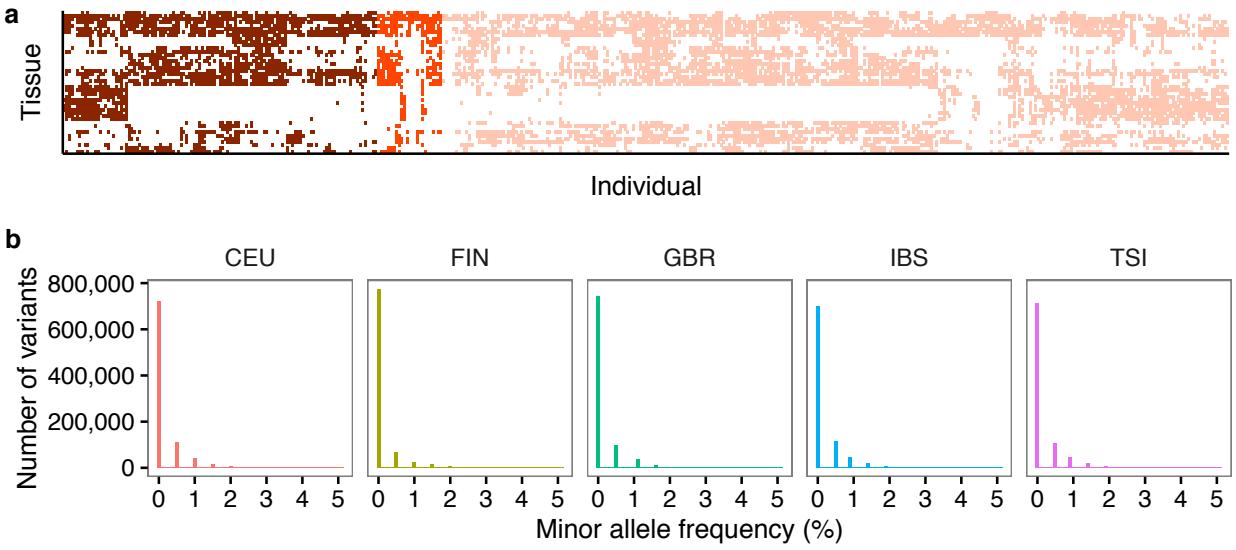
Extended data figures and tables

The impact of rare variation on gene expression across tissues

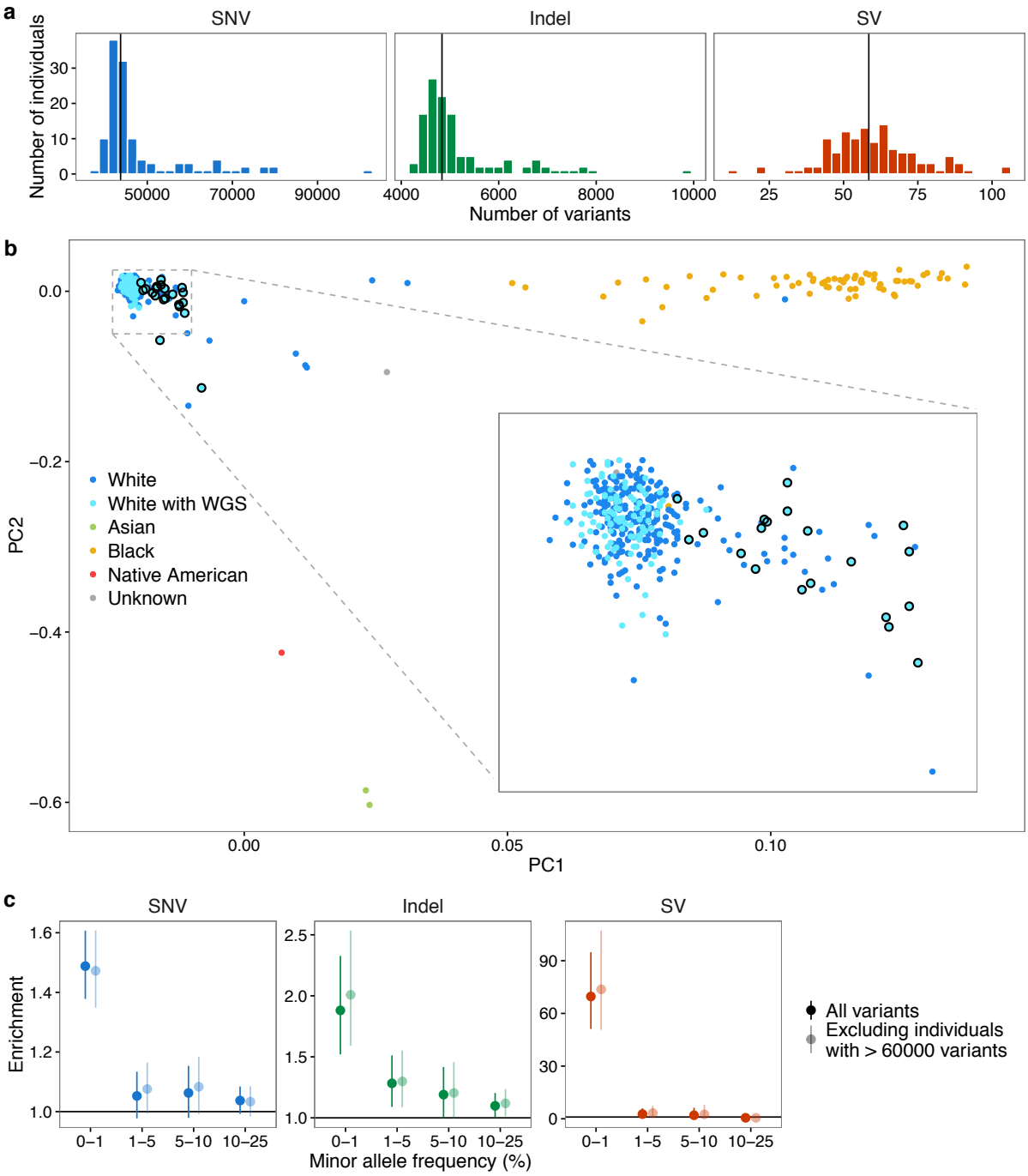
Xin Li[1†], Yungil Kim[2†], Emily K. Tsang[3†], Joe R. Davis[4†], Farhan N. Damani[2], Colby Chiang[5], Zachary Zappala[4], Benjamin J. Strober[6], Alexandra J. Scott[5], Andrea Ganna[7,8,9], Jason Merker[1], GTEx Consortium, Ira M. Hall[5,10,11], Alexis Battle[2*] and Stephen B. Montgomery[1,4*]

[1] Department of Pathology, Stanford University, Stanford, CA. [2] Department of Computer Science, Johns Hopkins University, Baltimore, MD. [3] Biomedical Informatics Program, Stanford University, Stanford, CA. [4] Department of Genetics, Stanford University, Stanford, CA. [5] McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO. [6] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD. [7] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA. [8] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA. [9] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA. [10] Department of Medicine, Washington University School of Medicine, St. Louis, MO. [11] Department of Genetics, Washington University School of Medicine, St. Louis, MO.

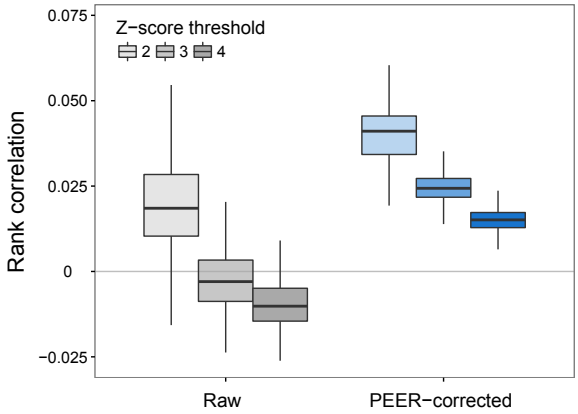
[†]equal contribution
*co-corresponding authors, alphabetical
Correspondence to ajbattle@cs.jhu.edu, smontgom@stanford.edu



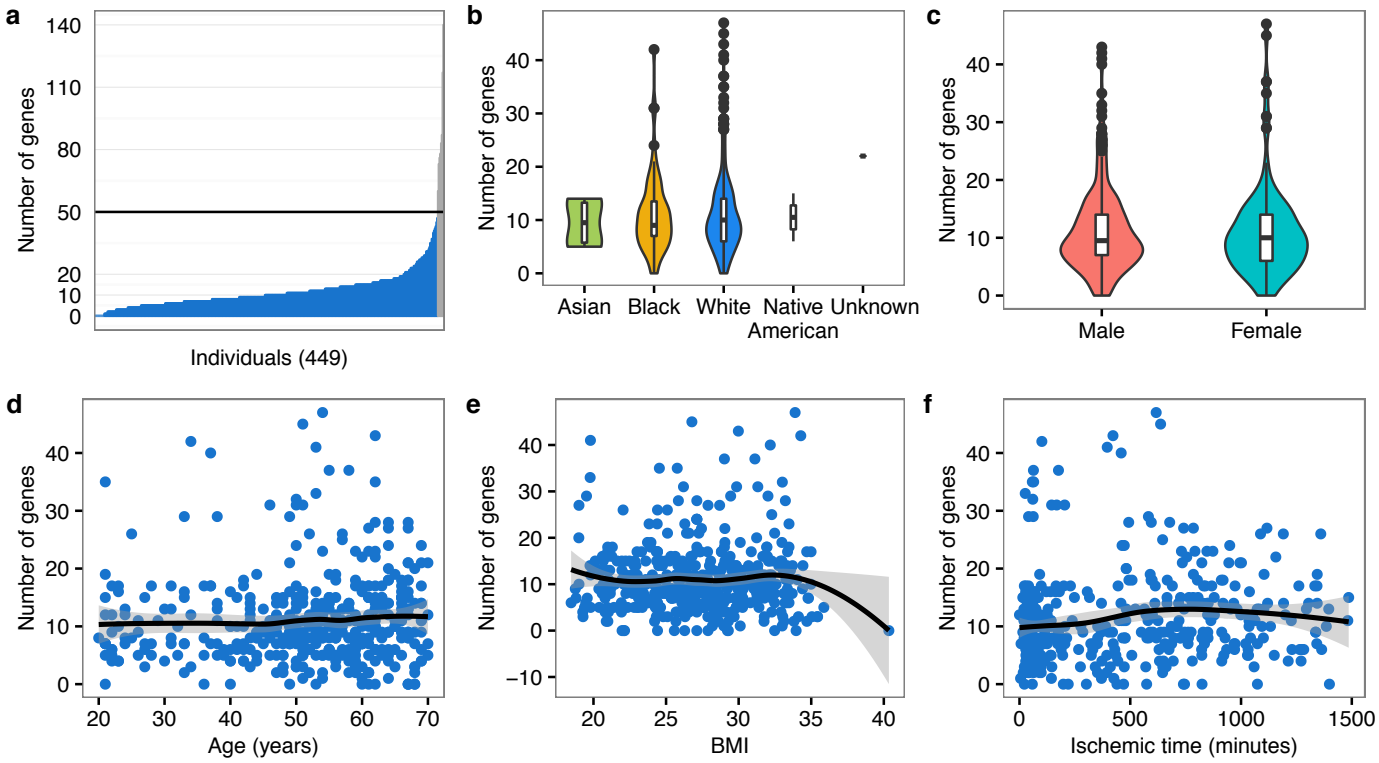
Extended Data Figure 1. Individuals sampled for each tissue and European population allele frequencies of rare variants included in the analysis. (a) Matrix of the 44 tissues and 449 individuals analyzed. Available tissue samples for each individual are depicted in red. The two highlighted groups of individuals on the left had whole genome sequencing data. The darker shade indicates the individuals of European descent, who were used for rare variant analyses. (b) European population allele frequency distributions in the 1000 Genomes European project of rare SNVs and indels analyzed (MAF ≤ 0.01 in GTEx individuals of European descent and in 1000 Genomes European super population).



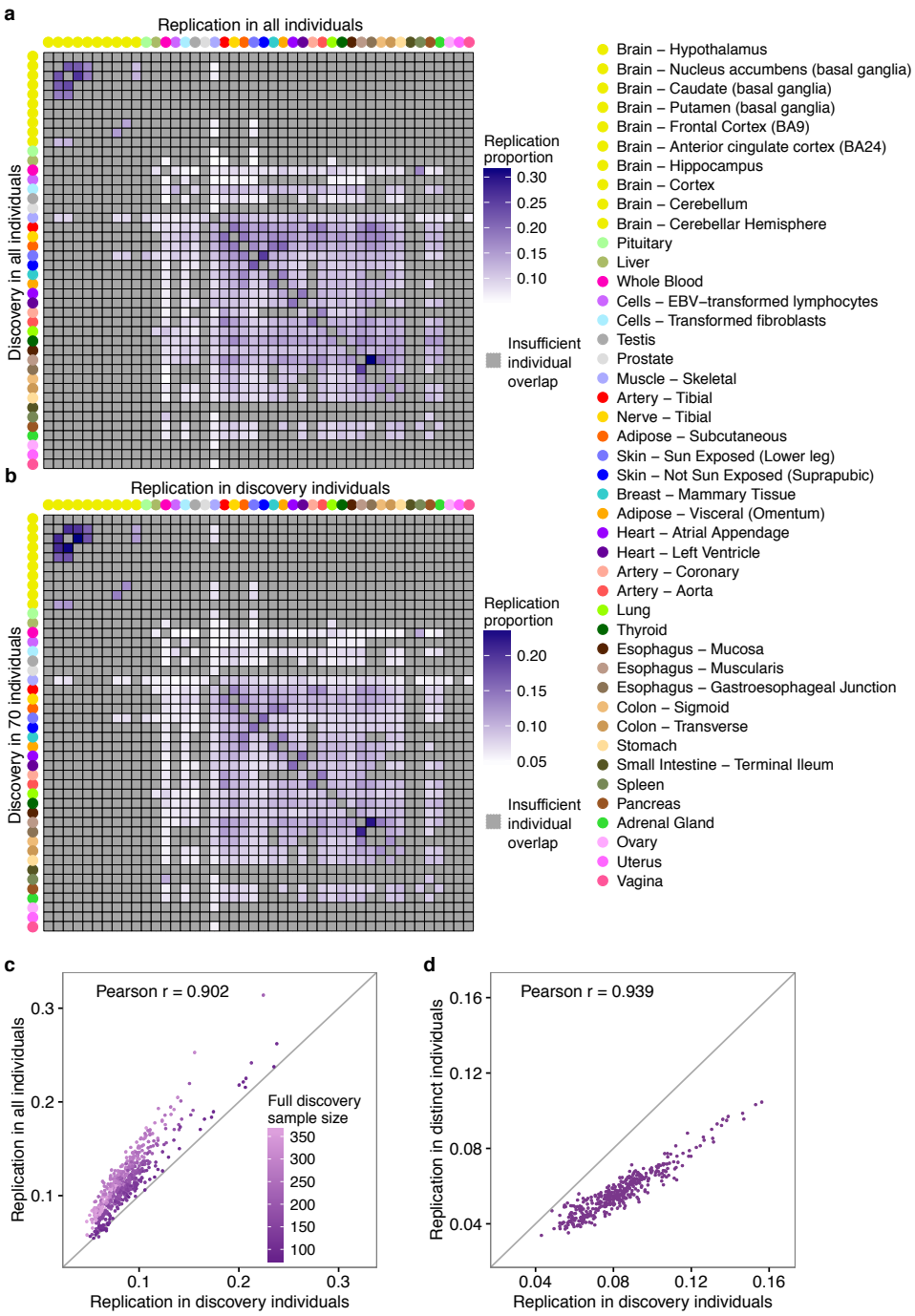
Extended Data Figure 2. Number of rare variants per individual. (a) The distribution of the number of variants of each type for individuals of European descent (reported as white). Certain individuals harbored many more rare variants than the population median (vertical black line). (b) Principal component analysis of all individuals. Individuals are plotted according to their first two genotype principal components (PCs) and colored by their reported ancestry. White individuals with whole genome sequencing data, included in (a), are colored in a lighter shade of blue and those with 60,000 or more rare variants are circled in black. The individuals with an excess of rare variants likely had African or Asian admixture. (c) Removing individuals with an excess of rare variants (circled in (b)) did not substantially affect the enrichment patterns.



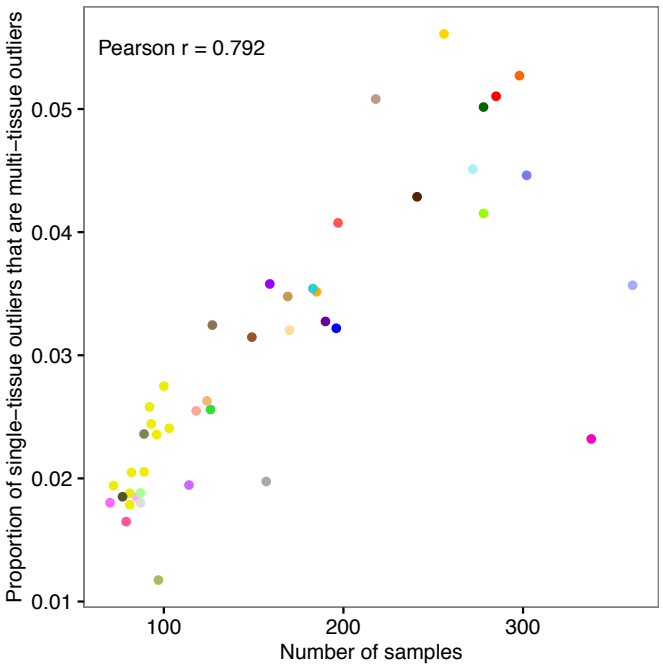
Extended Data Figure 3. PEER correction improves replication of outliers across tissues. Spearman rank correlation between outlier status in a set of four discovery tissues and the absolute expression in a replication tissue. We tested this correlation for three discovery |median Z-score| thresholds. We used each of the 27 tissues with at least 100 European individuals as a replication tissue and randomly selected four other tissues as the discovery set. We randomly sampled 10^5 individual and gene pairs. The same sets of tissues and individual and gene pairs were used for predicting outliers with both raw and PEER-corrected data.



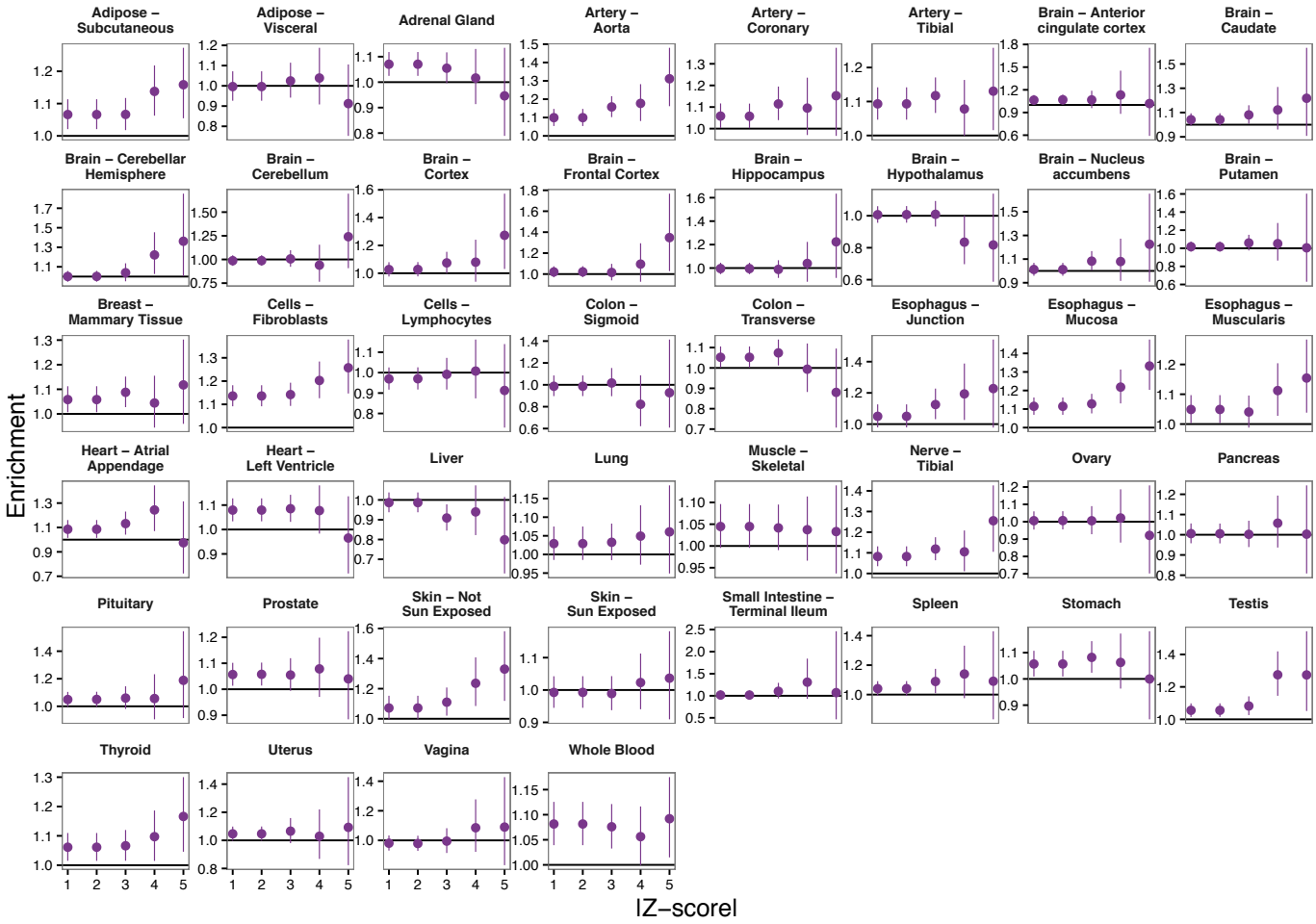
Extended Data Figure 4. Distribution of the number of genes with a multi-tissue outlier. (a) Distribution of the number of genes for which each individual was a multi-tissue outlier. Each individual was an outlier for a median of 10 genes. Individuals with 50 or more outliers are colored in grey and were excluded from downstream analyses as they may be driven by environmental or other non-genetic factors. (b–f) Distribution of the number of genes for which individuals, stratified by common covariates (race, sex, age, body mass index, and ischemic time), were multi-tissue outliers. For race and sex, we compared the distributions using an unsigned Wilcoxon rank sum test, while for the remaining covariates we used Spearman's ρ to test for association. Only age (Spearman's ρ = 0.101, P = 0.0333) and ischemic time (Spearman's ρ = 0.175, P = 0.000217) were nominally associated with the number of outlier genes per individual. The association with age fails to achieve significance after correcting for multiple testing using the Bonferroni method. Note that in (b) we only tested for a significant difference in the distribution of the number of outlier genes between White and Black individuals because there were too few individuals in the other groups.



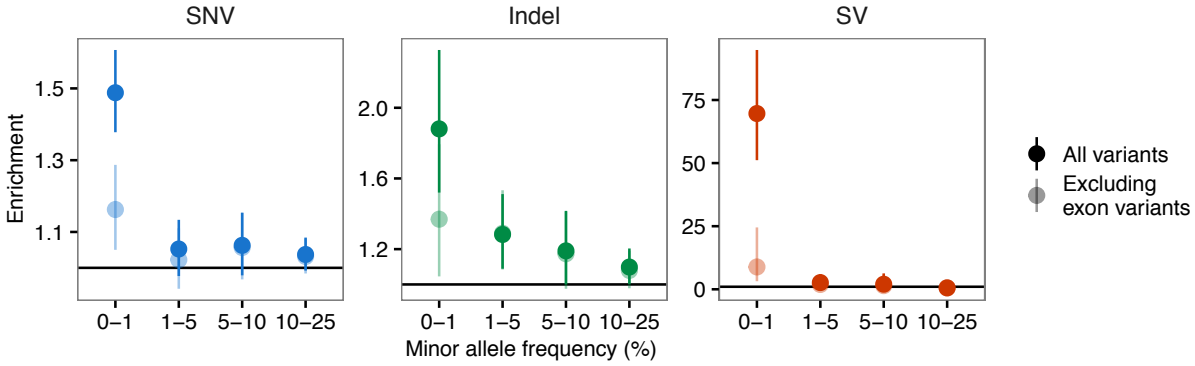
Extended Data Figure 5. Single-tissue outlier replication controlling for individual overlap in the discovery and replication sets. (a) Single-tissue outlier discovery and replication using all individuals, as in Fig. 1b, but data are only shown for pairs with at least 70 overlapping individuals. (b) For each pair of tissues with sufficient samples, outlier discovery and replication using only a set of 70 individuals that were sampled in both tissues. (c) Correlation between the replication values obtained from all samples and from a subset of 70 overlapping individuals per tissue pair. The replication rates decreased more when restricting to 70 individuals for discovery tissues with more samples in the full data set. (d) Correlation between replication in the 70 individuals used for discovery and replication assessed in a set of 70 individuals that included the outlier individual and 69 individuals excluded from the discovery set.



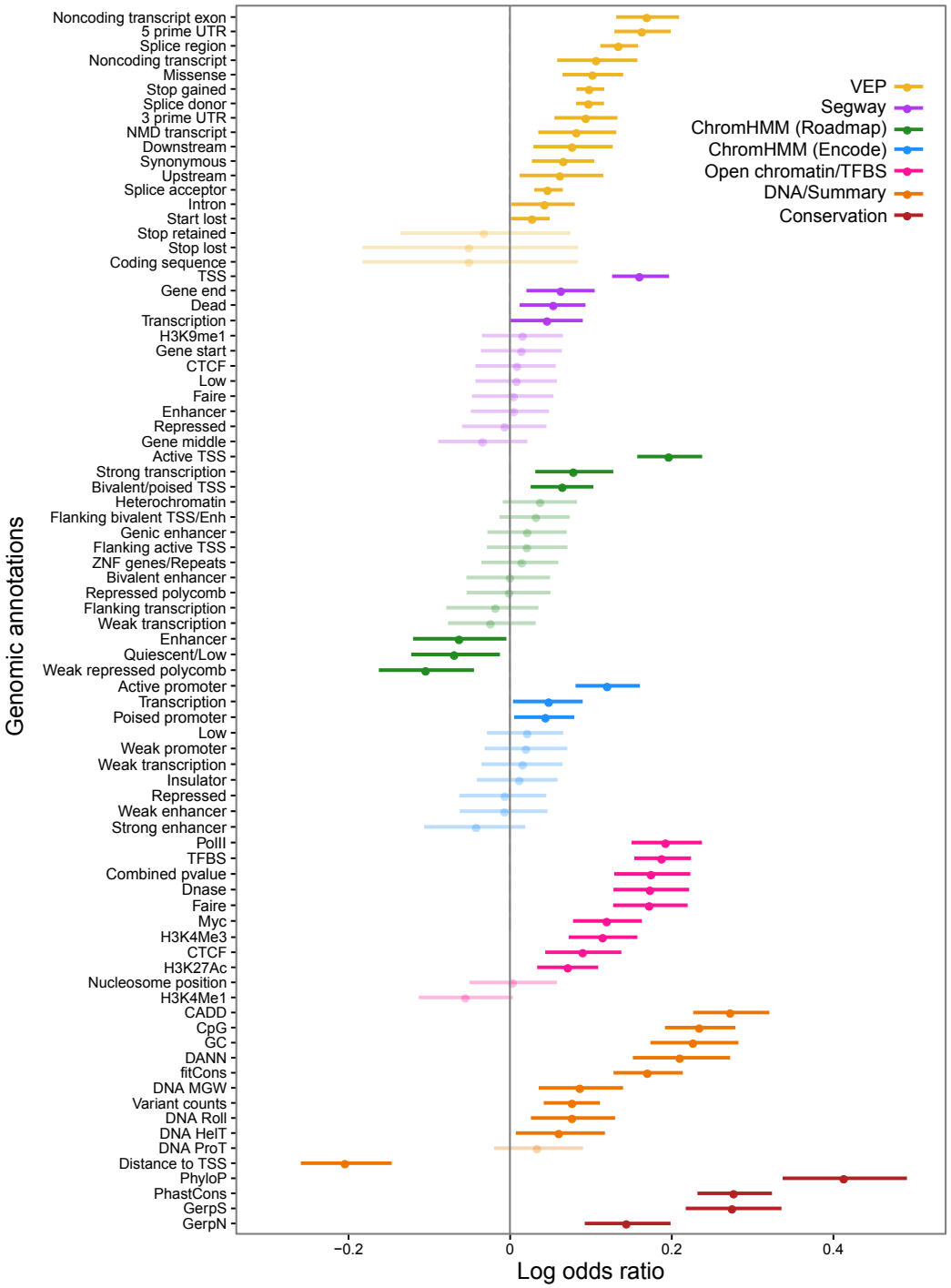
Extended Data Figure 6. Overlap between single-tissue and multi-tissue outliers. For each tissue, the proportion of (individual, gene) outlier pairs where the individual was also a multi-tissue outlier for the gene. This proportion increased with the tissue sample size. Points are colored by tissue following the convention in Fig 1.



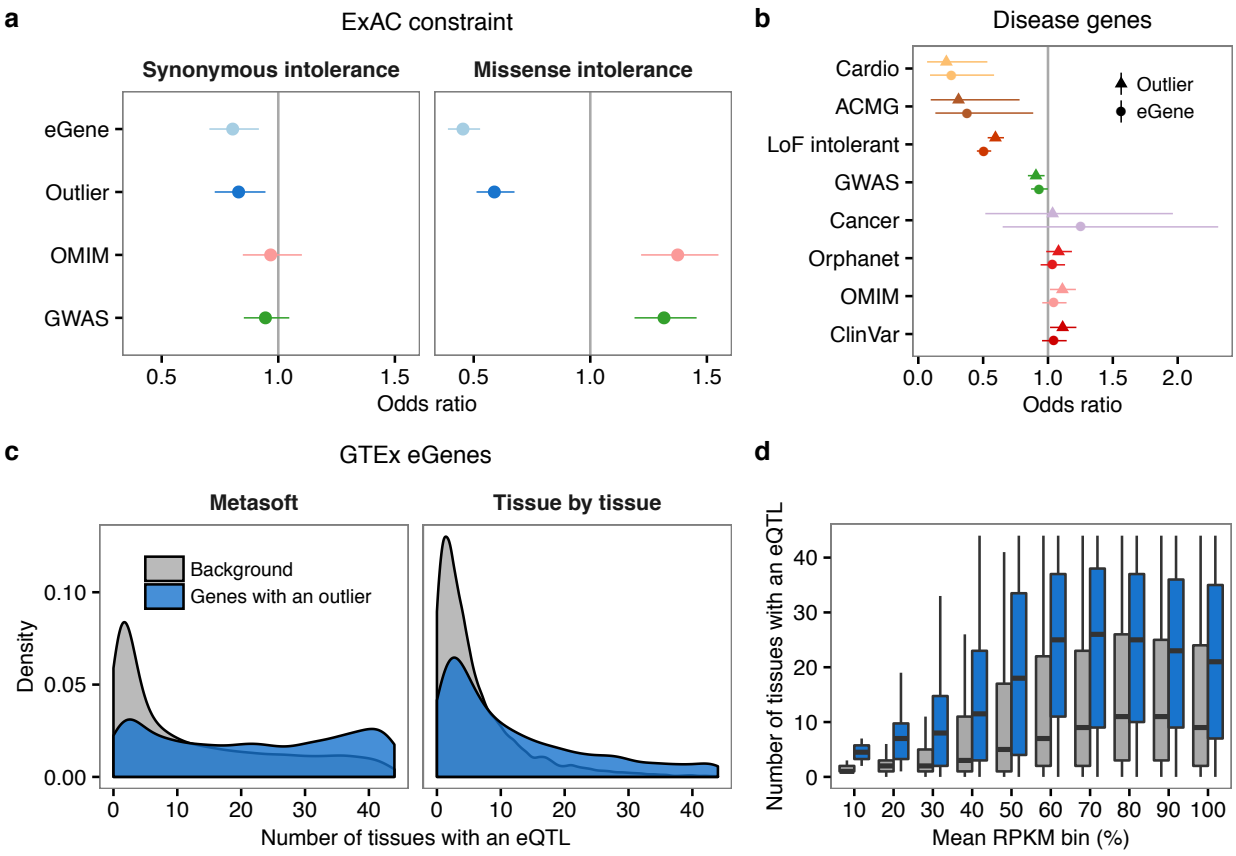
Extended Data Figure 7. Enrichment of rare variants in single-tissue outliers. For each tissue, rare SNV enrichment near genes with outliers in outlier compared with non-outlier individuals at increasing |Z-score| thresholds. Enrichments calculated as in Fig. 2. The rare variant enrichments varied between tissues though the overall pattern mirrored that of multi-tissue outliers when combining all the tissues (Fig. 2b). The high variance in the enrichments underscores the noise in single-tissue outlier discovery.



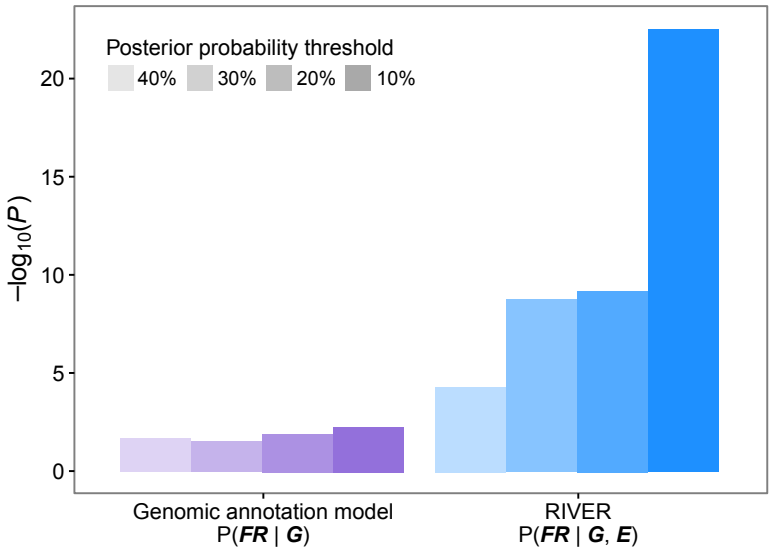
Extended Data Figure 8. Enrichment of rare variants when excluding exonic regions. As in Fig. 2a, enrichment of SNVs, indels, and SVs for outliers compared with the same genes in non-outliers either including all rare variants or excluding those overlapping protein-coding or lincRNA exons in Gencode v19 annotation. The enrichment of rare variants was weaker, but still significant, for all variant types when excluding exonic regions. The decreased enrichment was most striking for structural variants.



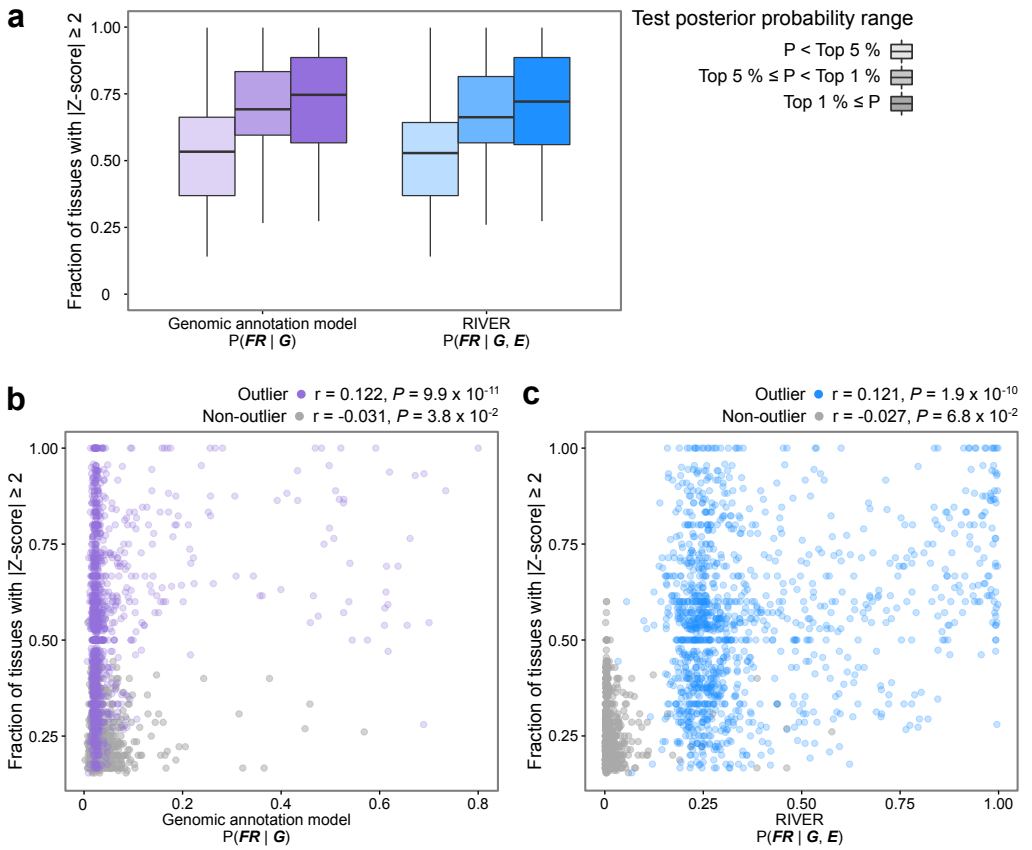
Extended Data Figure 9. Enrichment of functional genomic annotations among an expanded set of multi-tissue outliers. For outliers discovered with |median Z-score| ≥ 1.5 and allowing multiple outliers per genes, we calculated log odds ratios and 95% Wald confidence intervals from univariate logistic regressions modeling outlier status as a function of each genomic feature. When more than one feature corresponded to the same genomic annotation (e.g. the number or the presence of rare variants in a splice region; Extended Data Table 3), the feature with the highest enrichment is shown. Lighter shading indicates a non-significant log odds ratio (nominal $P > 0.05$).



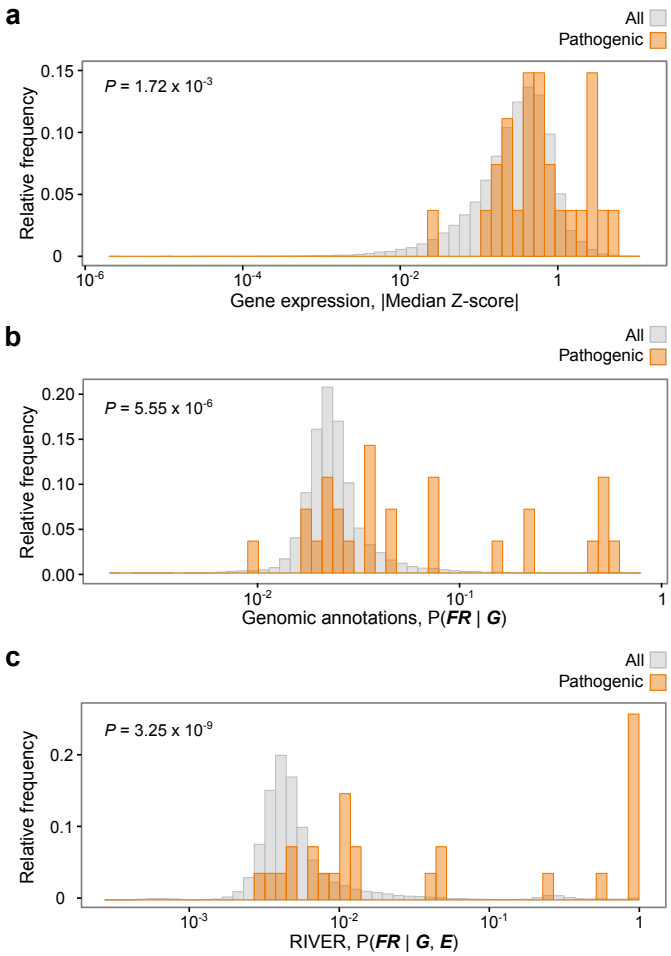
Extended Data Figure 10. Evolutionary constraint and regulatory control of multi-tissue outlier genes. (a) Odds ration of being intolerant to synonymous and missense variants for genes with multi-tissue outliers, eGenes, GWAS and OMIM genes. We used scores of synonymous and missense constraint provided by ExAC. As expected, GWAS and OMIM genes showed no enrichment or depletion for synonymous variation intolerant genes (synonymous Z-score above the 90th percentile). Genes with multi-tissue outliers and eGenes showed slight depletion for these genes. In contrast, genes with multi-tissue outliers and eGenes were strongly depleted for missense variation intolerant genes (missense Z-score above the 90th percentile) compared to OMIM and GWAS genes. (b) Comparison of the depletion of disease and loss-of-function (LoF) intolerant genes among genes with a multi-tissue outlier and eGenes. (c) Distribution of the number of tissues with an eQTL for genes with and without outliers. Genes with multi-tissue outliers had eQTLs in more tissues than genes without, which suggests that they are more susceptible to shared regulatory control. This result held whether we defined shared eQTLs with Metasoft (21 vs 6 tissues, Wilcoxon rank sum test $P < 2.2 \times 10^{-16}$) or through a tissue-by-tissue analysis (7 vs 3 tissues, $P < 2.2 \times 10^{-16}$). (d) This eGene enrichment was robust for different mean expression levels (RPKM) across tissues. The comparison between genes with and without outliers was nominally significant for all RPKM deciles (two-sided Wilcoxon rank sum tests, $P < 0.05$). Only the lowest decile was no longer significant after Bonferroni correction (all other $P < 5.74 \times 10^{-13}$).



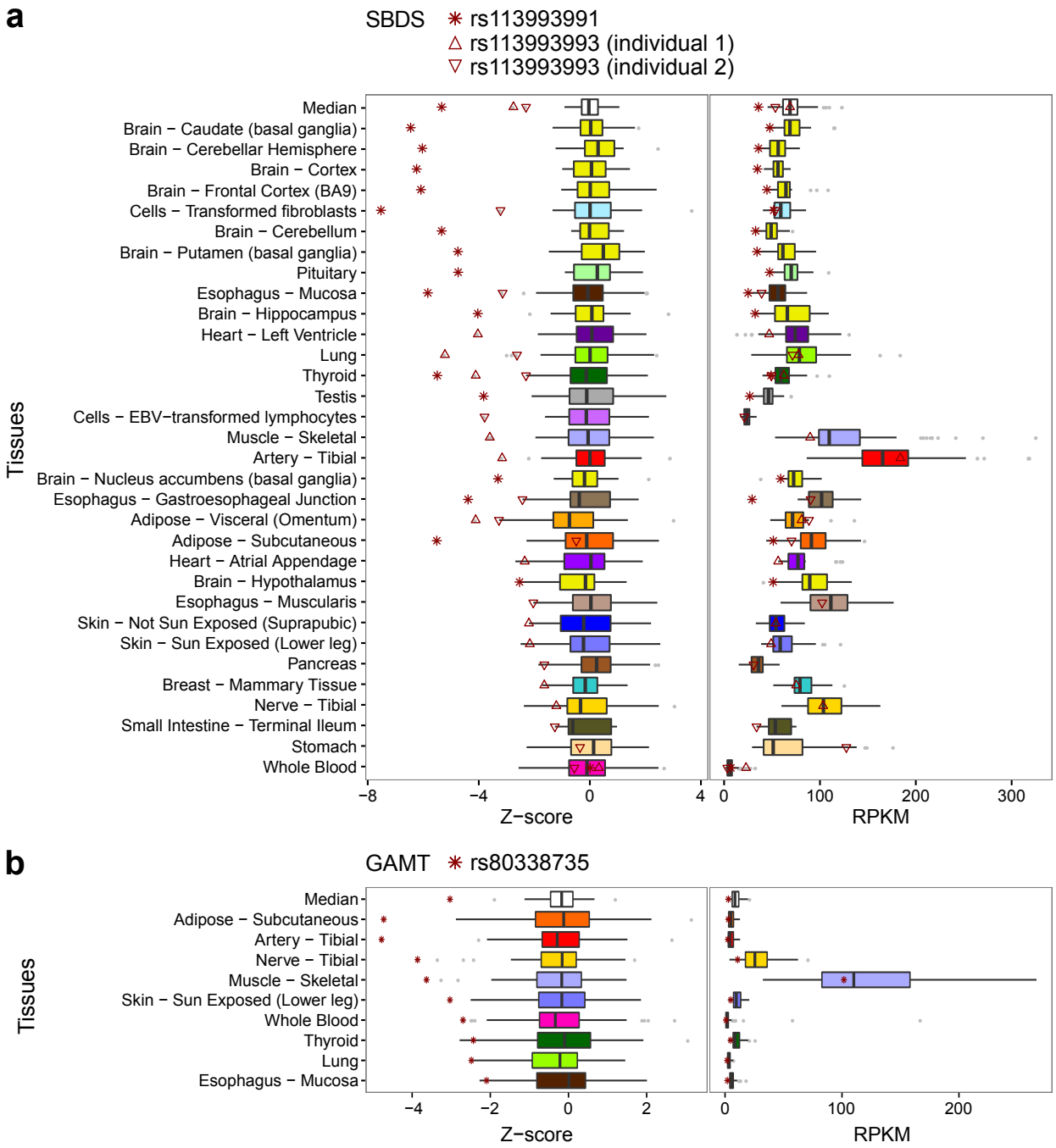
Extended Data Figure 11. RIVER scores were strongly associated with ASE. *P*-values from Fisher's exact test measuring the association between allelic imbalance and the posterior probability of a functional rare variant according to two models. We used four thresholds on the posterior probabilities (top 10%, 20%, 30% and 40%) from the two models. We evaluated ASE as the median across tissues of the absolute difference between the reference allele ratio and 0.5. We considered ASE in the top 10% of the empirical distribution to be allelic imbalance.



Extended Data Figure 12. The fraction of tissues with |Z-score| ≥ 2 in multi-tissue outliers was correlated with the posterior probability of a functional rare variant. (a) The fraction of tissues with |Z-score| ≥ 2 for three groups of multi-tissue outliers defined using thresholds on the test posterior probability of a functional rare variant. (b,c) Correlations, using Kendall's tau, between the fraction of tissues with |Z-score| ≥ 2 and the test probabilities from the genomic annotation model (b) and RIVER (c). We considered multi-tissue outliers and non-outliers separately for each model and calculated test posterior probabilities using 10-fold cross validation. Only individual and gene pairs with a fraction of tissues with |Z-score| ≥ 2 that was significantly different from 0.05 were considered (one-sided binomial exact test, Benjamini-Hochberg adjusted $P < 0.05$).



Extended Data Figure 13. Distributions of predictive scores for 27 individual and gene pairs with pathogenic variants compared to all variants. Relative frequency of (a) the |median Z-score|, (b) posterior probabilities from the genome annotation only model, and (c) posterior probabilities from RIVER for all individual and gene pairs (grey) and 27 pairs with pathogenic variants from ClinVar (orange). *P*-values were computed using a two-sided Wilcoxon rank sum test.



Extended Data Figure 14. Expression levels for genes proximal to pathogenic variants. *Z*-score and RPKM distributions for (a) *SBDS* and (b) *GA25* were compared to the values for four individuals carrying regulatory pathogenic variation (red asterisks and triangles). Three individuals carrying a total of two unique rare variants are shown for *SBDS*; one individual carrying one rare variant is shown for *GA25*. The median Z-score and RPKM values across tissues are shown at the top of each plot. Tissues are sorted in decreasing order of the difference between the average Z-score of individuals with a regulatory pathogenic variant and the median Z-score for the tissue.

Feature Name	Type	Description	Source
Duplication	binary	Presence of a rare duplication SV	Chiang et al.
CNV	binary	Presence of a rare CNV	Chiang et al.
Deletion	binary	Presence of a rare deletion SV	Chiang et al.
Breakend	binary	Presence of a rare breakend SV	Chiang et al.
Inversion	binary	Presence of a rare inversion SV	Chiang et al.
Splice	binary	Presence of a splice region, acceptor or donor variant	VEP
Frameshift	binary	Presence of a frameshift variant	VEP
Stop	binary	Presence of a start or stop lost or stop gained variant	VEP
TSS non-coding	binary	Presence of a rare, non-coding† SNV or indel between -250 and 750 bp from the TSS	VEP
Top 1% conserved non-coding	binary	Presence of a rare, non-coding† SNV or indel with a CADD or PhyloP score in the top 1% of all variants	VEP
Coding	binary	Presence of a rare missense, synonymous, stop retained, inframe deletion, inframe insertion	VEP
Non-conserved	binary	Presence of a rare non-coding† and non-conserved SNV or indel (not in the top 1% of CADD or PhyloP scores)	VEP
Distance to TSS	integer	Absolute distance (in bp) between the TSS and the closest rare variant	Gencode v19
Promoter	binary	Presence of a rare SNV or indel within a promoter of any of 19 tissue groups*	Epigenomics Roadmap
Enhancer	binary	Presence of a rare SNV or indel within an enhancer of any of 19 tissue groups*	Epigenomics Roadmap
TFBS	binary	Presence of a rare SNV or indel within any transcription factor binding site	CADD
CpG	numeric	Maximum percent CpG in a +/- 75 bp window over all rare variants	CADD
CADD	numeric	Maximum CADD score (SNVs only)	CADD
PhyloP	numeric	Maximum vertebrate PhyloP score	CADD
PhastCons	numeric	Maximum vertebrate PhastCons score	CADD
fitCons	numeric	Maximum fitCons score	CADD
GerpN	numeric	Maximum neutral Gerp score	CADD
GerpS	numeric	Maximum rejected substitution Gerp score	CADD

*The tissue groups were selected sets of tissues from the Epigenomics Roadmap project that matched at least one of the 44 GTEx tissues.
†Non-coding VEP categories include 3' UTR, 5' UTR, intron, upstream, downstream, intergenic, regulatory region, TFBS, and TFBS ablation.

Extended Data Table 1. Rare variant features tested for enrichment.

Gene List Name	Description	# Genes	Source
GWAS	Genes reported in the GWAS catalog to have an association with a complex trait or disease	9480	http://www.ebi.ac.uk/gwas/
OMIM	Genes in the OMIM Gene Map that are linked to a disorder with a known molecular cause	3576	http://www.omim.org/
OrphaNet	Genes associated with rare diseases as curated by OrphaNet	3451	http://www.orpha.net/
ClinVar	Genes reported in the ClinVar database to have an association with a disease	6279	http://www.ncbi.nlm.nih.gov/clinvar/
ACMG	Genes covered by the ACMG guidelines for incidental findings	58	http://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/
Cardio	Heritable cardiovascular disease genes	86	See Online Methods
Cancer	Genes implicated in heritable cancer predisposition	55	See Online Methods
LOF-intolerant	Genes in the ExAC database with a pLI score > 0.9	3230	http://www.exac.broadinstitute.org/

Extended Data Table 2. Disease gene sets tested for enrichment among genes with outliers.

Annotation	#	Type	Description	Category	Source
Noncoding transcript exon	2	binary/integer	Presence/number of rare noncoding transcript exon SNVs	VEP	VEP
5 prime UTR	2	binary/integer	Presence/number of rare 5' UTR SNVs	VEP	VEP
Splice region	2	binary/integer	Presence/number of rare splice region SNVs	VEP	VEP
Noncoding transcript	2	binary/integer	Presence/number of rare noncoding transcript SNVs	VEP	VEP
Missense	2	binary/integer	Presence/number of rare missense SNVs	VEP	VEP
Stop gained	2	binary/integer	Presence/number of rare stop gained SNVs	VEP	VEP
Splice donor	2	binary/integer	Presence/number of rare splice donor SNVs	VEP	VEP
3 prime UTR	2	binary/integer	Presence/number of rare 3' UTR SNVs	VEP	VEP
NMD transcript	2	binary/integer	Presence/number of rare NMD transcript SNVs	VEP	VEP
Downstream	2	binary/integer	Presence/number of rare downstream gene SNVs	VEP	VEP
Synonymous	2	binary/integer	Presence/number of rare synonymous SNVs	VEP	VEP
Upstream	2	binary/integer	Presence/number of rare upstream gene SNVs	VEP	VEP
Splice acceptor	2	binary/integer	Presence/number of rare splice acceptor SNVs	VEP	VEP
Intron	2	binary/integer	Presence/number of rare intron SNVs	VEP	VEP
Start lost	2	binary/integer	Presence/number of rare start lost SNVs	VEP	VEP
Stop retained	2	binary/integer	Presence/number of rare stop retained SNVs	VEP	VEP
Coding sequence	2	binary/integer	Presence/number of coding sequence SNVs	VEP	VEP
Stop lost	2	binary/integer	Presence/number of a stop lost SNVs	VEP	VEP
TSS	1	integer	Number of rare SNVs in TSS Segway segmentation	Segway	CADD
Gene end	1	integer	Number in GE0, GE1, and GE2 Segway segmentation	Segway	CADD
Dead	1	integer	Number in D Segway segmentation	Segway	CADD
Transcription	1	integer	Number in TF0, TF1, and TF2 Segway segmentation	Segway	CADD
Gene start	1	integer	Number in GS Segway segmentation	Segway	CADD
H3K9me1	1	integer	Number in H3K9me1 Segway segmentation	Segway	CADD
CTCF	1	integer	Number in C0 Segway segmentation	Segway	CADD
Low	1	integer	Number in L0 and L1 Segway segmentation	Segway	CADD
Enhancer	1	integer	Number in E/GM Segway segmentation	Segway	CADD
Faire	1	integer	Number in F0 and F1 Segway segmentation	Segway	CADD
Repressed	1	integer	Number in R0, R1, R2, R3, R4, and R5 Segway segmentation	Segway	CADD
Gene middle	1	integer	Number in GM0 and GM1 Segway segmentation	Segway	CADD
Active TSS	1	numeric	Proportion of 127 cell types in state "TSS" (maximum across rare SNVs)	ChromHMM (Roadmap)	CADD
Strong transcription	1	numeric	Proportion of 127 cell types in state "Tx" (maximum)	ChromHMM (Roadmap)	CADD
Bivalent/poised TSS	1	numeric	Proportion of 127 cell types in state "TssBiv" (maximum)	ChromHMM (Roadmap)	CADD
Heterochromatin	1	numeric	Proportion of 127 cell types in state "Het" (maximum)	ChromHMM (Roadmap)	CADD
Flanking bivalent TSS/Enh	1	numeric	Proportion of 127 cell types in state "BivFlnk" (maximum)	ChromHMM (Roadmap)	CADD
Flanking active TSS	1	numeric	Proportion of 127 cell types in state "TssAFlnk" (maximum)	ChromHMM (Roadmap)	CADD
Genic enhancer	1	numeric	Proportion of 127 cell types in state "EnhG" (maximum)	ChromHMM (Roadmap)	CADD
ZNF genes/Repeats	1	numeric	Proportion of 127 cell types in state "ZNF/Rpts" (maximum)	ChromHMM (Roadmap)	CADD
Bivalent enhancer	1	numeric	Proportion of 127 cell types in state "EnhBiv" (maximum)	ChromHMM (Roadmap)	CADD
Repressed polycomb	1	numeric	Proportion of 127 cell types in state "ReprPC" (maximum)	ChromHMM (Roadmap)	CADD
Flanking transcription	1	numeric	Proportion of 127 cell types in state "TxFlnk" (maximum)	ChromHMM (Roadmap)	CADD
Weak transcription	1	numeric	Proportion of 127 cell types in state "TxWk" (maximum)	ChromHMM (Roadmap)	CADD
Quiescent/Low	1	numeric	Proportion of 127 cell types in state "Quies" (maximum)	ChromHMM (Roadmap)	CADD
Enhancer	1	numeric	Proportion of 127 cell types in state "Enh" (maximum)	ChromHMM (Roadmap)	CADD
Weak repressed polycomb	1	numeric	Proportion of 127 cell types in state "ReprPCWk" (maximum)	ChromHMM (Roadmap)	CADD
Active promoter	1	integer	Number of rare SNVs in active promoter state in NA12878	ChromHMM (Encode)	ENCODE
Transcription	1	integer	Number of rare SNVs in transcriptional transition and transcriptional elongation states in NA12878	ChromHMM (Encode)	ENCODE
Poised promoter	1	integer	Number of rare SNVs in inactive/poised promoter state in NA12878	ChromHMM (Encode)	ENCODE
Low	1	integer	Number of rare SNVs in heterochromatin/low signal state in NA12878	ChromHMM (Encode)	ENCODE
Weak promoter	1	integer	Number of rare SNVs in weak promoter state in NA12878	ChromHMM (Encode)	ENCODE
Weak transcription	1	integer	Number of rare SNVs in weakly transcribed region state in NA12878	ChromHMM (Encode)	ENCODE

Insulator	1	integer	Number of rare SNVs in insulator state in NA12878	ChromHMM (Encode)	ENCODE
Weak enhancer	1	integer	Number of rare SNVs in two weak/poised enhancer states in NA12878	ChromHMM (Encode)	ENCODE
Repressed	1	integer	Number of rare SNVs in two repetitive/copy number variation states in NA12878	ChromHMM (Encode)	ENCODE
Strong enhancer	1	integer	Number of rare SNVs in two strong enhancer states in NA12878	ChromHMM (Encode)	ENCODE
PolII	2	numeric	Maximum PRHED-scale *P*-value/peak signal of polII evidence for open chromatin (across rare SNVs)	Open chromatin/TFBS	CADD
TFBS	1	integer	Number of overlapping TFBS from ChIP-seq (maximum across rare SNVs)	Open chromatin/TFBS	CADD
TFBSPeaks	1	Integer	Number of overlapping TFBS peaks from ChIP-seq summed over different cell types/tissue (maximum across rare SNVs)	Open chromatin/TFBS	CADD
TFBSPeaksMax	1	Integer	Number of maximum values of overlapping ChIP TFBS peaks across cell types/tissue (maximum across rare SNVs)	Open chromatin/TFBS	CADD
Combined pvalue	1	numeric	Maximum ENCODE combined PHRED-scale *P*-value of Faire, Dnase, polII, CTCF, Myc evidence for open chromatin	Open chromatin/TFBS	CADD
Dnase	2	numeric	Maximum PHRED-scale *P*-value/peak signal of Dnase evidence for open chromatin	Open chromatin/TFBS	CADD
Faire	2	numeric	Maximum PHRED-scale *P*-value/peak signal of Faire evidence for open chromatin	Open chromatin/TFBS	CADD
Myc	2	numeric	Maximum PHRED-scale *P*-value/peak signal of Myc evidence for open chromatin	Open chromatin/TFBS	CADD
H3K4Me3	1	numeric	Maximum ENCODE H3K4 trimethylation level	Open chromatin/TFBS	CADD
CTCF	2	numeric	Maximum PHRED-scale *P*-value/peak of CTCF evidence for open chromatin	Open chromatin/TFBS	CADD
H3K27Ac	1	numeric	Maximum ENCODE H3K27 acetylation level	Open chromatin/TFBS	CADD
Nucleosome position	1	numeric	Maximum ENCODE Nucelosome position track score	Open chromatin/TFBS	CADD
H3K4Me1	1	numeric	Maximum ENCODE H3K4 methylation level	Open chromatin/TFBS	CADD
CADD	1	numeric	Maximum PHRED-scale CADD score	DNA/Summary	CADD
CpG	1	numeric	Maximum percent CpG in a window of +/- 75bp	DNA/Summary	CADD
GC	1	numeric	Maximum percent GC in a window of +/- 75bp	DNA/Summary	CADD
DANN	1	numeric	Maximum DANN score	DNA/Summary	DANN
Distance to TSS	1	integer	Absolute distance (in bp) between the TSS and the closest rare SNV	DNA/Summary	Gencode
fitCons	1	numeric	Maximum fitCons score	DNA/Summary	CADD
DNA MGW		numeric	Maximum predicted local DNA structure effect on dnaMGW	DNA/Summary	CADD
Variant counts	1	numeric	Total number of rare SNVs	DNA/Summary	GTEx
DNA Roll	1	numeric	Maximum predicted local DNA structure effect on dnaRoll	DNA/Summary	CADD
DNA HelT	1	numeric	Maximum predicted local DNA structure effect on dnaHelT	DNA/Summary	CADD
DNA ProT	1	numeric	Maximum predicted local DNA structure effect on dnaProT	DNA/Summary	CADD
PhyloP	3	numeric	Maximum primate, mammalian, and vertebrate PhyloP conservation scores	Conservation	CADD
PhastCons	3	numeric	Maximum primate, mammalian, and vertebrate PhastCons conservation scores	Conservation	CADD
GerpS	1	numeric	Maximum rejected substitution score defined by GERP++	Conservation	CADD
GerpN	1	numeric	Maximum neutral evolution score defined by GERP++	Conservation	CADD

Extended Data Table 3. Genomic annotations used for RIVER. All annotations are across rare SNVs within 10 kb of the gene's TSS.

Genomic feature	Log odds ratio of an outlier status from one individual	*P*-value
DANN	2.79	5.7×10^{-30}
CADD	2.78	1.2×10^{-29}
Logistic	2.72	2.2×10^{-27}
Vertebrate PhyloP	2.74	1.5×10^{-28}
TFBS	2.77	1.2×10^{-29}

Extended Data Table 4. Assessment of the advantage of incorporating gene expression with genomic annotations by simplified, supervised models of outlier status.

Gene	Variant ID	P(FR|G)	P(FR|G,E)	Median Z-score	Clinical significance	Disease	Molecular Consequence
SBDS	rs113993991*†	0.447	0.985	-5.337	pathogenic	Shwachman syndrome	nonsense
TPP1	rs119455955*	0.619	0.995	-4.110	pathogenic	Ceroid lipofuscinosis neuronal 2, Neuronal ceroid lipofuscinosis, Inborn genetic diseases	nonsense
GAMT	rs80338735*†	0.162	0.929	-2.813	pathogenic	Deficiency of guanidinoacetate methyltransferase	synonymous
SBDS	rs113993993*†	0.526	0.989	-2.753	pathogenic, risk factor	Shwachman syndrome, susceptibility to aplastic anemia	splice donor
OGG1	rs104893751	0.213	0.963	-2.733	pathogenic	Clear cell carcinoma of kidney	missense
BBS2	rs121908176*	0.519	0.992	-2.560	pathogenic	Bardet-Biedl syndrome 2	nonsense
SBDS	rs113993993*†	0.520	0.988	-2.301	pathogenic, risk factor	Shwachman syndrome, susceptibility to aplastic anemia	splice donor
NAGA	rs121434529	0.047	0.563	-1.663	pathogenic	Schindler disease, type 1	missense
OGG1	rs104893751	0.213	0.239	-1.231	pathogenic	Clear cell carcinoma of kidney	missense
SLC25A11	rs140547520	0.009	0.004	-0.700	pathogenic	Amyotrophic lateral sclerosis 18	missense
DSTYK	rs200780796	0.077	0.049	-0.694	risk factor	Susceptibility to congenital anomalies of the kidney and urinary tract 1	missense
CLPTM1	rs120074114	0.027	0.006	-0.660	pathogenic	Apolipoprotein c-ii variant	missense
MUTYH	rs34612342	0.078	0.038	0.650	pathogenic	Endometrial carcinoma, MYH-associated polyposis, Carcinoma of colon, Hereditary cancer-predisposing syndrome	missense
IVD	rs28940889	0.074	0.045	0.573	pathogenic	Isovaleryl-CoA dehydrogenase deficiency	missense
GPR97	rs121908464	0.025	0.009	0.508	pathogenic	Bilateral frontoparietal polymicrogyria	missense
ZNF200	rs61732874	0.017	0.003	-0.431	pathogenic, likely pathogenic	Familial Mediterranean fever	missense, 3' UTR
APOC4	rs120074114	0.038	0.012	0.411	pathogenic	Apolipoprotein c-ii variant	missense
SLC7A9	rs79389353	0.044	0.014	-0.375	pathogenic, likely pathogenic	Cystinuria	missense
RPL29	rs121912698	0.023	0.008	-0.371	pathogenic	Aminoacylase 1 deficiency	missense
RPS19	rs147508369	0.018	0.013	0.304	pathogenic	Diamond-Blackfan anemia 1	missense
ABHD14B	rs121912698	0.035	0.011	0.224	pathogenic	Aminoacylase 1 deficiency	missense
ZNF200	rs104895091	0.022	0.005	0.218	pathogenic	Autosomal dominant familial Mediterranean fever	Inframe, 3' UTR
ABHD14B	rs121912701	0.020	0.004	0.206	pathogenic	Aminoacylase 1 deficiency	missense
ZNF200	rs28940579	0.025	0.006	0.175	pathogenic	Familial Mediterranean fever	missense, 3' UTR
RPL29	rs121912698	0.036	0.012	0.153	pathogenic	Aminoacylase 1 deficiency	missense
RPL29	rs121912701	0.021	0.005	0.142	pathogenic	Aminoacylase 1 deficiency	missense
ABHD14B	rs121912698	0.035	0.011	0.025	pathogenic	Aminoacylase 1 deficiency	missense

* Regulatory pathogenic variant.
† Mentioned in the main text.

Extended Data Table 5. 27 GTEx rare SNVs reported as disease variants in ClinVar.

Parameter	Initialization	Spearman ρ	Accuracy
β	10% noise	> .999	0.880
	25% noise	> .999	0.862
	50% noise	> .999	0.849
	100% noise	> .999	0.848
	200% noise	> .999	0.843
	400% noise	> .999	0.846
	800% noise	> .999	0.846
θ [P(E = 0|FR = 1), P(E = 1|FR = 1)]	[0.1, 0.9]	> .999	0.841
	[0.4, 0.6]	> .999	1.000
	[0.45, 0.55]	> .999	1.000

Extended Data Table 6. Stability analysis of estimated parameters with different parameter initializations.