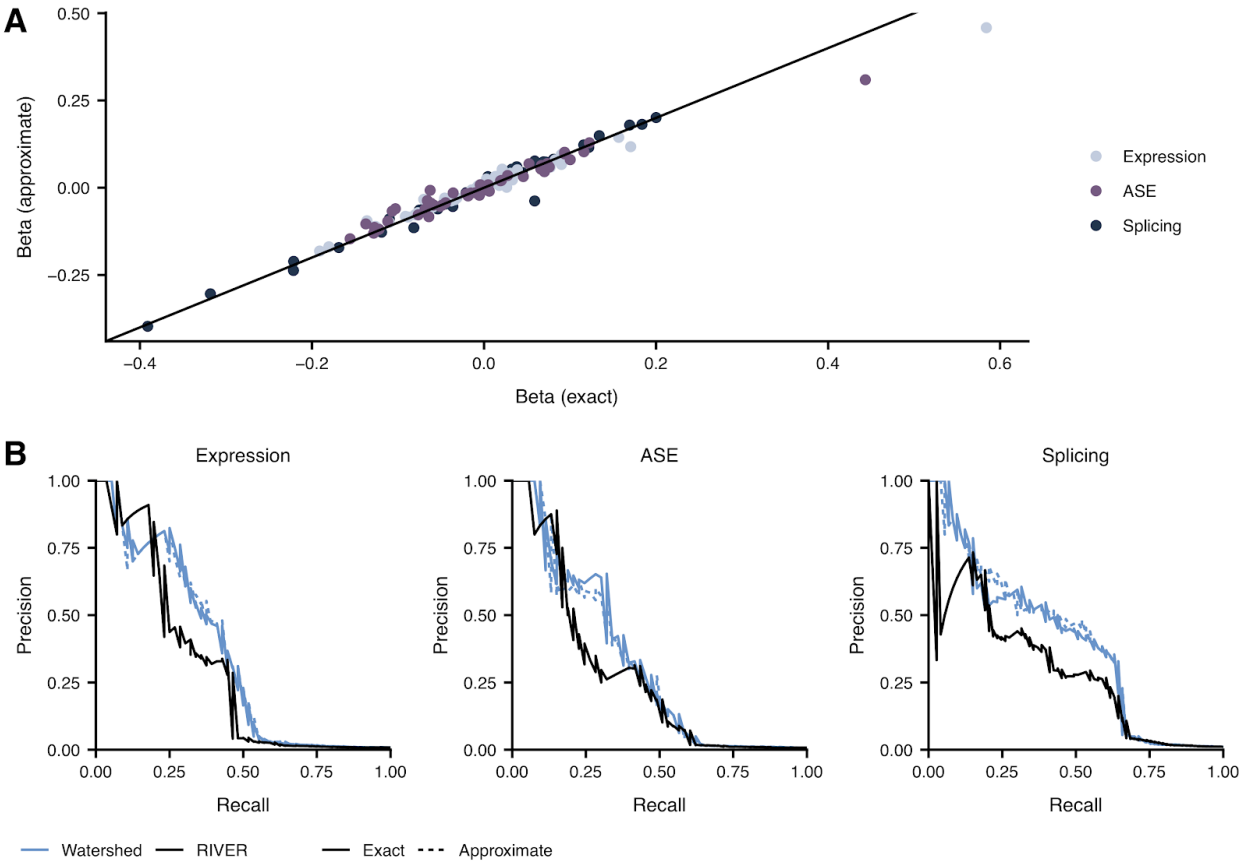

Figure S31. Comparison of exact and approximate inference in Watershed. (A) Scatterplot comparing Watershed (applied to median ASE, splicing, and expression outlier signals) genomic annotation coefficients (β) when model was optimized using exact inference (x-axis) compared to when model was optimized using approximate inference (y-axis) colored by which outlier signal the coefficient predicted. **(B)** Precision-recall curves comparing performance of RIVER, Watershed optimized via exact inference, and Watershed optimized via approximate inference (colors) using held out pairs of individuals for all three median outlier signals.

Figure S32. Tissue-Watershed edge weights. Learned tissue-Watershed edge weights (θ) between pairs of tissue- outlier signals after training tissue-Watershed on expression (top), ASE (middle), and splicing (bottom) outliers across single tissues.

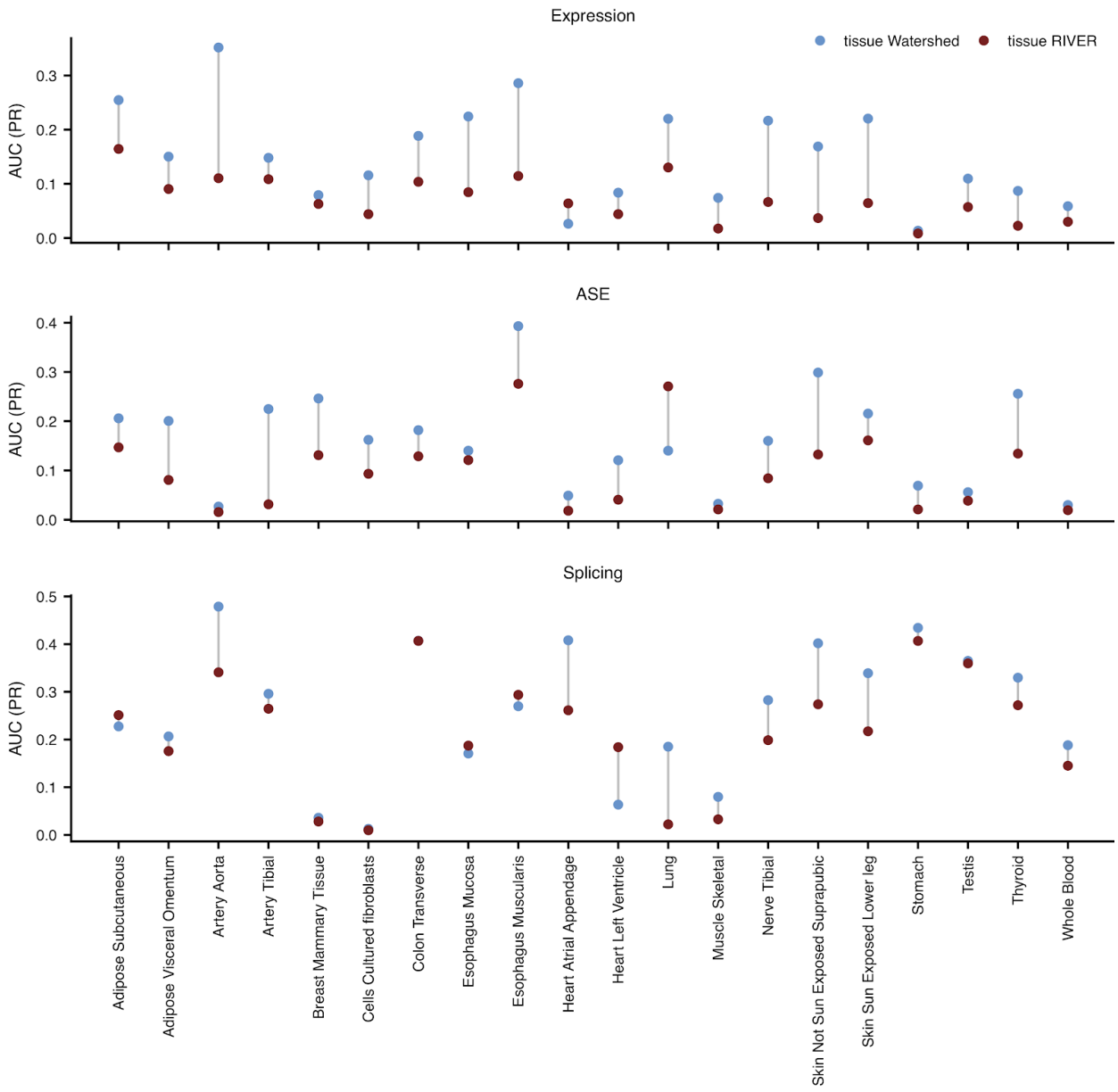

Figure S33. Area under precision recall curves in single tissues. Area under precision recall curves (AUC (PR); y-axis) in a single tissue (x-axis) for tissue-Watershed (blue) and tissue-RIVER (red) when applied outliers across single tissues for all 3 outlier types (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression.

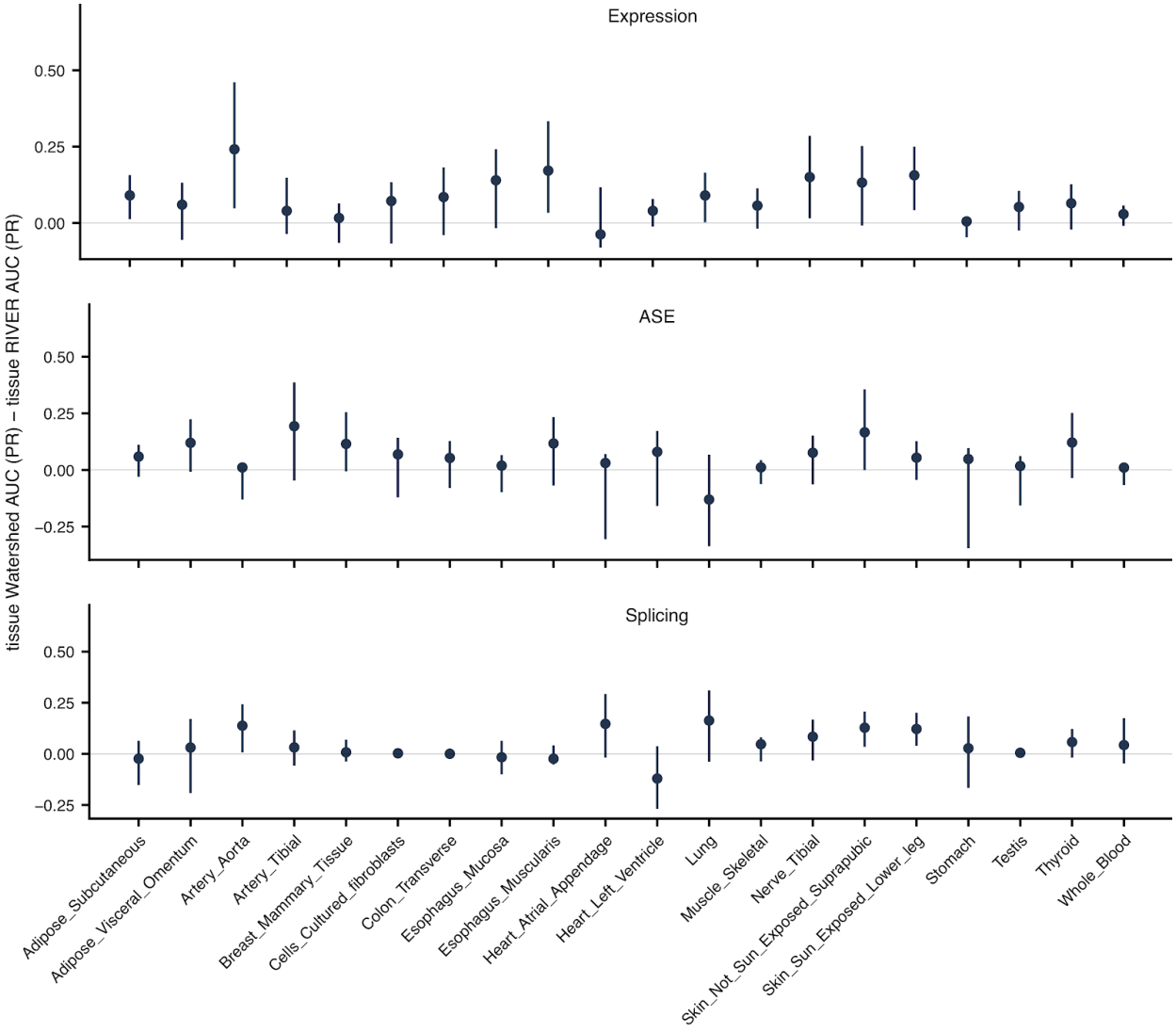

Figure S34. Difference in area under precision recall curves in single tissues. Difference in the area under the precision recall curves between tissue-Watershed and tissue-RIVER (y-axis) in a single tissue (x-axis), shown for expression, ASE, and splicing outlier signals (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression. Error bars (95% confidence interval) on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see Supplementary methods).

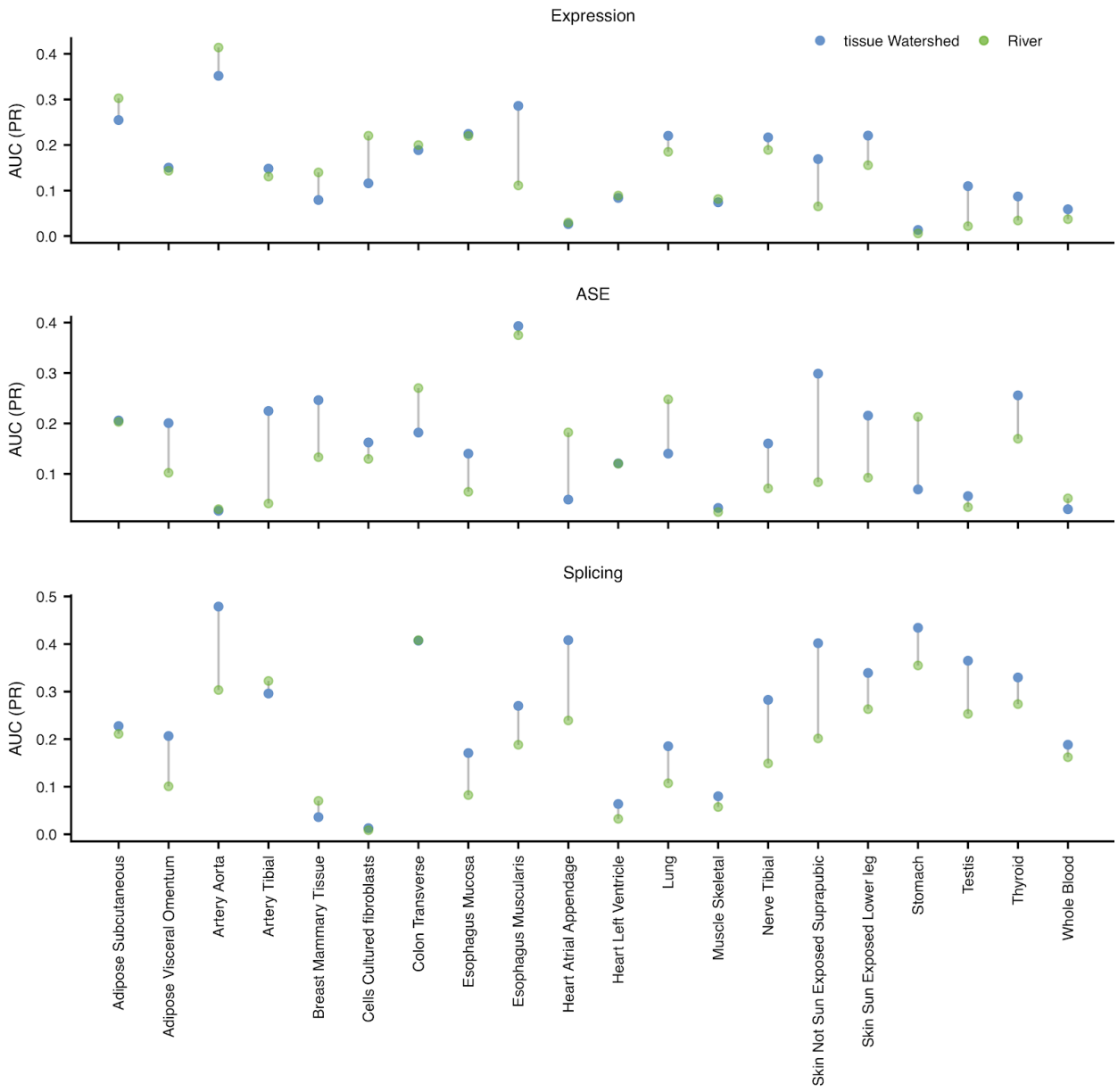

Figure S35. Area under precision recall curves in single tissues. Area under precision recall curves evaluated on outlier calls in a single tissue (x-axis) for each of the three outlier types (rows) based on a tissue-Watershed model trained across single tissues (blue) and a RIVER model trained on the median outlier signal (green). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression.

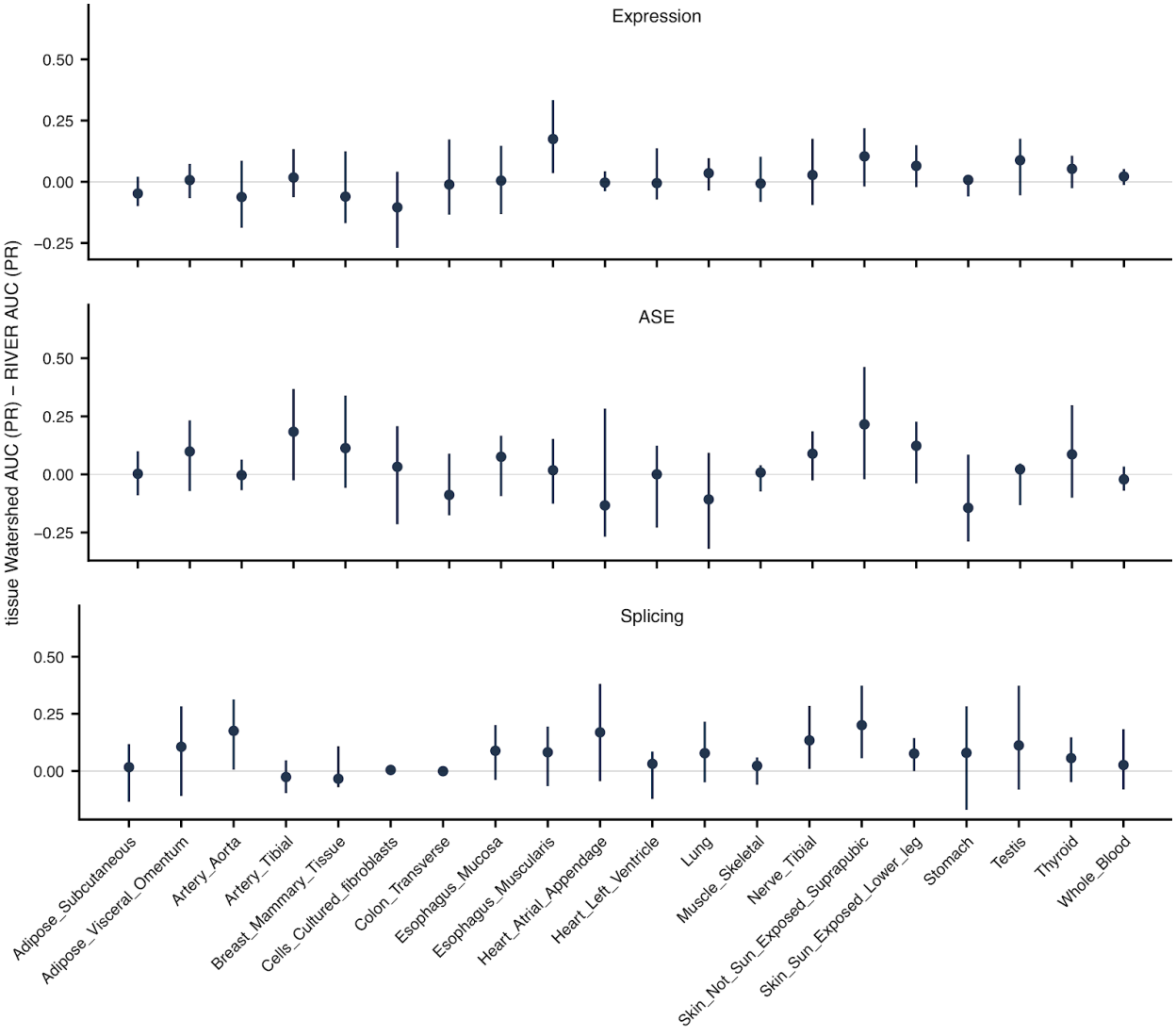

Figure S36. Difference in area under precision recall curves in single tissues. Difference in the area under the precision recall curves between tissue-Watershed and a RIVER model trained on the median outlier signal (y-axis) in a single tissue (x-axis), shown for expression, ASE, and splicing outlier signals (rows). Precision recall curves in each tissue generated using held out pairs of individuals where both individuals share the same rare variant and have observed outlier signal for the gene of interest. We limit to tissues that have at least 5 held out pairs of individuals that have outlier labels in ASE, splicing, and expression. Error bars (95% confidence interval) on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see Supplementary methods).

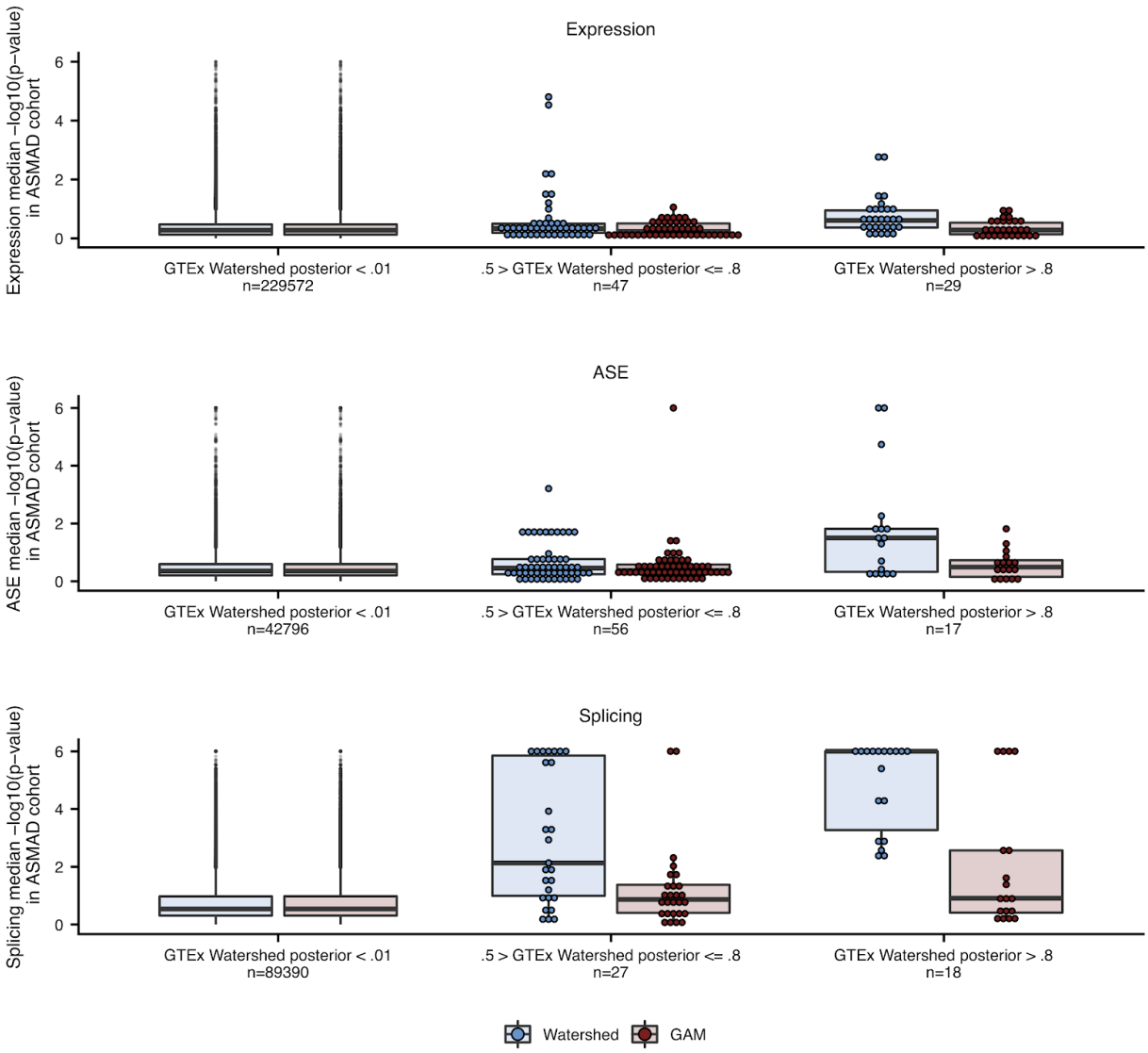

Figure S37. Replication in ASMAD cohort. Expression, ASE, and splicing outlier $-\log_{10}(\text{p-value} + 1\text{x}10^{-6})$ in ASMAD cohort of genes nearby rare variants binned by: GTEx Watershed posterior probability in the corresponding outlier type (blue), and GTEx GAM posterior probability in the corresponding outlier type greater than a threshold set to match the number Watershed variants in the corresponding bin (red). This analysis is limited to GTEx rare variants present in the ASMAD cohort. The number of variant-gene pairs in each bin (n) is shown beneath the posterior threshold labels on the x-axis. If multiple GTEx individuals have the same rare variant, we report the median posterior probability across individuals. If multiple ASMAD individuals have the same rare variant, we report the median p-value across individuals. There are 10 variant-gene pairs in the GTEx Watershed posterior > .8 bin that have ASMAD splicing outlier p-value exactly equal to 0 (or equivalently $-\log_{10}(\text{p-value} + 1\text{x}10^{-6})$ equal

to 6). This p-value point mass at 0 is a result of SPOT calculating p-values from an empirical distribution.

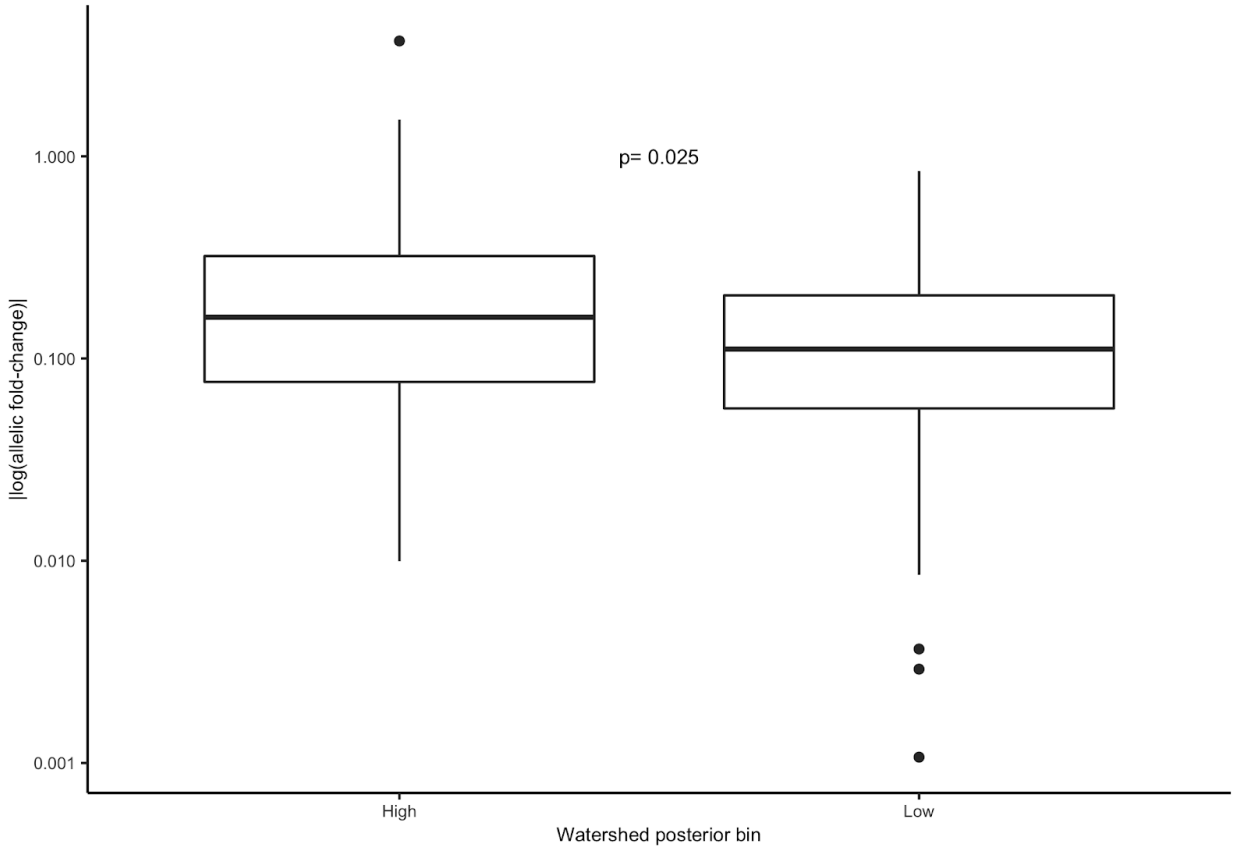

Figure S38. MPRA results. For 52 high Watershed expression (score >= 0.5) rare variants and 98 low Watershed expression (score < 0.5) variants nearby 62 eOutlier genes, the log fold-change in expression between the reference and edited alleles. p-value for the difference between Watershed bins is calculated from a one-sided Wilcoxon rank sum test.

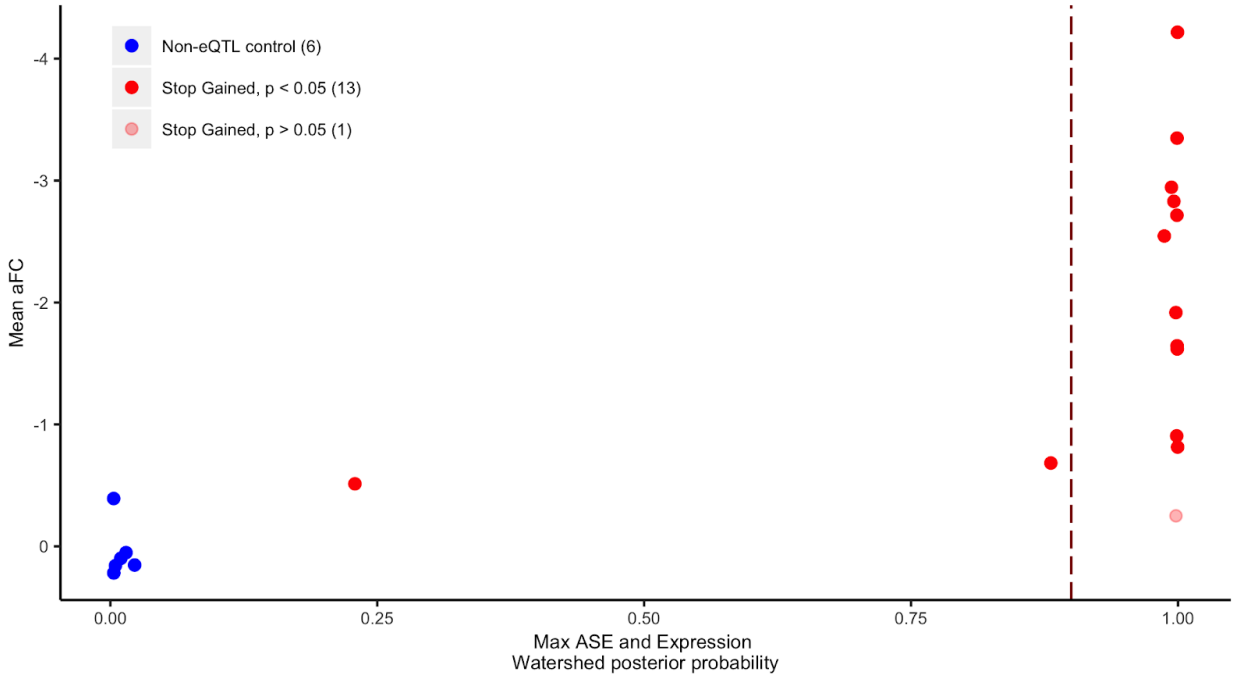

Figure S39. Experimental validation by editing 20 variants into inducible-Cas9 293T cell lines. 14 stop-gained variants were edited into cell lines, and their effect was evaluated using allelic fold change (aFC), shown on the y-axis, with the variant's maximum of ASE or expression Watershed score along the x-axis. When compared to negative control variants, 13 of the 14 edited variants caused significant aFC of their target genes (dark red). Non-eQTL control variants shown here are the 6 with Watershed scores available out of the 30 edited in total, and are not expected to induce an aFC effect.

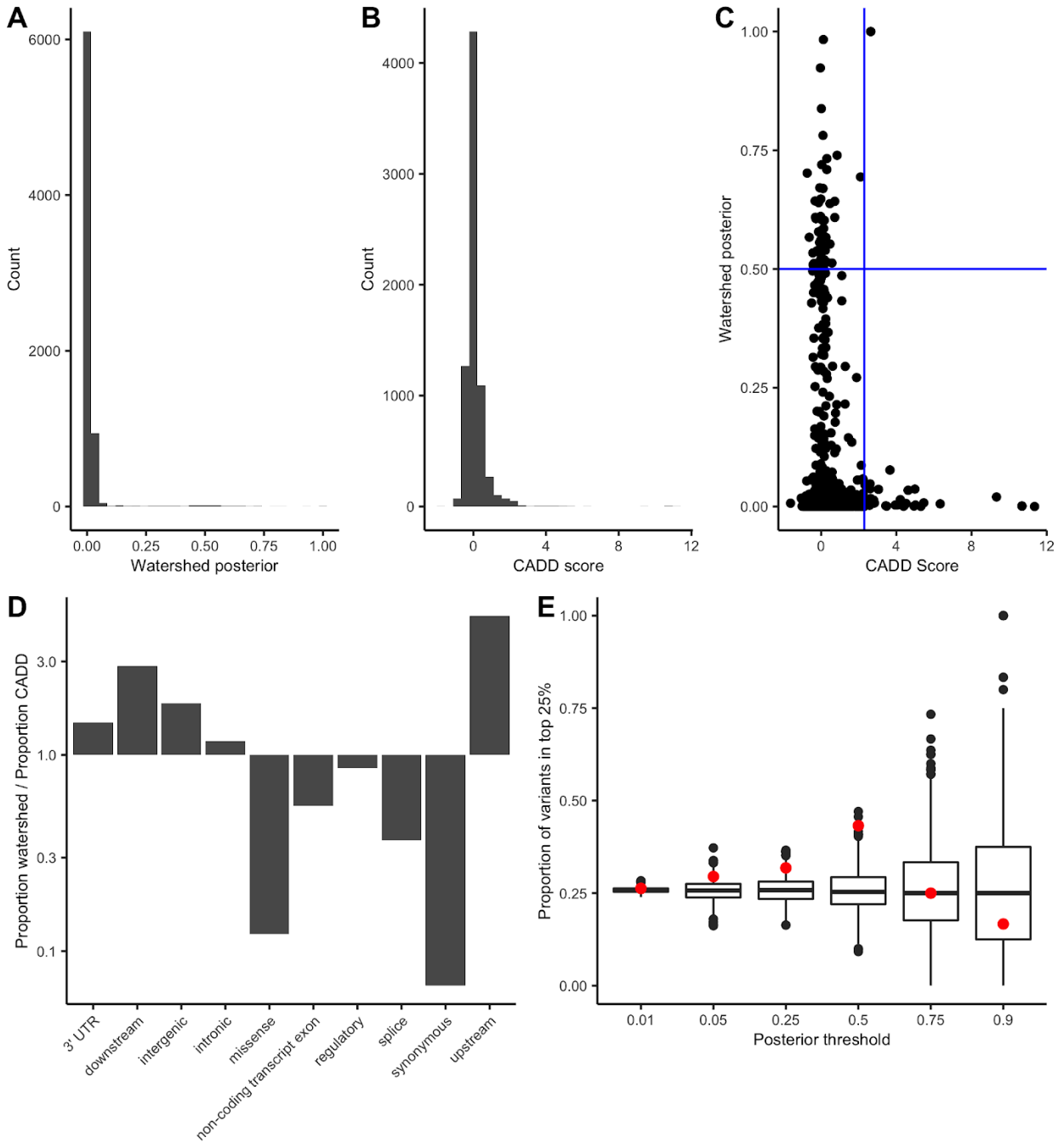

Figure S40. High CADD and Watershed variants in UKBB. (A) Distribution of the maximum Watershed posterior per variant for the set of variants in co-localized regions tested by Watershed and in UKBB. **(B)** Distribution of CADD scores per variant for the same set of variants in co-localized regions tested by Watershed and in UKBB. **(C)** The maximum Watershed posterior vs. CADD score for the tested variants in UKBB. The blue lines represent cut-offs of watershed posterior > 0.5, and the matching CADD threshold, 2.3, to obtain the same number of variants. **(D)** Of the high watershed and CADD variants in colocalized regions, the proportion of Watershed variants belonging to a specific category over the proportion of CADD

variants in the same category. The y-axis is log-scaled, so bars below 1 indicate the category is more common in high CADD variants, and vice versa. **(E)** Filtering by the CADD score that returns the same number of variants as the Watershed posterior on the x-axis, and returning the proportion that fall in the top 25% of effect sizes across traits in co-localized regions (red), and the proportion obtained by selecting a random set of tested variants equal in size (black).

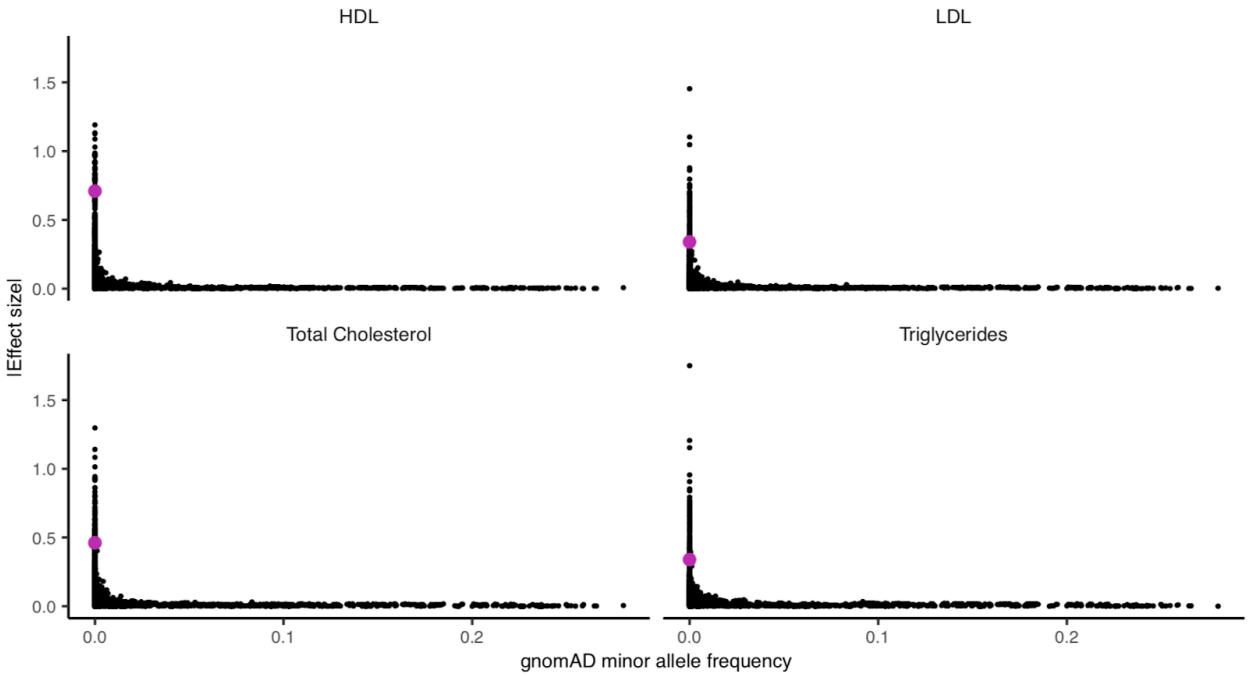

Figure S41. Distribution of rs564796245 effect sizes in MVP. All variants within a 250kb window of the high Watershed variant, in pink, rs564796245, tested for four related traits in the MVP cohort. The variant has a minor allele count of 11 in MVP, and for the set of rare variants tested in this window with a gnomAD non-Finnish European AF < 0.1%, it falls in the 99th percentile for HDL, 95th for LDL, 97th for Total Cholesterol, and 95th for Triglycerides.

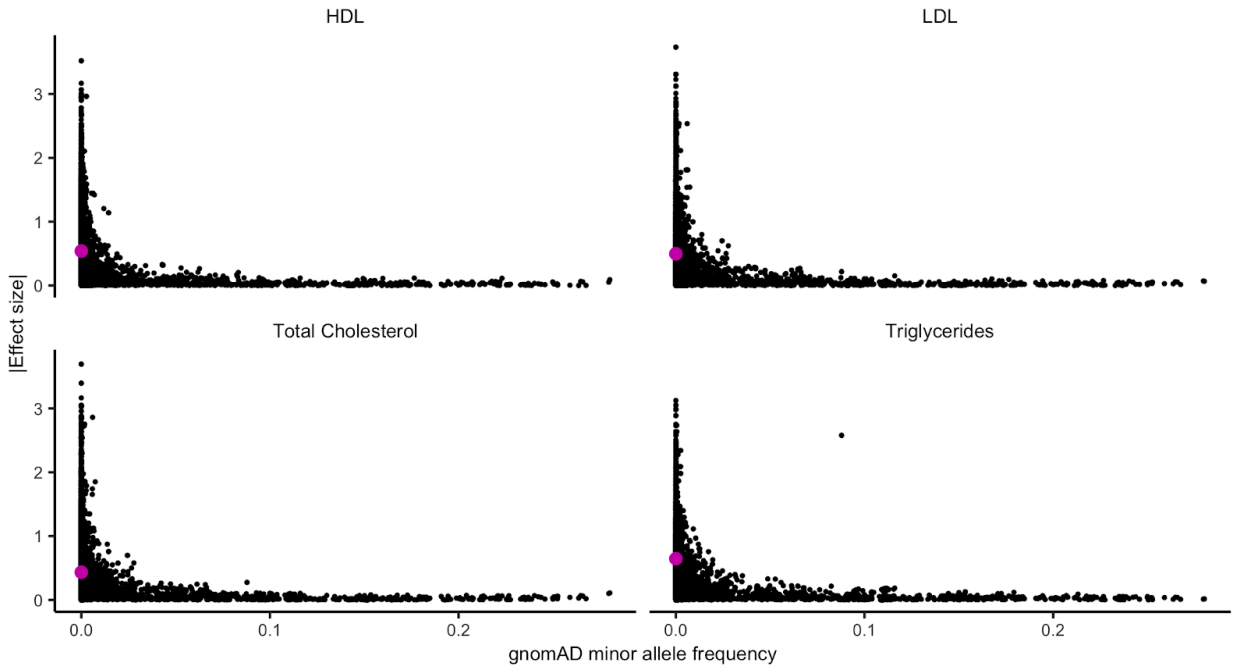

Figure S42. Distribution of rs564796245 effect sizes in JHS. All variants within a 250kb window of the high Watershed variant, in pink, rs564796245, tested for four related traits in the JHS cohort. The variant has a minor allele count of 4 in JHS, and for the set of rare variants tested in this window with a gnomAD non-Finnish European AF < 0.1%, it falls in the 69th percentile for HDL, 66th for LDL, 62nd for Total Cholesterol, and 72nd for Triglycerides.

Captions for Table S1 to S3

Table S1. Rare SVs impacting multiple genes. Table listing all rare SVs that are associated with outlier expression in more than one gene in the same individual, including the gene, median Z-score, aseOutlier p-value, sOutlier p-value, SV type and GTEx allele frequency. This table can be found online as a separate excel file.

Table S2. Tissue mapping for Roadmap to GTEx. Table mapping tissues collected in GTEx to equivalent tissues assayed in the Epigenomics Roadmap project. This includes 12 unique Roadmap tissues and 14 unique GTEx tissues, with some different GTEx tissues mapping to the same Roadmap tissue. This table can be found online as a separate excel file.

Table S3. Watershed genomic annotations. Table summarizing the 47 genomic annotations used in Watershed. This includes a description of each annotation, the source of each annotation, the imputation value used for each annotation (if the annotation was undefined for a particular variant), and the transformation used to aggregate across all SNVs mapped to (gene, individual) pair for each annotation (only applicable if a (gene, individual) pair had more than one SNV mapped to the gene). This table can be found online as a separate excel file.

Watershed AUC (PR) - RIVER (AUC) (PR) and corresponding 95% Confidence intervals			
Training and evaluation data	**Expression**	**ASE**	**Splicing**
Standard training data Standard evaluation data (Fig 4D)	**0.050** [-0.012, 0.11]	**0.049** [-0.045, 0.14]	**0.097** [0.034, 0.16]
Training data filter 1 Standard evaluation data (Fig S28A)	**0.056** [-0.0011, 0.11]	**0.043** [-0.046, 0.16]	**0.069** [-0.0045, 0.13]
Training data filter 2 Standard evaluation data (Fig S28B)	**0.046** [-5×10^{-5}, 0.087]	**-0.0096** [-0.087, 0.065]	**0.024** [0.0037, 0.042]
Training data filter 3 Standard evaluation data (Fig S28C)	**0.045** [0.00011, 0.085]	**0.0056** [-0.067, 0.075]	**0.024** [0.0062, 0.037]
Evaluation data filter 1 Standard training data (Fig S28D)	**0.033** [0.0031, 0.059]	**0.032** [0.0015, 0.054]	**0.066** [0.03, 0.099]
Evaluation data filter 2 Standard training data (Fig S28E)	**0.05** [0.028, 0.07]	**0.047** [0.022, 0.068]	**0.066** [0.039, 0.091]
Evaluation data filter 3 Standard training data (Fig S28F)	**0.066** [0.043, 0.088]	**0.033** [0.016, 0.049]	**0.076** [0.049, 0.1]

Table S4. Change in area under precision recall curves between Watershed and RIVER. Table summarizing the difference in area under the precision recall curves (AUC (PR)) between Watershed and RIVER for each of the three outlier types. 95% confidence intervals on these statistics generated using non-parametric bootstrapping with 20,000 bootstrapped samples (see Supplementary methods). Results shown across 7 different filters placed of Watershed training training or evaluation data (rows of table; See Supplementary methods) corresponding to 7 precision recall curves described in Fig 4D and Fig S28. Standard data corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.01). Filter 1 corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an outlier (median p-value < 0.05). Filter 2 corresponds to filtering to genes where all 3 outlier signals have at least one individual that is an

outlier (median p-value < 0.1). Filter 3 corresponds to filtering to genes where at least 1 outlier signals has at least one individual that is an outlier (median p-value < 0.01).

Gene:variant pair	Median expression p-value	Median Watershed expression posterior
P2RX7: chr12:g.121133096G>T	0.0105	0.996
ZNF350: chr19:g.51986869G>A	0.0619	0.925
CADM1: chr11:g.115500916G>A	0.779	0.00249
TSSC1: chr2:g.3377790T>C	0.0323	0.757
ARMC5: chr16:g.31460010C>T	0.0186	0.973

Table S5. Replication of SardiNIA Project "candidate causal rare variants". The SardiNIA Project (*46*) identified 30 "candidate causal rare variants" (and corresponding regulated genes). The above table shows 5 of the 30 "candidate causal rare variants" that were also present in an individual in GTEx v8, along with corresponding expression outlier p-value and Watershed expression posterior in GTEx v8 individuals. If multiple GTEx v8 individuals harbor the rare variant, we computed the median expression outlier p-value and median Watershed expression posterior across those individuals. SardiNIA Project rare variant calls were lifted to the hg38 genome build from the hg19 genome build using the Genome Browser (*5*). The variants from the SardiNIA Project were prioritized with expression outliers, followed by filtering based on genomic annotations. It is important to note that some of the genomic annotations used as input to Watershed were the same genomic annotations used by the SardiNIA Project to generate their list of "candidate causal rare variants".

Captions for Tables S6 to S11

Table S6. MPRA summary statistics. Table of summary statistics from the MPRA of 150 variants shown in Fig S38. Each row includes the affected gene, variant chromosome and position, the reference and alternative alleles, the base mean expression, log2 fold-change in expression, nominal and adjusted expression p-values, the base mean allelic expression, log2 fold-change in allelic expression, nominal and adjusted allelic expression p-values, eOutlier and aseOutlier p-values, Watershed scores for total expression and ASE, and the variant Watershed score bin. This table can be found online as a separate excel file.

Table S7. CRISPR summary statistics. Table including non-eQTL control variants and rare stop-gained variants with resulting summary statistics obtained from editing each variant into inducible-Cas9 293T cell lines. Columns include the variant type, chromosome and position, allelic fold-change, p-value, Bonferroni corrected p-values, outlier statistics, and Watershed scores for all outlier types. This table can be found online as a separate excel file.

Table S8. Rare variants impacting well-studied genes. Table listing outliers and associated rare variants for 11 well-studied genes. Each row is an individual-gene-variant combination and includes the gene, outlier value and type, Watershed scores for splicing, expression and ASE, and any nearby rare SVs, if applicable. This table can be found online as a separate excel file.

Table S9. UKBB traits and colocalizations. Table of the 34 UKBB traits included in our analysis and the number of colocalized genes and rare GTEx variants associated with each trait that overlap those tested in the UKBB dataset. This table can be found online as a separate excel file.

Table S10. High Watershed variants with high effect sizes. Table of the rare GTEx variants that had both high Watershed scores and high trait effect sizes for the set of UKBB traits tested. This includes the variant, gene, Watershed score, trait, effect size, and the effect size percentile. This table can be found online as a separate excel file.

Table S11. Asthma and cholesterol variant information. Table including outlier values and Watershed scores for the trait-associated variants shown in Fig 5D-E. The two asthma associated variants are found in six individuals and the high cholesterol associated variant is found in one individual. The table includes sOutlier, eOutlier and aseOutlier p-values and Watershed scores as well as the trait effect size and p-value for each individual-variant combination.Thistablecan foundonlineasaseparateexcelfile.

Study	Trait	MAC	Beta	SE	p-value
MVP	HDL	11	0.7098	0.4463	0.1118
MVP	LDL	11	-0.3401	0.4460	0.4457
MVP	Total cholesterol	11	-0.4618	0.4460	0.3005
MVP	Triglycerides	11	-0.3399	0.4464	0.4463
JHS	HDL	4	0.5394	0.4687	0.2499
JHS	LDL	4	-0.4973	0.4916	0.3118
JHS	Total cholesterol	4	-0.4335	0.4854	0.3719
JHS	Triglycerides	4	-0.6451	0.4877	0.186

Table S12. Cholesterol associations for rs564796245 in MVP and JHS. Table including outlier values and Watershed scores for the trait-associated variants shown in Fig 5D-E. The two asthma associated variants are found in six individuals and the high cholesterol associated variant is found in one individual. The table includes sOutlier, eOutlier and aseOutlier p-values and Watershed scores as well as the trait effect size and p-value for each individual-variantcombination.MAC=minorallelecount,SE=standarderror.

References and Notes

1. A. Keinan, A. G. Clark, Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012). doi:10.1126/science.1217283 Medline

2. C. F. Wright, D. R. FitzPatrick, H. V. Firth, Paediatric genomics: Diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 325 (2018). doi:10.1038/nrg.2018.12 Medline

3. L. Bomba, K. Walter, N. Soranzo, The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017). doi:10.1186/s13059-017-1212-4 Medline

4. S. B. Montgomery, T. Lappalainen, M. Gutierrez-Arcelus, E. T. Dermitzakis, Rare and common regulatory variation in population-scale sequenced human genomes. *PLOS Genet.* **7**, e1002144 (2011). doi:10.1371/journal.pgen.1002144 Medline

5. Y. Zeng, G. Wang, E. Yang, G. Ji, C. L. Brinkmeyer-Langford, J. J. Cai, Aberrant gene expression in humans. *PLOS Genet.* **11**, e1004942 (2015). doi:10.1371/journal.pgen.1004942 Medline

6. M. Pala, Z. Zappala, M. Marongiu, X. Li, J. R. Davis, R. Cusano, F. Crobu, K. R. Kukurba, M. J. Gloudemans, F. Reinier, R. Berutti, M. G. Piras, A. Mulas, M. Zoledziewska, M. Marongiu, E. P. Sorokin, G. T. Hess, K. S. Smith, F. Busonero, A. Maschio, M. Steri, C. Sidore, S. Sanna, E. Fiorillo, M. C. Bassik, S. J. Sawcer, A. Battle, J. Novembre, C. Jones, A. Angius, G. R. Abecasis, D. Schlessinger, F. Cucca, S. B. Montgomery, Population- and individual-specific regulatory variation in Sardinia. *Nat. Genet.* **49**, 700–707 (2017). doi:10.1038/ng.3840 Medline

7. X. Li, A. Battle, K. J. Karczewski, Z. Zappala, D. A. Knowles, K. S. Smith, K. R. Kukurba, E. Wu, N. Simon, S. B. Montgomery, Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *Am. J. Hum. Genet.* **95**, 245–256 (2014). doi:10.1016/j.ajhg.2014.08.004 Medline

8. R. D. Hernandez, L. H. Uricchio, K. Hartman, C. Ye, A. Dahl, N. Zaitlen, Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* **51**, 1349–1355 (2019). doi:10.1038/s41588-019-0487-7 Medline

9. A. Battle, Z. Khan, S. H. Wang, A. Mitrano, M. J. Ford, J. K. Pritchard, Y. Gilad, Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015). doi:10.1126/science.1260793 Medline

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

11. L. Frésard, C. Smail, N. M. Ferraro, N. A. Teran, X. Li, K. S. Smith, D. Bonner, K. D. Kernohan, S. Marwaha, Z. Zappala, B. Balliu, J. R. Davis, B. Liu, C. J. Prybol, J. N. Kohler, D. B. Zastrow, C. M. Reuter, D. G. Fisk, M. E. Grove, J. M. Davidson, T. Hartley, R. Joshi, B. J. Strober, S. Utiramerur, L. Lind, E. Ingelsson, A. Battle, G. Bejerano, J. A. Bernstein, E. A. Ashley, K. M. Boycott, J. D. Merker, M. T. Wheeler, S. B. Montgomery; Undiagnosed Diseases Network; Care4Rare Canada Consortium,

Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat. Med.* **25**, 911–919 (2019). Medline

12. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). doi:10.1038/s41586-018-0579-z Medline

13. J. M. Gaziano, J. Concato, M. Brophy, L. Fiore, S. Pyarajan, J. Breeling, S. Whitbourne, J. Deen, C. Shannon, D. Humphries, P. Guarino, M. Aslan, D. Anderson, R. LaFleur, T. Hammond, K. Schaa, J. Moser, G. Huang, S. Muralidhar, R. Przygodzki, T. J. O'Leary, Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016). doi:10.1016/j.jclinepi.2015.09.016 Medline

14. H. A. Taylor Jr., J. G. Wilson, D. W. Jones, D. F. Sarpong, A. Srinivasan, R. J. Garrison, C. Nelson, S. B. Wyatt, Toward resolution of cardiovascular health disparities in African Americans: Design and methods of the Jackson Heart Study. *Ethn. Dis.* **15** (Suppl 6), S6–S4, 17 (2005). Medline

15. X. Li, Y. Kim, E. K. Tsang, J. R. Davis, F. N. Damani, C. Chiang, G. T. Hess, Z. Zappala, B. J. Strober, A. J. Scott, A. Li, A. Ganna, M. C. Bassik, J. D. Merker, I. M. Hall, A. Battle, S. B. Montgomery; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017). doi:10.1038/nature24267 Medline

16. Materials and methods are available as supplementary materials.

17. P. Mohammadi, S. E. Castel, B. B. Cummings, J. Einson, C. Sousa, P. Hoffman, S. Donkervoort, Z. Jiang, P. Mohassel, A. R. Foley, H. E. Wheeler, H. K. Im, C. G. Bonnemann, D. G. MacArthur, T. Lappalainen, Genetic regulatory variation in populations informs transcriptome analysis in rare disease. *Science* **366**, 351–356 (2019). doi:10.1126/science.aay0256 Medline

18. F. Aguet, A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel, B. Jo, S. Kasela, S. Kim-Hellmuth, Y. Liang, M. Oliva, P. E. Parsana, E. Flynn, L. Fresard, E. R. Gaamzon, A. R. Hamel, Y. He, F. Hormozdiari, P. Mohammadi, M. Muñoz-Aguirre, Y. Park, A. Saha, A. V. Segrć, B. J. Strober, X. Wen, V. Wucher, S. Das, D. Garrido-Martín, N. R. Gay, R. E. Handsaker, P. J. Hoffman, S. Kashin, A. Kwong, X. Li, D. MacArthur, J. M. Rouhana, M. Stephens, E. Todres, A. Viñuela, G. Wang, Y. Zou, The GTEx Consortium, C. D. Brown, N. Cox, E. Dermitzakis, B. E. Engelhardt, G. Getz, R. Guigo, S. B. Montgomery, B. E. Stranger, H. K. Im, A. Battle, K. G. Ardlie, T. Lappalainen, The GTEx Consortium

atlas of genetic regulatory effects across human tissues. bioRxiv 787903 [Preprint]. 3 October 2019. https://doi.org/10.1101/787903.

19. F. Spitz, Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin. Cell Dev. Biol.* **57**, 57–67 (2016). doi:10.1016/j.semcdb.2016.06.017 Medline

20. P. H. L. Krijger, W. de Laat, Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771–782 (2016). doi:10.1038/nrm.2016.138 Medline

21. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski, B. M. Neale, M. J. Daly, D. G. MacArthur; Genome Aggregation Database Consortium, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). doi:10.1038/s41586-020-2308-7 Medline

22. Z.-F. Yang, S. Mott, A. G. Rosmarin, The Ets transcription factor GABP is required for cell-cycle progression. *Nat. Cell Biol.* **9**, 339–346 (2007). doi:10.1038/ncb1548 Medline

23. Y. Takahashi, J. B. Rayman, B. D. Dynlacht, Analysis of promoter binding by the E2F and pRB families in vivo: Distinct E2F proteins mediate activation and repression. *Genes Dev.* **14**, 804–816 (2000). Medline

24. S. Gordon, G. Akopyan, H. Garban, B. Bonavida, Transcription factor YY1: Structure, function, and therapeutic implications in cancer biology. *Oncogene* **25**, 1125–1142 (2006). doi:10.1038/sj.onc.1209080 Medline

25. T. Han, S. Oh, K. Kang, ETS family protein GABP is a novel co-factor strongly associated with genomic YY1 binding sites in various cell lines. *Genes Genomics* **38**, 119–125 (2016). doi:10.1007/s13258-015-0358-2

26. A. Saha, Y. Kim, A. D. H. Gewirtz, B. Jo, C. Gao, I. C. McDowell, B. E. Engelhardt, A. Battle, Co-expression networks reveal the tissue-specific regulation of transcription and splicing. *Genome Res.* **27**, 1843–1858 (2017). doi:10.1101/gr.216721.116 Medline

27. V. K. Mittal, J. F. McDonald, De novo assembly and characterization of breast cancer transcriptomes identifies large numbers of novel fusion-gene transcripts of potential functional significance. *BMC Med. Genomics* **10**, 53 (2017). doi:10.1186/s12920-017-0289-7 Medline

28. P. López-Nieva, P. Fernández-Navarro, O. Graña-Castro, E. Andrés-León, J. Santos, M. Villa-Morales, M. Á. Cobos-Fernández, L. González-Sánchez, M. Malumbres, M. Salazar-Roa, J. Fernández-Piqueras, Detection of novel fusion-transcripts by RNA-Seq in

T-cell lymphoblastic lymphoma. *Sci. Rep.* **9**, 5179 (2019). doi:10.1038/s41598-019-41675-3 Medline

29. C. Neckles, S. Sundara Rajan, N. J. Caplen, Fusion transcripts: Unexploited vulnerabilities in cancer? *Wiley Interdiscip. Rev. RNA* **11**, e1562 (2020). doi:10.1002/wrna.1562 Medline

30. F. Baty, M. Brutsche, Fusion transcripts in lung cancer. *Lung Cancer* (2017).

31. S. Chen, J. Li, P. Zhou, X. Zhi, SPTBN1 and cancer, which links? *J. Cell. Physiol.* **235**, 17–25 (2020). doi:10.1002/jcp.28975 Medline

32. A. M. Fry, L. O'Regan, J. Montgomery, R. Adib, R. Bayliss, EML proteins in microtubule regulation and human disease. *Biochem. Soc. Trans.* **44**, 1281–1288 (2016). doi:10.1042/BST20160125 Medline

33. M. Burset, I. A. Seledtsov, V. V. Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364–4375 (2000). doi:10.1093/nar/28.21.4364 Medline

34. S. Zhang, K. E. Samocha, M. A. Rivas, K. J. Karczewski, E. Daly, B. Schmandt, B. M. Neale, D. G. MacArthur, M. J. Daly, Base-specific mutational intolerance near splice sites clarifies the role of nonessential splice nucleotides. *Genome Res.* **28**, 968–974 (2018). doi:10.1101/gr.231902.117 Medline

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

36. M. B. Shapiro, P. Senapathy, RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155–7174 (1987). doi:10.1093/nar/15.17.7155 Medline

37. C. J. Coolidge, R. J. Seely, J. G. Patton, Functional analysis of the polypyrimidine tract in pre-mRNA splicing. *Nucleic Acids Res.* **25**, 888–896 (1997). doi:10.1093/nar/25.4.888 Medline

38. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014). doi:10.1038/ng.2892 Medline

39. P. Rentzsch, D. Witten, G. M. Cooper, J. Shendure, M. Kircher, CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47** (D1), D886–D894 (2019). doi:10.1093/nar/gky1016 Medline

40. B. Georgi, D. Craig, R. L. Kember, W. Liu, I. Lindquist, S. Nasser, C. Brown, J. A. Egeland, S. M. Paul, M. Bućan, Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLOS Genet.* **10**, e1004229 (2014). doi:10.1371/journal.pgen.1004229 Medline

41. M. Brandt, A. Gokden, M. Ziosi, T. Lappalainen, A polyclonal allelic expression assay for detecting regulatory effects of transcript variants. bioRxiv 794081 [Preprint]. 7 October 2019. https://doi.org/10.1101/794081.

42. D. A. Forero, S. López-León, Y. González-Giraldo, D. R. Dries, A. J. Pereira-Morales, K. M. Jiménez, J. E. Franco-Restrepo, APOE gene and neuropsychiatric disorders and endophenotypes: A comprehensive review. *Am. J. Med. Genet. B. Neuropsychiatr. Genet.* **177**, 126–142 (2018). doi:10.1002/ajmg.b.32516 Medline

43. A. M. Habib, A. L. Okorokov, M. N. Hill, J. T. Bras, M.-C. Lee, S. Li, S. J. Gossage, M. van Drimmelen, M. Morena, H. Houlden, J. D. Ramirez, D. L. H. Bennett, D. Srivastava, J. J. Cox, Microdeletion in a FAAH pseudogene identified in a patient with high anandamide concentrations and pain insensitivity. *Br. J. Anaesth.* **123**, e249–e253 (2019). doi:10.1016/j.bja.2019.02.019 Medline

44. H. Kim, D. P. Mittal, M. J. Iadarola, R. A. Dionne, Genetic predictors for acute experimental cold and heat pain sensitivity in humans. *J. Med. Genet.* **43**, e40 (2006). doi:10.1136/jmg.2005.036079 Medline

45. A. N. Barbeira, R. Bonazzola, E. R. Gamazon, Y. Liang, Y. Park, S. Kim-Hellmuth, G. Wang, Z. Jiang, D. Zhou, F. Hormozdiari, B. Liu, A. Rao, A. R. Hamel, M. D. Pividori, F. Aguet, GTEx GWAS Working Group, L. Bastarache, D. M. Jordan, M. Verbanck, R. Do, GTEx Consortium, M. Stephens, K. Ardlie, M. McCarthy, S. B. Montgomery, A. V. Segrè, C. D. Brown, T. Lappalainen, X. Wen, H. K. Im, Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. bioRxiv 814350 [Preprint]. 23 May 2020. https://doi.org/10.1101/814350.

46. D. Préfontaine, J. Nadigel, F. Chouiali, S. Audusseau, A. Semlali, J. Chakir, J. G. Martin, Q. Hamid, Increased IL-33 expression by epithelial cells in bronchial asthma. *J. Allergy Clin. Immunol.* **125**, 752–754 (2010). doi:10.1016/j.jaci.2009.12.935 Medline

47. N. S. Grotenboer, M. E. Ketelaar, G. H. Koppelman, M. C. Nawijn, Decoding asthma: Translating genetic variation in IL33 and IL1RL1 into disease pathophysiology. *J. Allergy Clin. Immunol.* **131**, 856–865 (2013). doi:10.1016/j.jaci.2012.11.028 Medline

48. D. Smith, H. Helgason, P. Sulem, U. S. Bjornsdottir, A. C. Lim, G. Sveinbjornsson, H. Hasegawa, M. Brown, R. R. Ketchem, M. Gavala, L. Garrett, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, O. T. Magnusson, G. I. Eyjolfsson, I. Olafsson, P. T. Onundarson, O. Sigurdardottir, D. Gislason, T. Gislason, B. R. Ludviksson, D. Ludviksdottir, H. M. Boezen, A. Heinzmann, M. Krueger, C. Porsbjerg, T. S. Ahluwalia, J. Waage, V. Backer, K. A. Deichmann, G. H. Koppelman, K. Bønnelykke, H. Bisgaard, G. Masson, U. Thorsteinsdottir, D. F. Gudbjartsson, J. A. Johnston, I. Jonsdottir, K. Stefansson, A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. *PLOS Genet.* **13**, e1006659 (2017). doi:10.1371/journal.pgen.1006659 Medline

49. A. Mousas, G. Ntritsos, M.-H. Chen, C. Song, J. E. Huffman, I. Tzoulaki, P. Elliott, B. M. Psaty, Blood-Cell Consortium; P. L. Auer, A. D. Johnson, E. Evangelou, G. Lettre, A. P. Reiner, Rare coding variants pinpoint genes that control human hematological traits. *PLOS Genet.* **13**, e1006925 (2017). doi:10.1371/journal.pgen.1006925 Medline

50. T. A. Olafsdottir, F. Theodors, K. Bjarnadottir, U. S. Bjornsdottir, A. B. Agustsdottir, O. A. Stefansson, E. V. Ivarsdottir, J. K. Sigurdsson, S. Benonisdottir, G. I. Eyjolfsson, D. Gislason, T. Gislason, S. Guðmundsdóttir, A. Gylfason, B. V. Halldorsson, G. H. Halldorsson, T. Juliusdottir, A. M. Kristinsdottir, D. Ludviksdottir, B. R. Ludviksson, G.

Masson, K. Norland, P. T. Onundarson, I. Olafsson, O. Sigurdardottir, L. Stefansdottir, G. Sveinbjornsson, V. Tragante, D. F. Gudbjartsson, G. Thorleifsson, P. Sulem, U. Thorsteinsdottir, G. L. Norddahl, I. Jonsdottir, K. Stefansson, Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. *Nat. Commun.* **11**, 393 (2020). doi:10.1038/s41467-019-14144-8 Medline

51. D. Klarin, S. M. Damrauer, K. Cho, Y. V. Sun, T. M. Teslovich, J. Honerlaw, D. R. Gagnon, S. L. DuVall, J. Li, G. M. Peloso, M. Chaffin, A. M. Small, J. Huang, H. Tang, J. A. Lynch, Y.-L. Ho, D. J. Liu, C. A. Emdin, A. H. Li, J. E. Huffman, J. S. Lee, P. Natarajan, R. Chowdhury, D. Saleheen, M. Vujkovic, A. Baras, S. Pyarajan, E. Di Angelantonio, B. M. Neale, A. Naheed, A. V. Khera, J. Danesh, K.-M. Chang, G. Abecasis, C. Willer, F. E. Dewey, D. J. Carey, J. Concato, J. M. Gaziano, C. J. O'Donnell, P. S. Tsao, S. Kathiresan, D. J. Rader, P. W. F. Wilson, T. L. Assimes; Global Lipids Genetics Consortium; Myocardial Infarction Genetics (MIGen) Consortium; Geisinger-Regeneron DiscovEHR Collaboration; VA Million Veteran Program, Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018). doi:10.1038/s41588-018-0222-9 Medline

52. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012). doi:10.1038/nprot.2011.457 Medline

53. T. Hastie, R. Tibshirani, B. Narasimhan, G. Chu, impute: Imputation for microarray data. *Bioinformatics* **17**, 520–525 (2001).

54. P. Mohammadi, S. E. Castel, A. A. Brown, T. Lappalainen, Quantifying the regulatory effect size of *cis*-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017). doi:10.1101/gr.216747.116 Medline

55. R. Tewhey, D. Kotliar, D. S. Park, B. Liu, S. Winnicki, S. K. Reilly, K. G. Andersen, T. S. Mikkelsen, E. S. Lander, S. F. Schaffner, P. C. Sabeti, Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016). doi:10.1016/j.cell.2016.04.027 Medline

56. T. Magoč, S. L. Salzberg, FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011). doi:10.1093/bioinformatics/btr507 Medline

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

58. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014). doi:10.1186/s13059-014-0550-8 Medline

59. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan,

GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, Reference variance estimates and blacklisted genes for all GTEx v8 tissues for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://doi.org/10.5281/zenodo.3899574.

60. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, ANEVA-DOT code for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://doi.org/10.5281/zenodo.3406690.

61. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, SPOT code for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://zenodo.org/badge/latestdoi/209325700.

62. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, eOutlier code for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://zenodo.org/badge/latestdoi/210649448.

63. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, Watershed model for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://zenodo.org/badge/latestdoi/210165360.

64. N. M. Ferraro, B. J. Strober, J. Einson, N. S. Abell, F. Aguet, A. N. Barbeira, M. Brandt, M. Bucan, S. E. Castel, J. R. Davis, E. Greenwald, G. T. Hess, A. T. Hilliard, R. L. Kember, B. Kotis, Y. Park, G. Peloso, S. Ramdas, A. J. Scott, C. Smail, E. K. Tsang, S. M. Zekavat, M. Ziosi, Aradhana, TOPMed Lipids Working Group, K. G. Ardlie, T. L. Assimes, M. C. Bassik, C. D. Brown, A. Correa, I. Hall, H. K. Im, X. Li, P. Natarajan, GTEx Consortium, T. Lappalainen, P. Mohammadi, S. B. Montgomery, A. Battle, Code

used in all figures for: Transcriptomic signatures across human tissues identify functional rare genetic variation, Zenodo (2020); https://zenodo.org/badge/latestdoi/265935957.

65. C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, S. B. Montgomery, A. Battle, D. F. Conrad, I. M. Hall; GTEx Consortium, The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017). doi:10.1038/ng.3834 Medline

66. R. E. Handsaker, V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, S. A. McCarroll, Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015). doi:10.1038/ng.3200 Medline

67. E. J. Gardner, V. K. Lam, D. N. Harris, N. T. Chuang, E. C. Scott, W. S. Pittard, R. E. Mills, S. E. Devine; 1000 Genomes Project Consortium, The Mobile Element Locator Tool (MELT): Population-scale mobile element discovery and biology. *Genome Res.* **27**, 1916–1929 (2017). doi:10.1101/gr.218032.116 Medline

68. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi:10.1101/gr.229102 Medline

69. D. S. DeLuca, J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, G. Getz, RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012). doi:10.1093/bioinformatics/bts196 Medline

70. S. E. Castel, F. Aguet, P. Mohammadi, GTEx Consortium, K. G. Ardlie, T. Lappalainen, A vast resource of allelic expression data spanning human tissues. bioRxiv 792911 [Preprint]. 3 October 2019; https://doi.org/10.1101/792911.

71. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011). doi:10.1038/nbt.1754 Medline

72. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013). doi:10.1093/bib/bbs017 Medline

73. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi:10.1093/bioinformatics/btp352 Medline

74. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A.

Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis; Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi:10.1038/nature14248 Medline

75. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock; The Gene Ontology Consortium, Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). doi:10.1038/75556 Medline

76. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47** (D1), D330–D338 (2019). doi:10.1093/nar/gky1055 Medline

77. H. Mi, A. Muruganujan, D. Ebert, X. Huang, P. D. Thomas, PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47** (D1), D419–D426 (2019). doi:10.1093/nar/gky1038 Medline

78. J. Lafferty, A. McCallum, F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data" (2001); https://repository.upenn.edu/cis_papers/159/.

79. D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **112**, 859–877 (2017). doi:10.1080/01621459.2017.1285773

80. G. Gong, F. J. Samaniego, Pseudo maximum likelihood estimation: Theory and applications. *Ann. Stat.* **9**, 861–869 (1981). doi:10.1214/aos/1176345526

81. J. O'Connell, D. Gurdasani, O. Delaneau, N. Pirastu, S. Ulivi, M. Cocca, M. Traglia, J. Huang, J. E. Huffman, I. Rudan, R. McQuillan, R. M. Fraser, H. Campbell, O. Polasek, G. Asiki, K. Ekoru, C. Hayward, A. F. Wright, V. Vitart, P. Navarro, J.-F. Zagury, J. F. Wilson, D. Toniolo, P. Gasparini, N. Soranzo, M. S. Sandhu, J. Marchini, A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genet.* **10**, e1004234 (2014). doi:10.1371/journal.pgen.1004234 Medline

82. B. N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genet.* **5**, e1000529 (2009). doi:10.1371/journal.pgen.1000529 Medline

83. S. Andrews, FastQC, GitHub (2010); https://github.com/s-andrews/FastQC.

84. F. Krueger, Trim Galore (2015); http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

85. B. van de Geijn, G. McVicker, Y. Gilad, J. K. Pritchard, WASP: Allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015). doi:10.1038/nmeth.3582 Medline

86. T. Schütze, F. Rubelt, J. Repkow, N. Greiner, V. A. Erdmann, H. Lehrach, Z. Konthur, J. Glökler, A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Anal. Biochem.* **410**, 155–157 (2011). doi:10.1016/j.ab.2010.11.029 Medline

87. D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison 3rd, H. O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009). doi:10.1038/nmeth.1318 Medline

88. A. Yahi, P. Hoffman, M. Brandt, P. Mohammadi, N. P. Tatonetti, T. Lappalainen, EdiTyper: A high-throughput tool for analysis of targeted sequencing data from genome editing experiments. bioRxiv 229088 [Preprint]. 30 July 2020. https://doi.org/10.1101/2020.07.30.229088.

89. A. Yahi, T. Lappalainen, P. Mohammadi, N. P. Tatonetti, RecNW: A fast pairwise aligner for targeted sequencing. bioRxiv 371989 [Preprint]. 19 July 2018. https://doi.org/10.1101/371989.

90. X. Wen, R. Pique-Regi, F. Luca, Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genet.* **13**, e1006646 (2017). doi:10.1371/journal.pgen.1006646 Medline

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