www.sciencemag.org/content/348/6235/648/suppl/DC1

Supplementary Materials for

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*

*Corresponding author: Kristin G. Ardlie (kardlie@broadinstitute.org) or Emmanouil T. Dermitzakis (emmanouil.dermitzakis@unige.ch)

Published 8 May 2015, *Science* **348**, 648 (2015)
DOI: 10.1126/science.1262110

This PDF file includes:

Materials and Methods

Box S1

Figs. S1 to S34

Tables S1 to S15

References

Materials and Methods

List of Figures

List of Tables

Materials and Methods

S1 Biospecimen Collection and Processing

S1.1 Biospecimen Collection

The GTEx pilot study collected a total of 9365 tissue samples targeting more than 30 distinct tissues from 237 post-mortem donors. Briefly, two adjacent aliquots were prepared from each sampled tissue and preserved in PAXgene® tissue kits (PreAnalytiX®). Samples were fixed for a minimum of 2-4 hours, and then placed in the stabilizer buffer for shipment. Samples were then shipped to the GTEx Comprehensive Biospecimen Resource (CBR). One of each sample pair was then shipped, still in stabilizer buffer, to the GTEx Laboratory Data Analysis and Coordinating Center (LDACC) at the Broad Institute for processing, molecular analyses, transfer to cryovials, and long-term storage at -80°C. The remaining paired sample was retained at the CBR where it was embedded in paraffin (PFPE) to enable histopathological review of each tissue by the Pathology Resource Center (PRC). Brains were collected for only a subset of donors when conditions (donors could not have been on a ventilator for the 24 hours prior to death) and consent allowed. When available, brains were immediately removed from the body and placed on wet ice. Two areas of the brain (cortex and cerebellum) were sampled at the same time as the other tissues in PAXgene® kits as described above. The brain was then shipped on wet ice to the Brain Bank at the University of Miami. Up to 11 regions of brain were then sampled by the brain bank upon receipt, including a second sampling of each of the cortex and cerebellum regions previously sampled in PAXgene®, providing duplicate samplings of those two regions. All tissues sampled by the brain bank were placed in to cryovials and flash frozen in Liquid N2. These samples were shipped approximately monthly to the LDACC at the Broad institute for processing and analysis. In addition, whole blood preserved in ACD and PAXgene® blood tubes, as well as fresh skin samples were collected from each donor for DNA genotyping, RNA expression, and culturing of lymphocyte and fibroblast cell lines, respectively. These samples were shipped directly overnight to the LDACC for immediate processing.

Complete descriptions of the donor enrollment and consent process, as well as biospecimen procurement methods, sample fixation, and histopathological review procedures are being described elsewhere. In brief, a robust quality management program was established and implemented for the collection and handling of GTEx specimens. This included establishing methodology for data management, Standard Operating Procedure (SOP) development, and auditing of the program. A total of four biospecimen source sites (BSS's) were engaged in collecting the samples during the pilot phase of the project. While a single SOP for biospecimen procurement was difficult to implement at all sites, due to institutional differences, the BSS's were consulted when new SOPs were developed to ensure they would be operationally feasible at all sites. Document control software was used to ensure all sites used current versions of SOPs, and training was conducted prior to implementation of all new procedures. Supporting quality documents were developed to provide consistency and clarity to the program, and many of those documents, such as the SOPs used, and workflows for the project, are available to the public (http://biospecimens.cancer.gov/resources/sops/default.asp).

In addition to the postmortem donors, a small number (13) of surgical donors were enrolled during the pilot. Blood samples and up to 5 tissues [adipose, subcutaneous; muscle, skeletal; nerve, tibial; skin, artery, tibial] were collected, in the manner described above, to provide a reference comparison for the postmortem tissues. Tissue samples were processed and RNA selected for sequencing in the same manner as the postmortem samples (see below).

S1.2 Molecular Analyte Extraction and QC

S1.2.1 DNA isolation from blood

DNA was isolated from whole blood using the Qiagen Gentra Puregene method (QIAGEN) following the manufacturers specifications. Briefly, blood samples were placed in an Input tube and RBC lysis solution was added. The samples were inverted to mix, and incubated for 5 minutes at room temperature. Samples were inverted to mix again, and then centrifuged at 2000 RCF for 5 mins. Supernatants were discarded and the white blood cell pellets were vortexed vigorously for 5 seconds. Cell Lysis Solution, containing RNase and Proteinase K, was then added to the cell pellets. Samples were vortexed vigorously for 20 seconds then incubated at 55^{o}C for 2hrs or overnight if needed. When no solid precipitate remained, samples were cooled to room temperature and Protein Precipitation Solution was added. Samples were vortexed vigorously for 20 seconds, then centrifuged at 2000 RCF for 10 minutes to precipitate proteins. The supernatants were poured into a new Output tube containing Isopropanol, and inverted 50 times to precipitate DNA then centrifuged at 2000 RCF for 3 minutes to pellet the DNA. The supernatants were discarded and pellets washed with 70% Ethanol. Samples were recentrifuged at 2000 RCF for 1 minute. Supernatants were discarded, and pellets were allowed to air dry for 5-10 minutes before hydration with 1XTE. After incubation at 65°C for 1 hour, viscosity was assessed and additional TE added as needed. Samples were then stored at 4°C to undergo QC. For long term storage, all samples were stored at -20°C. DNA samples were quantified in triplicate using Picogreen (ThermoScientific Varioskan Flash instrument). DNA samples were further qualified by agarose gel electrophoresis (Invitrogen 1% Agarose E-gel) to assess the molecular weight distribution of the material.

The identity of all DNA samples was assessed by genotyping with a multiplexed panel of 96 SNPs (Fluidigm 96.96 Array) designed to include multiple proxy SNPs overlapping common Affymetrix (Affymetrix, Santa Clara, CA) and Illumina (Illumina, San Diego, CA) genotype arrays, SNPs in key common housekeeping genes, and a gender specific SNP. Genotypes were detected on the BioMark HD System (Fluidigm, San Francisco, CA).

S1.2.2 RNA isolation from PAXgene® preserved blood

Total RNA was isolated from whole blood samples collected and preserved in PAXgene® blood RNA tubes following the manufacturers protocol (Qiagen). Blood samples were centrifuged at 4000g for 10 minutes and the supernatants discarded. Nuclease-free water was added to the pellets and vortexed to resupend, then centrifuged for an additional 10 minutes at 4000g. The supernatants were discarded. Buffer BR1 was added to the pellets and vortexed to dissolve the pellet. Buffer BR2 was the added and samples were vortexed briefly again. Samples were transferred to 1.5ml microcentrifuge tubes. Proteinase K was added and samples were vortexed for 5 seconds, then incubated at 55°C for 10 minutes in a Thermomixer at 1400rpm. Following incubation, lysates were pipetted directly onto PAXgene Shredder spin columns, and spun for 4 minutes at maximum speed. Eluted samples were transferred to fresh 1.5ml tubes and 100% ethanol was added, vortexed briefly and then spun for 1-2 seconds at 500g. Samples were then pipetted into PAXgene spin columns and centrifuged at maximum speed for 2.5 minutes. Flow through was discarded. Buffer BR3 was added to the spin columns and spun for 2.5 minutes at maximum speed; flow through was discarded. DNAse I was added directly to the PAXgene column membranes and incubated at room temperature for 15 minutes. A second aliquot of buffer BR3 was then added to each spin column and centrifuged for 2.5 minutes at max speed; flow through was discarded. This was repeated with two more aliquots of buffer BR4. Buffer BR5 was added directly to column membranes and spun at maximum speed for 2.5 minutes to elute the RNA. RNA was incubated at 65°C for 5 minutes to denature and the cooled on ice. Two aliquots were removed immediately for QC and samples were then stored at -80°C. RNA samples were quantified and purity was assessed using the Nanodrop 8000 spectrophotometer (ThermoScientific). RNA quality was further assessed using an Agilent 2100 Bioanalyzer to obtain a RNA Integrity Number (RIN) score.

S1.2.3 RNA isolation from Cell Pellets

For all successfully transformed lymphocyte and fibroblast cultured cells, RNA was isolated from frozen cell pellets containing 1-10 x 10^6 cells using Trizol (Invitrogen). Briefly, Trizol was added to each pellet and the pellet was resuspended by pipetting. The lysate suspension was transferred to a 1.5 ml microcentrifuge tube and incubated at room temperature for 5 minutes. Chloroform was then added to each sample and mixed by pipetting. Samples were centrifuged at 13,000 rpm for 10 minutes at 4°C. The aqueous (upper) phase was then carefully transferred to a fresh 1.5 ml tube and isopropanol was added. Samples were mixed and incubated at room temperature for 10 minutes, followed by centrifugation at 13,000 rpm for 10 minutes at 4°C. Supernatants were discarded. The RNA pellet was washed with 75% ethanol and centrifuged at 5,000 rpm for 5 minutes at 4°C. Ethanol was removed carefully and RNA pellets were allowed to dry for 10 minutes (room temperature). RNA pellets were then rehydrated in DEPC treated sterile water. RNA's were then incubated for 5 minutes at 65°C to denature and then cooled on ice. Two aliquots were removed immediately for QC and samples were then stored at -80°C. RNA samples were quantified and purity was assessed using the Nanodrop 8000 spectrophotometer (ThermoScientific). RNA quality was further assessed using an Agilent 2100 Bioanalyzer to obtain a RIN score.

S1.2.4 RNA isolation from PAXgene® fixed tissues

Total RNA was isolated from PAXgene® fixed tissue samples using the PAXgene® Tissue miRNA Kit from PreAnalytix® (Qiagen) following the manufacturers specifications. Samples were isolated using a manual protocol in batches of 12, which included both a range of tissue types and donors in each batch to minimize batch effects. Briefly, samples were removed from storage at -80°C, and placed on dry ice for further manipulations. For most tissue types, 10-12 mg of tissue was cut from each tissue as input material. Exceptions included breast and adipose tissues for which 20mg of input material was used, and spleen and pancreas for which no more than10mg and 5mg respectively were used. Cut samples were placed in labeled cryovials with 250µl of Buffer TM1 prepared with β-ME and a stainless steel bead, and homogenized using a TissueLyser II for 2 minutes at 20Hz. Inner and outer tubes were then swapped to ensure even coverage, and re-homogenized for an additional 2 minutes. After removal of the steel beads with a magnet, 480µl of RNase-free water and 20µl of proteinase K were added to each sample and vortexed to mix. Samples were then incubated for 15 minutes at 45°C using a shaker-incubator at 1400rpm. Tissue lysates were then centrifuged for 3 minutes at maximum speed (not exceeding 20,000xg) and the supernatant fraction was transferred to a new tube. 1100µl of isopropanol was added to each tube and mixed, then samples were loaded on to PAXgene® RNA MinElute spin columns and centrifuged for 1 min at 8000xg (discarding flow through). Columns were then washed with 350µl of Buffer TM2 and centrifuged for 20 seconds at 8000xg (discarding flow through). Samples were DNase 1 treated by adding 80µl of dilute DNase 1 on to each column, incubating at room temp for 15 minutes then spinning for 20 seconds at 8000xg (retaining flow through in a new tube). 350µl of Buffer TM2 was then added to the flow through, and mixed, then pipetted on to the spin column and centrifuged for 20 seconds at 8000xg (discarding flow through). 500µl of Buffer TM3 was pipetted on to the spin columns and centrifuged for 20 seconds at 8000xg, followed by 500ul 80% ethanol, centrifuged for 2 minutes at 8000xg (discarding flow through both times). After drying the spin column for 5 minutes, it was placed in to clean elution tube, and 20µl Buffer TM4 was added to the column and centrifuged for 1 minute at maximum speed to elute the RNA. This elution step was performed twice to increase yield. RNA's were then incubated for 5 minutes at 65°C to denature and then cooled on ice. As above, two aliquots were removed immediately for QC and samples were then stored at -80°C. RNA samples were quantified and purity was assessed using the Nanodrop 8000 spectrophotometer (ThermoScientific). RNA quality was further assessed using an Agilent 2100 Bioanalyzer to obtain a RIN score.

S1.2.5 RNA isolation from frozen tissues

Total RNA was isolated from the GTEx Brain tissues frozen in Liquid N2, using the miRNeasy Mini Kit (Qiagen) following the manufacturers specifications. Samples were isolated using a manual protocol in batches of 12 - 24, which included both a range of brain subtypes and donors in each batch to minimize batch effects. Briefly, samples were removed from storage at -80°C, and placed on dry ice for further manipulations. From each tissue 25mg was cut for use as input material. Cut samples were placed in labeled cryovials with 700µl of Qiazol Lysis reagent and a stainless steel bead, and homogenized using a TissueLyser II for 2.5 minutes at 25Hz. Inner and outer tubes were then swapped to ensure even coverage, and re-homogenized for an additional 2.5 minutes. After removal of the steel beads with a magnet, samples were incubated at room temperature for 5 minutes and 140µl of chloroform was added to each sample and mixed. After incubation at room temperature for 2-3 minutes, samples were then centrifuged for 15 minutes at 12,000xg at 4°C. Following centrifugation, the upper aqueous phase was then transferred to a new tube. 525µl of 100% ethanol was then added and mixed, and samples were then pipetted on to RNeasy mini spin columns and centrifuged for 15 seconds at 8000xg (discarding flow through). Columns were then washed with 350µl Buffer RWT and centrifuged for 15 seconds at 8000xg (discarding flow through). Samples were DNase 1 treated by adding 80µl of dilute DNase 1 on to each column, incubating at room temp for 15 minutes then washed with 350µl Buffer RWT and centrifuged for 15 seconds at 8000xg (discarding flow through). Samples were then washed with 700µl Buffer RWT, followed by 500µl Buffer RPE, each time centrifuging for 15 seconds at 8000xg (discarding flow through). After a final wash with 500µl Buffer RPE, followed by centrifuging for 2 minutes at 8000xg (discarding flow through), the column was dried. Total RNA was then eluted by pipetting 40µl of RNase-free water on to the spin column and centrifuging at 80000xg for 1 minute. RNA samples were then quantified and QC'd as described for PAXgene tissue samples above.

S2 Genotyping and Imputation

S2.1 Genotyping Arrays

DNA isolated from the blood samples collected for each donor was the primary source of DNA used for genotyping. In a few instances where blood was not available, however, or if the blood DNA sample failed QC metrics, then DNA isolated from one of the other PAXgene tissue samples was substituted. Genomic DNA samples, including HAPMAP controls and several duplicates, were genotyped on two separate arrays, Illumina's Human Omni5-Quad (>360 ng DNA) and Infinium ExomeChip (>200 ng DNA), to maximize genotype density and allele frequency spectrum. The Human Omni5-Quad provides content on a 4-sample Infinium array of 4,301,331 fixed genome-wide markers that can interrogate genetic variation as low as 1% minor allele frequency (MAF). A total of 191 GTEx samples that correspond to 185 unique individuals were genotyped during the pilot phase by the Broad Institute's Genetic Analysis Platform using the HumanOmni5-Quad Array (Table S3). One HapMap individual (NA12878) was genotyped for positive control. Genotypes were called using Illumina's GeneTrain calling algorithm (Autocall).

The Infinium ExomeChip provides content on a 12-sample Infinium array of a total of >250,000 functional exonic markers, enabling high-throughput and robust genotypes. A total of 190 GTEx samples (corresponding to 184 unique individuals) and 1 HapMap individual (NA12878) were genotyped on this array during the pilot phase (Table S4), by the Broad Institute's Genetic Analysis Platform (1 GTEx sample failed: Collaborator Sample ID, GTEx-N7MT-0009). Genotypes were called using the Illumina Autocall algorithm and the rare variant genotype caller, zCall (*50*). On average our genotyping call rates exceeded 98% on both genotyping platforms.

To confirm sample identity of both the DNA genotyping arrays and the tissue samples, a set of 96 markers, including a gender confirmation assay was genotyped on the Fluidigm fingerprint panel using 25 ng of input DNA. Concordance of these 96 markers with both the genotyping arrays and RNA-Seq data from the different individuals and tissues was verified.

S2.2 Quality control

S2.2.1 Human Omni5-Quad array

Quality control (QC) steps were performed using the software PLINK (*51*). The QC steps and the number of samples and SNPs removed, flagged or retained at each QC step are summarized in Table S3. SNPs that mapped to more than one genomic location were removed from the analysis. To identify poor DNA quality or sample contamination, a heterozygosity test was applied to the autosomal chromosomes using the inbreeding coefficient. To further identify sample contamination, as well as cryptic relationships and sample duplicates, identity-by-descent (IBD) was computed between all pairwise sample combinations using genome-wide genotype data. SNP missing rate was predicted using the plink function --test-mishap. Deviation from

Hardy-Weinberg equilibrium (HWE) *P*-values was assessed for all SNPs using the subset of European individuals. Association with chemistry plate ID was estimated using the plink function --loop-assoc. Sex check was done using SNP genotypes on X and Y chromosomes. Two individuals believed to have Klinefelter Syndrome, were identified genotypically as being XXY. Both were subsequently confirmed by Pathology review of the testis H&E slides. All 8 flagged samples, including the two Klinefelter Syndrome individuals, were removed from downstream eQTL and expression analyses in our study (although these data were of otherwise good quality and hence were included in the release of data to dbGAP).

S2.2.2 Infinium ExomeChip

Similar QC steps were performed on the Exome Chip as on the OMNI 5M array, using different cutoffs (Table S4). In particular, the initial SNP Autocall call rate cutoff on the Exome Chip was < 80% and after multiple QC steps, a more stringent SNPs call rate of < 99% (instead of 95% as on the OMNI 5M array) was applied.

S2.3 Population Structure

We used Principal Component Analysis (PCA) implemented in EIGENSTRAT (*52*) on the QC'd GTEx Illumina 5M genotype data, to verify that the genotypes of the GTEx individuals cluster as expected based on their ancestral backgrounds. PCA was first applied to the 185 GTEx and 1 HapMap sample in aggregate with genotype data from 455 HapMap2 samples (183 CEU, 47 CHB, 42 JPT, 93 TSI, 90 YRI) on Illumina's Omni5-Quad. In this analysis, we wanted to check that GTEx samples and HapMap samples cluster well with respect to their ancestral background using the same genotyping platform (Fig. S5A). To better understand the genetic background of GTEx samples, we utilized the HapMap3 samples that span a wider range of ancestral backgrounds. These HapMap samples, which include 1184 samples from 11 regions worldwide, served as our reference populations. Plots of the two main principal components (PCs) for each dataset are shown in Figure S5. The PCA was also used to identify samples that are outliers in genetic ancestry, and to test whether the geographical origin estimated from genotypes is consistent with the self-reported ancestry. Most GTEx individuals cluster amongst the European populations and a small fraction lies along the African or Asian PC axes (Fig. S5B). For eQTL analysis (see below), the first three principal components generated using HapMap3 as the reference panel with the GTEx samples, were used as covariates.

S2.4 Imputation

To increase power and coverage for eQTL discovery in the GTEx project, and to facilitate compatibility with other studies and assist with functional fine-mapping, we imputed autosomal variants from the 1000 Genomes Project genotypes into our 185 GTEx individuals' 5M genotype data. The imputation was performed using IMPUTE2 (*53,54*) and the 1,000

Genomes Phase 1 freeze (an updated version from 19 April 2012, release v3), as the reference panel, which was downloaded from the IMPUTE2 website**:** **http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference.**

Before imputation, we removed low-quality variants and duplicate samples identified in the 5M array QC steps, to ensure accuracy of the final results. SNPs were aligned to the 1000 Genomes Phase 1 freeze by chromosome position (build 37). We also applied additional QC steps, including removing indels, SNPs not present in the 1000 Genomes dataset, alleles incompatible between the array and reference panel, and SNPs mapped to more than one chromosome position in the reference dataset. All 185 GTEx samples (see Table S3) were phased together, imputed by segments on each chromosome and subsequently merged. VCF files were generated with genotype likelihoods (posterior probabilities) for each of the three possible genotypes. The distribution of SNPs by allele frequency (AF) and imputation quality score (INFO) is shown in Figure S6.

For eQTL analysis, following imputation, we filtered out the following SNPs: missing rate cutoff <95% for best-guessed genotypes at posterior probability >0.9, Hardy-Weinberg Equilibrium $p<1\text{x}10^{-6}$ (using the software tool SNPTEST (*53*), imputation confidence score, INFO<0.4, and minor allele frequency, MAF<5%. This yielded 6,820,471 genotyped and imputed autosomal SNPs. We also generated a VCF file with dosages of alternative allele counts used as input for the Matrix eQTL software package (*21*)**.**

S2.4.1 Evaluation of imputation accuracy

To evaluate the imputation accuracy we compared a set of imputed SNPs on OMNI 5M array across 184 GTEx individuals that had direct genotype calls on the Exome Chip with call rate >99% (90,157 SNPs) (Fig. S7). Four different accuracy metrics were tested as a function of the IMPUTE2 imputation confidence score, INFO and MAF that was computed from the direct SNP calls on Exome Chip: (i) Mean concordance, computed by calculating a correlation coefficient, r^2 of the imputed dosage of the minor allele per SNP across all individuals to their direct calls on the Exome Chip, and then averaging over the r^2 of all SNPs, (ii) Minor allele concordance rate, where concordance rate is defined as the fraction of 90,157 SNPs whose imputed genotype call at posterior probability > 0.9 is concordant with their direct call on Exome Chip, computed across all 184 individuals. If none of the three genotype groups per SNP had a posterior probability above 0.9 that SNP was not included in the analysis. (iii) Concordance rate of minor allele homozygotes, and (iv) Concordance rate of minor allele heterozygotes (Fig. S7). The accuracy estimates (mean concordance) of imputed GTEx data were similar to those in the 1000 Genomes Phase 1 paper (*55*). For common variants, the mean minor allele dosage correlations ranged from ~93% to 96%, and ~90 to 95% in the 1000 Genomes paper in non-African ancestry; for low-frequency variants (1-5%), the accuracies ranged from ~72% to 89%, and ~60% to 90% in the 1000 Genomes paper in all populations (Table S5). The other three concordance measures (mean r^2) showed similar trends.

S3 Exome Sequencing

We performed whole exome sequencing on DNA samples from 180 GTEx pilot phase donors at the Broad Institute's Genomics Platform, using Agilent Sure-Select Human All Exon v2.0, 44Mb baited target, and the Broad's in-solution hybrid selection process. For input DNA we used >250 ng of DNA, at >2ng/ul. Our exome-sequencing pipeline included sample plating, library preparation (2-plexing of samples per hybridization), hybrid capture, sequencing (76bp paired reads), sample identification QC check, and data storage. Our hybrid selection libraries cover >80% of targets at 20x and a mean target coverage of >80x. The exome sequencing data was de-multiplexed and each sample's sequence data were aggregated into a single Picard BAM file.

Exome sequencing data was processed through a pipeline based on Picard (http://picard.sourceforge.net/), using base quality score recalibration and local realignment at known indels. We used the BWA aligner (http://bio-bwa.sourceforge.net) for mapping reads to the human genome build 37 (hg19). SNPs and indels were jointly called across all 180 samples using GATK's UnifiedGenotyper package version 2.6 (*56*) and GATK's HaplotypeCaller version 2.8 (http://www.broadinstitute.org/gatk/gatkdocs/), Default filters were applied to SNP and indel calls using the GATK's Variant Quality Score Recalibration (VQSR) approach. We used the SNP calls from GATK's UnifiedGenotyper and indel calls from GATK's HaplotypeCaller due to its superior performance in indel calling.

Functional annotation was performed using the Variant Effect Predictor (VEP v2.5) tool from Ensembl (http://useast.ensembl.org/info/docs/tools/vep/). We modified it to produce custom annotation tags and additional Protein Truncating Variant (PTV) annotations. The additional PTV annotation was applied to variants that were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and the variants were flagged if any filters failed. A PTV variant was predicted as high confidence (HC) if there is one transcript that passes all filters, otherwise it is predicted as low confidence (LC). The custom annotation tag is a comma separated ordered list of features corresponding to each of the transcripts in GENCODE version 12 that overlapped the variant. This modified version of VEP was applied to the 180 whole exome sequences.

S4 RNA Expression

S4.1 Sample Selection

Following processing and QC, total RNA samples were available from blood, from cell lines (LCL and Fibroblast), from PAXgene-preserved, and from Frozen tissue samples. All samples that met criteria of having a RIN value of 6.0 or higher and at least 1uG of total RNA, were batched for RNA sequencing. To the extent possible, based on sample availability, batches for library construction were designed to include a range of samples from different tissues and to span multiple donors, so as to minimise donor and tissue batch effects. Given the limited scope of the pilot phase of the project a set of 9 tissues (adipose (subcutaneous), tibial artery, heart (left ventricle), lung, muscle (skeletal), tibial nerve, skin (sun exposed), thyroid, and whole blood) were prioritized for sequencing from as many donors as possible. These tissues were selected based on abundance (they were routinely sampled and received), and that they generally tended to meet RNA QC criteria. For donors on whom brain tissue samples were available, we additionally included all available RNA samples that met QC criteria so as to survey a broader range of tissues across a smaller number of donors (Figs. S1A, S1B). One control sample (K-562) was included in library construction and sequencing with each batch of 95 samples, and a set of GTEx RNA samples were also selected to be run in duplicate (processing replicates) across separate sequencing runs.

S4.2 RNA Sequencing

S4.2.1 Library preparation and sequencing

RNA sequencing was performed using a standard non-strand specific protocol with poly-A selection of mRNA. Non-strand specific RNA sequencing was performed at the Broad Institute using a large-scale, automated variant of the Illumina Tru Seq™ RNA Sample Preparation protocol (Illumina: TruSeq Protocol Info). Briefly, 200 ng of total RNA was used from each sample as the starting material. This method uses oligo dT beads to select poly-A mRNA from the total RNA sample. The selected RNA is then heat fragmented and randomly primed before cDNA synthesis from the RNA template. The resultant cDNA then goes through Illumina library preparation (end repair, base ‘A’ addition, adapter ligation, and enrichment) using Broad designed indexed adapters for multiplexing of samples. After enrichment, the samples are qPCR quantified and equimolar pooled before proceeding to Illumina sequencing which was done on the Illumina HiSeq 2000, with a sequence coverage goal of 50M 76bp paired-end reads (Median achieved was ~82M total reads). The entire process occurs in a 96-well format and all samples were electronically tracked through the process in real-time including reagent lot numbers, specific automation used, time stamps for each process step, and automatic registration.

S4.2.2 RNA Sequence data pipeline and QC

RNA-seq data were aligned with Tophat version v1.4.1 (*57*) to the UCSC human genome release version hg19 (Genome Reference Consortium GRCh37). Gencode version 12 (*15*) was used as a transcriptome model for the alignment as well as all gene and isoform quantifications. Unaligned reads were merged back in to create a final bam. Gencode V12 annotates a total of 53,934 genes, including 20,110 protein coding genes, 11,790 long non-coding RNA's (lncRNAs), and 12,869 pseudogenes.

Expression levels were produced at the gene and exon level in RPKM units (*58*) using RNA-SeQC (*59*). Exon coordinates per gene were derived from the Gencode GTF using the following isoform collapsing procedure: exons labeled as 'retained_intron' were excluded; overlapping intervals were merged; intervals associated with multiple genes were discarded; and the final gene level model was produced in GTF format.

For gene and exon level read count and gene level RPKM values, reads were filtered based on the requirements: (1) reads must be uniquely mapped (for Tophat this is mapping quality equal to 255); (2) reads must have proper pairs; (3) alignment distance must be <=6; (4) reads must be contained 100% within exon boundaries. Reads overlapping introns are not counted. For exon read counts, if a read overlaps multiple exons, then a fractional value equal to the portion of the read contained within that exon is allotted.

RNA-seq expression samples were scrutinized using several quality control measures before being included in the final analysis set. Samples with fewer than 10 million mapped reads were removed. Additional outliers were identified using a correlation-based statistic and using sex incompatibility checks, following methods described in (*60*). First, within-tissue sample-sample correlations of expression levels were computed, and for each sample we defined the statistic D_i as the mean correlation of sample *i* with the remaining samples. For this purpose, expression levels were computed as the read counts normalized by library size. For all tissues except whole blood, samples with $D<0.9$ were removed. Examination of the *D* statistic for whole blood indicated heterogeneity such that no clear outliers could be identified on this basis. Sex incompatibility checks were based on *XIST*, and on a set of chromosome Y genes that showed significant ($P<0.05$) differences between labeled males and females in all tissues (with non-significant genes presumably not expressed in males for some tissues). Scatterplots of *XIST* expression versus the average expression of the significant Y genes showed no gender mismatches, but a very few outliers (fewer than 5 across all tissues) for which either *XIST* or chromosome Y expression appeared outlying, and these samples were removed. In the case of processing replicates (same sample sequenced twice), the samples with the greater number of reads were retained for inclusion in the final analysis set (although all replicates were included in the data released to dbGAP). Samples derived from the two individuals with Klinefelters Syndrome (which failed the sex-specific expression check) and from one individual with multiple tissues that were D statistic outliers were excluded. The final pilot analysis data set comprised 1,641 samples from across 43 tissues and 175 donors. This included 18 samples from 4 surgical donors (SSA3, TMZS, VUSH, WCDI) and 1,623 samples from 171 postmortem donors. RNA-seq QC showed no difference between the surgical and postmortem samples, nor did a separate analysis of ischemic time and gene expression (see Mele *et al.* (17)), hence all samples were analyzed together.

S4.3 Covariates

To remove global effects on gene expression, that might mask or skew the effects of local genetic variation, we calculated the top genotype principal components (PCs) across the 185 GTEx samples (combined with 1184 HapMap3 samples as a reference panel) (*52,61*), as described in ***2.3*** above. The top three PCs were chosen because they account for 10% of the variation explained with diminishing returns (0.5% or smaller contribution) for subsequent PCs. These PCs are sufficient to represent the major population structure found in the GTEx dataset consisting of Caucasian, African American and Asian individuals (Table S1, Fig. S5).

To find hidden batch effects and other cofounders in the expression data, we also employed the Probabilistic Estimation of Expression Residuals (PEER) method (*62*). We included 15 PEER factors which we found maximized our sensitivity in the eQTL discovery process, capturing ~47% of the total variance in gene expression. Gender was considered as an additional covariate. To gain insight into the biological meaning of these factors, we correlated them with 12 known sample/donor characteristics, such as ischemic time, gender, collection site, and other possible sources of experimental batch effect (Fig. S8). The relationship between PEER factors and known covariates is complex, with variation in any given covariate distributed among several PEER factors and vice versa. The top PEER component was significantly correlated with ischemic time ($q=0.005$), and related dependent factors such as collection center ($q<0.05$). (Fig. S8). The second peer factor was most strongly associated with the RNA sequencing batch ($R^2 = 0.8$). In general the correlations between PEER factors and known covariates were moderate to low, with the effect of the known covariates captured by multiple PEER factors.

S4.4 Transcript isoform quantification

Employing the Gencode reference annotation (version 12, http://www.gencodegenes.org), we used the Flux Capacitor (version 1.2.3, http://flux.sammeth.net) to quantify the expression of several transcriptional elements. All quantifications are based on read pairs that were mapped to the genome by Tophat version v1.4.1 (*57*), see above, and match within the constraints of the selected Gencode reference annotation. The Flux quantifications distinguish 3 transcriptional elements: (1) splice junctions (gtf-feature

“SJ”): all read-pairs compatible with the annotation are considered, and those aligning immediately up- and downstream of an annotated intron are considered to quantify the corresponding splice junction; (2) introns (gtf-feature “intron”): all read mappings of which one mate agrees with the reference annotation and of which the other mate falls in a region that is not overlapping with any Gencode exon are considered to quantify the retention of the corresponding intronic region; (3) transcripts (gtf-feature “transcript”): according to a previously described deconvolution strategy (*63*), all read-pairs that comply with the reference annotation are represented as a system of linear equations. The mapping distribution of non-alternatively spliced loci allows estimation of the impact of intrinsic experimental biases. Considering these biases, the error of the observed distribution is minimized when segregating reads in common exonic areas into the single isoforms annotated for a locus. Based on read counts obtained by the deconvolution the RPKM measurement is then computed (*58*).

S4.5 Transcriptome Analysis

S4.5.1 Gene expression clustering

We have explored the gene expression similarity between tissues and across all samples, by performing hierarchical clustering (HC) using different settings in R statistical environment. RPKM values were used in log2-transformed scale. Distance between samples being defined as *dist = 1 – correlation*. Pearson and Spearman correlations showed similar results. Pearson correlation was used as the correlation measure. Average linkage method was used for all the tested settings. All the genes from the annotation were considered.

S4.5.2 Exon splicing clustering

Exon inclusion levels were calculated for all the internal exons of genes with three or more exons. We calculated the ‘Percent Spliced-in’ (PSI) as in Barbosa-Morais *et al*. (*18*). The PSI measure for each exon is defined as the ratio between the reads that support the inclusion of the exon and the sum of the reads that support the inclusion and the exclusion of the exon. PSI values range between 0 and 1, where 1 represents full inclusion of the exon and 0 full exclusion. For an internal exon C and its neighbor exons A1 and A2, *Inc* corresponds to reads that support the junction A1-C and *Inc’* the junction C-A2. *Exc* reads support the junction A1-A2. The PSI formula is then defined as PSI = avg(Inc,Inc’) / (avg(Inc,Inc’) + Exc). Only exons supported by a sufficient number of reads, Inc + Inc’ + Exc >= 10, were considered.

Hierarchical Clustering (HC) was performed using the same settings as the HC performed in expression clustering. We selected 54330 exons with PSI values in more than 90% of the samples. The "na.or.complete" parameter was used to handle missing values. Differential exon inclusion between groups was tested using the Wilcoxon test in R, with p-values corrected by the BH method (*64*). Exons are considered differentially included if FDR < 1% and difference in median PSI between groups > 0.1

S5 eQTL analysis

S5.1 Single tissue *cis*-eQTL analysis

Nine tissues all having greater than 80 samples were chosen for eQTL analysis: adipose, artery, blood, heart, lung, muscle, nerve, skin and thyroid. The *cis* window was defined as 1 megabase up- and down-stream of the transcriptional start site (±1Mb), and we tested between 1.5 x 10^8 and 1.7 x 10^8 gene-*cis*-SNP pairs depending on the tissue type and genes expressed. Nominal p-values were generated for SNP-gene pairs using Matrix eQTL (*21*) in linear regression mode. We corrected for the following covariates: the first three genotyping principal components (PC's), the first 15 expression PEER factors (Probabilistic Estimation of Expression Residuals) (*62*), and gender. Expression matrices were derived from RPKM values at the gene level. For a given tissue, genes having at least 0.1 RPKM in 2 or more individuals were retained. Expression values were quantile normalized across genes within a tissue. Finally, the expression values for each gene were transformed into a standard normal based on rank (to minimize the effects of outliers on the regression scores).

To identify eQTL-containing genes (**eGenes**), a permutation procedure was applied, correcting for the multiple hypothesis effect of many SNPs in LD for a given gene. The minimal p-value per gene (**min(p)**) was used as the test statistic. Permutations were performed by randomizing sample labels for the expression data. The same random indexes were applied to the PEER factors and gender covariates. Genotypes and Genotyping PCs were not randomized. A minimum of 1000 permutations and a maximum of 10,000 permutations were performed, with an exit criteria in between this range of having at least 15 permuted min(p) values less than the nominal min(p). Having derived an empirical p value for each gene, q-values were calculated using the Storey approach (*65*) and a q-value threshold of <= 0.05 was applied.

In addition to a list of eGenes it is desirable to produce a list of all significantly associated SNP-gene pairs. We do so here by using a permutation threshold based approach similar to previous studies (*63,66*). Permutations were performed as above. A permutation threshold is chosen and this value is mapped back to the equivalent nominal min(p) value among the permutation values. Here we chose the permutation threshold to equal the empirical p value of the eGene at the 0.05 q-value threshold. This produces a nominal min(p) value threshold for each gene. SNPs with p-values below or equal to this threshold are included in the final SNP-gene eQTL list. To estimate how the number of eGenes detected per tissue-type is expected to grow with more data, we reran the analysis on successively down-sampled donor subsets from each tissue. Bonferroni correction was used for downsampling (line in Fig 2A) to reduce computational burden. All tissues showed an approximately linear relationship, with the majority sharing the same slope of ~21 eGenes/sample (shown in Fig. 2A). The permutation-based calculations (♦) of the final data are offset from the line end points (downsampling), and exhibit slightly larger numbers of eGenes, due to differences in the multiple hypothesis correction between these two methods used.

To investigate the sensitivity and validity of our study, in particular since we are using tissues from deceased donors, we compared the GTEx eQTLs that we discovered in blood to eQTLs from two separate studies of whole blood samples: the Westra *et. al.* study of 5,311 samples *(7)*, and a study of 911 blood samples taken from the Estonian Biobank, that were analyzed in house. In both studies the expression levels were measured using microarrays and ~30,800 genes that had unique probe mapping were tested. For our replication analyses, we removed microarray probes that measured expression of more than one gene (e.g., due to genes that physically overlap on the chromosome). Out of 1984 GTEx blood eGenes only 1202 were tested for eQTLs in the Westra *et. al.* (*7*) and Estonia Biobank studies.

The *cis*-eQTLs from the Westra *et. al.* study were download from: http://genenetwork.nl/bloodeqtlbrowser/. A window of 250 kb was used around the transcript start sties (TSS), and the FDR correction was done at the probe-level *(7)*. We used the Westra *et. al.* eQTLs at FDR<5% to assess the replication rate of eGenes discovered in GTEx at FDR<5%. We did not assess the replication rate of SNP-gene pair eQTLs due to only partial overlap in variants tested between the two studies (in GTEx the genotypes were imputed using the 1000 genomes project as a reference and in Westra *et. al.* (*7*) HapMap 2 was used as reference), and since we did not have complete knowledge of the fraction of variants in GTEx also tested in Westra *et. al*. after variant filtering (QC).

To evaluate the goodness of discovery of our GTEx *cis*-eQTL SNP-gene pairs, we used a study of 911 blood samples taken from the Estonian Biobank, for which we had access to both genotype and expression data, and hence could apply a similar eQTL analysis pipeline to that used in GTEx. This study was one of nine studies included in the Westra *et. al.* eQTL meta-analysis, however for our purposes we applied a modified eQTL pipeline to Estonian samples. *Cis*-eQTL mapping was performed on the 911 samples from unrelated living individuals, with gene expression data measured from whole peripheral blood (Illumina HT12v3), using a ±1Mb window around the genes' TSS. Genotypes were cleaned as described before (*7*), data was imputed using the 1000 Genomes Phase I cosmopolitan haplotypes (version March 2012) as a reference, and imputation dosage values were used as genotypes for the eQTL analysis. The gene expression normalization procedure and eQTL-mapping framework is extensively described in (*7*). In short, gene expression data were quantile normalized to the median distribution and were subsequently log2 transformed. Probe and sample means were centered to zero. Gene expression data were then corrected for possible population structure and adjusted for technical artifacts. After normalization of the gene expression data, we correlated genotype dosages of SNPs (Hardy-Weinberg p-value > 0.001, minor allele frequency > 5%) with gene expression values, where the distance between SNP and the midpoint of the probe was smaller than 1 megabase (according to hg19 (genome build 37)). We then permuted the sample labels and repeated this analysis 10 times, in order to obtain the *p*-value distribution used to control the FDR at 5%, both at the probe- (N=33,280) and gene-level (N=18,904) (described in *7*). The gene-level correction was used for replication analyses as it is more similar to the correction applied in GTEx.

The replication rate of the GTEx eGenes in the Estonian study at FDR<5% was 56% for eGenes and 18% for SNP-gene pair eQTLs. Given that the GTEx study is much smaller than the Estonia study, we also tested for consistency in allelic direction and found that 98% of GTEx-significant eQTLs at FDR<5% showed consistent allelic direction with the Estonia study eQTLs. For the consistency test, we verified that the eQTLs' effect directions in the two studies were measured relative to the same effect allele.

Finally, to evaluate the potential of the GTEx study, we also evaluated the extent to which the GTEx eQTLs replicated the significant SNP-gene pair eQTLs found in the larger Estonia study (at FDR<5%). For this we used the π_1 statistic that provides an estimate of the fraction of true positive eQTLs ($\pi_{1=}1$-π_0, where π_0=estimated fraction of null eQTLs, estimated from the full distribution of p-values) (see Storey and Tibshirani *q*-value approach (*26*)). The π_1 statistic considers also sub-threshold GTEx eQTL *p*-values below the FDR<5% cutoff. About 97% of SNPs with significant eQTLs in the Estonian study were tested in GTEx. Despite the 6-fold difference in sample size, the overall estimated replication rate of the Estonia SNP-gene pair eQTLs in GTEx was $\pi_{1=}0.51$, i.e. the GTEx eQTL study captured about half the eQTLs detected in the Estonia study (Fig. S11). Furthermore, ~78% of the Estonia significant eQTLs showed consistent allelic direction with GTEx eQTLs that is significantly more than would be expected by chance (Binomial test $P < 1*10^{-200}$). This demonstrates that a substantial fraction of sub-threshold GTEx eQTLs are likely true but do not reach significance here due to our small sample sizes.

S5.2 Multi-tissue eQTL analyses

For permutation-based multi-tissue cis-eQTL analysis, the average expression level was first subtracted from each gene in each of the nine tissues used for multi-tissue modeling, and then combined in a single dataset. The 19 covariates (15 PEER factors, 3 genotype PC's, and gender) were removed from the expression and genotype data via residualization by linear regression prior to permutation. Permutation was performed on sample labels of the underlying genotype data compared to expression, and thus all of the linkage disequilibrium structure in the genotype data was preserved, as well as cross tissue expression correlations. For each gene, genotypes were permuted 10,000 times and the test statistic computed for each SNP, recording the most significant result for each gene in each permutation. The final (gene-level) empirical p-value was then computed by comparing the statistic of the most significant SNP in the original data to these 10,000 values. This approach is essentially the same as performed for single-tissue analysis described above. However, the multi-tissue version enabled more powerful permutation testing for combined evidence of activity in at least one tissue by computing the most significant statistic for all nine tissues in each permutation. After obtaining per-gene empirical p-values for individual tissues and for the joint set of nine tissues, the p-values were subjected to false discovery rate control using a MATLAB v 7.9.0.529 implementation of the R *qvalue()* function, and genes with $q<0.05$ were declared significant. Following the approach of Nica *et al.* (*22*), pairwise comparisons of multiple tissues were performed by identifying significant genes within each tissue, and for each such gene calculating the π_0 value (estimated proportion of null hypotheses) for the gene in the second tissue, using the *qvalue()* implementation. The procedure is not symmetric, so for the nine tissues a total of 9X8=72 π_0 values were computed.

To more fully utilize the wide range of tissues represented, the Bayesian multi-tissue model of Flutre *et al.* (*24*) was fit to the data, modified to accommodate the large number of tissues (9) and specifics of our study design. The model explicitly considers, for each gene-SNP pair, the probability for each of the $2^9=512$ binary configuration vectors of eQTL activity. The hierarchical model then borrows information across genes, using maximum likelihood (empirical Bayes) fitting to the model to then compute the probability of each configuration. The support in the data for each configuration is then assessed using a Bayes Factor which borrows information across tissues in which the eQTL is active. For the GTEx design, expression is available for each tissue from only a subset of the individuals, which creates complications in handling the correlation patterns. In the multivariate linear regression model of Flutre *et al.* (*24*), the residuals are allowed to be correlated, thereby inducing correlations between the estimated effect sizes of the genotype in each tissue. In order to handle residual correlation for individuals with missing data, we applied a missing-at-random assumption, such that the residuals correlations are estimated from the individuals for which both tissues are represented.

To compute gene-level posterior probabilities, the Bayesian model requires an estimate of the gene-level null prior $\pi_{0,}$ i.e. the probability that the gene does not contain any local eQTL in any tissue. As the permutation approach of Flutre *et al.* (*24*) is too computationally intensive for this setting, we developed a simple procedure using the fact that, under the gene-level null and assuming only one truly causal SNP per gene, the expectation of the Bayes Factor is equal to 1. This gene-level controlling procedure can be shown to conservatively estimate π_0. With the estimates in hand, for each gene posterior probabilities and Bayes Factors were computed for the SNP achieving the lowest posterior for the null hypothesis at the level of the gene-SNP pair. The conservativeness of this procedure in estimating gene-level π_0 was confirmed by performing direct permutation analysis on a subset of 4 tissues.

Finally, a *streamlined* Bayesian approach was implemented, which is based on multivariate normal mixture modeling for multi-tissue eQTL effects. The model considers gene-SNP pairs as the unit of analysis, and directly uses *z*-statistics from Fisher's transformation of the expression-genotype correlation (corrected for covariates) as the input to the model (*25*). Correlation of effects due to tissue similarity or due to overlapping individuals is handled in a common covariance framework. Like the Bayesian model of Flutre *et al*. (*24*), the streamlined model considers the 2^9 configuration profiles, and provides estimated posterior probabilities for each of the configurations. By working directly with gene-SNP pairs, the null probability is expressed at the gene-SNP level, and testing for gene-SNP pairs can be performed using the local false-discovery rate.

S6 Allele specific expression analysis

For heterozygous sites, we retrieved the counts of the two alleles in RNA-seq data by Samtools and custom scripts, separately for each individual. We excluded sites with potential allelic mapping bias: 1) 50bp mapability < 1 in the UCSC mapability track, and 2) simulations of RNA-seq reads show >5% allelic mapping bias (*67,68*). We used only uniquely mapping reads (MAPQ 225), and required base quality >10. In order to confirm the heterozygous genotype, we included only sites with RNA-seq reference allele ratio [0.02, 0.98].

We used a simple binomial test to estimate whether the allele counts deviate significantly from the expected. For the expected allelic ratio of the binomial test, instead of using 0.5, we used the expected allele ratio for each individual, for each base combination. This was calculated by summing up reads across all sites with down-sampling of the highest covered sites so that they would not have disproportionally large effect on the ratios. These ratios, generally only a few percentage points from 0.5, correct for subtle genome-wide mapping bias as well as GC bias. For further discussion of the data processing steps, see Lappalainen *et al*. (*67*).

For all the sites with >=8 reads, we performed a binomial test of the REF/NONREF allele counts in each individual, using the expected ratio as above. The master files available in dbGap include all these data – these files contain genotype data of heterozygous sites, and can thus require authorized access. In most analyses, unless otherwise specified, we used only sites with >=30 reads to ensure sufficient power and replicability of allelic ratios (*67*). Furthermore, since different coverage of sites – which depends on gene expression levels and is highly tissue-specific – affects the power to detect significant ASE and the variance of the allelic ratios in a manner that is difficult to account for perfectly, in most analyses we downsampled the reads to exactly 30 to avoid any confounding effects of differing read coverage.

In ASE quality control analyses, we detected slightly lower DNA-RNA genotype concordance for 3 individuals (GTEX-N7MT, GTEX-NPJ8, GTEX-PLZ5). Because ASE analysis is much more sensitive to this, we removed all samples from these individuals from subsequent analyses. Furthermore, 19 samples had slightly lower RNAseq coverage, which leads to a smaller number of sites left for ASE analysis, and we removed these samples from ASE analysis as well. Altogether, the ASE results are based on 1563 remaining samples. Figure S18 and Table S8A,B show basic statistics of the data used in ASE analysis.

S7 Analysis of isoform and splice-QTLs

We have developed two independent methods to identify SNPs associated with alternative splicing - Altrans (*35*) and sQLTseekeR (*36*).

S7.1 Altrans

Altrans (*35*) utilizes the paired end nature of the RNA-seq experimental design. It uses the mate pairs, where one mate maps to one exon and the other mate to a different exon, and split reads to count "links" between two exons. The first exon in a link is referred to as the "primary exon". We used exons from protein coding and long non coding RNAs (lncRNA) genes in the GENCODE v12 annotation (*15*). Overlapping exons are grouped into "exon groups" and unique portions of each exon, where there is no other overlapping exon is present, are identified. These unique portions are subsequently used to assign RNA-seq reads to an exon and count links between exons. The raw link counts are used to calculate 15 peer factors which are then subsequently used to normalize the raw link counts with linear regression. The normalized link counts ascertained from unique regions of exons, which can be derived from parts of the linked exons rather than the whole exons, are divided by the probability of observing such a link given the empirically determined insert size distribution for each sample and unique portions of the exons in question, which is referred to as "link coverage". Finally the quantitative metric produced is the fraction of one link's coverage over the sum of the coverages of all the links that the primary exon makes, which is between 0 and 1 representing the proportion of a given link among all the links the primary exon makes. We calculated this metric in 5'-to-3' (forward) and 3'-to-5' (reverse) directions to capture splice acceptor and donor effects respectively. We only included primary exons that made ≥10 links in 40% of the samples originating from exon groups that made ≥ 15 links in 90% of the samples. We ran Altrans independently in each of the nine tissues using a *cis* region for the associations of one megabase (±1Mb) flanking the transcription start site, the same distance as used for the eQTL analyses. In order to identify sQTLs we ran a Spearman's rank correlation test with imputed genotype dosages that were corrected for population structure with the first 3 eigenvectors, and Altrans quantifications. The p-values attained were corrected for multiple testing using the Benjamini–Hochberg method (*64*) and final results have an FDR < 0.05. The software, manual, and more in depth description of the Altrans algorithm are available at http://sourceforge.net/p/altrans/wiki/Home/.

S7.2 sQTLseekeR

SQTLseekeR (*36*) identifies SNPs that are associated with changes in the relative abundances of a gene transcript isoforms, which we refer to as splicing ratios. This is a multivariate phenotype—with as many values as there are transcript isoforms annotated for a given gene. We used a non-parametrical approach inspired from MANOVA theory and introduced by Anderson (*69*) to detect association between a given SNP and a gene splicing ratios. We used the Hellinger distance to compute the variability of splicing ratios across samples We compare the variability within genotypes with the variability between genotypes to compute an F score using the Anderson's method. To estimate the significance of these scores and obtain P-values, we compute a null distribution from permutations for each gene, and compare it with the distribution of the true scores. We used the *Vegan* R package (*70*) to compute the true and permuted F scores. Finally, we correct the P-values for multiple-testing using *qvalue* R Package (*71*). The characteristics of this approach, more precisely the distance-based scores and permutation support, are critical here to robustly integrate genes with different splicing profiles (e.g. number of isoform) without the need for complex or reductive models. The software, manual, and more in depth description of sQTLseekeR are available at http:// http://big.crg.cat/bioinformatics_and_genomics/sQTLseekeR.

We used the isoform quantification by the Flux Capacitor on GENCODE v12 (see above). We considered, only genes expressing at least two isoforms (>=0.01 RPKM), and since we were searching for direct effects on splicing, we used a smaller *cis*-window and tested for association with a gene using only SNPs within the gene body plus 5Kb upstream and downstream (±5Kb) from the gene.

To characterize the alternative splicing event resulting from a variant effect, we first identify the two transcripts whose relative expression changes the most between the genotype groups. The exon structure of these two transcripts was then compared using the AStalavista software (*72*). AStalavista provide an extensive characterisation of the splicing event differentiating the two transcripts. We investigated the presence of some simple events in this comprehensive characterization. Skipped exon, intron retention, alternative 5' or 3' splice site, mutually exclusive exon. We also identified event affecting transcription initiation or termination as alternative first or last exons and alternative 5' or 3' UTR. It is noteworthy that mixed and complex events are identified by AStalavista and, occasionally, no simple events or several ones are attributed to a variant.

To assess what fraction of sQTLs were also detected by eQTL analysis (i.e. also associated with the same gene target's overall expression changes), we estimated the proportion of true positive eQTLs at q<0.05 amongst the best sQTL per gene (FDR<0.05) detected by either Altrans or sQTLseekeR, for each of the 9 tissues ($\pi_1=1-\pi_0$), using the Storey and Tibshirani *q*-value approach (*26*). On average 20% of sQTLs associated with changes in exon junction abundance detected by Altrans, were predicted to be eQTLs, with a range of 14-27% across the 9 tissues ($\hat{\boldsymbol{\pi}}_1$=0.20, $\boldsymbol{\pi}_1$=0.14-0.27; Table S10). An even larger fraction of sQTLs detected by sQTLseekeR, associated with changes in relative abundances of gene transcript isoforms, were predicted to be eQTLs (48% on average), ($\hat{\boldsymbol{\pi}}_1$=0.48, $\boldsymbol{\pi}_1$=0.13-0.70; Table S10). While the enrichment of eQTLs amongst sQTLs is much larger than the expected 5% at FDR<0.05, a substantial fraction of sQTLs do not appear to be detected by standard eQTL analysis. This highlights the added value of searching for QTLs associated with other types of molecular phenotypes, in addition to expression variation. Lists of significant sQTLs detected by both methods can be found on the GTEx Portal (http://www.gtexportal.org/).

S7.3 Enrichment of biologically relevant features

The Ensemble regulatory build was used to determine locations of biologically annotated functional elements and all the variants were overlapped with these coordinates to obtain functional annotation per variant. The variant effect predictor was used to find the functional impacts (missense, splice region, etc.) of each variant. All sQTLs were then assigned to functional groups according to these annotations. A null distribution of variants, which were distance to TSS and allele frequency matched to sQTLs (margin of error was for distance 5kb and for allele frequency 2%), was created. This was repeated for the 1st, 2nd, 5th, and 10th most significant variant for each sQTL gene. The enrichment was calculated as the ratio of the frequency of a certain annotation group in the sQTLs to the frequency of the same group in the null distribution. The significance of each enrichment was calculated by a Fisher test.

S7.4 Splice site strength analysis

For every variant located in splice sites regions, we estimated the strength of the site using standard Position Weight Matrices. For each SNP, two scores were computed: for the reference and the alternative allele. Then, we measured the change in strength between the two. The change is considered consistent if the increase in the strength of the site is paired with an increase in site usage. We measured the usage of a given splice site as the sum of the relative abundances of all the transcripts including site. We regressed the transcript relative abundance across the three genotype groups and required a minimum regression slope (minimum 5% change in the site usage from one genotype group to another) along with a minimum strength score change (0.1) in the relevant direction to declare the changes consistent. Splice sites used by all or none of the expressed transcripts were not included here because they could not show any informative variation.

S8 Functional annotation of eQTLs

S8.1 Selection of functional annotations

Enhancers and promoters were defined using the chromatin state segmentation of (*73*). Specifically, promoters were defined as any region that in any of the assayed cell types was assigned to states 1 or 10. Similarly, enhancers were defined as the union of all regions in all cell types assigned to states 6, 7, or 12. Open chromatin regions were defined by taking the union of all DNase I hypersensitive sites from the Roadmap data, using the narrowPeak method. Regulatory protein-bound regions were defined by taking the union of all ChIP-Seq peaks from the ENCODE data (*10*). The final regions included regulator-bound locations from 472 ChIP Seq experiments of general and sequence specific transcription factors, open chromatin regions from DNaseI hypersensitivity experiments in 53 cell types, and maps of histone modifications across 126 Roadmap and ENCODE cell types resulting in chromatin state annotations (*9*) for proximal regulatory regions (promoters characterized primarily by H3K4me3 and H3K9ac) and distal regulatory regions (enhancers characterized by H3K4me1 and H3K27ac).

S8.2 Functional element enrichment analysis

To ask whether the eQTLs discovered across all tissues were enriched in regulatory regions, we chose a top significant SNP per gene from the single tissue eQTL analysis for each tissue (including multiple SNPs when there was a tie). We assembled a set of 14,431 eQTL SNPs. We discarded all SNPs that were within annotated genes, and in order to exclude SNPs that may potentially act post-transcriptionally, we only considered intergenic SNPs. This resulted in 4,085 intergenic eQTL SNPs that we compared to our regulatory annotations. For each tested eQTL subset (as described in the main text), a window 5kb in size (2.5kb on each side) was defined around each of the eQTLs, and these windows were merged if overlapping. The density of functional elements across the union of those windows was then calculated and used as the background frequency to determine the fold enrichment and hypergeometric *p*-value for functional annotation overlaps with the eQTL set. These operations on genomic intervals were performed using BEDTools (*74*). To estimate the enrichment signal for specific expression Quantitative Trait Nucleotides (eQTNs) that are more likely to be unambiguous, we also calculated our regulatory enrichments specifically for a subset of 91 intergenic eQTLs for which a single SNP is the best-associated with expression of a gene in *cis* and the second-best eQTL has a P-value at least 60-fold less significant (-$\log_{10}$ P value 1.5 higher).

To ask whether the specific SNP-gene links ('genetic links') predicted by our eQTL analysis were supported by enhancer-gene links based on functional correlation ('functional links'), we computed functional links as the Pearson correlation between open chromatin and gene expression at neighboring genes across 110 ENCODE and Epigenomics Roadmap datasets (*73*). We compared open chromatin regions containing an eQTL SNP to those without an eQTL.

S9 Co-expression network analysis and systems biology of eQTLs

S9.1 Constructing co-expression networks

For each of the 9 tissues used in the eQTL analysis above (those with > 80 samples), a co-expression network was constructed using Pearson correlation coefficient to measure pairwise similarities between genes expression levels. Each network was then sparsified to only include the top 1% strongest co-expressed links. This analysis only considered genes that were expressed (RPKM > 0.1 in 80% of samples) in all 9 tissues, and was performed on "residual" expression data where, as done for the eQTL analysis, the effect of 15 PEER factors, gender, and 3 genotype PCs were removed for each tissue.

As described in the multi-tissue eQTL analyses, following the approach of Nica et al. (*22*), quantification of sharing of gene-gene "co-expression links (or edges)" among pairs of tissues were performed using the π_1 statistic. In this approach, given a "discovery" co-expression network A and a "validation" co-expression network B, p-values for all links discovered in A were computed in B (as the p-value associated with correlation coefficient for each link). These p-values were then used to compute Storey's π_1 value, quantifying the proportion of links in A that are replicable in B. Note that this procedure is not symmetric, so for the nine tissues a total of 9X8=72 π_1 values were computed.

<u>S9.2 WGCNA co-expression networks and module annotation</u>

For each of the 9 tissues used in the eQTL analysis above (those with > 80 samples), the effect of top 3 PEER factors, gender, and 3 genotype PCs were removed. Weighted Gene Co-Expression Network (WGCNA) was then applied to each individual tissue to obtain co-expression network structures (*41,75*). The WGCNA assigns genes into different modules based on the correlation among genes. The weighted network analysis begins with a matrix of the Pearson correlations between all gene pairs, then converts the correlation matrix into an adjacency matrix using a power function f(x)=x^β. To explore the modular structures of the co-expression network, the adjacency matrix is further transformed into a topological overlap matrix (*41*). As the topological overlap between two genes reflects not only their direct interaction, but also their indirect interactions through all the other genes in the network, previous studies (*76*) have shown that topological overlap leads to more cohesive and biologically meaningful modules. To identify modules of highly co-regulated genes, we used average linkage hierarchical clustering to group genes based on the topological overlap of their connectivity, followed by a dynamic cut-tree algorithm to dynamically cut clustering dendrogram branches into gene modules (*77*). To distinguish between modules, each module was assigned a unique color identifier, with the remaining, poorly connected genes colored grey.

ENCODE ChIP-seq data was accessed in September 2012, and we used peak calls based on spp peak calling program. We merged all ChIP-seq data which contained information covering 135 unique TFs across 90 cell lines. In optimal files, default FDR qscore cut-off (10^{-3}) was used, and UCSC refseq 2012 was used to annotate gene-TF binding information. Given a significant peak was centered within 10kb of the transcription start position, we established a "regulation" between the TF and the corresponding gene. For a module of genes as output from

WGCNA, we then used Fisher exact test to evaluate the significance of the enrichment of certain TF-gene interactions (FDR <0.1%). We used Fisher Exact Test to annotate the GO function enrichment of module genes.

To identify cross tissue module interactions, we considered both cross-tissue module overlap and PC correlation. We tested whether modules from different tissues shared significantly more genes than by random chance using Fisher exact test. In addition, an overlap score $S(\Omega_1,\Omega_2)$ was defined between modules Ω_1 and Ω_2 from two tissues as $S(\Omega_1,\Omega_2)=|\Omega_1 \cap \Omega_2|/\min(|\Omega_1|,|\Omega_2|)$ to measure the fraction of overlap. We claimed two modules were significantly overlapping with each other if $S(\Omega_1,\Omega_2)>0.2$ and Bonferroni corrected Fisher exact test p-value < 0.001.

For module PC correlation calculation, samples from two tissues (T_1,T_2) were first paired based on donor IDs and samples only available in one of the tissues were excluded. The first principal components PC^1 of module M in tissue T_1 and $M^{'}$ in T_2 were calculated using R princomp. Pearson correlation coefficient ρ between the two PCs was then calculated. The two modules were considered to be correlated if $|\rho(\mathrm{PC}^1_{T_1,M},\mathrm{PC}^1_{T_2,M'})|\geq 0.58$. These two measurements enabled us to quantify the cross tissue module interactions and estimate how many of them could simply result from membership sharing.

S9.3 Module-Switching QTLs

Within each gene and using the complete set of normalized RPKM expression values available, we used a variation of inverse distance weighting to impute all the missing expression values, obtaining a cross-validation correlation of 0.98 after 1,000 random deletions and re-imputations of the observed expression values (Fig. S31A**).** Our imputation technique was particularly tailored to capture multi-tissue bimodal distributions in expression, which lie at the heart of systematic multi-tissue expression variation, and generally detected even small changes in expression as in HLA-G. This allowed us to discover 688 genes whose expression depends highly on gender, which we discarded from the rest of the analysis.

To impute expression, we considered per gene the 45 tissues x 175 individuals matrix of observed expression and for each missing entry (i,j) of this matrix, we computed the mean squared distance d_{row}(i,k) between row i and all other rows k for which entry (k,j) was available, and analogously the mean squared distance d_{col}(j,k) of column j and all other columns k for which entry (i,k) was available, averaging based on the number of common observed entries between each pair of rows and columns. We calculated an initial value for entry (i,j) as a weighted sum of the observed entries in row i and column j, where observed entries (k,j) were weighted with d_{row}(i,k)/S and observed entries (i,k) with d_{col}(j,k)/S, S chosen so that the sum of all weights was 1. This process was repeated 200 times, at each step recalculating the missing entries from the information available from the previous step and using a damping factor of 0.9.

Using imputed data, we grouped the remaining 19k protein-coding genes into 117 modules based on their average cross-tissue expression patterns. To discover modules, we first projected the average expression patterns of genes onto low-dimensional space and then learned extremal points or vertices of this simplified expression space using a convex hull approximation, further requiring that vertices were supported by at least other 15 genes by correlation-similarity to handle outliers. This procedure gave us 117 characteristic patterns of expression, and then global module membership was decided after finding the closest characteristic pattern to each gene's mean pattern. Notably, since our modules and their characteristic patterns were constructed from absolute mean expression across individuals, both are robust to global non-genetic confounders of gene expression. Modules varied in size from 25 to 400 genes each, and were strongly enriched (Storey Q < 3e-3 after permutation randomization) for common gene functions highly relevant to the corresponding tissues**.**

To measure coordinated multi-tissue variation across individuals, we used the 117 characteristic patterns as fixed pinpoints in the multi-dimensional expression space, and then computed proximity scores of the imputed cross-tissue patterns of expression of each gene and individual to these points, at each step narrowing the calculation of scores only to the most relevant and independent characteristic patterns of each gene.

To find modules, we considered the imputed 45x175 matrices of each gene and calculated the mean expression across rows. We then standardized these vectors, giving us a list of 19k mean correlation-patterns of expression across tissues, each corresponding to a protein-coding gene. We called these set of 19k points in 45-dimensional space the *space of expression*. We then mapped these vectors onto 20-dimensional space after taking their projection onto the first 20 Principal Components of the space of expression, curating the signal from possible Gaussian noise. Within each of these 20-dimensional vectors, we over-expressed entries with values > 2 and down-expressed entries with values < -2, multiplying them times a large positive factor, and then re-standardized the vectors. Lastly, we used a combination of affinity propagation and k-means on our modified vectors to discover vertices and re-mapped those to 45-dimensional space, producing the list of 117 characteristic patterns of expression. Modules where decided by selecting, for each of the 19k mean patterns of expression of genes, the closest characteristic vector among the 117 exemplars. Characteristic patterns with less than 15 associated genes where discarded. Notably, the traditional routine of clustering for centroid-discovery was modified into a vertex-discovery framework by considering correlation space and an over-representation of the signal on each gene.

Gene ontology (GO) enrichment was evaluated using only Biological Process (BP) terms which had at least 10 and at most 1000 annotations among our 19k genes, for a total of 4,206 BP terms used, and then using only those genes annotated in at least one of these terms, for a total of 13,013 genes. Electronic annotations were also considered as they added important validation information to patterns. Results were assessed after considering 200 random permutations of the module-membership list to compute the null p-value distribution.

To discover module-switching we calculated, per gene and individual, probabilities of membership to each of the 117 modules based on the Euclidean distance between the gene and each characteristic pattern. Then, the rows of the 117x175 matrix of probabilities of each gene were each projected onto the first 5 first principal components of the 45x175 expression matrix of the gene, and only the top scoring rows (i.e. the rows whose projection had the highest Euclidean norm along each PC) for each PC were selected as quantitative traits for modQTLs (module-switching QTLs) discovery. This gave at most 5 modules to consider for each gene, which clearly represented the most relevant and independent module-switching signals of the gene. We then performed the modQTLs discovery running Matrix eQTL (*21*) in *cis*, simultaneously permuting 1,000 times the entries of the module-switching vectors of genes to calculate Storey Q-values from the null p-value distribution.

S10 Personal transcriptomics and implications for human disease

S10.1 Analysis of Protein Truncating Variants (PTV's)

Annotation of all coding variants was first performed using Variant Effect Predictor (VEP) and PLINK/SEQ (v0.09, http://atgu.mgh.harvard.edu/plinkseq/) using the GENCODE V12 reference transcript set data set. Rare functional variants were identified based on stop, frameshift indel, nonsynonymous (SNV or 3n indel) or splice predictions. An additional layer of annotation for high confidence loss of function mutations used the methods described in MacArthur *et al*. (*43*). The Variant Effect Predictor (VEP v2.5) tool from Ensembl was modified to produce custom annotation tags and additional protein truncating variant (PTV) annotations. The additional PTV annotation was applied to variants which were annotated as STOP_GAINED, SPLICE_DONOR_VARIANT, SPLICE_ACCEPTOR_VARIANT, and FRAME_SHIFT and flagged if any filters failed. Filters included: PTV is the ancestral allele; exon is surrounded by non- canonical splice site (that is not AG/GT); PTV removes less than 5% of remaining protein; PTV is rescued by nearby start codon which results in less than 5% of protein truncated; transcript only has one coding exon; splice-site mutation within intron smaller than 15 bp; splice site is non-canonical OR other splice site within same intron is non-canonical; unable to determine exon/intron boundaries surrounding variant. A PTV variant is predicted as high confidence if there is one transcript that passes all filters, otherwise it is predicted as low confidence. We used PLINK/SEQ to generate predictions of nonsense-mediated decay based on

Maquat *et al*. (*78*) and Nagy and Maquat (*79*). RNA-seq isoform informed annotation and visualization of variants were performed using MAMBA (http://www.well.ox.ac.uk/~rivas/mamba/) (*80*). Additional methods for PTV analyses are presented in a companion manuscript (*31*).

S10.2 GWAS and eQTL integration analyses

We combined NCBI's Phenotype-Genotype Integrator (PheGenI) (downloaded March 2013, and publicly available from http://www.ncbi.nlm.nih.gov/gap/phegeni) (*48*), and the NHGRI Catalog of curated GWAS results (downloaded on November 2013 and publicly available from https://www.genome.gov/26525384) (*49*). The resulting catalog of SNP-trait associations included 10,129 genome-wide significant associations (defined as $p<5x10^{-8}$) and about 630 distinct phenotypes. This set formed the basis of the eQTL analyses of GWAS-identified associations.

Because most of the GWAS-identified variants were identified in samples of European ancestry, we used LD information from the 1000 Genomes Project samples of European descent (CEU) to arrive at a subset of independent variants. To this end, we pruned the full set of 10,129 genome-wide significant SNPs collectively using a $r^2 \geq 0.80$ LD cutoff, counting pleiotropic loci (all SNPs within $r^2 \geq 0.8$ associated with more than one complex trait) as a single instance. We also annotated each GWAS SNP with eQTLs (defined as FDR<0.05) in strong LD ($r^2 \geq 0.80$) with the GWAS SNP, considering the "best eQTL per gene" derived from the single-tissue and/or multi-tissue analyses in at least one of the 9 tissues examined. To assess the importance of regulatory effects (vs coding effects), we annotated each GWAS SNP with putative deleterious coding variants (non-synonymous or splice variant) in LD ($r^2 \geq 0.80$) with the GWAS SNP, using NCBI's dbSNP version 137 (http://www.ncbi.nlm.nih.gov/SNP/). We quantified the proportion of eQTLs in LD with GWAS SNPs that had been detected using only the single-tissue analyses or only the multi-tissue eQTL methods or the union of both. The frequencies of various genomic contexts of the GWAS SNPs in LD with a GTEx eQTL were estimated using genomic context annotations from the PheGenI database (listed in Table S13).

We compared proximity-based and eQTL-based gene assignment for GWAS SNPs and generated a list of GWAS SNPs in LD with an eQTL, for which the eQTL target gene and the physically nearest or host gene are discordant. We considered the distribution of the distance to the target gene of trait-associated GTEx eQTLs, including the case where the target genes are restricted to protein-coding genes (Table S14).

We evaluated the relevance of the identified eQTLs in disease mapping studies by analyzing the Wellcome Trust Case Control Consortium (WTCCC) studies of seven complex diseases (*45*). We evaluated enrichment for association with each of the 7 disease phenotypes among the eQTLs identified in each of the tissues as well as those from the multi-tissue eQTLs. Separately, we also considered enrichment for association with the autoimmune disorders (defined here as Crohn's disease, Rheumatoid arthritis, and Type 1 diabetes) among the eQTLs from the single-tissue and multi-tissue analyses. We generated a quantile-quantile (Q-Q) plot showing the associations of the single-tissue eQTLs, by tissue, with each of the disease traits. For each tissue, an enrichment p-value was derived from an application of the Kolmogorov-Smirnov (KS) test and reflects an enrichment of low p-values in association with the trait under investigation comparing the GWAS p-values of the eQTLs to the rest of the GWAS SNPs.

S11 Online Resources

All protected data including sequencing BAMs and clinical data are available through dbGaP by application (http://www.ncbi.nlm.nih.gov/gap).

All eQTL results are browseable on the GTEx portal, and analysis files, including significant single and multi-tissue eQTLs, and s-QTLs are available for download on the GTEx portal (www.gtexportal.org).

MAMBA software is available at http://www.well.ox.ec.uk/~rivas/mamba.

Box S1. Glossary of terms

Expression quantitative trait locus (eQTL): A genetic locus where the genotype of a variant is significantly associated with gene expression levels of one or more genes. Due to linkage disequilibrium (see below), an eQTL usually contains multiple DNA variants that tag the actual causal variant, which is responsible for the genotype-dependent gene expression. While in this study we use single-nucleotide polymorphisms (SNPs) as markers for eQTL discovery, the causal variant may be any type of DNA variant, including SNPs (**eQTNs**), indels or copy number variants. *There are two types of eQTLs:* (i) ***Cis*-eQTL**: A genetic variant that influences the expression levels of a proximal gene on the same chromosome in an allele-specific manner. (ii) ***Trans*-eQTL:** A genetic variants that affects gene expression through an intermediate *trans* factor, such as a protein or RNA regulator. Trans-eQTLs usually lie far away from the target gene or on a separate chromosome.

eGene: Genes with at least one significant *cis*-eQTL. In the current study, we considered only the most significant eQTL per gene due to power limitations. Multiple independent eQTLs per gene have been shown to exist.

Tissue-significant eQTLs: All eQTLs that are significant in a given tissue (here at FDR<5%) irrespective of their activity in other tissues tested.

Tissue-specific eQTLs: Subset of eQTLs active solely in one tissue out of *n* tissues tested, assessed using joint eQTL discovery methods that consider the association of a given variant across multiple tissues simultaneously.

Ubiquitous eQTLs: eQTLs that are significant in all tissues tested.

Splicing quantitative trait locus (sQTL): A genetic locus where the genotype of a variant is associated with differential alternative splicing activity or differential isoform abundance. We distinguish here between two different types of sQTLs that are examined in this paper: (a) **Splice-junction quantitative trait locus (sjQTL):** A genetic variant associated with changes in exon junction abundance (detected in this study using Altrans, *35*). (b) **Splicing-isoform ratio quantitative trait locus (srQTL)**: A genetic variant associated with changes in the relative abundances of gene transcript isoforms (detected here using sQTLseeker, *36*). This is distinct from a **transcript QTL** (**trQTL**), which is a genetic variant associated with the absolute abundance of a single isoform of a gene.

Module-switching QTL (modQTL): A genetic variant associated with membership switching of genes among dissimilar gene co-expression networks (modules) between different individuals. Co-expression modules were inferred from gene expression variation across tissues within individuals.

Allele-specific expression (ASE): Also known as ***Allelic imbalance*** of gene expression**,** ASE refers to significant differential expression between two allelic transcripts within a given heterozygous individual. A notable strength of ASE analysis compared to eQTL analysis is that it can be applied to a single sample versus a set of samples that are needed to detect eQTLs. ASE can be used as an independent replication of eQTL effects.

Genome-wide association studies (GWAS): Population- or case-control-based studies aimed at identifying in an unbiased manner, DNA variants (i.e., genomic loci) associated with complex disease risk or quantitative trait variation. Hundreds of thousands to millions of SNPs (>1% frequency in the population) are genotyped in these studies. To increase genomic coverage, missing genotypes are imputed into the study using a reference panel (e.g., HapMap or 1000 genomes project). To increase power, meta-analyses are often performed across multiple GWAS for a given disease or trait (individual GWAS typically contain hundreds to thousands of individuals).

Linkage disequilibrium (LD): The non-random association of alleles at two neighboring loci that descend from a single, ancestral chromosome. LD is disproportionately correlated with recombination rate between the two loci and is affected by other factors, such as selection and genetic drift. The correlation coefficient, r^2 between two alleles at two loci is a common measure of LD, and is dependent on the population frequencies of these alleles.

Fig. S1. Summary of tissues collected, sequenced, and analyzed during the GTEx pilot. (**A**) Distribution of the number of tissues collected per donor during the pilot phase. 52 tissue sites were represented (plus two cell line samples for a total of 54), but on average only 28.3 (median=26) tissues were collected for any given donor. The bimodal distribution is due to the subset of donors on whom brains were collected, adding an additional 9–11 samples for those donors. (**B**) The distribution of the number of RNA sequenced tissues per donor represented in in the final pilot analysis freeze (analyzed in this paper). (**C**) The distribution of samples received per tissue during the pilot phase. Tissue site abbreviations are described in Table 1. Tissues such as blood (WHLBLD), skeletal muscle (MSCLSK), and skin (SKINS) were collected on almost all donors, whereas brains (BRNxxx) and transplanted organs, such as kidneys (KDNxxx) and liver (LIVER) were obtained less frequently. The gender-specific tissues are also less frequent. Several tissues were preserved poorly overall and collection of those was ceased during the pilot (e.g., bladder [BLDDER] and fallopian tube [FLLPNT]). All samples that are included in the final analysis data set are shown in dark blue and samples that were sequenced, but excluded from analysis, are shown in pale blue. Sequenced samples were excluded from the final pilot analysis freeze for QC reasons (samples were excluded due to low tissue identity or donor issues, e.g., two Kleinfelter's syndrome donors were excluded), if they were duplicate samples, or due to data lag (e.g., no matching DNA sample was available at the time of analysis; these latter samples will be included in subsequent analysis freezes when DNA samples become available). Samples in pale green represent samples that were collected during the period, but for whom processing was either ongoing, or if processed, failed to meet RNA QC criteria.

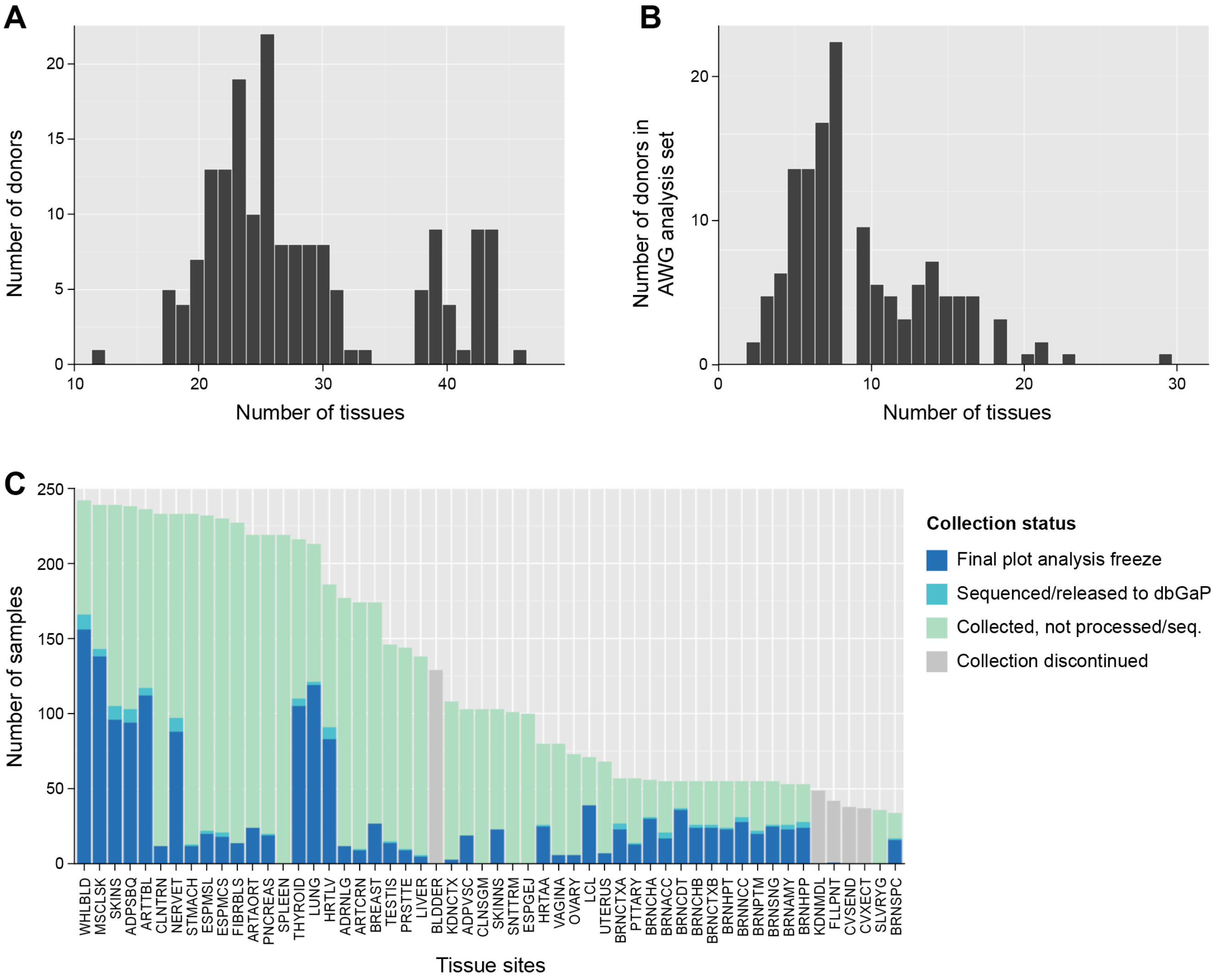


Fig. S2. RNA quality is influenced by tissue site and ischemic time. Using a best subset generalized linear model, ischemic time and tissue site had the largest effect on RNA quality, followed by autolysis score, which collectively accounted for ~40% of the variance of RNA quality ($p < 1 \times 10^{-27}$ for correlations between each significant predictor variable and RNA quality or its residual). The remaining variables tested, such as cause of death or PAXgene fixative time, showed very small contributions to RNA quality (<1% each). (**A**) Shows box plots of RNA quality as measured by RIN for all the PAXgene preserved tissues sampled during the pilot phase. A RIN value of 6.0 was the cutoff used for inclusion in RNA sequencing. Some tissues, such as muscle, lung, and esophagus, have well-preserved RNA. Other tissues, such as kidney, colon, and spleen, generally have poorly preserved RNA and few samples that meet the cutoff for inclusion. (**B**) Shows that the effect of ischemic time on RNA degradation is tissue specific. RNA quality in a tissue, such as the skin, is well preserved over a long ischemic time interval, whereas in pancreas RNA degrades very rapidly with increasing ischemic time.

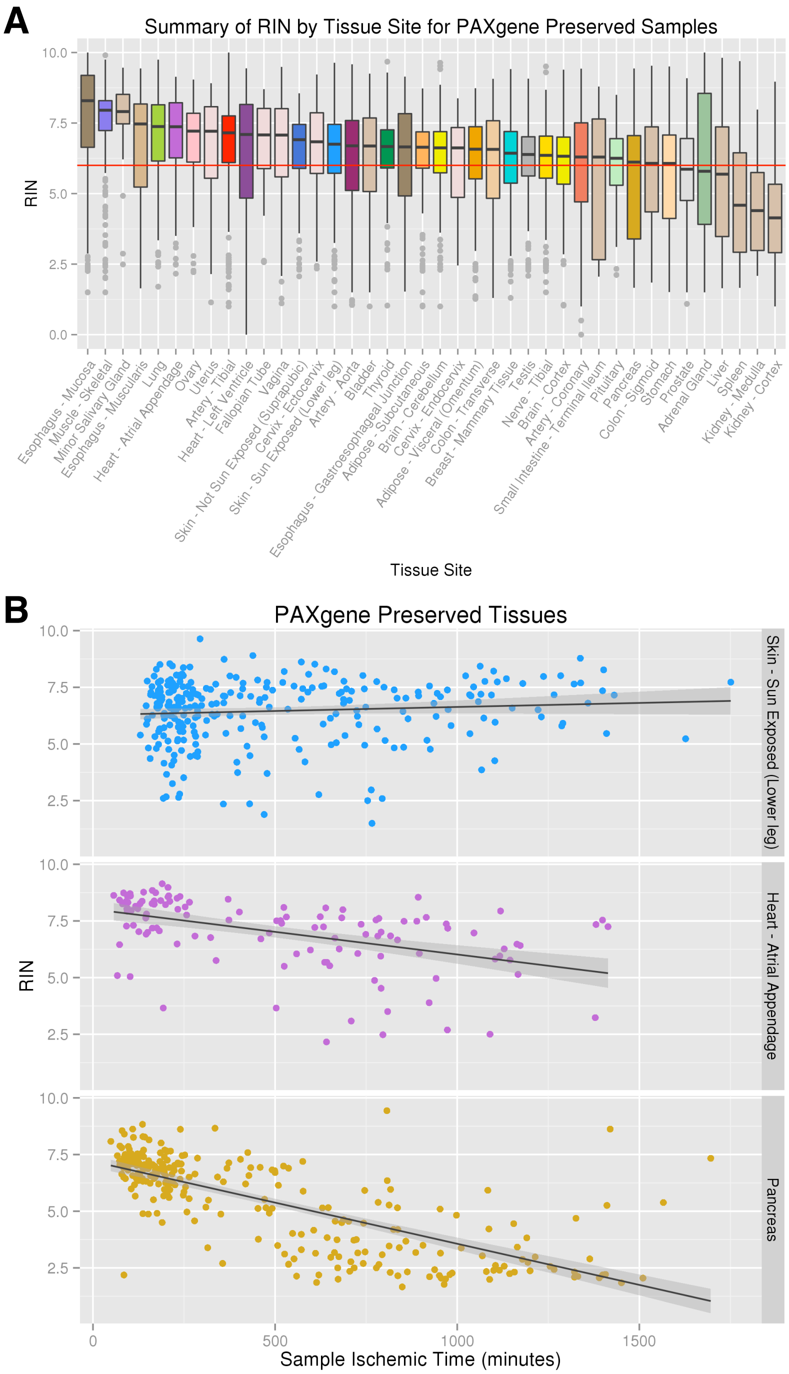


Fig. S3. RNA sequencing summary statistics. (**A**) The number of mapped reads obtained for each sample ranges from 12–200 million (average 80 million). Tissues sequenced early in the project had more variable read depths, while those sequenced later have higher numbers of mapped reads. A lower cutoff for inclusion was placed at 10 million mapped reads. (**B**) The number of genes detected based on all transcript encoding entries in GENCODE, which total ~55,000. Across tissues, we detect expression (RPKM > 0.1) for almost half of the genes. Notable outliers include testis, which expresses an average of 32,604 gene transcripts, and whole blood, muscle and kidney, which are dominated by a few highly expressed species resulting in fewer detectable gene transcripts. Tissue site abbreviations are described in Table 1.

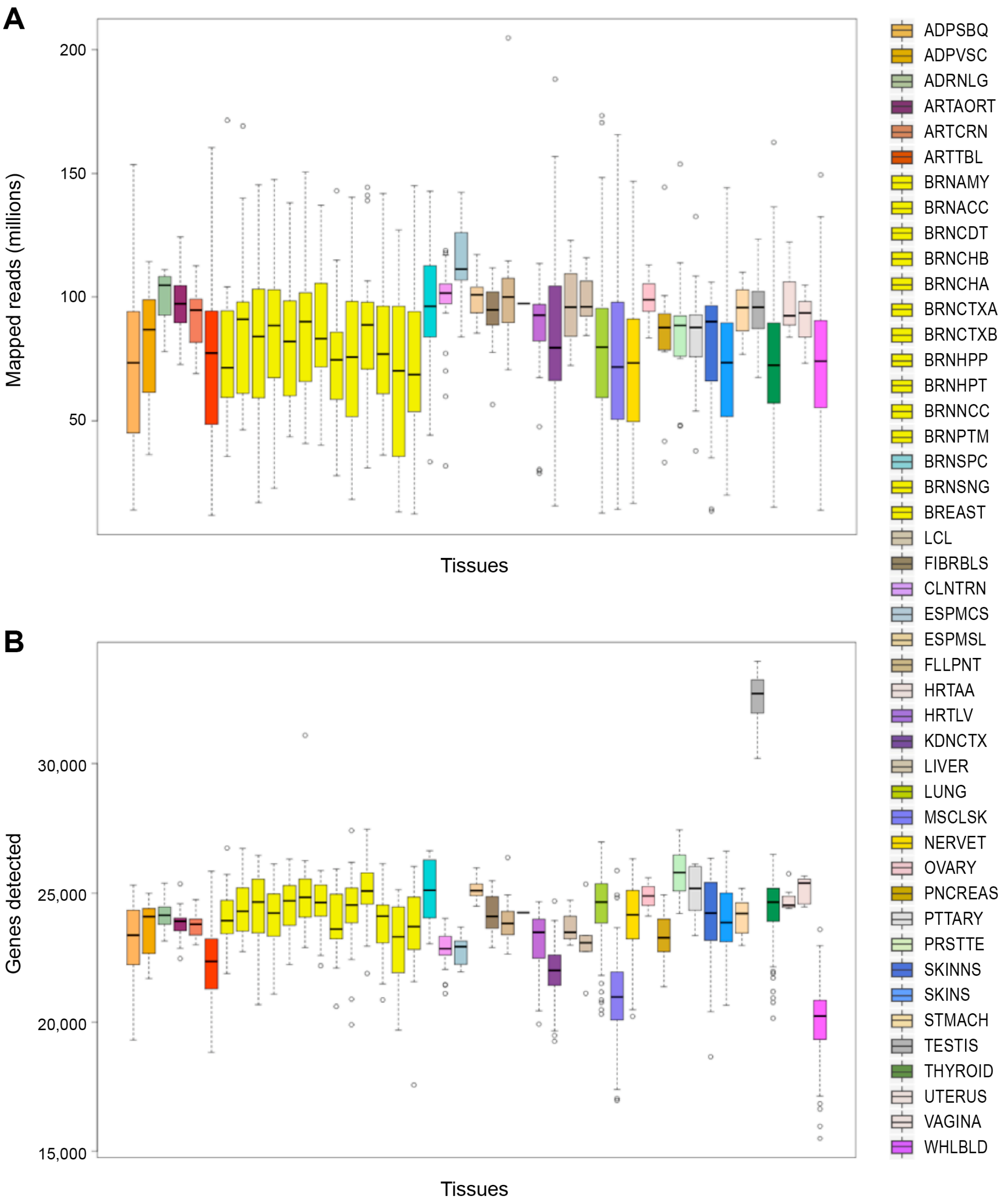


Fig. S4. Hierarchical clustering of brain region samples (**A**) Clustering of the 324 brain samples shows that the cerebellum region (represented by duplicate samples 'Cerebellum'=PAXgene preserved, and 'Cerebellar hemisphere'=Fresh Frozen) is the most differentiated among the brain region sub-samples. This is a close up of the same region as shown in Figure 1A. (**B**) Hierarchical clustering of the 76 cerebellum and cortex duplicate-sample pairs taken from the same donor brains (cerebellar hemisphere/cerebellum and cortex/frontal cortex) (*14*). Redundancy is observed for 11 of the 21 cerebellum pairs which cluster tightly together by donor, and 5 of the 17 cortex sample pairs which cluster by donor. The incomplete concordance among the remaining sample pairs might be due to the differing post-mortem ischemic intervals among the pairs or to slight location differences in the repeated sampling of the duplicate regions. The dendrogram overall reiterates the clear separation between the cortex and cerebellar regions, but little impact (clustering) of preservation method (PAXgene in black, Fresh Frozen in red)**.**

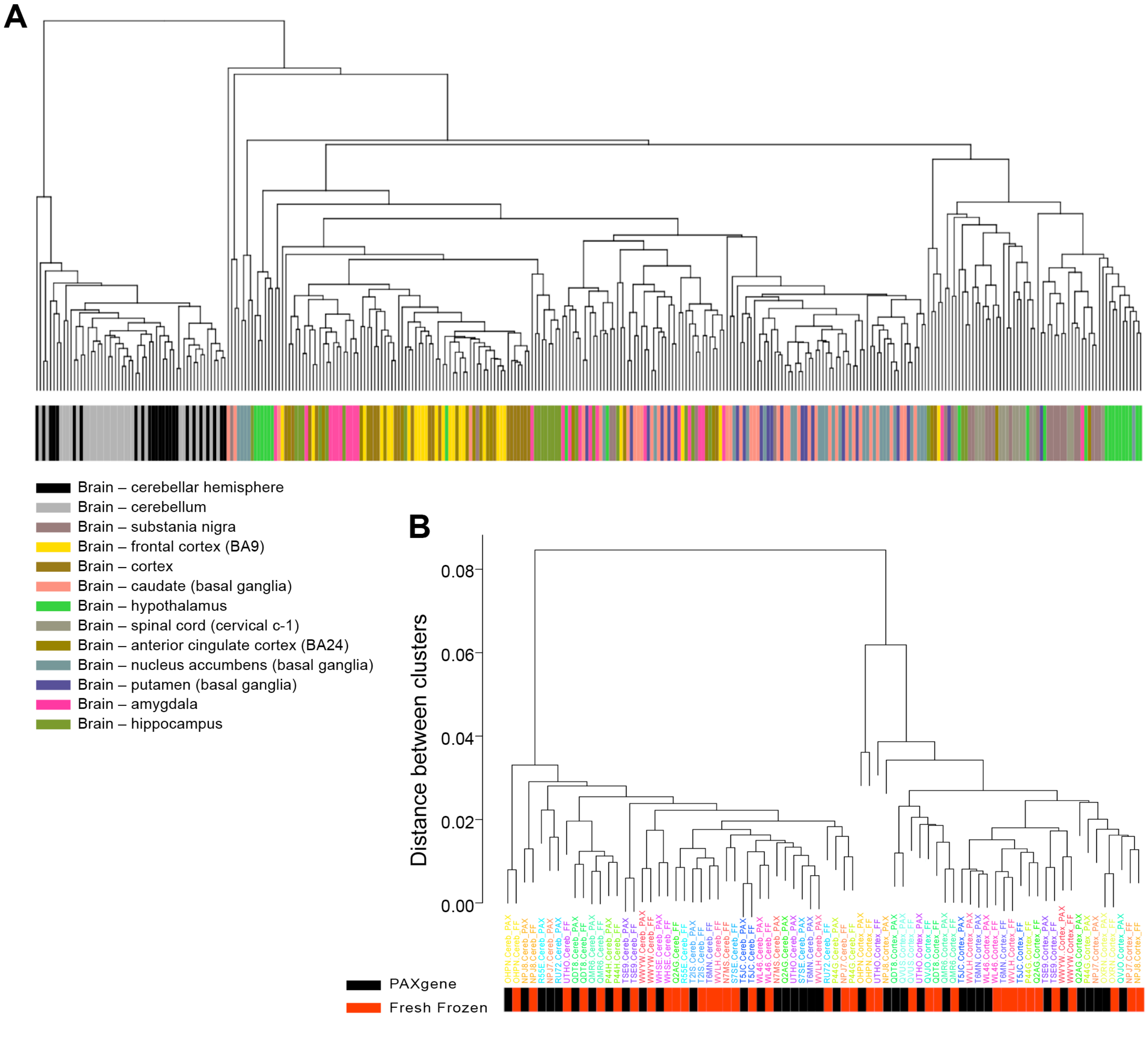


Fig. S5. **Principal component analysis (PCA) of the 185 GTEx DNA samples** (**A**) with 1 HapMap, and 455 HapMap2 samples genotyped on the Illumina Omni 5M array. (**B**) PCA of 185 GTEx samples and 1 HapMap sample genotyped on the Illumina Omni 5M array and 1184 HapMap3 samples with 1.65M SNPs. *x*-axis = principal component 1; *y*-axis = principal component 2.

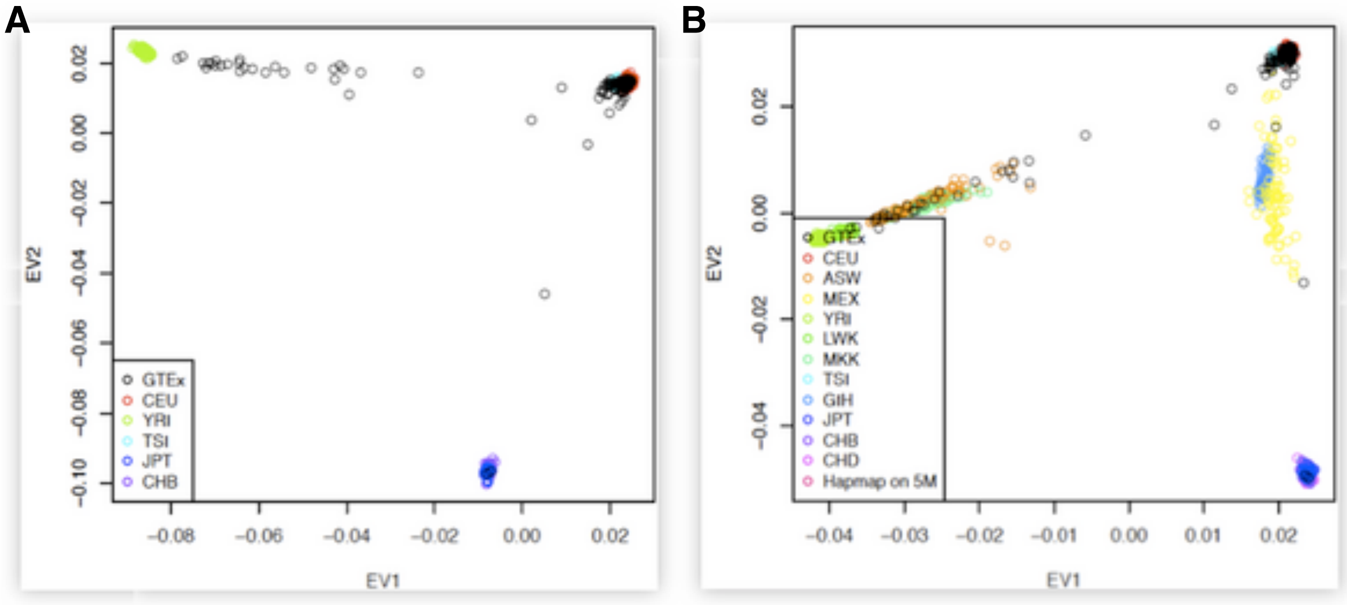


Fig. S6. **Distribution of imputed SNPs by allele frequency (AF) and imputation quality score (INFO).** The imputation was performed for 185 GTEx individuals genotyped on Illumina's Omni 5M array using the 1000 Genomes Project Phase I as the reference panel.

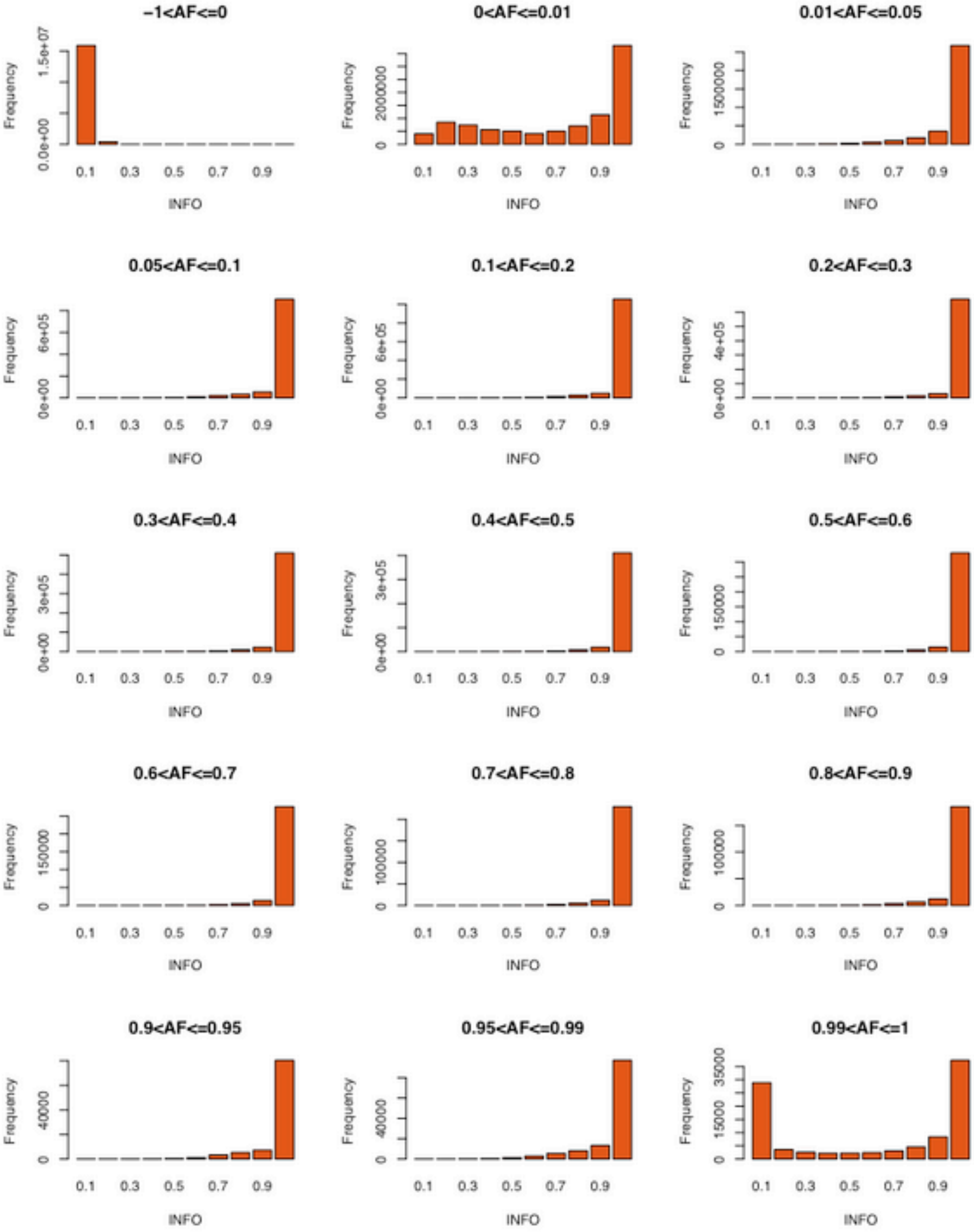


Fig. S7. **Evaluation of imputation accuracy.** Comparison of imputed calls on Illumina's OMNI 5M array to direct genotype calls on the Exome Chip using 184 GTEx samples. (**A**) Mean r^2 of imputed dosage of the minor allele refers to the correlation coefficient (r^2) between the imputed calls on the 5M array and the direct calls on the exome chip per SNP across all individuals, averaged across all SNPs and computed in each IMPUTE2 INFO and MAF bin, separately. (**B–D**) Concordance rate refers to the fraction of all SNPs whose imputed genotype call at a posterior probability above 0.9 is concordant with the direct calls on Exome Chip, computed separately for each INFO cutoff and MAF bin. IMPUTE2 info = a measure of confidence of imputation; af = allele frequency.

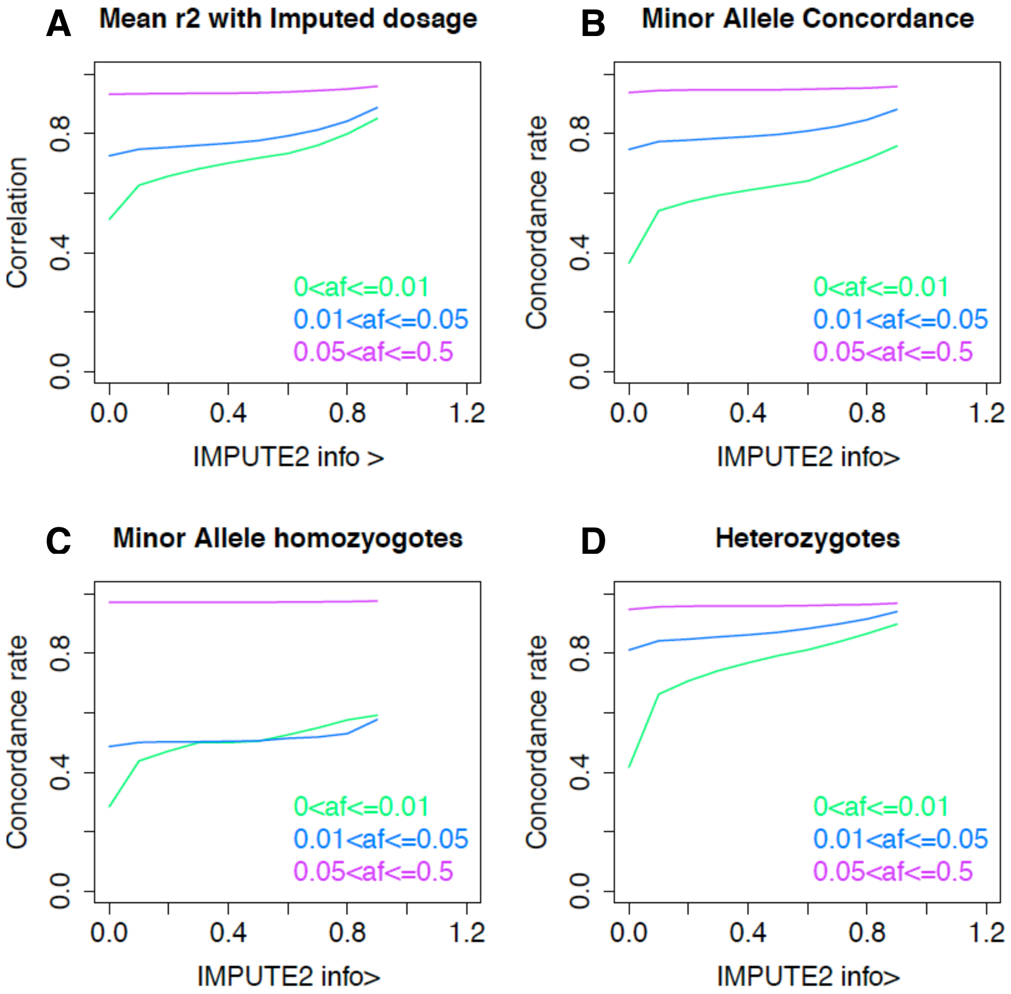
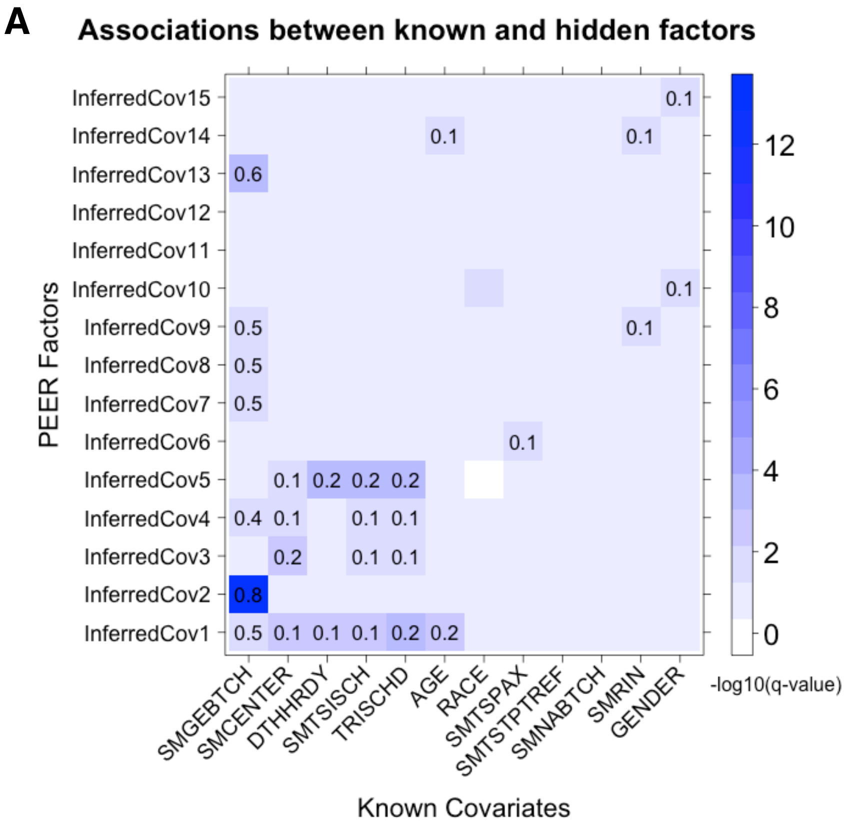


Fig. S8. **Analysis of PEER factors**. (**A**) An assessment of correlations between inferred PEER factors and known covariates in adipose tissue as a representative example. The color signifies significance of the association. For significant associations (*q* value $\leq$ 0.05), the r^2 value is reported in the cell. (**B**) The meaning of each covariate abbreviation is given in the table.



B

Code	Meaning
SMGEBTCH	Expression batch ID
SMCENTER	Collection center
DTHHRDY	Hardy scale
SMTSISCH	Ischemic time for sample
TRISCHD	Ischemic time for individual
AGE	Age of individual
RACE	Self reported race
SMTPAX	Time spent in fixative
SMTSTPTREF	Procurement reference point
SMNABTCH	Nucleic acid isolation batch
SMRIN	RNA quality score (RIN)
GENDER	Gender of individual

Fig. S9. Density and *p*-value distributions of best *cis*-eQTL per gene relative to transcript start site (TSS). (**A**) The fraction of most significant eQTLs per gene is plotted as a function of distance from the TSS, measured in kilobases (kb), for the significant eGenes at FDR < 5% (blue line), the non-significant genes at FDR > 5% (red line), and all SNP–gene pairs tested for eQTL association (black line), shown in whole blood. Negative distance refers to eQTLs upstream of the TSS and positive distance refers to eQTLs downstream of TSS, taking into consideration the strand orientation of the gene. eQTLs were plotted within a ± 1-Mb window around the TSS using 10-kb bins. (**B**) The density distribution of the most significant eQTL per gene for the significant eGenes (FDR < 5%) as a function of distance from the TSS is overlaid for all 9 tissues tested. The density plot is based on 10-kb bins.

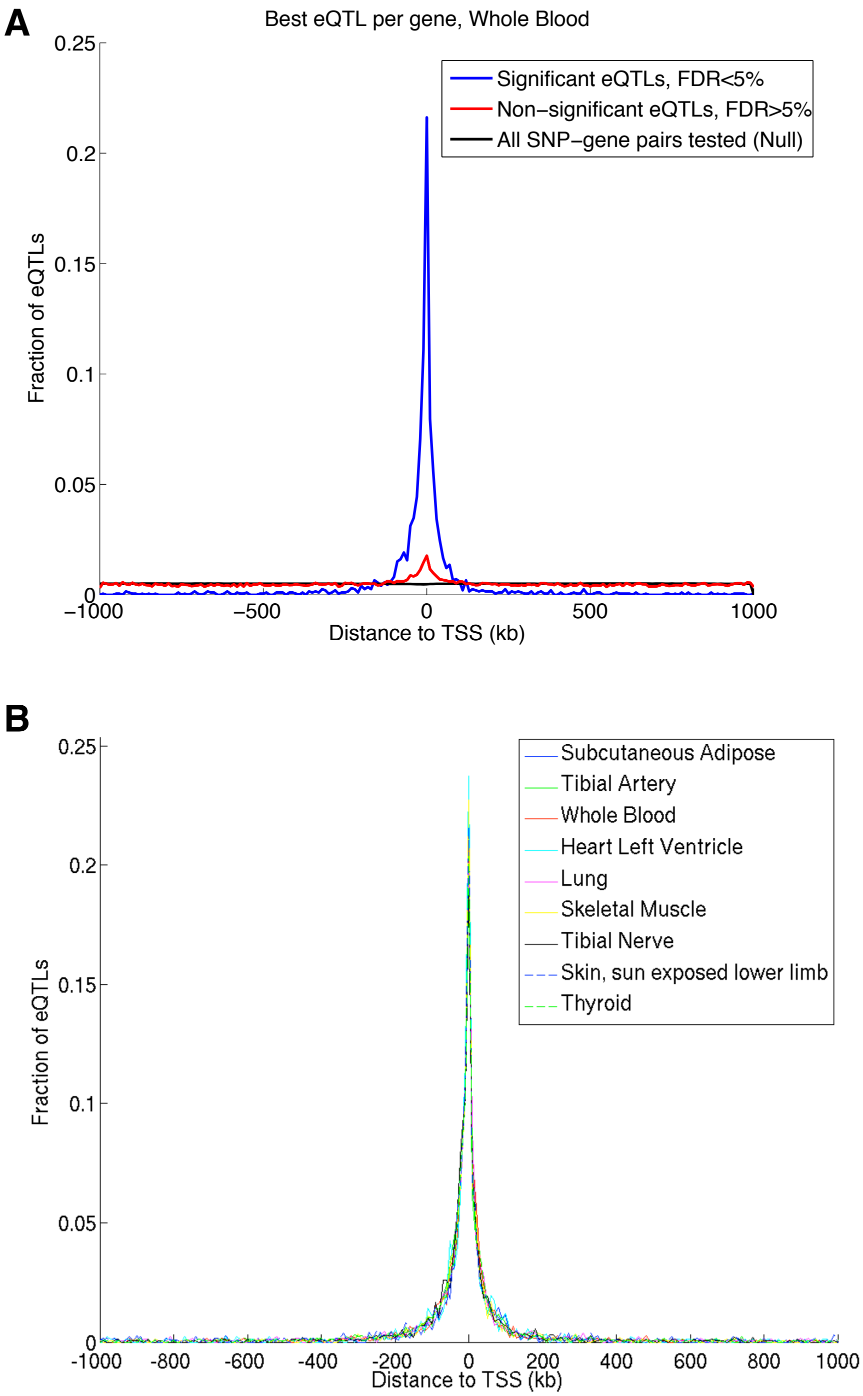
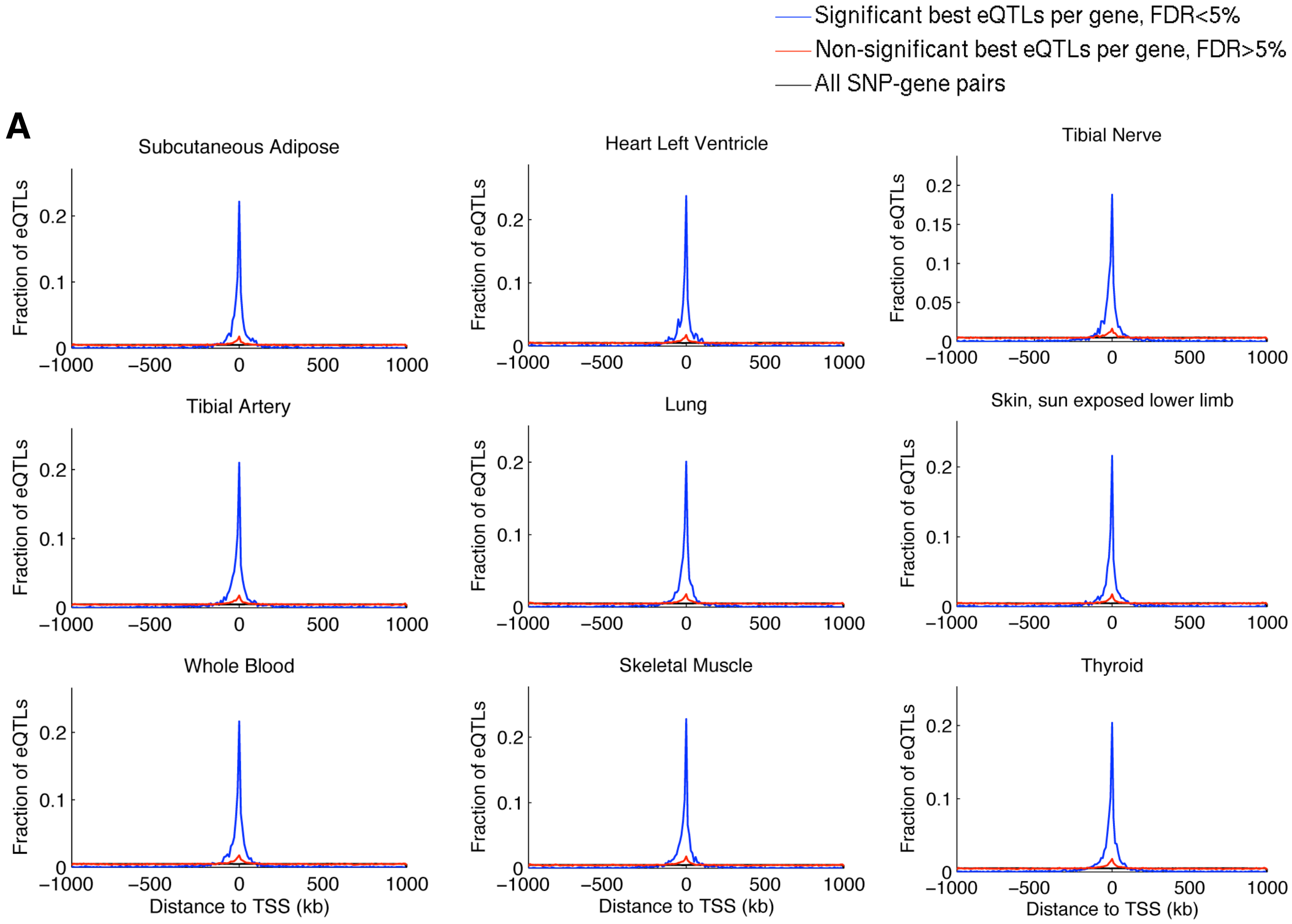


Fig. S10. Density and *p*-value distributions of best *cis*-eQTLs per gene relative to TSS for 9 tissues. This is an expansion of Figure S10A,B for all 9 tissues tested in the pilot phase of GTEx. (**A**) The panels display the fraction of the most significant eQTLs per gene plotted as a function of distance from the TSS, measured in kilobases (kb), for the significant eGenes at FDR < 5% (blue line), the non-significant genes at FDR > 5% (red line), and all SNP-gene pairs tested for association (black line) for all 9 tissues analyzed. Negative and positive distances refer to eQTLs upstream or downstream of the TSS, respectively, taking into consideration the strand orientation of the gene. *cis*-eQTLs were plotted within a ± 1-Mb window around the TSS, using a 10-kb bin. (**B**) The panels contain scatter plots of the negative log10 of eQTL *p*-values of the most significant SNP per gene for the significant eGenes at FDR < 5% as a function of distance of the associated SNP to the TSS (in kilobases).



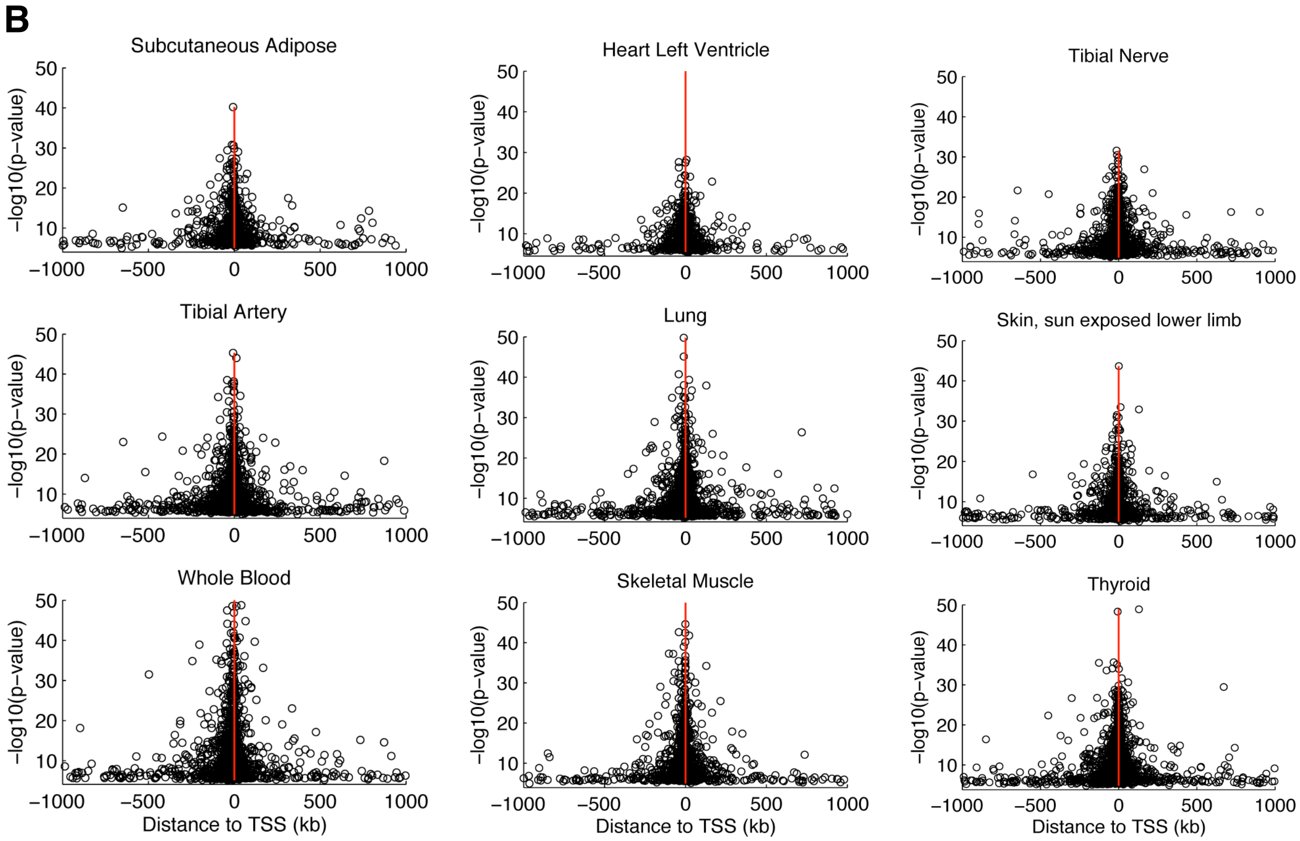

B
Subcutaneous Adipose
Heart Left Ventricle
Tibial Nerve
Tibial Artery
Lung
Skin, sun exposed lower limb
Whole Blood
Skeletal Muscle
Thyroid
−log10(p−value)
50
40
30
20
10
−1000
−500
0
500
1000
Distance to TSS (kb)

Fig. S11. Replication rate of whole blood *cis*-eQTLs in GTEx. The histogram represents the distribution of GTEx whole blood eQTL *p*-values for 644,043 significant SNP-gene pair *cis*-eQTLs (FDR < 0.05) discovered in a separate study of 911 whole blood samples from unrelated individuals in Estonia (*14*). We used the GTEx *P*-value distribution of the Estonia eQTLs to estimate the π1 statistic, which is the proportion of replicated significant *cis*-eQTLs in the GTEx study (see Storey and Tishirani *q*-value method (*26*)). The π1 statistic is computed from π0 (π1=1-π0) that is the estimated overall proportion of true null hypotheses among all tests performed. The GTEx replication rate of the Estonia eQTLs was: π1=51%. The histogram bin size used was *p* = 0.01.

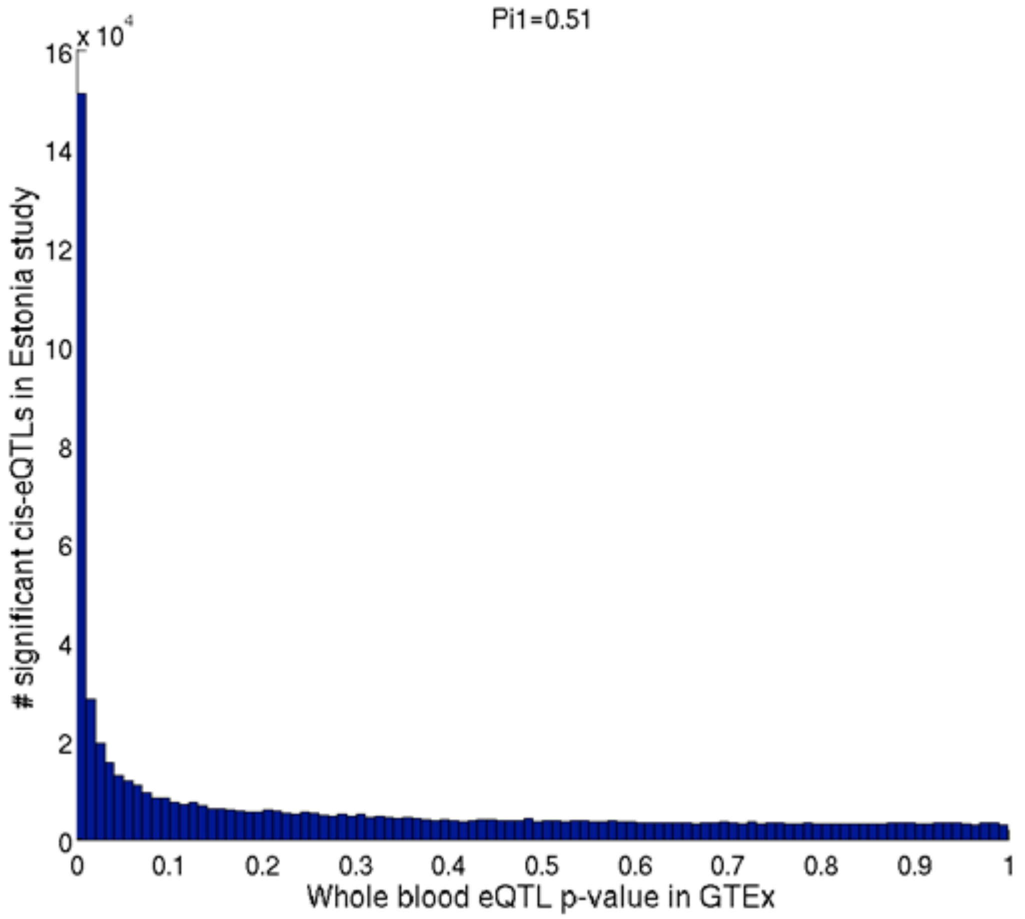


Fig. S12. **Heatmaps for expression and eQTL sharing** (**A**) Mean expression dendrogram and correlation heat map showing the pairwise similarities between mean gene-expression profiles of different tissues. The heatmap was constructed by computing the pairwise Spearman correlation coefficient between vectors of mean gene expression levels for each pair of tissues, and shows approximate correspondence with Fig. 2B. (**B**) Dendrogram and heatmap of eQTL sharing obtained from pairwise sharing of eQTLs inferred by appropriate marginalizing from the gene-based Bayesian multi-tissue eQTL model (*24*). Each element (*i,j*) is the probability that a gene has an eQTL in tissue *j*, given that the gene has an eQTL in tissue *i*.

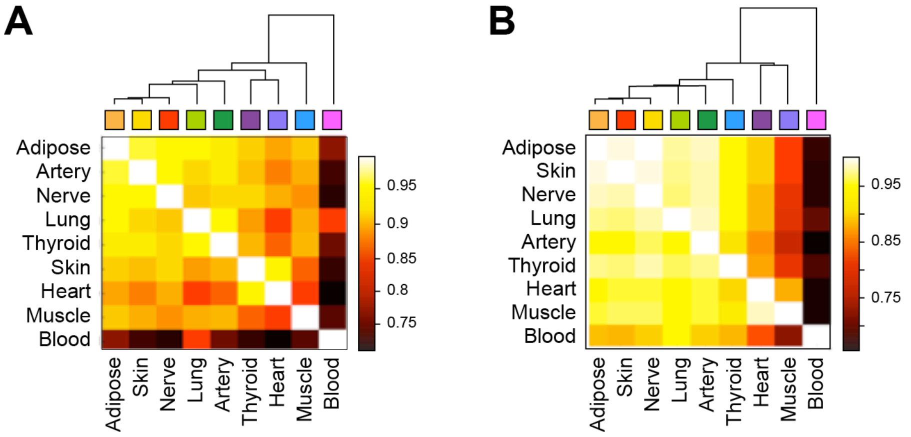


Fig. S13. **Pairwise sharing of eQTLs** (**A**) The *x*-axis shows the pairwise sharing for each pair of tissues (9 × 8 = 72 points) using the permutation-based method of Nica *et al*. (*22*), as shown in Fig. 2B. The *y*-axis shows the results from marginalization of the 9-tissue gene-based Bayesian model (*24*) to all tissue pairs, as shown in Figure S12B ($r = 0.63$ for the comparison). (**B**) The analogous plot for the SNP-based Bayesian model (*25*) ($r = 0.72$).

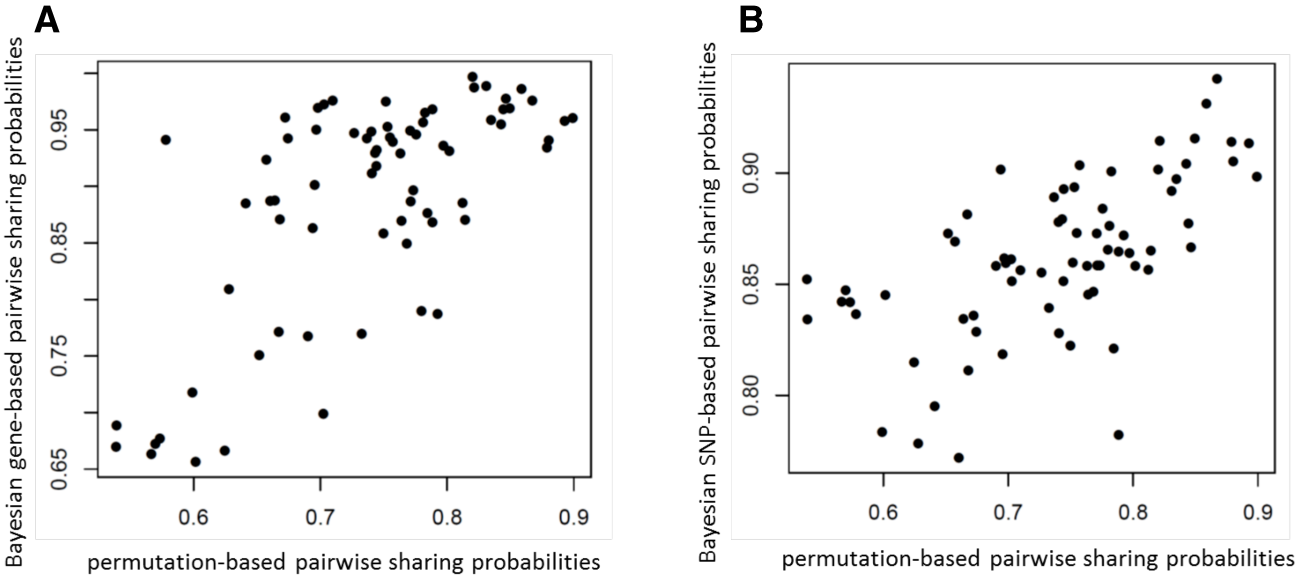


Fig. S14. Patterns of eQTL tissue-specificity and potential disease relevance. (**A**) Chromosome plot of *NDRG4* expression-genotype association in heart tissue, highlighting the SNP rs37062 in intron 40 of *CNOT1*, which is significantly associated with QT interval duration (*30*), and is in strong linkage disequilibrium with rs37055 (r^2=0.93), the GTEx *cis* eQTL identified with highest posterior probability for expression levels of *NDRG4*. (**B**) The G/A SNP rs37055 on chr16q21 is highly tissue specific as a local eQTL for *NDRG4* in Heart (left ventricle, r^2 = 0.32 for expression vs. genotype), with Bayesian model consensus posterior probability 0.98 for heart.

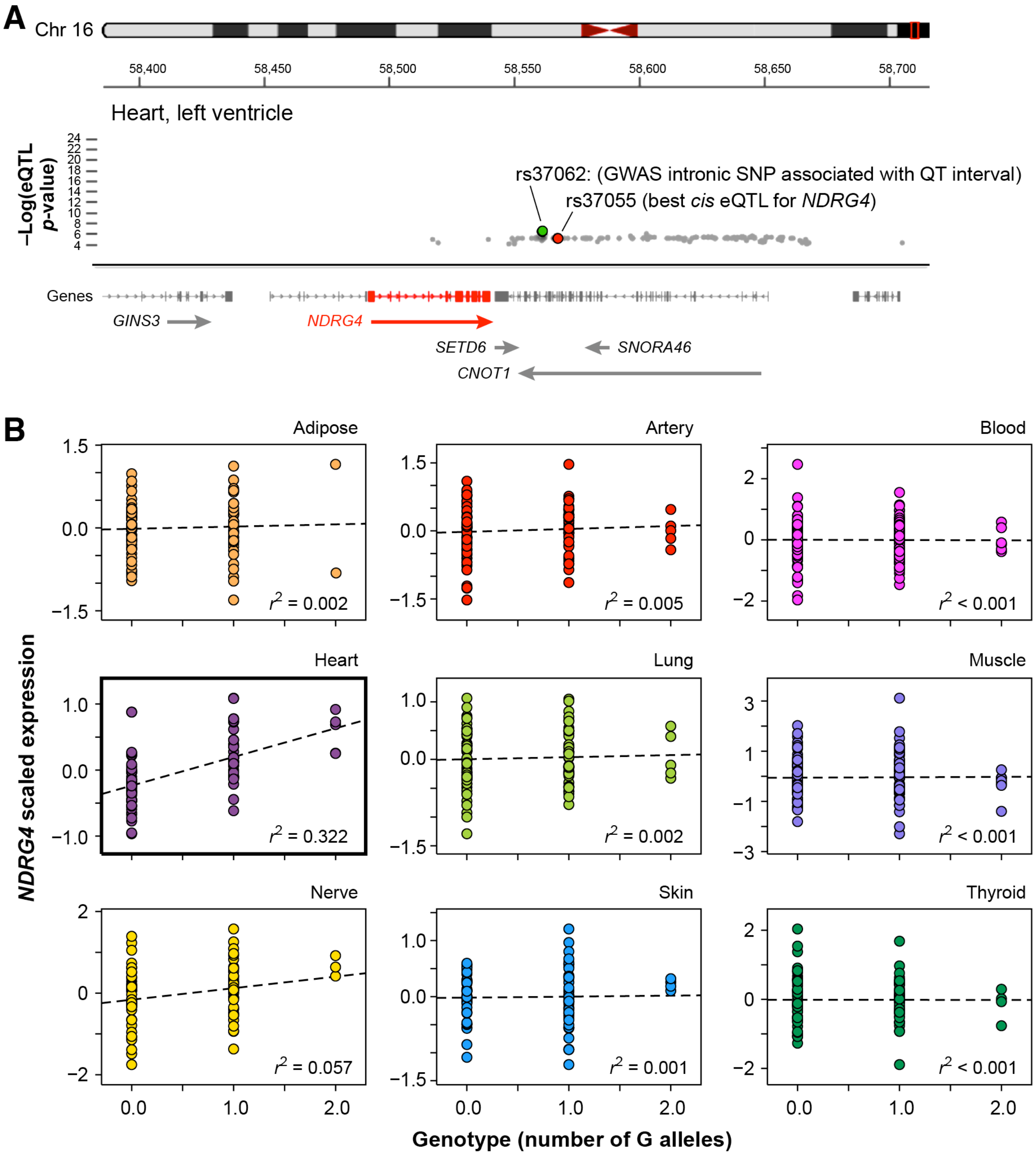
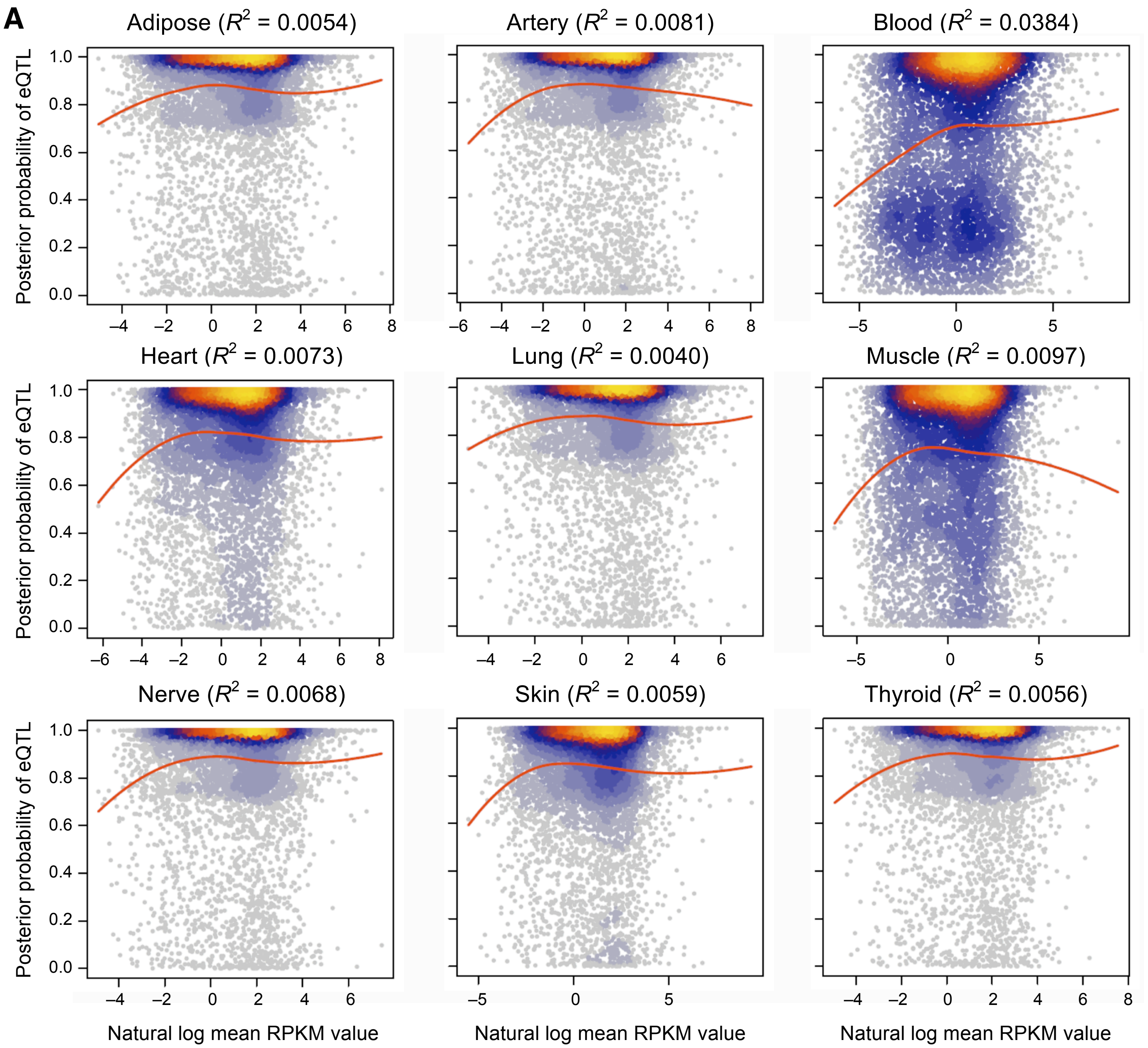
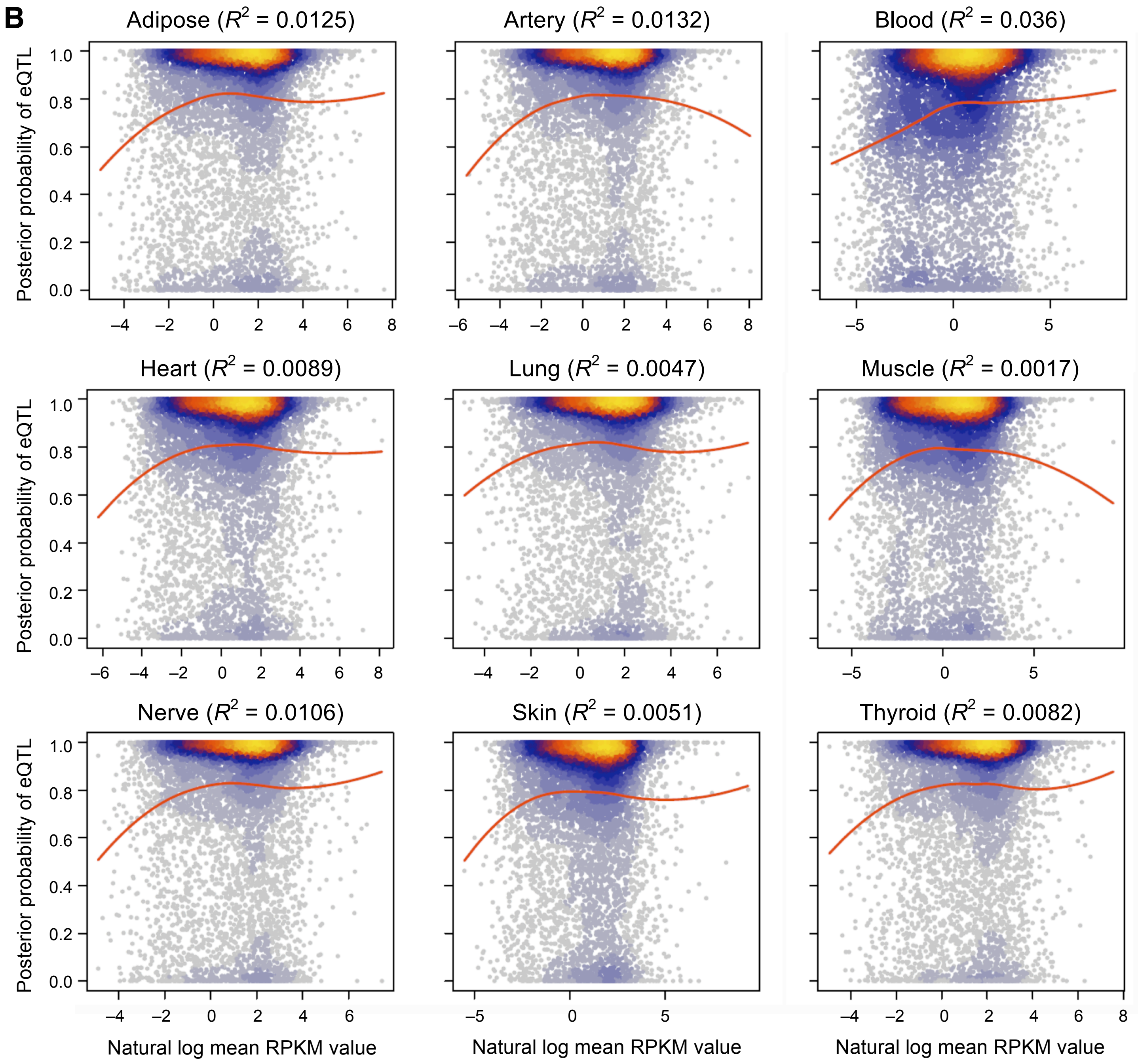
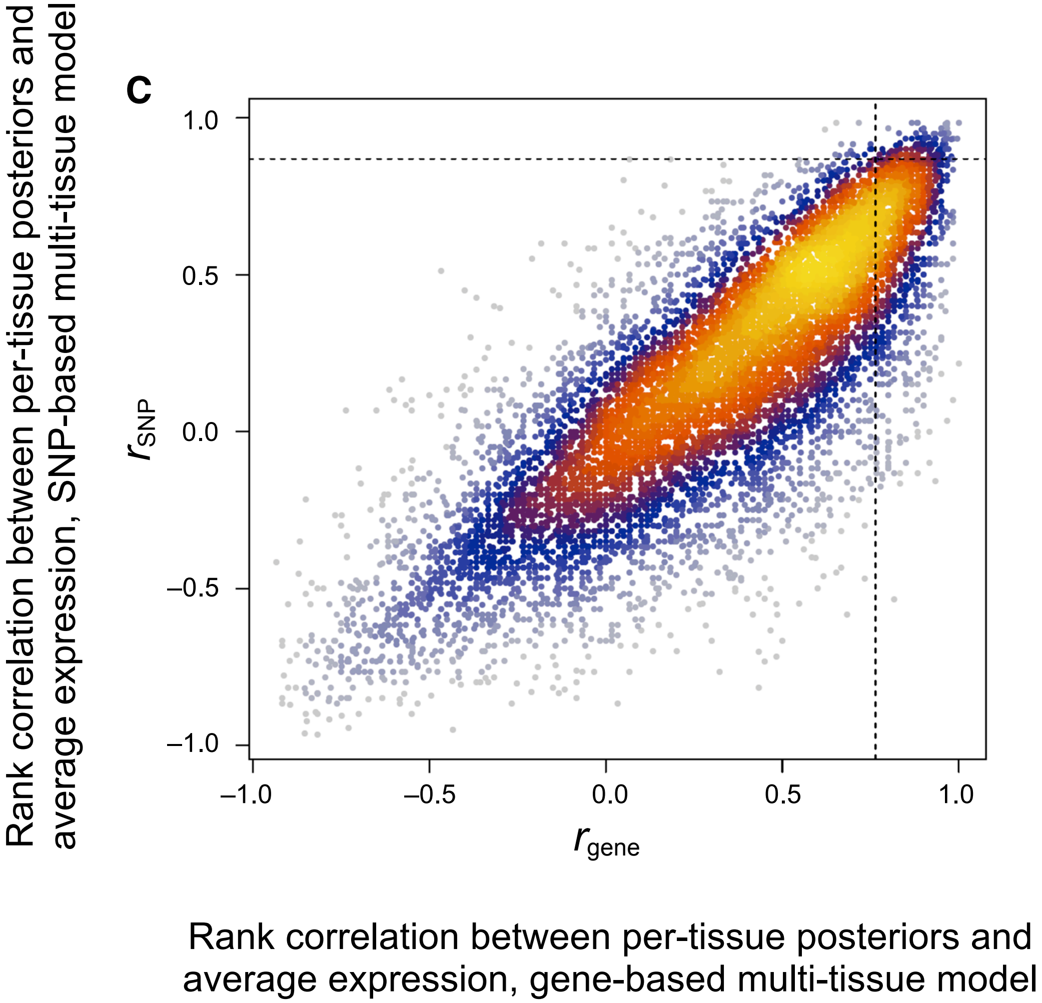
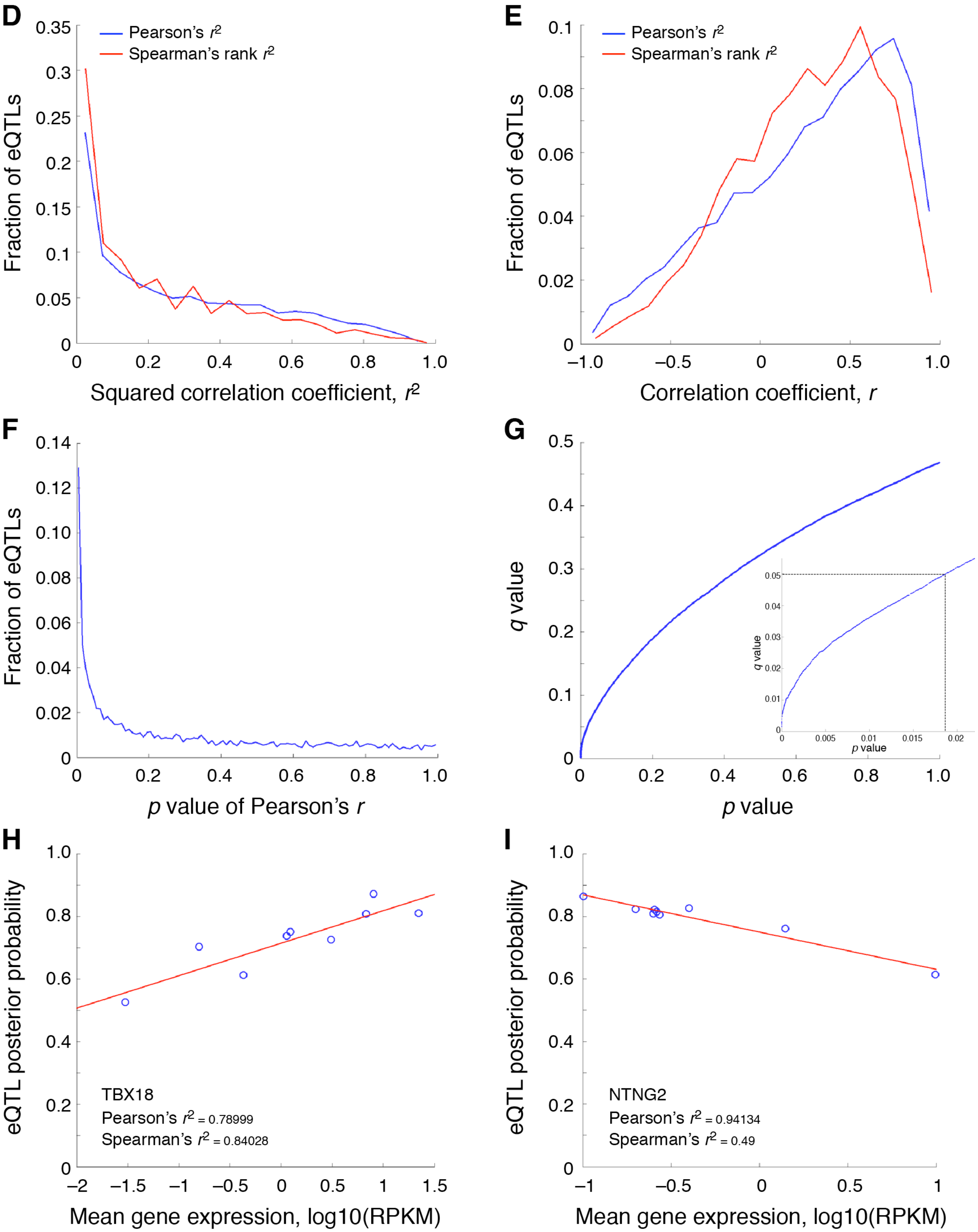


Fig. S15. Correlation between eQTL significance and the gene target's mean expression across 9 tissues. eQTL significance depends modestly on levels of expression. (**A**) For the 10,030 genes with a significant eQTL, marginal posterior probabilities from the nine-tissue Bayesian gene-based model are plotted vs. the expression level in the tissue. Loess curves (red) show very modest dependence, with most R^2 values <0.01. Beyond an initial rising relationship seen in all tissues, and only blood shows a consistently positive relationship across the range of expression. (**B**) The similar plots using the SNP-based Bayesian multi-tissue model, reaching the same conclusion. (**C**) For each gene, the rank correlation was computed between the nine tissue posteriors (both gene-based and SNP-based models) and the nine average expression levels per tissue, and the two sets of correlations (r_{gene} and r_{SNP}) and compared to each other. The average gene-based posterior vs. expression correlation was mean(r_{gene}) = 0.32, with 421 significantly correlated genes (FDR < 0.05, threshold shown as vertical dashed line). For the SNP-based posterior, the association was less strong, mean(r_{SNP}) = 0.19 and 49 significant genes (threshold shown as horizontal dashed line). No gene showed a significant negative correlation for either Bayesian multi-tissue model. (**D–N**) Linear correlation was computed between eQTL posterior probabilities (an average between two multi-tissue eQTL analysis methods used in this paper) and the mean expression of the eQTL's target gene across 9 tissues for 9875 genes that had a significant eQTL at FDR < 5% in at least one of 9 tissues tested. Only the most significant eQTL per gene were considered here. In all panels the mean gene expression values are on a log10 scale. Pearson's correlation coefficient (blue line in D–G) and Spearman's rank correlation coefficient (red line in D–E) were computed. (**D**) The density plot of the squared correlation coefficient (r^2) represents the fraction of variance of eQTL significance that can be explained by the mean expression profile of the eQTL's target gene across the 9 tissues. Based on this distribution, only a small fraction of eGenes (area under the curve) have a high r^2, i.e., for most eGenes it is not trivial to predict the eQTL tissue pattern solely from the target's mean expression profile (e.g., at FDR < 0.05, ~17% of eQTLs have a Pearson's $r^2 > 0.6$). (**E**) The density plot of the correlation coefficient (r) shows that there are many more positive correlations (~73%) between eQTL posterior probabilities and the gene target's mean expression levels than negative ones (~27%). (**F**) The distribution of Pearson's correlation p-values shows an excess of nominal p-values compared to an expected uniform distribution under the null (5.6-fold enrichment at $p < 0.05$). (**G**) Estimated q-values based on Storey and Tibshirani's method (*26*), are plotted as a function of Pearson's correlation p-values (Pearson's p-value cutoff at FDR < 0.05 is $p < 0.0186$; see insert). (**G–N**) We provided several representative examples (though not exhaustive) that demonstrate the different types of correlations observed between eQTL posterior probabilities and their target's mean expression levels: (**H**) strong positive correlation, **I)** strong negative correlation**,** (**J**) weak positive correlation, (**K**) weak negative correlation**,** (**L**) tissue-specific in lowest expressed tissue, (**M**) tissue-specific in highest expressed tissue, (**N**) constitutive eQTL irrespective of its target's expression levels (no correlation). This analysis is based on the assumption that a linear relationship exists between eQTL probabilities and their target gene expression levels, however in certain cases a nonlinear correlation may better capture the relationship between eQTLs and their target expression levels.


A
Adipose (R^2 = 0.0054)
Artery (R^2 = 0.0081)
Blood (R^2 = 0.0384)
Heart (R^2 = 0.0073)
Lung (R^2 = 0.0040)
Muscle (R^2 = 0.0097)
Nerve (R^2 = 0.0068)
Skin (R^2 = 0.0059)
Thyroid (R^2 = 0.0056)
Posterior probability of eQTL
1.0
0.8
0.6
0.4
0.2
0.0
−6
−5
−4
−2
0
2
4
5
6
8
Natural log mean RPKM value


B
Adipose (R^2 = 0.0125)
Artery (R^2 = 0.0132)
Blood (R^2 = 0.036)
Heart (R^2 = 0.0089)
Lung (R^2 = 0.0047)
Muscle (R^2 = 0.0017)
Nerve (R^2 = 0.0106)
Skin (R^2 = 0.0051)
Thyroid (R^2 = 0.0082)
Posterior probability of eQTL
1.0
0.8
0.6
0.4
0.2
0.0
–4 –2 0 2 4 6 8
–6 –4 –2 0 2 4 6 8
–5 0 5
–6 –4 –2 0 2 4 6 8
–4 –2 0 2 4 6
–5 0 5
–4 –2 0 2 4 6
–5 0 5
–4 –2 0 2 4 6 8
Natural log mean RPKM value


C
1.0
0.5
0.0
−0.5
−1.0
r_{SNP}
−1.0
−0.5
0.0
0.5
1.0
r_{gene}
Rank correlation between per-tissue posteriors and average expression, SNP-based multi-tissue model
Rank correlation between per-tissue posteriors and average expression, gene-based multi-tissue model


D
Pearson's r^2
Spearman's rank r^2
Fraction of eQTLs
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1.0
Squared correlation coefficient, r^2
E
Pearson's r^2
Spearman's rank r^2
Fraction of eQTLs
0.1
0.08
0.06
0.04
0.02
0
−1.0
−0.5
0
0.5
1.0
Correlation coefficient, r
F
Fraction of eQTLs
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
0
0.2
0.4
0.6
0.8
1.0
p value of Pearson's r
G
q value
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1.0
p value
0.05
0.04
0.03
0.02
0.01
0
q value
0
0.005
0.01
0.015
0.02
p value
H
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
−2
−1.5
−1
−0.5
0
0.5
1
1.5
Mean gene expression, log10(RPKM)
TBX18
Pearson's r^2 = 0.78999
Spearman's r^2 = 0.84028
I
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
−1
−0.5
0
0.5
1
Mean gene expression, log10(RPKM)
NTNG2
Pearson's r^2 = 0.94134
Spearman's r^2 = 0.49

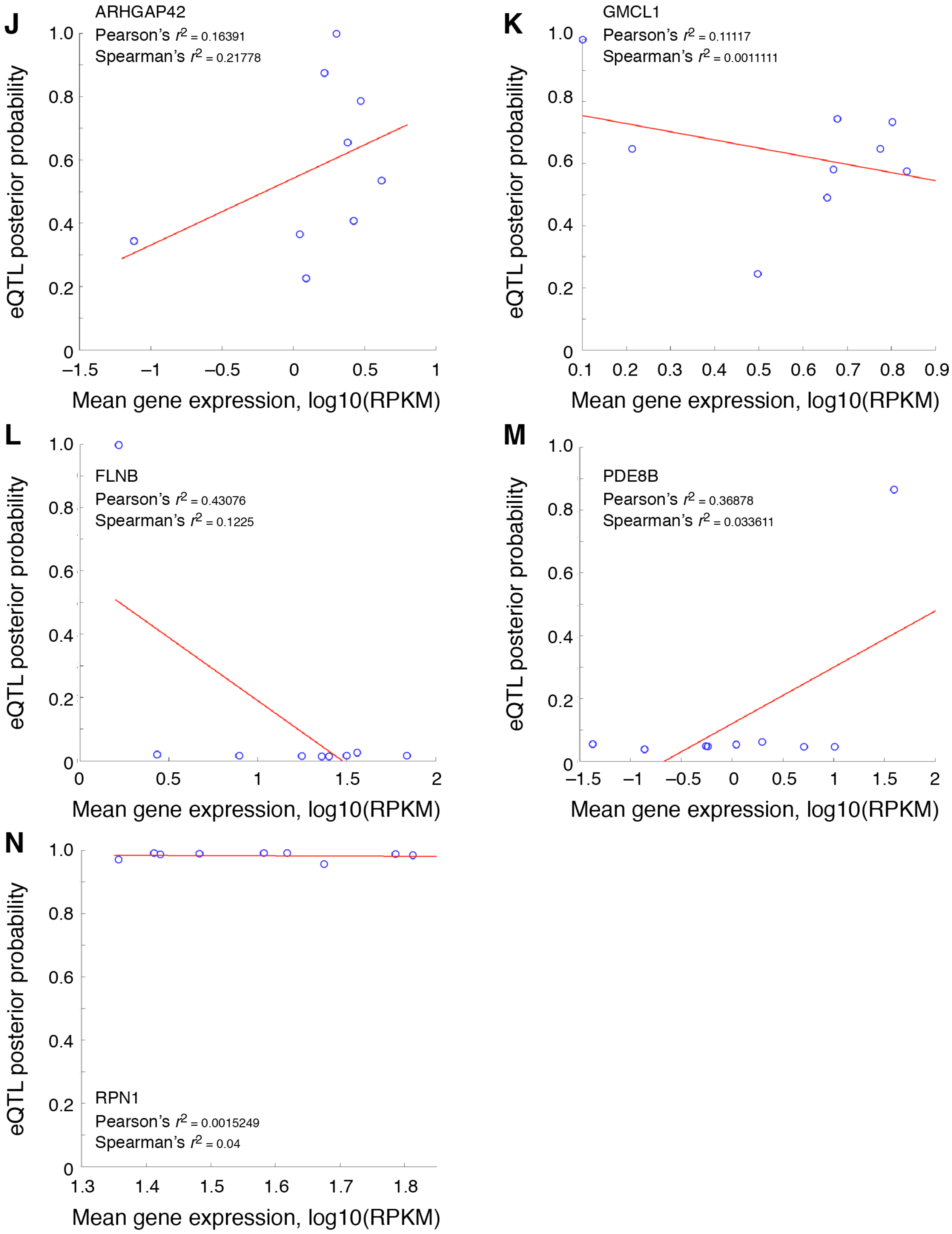

J
ARHGAP42
Pearson's r^2 = 0.16391
Spearman's r^2 = 0.21778
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
–1.5
–1
–0.5
0
0.5
1
Mean gene expression, log10(RPKM)
K
GMCL1
Pearson's r^2 = 0.11117
Spearman's r^2 = 0.0011111
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Mean gene expression, log10(RPKM)
L
FLNB
Pearson's r^2 = 0.43076
Spearman's r^2 = 0.1225
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
0
0.5
1
1.5
2
Mean gene expression, log10(RPKM)
M
PDE8B
Pearson's r^2 = 0.36878
Spearman's r^2 = 0.033611
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
–1.5
–1
–0.5
0
0.5
1
1.5
2
Mean gene expression, log10(RPKM)
N
RPN1
Pearson's r^2 = 0.0015249
Spearman's r^2 = 0.04
eQTL posterior probability
1.0
0.8
0.6
0.4
0.2
0
1.3
1.4
1.5
1.6
1.7
1.8
Mean gene expression, log10(RPKM)

Fig. S16. Bayesian model analysis of tissue sharing of eQTLs. Illustration of the Bayesian models applied to pairwise analysis of whole blood vs. skeletal muscle. (**A**) For each gene, the SNP declared by the Bayesian gene-based model as most likely to be “causal” is displayed in terms of the *z*-statistic for association of expression to genotype in each of the two tissues. Only SNPs for genes with posterior probability >0.95 of having an eQTL are shown, producing the open region near the origin for genes non-significant by this criterion. Using I_{blood}=1 as an indicator to represent eQTL status in blood (= 0 otherwise), and similarly for the indicator I_{muscle}, the model provides explicit probabilities for the four possible outcomes of $\{I_{blood,}\ I_{muscle}\}$. Genes are declared tissue-specific (i.e., the SNP is an eQTL for one tissue and not the other), if the posterior probability for either of the outcomes {0,1} (musce-specific) or {1,0} (blood-specific) is greater than for the other possibilities, and tissue-common if the posterior for {1,1} is greatest. The color scheme in the figure shows the results among significant eQTLs. (**B**) The analogous Bayesian inference using the SNP-based model. The model uses a slight modification of the *z*-statistics (*25*) and has many more plotted values, because each significant SNP is shown. However, otherwise the statistics and the resulting inference are highly concordant between the models. Note, for both models, the inference regarding “opposite-effect” eQTLs, many of which may be artifactual results of regional linkage disequilibrium patterns, as illustrated in more detail in Fig. S17. However, the overall trend for both gene-based inference and SNP-based inference is a strong positive correlation of findings between tissues, and examination of the estimated model parameters indicates the correlation is largely due to underlying commonality of eQTLs rather than overlapping subjects.

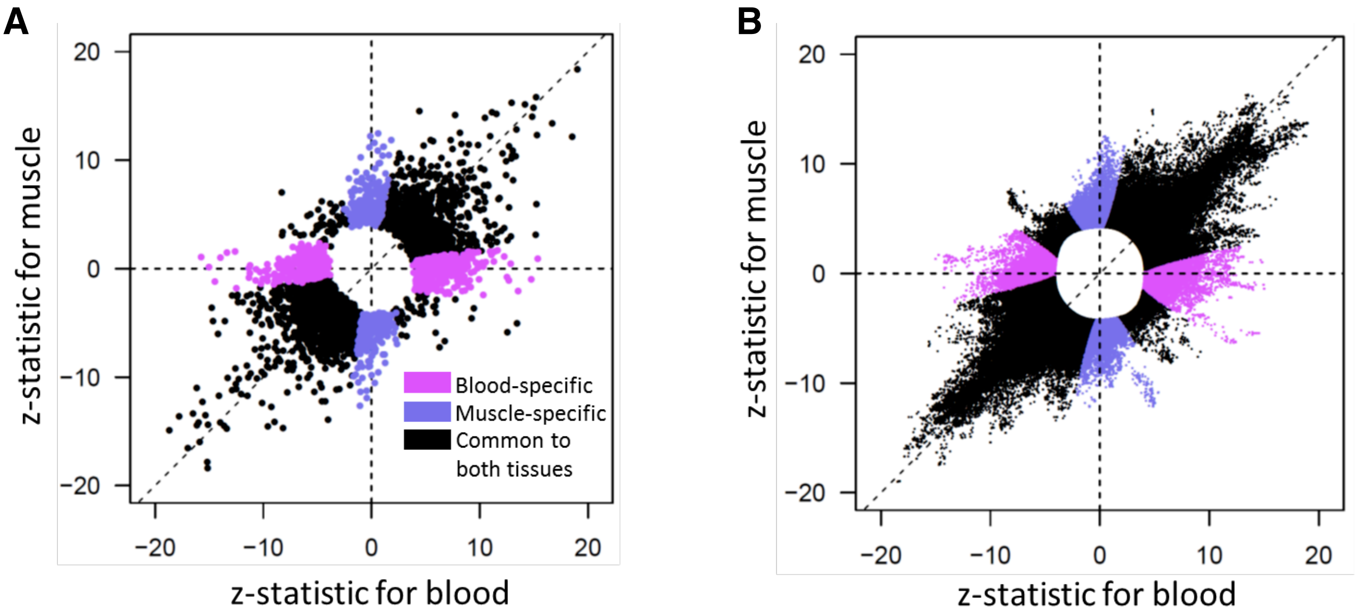


Fig. S17. Opposite effect eQTLs. Example of how multiple eQTLs may produce spurious signals of apparent ``opposite effects” in SNP-by-SNP analyses. (**A**) Results from Bayesian multi-SNP analysis of gene ENSG00000167528 for two tissues (whole blood and skeletal muscle) indicate two different tissue-specific eQTLs in this region, one in each tissue. The *x*-axis labels the positions of interrogated SNPs relative to TSS; the *y*-axis shows the posterior probability assigned to each SNP being a casual eQTN, color-coded according to tissue specificity. (Due to LD, there is uncertainty about which SNP is the causal eQTN in each tissue — no posterior probability is close to 1, but the sums of the posterior probabilities from the blood-specific eQTLs and the skeletal muscle-specific eQTLs are both very close to 1.) (**B**) Plot of pairwise LD (r^2) shows two blocks of SNPs: the SNPs showing strong signal in blood are in high LD, and the SNPs showing strong signal in skeletal muscle are in high LD. The two clusters are also moderately correlated ($r^2 \sim 0.4$). (**C**) Estimated effect sizes in each tissue from a SNP-by-SNP analysis for the top SNP in each tissue in results from top panel. The SNPs show a significant effect in both tissues, but with effects in the opposite direction because the allele increasing expression in blood is positively correlated with the allele decreasing expression in skeletal muscle. (**D**) Effect size estimates from a standard linear regression of expression in each tissue on the two top SNPs. This joint analysis accounts for the modest LD between the two SNPs, and the apparent opposite effects from the SNP-by-SNP analysis disappear. Taken together, the results suggest that this region contains two eQTLs, one active in blood and the other active in skeletal muscle, and that LD between the two eQTLs produces a spurious pattern of opposite effects in the SNP-by-SNP analysis.

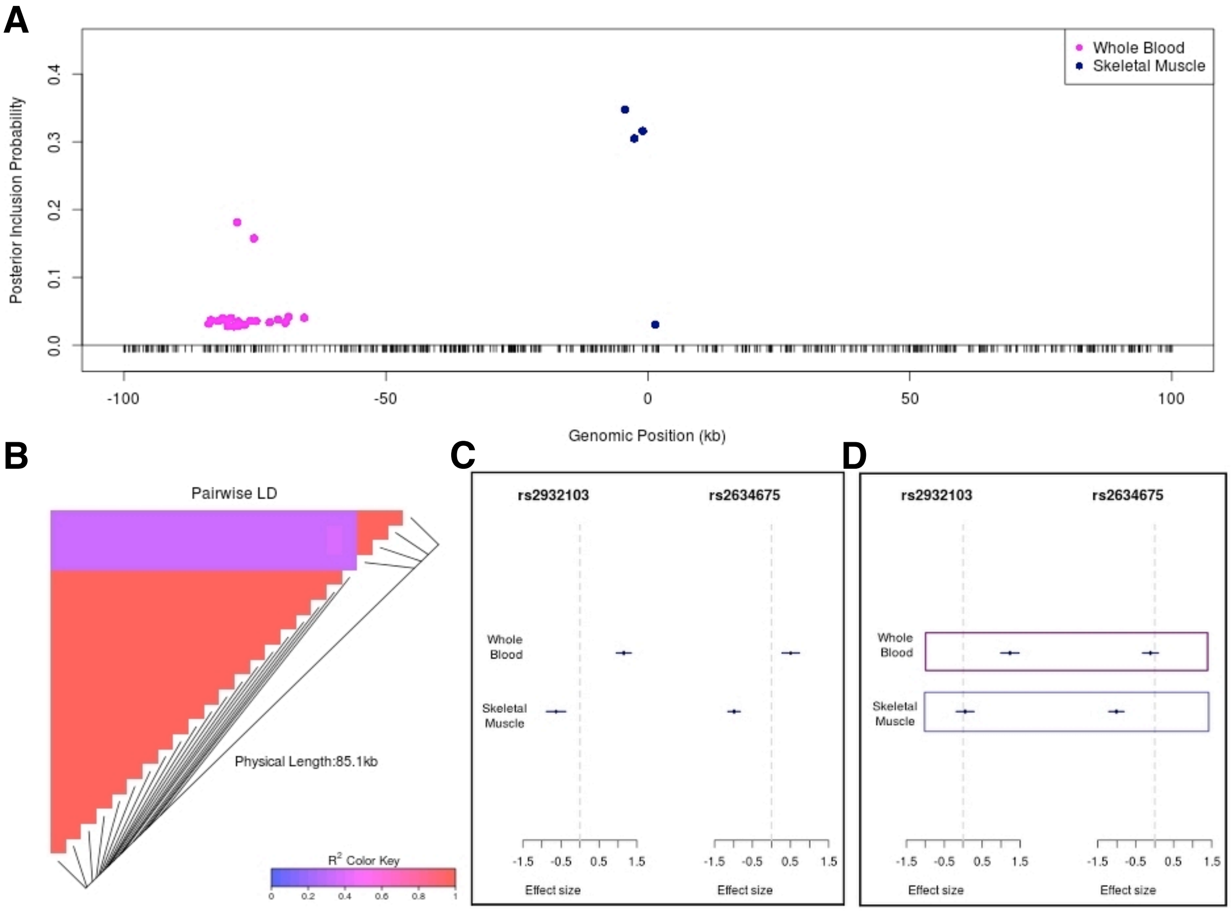


Fig. S18. Basic statistics for the allele-specific expression data. The number of measured sites per sample in ASE analysis for ≥8 reads as in the master data (**A, C**) and for ≥ 30 reads as in most ASE analyses (**B, D**) as histograms of total site count (**A, B**) and as scatterplots showing also the number of sites with significant ASE ($p < 0.005$; **C, D**).

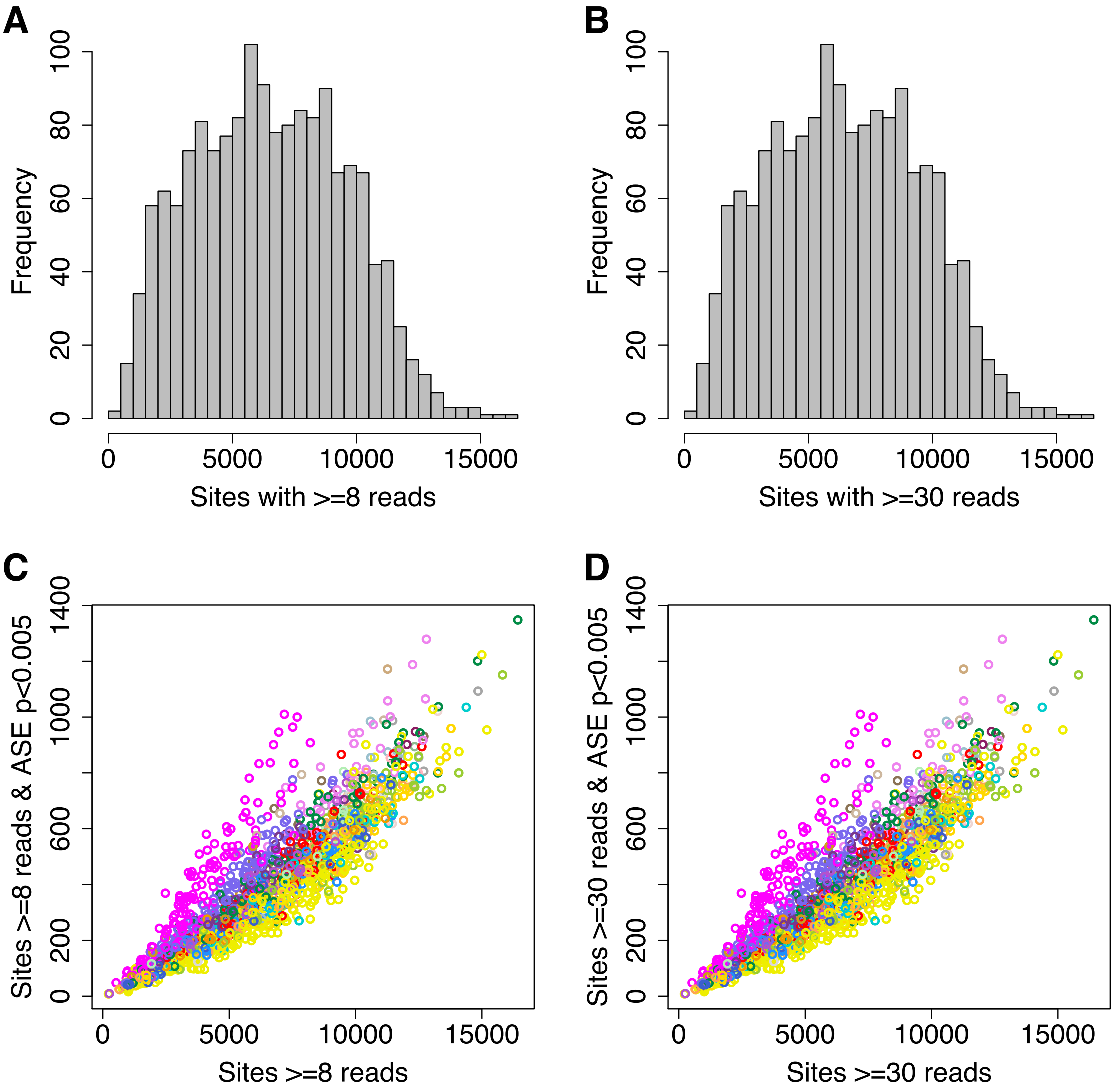


Fig. S19. **Tissue relatedness in ASE.** From the pairwise Spearman rank correlations between all the samples using allelic ratios over sites sampled to exactly 30 reads each (**A**), and total read counts over the same sites (**B**), we calculated median correlation between tissues. These statistics capture similarity in allelic expression and total gene expression levels, respectively. The two matrices are highly correlated (Mantel test, *r* = 0.766), which shows that while tissue-specific patterns in allelic expression are weaker, it captures similar, biologically sound tissue relationships as gene expression levels.

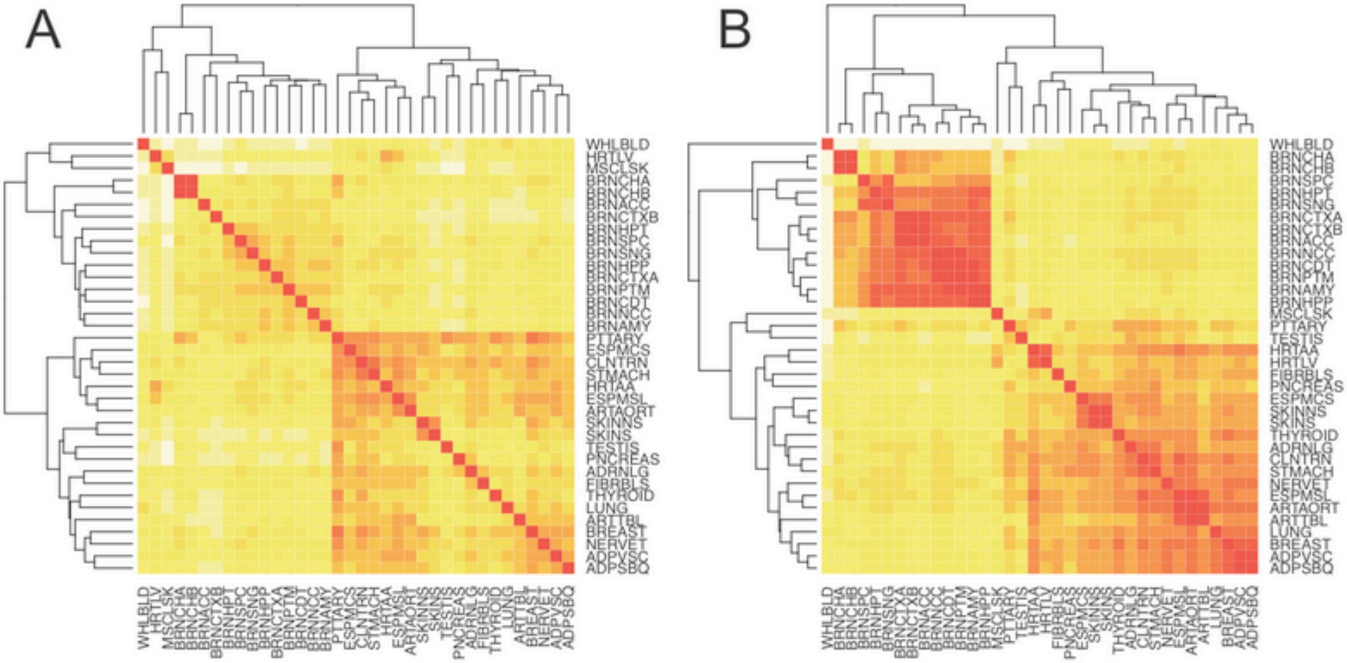


Fig. S20. **eQTL replication by ASE analysis across tissues.** The plots show the activity of eQTLs discovered in each of the 9 tissues, measured as the odds ratio of observing significant ASE in eQTL heterozygote individuals and compared to the null of homozygotes.

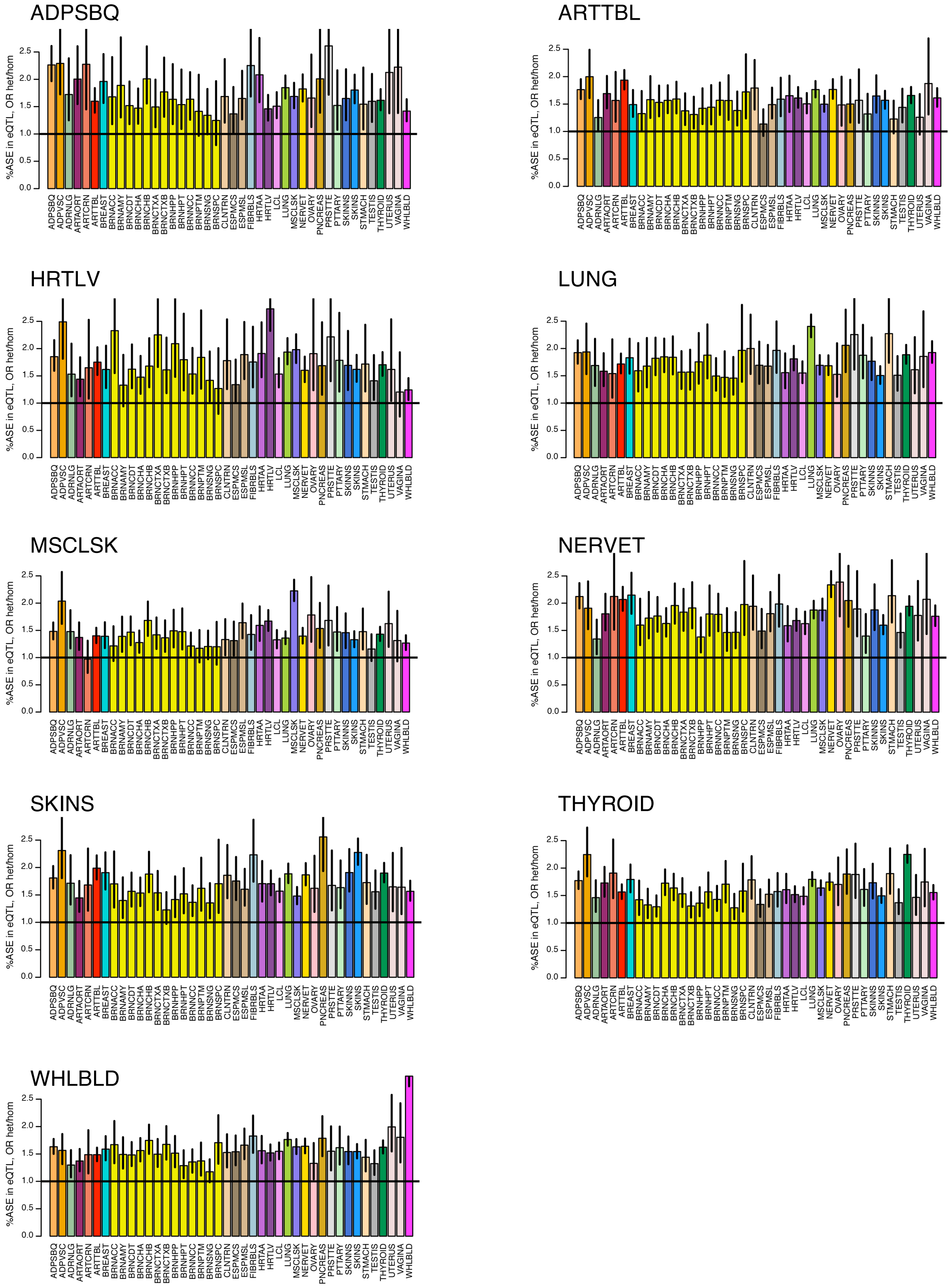


Fig. S21. eQTL replication by ASE analysis across tissues. This figure shows the same statistic as in Figure S20, plotted here as a heat map using (**A**) all estimated and (**B**) those that are significantly different from 1 ($p < 0.05$ with Bonferroni correction).

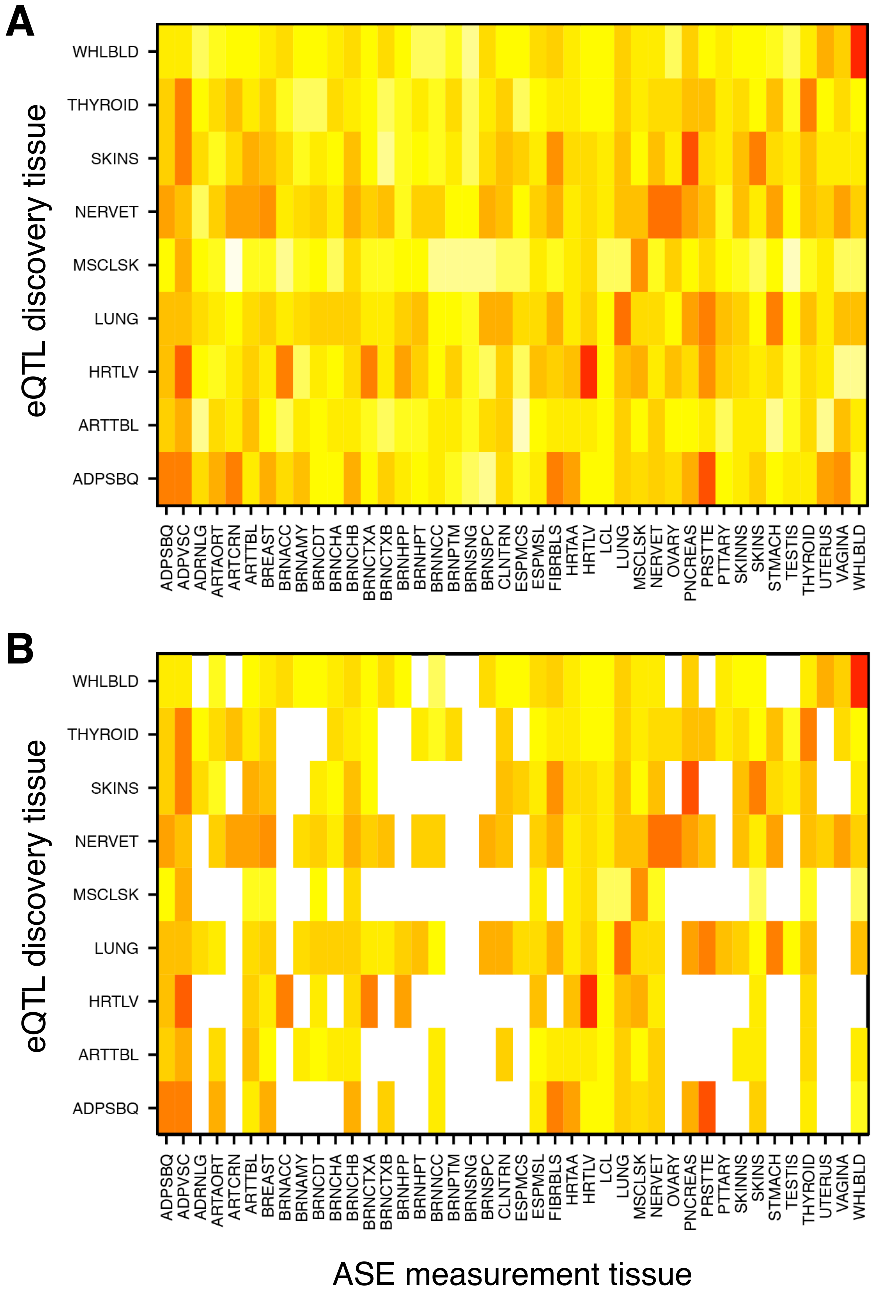


Fig. S22. Quantifying sharing of regulatory effects between tissues by ASE analysis. Each row shows how significant ASE sites ($p < 0.005$) discovered in each of the tissues are detected in other tissues of the same individual. (**A**) Shows the proportion of ASE sites in one tissue which are nominally significant ($p < 0.05$) in the second tissue, for only the sites that are measured in both tissues as a result of the gene being expressed in both. The relatively high degree of sharing is consistent with eQTL results. (**B**) Shows the proportion of ASE sites that are nominally significant ($p < 0.05$) in the second tissue, out of all the sites in the first tissue that were measured in the second tissue. This captures the total probability of detecting a regulatory effect in another tissue, and the difference between A and B shows that while regulatory variants can have tissue-specific effects even in ubiquitously expressed genes, being unable to find an effect of a regulatory variant in another tissue is dominated by the gene simply not being expressed in the other tissue.

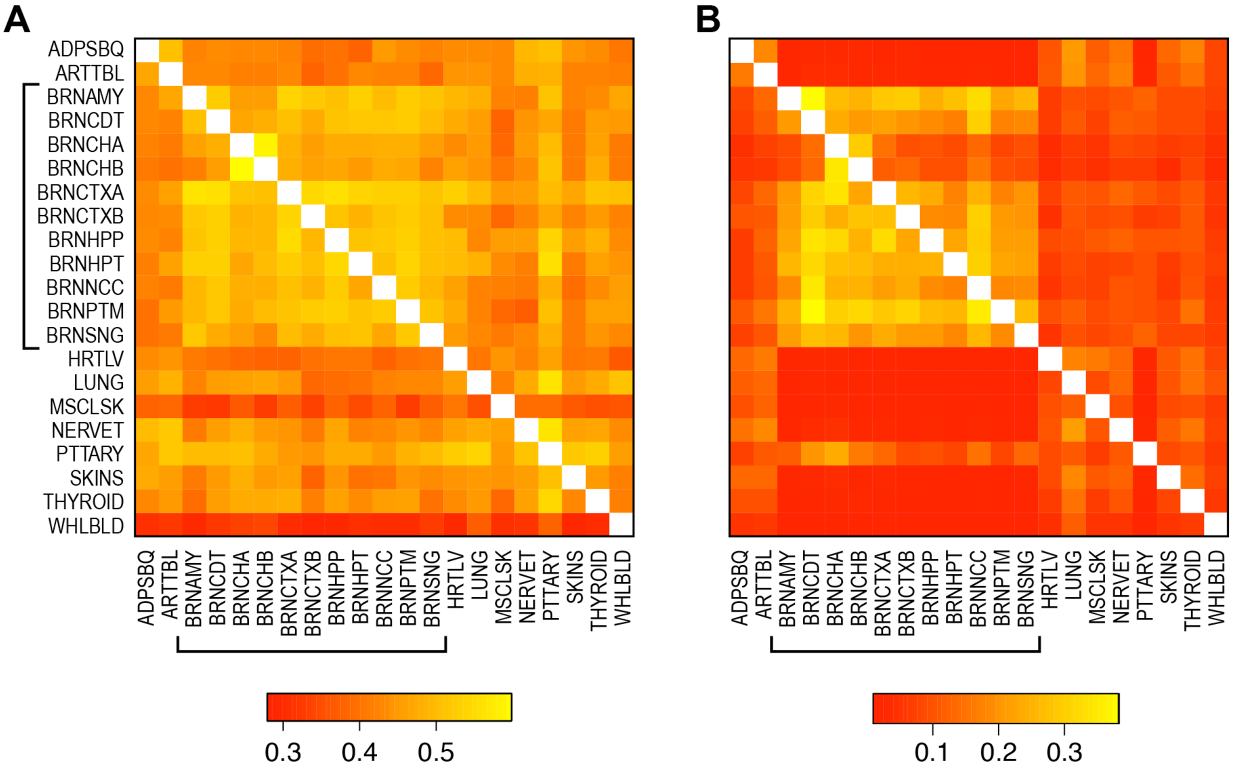


Fig. S23. Cartoon definitions of splicing events analyzed.

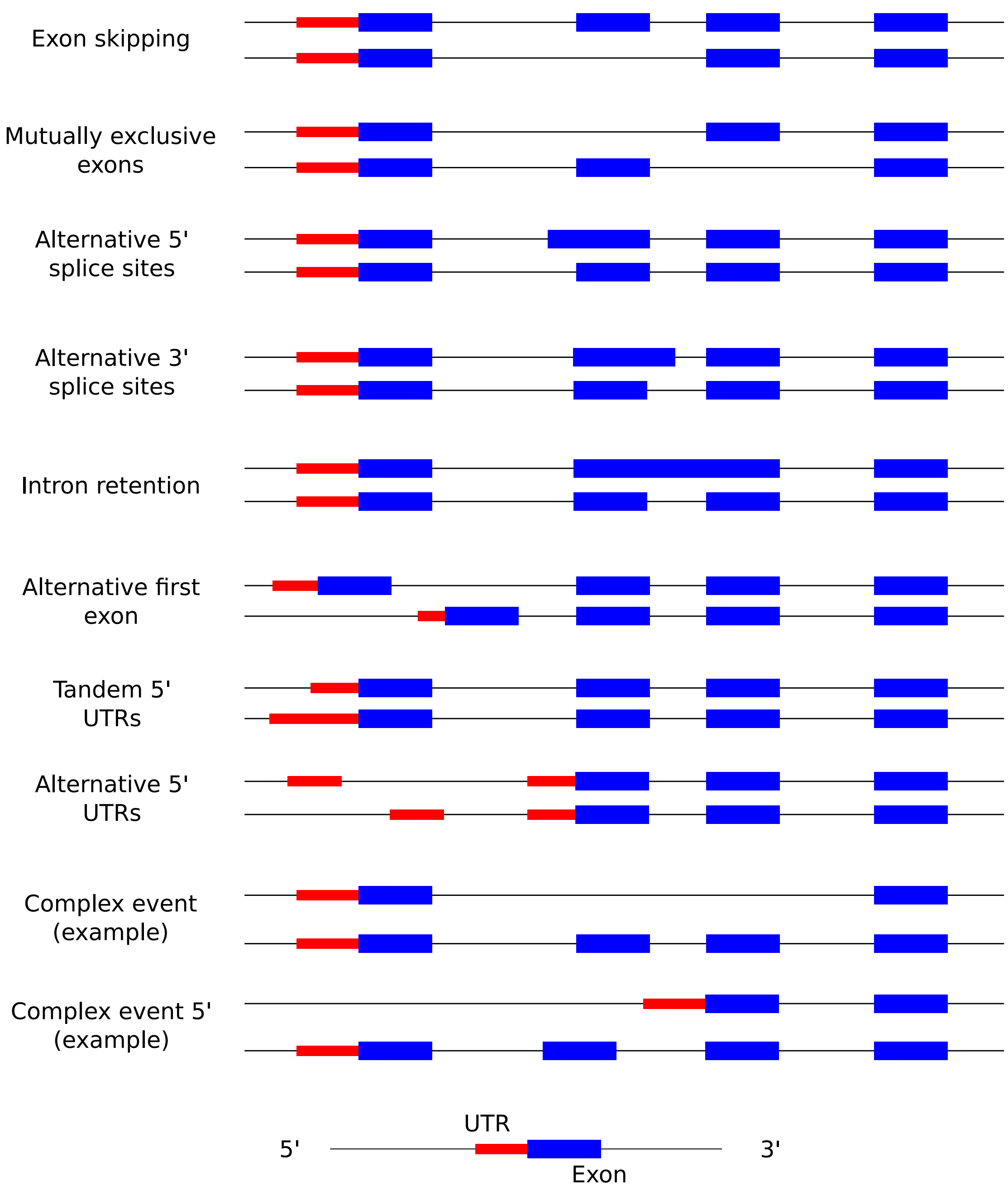


Fig. S24. Expression profiles of significant (FDR 5%) sQTLs identified by Altrans and sQTLseekeR. Plotted are the density functions of log10 (RPKM) distributions of different classes of genes. The genes that are significant in Altrans (orange line) (*35*) are significantly more expressed (Mann Whitney U $P < 2.2 \times 10^{-16}$) than those that are found to be significant in sQTLseekeR (magenta line) (*36*). This reflects the fact that Altrans quantifications are filtered for highly expressed links specifically to filter out noise in quantifications, whereas sQTLseekeR quantifications are not filtered (*14)*, which is one of the main differences between the two methods.

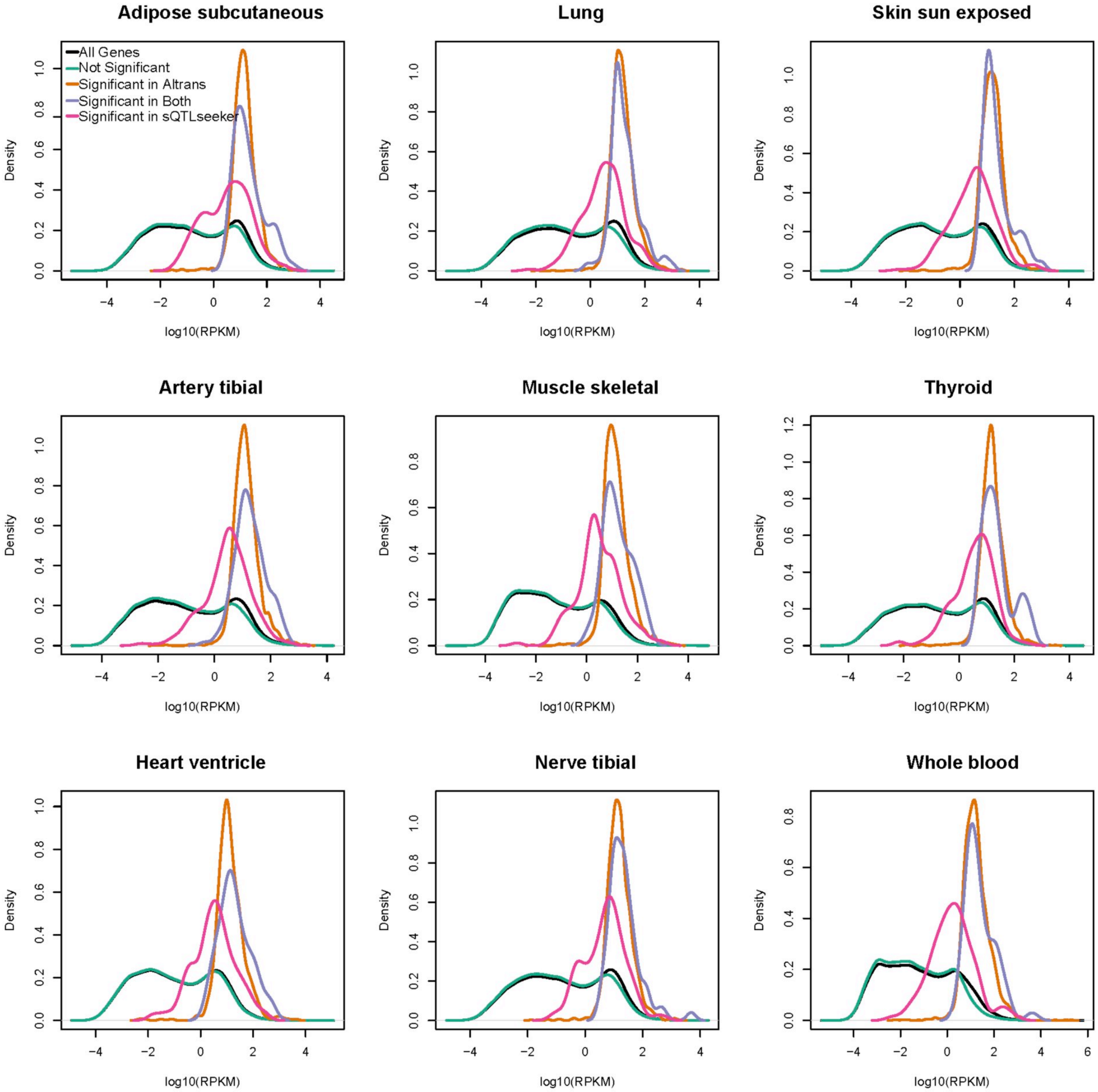


Fig. S25. Concordance between Altrans and sQTLseekeR. π1 estimates of the concordance between Altrans and sQTLseekeR considering only exon-skipping events. We examined cases where there was a significant variant for an exon-skipping event in sQTLseekeR and compared the *p*-value achieved for the same exon–variant pair in Altrans. We find a significant enrichment of low *p*-values indicating strong agreement between methods when overlapping types of splicing events are considered.

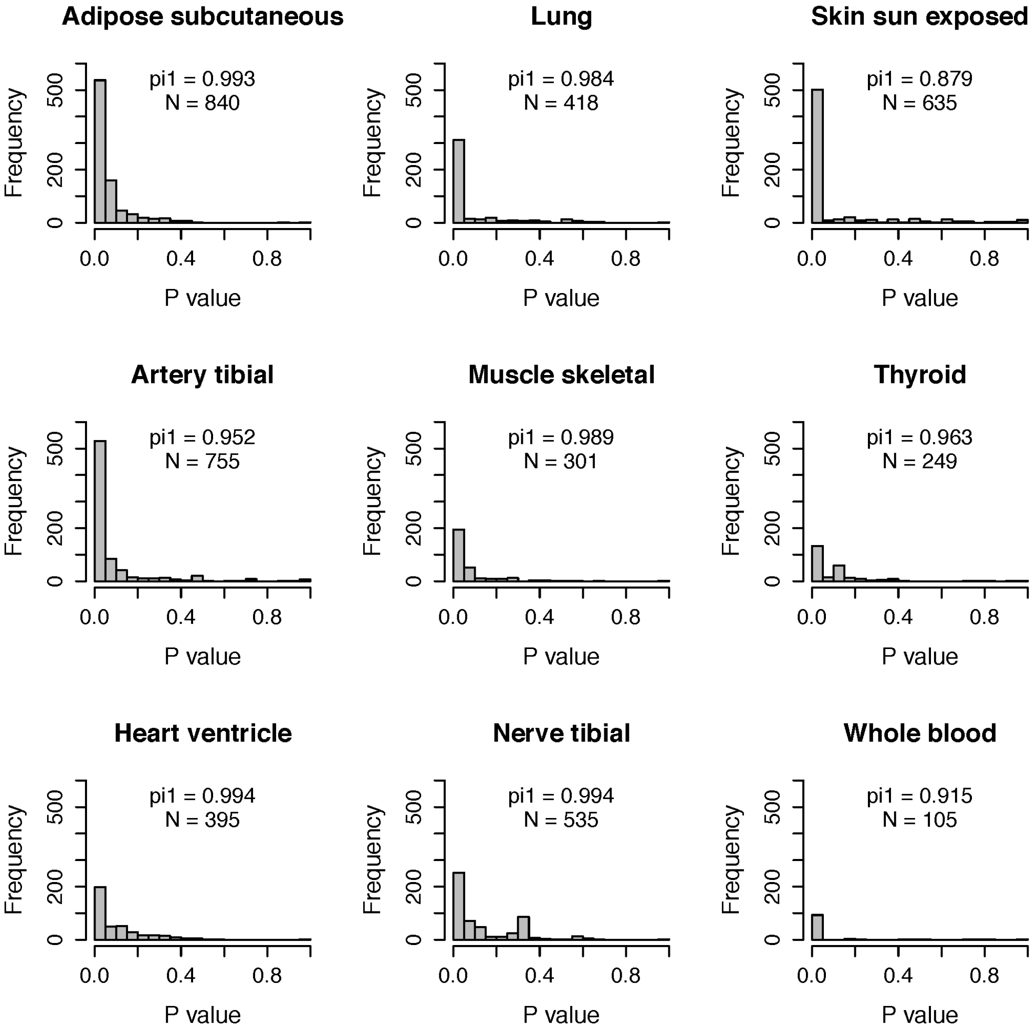


Fig. S26. Examples of sQTLs. In the left panels we represent as box-plots the distribution of the splicing ratios of the candidate gene in a selected population. The distributions are given separately for each genotype and the number of tested individuals in each genotype is given in parenthesis next to the genotype. Each splicing isoform is represented by a different color. When there are more than six transcripts, those with low expression levels are merged into a single transcript for the sake of clarity. The right panels show the exonic structure of the transcripts along with the location of the sQTL SNP (shown as the dotted line). (**A**) In blood, we found the SNP rs3865444 associated with a change of the major isoform of CD33 gene. This gene has four annotated isoforms. In the individuals with the CC genotype, the green isoform dominates, while the blue isoform is the dominant one in the individuals with the AA genotype. Heterozygotes use the two isoforms with similar frequency. SNP rs3865444 has been previously associated with Alzheimer disease but no eQTLs have been found. The SNP, however, has been recently found to be a sQLT in Battle *et al*. (*39*) (**B**) Shows an example in adipose tissue. The SNP rs116179804 is associated with changes in the relative usage of isoforms of the gene MICA, which has seven annotated isoforms. The same association was observed in 6 other tissues. This SNP is also associated with changes in gene expression in the GTEx data, but had not previously been described as an eQTL.

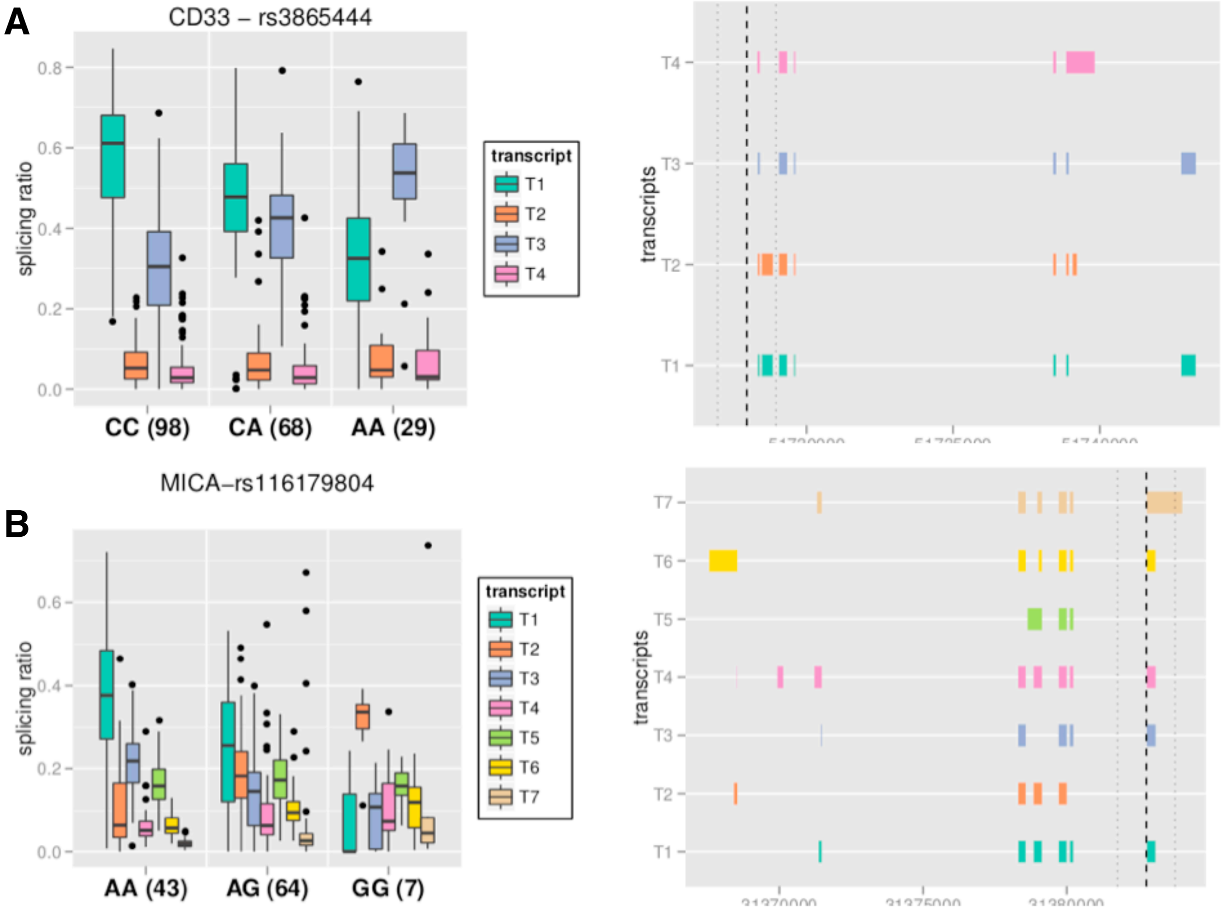
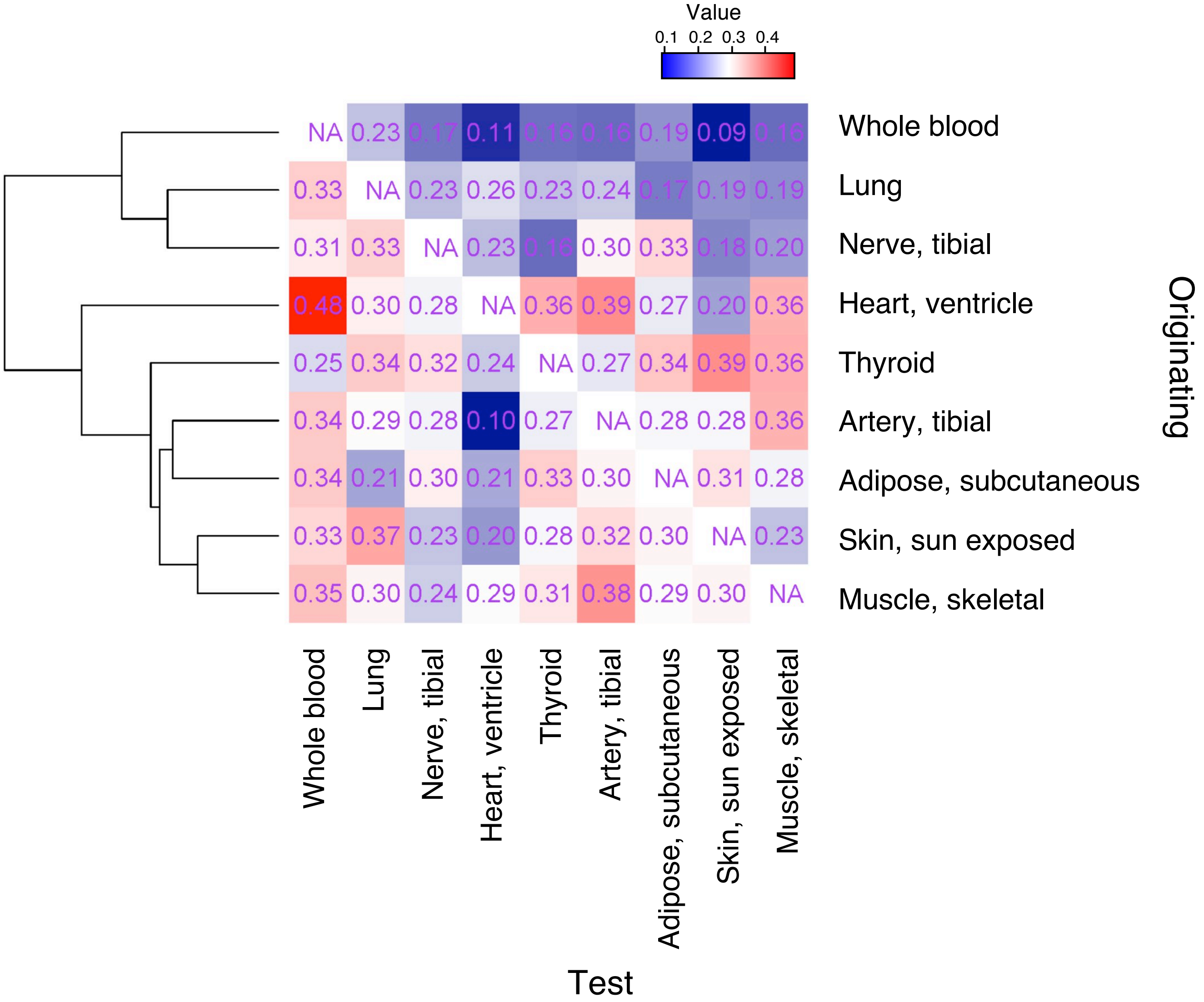


Fig. S27. **π1 estimates of tissue sharing for Altrans sQTLs.** Each pairwise estimate is plotted as a heat map, with the numbers in the boxes showing the actual π1 estimate. The tree represents a hierarchical clustering of the samples based on these estimates. The π1 estimates are based on the *p*-value distributions obtained by testing the significance of originating tissues' significant SNP-link pairs in the test tissue.



Originating \ Test	Whole blood	Lung	Nerve, tibial	Heart, ventricle	Thyroid	Artery, tibial	Adipose, subcutaneous	Skin, sun exposed	Muscle, skeletal
Whole blood	NA	0.23	0.17	0.11	0.16	0.16	0.19	0.09	0.16
Lung	0.33	NA	0.23	0.26	0.23	0.24	0.17	0.19	0.19
Nerve, tibial	0.31	0.33	NA	0.23	0.16	0.30	0.33	0.18	0.20
Heart, ventricle	0.48	0.30	0.28	NA	0.36	0.39	0.27	0.20	0.36
Thyroid	0.25	0.34	0.32	0.24	NA	0.27	0.34	0.39	0.36
Artery, tibial	0.34	0.29	0.28	0.10	0.27	NA	0.28	0.28	0.36
Adipose, subcutaneous	0.34	0.21	0.30	0.21	0.33	0.30	NA	0.31	0.28
Skin, sun exposed	0.33	0.37	0.23	0.20	0.28	0.32	0.30	NA	0.23
Muscle, skeletal	0.35	0.30	0.24	0.29	0.31	0.38	0.29	0.30	NA

Fig. S28. Scatter plot of sQTL variant distance to the transcription start site (TSS) vs. FDR corrected –log10(*P*). The points are colored based on the number of tissues a variant is significant in. There is a strong enrichment for sQTLs that are closer to the TSS. Further, sQTLs that are observed in multiple tissues are closer to the TSS than those that are tissue specific.

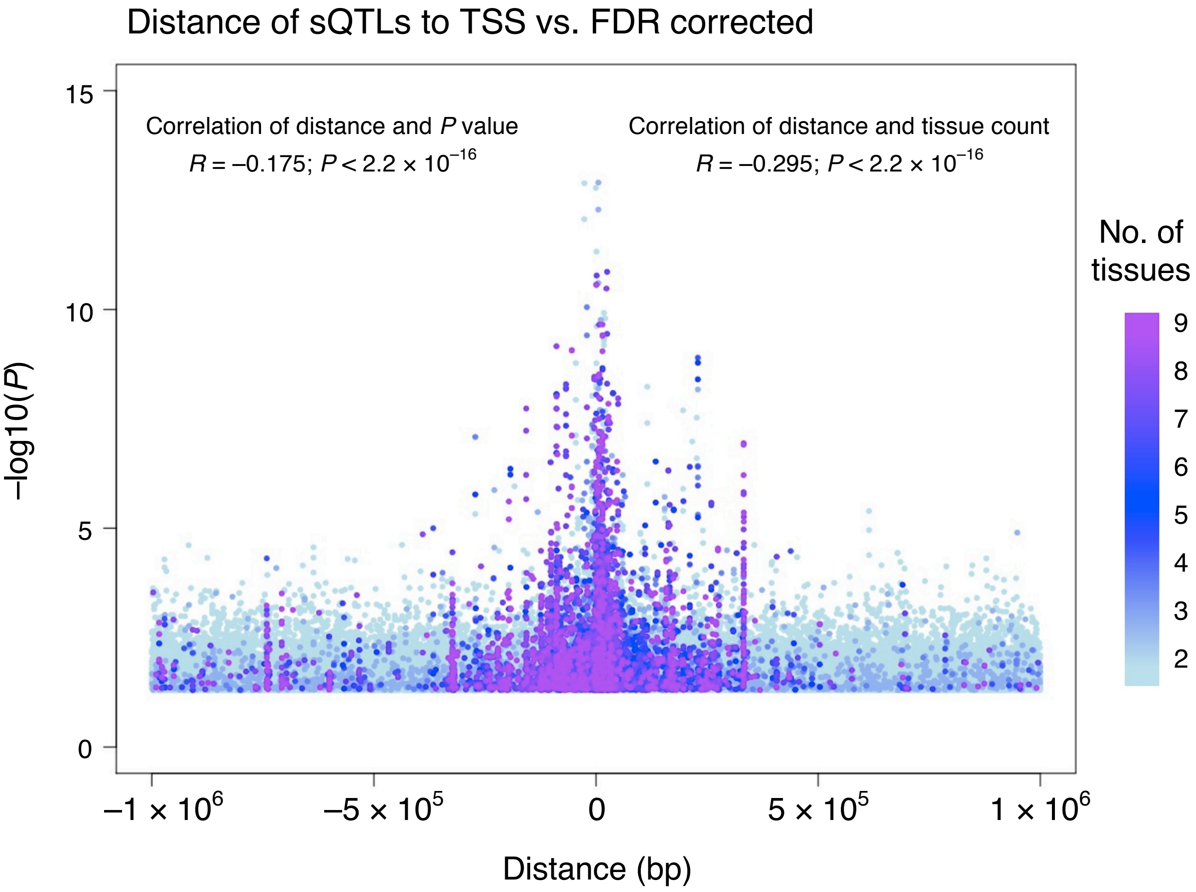


Fig. S29. Enrichment of regulatory element annotations for eQTLs. Each group of bars represents a regulatory annotation (in decreasing genomic prevalence: Enhancer, union of all enhancer regions defined by histone modification ChIP in the Roadmap cell types; DNase, union of all open chromatin regions defined by DNaseI hypersensitive peaks in the Roadmap cell types; ChIP, union of all protein binding sites identified by ChIP peaks in the ENCODE experiments; Enh+DNase, regions that are both identified as an enhancer by histone modification and open chromatin by DNase in the same cell type; Promoter, union of all promoter regions defined by histone modification ChIP in the Roadmap cell types). Each group of four bars represents the frequency of the regulatory annotation in four sets, from left to right: (a) 91 unambiguous (putatively causal) intergenic eQTLs from the single-tissue analysis; (b) 4085 intergenic eQTLs (best per gene/tissue) from the single-tissue analysis; (c) 14.7 Mb of intergenic sequence within 2.5 kb of the intergenic eQTLs; (d) 1.2 Gb of the total intergenic genome. In each case, the unambiguous eQTLs are strongly enriched relative to the set of all eQTLs, the local genomic context, and the intergenic genome.

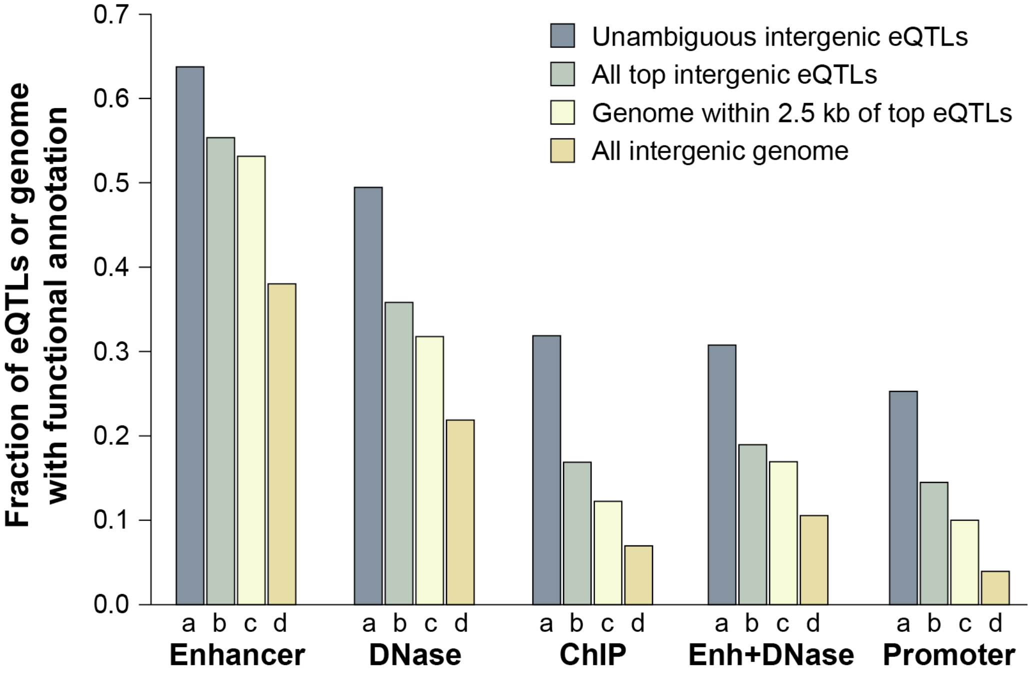


Fig. S30. WGCNA co-expression networks and module annotation (**A**) Transcription Factor TF-Module connectivity map in heart tissue. Each green circle node is a TF measured by ENCODE, and each orange hexagon node is a heart co-expression module. The edge between a TF and a module indicates that the TF is significantly enriched for binding to the transcription start sites of the genes in the module (BH corrected p<0.01). Some modules are potentially regulated by a large number of TFs (those in the center). (**B**) Correspondence of modules, based on weighted gene co-expression network analysis (WGCNA), learned independently in each tissue based on cross-individual expression similarity. Each slice in the circos plot is one module within a tissue. Edges between modules learned in different tissues represent the overlap in genes (green edges, overlap/enrichment approach, proportion overlap > 20% and Bonferroni corrected p<0.001), and the similarity in expression vectors across individuals (red edges, module PC correlation approach, FDR cutoff=0.1%). The thick blue edges indicate interactions identified by both methods. Only modules with at least one cross tissue interaction are shown.

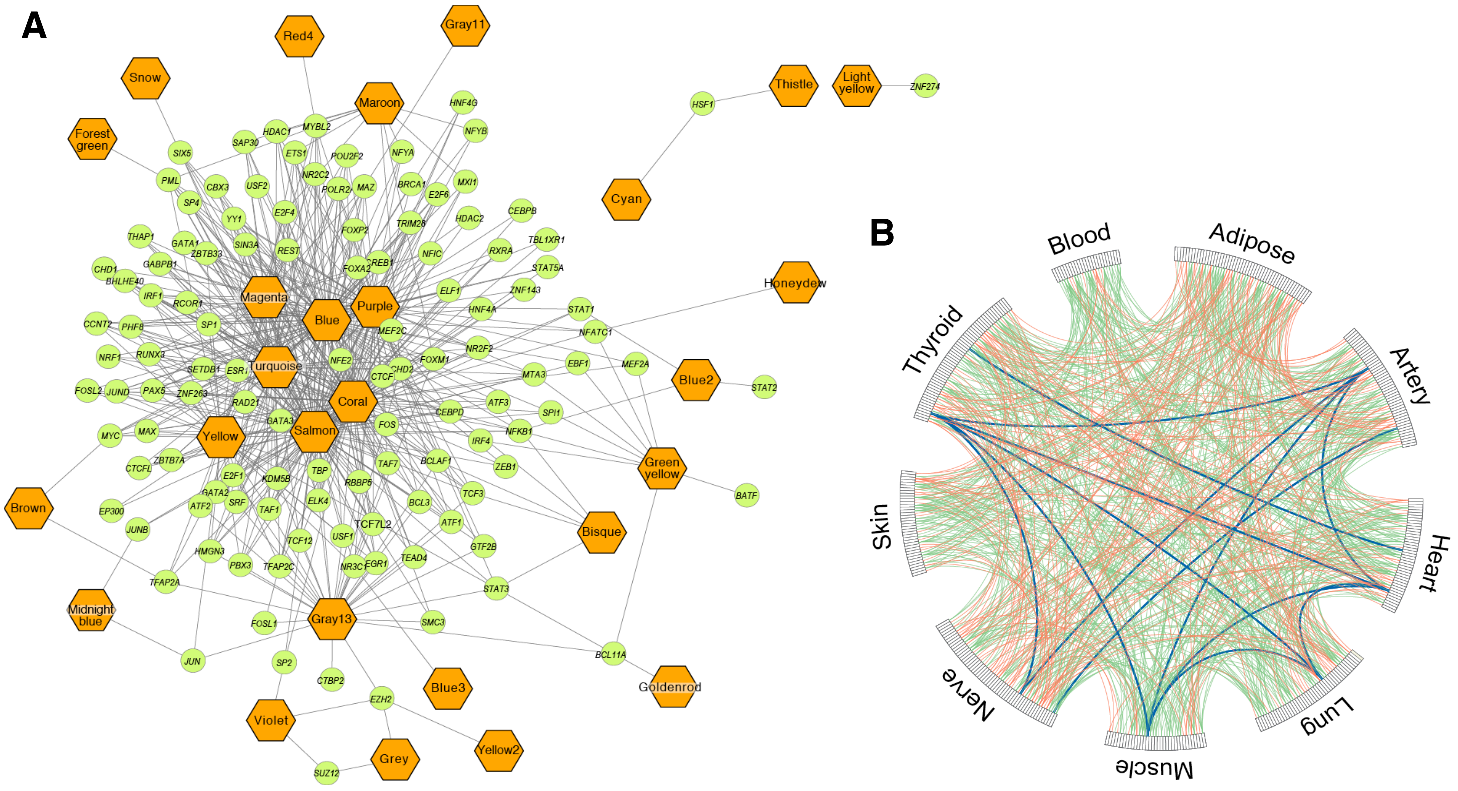
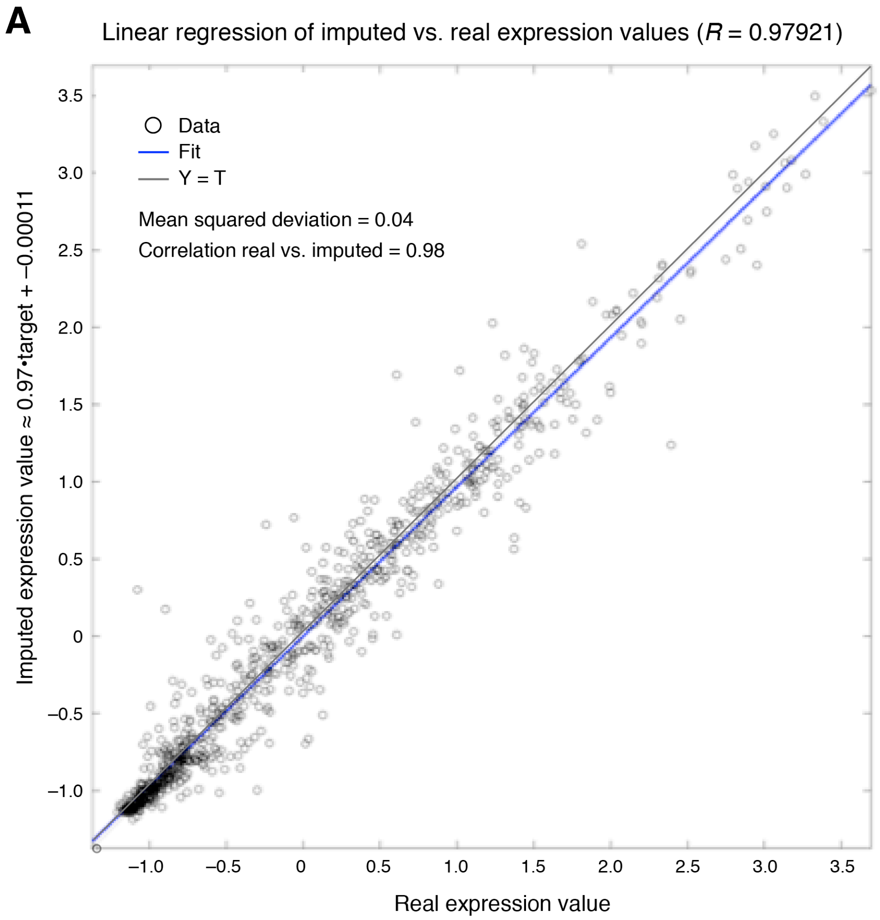


Fig. S31. Module switching QTLs (**A**) Scatter plot of imputed vs. real values in cross-validation of imputation methodology. A total number of 1000 random observed samples were deleted and then re-imputed. A Pearson correlation of 0.98 was found between the vectors of real and imputed data. (**B**) A total of 65,000 module-switching instances were called genome-wide based on gene expression variation. These varied on the strength of their effects on gene expression (measured as the correlation distance between the pair of characteristic patterns involved at each case). The red curve shows the fraction of all these module switches that have strength above a given threshold. The green curve shows the fraction of module switches above a certain given threshold of strength that are associated with a modQTL at $q < 0.05$. An increase in effect sizes is clearly positively correlated with an increase in the rate of modQTL discoveries. (**C**) Direct comparison between *strong-effect* modQTLs at q <0.05 that cause module switches of strength >0.5 (in set A), all modQTLs at q <0.05 (in set B), single tissue eQTLs (in set C), and multi-tissue eQTLs (in set D), discovered within the pilot release and analyses of GTEx. Only sufficiently significant discoveries of each group were considered for the comparison, but measures of statistical significance varied between different analyses.



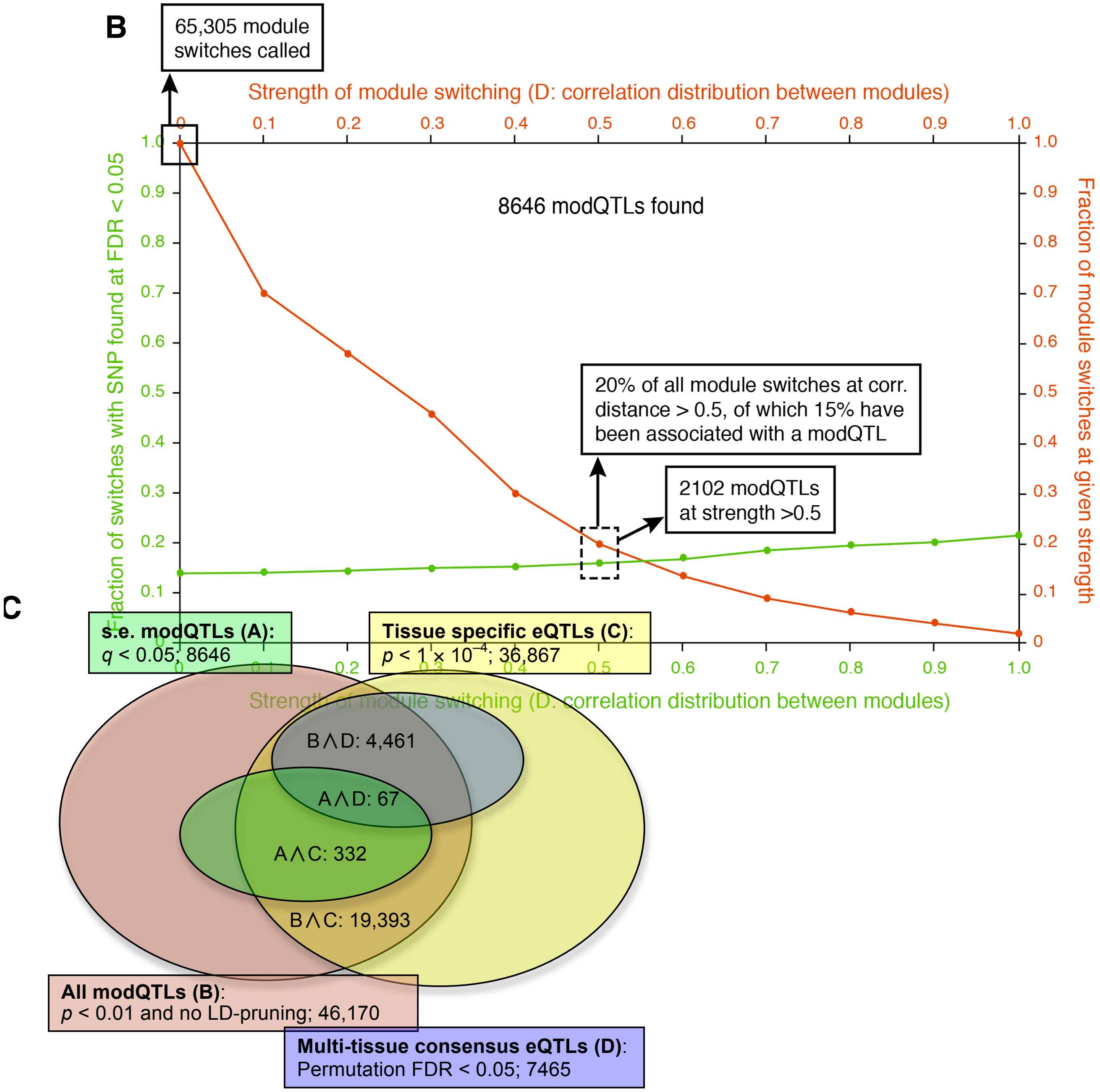
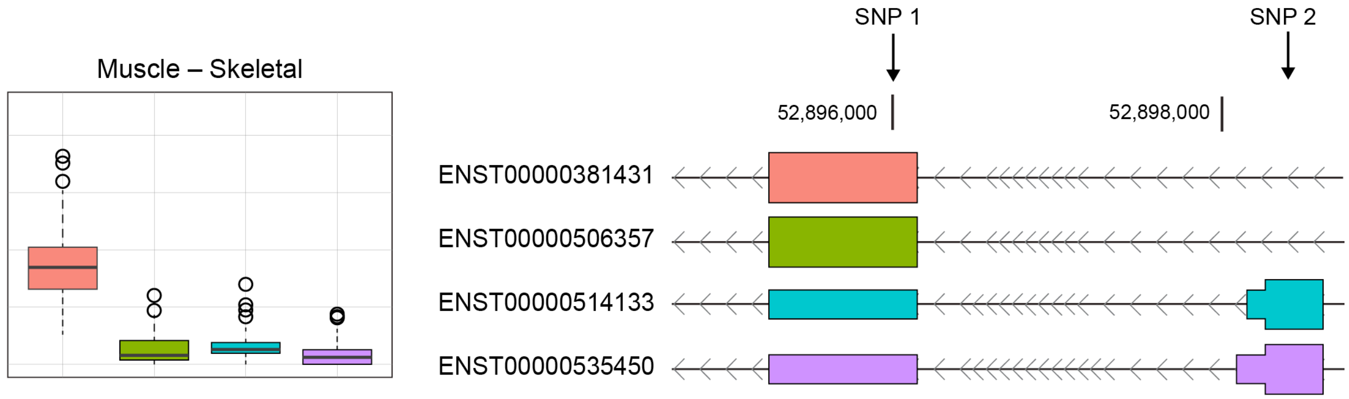

B
65,305 module switches called
Strength of module switching (D: correlation distribution between modules)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Fraction of switches with SNP found at FDR < 0.05
Fraction of module switches at given strength
8646 modQTLs found
20% of all module switches at corr. distance > 0.5, of which 15% have been associated with a modQTL
2102 modQTLs at strength >0.5
Strength of module switching (D: correlation distribution between modules)
C
s.e. modQTLs (A): q < 0.05; 8646
Tissue specific eQTLs (C): p < 1 × 10−4; 36,867
B⋀D: 4,461
A⋀D: 67
A⋀C: 332
B⋀C: 19,393
All modQTLs (B): p < 0.01 and no LD-pruning; 46,170
Multi-tissue consensus eQTLs (D): Permutation FDR < 0.05; 7465

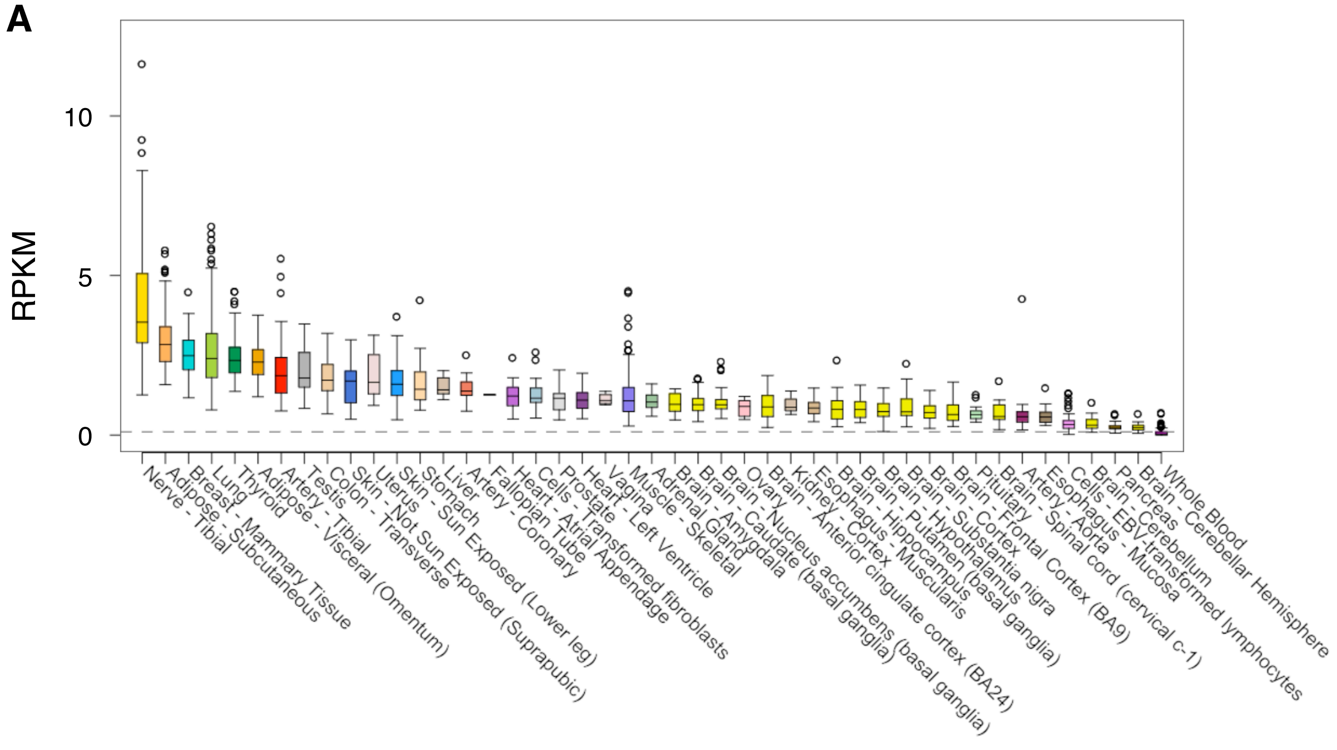
Fig. S32. Use of transcriptome data to improve clinical variant annotation. Tissue-specific isoform annotation from RNA-seq can impact the clinical interpretation of putative novel alleles in the gene *SGCB* which is associated with limb girdle muscular dystrophy type 2E (*81*). Four annotated transcript isoforms are reported for this gene in GENCODE v12 (*15*). In skeletal muscle ENST00000381431 is the dominant isoform expressed, with the other protein-coding isoforms only minimally expressed. SNP1 is predicted to be a stop-gain variant across all four protein-coding transcript isoforms, whereas SNP2 is predicted to be a nonsense variant across only two protein-coding transcript isoforms and an intronic variant across the remaining isoforms.



In silico	Stop gain	Stop gain
Transcript ⟵ RNA-Seq*	Stop gain	Intronic

* DNA variant annotation informed by tissue RNA-Seq quantification.

Fig. S33. Cross-tissue expression pattern of two candidate genes in a blood pressure GWAS locus. A–B) Expression box plots for (**A**) *ARHGAP42* and (**B**) *TMEM133* across all 45 tissues that have RNA sequencing data in GTEx. The box plots display the median and upper and lower quartiles of the expression values in each tissue. Circles denote outliers. The GWAS SNP rs633185 associated with systolic blood pressure ($p = 1.2 \times 10^{-17}$) and diastolic blood pressure ($p = 2 \times 10^{-15}$) lies in an intron of *ARHGAP42* and is a significant Tibial Artery eQTL acting on *ARHGAP42* and on the neighboring gene, *TMEM133*, based on single tissue eQTL analysis and FDR < 0.05 (see Fig. 9). The different sample sizes per tissue are listed in Table 1. RPKM refers to reads per kilobase per million. (**C**) Comparison of the expression box plots of *ARHGAP42* and *TMEM133* across the subset of 9 tissues with sufficient sample size for eQTL detection. *ARHGAP42* and *TMEM133* show similar expression patterns across the 9 tissues tested. In all panels, tissues are ordered according to median expression values in descending order.



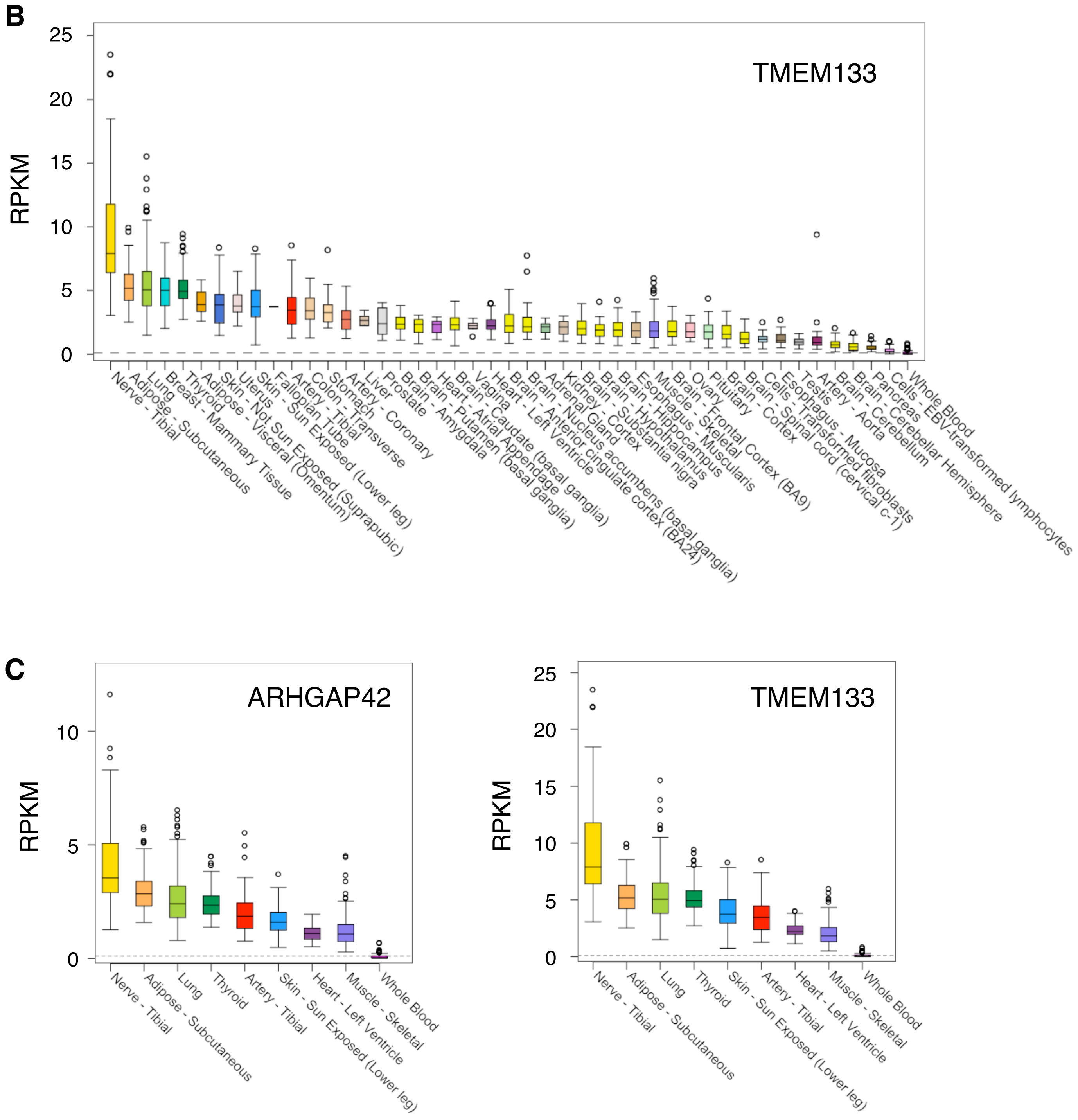
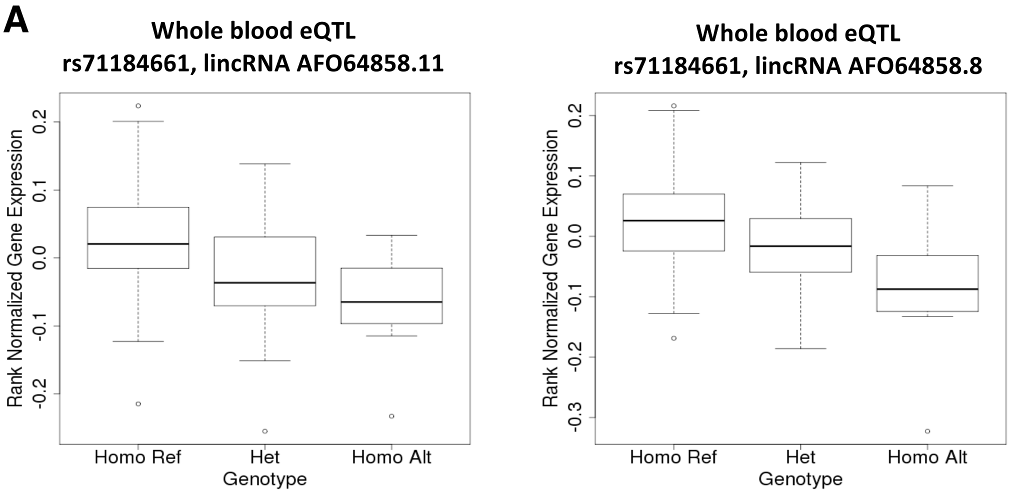

B
TMEM133
RPKM
25
20
15
10
5
0
Nerve - Tibial
Adipose - Subcutaneous
Lung
Breast - Mammary Tissue
Thyroid
Adipose - Visceral (Omentum)
Skin - Not Sun Exposed (Suprapubic)
Uterus
Skin - Sun Exposed (Lower leg)
Fallopian Tube
Artery - Tibial
Colon - Transverse
Stomach
Artery - Coronary
Liver
Prostate
Brain - Amygdala
Brain - Putamen (basal ganglia)
Heart - Atrial Appendage
Brain - Caudate (basal ganglia)
Vagina
Heart - Left Ventricle
Brain - Anterior cingulate cortex (BA24)
Brain - Nucleus accumbens (basal ganglia)
Adrenal Gland
Kidney - Cortex
Brain - Substantia nigra
Brain - Hypothalamus
Brain - Hippocampus
Esophagus - Muscularis
Muscle - Skeletal
Brain - Frontal Cortex (BA9)
Ovary
Pituitary
Brain - Cortex
Brain - Spinal cord (cervical c-1)
Cells - Transformed fibroblasts
Esophagus - Mucosa
Testis
Artery - Aorta
Brain - Cerebellum
Brain - Cerebellar Hemisphere
Pancreas
Cells - EBV-transformed lymphocytes
Whole Blood
C
ARHGAP42
RPKM
10
5
0
Nerve - Tibial
Adipose - Subcutaneous
Lung
Thyroid
Artery - Tibial
Skin - Sun Exposed (Lower leg)
Heart - Left Ventricle
Muscle - Skeletal
Whole Blood
TMEM133
RPKM
25
20
15
10
5
0
Nerve - Tibial
Adipose - Subcutaneous
Lung
Thyroid
Skin - Sun Exposed (Lower leg)
Artery - Tibial
Heart - Left Ventricle
Muscle - Skeletal
Whole Blood

Fig. S34. Blood-specific *cis*-eQTL indicates two lncRNAs as putative causal genes in a pleiotropic GWAS locus. (**A**) The intergenic GWAS SNP, rs2836878, associated with inflammatory bowel disease (p =5×10^{-48}) (*82*), pediatric-onset inflammatory bowel disease (p =4×10^{-12}) (*83*), and ulcerative colitis (p =2×10^{-22}) (*84*) is in high LD (r^2 = 0.90) with a whole blood eQTL, rs71184661, acting on the lincRNAs *AF064858.11* (single-tissue eQTL p =1.1×10^{-6}) and *AF064858.8* (single-tissue eQTL p =2×10^{-6}). The SNP rs71184661 is the most significant whole blood *cis* eQTL for *AF064858.11* and *AF064858.8* based on the single tissue eQTL analysis at FDR < 0.05. The start site of *AF064858.11* (on the reverse strand) is 83.5 kb upstream of the GWAS SNP, rs2836878, and the start site of *AF064858.8* (also on the reverse strand) is 87.5 kb upstream of the GWAS SNP. (**B**) An average eQTL posterior probability between two multi-tissue eQTL detection methods used in this paper is presented for *AF064858.11* and *AF064858.8* across the 9 tissues tested for eQTLs, ordered based on the gene's median expression values in descending order (same order as in panel **C)**. This GWAS-eQTL is highly whole blood-specific based on the multi-tissue eQTL analyses (*AF064858.11*: mean posterior probability of whole blood eQTL, P = 0.96, and mean posterior probability of blood-specific configuration, P = 0.84; *AF064858.8*: mean posterior probability of whole blood eQTL, P = 0.90, and mean posterior probability of blood-specific configuration, P = 0.59). The eQTL tissue-specificity is a bit stronger for *AF064858.11* compared to *AF064858.8.* The most significant *cis* eQTL for the lincRNA *AF064858.11* based on the multi-tissue eQTL analysis was rs2836883 (r^2 = 0.99 with GWAS SNP, rs2836878), and for *AF064858.8* was rs71184661 (r^2 = 0.90 with GWAS SNP). (**C**) Box plot expression profiles are presented for *AF064858.11* and *AF064858.8* across the 9 tissues analyzed for eQTLs. Tissues were ordered in descending order based on the gene's median expression values. (**D)** Box plot expression profile of *AF064858.11* and *AF064858.8* across all tissues collected in GTEx that have RNA-sequencing data. Tissues were ordered in descending order according to the gene's median expression values.



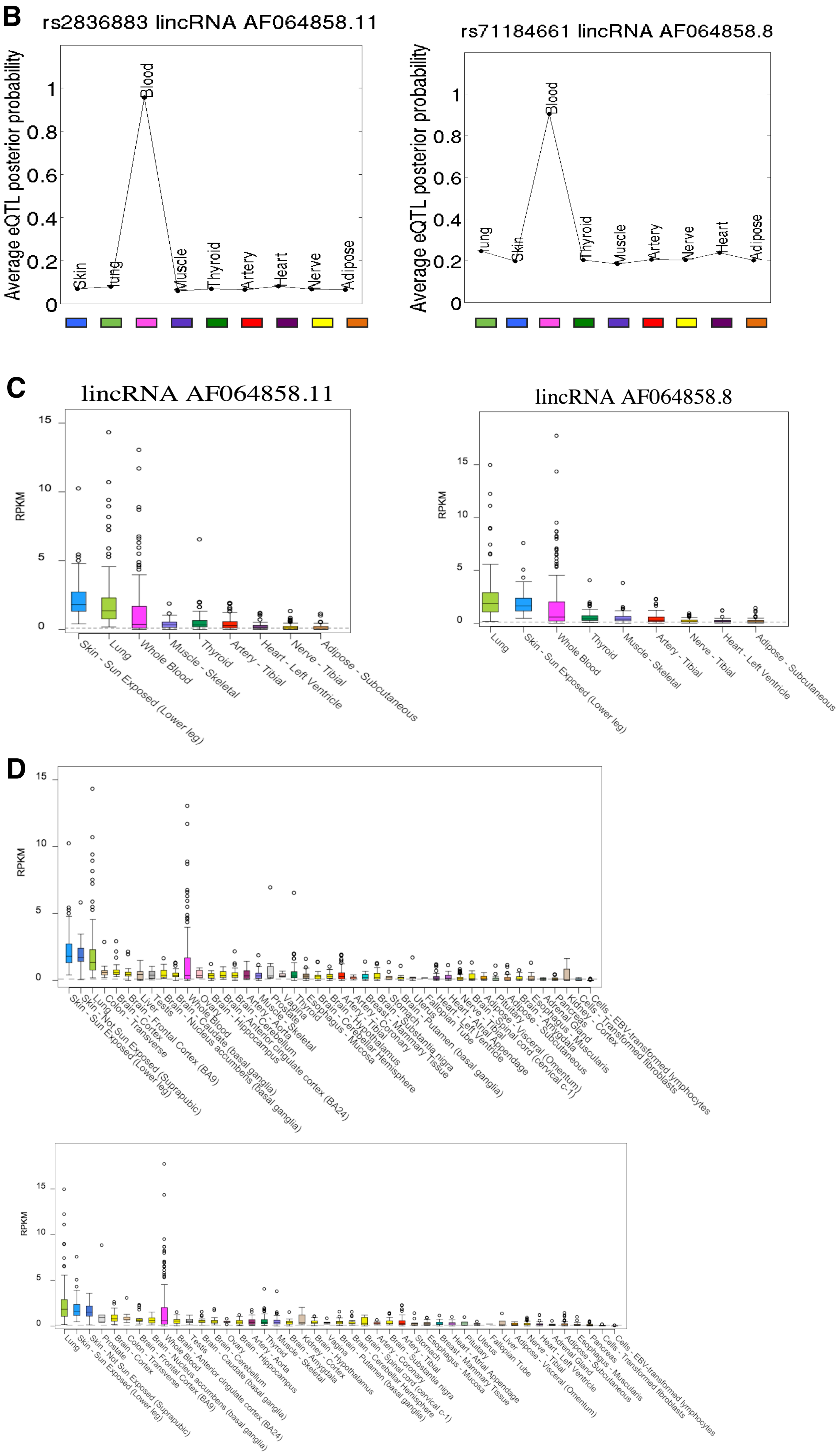

B
rs2836883 lincRNA AF064858.11
Average eQTL posterior probability
0
0.2
0.4
0.6
0.8
1
Skin
lung
Blood
Muscle
Thyroid
Artery
Heart
Nerve
Adipose
rs71184661 lincRNA AF064858.8
Average eQTL posterior probability
0
0.2
0.4
0.6
0.8
1
lung
Skin
Blood
Thyroid
Muscle
Artery
Nerve
Heart
Adipose
C
lincRNA AF064858.11
RPKM
0
5
10
15
Skin - Sun Exposed (Lower leg)
Lung
Whole Blood
Muscle - Skeletal
Thyroid
Artery - Tibial
Heart - Left Ventricle
Nerve - Tibial
Adipose - Subcutaneous
lincRNA AF064858.8
RPKM
0
5
10
15
Lung
Skin - Sun Exposed (Lower leg)
Whole Blood
Thyroid
Muscle - Skeletal
Artery - Tibial
Nerve - Tibial
Heart - Left Ventricle
Adipose - Subcutaneous
D
RPKM
0
5
10
15
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Lung
Colon - Transverse
Brain - Cortex
Liver
Brain - Frontal Cortex (BA9)
Testis
Brain - Nucleus accumbens (basal ganglia)
Whole Blood
Brain - Caudate (basal ganglia)
Ovary
Brain - Hippocampus
Brain - Anterior cingulate cortex (BA24)
Artery - Aorta
Brain - Cerebellum
Muscle - Skeletal
Prostate
Vagina
Thyroid
Esophagus - Mucosa
Brain - Cerebellar Hemisphere
Artery - Tibial
Brain - Hypothalamus
Artery - Coronary
Breast - Mammary Tissue
Stomach
Brain - Substantia nigra
Uterus
Brain - Putamen (basal ganglia)
Fallopian Tube
Heart - Left Ventricle
Heart - Atrial Appendage
Nerve - Tibial
Brain - Spinal cord (cervical c-1)
Adipose - Visceral (Omentum)
Pituitary
Adipose - Subcutaneous
Brain - Amygdala
Esophagus - Muscularis
Adrenal Gland
Pancreas
Kidney - Cortex
Cells - Transformed fibroblasts
Cells - EBV-transformed lymphocytes
RPKM
0
5
10
15
Lung
Skin - Sun Exposed (Lower leg)
Skin - Not Sun Exposed (Suprapubic)
Prostate
Brain - Cortex
Colon - Transverse
Brain - Frontal Cortex (BA9)
Whole Blood
Brain - Nucleus accumbens (basal ganglia)
Testis
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Ovary
Brain - Cerebellum
Artery - Aorta
Brain - Hippocampus
Thyroid
Muscle - Skeletal
Brain - Amygdala
Kidney - Cortex
Vagina
Brain - Hypothalamus
Brain - Putamen (basal ganglia)
Brain - Cerebellar Hemisphere
Artery - Coronary
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Artery - Tibial
Stomach
Esophagus - Mucosa
Breast - Mammary Tissue
Heart - Atrial Appendage
Pituitary
Uterus
Fallopian Tube
Liver
Adipose - Visceral (Omentum)
Nerve - Tibial
Heart - Left Ventricle
Adrenal Gland
Adipose - Subcutaneous
Esophagus - Muscularis
Pancreas
Cells - Transformed fibroblasts
Cells - EBV-transformed lymphocytes

Table S1. **Demographic summary of the 175 donors in the pilot analysis freeze.** BMI = body mass index, SD = standard deviation, AI = American Indian, AA = African American, EA = European American. Donor ischemic time is calculated as the number of minutes since the donors' tissues were last perfused (i.e., onset of ischemia) until the first tissue is put in the PAXgene stabilizer.

		Age		BMI		Donor ischemic time (min)		Race				Ethnicity			
Gender	***n***	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**AI**	**Asian**	**AA**	**EA**	**Hispanic**	**Not Hispanic**	**Not reported**	**Unknown**
Female	64	49.4	14.3	26.2	4.3	394.3	399	0	1	8	55	0	35	8	21
Male	111	50	12.6	27.5	3.8	365.2	392.5	1	1	16	93	1	59	14	37
All	**175**	**49.8**	**13.2**	**27**	**4**	**375.8**	**394**	**1**	**2**	**24**	**148**	**1**	**94**	**22**	**58**

Table S2. **Cause of death categories for donors.** Coding for the Hardy scale (*85,86*) score is as follows: 0 = subject as on a ventilator when death was pronounced (note: not part of the original Hardy Scale), 1 = deaths due to accident, blunt force trauma or suicide, terminal phase estimated at < 10 minutes, 2 = sudden unexpected deaths of people who had been reasonably healthy, after a terminal phase estimated at < 1 hour, 3 = death after a terminal phase of 1–24 hours (not classifiable as 2 or 4), 4 = death after a long illness, with a terminal phase longer than 1 day. 4 surgical donors are not included in the table.

		Age		**Donor ischemic time**		**Gender**		**Hardy score**				
Cause of death	***n***	**Mean**	**SD**	**Mean**	**SD**	**Female**	**Male**	**0**	**1**	**2**	**3**	**4**
Asphyxiation	10	40.6	12.5	382.2	264.8	3	7	7	3	0	0	0
Blunt injury	25	40.2	16.5	344	272.3	9	16	20	3	0	1	0
Burn	1	62	NA	270	NA	0	1	1	0	0	0	0
Cardiovascular disease	39	52.9	11.1	608.2	438.6	14	25	20	0	13	4	1
Cerebrovascular accident	52	53	9.3	298.5	275.4	20	32	45	0	0	3	3
Drug overdose	7	38.3	16.5	404.6	241	3	4	3	1	2	1	0
Gun shot wound	6	41.5	10.3	364	451.1	0	6	5	1	0	0	0
Liver disease	5	47	13.7	612.4	330.1	2	3	0	0	0	0	5
Neurological disorder	4	53.5	21.9	643	328.6	3	1	1	0	0	0	3
Renal failure	5	60	8.6	773.4	413.2	4	1	1	0	0	0	4
Respiratory disease	5	58	7.4	800	414	2	3	3	1	1	0	0
Traumatic brain injury	10	52.8	9.8	501.6	290.8	3	7	6	1	0	1	2
Viral infection	2	62.5	2.1	898	455.4	1	1	1	0	0	0	1
All	**171**	**49.8**	**13.2**	**442.9**	**364.2**	**64**	**111**	**113**	**10**	**16**	**10**	**19**

Table S3. Sample and SNP QC summary for the Illumina Omni 5M genotyping. A total of 191 DNA samples were submitted for genotyping, consisting of 185 unique GTEx individuals. All 191 samples were released to dbGaP. Imputation was run on only the 185 unique samples, after excluding the duplicates and the HapMap individual, but not the Klinefelter samples. Of those, only 175 were included in the final eQTL and transcriptome analysis freeze, since donors required both genotype and RNA sequence data to be included in the freeze. None of the 8 flagged samples were included in the final analysis freeze. Flags for potentially bad SNPs were set at Hardy–Weinberg equilibrium (HWE) $P<1\times10^{-6}$, genotype missingness $P<1\times10^{-8}$, and batch association $P<1\times10^{-3}$.

QC steps	**# unique GTEx individuals**	**# flagged samples**	**# SNPs kept**	**# SNPs removed**
Original data	191		4,276,680	
1. Exclude monomorphic SNPs and SNPs with < 90% genotyping rate	191		4,276,680	671,702
2. Exclude individuals with call rate < 95%	191		3,604,978	
3. Sex check	191	2 Klinefelter individuals (represented by 3 samples)	3,604,978	
4. Heterozygosity test	191		3,604,978	
5. Genome identity-by-descent (IBD)	191		3,604,978	
Sample contamination	191		3,604,978	
Cryptic relationships	191		3,604,978	
Sample duplicates	191	5	3,604,978	
6. HapMap individual	186	1	3,604,978	
7. SNPs test statistics	185		3,604,978	
Testing HWE using 156 Europeans ($p < 1 \times 10^{-6}$)	185		3,604,978	312
Genotype missingness predicted using surrounding haplotypes ($p < 1 \times 10^{-8}$)	185		3,604,978	1431
Testing for association with plates ($p < 1 \times 10^{-3}$)	185		3,604,978	75
Subtotal	185		3,604,978	1757
8. SNP call rate < 95%	185		3,603,221	24,797
9. SNPs with heterozygous haploid genotypes on sex chromosomes	185		3,578,424	2,547
Total	**185**	**8**	**3,575,877**	**700,803**

Table S4. Sample and SNP QC summary for the Illumina HumanExome genotyping. Thresholds for potentially bad SNPs were set at Hardy–Weinberg equilibrium (HWE) $P < 1\text{x}10^{-6}$, genotype missingness $P < 1\text{x}10^{-8}$, and batch association $P < 1\text{x}10^{-3}$. All 8 flagged samples were not included in the final analysis freeze, which included data for only 175 donors (see legend to Table S3). All 190 samples were released to dbGaP.

QC steps	**# unique GTEx individuals**	**# flagged samples**	**# SNPs kept**	**# SNPs removed**
Original data	190		242,040	
1. Exclude SNPs with autocall call rate <80%	190		242,040	536
1. Exclude monomorphic SNPs and SNPs with < 99% genotyping rate	190		242,040	163,005
2. Exclude Individuals with call rate < 95%	190		78,499	
3. Sex check	190	2 Klinefelters		
4. Heterozygosity	190			
5. Genome IBD	190			
Contamination	190		78,499	
Sample duplicates	190	5	78,499	
6. HapMap individual	185	1	78,499	
7. SNP filter	184		78,499	
Testing HWE ($p < 1 \times 10^{-6}$) with 155 Europeans				347
Genotype missingness predicted using surrounding haplotypes ($p < 1 \times 10^{-8}$)	184			
Testing for association with plates ($p < 1 \times 10^{-3}$)	184			1
Subtotal	184		78,499	348
8. SNP call rate < 99%	184		78,151	0
9. SNPs with heterozygous haploid	184		78,151	21
Total	**184**	**8**	**78,130**	**163,910**

Table S5.Evaluation of imputation accuracy. Evaluation of accuracy of imputed data on Illumina's OMNI 5M array by comparison with direct genotype calls on Exome Chip from the same overlapping 184 GTEx samples.

MAF	Info	Minor allele concordance	Counts (equal/all)	Minor allele hom	Counts (equal/all)	Het	Counts (equal/all)	# CNPs	Mean r^2	Median r^2	Complete pair # CNP
0	≥0	0.0000	0/10301	NaN	0/0	NaN	0/0	69913	NA	NA	0
0	≥0.1	0.0000	0/9573	NaN	0/0	NaN	0/0	15790	NA	NA	0
0	≥0.2	0.0000	0/9573	NaN	0/0	NaN	0/0	11835	NA	NA	0
0	≥0.3	0.0000	0/9572	NaN	0/0	NaN	0/0	9550	NA	NA	0
0	≥0.4	0.0000	0/9564	NaN	0/0	NaN	0/0	7982	NA	NA	0
0	≥0.5	0.0000	0/9513	NaN	0/0	NaN	0/0	6629	NA	NA	0
0	≥0.6	0.0000	0/9186	NaN	0/0	NaN	0/0	5664	NA	NA	0
0	≥0.7	0.0000	0/8066	NaN	0/0	NaN	0/0	4619	NA	NA	0
0	≥0.8	0.0000	0/6432	NaN	0/0	NaN	0/0	3387	NA	NA	0
0	≥0.9	0.0000	0/4890	NaN	0/0	NaN	0/0	2058	NA	NA	0
0 < a ≤ 0.01	≥0	0.3654	7345/20104	0.2841	50/176	0.4173	7295/17482	14991	0.5121	0.4996	12686
0 < a ≤ 0.01	≥0.1	0.5414	7345/13566	0.4386	50/114	0.6628	7295/11006	9473	0.6272	0.8065	9463
0 < a ≤ 0.01	≥0.2	0.5712	7345/12859	0.4717	50/106	0.7078	7295/10307	8707	0.6578	0.8802	8700
0 < a ≤ 0.01	≥0.3	0.5933	7345/12380	0.5000	50/100	0.7418	7295/9834	8101	0.6821	0.9206	8094
0 < a ≤ 0.01	≥0.4	0.6101	7345/12039	0.5000	50/100	0.7685	7295/9493	7636	0.7015	0.9500	7629
0 < a ≤ 0.01	≥0.5	0.6257	7344/11737	0.5051	50/99	0.7924	7294/9205	7168	0.7186	0.9668	7161
0 < a ≤ 0.01	≥0.6	0.6411	7301/11388	0.5263	50/95	0.8119	7251/8931	6734	0.7339	0.9782	6728
0 < a ≤ 0.01	≥0.7	0.6786	7113/10482	0.5495	50/91	0.8378	7063/8430	6110	0.7611	0.9916	6105
0 < a ≤ 0.01	≥0.8	0.7152	6548/9155	0.5765	49/85	0.8666	6499/7499	5181	0.8000	0.9989	5177
0 < a ≤ 0.01	≥0.9	0.7589	5273/6948	0.5921	45/76	0.8984	5228/5819	3837	0.8515	1.0000	3834
0.01 < a ≤ 0.05	≥0	0.7474	11203/14989	0.4869	335/688	0.8113	10868/13395	2379	0.7259	0.8408	2351
0.01 < a ≤ 0.05	≥0.1	0.7738	11203/14477	0.5007	335/669	0.8424	10868/12902	2273	0.7482	0.8548	2272
0.01 < a ≤ 0.05	≥0.2	0.7786	11203/14389	0.5030	335/666	0.8479	10868/12817	2252	0.7542	0.8595	2251
0.01 < a ≤ 0.05	≥0.3	0.7847	11203/14276	0.5030	335/666	0.8555	10868/12704	2225	0.7612	0.8639	2224
0.01 < a ≤ 0.05	≥0.4	0.7904	11203/14174	0.5045	335/664	0.8623	10868/12604	2198	0.7679	0.8684	2197
0.01 < a ≤ 0.05	≥0.5	0.7976	11200/14042	0.5060	335/662	0.8708	10865/12477	2163	0.7770	0.8766	2162
0.01 < a ≤ 0.05	≥0.6	0.8094	11157/13784	0.5146	335/651	0.8834	10822/12251	2091	0.7929	0.8869	2090
0.01 < a ≤ 0.05	≥0.7	0.8249	11023/13363	0.5187	333/642	0.8982	10690/11901	1986	0.8131	0.9031	1985
0.01 < a ≤ 0.05	≥0.8	0.8472	10626/12543	0.5304	332/626	0.9158	10294/11240	1816	0.8427	0.9304	1815
0.01 < a ≤ 0.05	≥0.9	0.8821	9252/10489	0.5777	305/528	0.9405	8947/9513	1433	0.8882	0.9841	1432
0.05 < a ≤ 0.5	≥0	0.9387	197757/210678	0.9720	39716/40861	0.9478	158041/166745	2874	0.9327	0.9866	2867
0.05 < a ≤ 0.5	≥0.1	0.9455	197757/209150	0.9721	39716/40857	0.9565	158041/165221	2863	0.9344	0.9867	2862
0.05 < a ≤ 0.5	≥0.2	0.9473	197757/208769	0.9721	39716/40854	0.9587	158041/164843	2859	0.9353	0.9867	2859
0.05 < a ≤ 0.5	≥0.3	0.9474	197757/208737	0.9722	39716/40853	0.9589	158041/164815	2857	0.9360	0.9867	2857
0.05 < a ≤ 0.5	≥0.4	0.9476	197757/208688	0.9722	39716/40850	0.9592	158041/164769	2856	0.9362	0.9867	2856
0.05 < a ≤ 0.5	≥0.5	0.9480	197733/208570	0.9723	39716/40849	0.9596	158017/164669	2851	0.9378	0.9868	2851
0.05 < a ≤ 0.5	≥0.6	0.9497	197627/208103	0.9725	39670/40791	0.9609	157957/164388	2841	0.9405	0.9870	2841
0.05 < a ≤ 0.5	≥0.7	0.9520	197453/207409	0.9733	39648/40735	0.9629	157805/163878	2823	0.9452	0.9871	2823
0.05 < a ≤ 0.5	≥0.8	0.9538	196612/206143	0.9742	39508/40553	0.9643	157104/162916	2788	0.9504	0.9877	2788
0.05 < a ≤ 0.5	≥0.9	0.9588	192715/201001	0.9761	38849/39801	0.9685	153866/158869	2680	0.9596	0.9888	2680

Column	**Detail**
MAF	Minor allele frequency stratified by direct calls on the Exome Chip.
Info	IMPUTE2 info score; measure of imputation confidence.
Minor allele concordance	Minor allele concordance of all SNPs.
Counts (equal/all)	All: the number of complete obs, equal: the number of equal obs, equal/all = minor allele concordance.
Minor allele hom	Minor allele homozygotes concordance.
Counts (equal/all)	All: the number of complete obs, equal: the number of equal obs, equal/all=minor allele homozygotes concordance.
het	Heterozygotes concordance.
Counts (equal/all)	All: the number of complete obs, equal: the number of equal obs, equal/all=heterozygote concordance.
#CNPs	The number of CNPs that are in each category by MAF and INFO, regardless of how many calls are missing.
Mean r^2	Correlation coefficient r^2 is calculated per SNP, and then the r^2s of all SNPs are averaged.
Median r^2	Correlation coefficient r^2 is first calculated per SNP, and then the median of r^2s of all SNPs is taken.
Complete pair #CNP	For mean r^2 and median r^2, the # of SNPs whose r^2 can be calculated, sd!=0.

Table S6. Distribution of *cis*-eQTLs around the TSS. Distribution of the most significant *cis*-eQTLs per gene around the gene target's transcript start site.

Tissue	**% of *cis*-eQTLs* upstream of TSS**	**% of *cis*-EQTLs downstream of TSS**	**% of *cis*-eQTLs within ±100 kb from TSS**	**# of *cis*-EQTLs within ±100 kb from TSS**
Subcutaneous adipose	59	41	83	966
Tibial artery	62	38	80	1408
Whole blood	59	41	83	1648
Heart. left ventricle	61	39	82	753
Lung	59	41	80	1480
Skeletal muscle	62	39	81	1250
Tibial nerve	61	39	81	1268
Skin, sun exposed	60	40	79	1079
Thyroid	60	40	81	1812
Average	**60**	**40**	**81**	**1296**

*Fraction of eQTLs were computed out of all significant *cis*-eQTLs in a ± 1-Mb window around the gene target's transcript start site (TSS), considering only the most significant *cis*-eQTL per gene at FDR < 5%. On average about 50% of all the SNPs tested for eQTLs lie upstream and 50% downstream the target genes' TSS, for each of the 9 tissues tested.

Table S7. Distance of significant *cis*-eQTLs from their target TSS. Distance of the most significant *cis*-eQTL per gene from their target transcript start site at different percentiles of the distance distribution.

A. Distance was computed for the different percentiles, considering both upstream and downstream eQTLs in the percentile ranking based on absolute distance.

(Distance boundaries are symmetrical around transcript start site)

Percentiles:

Tissue	25th	50th	75th	90th	95th	99th
Subcutaneous Adipose	6,190	21,862	66,253	183,560	435,790	860,370
Tibial Artery	7,012	26,028	77,507	242,320	483,700	820,290
Whole Blood	6,424	22,104	65,978	191,700	401,960	861,200
Heart Left Ventricle	5,597	21,247	65,414	168,590	368,360	872,180
Lung	7,078	24,446	72,257	235,100	500,070	880,680
Skeletal Muscle	5,969	23,170	68,100	194,480	405,420	850,160
Tibial Nerve	7,862	25,863	73,712	217,000	432,160	871,140
Skin, sun exposed lower limb	6,670	24,350	78,303	251,480	515,110	950,740
Thyroid	7,048	25,360	73,382	215,690	432,710	878,560
Average:	6,650	23,826	71,212	211,102	441,698	871,702

B. Distance was computed considering only the upstream eQTLs in the percentile ranking.

Percentiles:

Tissue	25th	50th	75th	90th	95th	99th
Subcutaneous Adipose	6,054	21,453	63,078	177,440	369,380	885,280
Tibial Artery	7,697	26,222	69,722	209,980	383,130	762,550
Whole Blood	6,309	22,876	68,586	184,520	361,440	845,890
Heart Left Ventricle	5,107	21,159	58,851	152,900	343,690	916,440
Lung	7,138	24,725	69,847	207,990	461,710	858,030
Skeletal Muscle	6,142	24,051	65,957	171,340	329,570	850,010
Tibial Nerve	7,806	25,560	67,976	175,350	334,770	866,200
Skin, sun exposed lower limb	6,880	24,118	74,454	233,310	502,840	937,910
Thyroid	7,481	25,250	74,816	202,590	404,080	847,180
Average:	6,735	23,935	68,143	190,602	387,846	863,277

C. Distance was computed considering only the downstream eQTLs in the percentile ranking.

Percentiles:

Tissue	25th	50th	75th	90th	95th	99th
Subcutaneous Adipose	6,534	22,279	72,478	229,280	468,980	803,230
Tibial Artery	5,896	25,664	93,665	292,130	583,820	869,020
Whole Blood	6,714	20,317	61,260	213,270	448,540	869,030
Heart Left Ventricle	5,754	21,911	73,124	203,540	441,710	836,000
Lung	7,064	23,401	79,245	288,640	558,830	902,070
Skeletal Muscle	5,638	22,441	72,671	231,060	462,700	856,160
Tibial Nerve	8,097	27,067	90,002	298,180	611,020	878,250
Skin, sun exposed lower limb	6,051	24,814	83,115	296,820	543,530	956,170
Thyroid	6,523	25,649	70,387	243,480	487,820	901,990
Average:	6,475	23,727	77,327	255,156	511,883	874,658

(A–C) Distance is in base pair units. A cutoff of FDR < 0.05 was used her for calling a *cis*-eQTLs significant.

Table S8. ASE QC statistics. (**A**) Basic statistics of ASE analysis. (**B**) Partitioning of the variance of pairwise correlation matrices based on allelic ratios, or total read counts per site, according to whether the sample pairs come from same or different individuals or tissues.

A

	Total sites ≥30 reads	**Sites 30 reads ASE $p < 0.005$**	**Sites 30 reads ASE $p < 0.005$ (%)**
Minimum	221	8	1.59%
Median	6383.5	389.5	5.99%
Maximum	16422	1349	15.0%

B

	Tissue	**Subject**	**Residual**
Allelic ratio	0.08625	0.17871	0.73504
Expression level	0.85577	0.09979	0.14444

Table S9. Splice-QTLs identified. The number of splice-QTLs (sQTLs) detected by Altrans and sQTLseekeR.

Tissue	**Altrans (+/– 1 MB from TSS)**		**sQTLseekeR (gene +/– 5 kb from TSS)**	
	Genes tested	**sQTLs**	**Genes tested**	**sQTLs**
Adipose, subcutaneous	8883	1869	14,615	335
Artery, tibial	8407	2347	14,583	435
Heart, left ventricle	7288	656	14,094	242
Lung	9010	2877	15,262	289
Muscle, skeletal	6770	1557	13,720	200
Nerve, tibial	9104	2083	14,534	234
Skin, sun-exposed	8495	1494	15,124	315
Thyroid	8902	1885	14,900	226
Whole blood	5098	2735	13,751	153

Table S10. Proportion of splice-QTLs also detected as significant eQTLs. The estimated fraction of true positive eQTLs amongst the significant sQTLs (π_1) for each of the nine pilot tissues was computed at a false discovery rate (FDR) cutoff of 0.05, using the Storey and Tibshirani *q*-value method (*26*). Only the most significant sQTL per gene, among multiple significant sQTLs per gene at FDR < 0.05 was considered, to satisfy the independence assumption of the π_0 calculation**.** This was computed for both the Altrans splice QTL method (which identifies SNPs associated with differences in expression levels of exon junctions), and the sQTLseekeR method (that identifies SNPs associated with differences in relative abundances of gene transcript isoforms). To estimate π_1, we extracted the nominal eQTL *p*-values from Matrix eQTL for each of the significant sQTLs considered at FDR < 0.05 for a given tissue, and estimated π_0 that is the proportion of true null eQTL associations amongst all sQTLs tested. π_1 is defined as $1 - \pi_0$, with values between [0–1]. *Number of genes with at least one significant sQTL at FDR < 0.05, whose most significant sQTL SNP had a matching eQTL *p*-value from Matrix eQTL. Hence, the sQTL gene numbers here might be slightly smaller than those in Table S9. TSS, transcript start site.

	Altrans (+/– 1 Mb from TSS)		**sQTLseekeR (gene +/– 5 kb)**	
Tissue	**# sQTL genes***	π_1	**# sQTL genes***	π_1
Adipose, subcutaneous	1806	0.14	332	0.50
Artery, tibial	2254	0.19	417	0.70
Heart, left ventricle	629	0.18	238	0.13
Lung	2793	0.27	265	0.43
Muscle, skeletal	1505	0.27	198	0.57
Nerve, tibial	2008	0.16	231	0.55
Skin, sun-exposed	1432	0.16	309	0.40
Thyroid	1807	0.24	223	0.40
Whole blood	2625	0.18	151	0.66
Average [range]:	**1873 [629–2793]**	**0.20 [0.14–0.27]**	**263 [151–417]**	**0.48 [0.13–0.70]**

Table S11. **Enrichment of regulatory annotations among eQTLs.** Two sets of eQTLs (unambiguous intergenic, and all top intergenic) and two background genomic regions (genome within 2.5 kb of all top intergenic, and entire intergenic genome) were intersected with five regulatory annotations (see Fig. S29). Hypergeometric *p*-values were calculated for enrichment of each regulatory annotation among unambiguous vs. all eQTLs, and all eQTLs vs. genome within 2.5 kb.

	Base pairs (bp)				**Percent of all (%)**					
	Unambiguous inter-genic eQTLs	All top inter-genic eQTLs	Genome within 2.5 kb of all top intergenic eQTLs	All intergenic genome	Unambiguous intergenic eQTLs	All top inter-genic eQTLs	Genome within 2.5 kb of all top inter-genic eQTLs	All inter-genic genome	Hyper-geometric *p* (unambiguous vs. all top eQTLs)	Hyper-geometric *p* (all top eQTLs vs. proximal genome)
All	91	4085	14,673,356	1,208,196,281						
ChIP	29	690	1,796,918	84,244,717	31.9	16.9	12.2	7.0	0.0002902	4.28×10^{-18}
DNase	45	1463	4,662,689	264,353,648	49.5	35.8	31.8	21.9	0.0047993	2.88×10^{-8}
Enh+DNase	28	774	2,486,138	127,579,086	30.8	18.9	16.9	10.6	0.0041887	0.00045102
Promoter	23	592	1,469,856	47,769,552	25.3	14.5	10.0	4.0	0.0042916	1.68×10^{-19}
Enhancer	58	2261	7,797,839	459,415,816	63.7	55.3	53.1	38.0	0.0632879	0.00302844

Table S12. GWAS SNPs in LD with eQTLs. The number of genome-wide significant GWAS SNPs in linkage disequilibrium (LD) with at least one eQTL among the 9 tissues tested in GTEx.

SNP category	# (%) pruned GWAS SNPs*	# GWAS SNPs not in LD with non-synonymous or splice variants**	% Coding	% Non-coding
Total # pruned SNP associations with ~630 traits	5,195 (100%)	4,562 (87.8%)	4.6%	95.4%
# GWAS SNPs in LD with best-eQTL-per-gene in at least one GTEx tissue$^{\Psi}$**:**				
Union of single and multi-tissue eQTL methods	308 (6.0%)	211	11.0%	89.0%
Single tissue eQTL method$^{\xi}$	214 (4.1%)	146	9.0%	91.0%
Multi-tissue eQTL methods$^{\Phi}$	208 (4.0%)	144	12.1%	87.9%
Overlap between single- and multi-tissue methods	114 (2.8%)	79	8.8%	91.2%

* GWAS SNP associations at genome-wide significance (p<5x10^{-8}) from PheGenI and the NHGRI GWAS catalog were collectively pruned using a linkage disequilibrium (LD) cutoff of r^2≥0.8. Percentages (%) were computed relative to the 5,195 pruned SNPs.

** Functional annotations were taken from dbSNP version 137.

$^{\Psi}$ GWAS SNPs in LD with eQTLs significant in more than one tissue were counted as a single instance.

$^{\xi}$ The single tissue eQTL method refers to Matrix eQTL analysis. The significance threshold used was a permutation-based FDR<0.05.

$^{\Phi}$ The multi-tissue eQTL probabilities were computed as the average posterior probabilities between the two methods used in this paper, for eQTLs that passed an FDR<0.05 cutoff.

Table S13. Integration of GWAS SNPs with GTEx eQTLs. Genomic context of all genome-wide significant GWAS SNPs compared to those in LD to a GTEx eQTL.

		Subset of pruned GWAS SNPs in LD with GTEx eQTL*		All pruned GWAS SNPs	
Genomic Region	**Genomic context**	**# pruned GWAS SNPs in LD with GTEx eQTL***	**% pruned GWAS SNPs in LD with GTEx eQTL**	**All pruned GWAS SNPs**	**% pruned GWAS SNPs**
Noncoding	Intron	155	50.8%	2221	43.5%
Noncoding	Intergenic	76	24.9%	2372	46.5%
Noncoding	Near Gene 5'	22	7.2%	142	2.8%
Coding	Missense	18	5.9%	170	3.3%
Coding	CDs-synonymous	16	5.2%	61	1.2%
Noncoding	3' UTR	14	4.6%	82	1.6%
Noncoding	Near Gene 3'	3	1.0%	31	0.6%
Noncoding	5' UTR	1	0.3%	13	0.3%
Coding	frameshift	0	0%	2	0.04%
Noncoding	ncRNA	0	0%	7	0.14%
Coding	STOP-GAIN	0	0%	5	0.10%
Total number**:		305	100%	5106	100%

*LD pruning was performed on all GWAS SNPs from PheGenI and the NHGRI GWAS catalog, collectively, using an $r^2 > 0.8$ cutoff, to generate a rough set of independent SNPs. The genomic context refers to the GWAS SNPs. Significant eQTLs refer to a union of results from the single- and multi-tissue detection methods at FDR < 0.05. **A small number of SNPs did not have a genomic context annotation in these databases. “CDs” refers to coding sequence.

Table S14. **Proximity-based versus eQTL-based gene assignment for GWAS SNPs.** The tables below report the number of GWAS loci at genome-wide significance ($p<5\text{x}10^{-8}$) that show disconcordance between the candidate causal genes in GWAS loci proposed based on eQTL annotation or physical proximity to the GWAS SNP. The discordant counts where computed excluding (**A-B**) or including non-protein coding genes (**C-D**), and using eQTLs either from only the single tissue Matrix eQTL method (**A, C**) or from a union of both the single and multi-tissue eQTL analysis methods (**B, D**). For the analysis we used all genome-wide significant GWAS SNPs collated from the NHGRI catalog and the PheGenI database for >600 complex diseases and traits, that we found to be in high linkage disequilibrium (LD) (r^2≥0.80) with ≥1 significant eQTL (at FDR<0.05) across the 9 GTEx pilot tissues. The numbers in the row headers, *x* (gray), corresponding to each column, refer to the number of target genes regulated by eQTLs from ≥1 of 9 pilot tissues that are in LD (r^2≥0.80) with a GWAS SNP. The numbers in the column labels, *y* (gray), corresponding to each row, refer to the number of eQTL targets that lie in a GWAS locus whose transcript start sites (TSS) are not the closest TSS to the GWAS SNP or any SNPs in LD to it (r^2≥0.80; called proxy SNPs). Hence, the numbers in the table cells refer to the number of GWAS loci for which *y* out of the *x* eQTL target genes in its locus are not the most proximal gene to the GWAS SNP (distance was computed between the TSS of the target genes to any SNP in LD to the GWAS SNP). The pink cells refer to the number of GWAS loci with a subset of novel targets suggested by GTEx eQTLs and the yellow cells refer to number of GWAS loci with totally novel targets suggested by eQTLs.

A

Single-tissue analysis, protein-coding target genes

	# gene targets regulated by eQTLs in LD to GWAS SNP (x*)*					
# eQTL target genes whose TSS are not the most proximal to GWAS SNPs or proxy SNPs (*y)*	1	2	3	4	5	6
0	132	6	0	0	0	0
1	55	11	0	0	0	0
2		6	3	1	0	0
3			1	2	1	0
4				1	4	0
5					0	6
6						0

B

Single-tissue analysis, protein-coding target genes

	# gene targets regulated by eQTLs in LD to GWAS SNP (x*)*					
# eQTL target genes whose TSS are not the most proximal to GWAS SNPs or proxy SNPs (*y)*	1	2	3	4	5	6
0	132	6	0	0	0	0
1	55	11	0	0	0	0
2		6	3	1	0	0
3			1	2	1	0
4				1	4	0
5					0	6
6						0

C

Single-tissue analysis, all target genes (including non-protein coding)

# eQTL target genes whose TSS are not the most proximal to GWAS SNPs or proxy SNPs (*y*)	# gene targets regulated by eQTLs in LD to GWAS SNP (x*)*								
	1	2	3	4	5	6	7	8	9
0	125	17	4	0	0	0	0	0	0
1	65	24	5	2	0	0	0	0	0
2		3	3	3	1	0	0	0	0
3			0	1	2	2	0	0	0
4				0	1	1	0	0	0
5					0	1	2	1	1
6						0	0	0	4
7							0	0	0
8								0	0
9									0

D

Union of eQTLs from single-tissue analysis and multi-tissue analysis, all target genes (including non-protein coding)

# eQTL target genes whose TSS are not the most proximal to GWAS SNPs or proxy SNPs (*y*)	# gene targets regulated by eQTLs in LD to GWAS SNP (*x)*								
	1	2	3	4	5	6	7	8	9
0	133	19	4	0	0	0	0	0	0
1	98	22	3	1	1	1	0	0	0
2		17	6	2	0	0	0	0	0
3			2	0	1	0	0	0	0
4				4	2	1	0	0	0
5				0	2	2	0	0	0
6						1	2	0	0
7							0	0	1
8								0	5
9									0

Table S15. Fraction of gene biotypes implicated by eQTLs in LD to GWAS SNPs.

Biotype	**Gene type**	**# candidate genes proposed by single tissue eQTLs in GWAS loci***	**% Biotypes**	**# candidate genes proposed by multi-tissue eQTLs in GWAS loci***	**% Biotypes**	**# of all eGenes across 9 tissues**	**% Biotypes**
Protein coding	Protein coding	112	77.6%	114	79.4%	4422	70.5%
	Polymorphic pseudogene	0.5		0		11	
Long noncoding	lincRNA	10	13.4%	6	12.2%	531	18.4%
	Antisense	6		9		446	
	Processed transcript	3		2		173	
	3prime overlapping ncrna	1		1		6	
Pseudogene	Pseudogene	13	9%	12	8.4%	697	11.1%
Total:		**146**		**144**		**6,286**	

* The number of gene types of the target genes of eQTLs in linkage disequilibrium (LD) to genome-wide significant GWAS SNPs were normalized to the number of eQTL target genes per GWAS locus, rounded to nearest one. 211 genome-wide significant GWAS SNPs taken from the NHGRI GWAS catalog and NCBI's PheGenI database that were in LD (r2>0.8) with at least one best eQTL per gene (at FDR<5%) from either the single tissue analysis or multi-tissue analyses across the 9 tissues (in Table S12) were analyzed. The full list of significant eGenes from all 9 tissues contains additional gene types, however, only gene types of eQTL target genes in GWAS loci were included here (total unique number of eGenes is 6,486). For the full list of gene types in GENCODE v12 see: http://useast.ensembl.org/Help/Glossary?id=275.

References

1. D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, H. Parkinson, The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014). Medline doi:10.1093/nar/gkt1229

2. P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012). Medline doi:10.1016/j.ajhg.2011.11.029

3. B. E. Stranger, E. A. Stahl, T. Raj, Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–383 (2011). Medline doi:10.1534/genetics.110.120907

4. L. D. Ward, M. Kellis, Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012). Medline doi:10.1038/nbt.2422

5. M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, J. A. Stamatoyannopoulos, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012). Medline doi:10.1126/science.1222794

6. M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, V. G. Cheung, Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004). Medline doi:10.1038/nature02797

7. H. J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, J. E. Powell, A. Zhernakova, D. V. Zhernakova, J. H. Veldink, L. H. Van den Berg, J. Karjalainen, S. Withoff, A. G. Uitterlinden, A. Hofman, F. Rivadeneira, P. A. ’t Hoen, E. Reinmaa, K. Fischer, M. Nelis, L. Milani, D. Melzer, L. Ferrucci, A. B. Singleton, D. G. Hernandez, M. A. Nalls, G. Homuth, M. Nauck, D. Radke, U. Völker, M. Perola, V. Salomaa, J. Brody, A. Suchy-Dicey, S. A. Gharib, D. A. Enquobahrie, T. Lumley, G. W. Montgomery, S. Makino, H. Prokisch, C.

Herder, M. Roden, H. Grallert, T. Meitinger, K. Strauch, Y. Li, R. C. Jansen, P. M. Visscher, J. C. Knight, B. M. Psaty, S. Ripatti, A. Teumer, T. M. Frayling, A. Metspalu, J. B. van Meurs, L. Franke, Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013). Medline doi:10.1038/ng.2756

8. E. Grundberg, K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T. P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S. Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, T. D. Spector, Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012). Medline doi:10.1038/ng.2394

9. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011). Medline doi:10.1038/nature09906

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

11. B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, L. J. McCulloch, T. Ferreira, H. Grallert, N. Amin, G. Wu, C. J. Willer, S. Raychaudhuri, S. A. McCarroll, C. Langenberg, O. M. Hofmann, J. Dupuis, L. Qi, A. V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A. J. Bennett, R. Blagieva, E. Boerwinkle, L. L. Bonnycastle, K. Bengtsson Boström, B. Bravenboer, S. Bumpstead, N. P. Burtt, G. Charpentier, P. S. Chines, M. Cornelis, D. J. Couper, G. Crawford, A. S. Doney, K. S. Elliott, A. L. Elliott, M. R. Erdos, C. S. Fox, C. S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C. J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A. U.

Jackson, P. R. Johnson, T. Jørgensen, W. H. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C. M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M. A. Morken, N. Narisu, P. Nilsson, K. R. Owen, F. Payne, J. R. Perry, A. K. Petersen, C. Platou, C. Proença, I. Prokopenko, W. Rathmann, N. W. Rayner, N. R. Robertson, G. Rocheleau, M. Roden, M. J. Sampson, R. Saxena, B. M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H. M. Stringham, Q. Sun, A. J. Swift, B. Thorand, J. Tichet, T. Tuomi, R. M. van Dam, T. W. van Haeften, T. van Herpt, J. V. van Vliet-Ostaptchouk, G. B. Walters, M. N. Weedon, C. Wijmenga, J. Witteman, R. N. Bergman, S. Cauchi, F. S. Collins, A. L. Gloyn, U. Gyllensten, T. Hansen, W. A. Hide, G. A. Hitman, A. Hofman, D. J. Hunter, K. Hveem, M. Laakso, K. L. Mohlke, A. D. Morris, C. N. Palmer, P. P. Pramstaller, I. Rudan, E. Sijbrands, L. D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N. J. Wareham, R. M. Watanabe, G. R. Abecasis, B. O. Boehm, H. Campbell, M. J. Daly, A. T. Hattersley, F. B. Hu, J. B. Meigs, J. S. Pankow, O. Pedersen, H. E. Wichmann, I. Barroso, J. C. Florez, T. M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J. F. Wilson, T. Illig, P. Froguel, C. M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M. I. McCarthy, Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010). Medline doi:10.1038/ng.609

12. K. S. Small, A. K. Hedman, E. Grundberg, A. C. Nica, G. Thorleifsson, A. Kong, U. Thorsteindottir, S. Y. Shin, H. B. Richards, N. Soranzo, K. R. Ahmadi, C. M. Lindgren, K. Stefansson, E. T. Dermitzakis, P. Deloukas, T. D. Spector, M. I. McCarthy, Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**, 561–564 (2011). Medline doi:10.1038/ng.833

13. GTEx Consortium, The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013). Medline doi:10.1038/ng.2653

14. See supplementary materials on *Science* Online.

15. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M.

Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, T. J. Hubbard, GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012). Medline doi:10.1101/gr.135350.111

16. H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, N. Sestan, Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011). Medline doi:10.1038/nature10523

17. M. Melé *et al*., Human transcriptional variation across individuals and tissues. *Science* **348**, 660 (2015). doi:10.1126/science.aaa0355

18. N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, L. J. Lee, V. Slobodeniuc, C. Kutter, S. Watt, R. Colak, T. Kim, C. M. Misquitta-Ali, M. D. Wilson, P. M. Kim, D. T. Odom, B. J. Frey, B. J. Blencowe, The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587–1593 (2012). Medline doi:10.1126/science.1230612

19. G. Yeo, D. Holste, G. Kreiman, C. B. Burge, Variation in alternative splicing across human tissues. *Genome Biol.* **5**, R74 (2004). Medline doi:10.1186/gb-2004-5-10-r74

20. A. S. Dimas, S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel, H. Attar-Cohen, C. Ingle, C. Beazley, M. Gutierrez Arcelus, M. Sekowska, M. Gagnebin, J. Nisbett, P. Deloukas, E. T. Dermitzakis, S. E. Antonarakis, Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009). Medline doi:10.1126/science.1174148

21. A. A. Shabalin, Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012). Medline doi:10.1093/bioinformatics/bts163

22. A. C. Nica, L. Parts, D. Glass, J. Nisbet, A. Barrett, M. Sekowska, M. Travers, S. Potter, E. Grundberg, K. Small, A. K. Hedman, V. Bataille, J. Tzenova Bell, G. Surdulescu, A. S. Dimas, C. Ingle, F. O. Nestle, P. di Meglio, J. L. Min, A. Wilk, C. J. Hammond, N.

Hassanali, T. P. Yang, S. B. Montgomery, S. O'Rahilly, C. M. Lindgren, K. T. Zondervan, N. Soranzo, I. Barroso, R. Durbin, K. Ahmadi, P. Deloukas, M. I. McCarthy, E. T. Dermitzakis, T. D. Spector, The architecture of gene regulatory variation across multiple human tissues: The MuTHER study. *PLOS Genet.* **7**, e1002003 (2011). Medline doi:10.1371/journal.pgen.1002003

23. J. B. Veyrieras, S. Kudaravalli, S. Y. Kim, E. T. Dermitzakis, Y. Gilad, M. Stephens, J. K. Pritchard, High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLOS Genet.* **4**, e1000214 (2008). Medline doi:10.1371/journal.pgen.1000214

24. T. Flutre, X. Wen, J. Pritchard, M. Stephens, A statistical framework for joint eQTL analysis in multiple tissues. *PLOS Genet.* **9**, e1003486 (2013). Medline doi:10.1371/journal.pgen.1003486

25. G. Li, A. A. Shabalin, I. Rusyn, F. A. Wright, A. B. Nobel, http://arxiv.org/abs/1311.2948 (2013).

26. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003). Medline doi:10.1073/pnas.1530509100

27. X. Wen, http://arxiv.org/abs/1311.3981 (2013).

28. X. Wen, Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70**, 73–83 (2014). Medline doi:10.1111/biom.12112

29. J. H. Sul, B. Han, C. Ye, T. Choi, E. Eskin, Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLOS Genet.* **9**, e1003491 (2013). Medline doi:10.1371/journal.pgen.1003491

30. C. Newton-Cheh, M. Eijgelsheim, K. M. Rice, P. I. de Bakker, X. Yin, K. Estrada, J. C. Bis, K. Marciante, F. Rivadeneira, P. A. Noseworthy, N. Sotoodehnia, N. L. Smith, J. I. Rotter, J. A. Kors, J. C. Witteman, A. Hofman, S. R. Heckbert, C. J. O'Donnell, A. G. Uitterlinden, B. M. Psaty, T. Lumley, M. G. Larson, B. H. Stricker, Common variants at ten loci influence QT interval duration in the QTGEN Study. *Nat. Genet.* **41**, 399–406 (2009). Medline doi:10.1038/ng.364

31. M. A. Rivas *et al.*, Impact of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 665 (2015).

32. E. Khurana, Y. Fu, V. Colonna, X. J. Mu, H. M. Kang, T. Lappalainen, A. Sboner, L. Lochovsky, J. Chen, A. Harmanci, J. Das, A. Abyzov, S. Balasubramanian, K. Beal, D. Chakravarty, D. Challis, Y. Chen, D. Clarke, L. Clarke, F. Cunningham, U. S. Evani, P. Flicek, R. Fragoza, E. Garrison, R. Gibbs, Z. H. Gümüs, J. Herrero, N. Kitabayashi, Y. Kong, K. Lage, V. Liluashvili, S. M. Lipkin, D. G. MacArthur, G. Marth, D. Muzny, T. H. Pers, G. R. Ritchie, J. A. Rosenfeld, C. Sisu, X. Wei, M. Wilson, Y. Xue, F. Yu, E. T. Dermitzakis, H. Yu, M. A. Rubin, C. Tyler-Smith, M. Gerstein, Integrative annotation of variants from 1092 humans: Application to cancer genomics. *Science* **342**, 1235587 (2013). Medline doi:10.1126/science.1235587

33. A. Chess, Mechanisms and consequences of widespread random monoallelic expression. *Nat. Rev. Genet.* **13**, 421–428 (2012). Medline doi:10.1038/nrg3239

34. A. Buil, A. A. Brown, T. Lappalainen, A. Viñuela, M. N. Davies, H. F. Zheng, J. B. Richards, D. Glass, K. S. Small, R. Durbin, T. D. Spector, E. T. Dermitzakis, Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015). Medline doi:10.1038/ng.3162

35. H. Ongen, E.T. Dermitzakis, http://biorxiv.org/content/early/2015/01/22/014126 (2015).

36. J. Monlong, M. Calvo, P. G. Ferreira, R. Guigó, Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.* **5**, 4698 (2014). Medline doi:10.1038/ncomms5698

37. L. D. Ward, M. Kellis, HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012). Medline doi:10.1093/nar/gkr917

38. D. J. Gaffney, J. B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, J. K. Pritchard, Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012). Medline doi:10.1186/gb-2012-13-1-r7

39. A. Battle, S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman, C. McCormick, C. D. Haudenschild, K. B. Beckman, J. Shi, R. Mei, A. E. Urban, S. B. Montgomery, D. F. Levinson, D. Koller, Characterizing the genetic basis of transcriptome diversity through

RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014). Medline doi:10.1101/gr.155192.113

40. B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, J. A. Thomson, The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010). Medline doi:10.1038/nbt1010-1045

41. B. Zhang, C. Gaiteri, L. G. Bodea, Z. Wang, J. McElwee, A. A. Podtelezhnikov, C. Zhang, T. Xie, L. Tran, R. Dobrin, E. Fluder, B. Clurman, S. Melquist, M. Narayanan, C. Suver, H. Shah, M. Mahajan, T. Gillis, J. Mysore, M. E. MacDonald, J. R. Lamb, D. A. Bennett, C. Molony, D. J. Stone, V. Gudnason, A. J. Myers, E. E. Schadt, H. Neumann, J. Zhu, V. Emilsson, Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707–720 (2013). Medline doi:10.1016/j.cell.2013.03.030

42. M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, M. Snyder, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012). Medline doi:10.1038/nature11245

43. D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C.

Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, C. Tyler-Smith, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012). Medline doi:10.1126/science.1215040

44. F. E. Dewey, M. E. Grove, C. Pan, B. A. Goldstein, J. A. Bernstein, H. Chaib, J. D. Merker, R. L. Goldfeder, G. M. Enns, S. P. David, N. Pakdaman, K. E. Ormond, C. Caleshu, K. Kingham, T. E. Klein, M. Whirl-Carrillo, K. Sakamoto, M. T. Wheeler, A. J. Butte, J. M. Ford, L. Boxer, J. P. Ioannidis, A. C. Yeung, R. B. Altman, T. L. Assimes, M. Snyder, E. A. Ashley, T. Quertermous, Clinical interpretation and implications of whole-genome sequencing. *JAMA* **311**, 1035–1045 (2014). Medline doi:10.1001/jama.2014.1717

45. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007). Medline doi:10.1038/nature05911

46. D. L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M. E. Dolan, N. J. Cox, Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLOS Genet.* **6**, e1000888 (2010). Medline doi:10.1371/journal.pgen.1000888

47. A. C. Nica, S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley, I. Barroso, E. T. Dermitzakis, Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLOS Genet.* **6**, e1000895 (2010). Medline doi:10.1371/journal.pgen.1000895

48. E. M. Ramos, D. Hoffman, H. A. Junkins, D. Maglott, L. Phan, S. T. Sherry, M. Feolo, L. A. Hindorff, Phenotype-Genotype Integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22**, 144–147 (2014). Medline doi:10.1038/ejhg.2013.96

49. L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, T. A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 9362–9367 (2009). Medline doi:10.1073/pnas.0903103106

50. J. I. Goldstein, A. Crenshaw, J. Carey, G. B. Grant, J. Maguire, M. Fromer, C. O’Dushlaine, J. L. Moran, K. Chambert, C. Stevens, P. Sklar, C. M. Hultman, S. Purcell, S. A.

McCarroll, P. F. Sullivan, M. J. Daly, B. M. Neale, zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012). Medline doi:10.1093/bioinformatics/bts479

51. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). Medline doi:10.1086/519795

52. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006). Medline doi:10.1038/ng1847

53. J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007). Medline doi:10.1038/ng2088

54. 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). Medline doi:10.1371/journal.pgen.1000529

55. 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, G. A. McVean, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010). Medline

56. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). Medline

57. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009). Medline doi:10.1093/bioinformatics/btp120

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

59. 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). Medline doi:10.1093/bioinformatics/bts196

60. F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y. H. Zhou, A. Abdellaoui, S. Batista, C. Butler, G. Chen, T. H. Chen, D. D’Ambrosio, P. Gallins, M. J. Ha, J. J. Hottenga, S. Huang, M. Kattenberg, J. Kochar, C. M. Middeldorp, A. Qu, A. Shabalin, J. Tischfield, L. Todd, J. Y. Tzeng, G. van Grootheest, J. M. Vink, Q. Wang, W. Wang, W. Wang, G. Willemsen, J. H. Smit, E. J. de Geus, Z. Yin, B. W. Penninx, D. I. Boomsma, Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014). Medline doi:10.1038/ng.2951

61. International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010). Medline doi:10.1038/nature09298

62. 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). Medline doi:10.1038/nprot.2011.457

63. S. B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle, J. Nisbett, R. Guigo, E. T. Dermitzakis, Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010). Medline doi:10.1038/nature08903

64. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

65. J. D. Storey, A direct approach to false discovery rates. *J. R. Stat. Soc. B* **64**, 479–498 (2002). doi:10.1111/1467-9868.00346

66. B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, E. T. Dermitzakis, Patterns of cis regulatory

variation in diverse human populations. *PLOS Genet.* **8**, e1002639 (2012). Medline doi:10.1371/journal.pgen.1002639

67. T. Lappalainen, M. Sammeth, M. R. Friedländer, P. A. ’t Hoen, J. Monlong, M. A. Rivas, M. Gonzàlez-Porta, N. Kurbatova, T. Griebel, P. G. Ferreira, M. Barann, T. Wieland, L. Greger, M. van Iterson, J. Almlöf, P. Ribeca, I. Pulyakhina, D. Esser, T. Giger, A. Tikhonov, M. Sultan, G. Bertier, D. G. MacArthur, M. Lek, E. Lizano, H. P. Buermans, I. Padioleau, T. Schwarzmayr, O. Karlberg, H. Ongen, H. Kilpinen, S. Beltran, M. Gut, K. Kahlem, V. Amstislavskiy, O. Stegle, M. Pirinen, S. B. Montgomery, P. Donnelly, M. I. McCarthy, P. Flicek, T. M. Strom, H. Lehrach, S. Schreiber, R. Sudbrak, A. Carracedo, S. E. Antonarakis, R. Häsler, A. C. Syvänen, G. J. van Ommen, A. Brazma, T. Meitinger, P. Rosenstiel, R. Guigó, I. G. Gut, X. Estivill, E. T. Dermitzakis, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013). Medline doi:10.1038/nature12531

68. N. I. Panousis, M. Gutierrez-Arcelus, E. T. Dermitzakis, T. Lappalainen, Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. *Genome Biol.* **15**, 467 (2014). Medline doi:10.1186/s13059-014-0467-2

69. M. J. Anderson, A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26**, 32–46 (2001).

70. J. Oksanen *et al.*, Vegan: Community Ecology Package. R package version 1.15-1 (2008) (http://CRAN.Rproject. org/package=vegan).

71. J. Storey, *qvalue: Q-value estimation for false discovery rate control*, R package version 2.0.0 (2015); http://qvalue.princeton.edu..

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

73. 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. 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, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). Medline doi:10.1038/nature14248

74. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

75. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, e17 (2005). Medline doi:10.2202/1544-6115.1128

76. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabási, Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002). Medline doi:10.1126/science.1073374

77. P. Langfelder, B. Zhang, S. Horvath, Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008). Medline doi:10.1093/bioinformatics/btm563

78. L. E. Maquat, W. Y. Tarn, O. Isken, The pioneer round of translation: Features and functions. *Cell* **142**, 368–374 (2010). Medline doi:10.1016/j.cell.2010.07.022

79. E. Nagy, L. E. Maquat, A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. *Trends Biochem. Sci.* **23**, 198–199 (1998). Medline doi:10.1016/S0968-0004(98)01208-0

80. M. Pirinen, T. Lappalainen, N.A. Zaitlen, GTEx Consortium, E.T. Dermitzakis, P. Donnelly, M. I. McCarthy, M. A. Rivas, BioRxiv 10.1101/007211 (2014).

81. L. E. Lim, F. Duclos, O. Broux, N. Bourg, Y. Sunada, V. Allamand, J. Meyer, I. Richard, C. Moomaw, C. Slaughter, F. M. S. Tomé, M. Fardeau, C. E. Jackson, J. S. Beckmann, K. P. Campbell, β-Sarcoglycan: Characterization and role in limb-girdle muscular dystrophy linked to 4q12. *Nat. Genet.* **11**, 257–265 (1995). Medline doi:10.1038/ng1195-257

82. L. Jostins, S. Ripke, R. K. Weersma, R. H. Duerr, D. P. McGovern, K. Y. Hui, J. C. Lee, L. P. Schumm, Y. Sharma, C. A. Anderson, J. Essers, M. Mitrovic, K. Ning, I. Cleynen, E. Theatre, S. L. Spain, S. Raychaudhuri, P. Goyette, Z. Wei, C. Abraham, J. P. Achkar, T. Ahmad, L. Amininejad, A. N. Ananthakrishnan, V. Andersen, J. M. Andrews, L. Baidoo, T. Balschun, P. A. Bampton, A. Bitton, G. Boucher, S. Brand, C. Büning, A. Cohain, S. Cichon, M. D'Amato, D. De Jong, K. L. Devaney, M. Dubinsky, C. Edwards, D. Ellinghaus, L. R. Ferguson, D. Franchimont, K. Fransen, R. Gearry, M. Georges, C. Gieger, J. Glas, T. Haritunians, A. Hart, C. Hawkey, M. Hedl, X. Hu, T. H. Karlsen, L. Kupcinskas, S. Kugathasan, A. Latiano, D. Laukens, I. C. Lawrance, C. W. Lees, E. Louis, G. Mahy, J. Mansfield, A. R. Morgan, C. Mowat, W. Newman, O. Palmieri, C. Y. Ponsioen, U. Potocnik, N. J. Prescott, M. Regueiro, J. I. Rotter, R. K. Russell, J. D. Sanderson, M. Sans, J. Satsangi, S. Schreiber, L. A. Simms, J. Sventoraityte, S. R. Targan, K. D. Taylor, M. Tremelling, H. W. Verspaget, M. De Vos, C. Wijmenga, D. C. Wilson, J. Winkelmann, R. J. Xavier, S. Zeissig, B. Zhang, C. K. Zhang, H. Zhao, M. S. Silverberg, V. Annese, H. Hakonarson, S. R. Brant, G. Radford-Smith, C. G. Mathew, J. D. Rioux, E. E. Schadt, M. J. Daly, A. Franke, M. Parkes, S. Vermeire, J. C. Barrett, J. H. Cho, Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012). Medline doi:10.1038/nature11582

83. S. Kugathasan, R. N. Baldassano, J. P. Bradfield, P. M. Sleiman, M. Imielinski, S. L. Guthery, S. Cucchiara, C. E. Kim, E. C. Frackelton, K. Annaiah, J. T. Glessner, E. Santa, T. Willson, A. W. Eckert, E. Bonkowski, J. L. Shaner, R. M. Smith, F. G. Otieno, N. Peterson, D. J. Abrams, R. M. Chiavacci, R. Grundmeier, P. Mamula, G. Tomer, D. A. Piccoli, D. S. Monos, V. Annese, L. A. Denson, S. F. Grant, H. Hakonarson, Loci on

20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40**, 1211–1215 (2008). Medline doi:10.1038/ng.203

84. C. A. Anderson, G. Boucher, C. W. Lees, A. Franke, M. D'Amato, K. D. Taylor, J. C. Lee, P. Goyette, M. Imielinski, A. Latiano, C. Lagacé, R. Scott, L. Amininejad, S. Bumpstead, L. Baidoo, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Büning, J. F. Colombel, L. A. Denson, M. De Vos, M. Dubinsky, C. Edwards, D. Ellinghaus, R. S. Fehrmann, J. A. Floyd, T. Florin, D. Franchimont, L. Franke, M. Georges, J. Glas, N. L. Glazer, S. L. Guthery, T. Haritunians, N. K. Hayward, J. P. Hugot, G. Jobin, D. Laukens, I. Lawrance, M. Lémann, A. Levine, C. Libioulle, E. Louis, D. P. McGovern, M. Milla, G. W. Montgomery, K. I. Morley, C. Mowat, A. Ng, W. Newman, R. A. Ophoff, L. Papi, O. Palmieri, L. Peyrin-Biroulet, J. Panés, A. Phillips, N. J. Prescott, D. D. Proctor, R. Roberts, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, P. Schumm, F. Seibold, Y. Sharma, L. A. Simms, M. Seielstad, A. H. Steinhart, S. R. Targan, L. H. van den Berg, M. Vatn, H. Verspaget, T. Walters, C. Wijmenga, D. C. Wilson, H. J. Westra, R. J. Xavier, Z. Z. Zhao, C. Y. Ponsioen, V. Andersen, L. Torkvist, M. Gazouli, N. P. Anagnou, T. H. Karlsen, L. Kupcinskas, J. Sventoraityte, J. C. Mansfield, S. Kugathasan, M. S. Silverberg, J. Halfvarson, J. I. Rotter, C. G. Mathew, A. M. Griffiths, R. Gearry, T. Ahmad, S. R. Brant, M. Chamaillard, J. Satsangi, J. H. Cho, S. Schreiber, M. J. Daly, J. C. Barrett, M. Parkes, V. Annese, H. Hakonarson, G. Radford-Smith, R. H. Duerr, S. Vermeire, R. K. Weersma, J. D. Rioux, Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011). Medline doi:10.1038/ng.764

85. J. A. Hardy, P. Wester, B. Winblad, C. Gezelius, G. Bring, A. Eriksson, The patients dying after long terminal phase have acidotic brains; implications for biochemical measurements on autopsy tissue. *J. Neural Transm.* **61**, 253–264 (1985). Medline doi:10.1007/BF01251916

86. H. Tomita, M. P. Vawter, D. M. Walsh, S. J. Evans, P. V. Choudary, J. Li, K. M. Overman, M. E. Atz, R. M. Myers, E. G. Jones, S. J. Watson, H. Akil, W. E. Bunney Jr., Effect of agonal and postmortem factors on gene expression profile: Quality control in microarray

analyses of postmortem human brain. *Biol. Psychiatry* **55**, 346–352 (2004). Medline doi:10.1016/j.biopsych.2003.10.013