Supplementary Materials

Table Of Contents

1. Recruitment and sample collection 3
2. Quality control 3
2.1 SNP array quality control 3
2.2 RNA-sequencing quality control 4
3. Quantifying ASE and detecting mapping bias 4
4. Correcting for potential confounders using HCP 5
4.1 The HCP model 5
4.2 Known covariates used in HCP 6
5. Replication of eQTLs 7
5.1 Replication of *cis*-eQTLs 7
5.2 Replication of *trans*-eQTLs 9
6. GWAS associated SNPs and *cis*-eQTLs 9
7. Effect size computation for eQTLs and ASE 10
8. *Trans*-eQTL detection and filtering 10
9. Intra-chromosomal analysis 11
10. Learning *trans*-eQTL regulatory network structure 12
11. Comparison of LRVM to other models 12

List of Figures

Figure S1. RNA-sequencing quality control.
Figure S2. Concordance between SNP array and RNA-seq called genotypes.
Figure S3. Mappability and distribution of mapped bases.
Figure S4. Ancestry and Principal Components of genotype data.
Figure S5. Correlation of genotype Principal Components and expression levels.
Figure S6. Evaluation of case/control status on eQTL detection.
Figure S7. Correlation of HCP factors with known covariates.
Figure S8. Removing the effects of confounding factors using HCP.
Figure S9. Bonferroni corrected versus permutation eQTL p-values.
Figure S10. Sample size and number of detected QTLs.
Figure S11. Replication of eQTLs in independent cohorts.
Figure S12. Variance in expression explained by genotype and phenotype.
Figure S13. *Cis*-eQTL allelic effects.
Figure S14. eQTL and ASE effect size.
Figure S15. *cis*-eQTLs and allelic imbalance.
Figure S16. Factors affecting validation of *cis*-eQTLs through ASE.
Figure S17. Co-expression of genes with a shared regulatory variant.
Figure S18. Example eQTL module, detail.
Figure S19. Co-regulation in topological domains.
Figure S20. Filtering for accurate identification of *trans*-eQTLs.
Figure S21. *Trans*-eQTL SNPs' effects on proximal genes.
Figure S22. Enrichment for low p-values for association of *rs2759386* with distal isoform ratios.
Figure S23. Example of a paradoxical *trans*-eQTL relationship.
Figure S24. Potential confounding factors in *trans*-eQTL detection.
Figure S25. Relationship between population structure and *trans*-eQTL detection.
Figure S26. eQTLs and selection pressure.
Figure S27. Genomic position of QTLs.
Figure S28. Latent Regulatory Variant Model (LRVM) for *cis*-eQTL prediction.
Figure S29. Application of LRVM to splicing QTLs.
Figure S30. LRVM sQTL prediction for GWAS SNPs.
Figure S31. Effect of distance from TSS on LRVM.

List of Tables

Table S1. Covariates used in RNA-seq normalization.
Table S2. Disease and trait-associated variants among eQTL hits.
Table S3. Individual traits with eQTL enrichment in GWAS hits.
Table S4. Genes and pathways correlated with demographic factors.

1. Recruitment and sample collection

This study was performed on 922 European individuals, recruited through a survey company called Knowledge Networks Inc (KN), for the Depression Susceptibility Genes and Networks Project (NIMH Grant 5RC2MH089916). Through a process involving an online questionnaire (relevant sections of CIDI-SF), KN identified potential candidates for this study. From this pool, 1,259 individuals actually went on to have their blood drawn, filled out consent forms, and were then telephone interviewed (SCID interview--- Structured Clinical Interview for DSM-IV covering depressive, bipolar, psychotic, alcohol, substance and anxiety disorders as well as family history of mood disorders). After excluding some of the eligible individuals for reported non-European ancestry or medical comorbidities, which were thought to be too unusual and significant for the gene expression analysis, and performing quality control described below, we actually analyzed 463 cases of Major Depressive Disorder and 459 controls (total of 922) individuals. The case/control aspect of this cohort does not have a noticeable impact on eQTL detection (Figure S6), so for this study we do not consider or analyze depression status.

2. Quality control

2.1 SNP array quality control

Genotype data was filtered for genotype quality as follows. QC was carried out simultaneously for this dataset and a second dataset (Genetics of Recurrent Early-Onset Depression, phase 2, unpublished data) as they were genotyped by the same lab on the same platform a few months apart. Pairwise estimates of IBD were computed in PLINK (Purcell et al. 2007) , and any sample duplicates were excluded. Principal components analysis (PCA) was carried out for all individuals using every fifth autosomal SNP (to reduce the influence of LD among SNPs), and the principal component scores examined for relationship to geographical/ethnic origin based on self-report (Figure S1). As shown in previous studies (Shi et al. 2009), PC1 was interpretable as a North-South gradient (Anglo-Saxon and Scandinavian to the North, Mediterranean and Ashkenazi Jewish to the South) and PC2 as East-West (from Russian/Slavic to Western European), with PC3 further separating Ashkenazi Jewish ancestry from other Mediterranean (Italian, Greek) ancestry. Individuals were excluded who were obvious outliers (by visual inspection of the plot of PC1 vs. PC2, or because multiple smaller components were numerical outliers) and PCA was repeated to ensure that there were no outliers by visual inspection of plots (Figure S4). Additionally, samples were excluded for elevated rates of heterozygosity (over 34.5% of SNPs) or if genotypes could not be called for more than 1.4% of SNPs. For SNPs, we evaluated QC metrics in each study, and retained SNPs with a missingness rate below 0.012, a 10% Gencall score above 0.55, and a Hardy-Weinberg probability below 0.0001.

We corrected for the sub-population structure observed from the genotype PCs, by regressing the expression of each gene against the top three genotype PCs (from normalized expression data, Section 4). However, we observed that removal of population signal does not significantly influence *cis*-eQTL discovery, and overall we lost only 11 genes from *cis*-eQTL results through this correction (Figure S5).

2.2 RNA-sequencing quality control

If pooled RNA-sequencing libraries did not produce at least 180M reads, sequencing or library preparation was repeated for all three individuals in the lane (Figure S2a). For each individual with more than one sequencing run, runs with sufficient reads (> 20M), good mappability (> 40%), and good reproducibility of quantified expression data with the other runs ($r^2 > 0.9$), were merged. The first base of each read was trimmed to account for the stronger sequencing biases at the beginning cycles before mapping (Figure S2b). Genotype calls were made from RNA-seq data based on loci with sufficient read depth, and compared to genotypes from the SNP array. Individuals with concordance below 0.85 were removed from the study as potentially mislabeled (Figure S3). Additional quality metrics were evaluated to ensure high RNA Integrity Numbers (RIN) (Figure S2c), low percentage of hemoglobin reads (Figure S2d), and a high proportion of mapped reads in each individual (Figure S4). The results of quality control analysis were also utilized in a data normalization step described in Section 4.

3. Quantifying ASE and detecting mapping bias

Allele-specific expression was quantified by measuring allelic ratio at genotyped, heterozygous loci, after filtering for loci with significant mapping bias evident through simulation. For each candidate SNP, we simulated 4 sets of 50bp reads containing all 50 reads covering the locus, for both reference and alternative allele using both strands. We assigned base quality scores to the simulated reads based on the per-position distribution observed in our real data, and aligned reads with the same settings as our standard pipeline. Finally, we evaluated the alignments around each locus using samtools mpileup (Li et al. 2009), and compared the mapping rates between reference alleles and alternative alleles. As observed in past studies, there was an overall bias towards higher mapping rates for the reference allele, and any individual locus with evidence of mapping bias in either direction was removed from further analysis. At the remaining loci, allelic ratio was then computed for each individual from reads over each allele using samtools mpileup (Li et al. 2009), and used in aseQTL detection. In addition, statistically significant ASE was called per-locus, per-individual, assuming a binomial distribution with binomial parameter *p* computed using each individual's personal reference bias. ASE calls per-individual were restricted to loci with 30 or more reads, at least one read observed per allele, and a read fraction of at least 1e-3 for each allele.

4. Correcting for potential confounders using HCP

We and others have observed that major components of expression variability can often be attributed to technical factors that introduce unwanted, systematic variability in data (Leek and Storey 2007). Such unwanted, systematic variability leads to spurious correlations in the data (among genes or samples), and results in both false positive (Kang et al. 2008; Listgarten et al. 2010) and false negative associations (Stegle et al. 2012). Importantly, several eQTL studies have demonstrated that removing such unwanted, systematic variability dramatically improves the power to detect *cis*-eQTLs, and also leads to more consistent discoveries among different datasets (Stegle et al. 2012). Consistent with these previous observations, we also observed variability in RNA-sequencing data that could be attributed to measurable technical factors (Figure S8c). Therefore, we devised a method, called HCP (Hidden Covariates with Prior)(Mostafavi et al. 2013), for correcting for such systematic unwanted variability in RNA-seq data, capturing both known technical covariates and any major broad trend in the data even if it arises from an unobserved source. As shown in Figure S8a, we observed a dramatic increase in the number of *cis*-eQTL and validated co-expression associations after correcting for confounding structure using HCP.

4.1 The HCP model

Intuitively, as described in details below, HCP can potentially correct for two types of confounding factors: known and *latent* (hidden) confounding factors that introduce unwanted variability. A key component of HCP is an informed assumption about variability patterns introduced by hidden factors: given a set of potential confounding factors (or known covariates) (see Table S1 for the list of potential confounding factors in this study), HCP infers hidden factors that are informed by a prior based on the known factors. However, the strength of this prior can be adjusted (through parameter tuning) to remove different amounts of systematic variability attributed to unobserved factors, as appropriate for different analyses. For example, removing too much variability may impede the detection of *trans*-eQTLs that affect many genes, whereas aggressive normalization may improve detection of *cis*-eQTLs and other narrow effects. Therefore, as we describe below, we set HCP's parameter for *cis*- and *trans*-eQTL detection in two different ways.

In particular, given a matrix of *logarithm*-transformed read counts, represented by Y of size n x g where n is the number of individuals and g is the number of genes, HCP models *hidden* confounding factors Z and their effect per gene as follows:

$$Y_{:,i} \sim N\left(Z \times W_{:,i}, 1\right) \quad (1)$$

$$Z_{:,k} \sim N\left(F \times U_{:,k}, \sigma_1 I\right), U_{:,k} \sim N(0, \sigma_2 I), W_{:,i} \sim N(0, \sigma_3 I) \quad (2),$$

where $Y_{:,i}$ (column i of Y) is the log of expression values for gene i in all n individuals, Z is a matrix of inferred hidden factors and has k columns (k being the number of hidden factors, and $Z_{:,k}$ is the k^{th} column of Z), $W_{:,i}$ is a vector that represents the linear effect of hidden factors on gene i, and F is a matrix of known confounding factors. We estimate the unknown Z, W, and U using maximum likelihood, and set the hyper-parameters k, σ_1, σ_2, and σ_3, using the procedure described below.

The model hyper-parameters k, σ_1, σ_2, and σ_3, allow us to adjust the flexibility of HCP in identifying broad versus narrower effects in the expression data, and we set these parameters by optimizing two different criteria for *cis*- and *trans*-eQTL detection. In particular, one could set these hyper-parameters to optimize the number of *cis*- or *trans*-eQTLs directly, however, such optimization will likely lead to overfitting the data. Therefore, for *cis*-eQTL analysis, we set the model hyper-parameters by optimizing the total number of *cis*-eQTLs for genes on only one (training) chromosome (chromosome 18), and evaluate the model on a single test chromosome (chromosome 14) (Figure S8a). For *trans*-eQTL analysis, the scarcity of significant associations makes it difficult to directly tune parameters without risk of over-fitting. Thus, we optimize the parameter settings to maximize the number of significant associations in *expression profiles* among a set of transcription factors (TF) and their known targets (TF target information was obtained from the ChEA database(Lachmann et al. 2010)) (Figure S8b).

4.2 Known covariates used in HCP

We provided HCP with a set of 35 observed covariates (Table S1). We found that the inferred HCP factors summarize multiple correlated known confounding factors (Figure S7) with significant contribution to variability in the RNA-sequencing data. Generally, we observed that the top HCP factors largely correspond to technical factors specific to RNA-seq such as sequencing depth, percent duplicated reads, and others obtained from the Picard metrics and to a much lesser extent to biological factors such as time of blood drawn and estimates of cell type frequencies (described below) (Figure S7).

Whole blood is a mixture of cell-types, and differences in cell-type frequencies in individuals could potentially lead to unwanted expression variability. Therefore, we estimated blood cell type frequencies and used these estimates as covariates in HCP. To estimate cell type frequencies, we obtained cell type signatures for blood cell types as in (Abbas et al. 2009) (only 9 of the original 17 cell types from (Abbas et al. 2009) were assigned a non-zero frequency in any of our subjects (see Table S1 for the names of the 9 cell types)). We then estimate per-individual cell type frequencies by using non-negative least squares regression, and regressing the observed expression data that reflect a mixture of cells for a given individual onto the cell-type-specific expression signatures. In particular, given the observed log (un-normalized) expression values A_i of size s x 1 for s genes in individual i (defined below), we estimated cell type frequencies in individual i as follows:

$A_i \sim N(X \times C_i, I)$, s.t. $C_i \geq 0$ (3)

Where X is a matrix of size s x 9, s being the number of genes that are in the cell signatures: each column k of X represents the expression levels of the s genes in cell type k. The vector C_i then represents the frequencies of each of the nine cell types in individual i.

For the computation of percent variance explained (PVE) using genotype or phenotype data (Figure S12), we only correct the raw data for the known covariates and do not use the full HCP model. To do so, we perform linear regression for each gene using all the 35 known covariates, and use the residual of the regression.

5. Replication of eQTLs

5.1 Replication of *cis*-eQTLs

In order to examine overlap of our eQTLs with previous studies, we obtained two existing large *cis*-eQTL datasets: (1) the MuTHER study (Grundberg et al. 2012) provides the complete list of tested *cis* SNP-gene p-values (http://www.muther.ac.uk/), and (2) the Fehrmann study (Fehrmann et al. 2011) provides a list of significant SNP-gene pairs at 0.05 FDR.

Given the differences in statistical power in different datasets, we compare our results using the π_1 statistic (Storey and Tibshirani 2003), which estimates the total proportion of non-null hypotheses, and provides a more robust alternative to simple overlap of statistically significant results according to an arbitrary threshold. The Fehrmann et al. study (whole blood, 1469 European individuals) discovered 29,049 SNP-gene pairs (corresponding to 4,904 genes) significant at FDR 0.05 that were also *tested* in our data. Because we do not have the entire SNP-gene p-values for this study, we can only report π_1 of their *significant* eQTLs as assessed in our study. We estimate π_1=0.88, which is a value slightly higher than previously reported eQTL replication rates in microarray studies (Grundberg et al. 2012).

The MuTHER data consists of all tested SNP-gene p-values for adipose, LCL, and skin, obtained from a cohort of Caucasian female twins. We identified MuTHER *cis*-eQTLs using the same multiple testing correction procedure that was used in our study (0.05 FDR at gene level, applied after a Bonferroni correction for number of SNPs tested per gene). Replication rates of MuTHER eQTLs in our data are π_1=0.73, 0.74, and 0.66, for adipose, LCL, and skin, respectively, and we observe a higher value, π_1=0.89, when only considering eQTLs that are shared among the three tissues. The replication rates of our eQTLs into MuTHER are π_1=0.51, 0.58, 0.56, for adipose,

LCL, and skin, respectively. These replication rates are slightly lower than those reported by (Grundberg et al. 2012) (ranging between 0.6-0.7) for replication of their eQTLs in the same tissue. For a comparison, we also investigated the replication rate of (Fehrmann et al. 2011) in MuTHER data, and observed slightly higher replication rates: 0.64, 0.73, 0.68, for adipose, LCL, and fat, respectively. This may be expected given major differences in technologies (MuTHER and Fehrmann use microarrays, whereas we are using RNA-sequencing), though, again, the replication of Fehrmann et al. into our data at π_1=0.88, using the same tissue, is ultimately the highest of the measured rates.

Additionally, to better understand the factors that affect replication rate of genes we investigated (1) residual correlation of expression data with cell type proportions that may have been insufficiently corrected by our normalization procedure and could lead to spurious associations, (2) impact of effect size and number of individuals (summarized by regression coefficient), and (3) average expression levels of genes. For this analysis, we assigned a *reproducibility* number {0,1,2,3,4} to each eQTL, based on the number of "studies" in which the eQTL was reported as significant (using a 0.05 FDR threshold reported in the original studies), where the four investigated studies are the three tissues from Grundberg et al. (each counted as a separate study) and eQTLs from Fehrmann et al. We only considered gene-SNP pairs that were tested in our study and all three tissues in Grundberg et al., and considered the best SNP for each eQTL gene (according to this criteria 8208 genes had an eQTL in our study). First, to ensure that failure of replication is not influenced by residual correlations with cell type proportions, we estimated the spearman correlation coefficient between our cell type estimates and compared this estimate to reproducibility of eQTLs (Figure S11c). As shown, we did not observe a correlation between strength of correlation with cell-type and reproducibility ($p>0.1$). In contrast, we observed a significant correlation ($p<1e-20$) between average expression levels and reproducibility (Figure S11b), and between strength of the association (Spearman rho) and reproducibility ($p<1e-100$) (Figure S11a). Therefore, these results suggest that the slightly lower replication rate of our eQTLs in MuTHER data, compared to rates previously reported by (Grundberg et al. 2012) are likely due to differences in technology (better quantification of lowly expressed genes by RNA-seq, and very high depth of current study compared to previous RNA-seq studies), statistical power, and differences in tissues. Finally, we also investigated the relationship between reproducibility and whether or not and eQTL SNP is exonic, looking for evidence that mapping errors due to genetic variation within coding regions could drive spurious eQTLs in RNA-seq data. We found that exonic eQTLs do not show any signs of lower reproducibility between our data and microarray studies (in fact we observe a slightly higher replication rate, with only 3% of non-replicated eQTLs being exonic, compared with 5% of replicated eQTLs, $p < 1e-4$). This suggests mapping errors are not a major source of differences between RNA-seq and microarray eQTLs.

5.2 Replication of *trans*-eQTLs

Here, we compared our *trans*-eQTLs to those of Fehrmann, as both datasets have a relatively large number of samples, and involve the same tissue (whole blood), though with a major difference in measurement technology (RNA-seq versus microarray) and in subject recruitment criteria.

Fehrmann et al. reported 396 SNP-gene pairs (95 genes) with a *trans*-association that were also tested in our data. 21 of these SNP-gene pairs were deemed genome-wide significant in our data, corresponding to 11 genes with a shared *trans*-eQTLs in both studies (significantly more than expected by chance, hypergeometric $p < 1e\text{-}100$). (Recall that this study identified 138 genes with an associated *trans*-eQTL). The replication rate of Fehrmann et al *trans*-eQTLs in our study, as measured by π_1=0.74, is higher than the replication rate of previously reported *trans*-eQTLs (Grundberg et al. 2012). To avoid multiple non-independent tests because of the LD structure, we only considered the best SNP per eQTL gene reported in Fehrmann et al. To ensure that this π_1 estimate is not inflated due to spurious associations between random SNP-gene pairs in our data, we also computed the π_1 statistic for 500 random sets of 95 SNP-gene pairs. Among these random controls, we observe a median π_1 of 0.0, and 95% falling below π_1=0.26, indicating the replication of *trans*-eQTLs is significantly higher than would be expected by chance. (Similarly we see no evidence of overall inflation in *trans*-eQTL associations, Figure S25). Unfortunately, we could not estimate π_1 for replication rate of our *trans*-eQTLs in Fehrmann et al., as only their significant (FDR 0.05) results were made publically available.

6. GWAS associated SNPs and *cis*-eQTLs

To determine whether certain disease categories were overrepresented in our GWAS eQTL hits, we performed two tests. Here, we used the first definition of GWAS eQTL from Table S2. First, we computed the enrichment of eQTLs for each disease category using a hypergeometric test (Table S3). Second, we curated 8 broad categories of diseases, and computed the enrichment of these broad categories compared to other diseases. In order to ensure that our enrichment our not biased because of LD patterns, we only considered SNPs that are at least 50Kb in distance for each disease.

In total, 232 diseases from the GWAS catalog (Hindorff et al. 2009) were associated with at least one eQTL. As shown in Table S2, 3 diseases were significantly enriched in eQTLs at 0.1 FDR, and 11 were significantly enriched at 0.2 FDR. Next, we considered 8 broad categories of diseases: neurological disorders, body size, pigmentation, metabolic, autoimmune diseases, cardiovascular traits, cancer, and blood chemistry. For each of these disease categories, we

identified relevant SNPs, and computed enrichment compared to that of other diseases. We identified two nominally significantly enriched categories: autoimmune (p-value 0.03) and blood chemistry (p-value 0.024), whereas the rest of these categories were not nominally significant (p>0.05).

7. Effect size computation for eQTLs and ASE

For each candidate SNP, we subsampled 100 individuals to create a balanced dataset with no more than 50 individuals represented by any single genotype (thus at 50 individuals will have each allele). Within this sample, we estimated significance and effect size using linear regression on the raw (non-normalized) expression data. For each gene, we identified the strongest SNP according to these sampled estimates and, for eQTLs significant at FDR 0.05, we recorded the effect size (expression fold change) and allele frequency for that SNP. We also used subsampling to compute ASE effect size (defined here the average fold change between the highly expressed allele and the less expressed allele). We sample two individuals who are heterozygous at both the candidate regulatory variant and a coding locus in the corresponding gene. Within each individual we subsampled reads to avoid biases based on read depth, drawing 30 reads at random from the reads observed for each individual and computing allelic fold change from this sample. We then average the fold change observed in these two individuals as an estimate of allelic effect of the regulatory variant.

8. *Trans*-eQTL detection and filtering

To account for the possibility of mapping errors introducing spurious *trans*-associations, we applied a series of filters to our candidate *trans*-eQTLs. In particular, in the unfiltered data, we did observe a moderate enrichment of known pseudogenes (12%, compared to 1% of overall expressed genes, p<1e-4), and specific associations between regions of sequence similarity (enrichment p-value for eQTLs arising between regions of sequence similarity p<1e-80). Such associations may arise between a SNP impacting the expression of a nearby gene, but where reads are incorrectly mapped to a distant gene similar in sequence. Such mapping errors are possible even among uniquely mapped reads for regions of high sequence similarity due to factors such as reference genome errors, polymorphisms, and sequencing errors. As a first step, we simply removed all pseudogenes in our *trans* analysis, and removed associations occurring between a gene in a known paralog family and any *cis*-regulatory SNP for another paralog in the same family.

Then, we directly identify regions of the genome with potential for mapping errors. We simulated 52 million 50bp reads covering known gene coding/UTR and pseudogene annotations in sliding windows. These reads were then trimmed for the first base, combined with two Illumina sequencing adaptors that are unmappable to human reference genome, and given a set of sequencing quality scores sampled from a position dependent distribution drawn from our real reads. The simulated reads were then aligned with TopHat (Trapnell et al. 2009) allowing up to 4 mismatches and multiple alignment. For each gene, we eliminated from consideration SNPs within 5Mb of blocks where simulated reads from the gene mapped with these lenient parameters.

Finally, real associations should affect expression of an entire transcript, evident by reads distributed smoothly across an entire gene or set of exons, whereas mapping errors among uniquely aligned reads are likely confined to a small number of positions within a gene. Thus, we developed a method for evaluating the "smoothness" of an eQTL association signal across the expressed exons of a gene. An example of a suspicious association identified by this method is shown in Figure S20a, S20b. In particular, we computed the effect size α_g of each eQTL at the gene level using linear regression, in addition to estimates of the effect size α_i based on reads covering each individual exonic position *i* within the gene. Covering reads were identified using samtools mpileup(Li et al. 2009). We removed any association where the final effect size at the gene level was not well represented across positions (where $\alpha_g > 4\times$ median (α_i)). We note that the *trans* associations filtered by paralog and cross-mapping steps were very likely to also fail this test. Figures S20c and S20d summarize the number of associations that were eliminated by each of our filters.

9. Intra-chromosomal analysis

To evaluate the enrichment of eQTL sharing in gene-dense regions, we estimated the number of other genes within 1Mb of the transcription start site of each gene, and compared this estimate of density between 1) genes with a shared regulatory variant and 2) genes with only non-shared variants, and found density to be significantly higher among the first group ($p<1e-74$). We downloaded normalized Hi-C contact estimates along with a discretization of each chromosome into *topological domains* based on measurements in human fibroblasts (Dixon et al. 2012). We evaluated *sharing* of regulatory variants by pairs of genes in this analysis by identifying SNPs nominally associated with both genes using a stringent p-value threshold of 1e-7, not necessarily requiring that they share their best SNP. This is a more inclusive definition of sharing than using the results of stepwise regression, but the stepwise method suffers from arbitrary and independent selection of SNP for each gene particularly among SNPs in strongly linked, highly associated regions. Using this definition, we record for every pair of genes located on the same chromosome whether they share any strong candidate regulatory variant. We then evaluated whether membership in the same topological domain is predictive of sharing of regulatory variants, by computing the conditional odds multiplier after controlling for linear distance between the TSS of the two genes.

10. Learning *trans*-eQTL regulatory network structure

We used a series of likelihood ratio tests to distinguish between three types of regulatory network structures involving a SNP acting in *cis*- to nearby genes and *trans*- to a distant gene. We performed this analysis considering only the best *trans*-eQTL for each gene. In particular, we first identified 74 *trans* acting SNPs that had a local effect on any *cis*-gene (1Mb distance threshold) using a p-value threshold of 0.05. For each of these 74 cases, we classified the regulatory network structure consisting of the *trans*-eQTL SNP, nearby effected *cis*-genes, and the target *trans*-gene as (1) full mediation, (2) partial mediation, or (3) independent. We perform all analysis on ranked data, to ensure consistency with Spearman's rank correlation used to obtain *trans*-eQTLs.

We first identified *full mediation* of the effects of a SNP through the expression of its *cis*-genes by comparing the likelihood of the model $p(\boldsymbol{t}|\boldsymbol{s},\boldsymbol{c})\sim N(\boldsymbol{t}|\mu+\boldsymbol{s}\alpha+\boldsymbol{c}\beta)$ to that of $p(\boldsymbol{t}|\boldsymbol{c})\sim N(\boldsymbol{t}|\mu+\boldsymbol{c}\beta)$, where $\boldsymbol{t}$ is a vector representing the expression level of the target (*trans*) gene, $\boldsymbol{c}$ is a vector or matrix consisting of the expression levels of the *cis*-gene(s) (only considering those with a correlation p-value of 0.05 with $\boldsymbol{t}$), $\boldsymbol{s}$ is a vector representing genotype $\in \{0,1,2\}$ at the *trans*-eQTL SNP, and μ,α,β are parameters that are estimated using maximum likelihood for each model separately. We classified instances as full mediation whereby the full model $p(\boldsymbol{t}|\boldsymbol{s},\boldsymbol{c})$, does not provide a significantly better fit compared to $p(\boldsymbol{t}|\boldsymbol{c})$ (again using a p-value threshold of 0.05).

For the remaining *trans*-eQTLs, we identified *partial mediation* by comparing $p(\boldsymbol{t}|\boldsymbol{s},\boldsymbol{c})\sim N(\boldsymbol{t}|\mu+\boldsymbol{s}\alpha+\boldsymbol{c}\beta)$ to that of $p(\boldsymbol{t}|\boldsymbol{s})\sim N(\boldsymbol{t}|\mu+\boldsymbol{s}\beta)$, and identifying instances where $p(\boldsymbol{t}|\boldsymbol{s},\boldsymbol{c})$ provides a significantly better fit compared to $p(\boldsymbol{t}|\boldsymbol{s})$. Instances that are not classified as either full or partial mediation are classified as *independent* effects. Finally, we also identified instances where the *predicted* directionality of *trans*-SNP to target effect (based on the mediated relationships) is not consistent with the observed directionality (we call such instances *paradoxical effects*) (see Figure S23). In particular, for each *trans*-eQTL s, we identify all *cis* genes $\{c_1,\ldots,c_k\}$ using a threshold of 0.05. Among these *cis*-genes, we identify the gene that has the highest correlation with the target gene t, say c_b. Then the predicted sign of correlation between s and t can be obtained as sign$(corr(c_b,t))$ x sign$(corr(s,c_b))$, and compared to sig$(corr(\boldsymbol{s},\boldsymbol{t}))$.

11. Comparison of LRVM to other models

We compared the performance of LRVM to two reduced models that explore the impact of modeling assumptions made in our method. In the first, we predict associations a directly from

features F using a logistic function. In the second, we incorporate a correction for minor allele frequency by including raw features F along with corrected features $m \times F$ into a logistic function to predict associations a. Neither baseline model includes inference of the latent variables d, which are only necessary in LRVM to model the effects of linkage disequilibrium. The results of this comparison are shown in Figure S28 and S29.

Additionally, to assess the impact of SNP position on LRVM, we evaluate different sets of genomic features. We trained LRVM-DO using only positional features (including whether a SNP falls in a particular intronic, exonic or UTR location), LRVM-ND using only non-positional genomic annotation features, and of course the full LRVM model. We compare each of these to predictions made based on positional features alone, not in the LRVM framework, and thus not accounting for MAF or LD. Results are shown in Figure S31.

In all evaluations of LRVM and baseline predictive models, we constructed our training set (used to estimate model parameters), from all odd numbered chromosomes, and our test set, used for evaluation only, from even numbered chromosomes. This split was chosen in order to reduce the potential for the training set to contain SNPs or genes highly informative of the test set (such as containing SNPs in LD between the two), while still having a similar distribution of regulatory effects in the two sets.

Methods related to LRVM include the Bayesian hierarchical model (BHM) presented in (Gaffney et al. 2012), (Lee et al. 2009), or even the simple scores provided by RegulomeDB (Boyle et al. 2012), with Gaffney being the most similar. Both BHM and LRVM assume a logistic function over genomic annotations as the backbone of predicting regulatory potential of each SNP. However, BHM primarily *compares* candidate SNPs in a competitive model within each eQTL region, and without access to the expression and genotype information for the relevant region, does not attempt to ultimately predict eQTL status for unseen variants or genes. Further, because BHM is competitive, it will identify the single best SNP in a region, and unlike LRVM, BHM does not attempt to predict cases where there are multiple, independent eQTN in the same region, which have been observed to be common (Stranger et al. 2012). Thus, the two models are motivated differently and will have different potential applications. LRVM is unique in its ability to train on association statistics rather than raw expression and genotype data (often not made publically available), and its ability to make eQTL predictions on genes and variants not included in the training data. We expect this to provide compelling use cases beyond existing methods, including training tissue- or population-specific models even for eQTL studies without full publicly released data, and prioritizing non-coding variants identified in disease studies, even if those variants weren't available during model training.

Supplementary Table S1. Known covariates used in HCP. Table lists all the known covariates (technical and biological) that were used as input the HCP method. 17 of these are technical factors (colored in yellow) directly obtained from Picard QC metrics. There are two technical factors relating to sample preparation obtained from the technicians (colored in orange), four other technical factors that we estimated from the quantified reads (colored in purple), nine factors are the inferred cell type frequencies (see Supplementary Materials), and one factor representing the time of the day that the individual's blood was drawn. individual-specific exon length, and individual-specific GC are estimated as the proportion of read variance in each individual (mapped to exons) that can be explained by GC compositions of the exons or the length of the exons, respectively---these factors are estimated per individual, by correlating mapped reads to exons with a vector of exon GC composition or exon lengths.

1	Sequencing Depth
2	Number of Coding Bases
3	Number of UTR Bases
4	Number of PF Aligned Bases
5	Number of BF Bases
6	Percent Coding Bases
7	Percent MRNA Bases
8	Percent Usable Bases
9	Percent UTR Bases
10	Cell Frequency: Mono
11	Cell Frequency: DC
12	RNA Yield
13	Cell Frequency: Neutro
14	individual-Specific Exon Length
15	Median 5PRime Biase
16	Median 5Prime to 3Prime Bias
17	Cell Frequency: NK Cells
18	Cell Frequency: Th Cells
19	Cell Frequency: Platelet Cells
20	Cell Frequency: NK_act Cells
21	Cell Frequency: DC_act Cells
22	Cell Frequency: B Cells
23	Time of Day Blood Drawn
24	Percent Hemoglobin
25	individual-Specific GC
26	Percent Duplicated Reads
27	Median 3Prime Bias
28	Median CV Coverage
29	Cell Frequency: Tc Cells
30	Globin Flag (Technician)
31	Number of Intergenic Bases
32	Number of Intronic Bases
33	Percent Intergenic Bases
34	Percent Intronic Bases
35	Cell Frequency: Tc_act Cells

Supplementary Table S2: Disease and trait-associated variants among eQTL hits. We evaluate the presence of regulatory associations for 1,445 trait- and disease-associated variants (Hindorff et al. 2009) in three ways. In the first method ("Any association") we simply require that the (Bonferroni corrected) p-value observed for the trait-associated variant pass the global eQTL threshold (FDR 0.05). In the second method we look for nominal p-values below 1e-7, corresponding to the threshold used to identify trait associations. In the third method, we require the SNP both pass the global eQTL threshold, and be within two orders of magnitude of the best QTL SNP for the corresponding gene. The final column (any QTL) indicates the union among all QTL types, which do have some overlap.

	***cis*-eQTL**	**sQTL**	***trans*-eQTL**	**any QTL**
Any association	790	218	9	818
Association p <= 1e-7	655	159	9	680
Association near best per gene	184	54	9	224

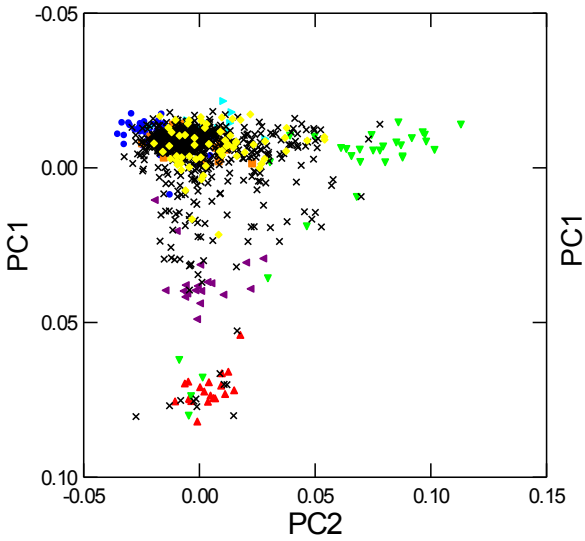
Supplementary Table S3. Individual traits with eQTL enrichment in GWAS hits. The table shows the p-values and q-values for enrichment of eQTLs in GWAS traits. The table only shows disease with an enrichment q-value < 0.2.

disease name	**p-value**	**q-value**
Plasma levels of liver enzymes	0.0005	0.0584
Ulcerative colitis	0.0003	0.0584
Primary biliary cirrhosis	0.0009	0.077
Hematological and biochemical traits	0.002	0.1117
Psoriasis	0.0027	0.1117
Triglycerides	0.0023	0.1117
Chronic kidney disease	0.0077	0.1755
Crohns disease	0.0057	0.1755
LDL cholesterol	0.0068	0.1755
Plasma levels of liver enzymes (gamma-glutamyl transferase)	0.0068	0.1755
Systolic blood pressure	0.0076	0.1755

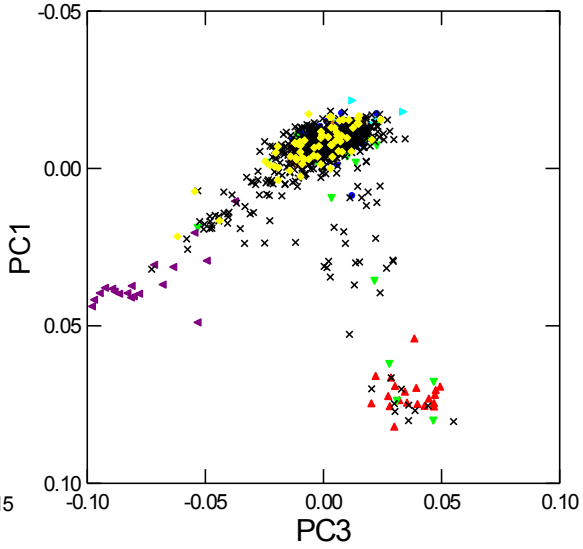
Supplementary Table S4. Genes and pathways correlated with demographic factors. The table provides the top twenty genes associated with age and sex in this cohort. We have also performed a functional enrichment analysis among the top 100, and 500, 1000 associated genes for age and sex separately using MSigDB (c2.cp.v3.1) annotations. In particular, we find two immune related pathways to be significantly enriched (0.05 FDR) in top 500 and 1000 sex-associated genes:
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM
REACTOME_INTERFERON_SIGNALING.
We also find four enriched pathways (0.05 FDR) among the top 1000 age-associated genes:
KEGG_ECM_RECEPTOR_INTERACTION
REACTOME_SIGNALING_BY_GPCR
REACTOME_GPCR_DOWNSTREAM_SIGNALING
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION.

Sex		Age	
Gene	**Q-value**	**Gene**	**Q-value**
OPLAH	1.20E-55	CD248	2.20E-71
C1orf93	5.90E-42	NRCAM	7.90E-49
MMEL1	3.70E-33	FBLN2	4.20E-45
CD274	7.70E-30	REG4	8.40E-45
GPR109B	5.30E-26	ROBO1	6.90E-43
GPR109A	1.20E-24	CACHD1	3.50E-38
NOS3	1.80E-23	ACCN2	5.50E-38
ADAMTSL2	3.00E-23	TSHZ2	1.60E-37
ULK1	6.40E-22	WNT10A	1.60E-37
DDX43	4.50E-21	PHLDA3	8.80E-37
GPR171	5.70E-21	SHANK1	2.50E-35
ADM	2.80E-20	FOXJ1	2.00E-33
C2orf55	1.10E-19	ZNF496	1.90E-32
NSD1	2.00E-19	SLC4A10	1.50E-30
MAN2A2	2.50E-19	AP3M2	3.40E-29
MPO	4.00E-19	ISM1	1.10E-28
GALNT3	4.00E-19	PTK7	1.10E-28
CHST10	4.80E-19	TTC24	2.60E-28
PER3	5.80E-19	CRTAM	1.00E-27
PDCD1LG2	6.70E-19	DDB2	1.50E-27

(a)



(b)



(c)

ANCGROUP	Frequency	Cumulative Frequency	Percent	Cumulative Percent
Anglo-Saxon	104	104	11.05	11.05
Ashkenazi	20	124	2.13	13.18
Eastern Eur	31	155	3.29	16.47
Mediterranean	19	174	2.02	18.49
Northern Eur	14	188	1.49	19.98
Native Amer	28	216	2.98	22.95
None	662	878	70.35	93.3
Western Eur	63	941	6.7	100

Figure S1. **Ancestry and Principal Components of genotype data.** The plot shows Principal Component (PC) 1 and 2 scores for 941 individuals with genotype data, of which 279 reported that 3 or 4 of their grandparents were of the same ethnic background, as shown in the table above; the predominant ancestry of these individuals is indicated in the legend, while the other 662 are labeled "None" (no known predominant ancestry). (a) PC1 reflects a North (here, more negative) to South gradient with Anglo-Saxons and Northern Europeans (Scandinavians) at the North end and Ashkenazi Jews at the South end, with Mediterranean (Italians, Greeks) in between. PC2 reflects West to East (non-Jewish Slavic/Russian). Note that, consistent with our previous observations in similar samples, individuals with self-reported predominantly Native American ancestry had PC scores in the main cluster of Western European ancestries, probably reflecting a reporting bias (i.e., over-estimation of the proportion of Native American ancestry in the family). (b) The plot shows that PC3 separated Ashkenazi from Mediterranean ancestry. PCs 1-3 were used to correct expression data for population structure. We note that the top genotype PCs had a very small impact on expression variability (Figure S5), and correcting for these PCs only resulted in losing 11 associations.

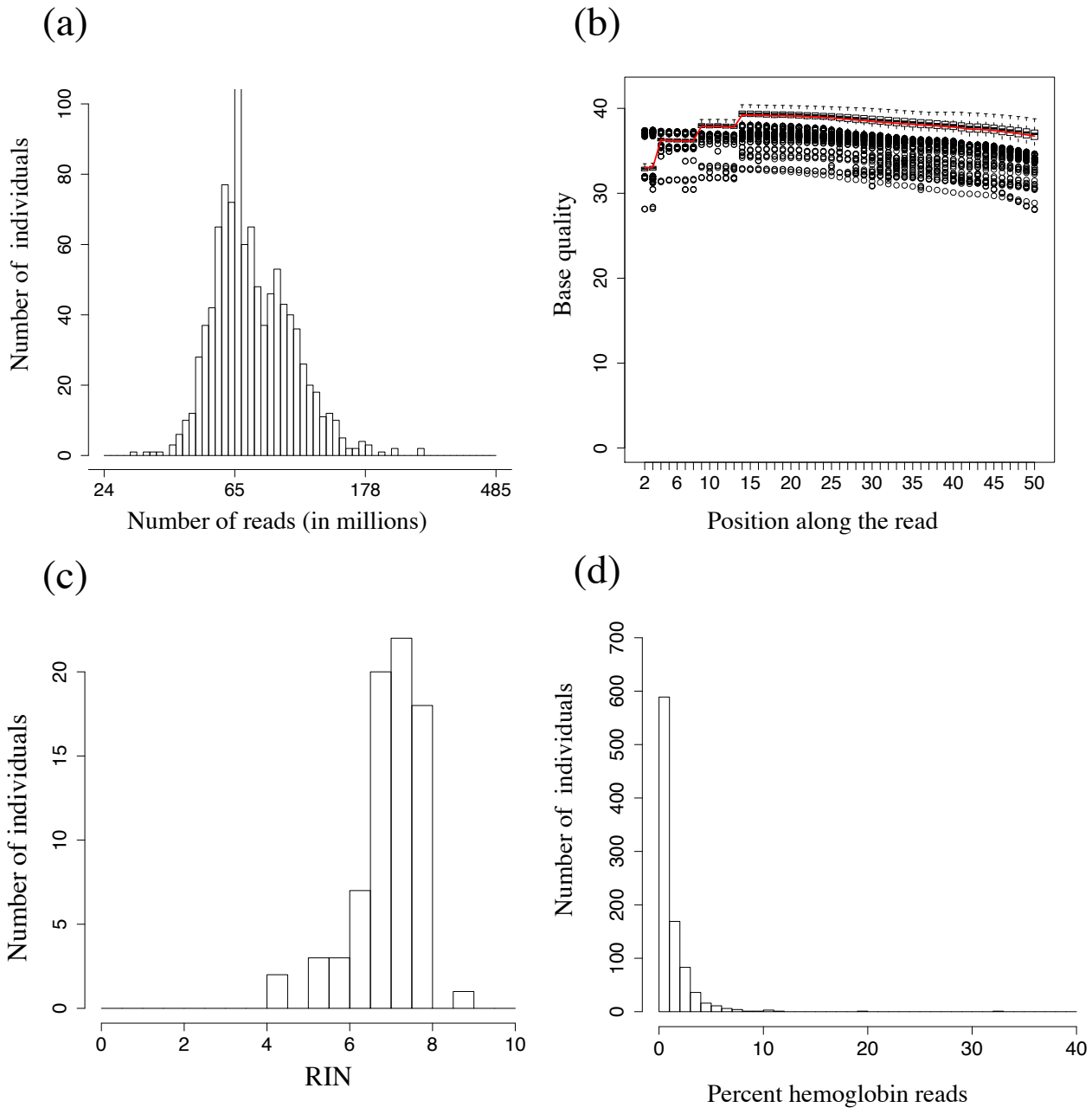


Figure S2. RNA-sequencing quality control. (a) The distribution of the number of sequenced reads is plotted in log scale. The distribution is skewed to the right because of the extra sequencing runs for poorly sequenced individuals. (b) Boxplot of base quality scores along the sequencing reads (from base 2 to base 50). Average score at each position is marked in red. The base quality reaches its maximum at base 14 and begins to decrease slowly after base 25. (c) RNA Integrity Numbers (RIN) for post-GLOBINclear™ (Invitrogen), RNA. We recorded the RINs for 12 samples from each 96 well plate containing RNAs. (d) Using the GLOBINclear™ (Invitrogen) protocol, hemoglobin RNA was removed from each sample before sequencing. A histogram of the percent reads coming from hemoglobin transcripts demonstrates the effectiveness of the GLOBINclear™ procedure amongst our individuals (median percent hemoglobin read is 0.7%) .

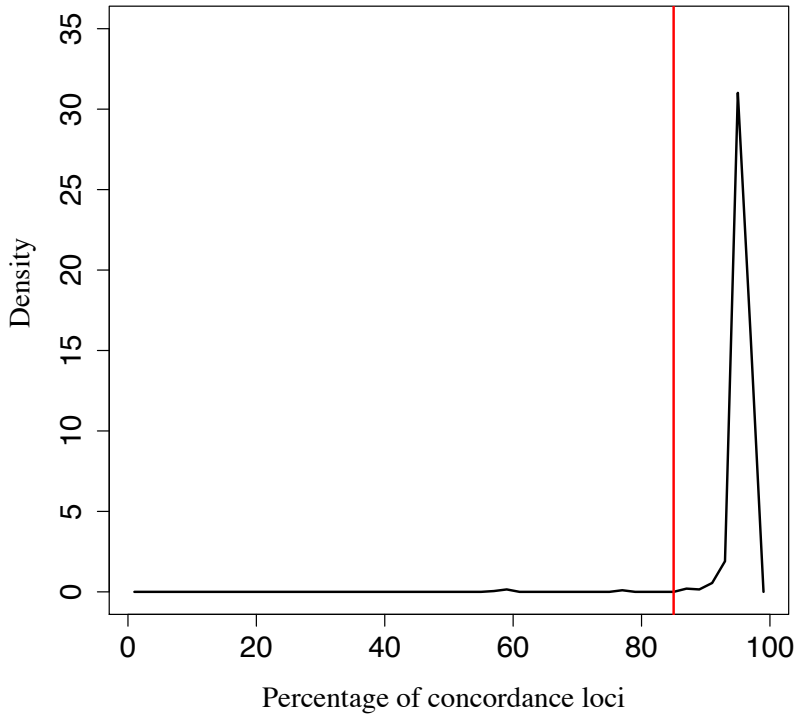


Figure S3. Concordance between SNP array and RNA-seq called genotypes. SNP genotypes were called using RNA-seq reads in deep covered regions and compared with the SNP array data. Low concordance (<85%, shown as a red line) suggests a potential labeling error, and such individuals were removed from this study. Most individuals show high estimates of concordance. We removed 6 subjects at the cutoff of 85% concordance.

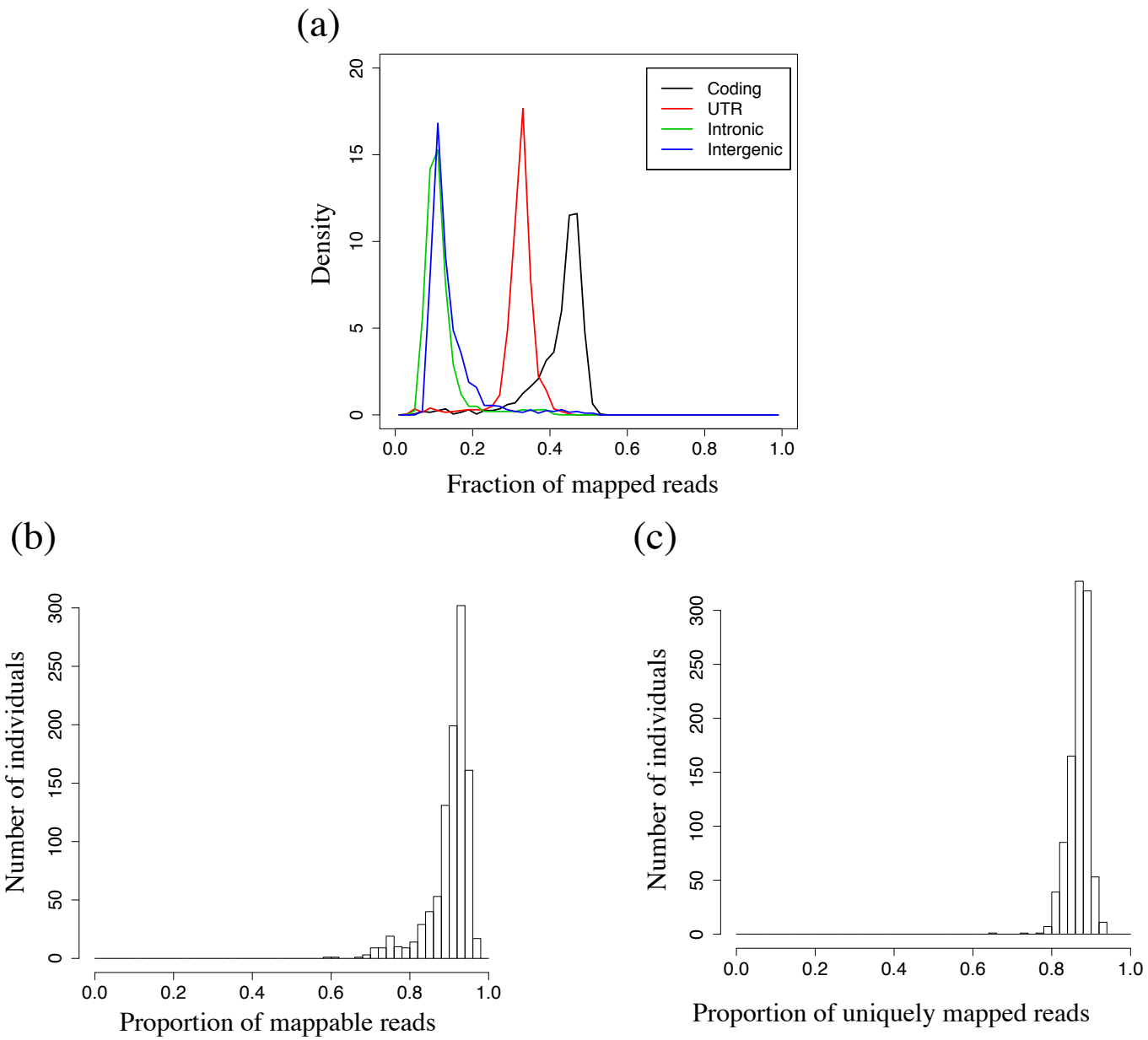


Figure S4. Mappability and distribution of mapped bases. (a) For each individual, we computed the fraction of mapped reads in coding regions (black), UTRs (red), introns (green) or intergenic regions (blue). This figure shows the distribution of fraction of mapped reads in each of these regions. As expected, majority of the mapped bases are within the coding regions or UTRs, while ~10% of the bases are within introns or intergenic regions. (b) Histogram of proportion of mappable reads in each individual. (c) Histogram of proportion of uniquely mapped reads (among the reads that were mapped) in each individual. As shown, in the majority of the individuals, at least 80% of the mapped reads were mapped uniquely.

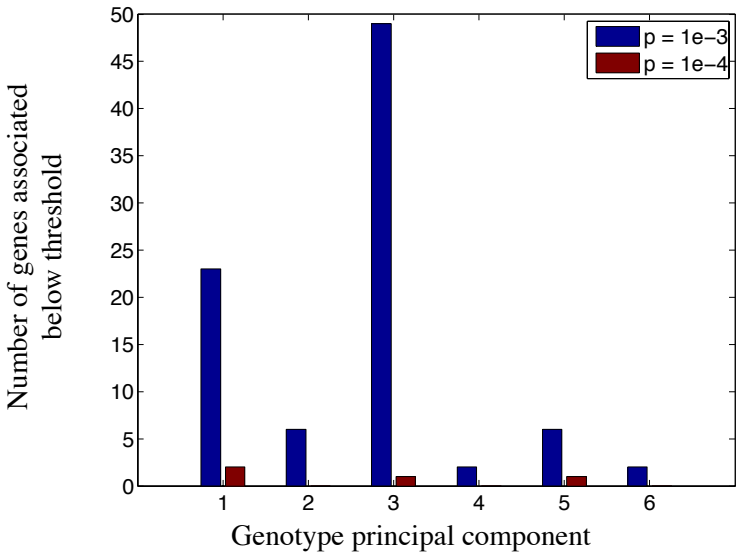


Figure S5. Correlation of genotype Principal Components and expression levels. Genotype principal components are not strongly associated with gene expression. In this figure, we show the number of genes correlated with each PC at two nominal significance thresholds, demonstrating very few genes with any correlation to one of these PCs. We also find that removal of population signal does not affect *cis*-eQTL discovery. We regress each gene against PCs 1-3, and remove the predicted component before eQTL discovery, but find that fewer than .2% of specific SNP to gene associations are changed, and overall we lose 11 genes from *cis*-eQTL results through this correction.

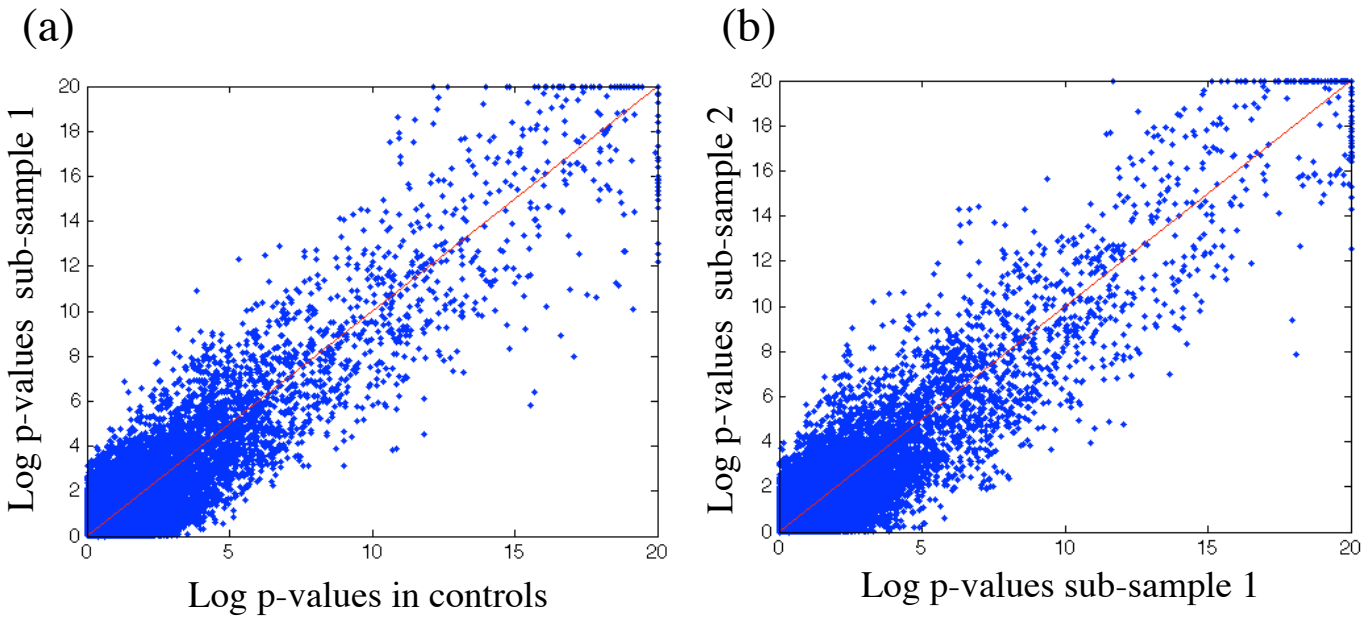


Figure S6. Evaluation of case control status on eQTL detection. To ensure that the cohort structure in this study does not introduce any bias in eQTL detection we computed *cis*-eQTL associations using the controls only (459 individuals), and compared it to *cis*-eQTLs detected using a random sub-samples of 459 individuals from the complete study (thus the sub-sample includes both cases and controls). Figure (a) shows the log p-values for all *cis*-SNPs on chromosome 18, computed using controls only (x-axis) or a sub-sample from the complete study that only includes 459 individuals. The variability in p-values shown in (a) is no more than the variability in two randomly chosen sub-samples of size 459 shown in (b).

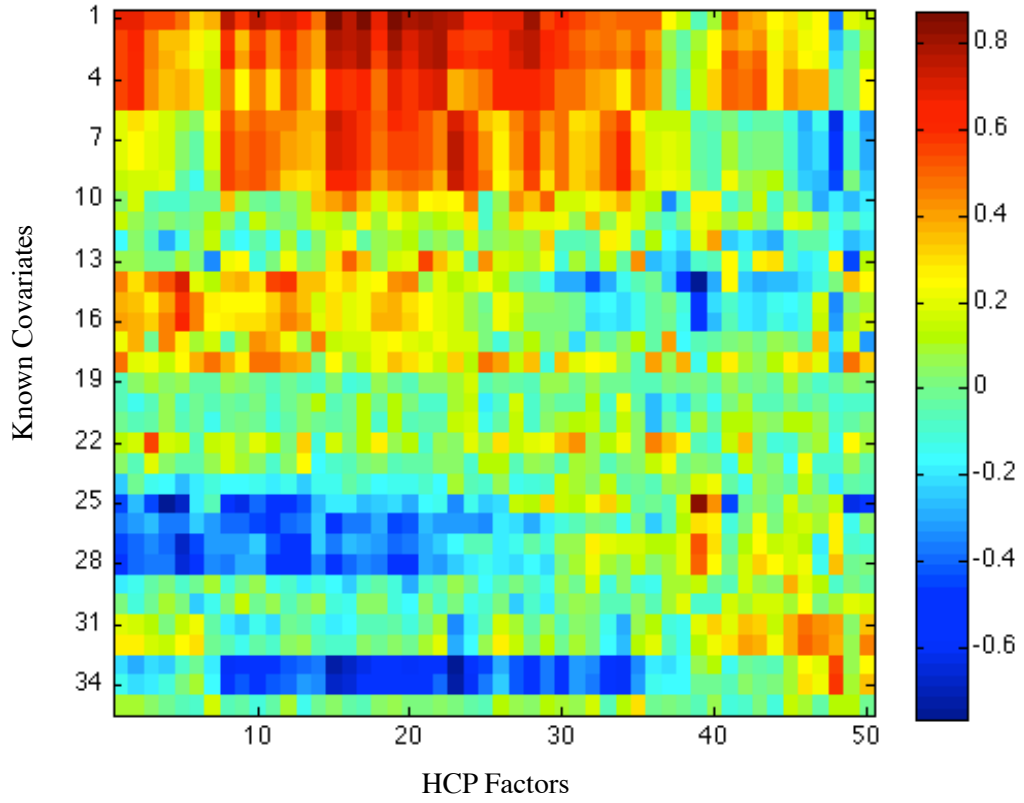


Figure S7. Correlation of HCP factors with known covariates. Figure shows the correlation coefficient (Spearman) between 35 known factors (x-axis) and 50 inferred HCP factors (y-axis). As shown, HCP factors summarizes multiple correlated known covariates, with technical covariates sequencing depth and other covariates that correspond to proportion of mapped reads with various annotations being particularly strong (Table S1 describes each of the 35 known covariates used to infer the HCP factors---ordering of the rows in this figure correspond to the covariate numbers in Table S1).

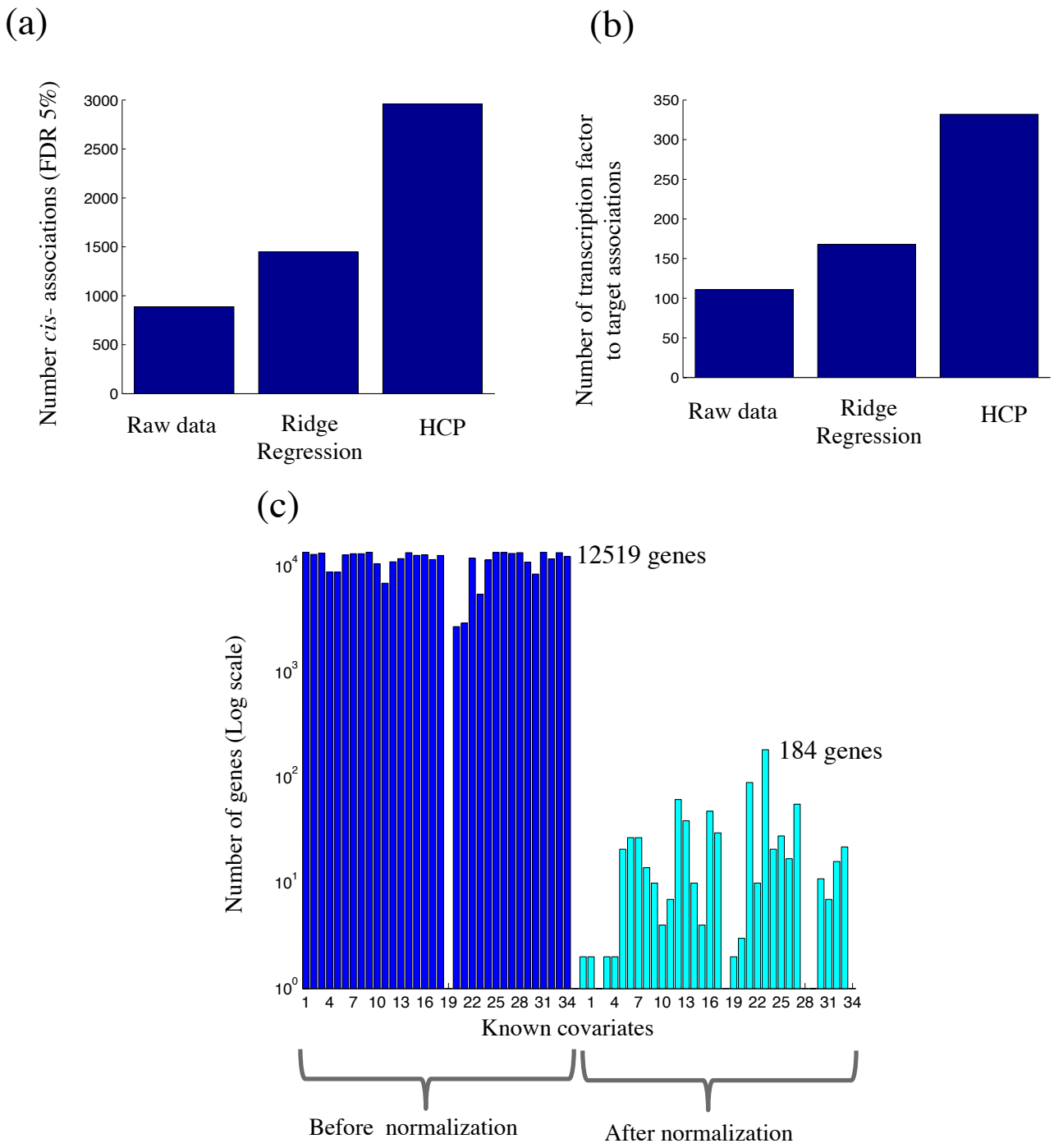


Figure S8. Removing the effects of confounding factors using HCP. (a) We optimized HCP's parameters for detecting *cis*-eQTL by using one chromosome for training (chromosome 18) and 1 chromosome for testing (chromosome 14). Figure shows the number of *cis*-eQTLs detected on *test* chromosome 14 using (i) raw data, (ii) ridge regression where we remove the effects of the 35 known covariates (see Table S1), and (iii) HCP. (b) We optimized HCP's parameters for *trans*-eQTL detection by optimizing the number of transcription factors and targets with significant expression correlation (FDR 5% using permutation p-values). Figure shows the number of such co-expressions detected using (i) raw data, (ii) ridge regression, and (iii) HCP. (c) Figure shows number of genes that are significantly correlated with the 35 known covariates at FDR 5% on RPKM data and HCP-normalized data. The technical covariates are ordered as in Supplementary Table S1.

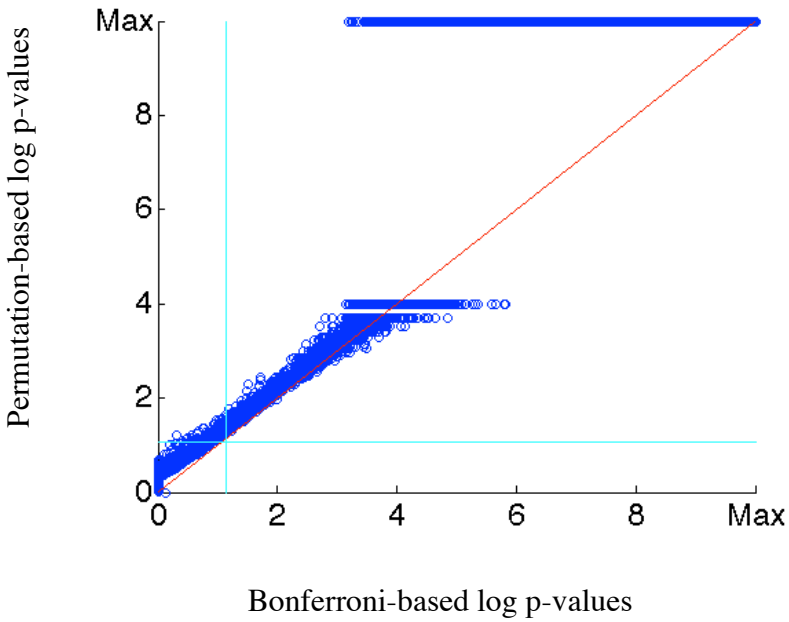


Figure S9. Bonferroni corrected versus permutation eQTL p-values. Comparison of *cis*-eQTL p-values (shown in log scale) obtained through Bonferroni correction per-gene (for the number of SNPs), and through permutation analysis using 10,000, for all genes. Each dot on the figure represents a gene. Bonferroni is somewhat conservative for large p-values (small log p-values), but more precise for very significant p-values since permutation p-value precision is limited by the number of permutations. No gene is called significant by Bonferroni which would not also be called significant according to permutation.

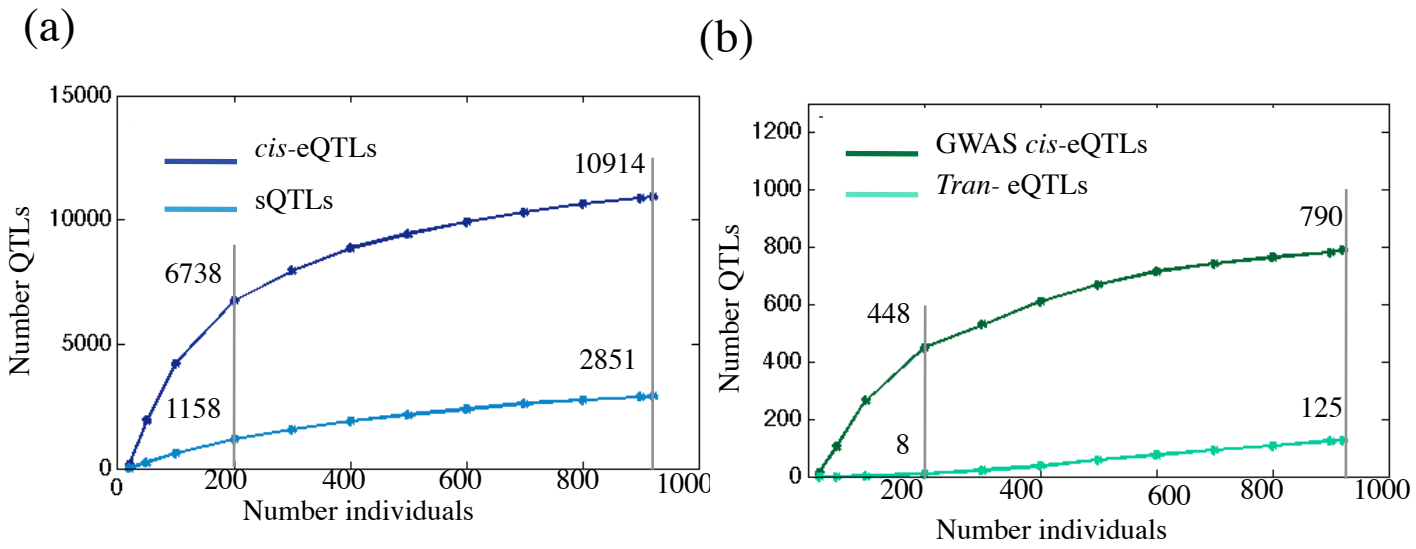


Figure S10. Sample size and number of detected QTLs. (a) Figure shows the number of *cis*-eQTLs and sQTLs detected with increasing sample size (number of individuals). For each sample size, we randomly selected the appropriate number of individuals from our study. As shown in this figure, as sample sizes increase above 500 individuals we begin to see some plateau effect in the number of discoveries. (b) Figure shows the number of trait and disease *cis*-eQTLs, where we define trait and disease SNPs as those reported in the CPGWAS (Hindorff et al. 2009), and *trans*-eQTLs detected with increasing sample size. With the size of the current study, we see a plateau effect in the number of detected GWAS *cis*-eQTLs. On the other, this figure shows that we haven't yet reached the plateau in detecting *trans*-eQTLs, and the number of *trans*-eQTLs detected will likely increase as sample sizes increase above 1000 individuals.

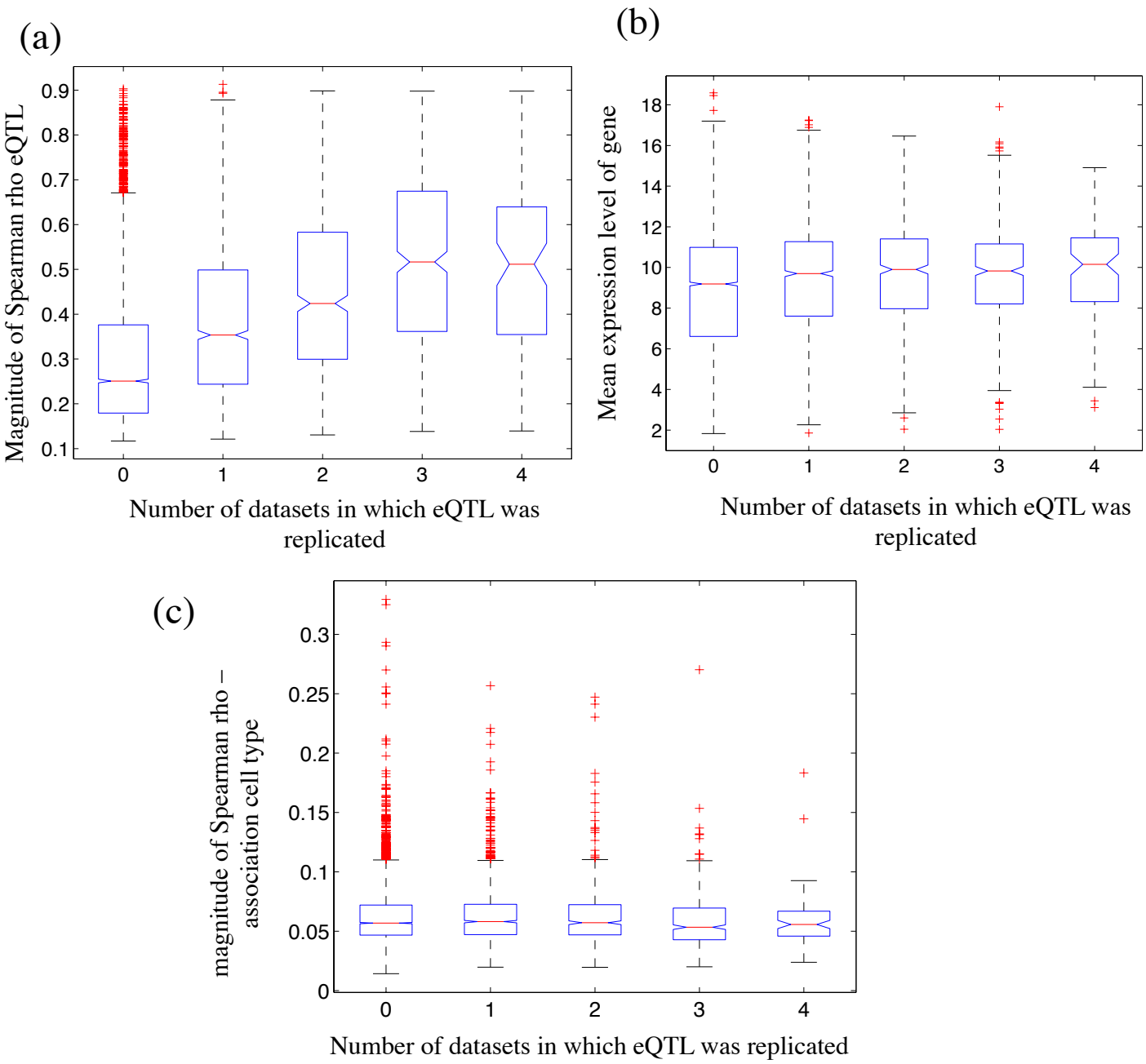


Figure S11. Replication of eQTLs in independent cohorts. We investigated the relationship between replication rate and (a) the strength of eQTLs (magnitude of Spearman rho), (b) mean expression levels, and (c) impact of cell type proportions (see Supplementary Material). For each associated gene, we only consider the best SNP per gene. We investigated replication rates in four datasets: MuTHER (Grundberg et al. 2012)) data which consists of three tissues (adipose, LCL, skin), and (Fehrmann et al. 2011) data (whole-blood). As shown, (a) more strongly associated eQTLs in this study, are significantly more likely to be replicated ($p < 1e\text{-}100$) , and (b) genes with higher expression levels are more likely to be replication ($p < 1e\text{-}20$), whereas (c) residual correlation between genes and cell type proportions do not impact the replication rate ($p > 0.1$).

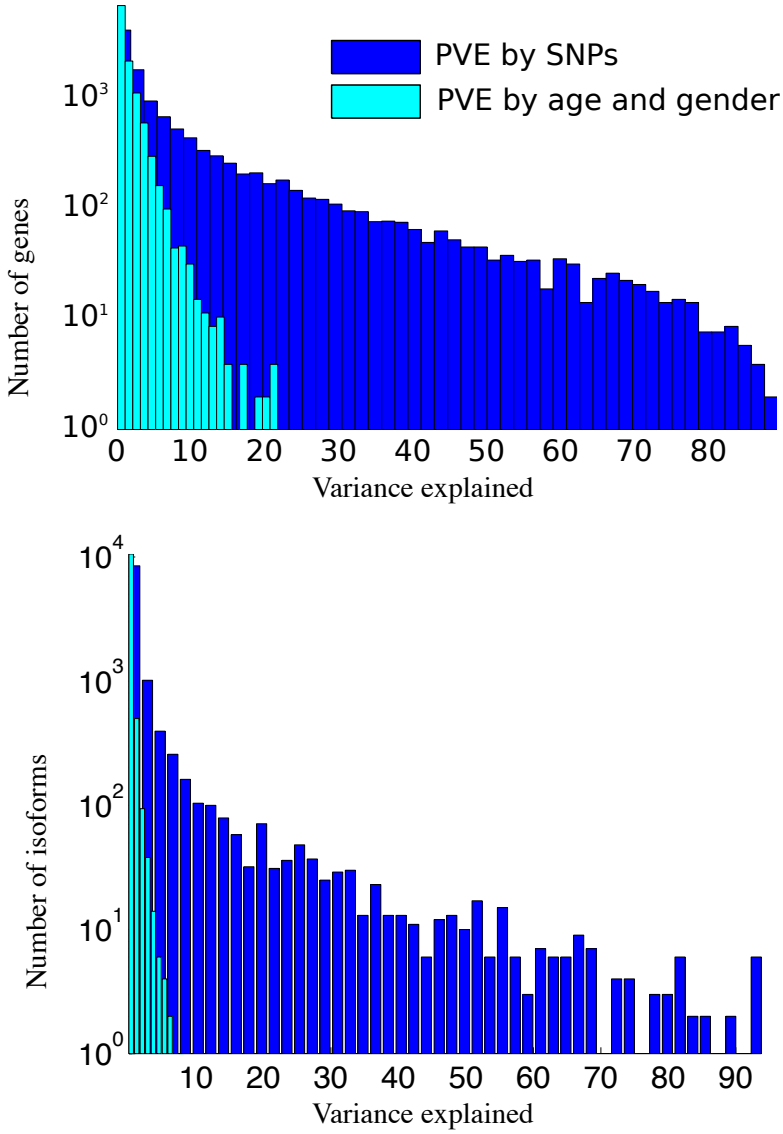


Figure S12. Variance in expression explained by genotype and demographic factors. Histograms of of percent total expression variance explained by genotype (dark blue) and both age and sex (light blue). We use stepwise linear regression to find all independently associated *cis*-SNPs (Bonferroni corrected threshold of 0.05), using expression data normalized only for known technical covariates rather than our full HCP procedure (Supplementary Materials). The proportion of variance explained for genes whose expression is not significantly associated with any of the corresponding factors is set to zero. (a) Figure shows variance explained for total gene expression. There are 9,263 genes whose expression has a significant genotype predictor in this analysis and 8,704 genes whose expression has a significant demographic predictor. (b) Figure shows the proportion of variance in isoform ratios. There was a significant genotypic predictor of isoform ratios for 3,168 isoforms, but there was a significant phenotypic (age or sex) predictor for only 1,443.

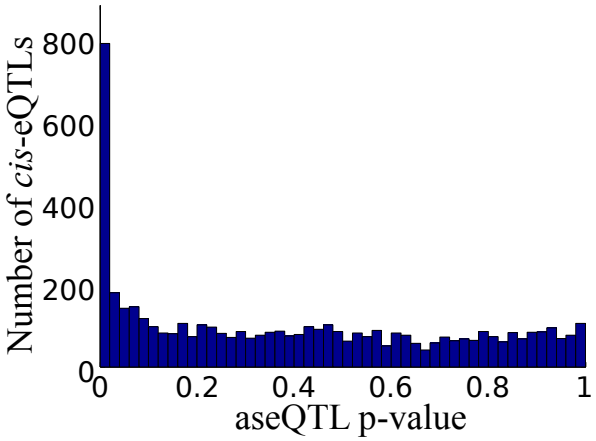


Figure S13. *Cis*-eQTL allele-specific expression effects. Distribution of aseQTL association statistics for SNP-gene pairs identified as *cis*-eQTLs. Only the top SNP was considered for each *cis*-eQTL, and we test for association with allelic imbalance at all candidate exonic loci in the gene, correcting for the number of tests using permutation analysis. We observe strong enrichment for low aseQTL p-values among *cis*-eQTL SNPs.

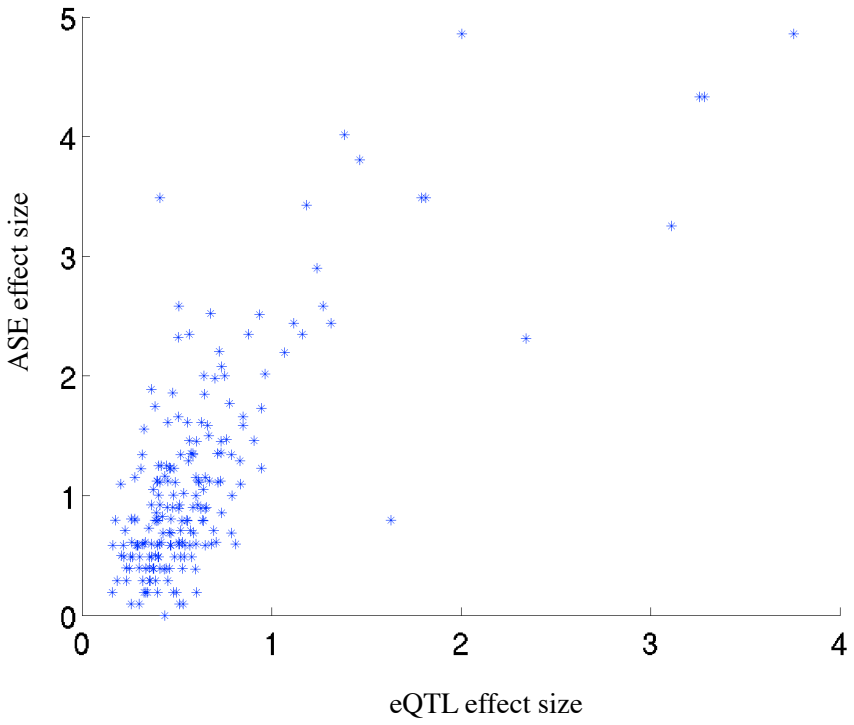


Figure S14. eQTL and ASE effect size. Figure shows a scatter plot of effect sizes computed for eQTLs and effect size computed for ASE for the corresponding SNP-gene. We used a sampling approach to compute effect sizes (see Supplementary Materials). The two estimates of variant impact are highly correlated (Spearman $p < 1e\text{-}21$).

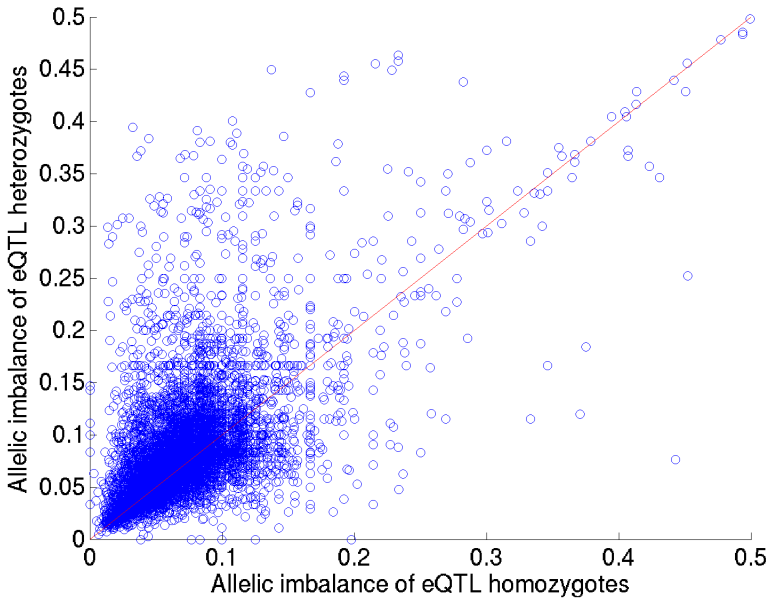


Figure S15. *cis*-eQTLs and allelic imbalance. Across all *cis*-eQTLs, allelic imbalance within the eQTL gene is much greater among eQTL SNP heterozygotes than homozygotes, evaluated only within heterozygous individuals at a coding locus within the gene.

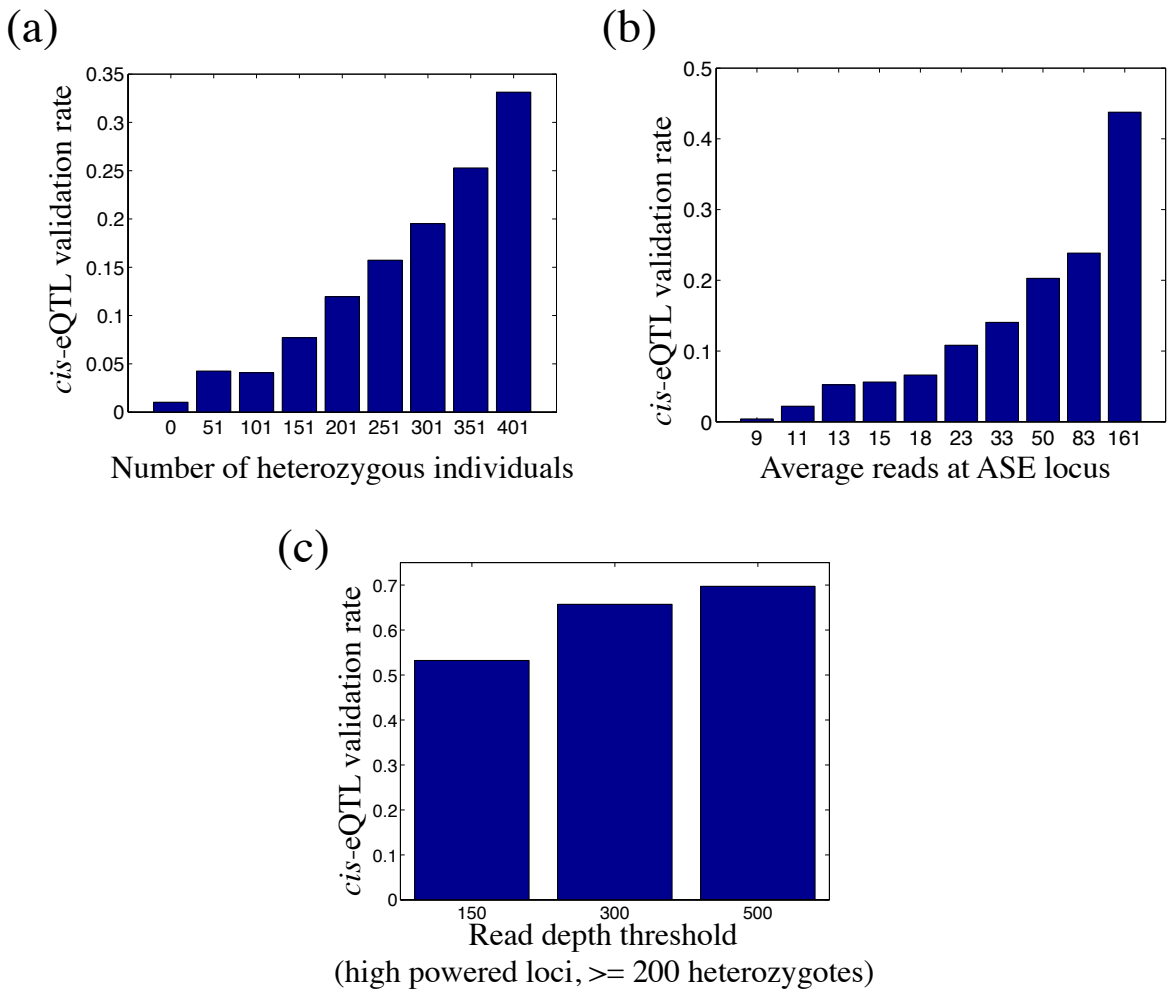


Figure S16. Factors affecting validation of *cis*-eQTLs through ASE. Many factors affect the validation of *cis*-eQTLs through aseQTL analysis. Of the 10,914 *cis*-eQTLs reported, 4,983 met a threshold of at least 10 compound heterozygous individuals (heterozygous at both the eQTL SNP and at some exonic locus within the gene), and were thus tested for ASE effects. Unsurprisingly, the strength of the *cis* association is the primary contributor ($p < 1e\text{-}148$) to validation through ASE. We do not observe a trend with respect to expression correlation with cell type markers. Other than *cis* effect, we identify two primary trends shown here, both related to statistical power. (a) Figure shows the rate of *cis*-eQTL validation broken down by buckets representing the number of heterozygous individuals available for aseQTL testing (the lower bound for the bucket is shown on the x-axis), with a strong relationship observed ($p < 1e\text{-}78$). (b) Figure shows rate of validation broken down by average read depth over the ASE locus, demonstrating that loci with greater read depth (and thus more accurate estimates of allelic expression) have higher validation rates ($p < 1e\text{-}92$). (c) Figure shows validation rate among the most highly powered eQTLs. Here, we restrict to SNPs with at least 200 heterozygous individuals, for different read depth thresholds. Overall, this analysis suggests that the validation rate of *cis*-eQTLs, given sufficient power, is between 50-70% (with no threshold on minimum effect size).

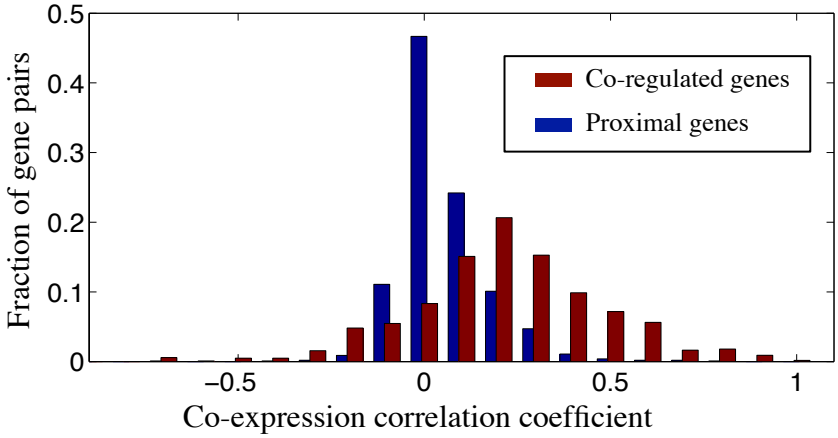


Figure S17. Co-expression of genes with a shared regulatory variant. Gene pairs with an identified shared *cis*-eQTLs are co-expressed to a higher degree than random gene pairs drawn from the same proximity distribution as the co-regulated genes ($p < 1e\text{-}87$). We also observe that the eQTL direction of effect agrees in 961 out of 1225 co-regulated pairs, and that 44.15% of the variance explained by the SNP is shared between co-regulated genes (estimated through partial correlation, likely an underestimate due to other factors affecting expression level of each gene).

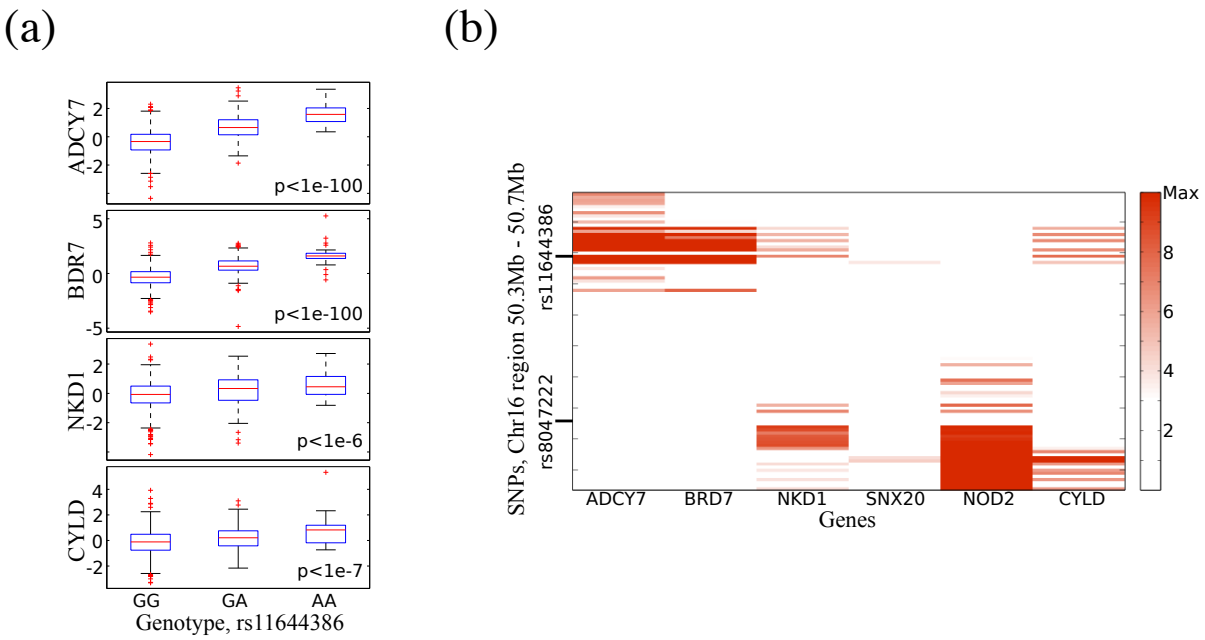
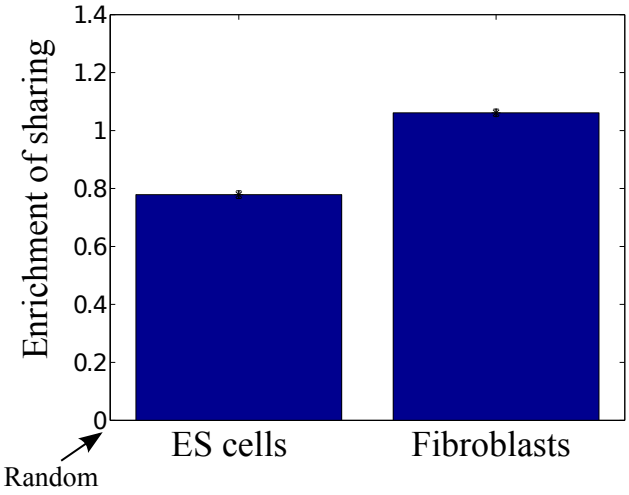


Figure S18. Example eQTL module, detail. a) The expression of four genes, *ADCY7*, *BRD7*, *NDK1*, and *CYLD* are all significantly associated with the SNP *rs11644386*. Box plots show expression levels of each gene grouped by genotype of *rs11644386*, annotated with significance of each association. b) Heatmap showing association results for each SNP in this genomic region with the genes depicted. This heatmap, with each colored rectangle depicting strength of association between one gene (on the x-axis) and one SNP (y-axis), is based on raw association statistics rather than the stepwise procedure, to illustrate the clear pattern of association sharing among these genes. The regulatory mechanisms that affect *NKD1* and *NOD2*, evident from the association with *rs8047222*, appear to be distinct from the mechanisms that regulate *ADCY7*, *BRD7*, *NKD1*, and *CYLD*. We note that the intermediate gene *SNX20*, which is not significantly associated with either SNP, is transcribed in the opposite direction from *NKD1* and *NOD2*.

(a)



(b)

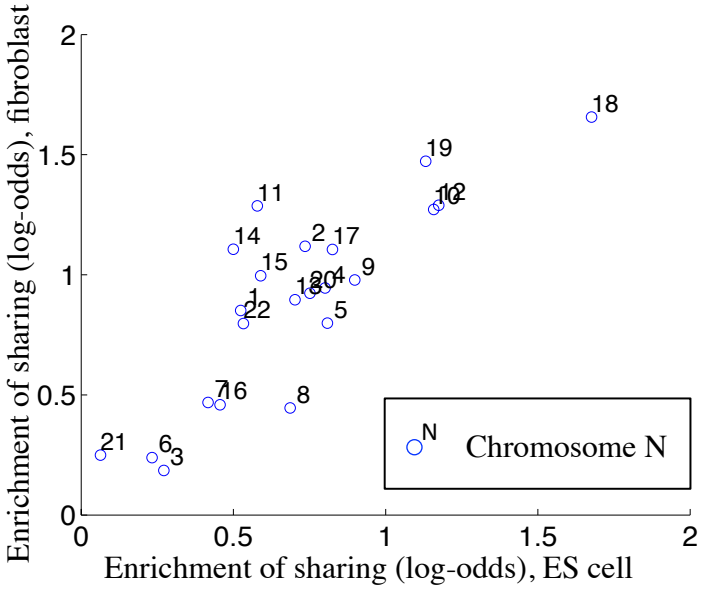
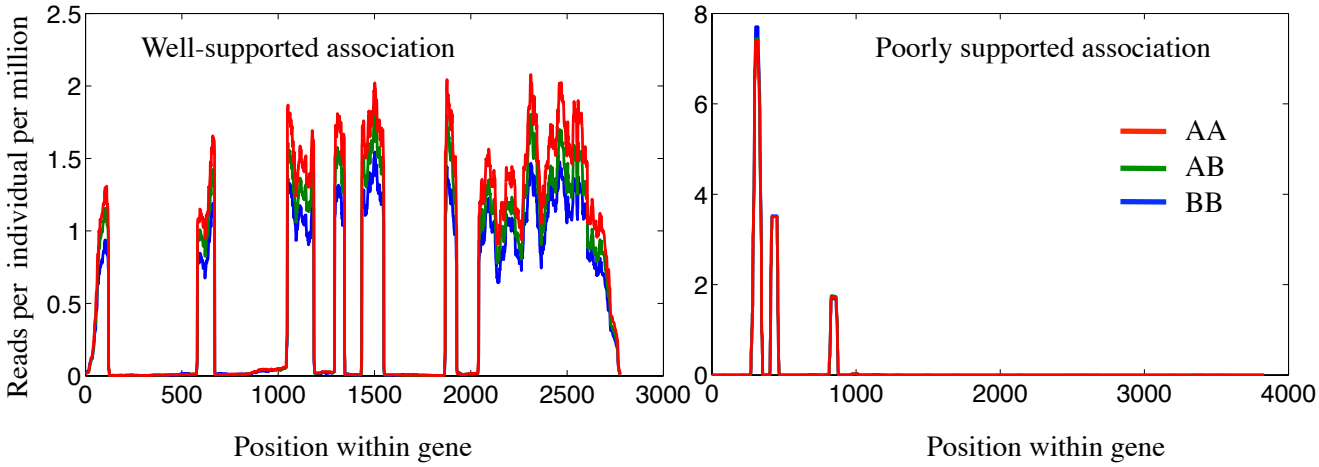


Figure S19. Co-regulation in topological domains. Enrichment of shared regulatory variants among genes within the same topological domain. Using all pairs of genes, we compute the log odds multiplier on a gene pair sharing some eQTL variant given that the TSS of both genes fall in the same topological domain as defined by regions of chromatin interaction, first controlling for linear proximity between the two TSSs (Supplementary Materials). As shown, topological domains are significantly predictive of shared eQTLs. a) Figure demonstrates enrichment according to topological domains derived Hi-C assays in two cell types: embryonic stem cells (ES cells) and adult fibroblasts. Both show significant enrichment. b) Enrichment of shared regulatory variants among genes within the same topological broken down by chromosome. The strength of enrichment varies significantly by chromosome, but the pattern between chromosomes is shown to be highly consistent between the two cell types topological domains were derived from.

(a) (b)



(c) Filtering *trans*-eQTLs

filtering step	# of genes	# of SNPs
Inter-chromosomal	175	2041
removed pseudo genes	163	953
removed cross-mapped regions	144	768
paralog filter	144	768
smoothness Filter	138	736

(d) Filtering *trans*-sQTLs

filtering step	# of isoforms	# unique genes	# of SNPs
Inter-chromosomal	9	6	1164
removed pseudo genes	7	4	10
removed cross-mapped regions	5	3	8
paralog filter	5	3	8
Smoothness filter	5	3	8

Figure S20. Filtering for accurate identification of *trans*-eQTLs. Figure (a) shows broad coverage of reads across exons of the gene and consistent association supported at each position, whereas in Figure (b) reads are only found covering a small number of positions along the gene, though at great depth. The association shown in Figure (b) is driven purely from few positions, and thus is filtered by our method. (c) and (d) Tables show the number of candidate *trans*-eQTLs and *trans*-sQTLs filtered at each stage.

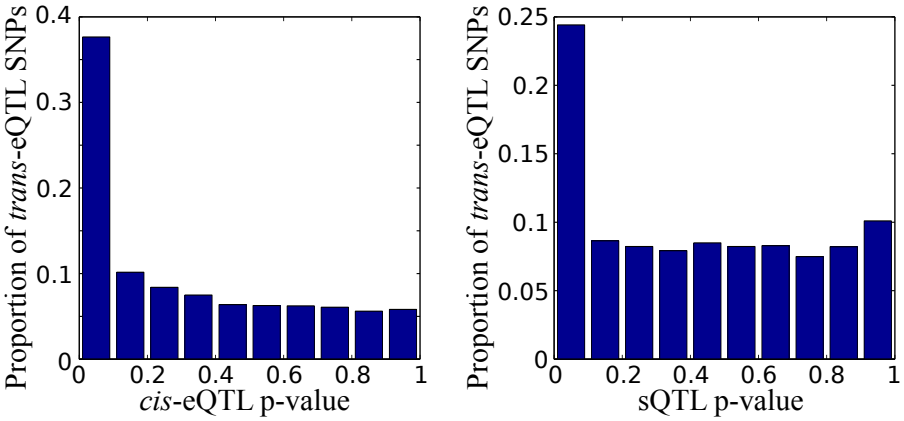


Figure S21. *Trans*-eQTL SNPs' effects on proximal genes. *Trans*-eQTL SNPs also have effects on expression and isoform ratio of proximal genes. The first histogram shows the distribution of *cis*-eQTL p-values for all *trans*-eQTL SNPs, considering all genes within 1Mb of each SNP. The second histogram shows the distribution of sQTL p-values, evaluated similarly.

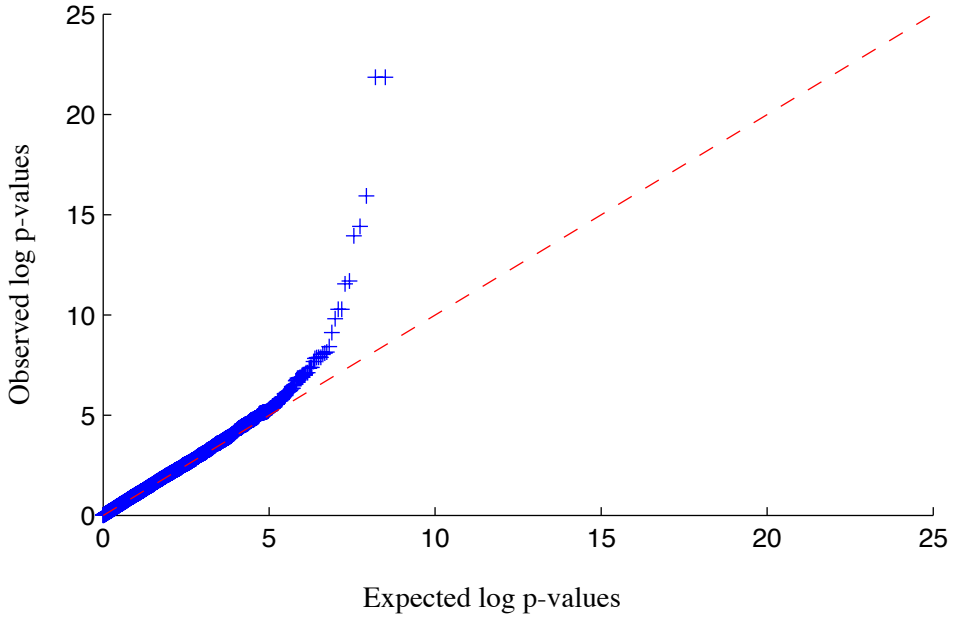


Figure S22. Enrichment for low p-values for association of *rs2759386* with distal isoform ratios. Figure shows the observed quantiles of log p-values of SNP *rs2759386* with all isoform ratios (y-axis) (a total of 12080 isoform ratio, corresponding to 4421 unique genes), compared to the expectation (x-axis) (each dot represents the log p-value for the association of this SNP with one isoform). *rs2759386* is a *cis*-eQTL for the splicing factor *QKI*, and it is more correlated with a large number of isoform ratios compared to the expectation.

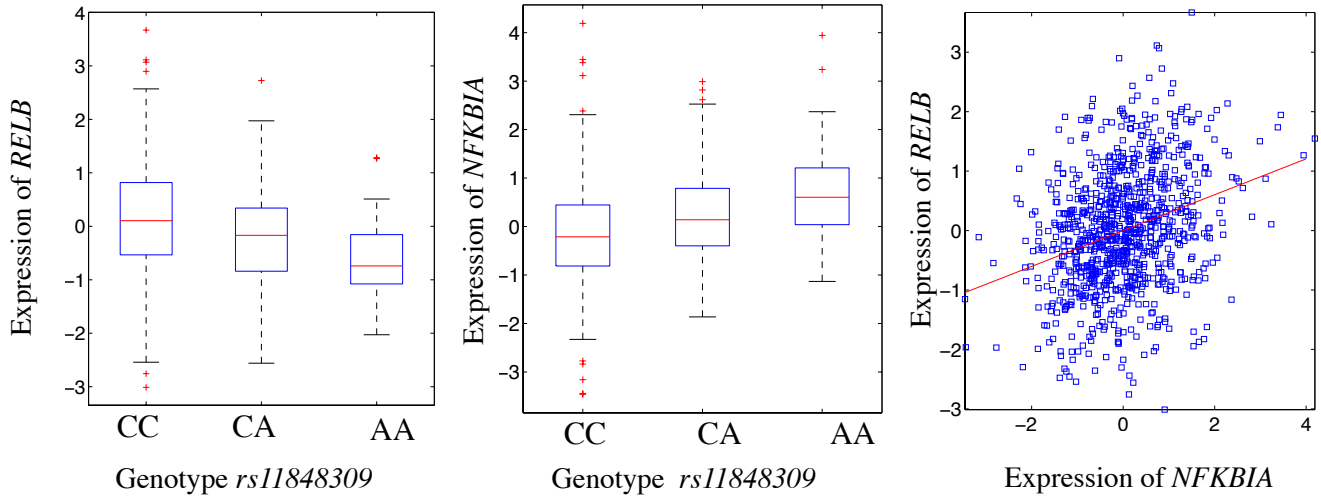


Figure S23. Example of a paradoxical *trans*-eQTL relationship. SNP *rs11848309* is a *trans*-eQTL for the gene *RELB*, and a *cis*-eQTL for the gene *NFKBIA*. The expression of *NFKBIA* and *RELB* are significantly correlated (p<1e-18), however the direction of the correlation is the opposite of that predicted based on the relationship between *rs11848309* and *RELB* (positive correlation), and the relationship between *rs11848309* and *NFKBIA* (negative correlation). Potential mechanisms could include *cis*-SNP effecting protein function of *NFKBIA* and autoregulatory feedback.

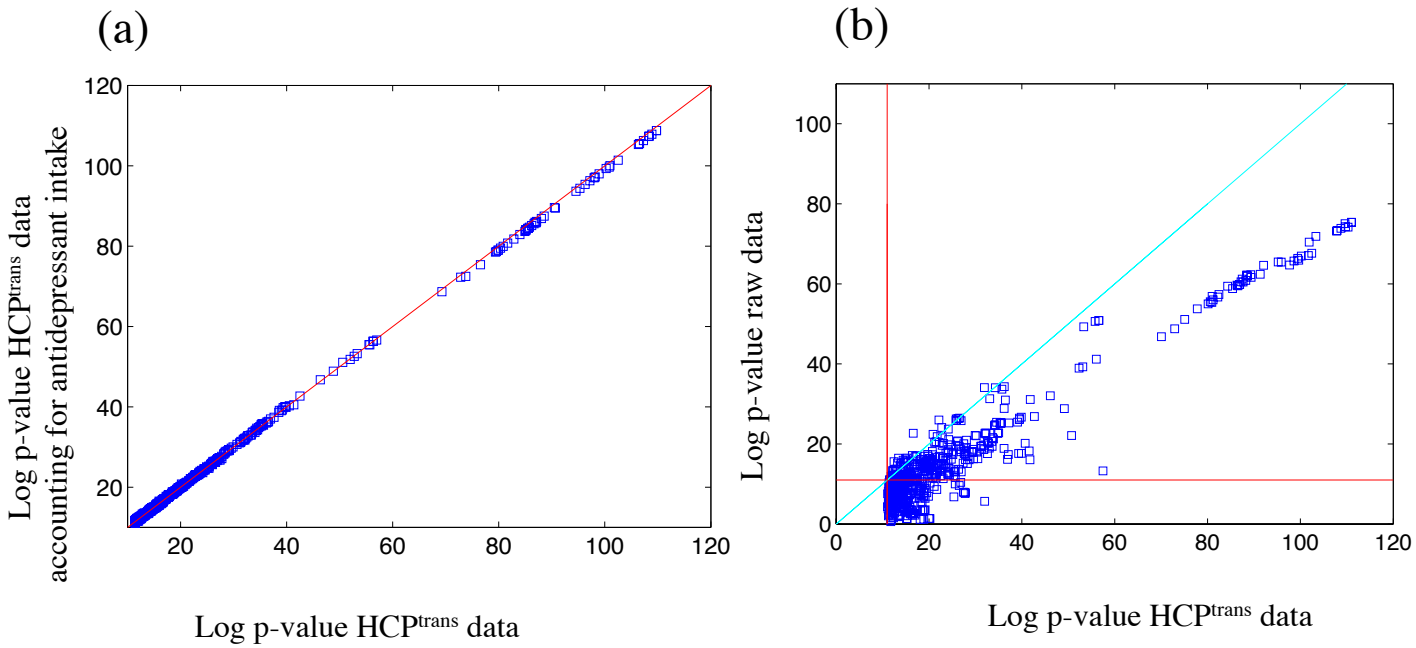
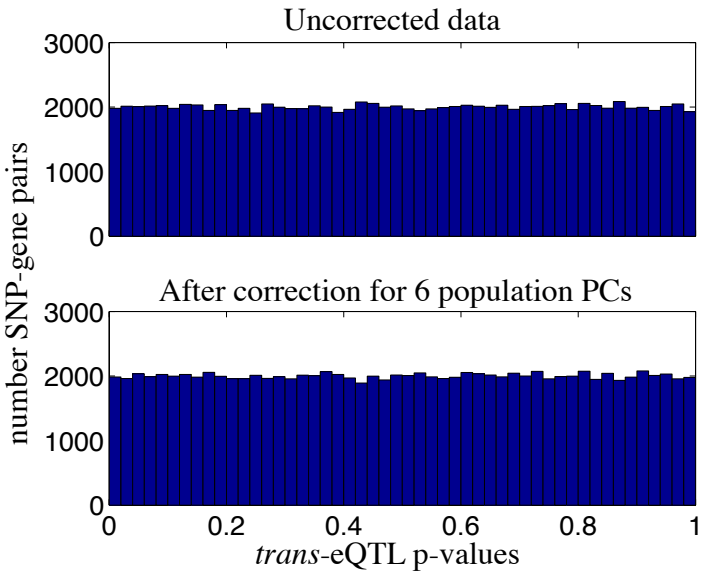


Figure S24. Potential confounding factors in *trans*-eQTL detection. **(a)** Figure shows the log p-values for discovered *trans*-eQTLs. X-axis shows p-values obtained from the HCP model, which corrects for cell-type proportions and technical covariates. Y-axis shows p-values obtained from HCP data that has been further adjusted (linear regression) for intake of antidepressant. As shown, the medication intake does not effect *trans*-eQTL p-values. (b) Figure shows the log p-values for discovered *trans*-eQTLs using quantile normalized data (y-axis) and HCP^{trans} data (x-axis). Quantile normalization was applied to raw data to account for variable sequencing depth. As shown in the figure, HCP normalized data identifies many more *trans*-eQTLs compared to the quantile normalized data (the red lines mark the Bonferroni threshold).

(a)



(b)

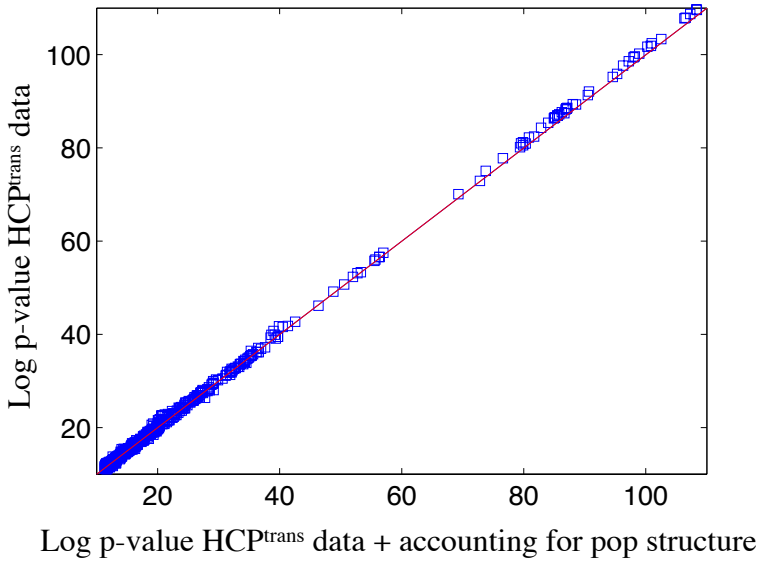


Figure S25. Relationship between population structure and *trans*-eQTL detection. (a) In the raw data (before any population correction), we observe no overall inflation of associations among cross-chromosomal SNP-gene pairs (top histogram), and after correcting for six population PCs, little to no change in distribution. This is consistent with Figure S5, which suggests that very few genes' expression levels are affected by population structure. Our final analysis is based on removal of 3 population PCs. (b) Figure shows the log p-values for our reported 138 *trans*-eQTLs before (y-axis) and after (x-axis) adjusting for population structure (regressing out three genotype PCs). As discussed in Supplementary Materials, we discovered *trans*-eQTLs on data adjusted for population structure, but this suggests that correction had very little effect. This figure shows that the population structure here does not result in a high inflation of *trans*-associations.

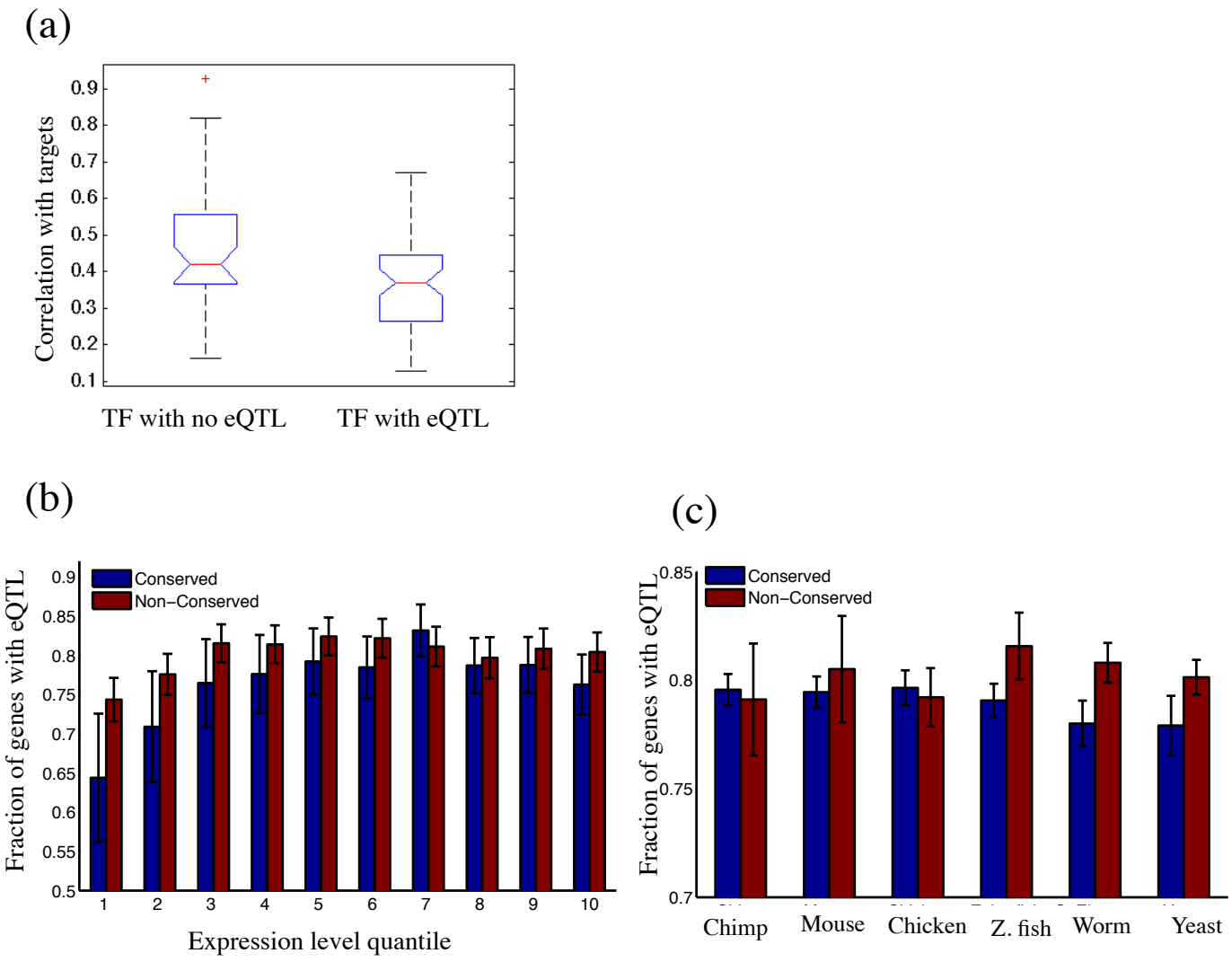


Figure S26. eQTLs and selection pressure. (a) transcription factors (TFs) with *cis*-eQTL are less correlated (by co-expression) to their known targets. Boxplot shows that TFs that have at least one *cis*-eQTL tend to have a lower expression correlation with their targets, compared to those TFs that have no *cis*-eQTLs. (b) Depletion of *cis*-eQTLs appears to hold among conserved genes (here in yeast) across all levels of gene expression, and thus is not explained simply by expression level of conserved genes. Although sub-dividing genes into expression-level buckets does result in statistical significance for every bucket, depletion is observed across expression levels. (c) The fraction of genes with an eQTL for non-conserved (red) genes is compared to the fraction among conserved genes (blue) for six species. While no significant result is observed for more recently diverged species (chimp, mouse, chicken), analysis for the three more distant species (zebrafish, c. elegans, yeast) all show a significant depletion of eQTLs among conserved genes. We show 95% confidence intervals for each estimate.

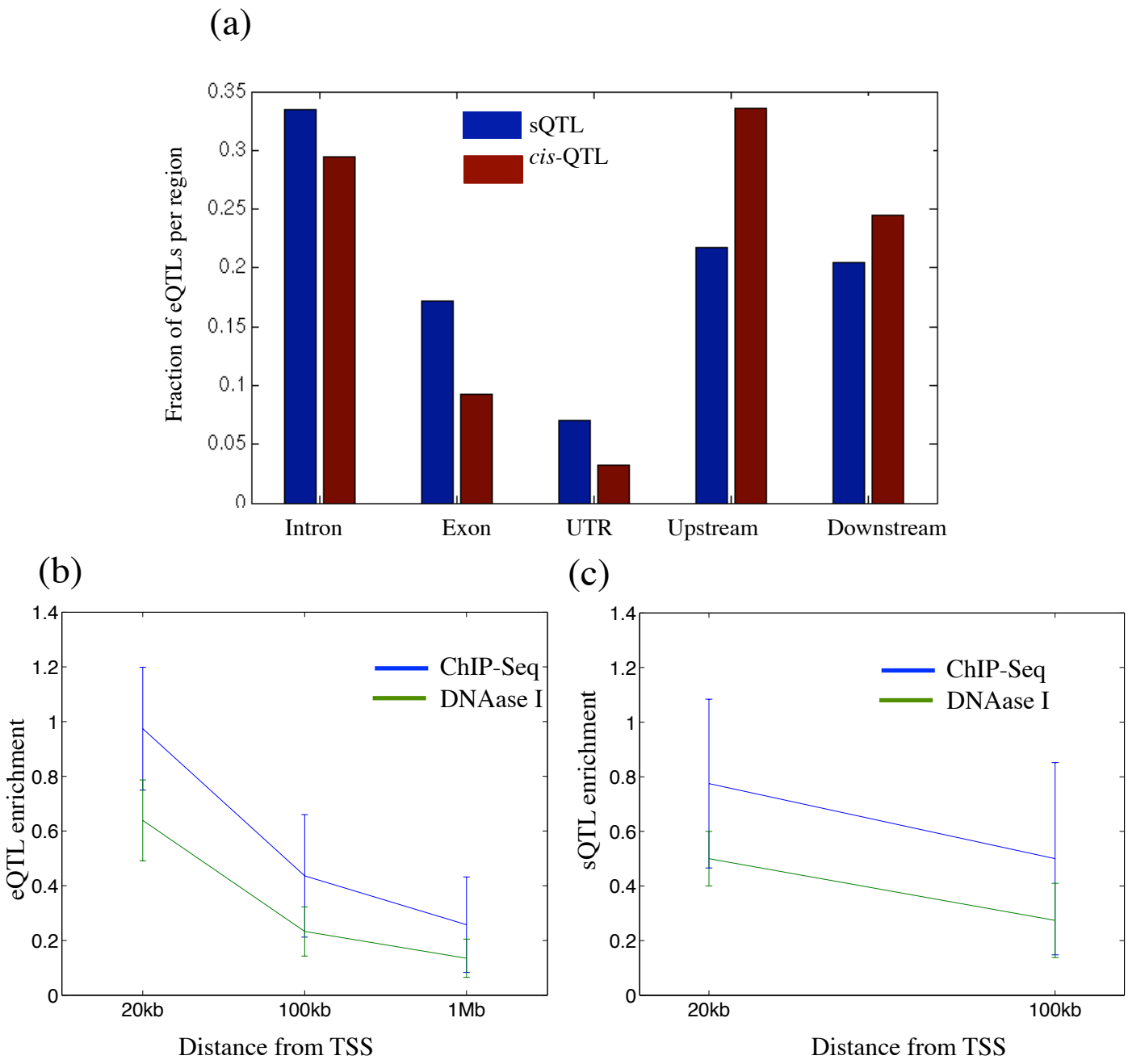
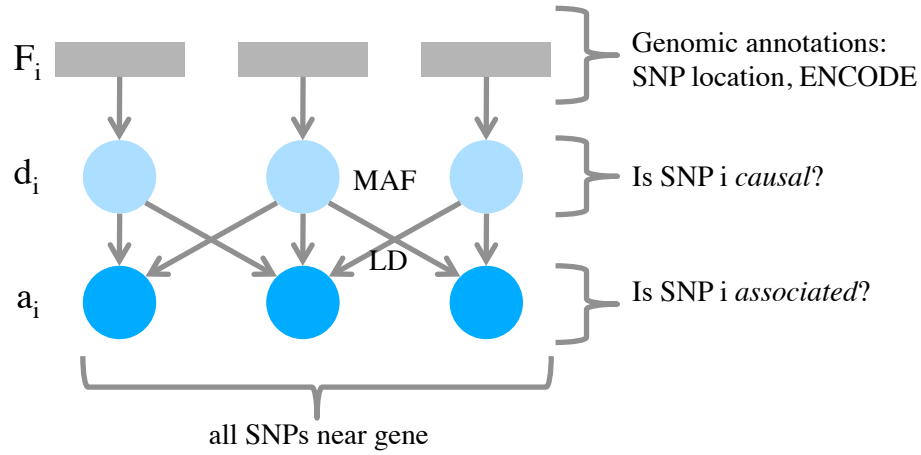


Figure S27. Genomic position of QTLs. (a) Comparison of *cis*-eQTLs (blue) and sQTLs (red) according to the location of the single strongest SNP for each gene, in order to highlight the differences between the two regulatory mechanisms. Each bar represents the fraction of these SNPs that fall into the labeled genomic region (thus, the bars for each QTL type sum to 1.0). We find that sQTLs are much more concentrated within gene boundaries compared to *cis*-eQTLs, which have a much stronger enrichment among upstream regions. (Using our logistic model that accounts for SNP position, we explored the enrichment of (b) eQTLs and (c) sQTLs in TF binding sites (ENOCDE ChIP-seq data) and open chromatin regions (ENCODE DNAase I hypersensitivity data), with increasing distance from TSS. As shown on these figures, we observed that with increasing distance from TSS the enrichment declines, indicating that farther ChIP-seq and DNAase sites are not as predictive of QTLs as those that are closer to the TSS. Figures shows results from a single ChIP-Seq and a single DNAase assay with the highest enrichment score. We only show results for up to 100kb for sQTLs as there is no enrichment beyond 100kb.

(a)



(b)

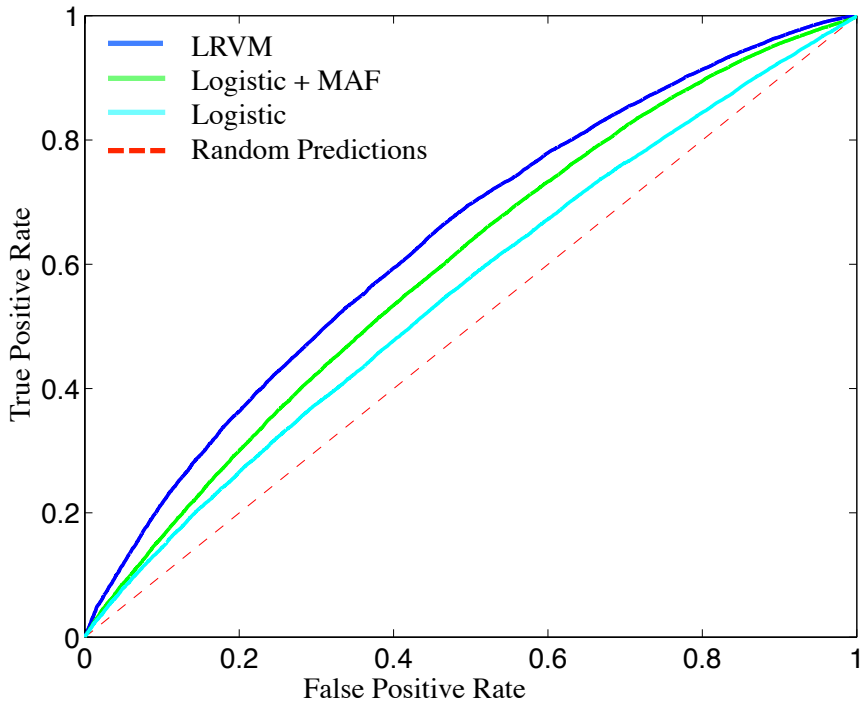


Figure S28. Latent Regulatory Variant Model (LRVM) and application to *cis*-eQTL prediction. (a) Schematic diagram of Bayesian network model of eQTL association. (b) ROC curve demonstrating the performance of LRVM on predicting eQTL associations. We compare 1) LRVM (AUC 0.6383) 2) logistic regression incorporating a multiplicative term for minor allele frequency (AUC 0.5971) and 3) plain logistic regression (AUC 0.5553).

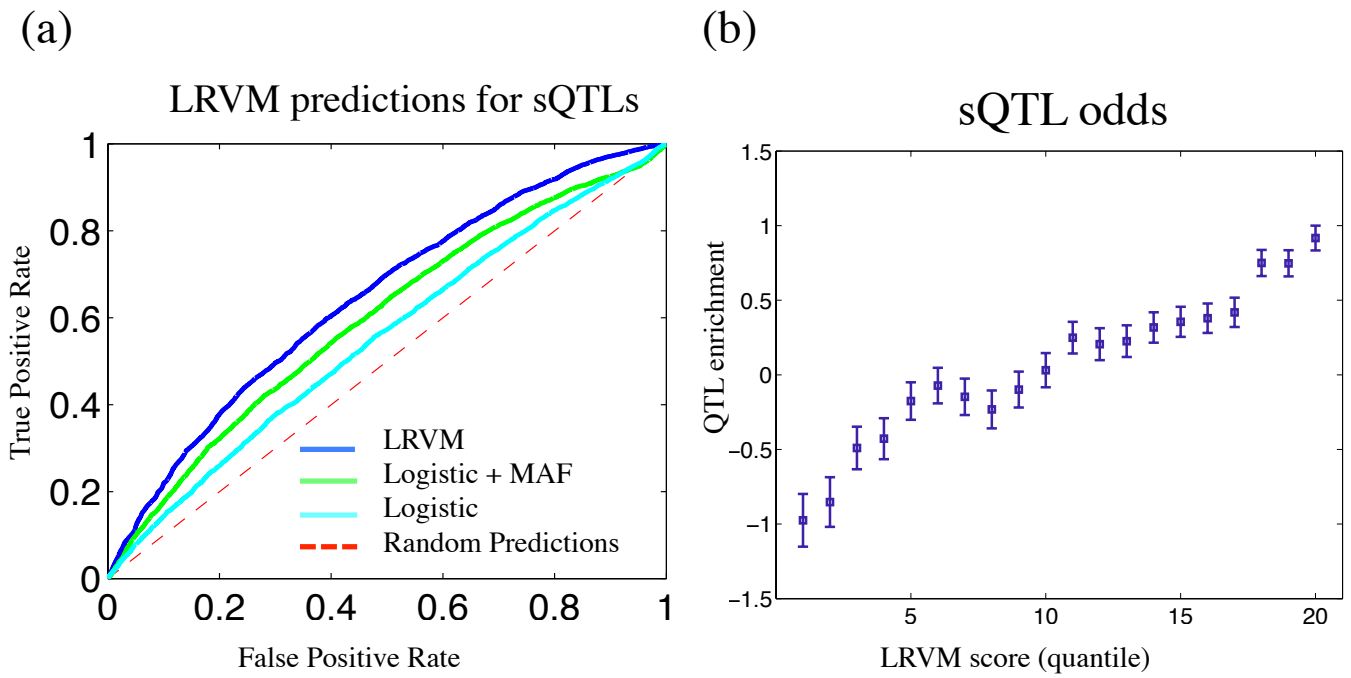


Figure S29. Application of LRVM to splicing QTLs. (a) ROC curve demonstrating the performance of LRVM on predicting sQTL associations. We compare 1) LRVM (AUC 0.6442) 2) logistic regression incorporating a multiplicative term for minor allele frequency (AUC 0.5965) and 3) plain logistic regression (AUC 0.5514). (b) LRVM scores are shown for candidate sQTLs (isoform/SNP pairs) which were not included in training LRVM. Each sQTL was scored by LRVM for predicted likelihood of association, and twenty quantiles were computed for the resulting scores. Then, enrichment of (observed) sQTL associations was computed for each quantile. Here, we use conditional odds to estimate enrichment, first correcting for SNP position in the same procedure used for genomic annotation enrichment analysis (Supplementary Materials) – although LRVM does incorporate position features, we highlight here the information captured by LRVM scores beyond SNP position.

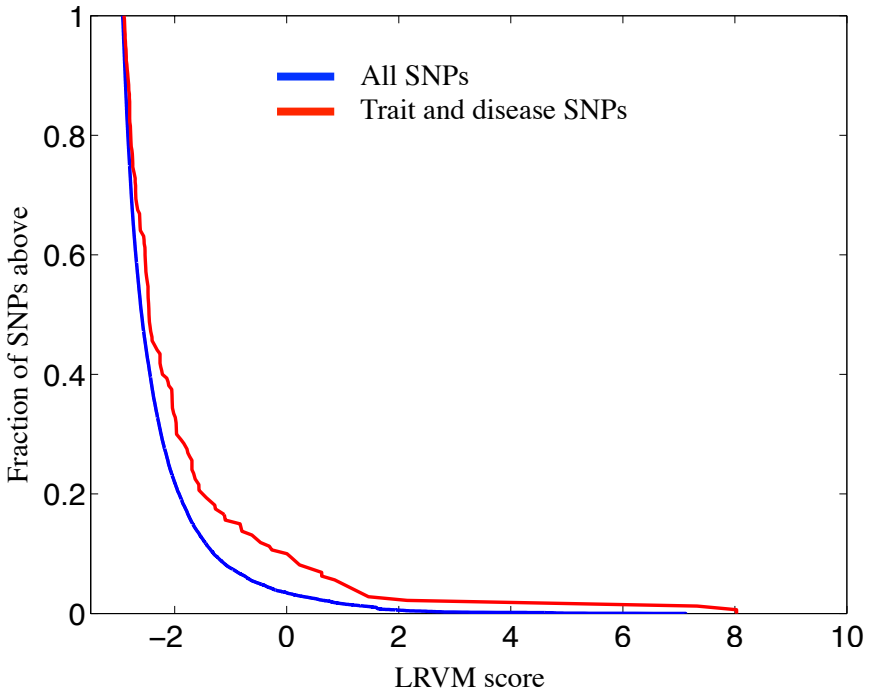


Figure S30. LRVM sQTL prediction for GWAS SNPs. LRVM assigns higher sQTL regulatory scores to trait and disease SNPs compared to background SNPs. Predicted splicing (sQTL) impact of traits and disease SNPs according to LRVM, for trait and disease variants not available during model training. We compute the score of each SNP for each of its proximal genes. Known trait- or disease-associated SNPs score more highly that expected at random ($p < 1e\text{-}7$).

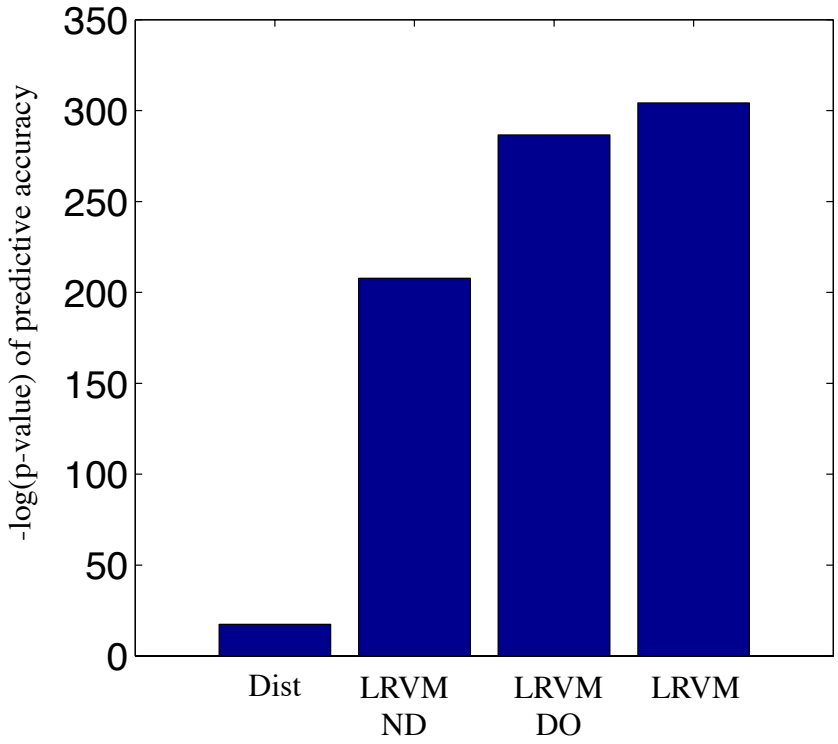


Figure S31. Effect of distance from TSS on LRVM. We explored the effects of SNP location on LRVM performance. Figure 5 demonstrates that SNP position (distance from TSS and within-gene location) are strong predictors of eQTL SNPs, and similarly we see that position is one of the strongest features used by LRVM. Here, we compare a purely distance-based prediction (left-most bar) and three versions of LRVM. In order from left to right: 1) Dist – a logistic model using only distance features without LRVM latent variable modeling, 2) LRVM-ND – LRVM with features only based on genomic annotations with no distance features, 3) LRVM-DO – LRVM using only distance features, and 4) the full LRVM model. Overall we find that the modeling refinements made by LRVM (accounting for MAF and LD) are the most significant contributors to its accuracy, with 'Dist' alone performing much worse than any of the LRVM models. Second, both distance and genomic annotations contribute to LRVM accuracy. Given significant correlation between SNP position and regulatory elements (TF binding sites are themselves enriched near the TSS), there is overlap in the information provided by these two signals, which cannot be causally disambiguated without interventional study, but we find ultimately both do contribute.

References:

Picard. http://picardsourceforgenet/.

Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. 2009. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* **4**(7): e6098.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**(9): 1790-1797.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): 376-380.

Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, Fu J, Deelen P, Groen HJ, Smolonska A et al. 2011. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* **7**(8): e1002197.

Gaffney DJ, Veyrieras JB, Degner JF, Pique-Regi R, Pai AA, Crawford GE, Stephens M, Gilad Y, Pritchard JK. 2012. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol* **13**(1): R7.

Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**(10): 1084-1089.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**(23): 9362-9367.

Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* **180**(4): 1909-1925.

Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. 2010. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**(19): 2438-2444.

Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D. 2009. Learning a prior on regulatory potential from eQTL data. *PLoS Genet* **5**(1): e1000358.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**(9): 1724-1735.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Listgarten J, Kadie C, Schadt EE, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A* **107**(38): 16465-16470.

Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, Koller D. 2013. Normalizing RNA-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One* **8**(7): e68141.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ et al. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**(7256): 753-757.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7**(3): 500-507.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**(16): 9440-9445.

Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M et al. 2012. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* **8**(4): e1002639.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.