Leveraging tissue specific gene expression regulation to identify genes associated with complex diseases

Wei Liu[1,#], Mo Li[2,#], Wenfeng Zhang[2], Geyu Zhou[1], Xing Wu[4], Jiawei Wang[1], Hongyu Zhao[1,2,3*]

[1] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA 06510

[2] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA 06510

[3] Department of Genetics, Yale School of Medicine, New Haven, CT, USA 06510

[4] Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA 06510

[#] These authors contributed equally to this work

* To Whom correspondence should be addressed:
Dr. Hongyu Zhao
Department of Biostatistics
Yale School of Public Health
60 College Street,
New Haven, CT, 06520, USA
hongyu.zhao@yale.edu

Key words: GWAS; gene expression imputation; Alzheimer's disease; gene-level association test

Abstract

To increase statistical power to identify genes associated with complex traits, a number of methods like PrediXcan and FUSION have been developed using gene expression as a mediating trait linking genetic variations and diseases. These methods first develop models for expression levels based on inferred expression quantitative trait loci (eQTLs) and then identify expression-mediated genetic effects on diseases by associating phenotypes with predicted expression levels. The success of these methods critically depends on the identification of eQTLs which are likely to be tissue specific. However, tissue-specific eQTLs identified from these methods do not always have biological functions in the corresponding tissue, due to linkage-disequilibrium (LD) and the correlation of gene expression between tissues. Here, we introduce a new method named T-GEN (**T**ranscriptome-mediated **i**dentification of disease-associated **G**enes with **E**pigenetic a**N**notation) to identify disease-associated genes leveraging tissue specific epigenetic information. Through prioritizing SNPs with tissue-specific epigenetics annotation in developing gene expression prediction models, T-GEN can identify SNPs that are more likely to be “true” eQTLs both statistically and biologically. A significantly higher percentage (an increase of 18.7% to 47.2%) of eQTLs identified by T-GEN have biological functions inferred by ChromHMM [1]. When we applied T-GEN to 208 traits from LD Hub [2], we were able to identify more disease-associated genes (ranging from an increase of 7.7 % to 102%), many of which were tissue-specific. Finally, we applied T-GEN to late-onset Alzheimer’s disease and identified 96 genes located in 15 loci, two loci out of which are not reported in previous GWAS findings.

Introduction

Since 2005, Genome-Wide Association Studies (GWASs) have been very successful in identifying disease-associated variants such as single nucleotide polymorphisms (SNPs) for human complex diseases like Schizophrenia [3]. However, most identified SNPs are located in non-coding regions, making it challenging to understand the roles of these SNPs in disease etiology. Several approaches have been developed recently in linking genes to identified SNPs to provide further insights for downstream analysis [4-8]. PrediXcan [9], FUSION [10] and UTMOST [11] have been applied for utilizing transcriptomic data, such as those from GTEx [12], to interpret identified GWAS non-coding signals and to identify additional associated genes for human diseases. These methods first build models to impute gene expression levels based on genotypes of inferred eQTLs and then identify disease-associated genes by associating phenotypes with predicted expression levels. At the gene level, these methods often implicate a number of genes in a region and it is difficult to distinguish their functional roles. At the SNP level, SNPs (i.e. eQTLs) used in gene expression imputation model are selected through statistical correlation between these SNPs' genotypes and gene expression levels.

Since SNPs in the same LD region are correlated, it is challenging to statistically differentiate regulatory SNPs from other SNPs that are in LD. Besides, as expression levels of a gene in multiple tissues are correlated, selecting eQTLs purely based on gene expression level may result in the same SNPs selected in multiple tissues [13]. This amplifies the correlation of the gene expression level in multiple tissues and leads to significant associations in irrelevant tissues. In general, these methods do not consider available tissue-specific cis-regulation information such as that from epigenetic marks when modeling gene expression, in that the gene expression imputation models are only developed on statistical associations between the observed expression levels and genotypes. There are often different expression regulations across tissues for the same gene, resulting from different epigenetics activities of regulatory SNPs across tissues [14] and tissue specific regulatory information may be available from epigenetic data, in addition to tissue specific expression levels.

To more accurately identify regulatory eQTLs that play functional roles, in this article, we hypothesize that SNPs with active epigenetic annotations are more likely to regulate tissue-specific gene expression, based on current understanding that gene expression regulatory regions are often enriched in regions with epigenetic marks like H3K4me3 and histone acetylation [15]. As for available tissue-specific epigenetic data provided in Roadmap epigenetics data [16], we consider epigenetic marks that have been reported as hallmarks for DNA regions with important functions, such as H3K4me1 signals in enhancers [17]. We note that these epigenetic marks have been used to infer regulatory regions and prioritize eQTLs in previous studies [18-21]. In this paper, we used these epigenetic signals to select effective SNPs out of all candidate cis-SNPs, which are SNPS located from 1MB upstream of gene transcription starting sites to 1MB downstream of gene transcription end sites, when modeling the relationship between gene expression levels and SNP genotypes. By prioritizing the SNPs likely having regulatory roles through epigenetic marks when building gene expression imputation models, our method can better identify eQTLs with functional roles (34% located in predicted functional motifs [22], compared to 17.4% by elastic net used in PrediXcan and FUSION).

We note that our method, T-GEN, can better capture tissue-specific regulatory effects than current methods, which tend to characterize shared genetic regulatory effects across tissues when modeling gene expression. Compared with the observed correlation of gene expression between tissues calibrated in GTEx, most current approaches have inflated gene expression correlation between tissues after gene expression imputation using genotypes. In contrast, our method is able to capture tissue-specific regulatory effects without inflations after expression imputation. Focused more on tissue-specific genetic effects than previous approaches like PrediXcan and FUSION, T-GEN can impute tissue-specific gene expressions for more genes with an increase of 2.6% to 55.3%.

After obtaining gene expression imputation models for each gene in different tissues using data from GTEx, we further studied associations between predicted tissue specific expression levels with GWAS summary statistics from 208 traits at the LD Hub [2] to identify genes associated with these traits. Compared to other gene expression imputation models, T-GEN identified more trait-associated genes. For example, for Alzheimer's disease (AD), with 96 genes located in 15 loci were identified as significantly associated with AD compared to 79 genes in 10 loci identified by PrediXcan, with 2 novel loci located beyond 1MB region near the GWAS significant SNPs. Overall, our method offers a more tissue-specific and interpretable approach to identifying disease-associated genes and to understanding disease mechanisms.

Results

Method overview

Similar to previous methods [10, 11, 23], we use two linear models to analyze gene-level genetic effects on traits mediated by gene expression regulation. Firstly, individual-level genotypes and gene expression data are used to build gene expression imputation models in each tissue. One novel feature of our method is the use of a logit layer to integrate Roadmap epigenetics data [16] for prioritizing regulatory SNPs in this step of developing gene expression imputation models. As a result, SNPs located in regions with certain epigenetic signals (like H3K4me3 and DNase-I hypersensitivity) are more likely to be selected. After obtaining tissue-specific gene expression imputation models for each gene, we combine them with GWAS summary statistics to identify gene-level associations with disease phenotypes. A schematic workflow of T-GEN is shown in **Fig 1**.

Individual genotype and tissue-specific gene expression data from GTEx were used to build gene expression imputation models in each tissue. By utilizing a spike-and-slab prior, SNPs regulating gene expression were selected not only based on their correlation with gene expression levels, but also based on their epigenetic signals in the following model.

$$Y = X\beta + \epsilon,$$

$$\epsilon \sim N(0, \sigma^2 I)$$

$$\beta_k | \gamma_k = 1 \sim \mathrm{N}\left(0, \sigma_\beta^2 \sigma^2 I\right),$$

$$\beta_k | \gamma_k = 0 \sim \delta_0,$$

$$\gamma_k \sim Bernoulli(\pi_k)$$

$$logit(\pi_k) = A_k \omega,$$

where Y is the gene expression vector for a gene in a specific tissue, X is the genotype matrix for SNPs located within 1MB from the upstream or downstream of the gene, β is the tissue-specific coefficient vector for effects of genotypes on gene expression, ϵ is the random effect vector following a normal distribution. For a cis-SNP k, its coefficient on gene expression levels follows a mixture prior of a normal distribution and a point

mass around 0. When the SNP k regulates expression level of its nearby gene, where the indicator γ_k equals 1, then β_k follows a normal distribution, otherwise, β_k is zero. Besides, the probability π_k of the SNP regulating gene expression level of its nearby gene is linked to its epigenetic annotation matrix A_k via a logit model, with ω being the annotation coefficient vector in the logit model. To identify the SNP coefficient vector β under the model setting, a variational Bayesian method [24] is used here. More details are shown in the Methods section.

For comparison, we also consider four other gene expression imputation methods, including elastic net (elnt) [25, 26], Bayesian sparse linear mixed-kind models solved by variational Bayes (vb) [27, 28], elastic net applied only to SNPs having active epigenetic signals (elnt.annot), and Bayesian sparse linear mixed-kind models only on SNPs having active epigenetic signals (vb.annot). For each method, we consider associations between traits with imputed gene expression levels when the imputation quality is good (False Discovery Rate/FDR < 0.05). Therefore, different numbers of genes are studied in association analysis for each method.

In general, we can describe the relationship between the imputed gene expression and selected SNPs in the form of $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is the SNP effect estimates in the imputation models. Different imputation methods lead to different sets of SNPs and effect size estimates. We then use a univariate regression model to test the association between traits and imputed gene expression levels:

$$\mathrm{T} = \mu + \hat{Y}\kappa_Y + \tau$$

The z-score of the gene coefficient κ_Y is denoted as $z = \frac{\widehat{\kappa_Y}}{se(\widehat{\kappa_Y})} = \hat{\beta}^T \Lambda \hat{Z}_X$, where Λ is a diagonal matrix with diagonal elements being the ratio of the standard deviation of SNP genotype over the standard deviation of the imputed gene expression, while $\hat{Z}_X$ is the z-score vector of SNP effects on traits provided in the GWAS results. We test the association between each tissue-gene pair and trait. Significant disease-associated tissue-gene pairs are then identified after multiple comparison adjustment.

Better identification of eQTLs with biological function potentials when imputing gene expression

To study whether our method can better identify functional SNPs as eQTLs in gene expression imputation models, we evaluated the function states of identified eQTLs using function annotations in the HaploReg database [22], which includes the chromHMM model-annotated SNP status [1]. ChromHMM-15/25 is a Hidden Markov Model used to annotate the human genome using ChIP-seq data. We compared the percentages of identified eQTLs with ChromHMM-predicted active status for our method and four other imputation methods.

Fig 2 shows that, among the five methods, our method is more likely to select SNPs with active chromatin states as eQTLs for gene expression imputation. More specifically, 47.2% of eQTLs identified by our method are annotated as having one of the 11 active states such as active transcription starting sites and enhancers, while only 25.2% of SNPs identified by elastic net models are annotated to have one of the 11 active states (**S1 Table**). In the ChromHMM-25 state model (**S1 Fig**), a higher average percentage of SNPs also have one of the 21 active states in our models with 35.1% for our method versus 18.7% for elastic models. Overall, both comparisons showed that our method can better identify SNPs with function potential of regulating gene expression, which indicates T-GEN setting can actually prioritize SNPs with active epigenetic marks since the same Rodemap data are used in both ChromHMM and T-GEN. (Further confirmation of T-GEN's ability in identifying SNPs with function potential may require an external epigenetics annotation database for multiple tissues/cell lines)

Better capture of tissue-specific regulatory effects when imputing gene expression

Many genes are expressed and functional in multiple tissues, and dysfunctions of the same group of genes in different tissues may lead to abnormality for different traits [29, 30]. Therefore, identifying the specific tissue responsible for the observed gene-trait associations is important, which requires us better capturing tissue-specific genetic regulatory effects on gene expression. We compared correlations between the imputed gene expression levels between tissues based on results obtained from the five methods. For a given gene, the correlation of imputed gene expression levels was calculated between each pair of tissues across individuals and correlation was also calculated for the observed expression levels between tissues from GTEx.

Compared with T-GEN, most of other imputation methods often led to higher inflation in gene expression correlation between tissues after expression imputation. Tissue-wise correlation of imputed expression by “elnt” model showed an increase of 165% over the observed correlation (mean value comparison). In comparison, T-GEN only showed the smallest changes, an increase of 3.3% (**Fig 3**). The variational Bayes-like methods had less inflation than elastic net-like methods (191% increase for elnt.annot method) in imputed tissue-wise expression correlations. After adding epigenetic annotations, T-GEN showed lower changes compared to the “true” correlation than vb (10.5% decrease) and lower inflation than vb.annot (13.9% increase). The less inflated correlation (p=1.22e-8, t-test) of results by T-GEN indicates that more dynamic prioritization of SNPs using epigenetic annotations would keep more tissue-specific genetic effects in modeling gene expression.

To study whether eQTL identification has a similar tissue-specific pattern, we checked the numbers of tissues where a SNP was identified as eQTL for the same gene. The percentage of eQTLs only identified in one tissue by T-GEN is 91.8%, which is higher than that by vb.annot (91%), elnt (90.2%) and elnt.annot (88%). Most eQTLs (95%) identified by vb were only identified in a single tissue. The overall tissue-specificity pattern is consistent with that of tissue-wise correlation after gene expression imputation.

In addition to inflation in tissue-wise expression correlations after expression imputation, inflated expression correlation between genes located in the same loci may also lead to incorrect identification of trait-associated genes [31]. Therefore, we compared the local correlation structure of the imputed gene expressions in all five models with the “true” correlation of whole gene expression from GTEx. After dividing the whole genome into pre-defined 825 cytogenetic bands [32], the mean values of pair-wise correlation between genes at a locus were calculated in the whole blood tissue. Across all five models, our model had the lowest mean values of imputed gene expression correlation (**S2 Fig**). These results also suggest that our model is more likely to capture different regulatory genetic effects for genes located in a local region instead of shared regulatory effects.

Larger number of genes whose expression levels can be imputed

The number of genes that can be identified to be associated with a trait of interest is strongly affected by the number of genes with confident expression imputation models. For all five imputation methods, gene expression imputation models were filtered based on the corresponding p values of associations between the imputed and observed expression levels provided in GTEx. Therefore, significant trait-gene associations can only be detected among genes with significant imputation models.

Under the same filtering criterion (an FDR cutoff of 0.05), the number of significant gene expression imputation models selected for each tissue was compared across all five methods (**Fig 4**). T-GEN had an increase in the range of 2.6% to 55.3% compared with the elastic net methods in the number of gene imputation models across 26 tissues. There was a smaller increase in tissues with a smaller sample size like Brain Cortex (4.8%, n=96) and Brain Frontal Cortex BA9 (2.6%, n=92) than those for tissues with larger sample sizes. It is worth to note that, by direct SNP filtering using epigenetic annotations, both “elnt.annot” and “vb.annot” had many fewer genes with high quality imputation models, implying that stringent SNP categorization might lead to loss in power of detecting genes with considerable expression components explained by cis-eQTLs.

To check whether the increased number of gene imputation models was caused by more genes passing the FDR cut-off or more genes getting imputed before FDR filtering, we considered the intersection of imputation models only identified by T-GEN before and after FDR filtering. We found that, many (up to 94.5%) of genes with significant imputation models only identified by T-GEN are genes with imputation models only by T-GEN before FDR filtering. This indicates that T-GEN can identify cis-eQTLs for genes with none cis-eQTLs detected by other methods. Compared with elastic net, among 6035 genes with expression imputation models only identified by our method in whole blood, identified eQTLs for 3256 (54%) genes have validated regulatory effects based on high through-put experiment results [33], which implies that the eQTLs specifically identified by T-GEN are more likely to have real biological functions. This indicates that

our method can identify cis-eQTLs of genes which cannot be detected merely based on statistical correlation between genotypes and gene expressions in elastic net methods.

Gene expression imputation accuracy was compared across all five methods through both five-fold cross validation analysis and prediction in an external dataset from the CommonMind Consortium (CMC; http://www.synapse.org/CMC). We compared the R^2 (squared correlation) between the observed and imputed expression levels of our method with that from the other methods. In five-fold cross-validation analysis using the GTEx data, our method showed R^2 decrease from 7.99% (compared with vb.annot and vb) to 22.06% (compared with elastic net), compared to the other four methods (**S3 Fig**). Among all the five methods, we note that T-GEN built the largest number of genes (about 1.6 folds over those of the other four methods) cross validation models. For prediction analysis in the CommonMind dataset, expression imputation models trained on the GTEx Brain Cortex BA9 tissue were used to predict gene expression levels based on individual genotype data of CommonMind. All five methods had similar R^2 (~0.007) between predicted gene expression and whole expression in the Brain Cortex BA9 tissue of CommonMind (**S4 Fig**) but we note that our method included more genes whose expression levels can be imputed. For all 5 methods, in-sample cross-validation performance was always better than that in out-sample prediction as expected.

To assess the influences of imputation accuracies on the identification of trait-associated genes, we compared the number of genes identified as trait-associated ones out of genes imputed by both T-GEN and one of the other four methods. When considering genes that could be imputed by both T-GEN and one of the three methods (elnt, vb, vb.annot), more genes could be identified as associated with one of 208 traits in LD Hub by T-GEN with an increase ranging from 1.6% (compared with elnt) to 84% (compared with vb.annot). On the other hand, all these three methods showed higher imputation accuracies, ranging from 6.8% (vb.logit) to 80% (elnt), than T-GEN for intersected gene models. Compared with elnt.annot, T-GEN identified fewer trait-associated genes (-17%) and also showed lower imputation accuracies (-42%) for genes with imputation models by both methods. These results suggest that higher

imputation accuracy does not necessarily lead to more trait-associated genes identified. Genes with imputation models only identified by our method have lower R2 (cross validation/ Prediction/ Whole R2) than those identified by both our method and elastic net-like methods, implying our method can detect predictive SNPs in genes whose cis eQTLs cannot be identified only based on genetic correlation with gene expression levels or stiff leverage of SNPs with epigenetic signals using direct filtering. Overall, although T-GEN had lower performance of imputation accuracy across genes than other methods, it is able to impute expression levels for more genes and provides larger gene search space (gene candidate pool) for later gene-trait association identification.

More genes identified to be associated with 208 traits in LD Hub

To evaluate the performance in identifying trait-gene associations, which is the ultimate goal of gene expression imputation, we applied T-GEN to GWAS summary statistics of 208 traits from LD Hub [2]. After Bonferroni correction, significant trait-associated genes were identified in each tissue. T-GEN showed a 13.6% (compared with elnt) to 112% (compared with vb.annot) increase in the average number of significant trait-associated genes across 208 traits (**Fig 5A**). More specifically, in 93 lipid traits (**S2 Table**), T-GEN showed significant (Wilcoxon rank test) increase in the number of identified associated-genes compared to all the other four methods (**Fig 5B**). After aggregating identified genes into pre-defined cytogenetic bands, we observed similar pattern in the numbers of identified trait-associated loci across 208 traits. T-GEN identified the largest number of associated loci (11.5 loci on average) across all five methods (**S5 Fig**), showing a 5.5% (compared with elnt) to 31% increase (compared with vb.annot).

To assess the ability of identifying associated genes in tissues most relevant to traits, we compared the numbers of identified genes in tissues with the highest heritability enrichment across five methods (**S3 Table**). Using LD score regression [34] and annotation from GenoSkyline-Plus [35], for each trait, we identified the tissue with the highest fold heritability enrichment among all tissues with significant ($p<0.05$) heritability enrichment . The number of significantly-associated genes identified in heritability-enriched tissues was compared across five methods (**Fig 6**). In the most-enriched tissues, T-GEN identified the largest numbers of significantly associated genes (Wilcoxon rank test; compared with vb p: 0.06; compared with vb.annot p: 2.1e-5; compared with elnt p: 0.5, compared with elnt.annot p: 0.6). When comparing the average number of genes identified in the most-enriched tissue and those in the other tissues, T-GEN also shows an increase of 68%, only smaller than that of elnt (69%) and vb.annot (78%).These results suggest that our method can better identify disease-associated genes in the trait-relevant tissues.

Overall, T-GEN showed consistent improvement not only in the total number of trait-gene associations, but also in identifying tissue-specific associated genes in tissues most relevant to a trait.

Genes identified for late-onset Alzheimer's disease (LOAD)

To further investigate the performance of our method in identifying trait-associated genes in detail, we are focused on AD-associated genes that were only identified by our method. Corrected by the total number of tissue-gene pairs (258,039), 96 genes in 15 loci (**Table 1**) were identified by our method as significantly associated with AD (N=54,162) using summary stats in LD Hub, with five loci identified in the brain tissues (brain caudate basal ganglia, brain anterior cingulate cortex BA24, brain hippocampus, brain cortex, and brain frontal cortex BA9) and four loci identified in the whole blood. Thirteen out of the 15 loci were identified as disease loci in previous GWAS studies for AD. In comparison, 79 genes in 10 loci were identified by models built using elastic net, which was used on PrediXcan, 81 genes in eight loci were identified by elnt.annot models, 61 genes in three loci were identified by vb.annot models, and 81 genes in 11 loci were identified by models built by vb method.

Compared with AD-associated genes identified by the other four methods, three loci were only identified by our method. One locus is located on 14q32.12 including *LGMN* (p=9.45e-8). *LGMN* is located about 200kb to previously identified GWAS significant SNP rs1049863 [36]. As nearest genes located to the GWAS hit, *SLC24A4* and *RIN3* are two potential genes contributing to the GWAS signal of this locus in previous GWAS studies, whose functions are yet understood. Our method identified *LGMN* at this locus. *LGMN* encodes protein AEP that cleaves inhibitor 2 of PP2A and can trigger tau pathology in AD brain [37], which makes *LGMN* a potential signal gene in this locus. Another locus is 16p22.3, where *COG4* was identified (p=1.35e-7). *COG4* encodes a protein involved in the structure and function of Golgi apparatus. Recent research has shown that defects in the Golgi complex are usually associated with neurodegenerative diseases like AD and Parkinson's disease by affecting function of cellular trafficking proteins like Rab-GTPase and SNAREs [38]. The third novel locus is 6p12.3, where *CD2AP* (p=5.70e-8) and *RP11-385F7.1* (p=1.27e-7) were identified. This locus was previously identified as a susceptibility locus in AD [36] with *CD2AP* reported as the locus signal genes affecting the amyloid precursor protein (APP) metabolism and production of amyloid-beta [39]. The additional *RP11-385F7.1* identified by our method

is a long non-coding RNA (lncRNA) located near the promoter region of *CD2AP*. The level of RP11-385F7.1 lncRNA has a strong Pearson correlation with gene expression level of *COQ4* [40], which was also the gene only identified by our method and encodes a protein that may serve as an antioxidant strategy target for AD [41].

Most of the loci identified in our method are reported GWAS loci, like *CLU* locus (8p21.2), *BIN1* locus (2q14.3), *CR1* locus (1q32.2), *MS4A6A* locus (11q12.1), *PTK2B* locus (8p21.1, 8p21.2), and *CELF1* locus (11p11.2). In many of those GWAS loci, there are additional genes identified by our method compared to the reported gene or the gene located near the GWAS risk SNP. In the *PTK2B* locus, we identified an additional gene *ADRA1A* (p=1.85e-8), which is involved in neuroactive ligand-receptor interaction and calcium signaling and has been implicated as potential gene in late-onset Alzheimer's disease via gene-gene interaction analysis [42]. In the *MS4A6A* locus, we identified an additional gene named *OSBP* (p=1.56e-7), which transports sterols to nucleus where the sterol would down-regulate gens for LDL receptor, which is important in the etiology of AD [43]. Compared with previously reported *MS4A6A* gene clusters with unclear functions in the locus, the *OSBP* gene is more likely to be the gene contributing to the GWAS signal in the locus.

Two out of 15 loci identified by our method are novel compared with published GWAS loci in previous studies [36, 44-46]. One locus is located on 16q22.3, where *COG4* was identified (p=1.35e-7), which is also the locus only identified by our method, compared to results of the other 4 methods. *COG4* is located beyond 1MB flanking regions of all reported GWAS hits. Two of the identified eQTLs (**S4 Table**) of *COG4* are potential GWAS hits (rs7196032, p=1.1e-4 and rs7192890: p=3.5e-3; **Fig 7**). The protein encoded by *COG4* is important in determining Golgi apparatus structures [47] and key functions like intracellular transportation and glycoprotein modification [48]. Studies have indicated that Golgi apparatus may play an important role in the transportation and mature of gamma-secretase, which helps the final step of generating normal amyloid β-protein [49]. Another gene important in Golgi complex and transportation, *TMEM115* [50], was also identified in the other novel significant AD-associated locus (p=1.80e-8)

by our method. Both identified genes underlie the putative importance of Golgi-involved transportation. The assoicated gene identified in the other novel loci, *TMEM115* is located 1MB (**Fig 8**) downstream of the identify GWAS locus (the *PICALM* locus), and two of the identified eQTLs of *TMEM115* by our method were located in the GWAS loci (rs536841: p=5.27e-14 and rs541458: p=5.67e-12). On the other hand, *TMEM135* is a target gene for liver X receptors (LXR), which is involved in removing excessive cholesterol from peripheral tissues. Increased cholesterol levels in brain tissue is associated with accumulated amyloid-β [51], suggesting the potential role of *LXR* and *TMEM115* in the etiology of AD.

To validate our results, we further tried to replicate our findings in an independent GWAS for Alzheimer's disease using inferred phenotypes based on family history (GWAX, N=114,564) [52]. Out of all 96 genes identified by our method using the AD summary statistics from LD Hub, 48 (50%) genes were successfully replicated in this independent GWAS study ($p < 5.2e-4$). For the other four methods, only 9.9% (8/81, vb) to 41% (33/81, elnt.annot) genes were replicated in the GWAX data. Results (**S5 Table**) replicated in an independent study can be used to rule out signals identified by chances and uncontrolled biases [53]. The high replication rate of results in an external study further confirmed the power of T-GEN.

Discussion

In this paper, we introduced a new method called T-GEN, which leverages epigenetic signals to identify trait-associated genes using expression-mediated genetic effects. Different from previous methods, T-GEN prioritizes SNPs with active epigenetic annotations when building gene expression imputation models using data from GTEx. Through epigenetic prioritization, we found that the imputation model was more likely to include a higher percentage of SNPs with functional potentials and more gene imputation models could be built than other methods. In addition, imputed gene expression levels showed lower inflated regional correlations and across-tissue correlations. Application of our tissue specific gene expression imputation method to more than 200 traits in the LD Hub identified more genes/loci associated with these traits. Our method can especially identify the largest number of significantly associated genes in tissues with the highest enrichments of trait heritability.

When applied to Alzheimer's disease, T-GEN identified the largest number of associated genes (96 genes in 15 loci) compared with all the other four methods. We found novel association signals in previously identified loci including *LGMN* in the *SLC24A4/RIN3* locus, *ADRA1A* in the *PTK2B* locus and *OSBP* in the *MS4A6A* locus. Besides, two novel loci were identified on chromosomes 3 and 16. Genes identified by our method suggest putative roles of several biological processes in Alzheimer's disease, including mitochondrial dysfunction (indicated by *UQCR11*, *MTCH2* and *TMEM135*; [54, 55]), cholesterol transportation (indicated by *TMEM135* and *OSBP*;[56]), and neuron activities (indicated by *ADRA1A)*. Fifty out of 96 identified associated-genes were replicated in an independent GWAS dataset. Most importantly, the genes identified from our method for the Alzheimer's disease showed higher replication rates than those identified from other methods.

Overall, compared with previous methods, T-GEN showed consistent improvement in identifying associated genes across multiple traits. Limited by individual-level epigenetic data availability and sample sizes of used data, T-GEN also has some limitations, such as the accuracies of predicting gene expression levels. Utilizing epigenetic annotations

in the reference panel from Roadmap would limit the advantage of leveraging epigenetic information in gene expression imputation process, since the epigenetic signals may vary across individuals just like gene expression levels. In addition, only cis-SNPs are used here for gene expression imputation, while a larger proportion of gene expression may be explained by trans-eQTLs [57-59]. Integrating trans-eQTLs into gene expression imputations would help to identify more trait-associated genes and also co-regulatory peripheral genes and core genes [60]. By using individual-level data with larger sample sizes and matched individual-level epigenetic annotation would better aid in identifying disease-associated genes and providing better insights for etiology of complex diseases using T-GEN in the future.

Methods

Bayesian variable selection model

To model the genetic effects of SNPs on the expression level of a gene in a single tissue, we used the following bi-level variation selection model to select SNPs:

where is the gene expression level vector (denotes the sample size in this tissue), denotes the genotype matrix for cis-SNPs for this gene, which is centered, but not scaled in practice, and p is the number of SNPs. In analysis, denotes the gene expression in a specific tissue while denotes the genotype matrix for SNPs located within 1MB from the upstream or downstream of the gene. We used the vector to represent the effects of those cis-SNPs on the gene expression level while is the random error vector, which is assumed to follow a normal distribution. In a typical data set with both genotype data and RNA-seq data for the same group of individuals, the sample size is usually much smaller than the number of cis-SNPs for a gene (p), some constraints are needed to allow for accurate SNP selection and effect size estimation. Therefore, to select effective SNPs from thousands of candidate cis-SNPs, we assume the effect size vector comes from a mixture of normal distribution and a point mass at zero. is the indicator variable denoting whether the kth SNP has an effect on gene expression level. When a SNP k has an effect on gene expression level (i.e.), we assume that its effect size follows a normal distribution with 0 mean and variance .

To more accurately identify eQTLs with potential biological functions, we integrate the epigenetics annotation to prioritize the SNPs with active epigenetic signals including H3K4me1, H3K4me3, and H3K9ac. To achieve this goal, we used a logit link to associate the epigenetic annotation with the probability of a SNP being an eQTL:

where is the epigenetic signal vector for the SNP and the vector is the epigenetic signal effect vector, where is the number of epigenetic signal

categories considered, such as DNA methylation and histone marker status. We assume that the variance of the coefficients is .

Prior assumptions on hyper parameters

We make the following assumptions on the hyper parameters in the model above:

,

i.e. we assume that these three parameters follow either inverse gamma or gamma distribution.

In practice, we do not have any prior information to select the values of and , especially for a_0 and b_0. Therefore, twenty grid points of are initially sampled from the following distribution:

Variational Bayes inference

For the convenience of defining selected SNPs, we firstly introduce PPS(k) to denote the posterior probability of a SNP being selected:

which is basically the weighted average of the posterior probability of this SNP being effective while the weight is the model likelihood. is defined as the prior parameter set .

To address the computational challenge in fitting our model to high-dimensional genotype data, we applied variational Bayes method, which is an alternative of the Markov Chain Monte Carlo (MCMC) method for statistical inference. Variational Bayes provides an analytical approximation of the posterior probability by selecting one from a family of distributions with the minimum Kullback-Leibler divergence to the exact posterior.

Under each set of chosen prior parameters variational Bayes is applied to update other parameters to get the optimal model. Besides, for a set of chosen priors, we define the posterior probability of a SNP being selected as:

When updating one parameter, we need to get the approximation ($Q^*(.)$) of its posterior distribution via taking the expectations of all the other parameters in the full probability equation:

$$P(Y|X,A,\theta) = \int\int\int P(Y|\beta,\epsilon,X)P\left(\beta|\sigma_\beta^2,\sigma^2,\gamma\right)P(\gamma|\pi)P(\pi|A,\omega)P(\omega|\eta)d\eta d\sigma^2 d\sigma_\beta^2.$$

Updating beta coefficients:

For updating the SNP coefficient β_k of the SNP k in the model, we would get:

$$\ln Q^*(\beta_k) = E_{rest}\left[\ln P(Y|\beta,\epsilon,X) + \ln P\left(\beta\middle|\sigma_\beta^2,\sigma^2,\gamma\right)\right] + constant$$

$$= E_{rest}\left[\sum_{i=1}^{n} -\frac{(y_i - x_i\beta)^2}{2\sigma^2} - \sum_{j=1}^{p}\frac{\gamma_j\beta_j^2}{2\sigma_{\beta,j}^2\sigma^2}\right] + constant$$

$$= E_{rest}\left[\sum_{i=1}^{n} -\frac{(y_i - x_i\beta)^2}{2\sigma^2} - \frac{\gamma_k\beta_k^2}{2\sigma_{\beta,k}^2\sigma^2}\right] + constant$$

Therefore, by taking the expectation of other parameters, $\beta_k$approximately follows a normal distribution $\beta_k|\gamma_k = 1 \sim N(\mu_k, s_k^2)$ where:

$$s_k^2 = Var(\beta_k|\gamma_k = 1) = \frac{\sigma^2}{(X^TX)_{kk} + 1/\sigma_\beta^2}$$

$$\mu_k = E(\beta_k|\gamma_k = 1) = \frac{s_k^2}{\sigma_k^2}[(X^TX)_k - \sum_{j\neq k}(X^TX)_{jk}\alpha_j\mu_j]$$

Updating the probability of being selected

The probability of the kth SNP getting selected is dependent not only on the relationship between the SNP genotype and phenotype (gene expression level here), but also on the epigenetic signals of the SNP, which can be described through the equation below:

$$\ln Q^*(\gamma_k) = E_{rest}\left[\ln P(Y|X,\beta,\sigma^2) + \ln P\left(\beta\middle|\gamma,\sigma_\beta^2,\sigma^2\right) + \ln P(\gamma|A,\omega)\right] + constant$$

After solving the equation above term by term, we can get (details shown in the **S1 Appendix**):

$$\ln Q^*(\gamma_k = 1) = \frac{\mu_k^2}{2s_k^2} + \frac{1}{2}\ln\sigma^2\sigma_\beta^2 + \ln\frac{1}{1+\exp(-A_k\omega)} + constant$$

$$\ln Q^*(\gamma_k = 0) = \ln\frac{1}{1+\exp(A_k\omega)} + constant$$

$$\frac{\alpha_k}{1-\alpha_k} = \exp\left(\frac{\mu_k^2}{2s_k^2}\right) \times \frac{1}{\sigma_\beta^2\sigma^2} \times \frac{1+\exp(A_k\omega)}{1+\exp(-A_k\omega)}$$

Updating coefficients in the logit link of epigenetic signals

Parameters in the logit link are related to both the SNP epigenetic signals and the status of the SNP being effective or not, which shown in the following:

$$\ln Q^*(\omega) = E_{rest}[\ln P(\gamma|A,\omega) + \ln P(\omega|\eta)] + constant$$

.

Similar to the idea of using variational Bayes method in a logistic regression introduced in the literature [61], ωalso has a normal distribution with parameters:

$$\omega_N = E(\omega|rest) = V_N \sum_{j=1}^{p} \frac{\gamma_j}{2} A_j^T$$

$$V_N^{-1} = \frac{1}{Var(\omega|rest)} = E_{rest}(\eta)I + 2\sum_{j=1}^{p} \lambda(\xi_j) A_j^T A_j$$

To calculate the overall PPS, we need to get the likelihood of the full model under each prior parameter setting. Since we actually approximate the largest value of the lower bound in the full model likelihood, the estimated optimal lower bound is used to substitute for the exact likelihood.

$$\ln P(Y|X,Q,A) \geq F(Q,\theta)$$

$$= \int\int\int \mathrm{q}(\beta,\gamma,\omega;\theta)\log \frac{P(Y,\beta,\gamma,\omega|X,A;\theta)}{q(\beta,\gamma,\omega;\theta)} \mathrm{d}\beta\mathrm{d}\gamma\mathrm{d}\omega$$

$$= \mathrm{E}_{\beta,\gamma,\omega}[\log P(Y,\beta,\gamma,\omega|X,A;\theta)] - E_{\beta,\gamma,\omega}[\log q(\beta,\gamma,\omega|\theta)]$$

where

$$\log P(Y,\beta,\gamma,\omega|X,A,\theta) = \log P(Y|\beta,\sigma^2) + \log P(\beta|\gamma,\sigma^2,\sigma_\beta^2) + \log P(\gamma|\omega,A) + \log P(\omega|\eta)$$

$$\log q(\beta,\gamma,\omega;\theta) = \log q(\beta|\gamma;\theta) + \log q(\gamma|\omega;\theta) + \log q(\omega;\theta)$$

Then, the lower bound becomes:

$$\mathrm{F}(\mathrm{Q},\theta) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{\left||Y - X\alpha\mu|\right|^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^{P}[(X^TX)_{jj}Var(\beta_j)]$$

$$-\sum_{j=1}^{P} \alpha_j \log\frac{\alpha_j}{\pi} - \sum_{j=1}^{P}(1-\alpha_j)\log\frac{1-\alpha_j}{1-\pi}$$

Model training and evaluation

Gene expression level imputation models for 26 tissues were trained using the RNA-seq and genotype data from the GTEx project [12] and epigenetic data from the Roadmap Epigenomics Project. Only tissues with both epigenomics data from the Roadmap Epigenomics Project and the GTEx data were considered. For genotype data, we first removed less common SNPs (minor allele frequency < 0.01) and SNPs with ambiguous alleles. For each gene, we considered SNPs located from 1 Mb upstream of its transcription starting site to 1 Mb downstream of its transcription end site. For RNA-seq data from GTEx, we first normalized these data and further adjusted for possible confounding factors including sequencing platform, top three principal components, sex and probabilistic estimating of expression residuals (PEER) factors [62]. More specific details of preprocessing expression data from GTEx were described in our previous publication [11].

We used five-fold cross validation to evaluate our gene expression imputation models. More specifically, for each tissue, samples were randomly divided into five groups with about the same size. We compared our method with four other methods via training models using four groups of data and then testing them on the fifth group. To compare performance among different methods, squared correlation between the observed and imputed expression levels ( ) was used. In our method, with the number of epigenetic categories existing in Roadmap varying across different tissues (**S6 Table**), all available epigenetic annotation categories were used for each tissue. For elastic net models (elnt), as what previous studies did [9], the parameter α was set to be 0.5 and the optimal λ was selected via the function cv.glmnet provided in the 'glmnet' package [63]. For elastic net models with epigenetic signals direct filtering (elnt.annot), we first removed cis-SNPs without H3K4me1 [64], H3K4me3 [65] or H3K9ac [66] signals, which have been reported as associated with gene expression regulation. After filtering, the remaining cis-SNPs were the used to build imputation models. As for methods directly using variational Bayes (vb), the model is similar to our method, without the logit link of epigenetic annotations. We also applied the direct variational Bayes method to SNPs with some epigenetic signals (same filtering process in elnt.annot). Paired Wilcoxon rank test was used to compare the performance of the models across different genes in

each tissue. For further validation, different models were used to predict gene expression in an external data set of brain expression data CommonMind Consortium (www.synapse.org/CMC). For gene-trait association identification, only gene expression imputation models with FDR < 0.05 were considered.

Association analysis

After estimating SNP effects on gene expression levels, we used published GWAS summary-level results to identify the gene-level genetic effects on different traits mediated by gene expression. Just as the imputation model introduced above, for a gene, in a single tissue, its gene expression can be modeled via the genotypes of cis-SNPs while the expression-trait relation can be modelled by , where is the intercept and is the error term in the model.

The test statistic for the effect of gene expression on trait T is

where is the estimated expression effect on the trait and stands for the stand error for the estimated effect. For a linear model, $\hat{\kappa}$ has the following form

where is the vector of estimated SNP effects on gene expression levels, the denote the vector of GWAS SNP-level effect size for those identified effective SNPs in the gene expression imputation model. is a diagonal matrix whose elements are the genotype variances for each SNP and is the variance of imputed expression levels for this gene in the tissue.

To calculate the test statistic gene-trait z-score, we also need to get the stand error of estimated

:

Besides, based on three linear models between any two out of these three variables (genotypes, imputed expression , and trait phenotypes), we can get the variance proportions of outcomes explained by predictors as:

where is the estimated SNP-level effect size presented in the GWAS result and is the estimated SNP variance diagonal matrix. Therefore, the standard error of $\hat{\kappa}_Y$ is

.

Then, the -score is , where is the SNP coefficient vector in the gene expression imputation model, the matrix is a diagonal matrix, whose diagonal elements are the ratio of standard errors of SNPs over the standard error of imputed gene expression. Besides, the is the SNP-trait -score vector provided in the GWAS results. Intuitively, the statistic is the weighted sum of GWAS z-scores for SNPs selected in gene expression imputation models and weights are proportional to the variance proportion of imputed expression levels explained by each SNP.

GWAS data analysis

The summary statistics files of 208 traits were downloaded from the LD Hub website (http://ldsc.broadinstitute.org/gwashare/), imputation models built by all five methods across 26 tissues were applied to identify trait-associated genes. For these five methods, individual-level genotypes from the GTEx database were used to calculate the standard deviations of SNPs and also the stand errors of imputed gene expressions. For example, in the case of Alzheimer's disease, the summary statistics file was originally from the International Genomics of Alzheimer's Project (IGAP) with a sample size of 74,046 [36].

Acknowledgements

This work was partially supported by NIH grants 3P30AG021342-16S2 and R01GM122078 and NSF grant DMS 1713120.

Competing financial interests

The authors declare no competing financial interests

Author contributions

W.L., M.L., and H.Z. conceived the study and developed the statistical model.
W.L. and M.L. implemented the method.
W. Z., G. Z., X. W. and J. W. assisted in building imputation models.
W.L., M.L., and W.Z. performed the statistical analyses.
W.L., M.L., and H.Z. wrote the manuscript.
H.Z. advised on statistical and genetic issues.
All authors read and approved the manuscript.

References

1. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nature Methods, 2012. **9**: p. 215.
2. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2017. **33**(2): p. 272-279.
3. Li, Z., et al., *Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia.* Nature Genetics, 2017. **49**: p. 1576.
4. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* American journal of human genetics, 2011. **89**(1): p. 82-93.
5. Chen, L.S., et al., *Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data.* The American Journal of Human Genetics, 2010. **86**(6): p. 860-871.
6. Hormozdiari, F., et al., *Colocalization of GWAS and eQTL Signals Detects Target Genes.* The American Journal of Human Genetics, 2016. **99**(6): p. 1245-1260.
7. Joehanes, R., et al., *Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies.* Genome biology, 2017. **18**(1): p. 16-16.
8. Dobbyn, A., et al., *Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS.* American journal of human genetics, 2018. **102**(6): p. 1169-1184.
9. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nature Genetics, 2015. **47**: p. 1091.
10. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nature Genetics, 2016. **48**: p. 245.
11. Hu, Y., et al., *A statistical framework for cross-tissue transcriptome-wide association analysis.* bioRxiv, 2018.
12. Carithers, L.J., et al., *A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project.* Biopreservation and Biobanking, 2015. **13**(5): p. 311-319.
13. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**: p. 204.
14. Sonawane, A.R., et al., *Understanding Tissue-Specific Gene Regulation.* Cell Reports, 2017. **21**(4): p. 1077-1088.
15. Li, B., M. Carey, and J.L. Workman, *The Role of Chromatin during Transcription.* Cell, 2007. **128**(4): p. 707-719.
16. Romanoski, C.E., et al., *Roadmap for regulation.* Nature, 2015. **518**: p. 314.
17. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**: p. 108.
18. Berger, S.L., *Histone modifications in transcriptional regulation.* Current Opinion in Genetics & Development, 2002. **12**(2): p. 142-148.
19. Cheng, C., et al., *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets.* Genome Biology, 2011. **12**(2): p. R15.
20. Cheng, C. and M. Gerstein, *Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.* Nucleic Acids Research, 2012. **40**(2): p. 553-568.
21. Dong, X., et al., *Modeling gene expression using chromatin features in various cellular contexts.* Genome Biology, 2012. **13**(9): p. R53.

22. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* Nucleic Acids Research, 2012. **40**(D1): p. D930-D934.
23. Barbeira, A.N., et al., *Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.* Nature Communications, 2018. **9**(1): p. 1825.
24. Carbonetto, P. and M. Stephens, *Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies.* Bayesian Anal., 2012. **7**(1): p. 73-108.
25. Zou, H. and T. Hastie, *Regularization and Variable Selection via the Elastic Net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 2005. **67**(2): p. 301-320.
26. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nature genetics, 2015. **47**(9): p. 1091-1098.
27. Carbonetto, P., X. Zhou, and M. Stephens, *varbvs: Fast Variable Selection for Large-scale Regression.* eprint arXiv:1709.06597, 2017: p. arXiv:1709.06597.
28. Zhou, X., P. Carbonetto, and M. Stephens, *Polygenic Modeling with Bayesian Sparse Linear Mixed Models.* PLOS Genetics, 2013. **9**(2): p. e1003264.
29. Carlo, A.-S., et al., *The pro-neurotrophin receptor sortilin is a major neuronal apolipoprotein E receptor for catabolism of amyloid-β peptide in the brain.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2013. **33**(1): p. 358-370.
30. Mortensen, M.B., et al., *Targeting sortilin in immune cells reduces proinflammatory cytokines and atherosclerosis.* The Journal of clinical investigation, 2014. **124**(12): p. 5317-5322.
31. Wainberg, M., et al., *Transcriptome-wide association studies: opportunities and challenges.* bioRxiv, 2018.
32. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic acids research, 2004. **32**(Database issue): p. D493-D496.
33. van Arensbergen, J., et al., *Systematic identification of human SNPs affecting regulatory element activity.* bioRxiv, 2018.
34. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nat Genet, 2015. **47**.
35. Lu, Q., et al., *Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease.* PLOS Genetics, 2017. **13**(7): p. e1006933.
36. Lambert, J.-C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nature Genetics, 2013. **45**: p. 1452.
37. Basurto-Islas, G., et al., *Activation of Asparaginyl Endopeptidase Leads to Tau Hyperphosphorylation in Alzheimer Disease.* Journal of Biological Chemistry, 2013. **288**(24): p. 17495-17507.
38. Climer, L.K., M. Dobretsov, and V. Lupashin, *Defects in the COG complex and COG-related trafficking regulators affect neuronal Golgi function.* Frontiers in Neuroscience, 2015. **9**(405).
39. Dunstan, M.L., et al., *THE ROLE OF CD2AP IN APP PROCESSING.* Alzheimer's & Dementia: The Journal of the Alzheimer's Association, 2016. **12**(7): p. P458-P459.
40. Amlie-Wolf, A., et al., *Inferring the molecular mechanisms of noncoding Alzheimer's disease-associated genetic variants.* bioRxiv, 2018.
41. Wadsworth, T.L., et al., *Evaluation of coenzyme Q as an antioxidant strategy for Alzheimer's disease.* Journal of Alzheimer's disease : JAD, 2008. **14**(2): p. 225-234.

42. Meda, S.A., et al., *Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative.* Neurobiology of aging, 2013. **34**(5): p. 1518.e9-1518.e1.518E18.
43. Claus, U.P. and J. Sebastian, *Functional Role of Lipoprotein Receptors in Alzheimers Disease.* Current Alzheimer Research, 2008. **5**(1): p. 15-25.
44. Need, A.C., et al., *A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB.* Human Molecular Genetics, 2009. **18**(23): p. 4650-4661.
45. Hong, C. and P. Tontonoz, *Liver X receptors in lipid metabolism: opportunities for drug discovery.* Nature Reviews Drug Discovery, 2014. **13**: p. 433.
46. Fishilevich, S., et al., *GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.* Database, 2017. **2017**: p. bax028-bax028.
47. Richardson, B.C., et al., *Structural basis for a human glycosylation disorder caused by mutation of the COG4 gene.* Proceedings of the National Academy of Sciences, 2009. **106**(32): p. 13329.
48. Pokrovskaya, I.D., et al., *Conserved oligomeric Golgi complex specifically regulates the maintenance of Golgi glycosylation machinery.* Glycobiology, 2011. **21**(12): p. 1554-1569.
49. Hu, Z., et al., *The Study of Golgi Apparatus in Alzheimer's Disease.* Neurochemical Research, 2007. **32**(8): p. 1265-1277.
50. Ong, Y.S., et al., *TMEM115 is an integral membrane protein of the Golgi complex involved in retrograde transport.* Journal of cell science, 2014. **127**(Pt 13): p. 2825-2839.
51. !!! INVALID CITATION !!!
52. Liu, J.Z., Y. Erlich, and J.K. Pickrell, *Case–control association mapping by proxy using family history of disease.* Nature Genetics, 2017. **49**: p. 325.
53. Kraft, P., E. Zeggini, and J.P.A. Ioannidis, *Replication in genome-wide association studies.* Statistical science : a review journal of the Institute of Mathematical Statistics, 2009. **24**(4): p. 561-573.
54. Onyango, I.G., J. Dennis, and S.M. Khan, *Mitochondrial Dysfunction in Alzheimer's Disease and the Rationale for Bioenergetics Based Therapies.* Aging and disease, 2016. **7**(2): p. 201-214.
55. Moreira, P.I., et al., *Mitochondrial dysfunction is a trigger of Alzheimer's disease pathophysiology.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2010. **1802**(1): p. 2-10.
56. Martins, I.J., et al., *Cholesterol metabolism and transport in the pathogenesis of Alzheimer's disease.* Journal of Neurochemistry, 2009. **111**(6): p. 1275-1308.
57. Price, A.L., et al., *Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans.* PLOS Genetics, 2008. **4**(12): p. e1000294.
58. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nature Genetics, 2012. **44**: p. 1084.
59. Liu, X., et al., *Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues.* The American Journal of Human Genetics, 2017. **100**(4): p. 605-616.
60. Liu, X., Y.I. Li, and J.K. Pritchard, *Trans effects on gene expression can drive omnigenic inheritance.* bioRxiv, 2018.
61. Drugowitsch, J., *Variational Bayesian inference for linear and logistic regression.* arXiv preprint arXiv:1310.5438, 2013.
62. Stegle, O., et al., *Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses.* Nature Protocols, 2012. **7**: p. 500.

63. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software; Vol 1, Issue 1 (2010), 2010.
64. Cheng, J., et al., *A Role for H3K4 Monomethylation in Gene Repression and Partitioning of Chromatin Readers.* Molecular Cell, 2014. **53**(6): p. 979-992.
65. Liang, G., et al., *Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(19): p. 7357.
66. Zhou, J., et al., *Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression.* Plant Molecular Biology, 2010. **72**(6): p. 585-595.

Figures

Fig 1. The general scheme of our method

Gene expression imputation models were built based on gene expression matrix Y of a specific gene in a tissue, the genotype matrix X and epigenetics annotation matrix A of cis-SNPs of the gene. The annotation matrix was used to select SNPs having regulatory effects on gene expression since we assumed that only part of cis-SNPs have effects on gene expression of the nearby gene. After getting SNP coefficient vectors β in each tissue for the gene, we combined β's with GWAS summary stats and then derived the gene-level association statistics for each disease in each tissue.

Fig 2. More SNPs with functional potentials were idenfied by T-GEN.

The figure compares the percentages of SNPs in gene expression imputation models having active ChromHMM15 annotated states across 3 different methods(elnt, vb and vb.logit) using gene expression and genotype data from GTEx in 26 tissues. The "elnt" model was built via elastic net, the "vb" model was built via a variational bayesian method while "T-GEN" models were built using our method (variational bayesian method with a logit link). Across all 26 tissues, imputation models built by our method have higher percentage of SNPs with active ChromHMM annotated states (indicated by blue bars). X axis denotes the mean percentage of SNPs in imputation models having ChromHMM15 annotated states in each tissue for each models. The dotted lines are the mean values of R^2 across 26 tissues for each method.

Fig 3. Less inflation in correlation of imputed gene expression between tissues was observed by our model.

Correlation of imputed gene expression levels for matched genes between tissues were compared across five methods and also for the total gene expression level from GTEx (indicated by "true" in the figure). The "elnt.annot" model was built by selecting only SNPs with active epigenetic signals (see **Methods**) using elastic net. The "vb.annot" model was built by only selecting SNPs with active epigenetic signals using the variational Bayesian method. Values located above red diamonds indicate the mean levels for boxplots.

Fig 4. T-GEN can build more gene expression imputation models than other methods across all tissues.

The difference in the number of genes with high confidence (FDR < 0.05) imputation models between that from T-GEN and that from other methods. Dotted lines show the mean values of the increase of gene models imputed by our method over other methods.

Fig 5. More genes were identified as trait-associated by T-GEN across 208 traits

from the LD Hub.
(**A**) Applied to 208 traits from LD Hub, significantly trait-associated genes were identified in 26 tissues (p-values threshold: 0.05 divided by the number of gene-tissue pairs). Each boxplot represents the distribution of the number of differences between that identified from our tissue-specific analysis and that identified from the four other methods. (**B**) In 93 lipid traits, numbers of identified associated genes were compared acorss all five methods. Numbers above each line show the corresponding p values (wilcoxon rank test) between two corresponding models. Y axis is truncated at the value of 3 times the third quarters of each boxplot for visualization.

Fig 6. More genes were identified by T-GEN as trait-associated in tissues most enriched for SNP signals.
In tissues with the highest heritability enrichment and also other tissues, the numbers of identified trait-associated genes were compared across all five methods. Each barplot shows the mean value of the numbers of identified trait-associated genes across 208 traits in the LD Hub.

Fig 7. Regional Manhanttan plot for SNPs near the identified associated-gene *COG4*.
The listed SNP (rs7196032) is one of eQTLs identified by T-GEN in the imputation model of *COG4*.

Fig 8. Regional Manhattan plot around *TMEM135*.
The listed SNP (rs541458) is one of the identifed eQTLs by T-GEN in the imputation model of *TMEM135*. Among all eQTLs of *TMEM135* identified by T-GEN, this SNP also has the strongest GWAS signal in the published AD GWAS study.

Table 1. Association result for 15 significant loci for AD

chr	start	end	gene	ensembl id	identified tissue	matched loci	p values
chr6	47445524	47594999	CD2AP	ENSG00000198087.7	Esophagus_Mucosa	p12.3	5.70e-08
chr19	44978644	44979403	AC069278.4 [a]	ENSG00000267242.1	Brain_Cortex Heart_Left_Ventricle Breast_Mammary_Tissue	q13.2	1.65e-07 2.13e-09 3.17e-09
chr19	44507099	44518078	ZNF230	ENSG00000159882.8	Heart_Atrial_Appendage Artery_Aorta	q13.2	3.13e-10 5.47e-43
chr19	45855466	45874176	ERCC2	ENSG00000104884.10	Brain_Anterior_cingulate_cortex_BA24	q13.2	1.90e-54
chr19	45682002	45682089	BLOC1S3	ENSG00000189114.6	Ovary Esophagus_Muscularis Stomach	q13.2	8.97e-26 6.35e-16 7.29e-08
chr7	143955699	143956815	OR2A7	ENSG00000243896.3	Esophagus_Muscularis	q35	1.49e-10
chr19	45754515	45808541	MARK4 [a]	ENSG00000007047.10	Brain_Frontal_Cortex_BA9 Brain_Hippocampus Small_Intestine_Terminal_Ileum Brain_Anterior_cingulate_cortex_BA24	q13.2	6.1e-76 3.35e-22 1.74e-08 9.01e-12
chr19	46268042	46272484	SIX5	ENSG00000177045.6	Brain_Frontal_Cortex_BA9 Small_Intestine_Terminal_Ileum	q13.2	6.03e-13 4.53e-15
chr19	45666185	45681495	TRAPPC6A	ENSG00000007255.6	Esophagus_Mucosa Brain_Caudate_basal_ganglia Small_Intestine_Terminal_Ileum Muscle_Skeletal	q13.2	1.19e-44 4.91e-15 1.96e-08 1.55e-08
chr19	46000488	46005768	PPM1N [a]	ENSG00000213889.6	Brain_Caudate_basal_ganglia Liver Breast_Mammary_Tissue	q13.2	1.47e-31 4.41e-24 1.28e-26
chr19	44754317	44772366	ZNF233 [a]	ENSG00000159915.8	Lung Adipose_Visceral_Omentum Adipose_Subcutaneous	q13.2	1.4e-31 5.07e-30 2.04e-32

					Stomach		9.03e-12
chr19	44711699	44733969	ZNF227 [a]	ENSG00000131115.11	Brain_Caudate_basal_ganglia Lung Spleen Artery_Aorta	q13.2	2.91e-30 2.41e-08 3.87e-28 1.42e-28
chr19	45147097	45164590	PVR [a]	ENSG00000073008.10	Spleen Pancreas Esophagus_Muscularis Brain_Anterior_cingulate_cortex_BA24 Breast_Mammary_Tissue Muscle_Skeletal Adipose_Subcutaneous Colon_Sigmoid	q13.2	5.15e-25 1.33e-28 1.76e-10 2.5e-08 6.29e-44 1.49e-13 5.07e-09 3.47e-09
chr19	45453300	45457264	CTB-129P6.11 [a]	ENSG00000267114.1	Spleen Esophagus_Muscularis	q13.2	2.4e-66 1.58e-60
chr19	46386865	46389376	IRF2BP1 [a]	ENSG00000170604.3	Brain_Cortex Muscle_Skeletal	q13.2	1.31e-14 5.45e-84
chr6	47444978	47445308	RP11-385F7.1	ENSG00000270761.1	Esophagus_Mucosa	p12.3	1.27e-07
chr19	45135499	45222031	CTB-171A8.1 [a]	ENSG00000266903.1	Esophagus_Mucosa Heart_Atrial_Appendage Brain_Caudate_basal_ganglia Brain_Frontal_Cortex_BA9 Lung Esophagus_Gastroesophageal_Junction Spleen Pancreas Brain_Cortex Ovary Esophagus_Muscularis Colon_Transverse	q13.2	8.16e-11 5.43e-10 4.5e-10 1.08e-10 5.12e-10 4.39e-10 3.34e-10 1.04e-09 4.04e-10 3.79e-10 3.87e-10 3.89e-10 4.19e-10

					Heart_Left_Ventricle Whole_Blood Adipose_Visceral_Omentum Artery_Aorta Brain_Anterior_cingulate_cortex_BA24 Liver Breast_Mammary_Tissue Adipose_Subcutaneous Stomach		5.65e-09 4.84e-10 3.42e-10 3.69e-10 4.05e-10 4.54e-10 4.22e-11 5.03e-11
chr8	26605666	26724790	ADRA1A [a]	ENSG00000120907.13	Liver	p21.2	1.85e-08
chr19	45202420	45213986	CEACAM16 [a]	ENSG00000213892.6	Breast_Mammary_Tissue	q13.2	2.19e-11
chr19	44598491	44612479	ZNF224 [a]	ENSG00000267680.1	Brain_Frontal_Cortex_BA9 Lung	q13.2	6.09e-11 6.52e-28
chr8	27168998	27316903	PTK2B [a]	ENSG00000120899.13	Whole_Blood Adipose_Visceral_Omentum	p21.2	6.31e-08 5.83e-16
chr19	44331443	44353307	ZNF283 [a]	ENSG00000167637.12	Breast_Mammary_Tissue	q13.2	2.36e-31
chr19	44376514	44388203	ZNF404	ENSG00000176222.7	Brain_Caudate_basal_ganglia Whole_Blood	q13.2	8.67e-09 5.94e-10
chr19	46093021	46103011	GPR4 [a]	ENSG00000177464.4	Brain_Frontal_Cortex_BA9 Lung Brain_Cortex	q13.2	7.41e-10 1.03e-14 2.35e-36
chr19	45409010	45412650	APOE [a]	ENSG00000130203.5	Brain_Caudate_basal_ganglia Muscle_Skeletal	q13.2	5.6e-15 1.5e-28
chr19	46333421	46334366	AC092301.3 [a]	ENSG00000269148.1	Esophagus_Muscularis	q13.2	4.25e-08
chr19	45450459	45450568	APOC2 [a]	ENSG00000234906.4	Brain_Caudate_basal_ganglia Lung Esophagus_Gastroesophageal_Junction	q13.2	4.34e-08 6.08e-17 2.25e-08 4.18e-12

					Spleen Ovary Heart_Left_Ventricle Colon_Sigmoid		1.63e-26 5.97e-12 4.28e-08
chr19	45393825	45406946	TOMM40 [a]	ENSG00000130204.8	Brain_Frontal_Cortex_BA9 Brain_Hippocampus Brain_Cortex Colon_Transverse Whole_Blood Artery_Aorta	q13.2	1.23e-21 1.85e-07 2.51e-18 4.57e-08 4.58e-18 1.92e-32
chr19	45170226	45187631	CEACAM19 [a]	ENSG00000186567.8	Pancreas Ovary Esophagus_Muscularis Small_Intestine_Terminal_Ileum Liver	q13.2	1.26e-10 6.4e-09 3.91e-11 4.84e-11 4.45e-10
chr7	142334162	142334913	TRBV20-1	ENSG00000211747.3	Spleen Whole_Blood	q34	2e-08 1.26e-09
chr19	45659042	45663408	NKPD1	ENSG00000179846.7	Cells_EBV-transformed_lymphocytes	q13.2	8.94e-23
chr19	45715878	45737469	EXOC3L2 [a]	ENSG00000130201.3	Adipose_Visceral_Omentum	q13.2	1.84e-12
chr7	143140545	143141502	TAS2R60	ENSG00000185899.1	Whole_Blood	q35	2.39e-09
chr19	46268042	46273064	AC074212.5	ENSG00000259605.3	Whole_Blood Small_Intestine_Terminal_Ileum	q13.2	4.42e-10 3.29e-09
chr19	44270684	44285409	KCNN4	ENSG00000104783.7	Colon_Sigmoid	q13.2	5.27e-26
chr19	45595049	45650543	PPP1R37	ENSG00000104866.6	Heart_Atrial_Appendage Brain_Frontal_Cortex_BA9 Pancreas	q13.2	1.55e-10 2.58e-20 2.41e-25
chr19	45430060	45434643	APOC1P1 [a]	ENSG00000214855.5	Brain_Caudate_basal_ganglia Liver	q13.2	2.03e-49 1.72e-82
chr1	5934187	5938361	OSBP	ENSG00000110004	Heart_Left_Ventricle	q12.1	1.56e-07

1	0	7		8.7			
chr1	207669491	207813992	CR1 [a]	ENSG00000203710.6	Brain_Frontal_Cortex_BA9	q32.2	1.81e-14
chr19	4466922 5	4468253 4	ZNF226 [a]	ENSG00000167380.12	Esophagus_Mucosa Brain_Frontal_Cortex_BA9 Whole_Blood Brain_Anterior_cingulate_cortex_BA24	q13.2	2.01e-28 7.33e-19 6.64e-21 6.73e-09
chr8	2822286 4	2826021 8	ZNF395	ENSG00000186918.9	Brain_Frontal_Cortex_BA9	p21.1	3.10e-09
chr19	782754	785080	AC006273.5	ENSG00000267530.2	Adipose_Visceral_Omentum Stomach	p13.3	1.62e-08 6.68e-09
chr7	143963793	143966381	CTAGE8	ENSG00000244693.1	Liver	q35	1.14e-10
chr19	4473288 1	4480919 9	ZNF235 [a]	ENSG00000159917.10	Esophagus_Gastroesophageal_Junction Small_Intestine_Terminal_Ileum Brain_Anterior_cingulate_cortex_BA24	q13.2	3.8e-09 1.16e-09 1.25e-41
chr19	4461633 3	4463702 7	ZNF225 [a]	ENSG00000256294.3	Pancreas Breast_Mammary_Tissue	q13.2	4.53e-76 1.91e-16
chr11	4763886 6	4766417 5	MTCH2	ENSG00000109919.5	Esophagus_Muscularis	p11.2	6.77e-08
chr11	6010230 3	6010844 1	MS4A6E	ENSG00000166926.4	Lung	q12.2	1.14e-09
chr19	4504104 4	4512411 9	CEACAM22P [a]	ENSG00000230666.1	Adipose_Subcutaneous	q13.2	1.15e-08
chr19	4534943 1	4539248 5	PVRL2 [a]	ENSG00000130202.5	Pancreas Esophagus_Muscularis Whole_Blood Small_Intestine_Terminal_Ileum Artery_Aorta Liver	q13.2	3.79e-15 6.83e-13 1.82e-24 2.76e-11 3.62e-08 3.52e-34

					Muscle_Skeletal		2.56e-13
chr19	44472013	44502477	ZNF155	ENSG00000204920.6	Ovary	q13.2	1.85e-16
chr19	44555519	44567719	ZNF223 [a]	ENSG00000178386.8	Cells_EBV-transformed_lymphocytes Brain_Hippocampus Brain_Anterior_cingulate_cortex_BA24	q13.2	4.74e-11 2.45e-10 1.49e-10
chr19	45882891	45908339	PPP1R13L [a]	ENSG00000104881.10	Brain_Caudate_basal_ganglia Spleen	q13.2	3.45e-15 2.45e-40
chr19	44979853	45004576	ZNF180 [a]	ENSG00000167384.6	Brain_Frontal_Cortex_BA9 Adipose_Visceral_Omentum Brain_Anterior_cingulate_cortex_BA24	q13.2	1.17e-10 1.29e-20 9.92e-08
chr19	45988546	46000319	RTN2	ENSG00000125744.7	Brain_Cortex Esophagus_Muscularis	q13.2	7.22e-09 1.53e-12
chr19	45504687	45541452	RELB	ENSG00000104856.9	Pancreas	q13.2	1.35e-108
chr19	46271053	46275762	AC074212.6	ENSG00000267395.1	Whole_Blood	q13.2	2.89e-50
chr19	46171501	46186982	GIPR [a]	ENSG00000010310.4	Brain_Caudate_basal_ganglia Ovary Heart_Left_Ventricle	q13.2	3.33e-19 1.34e-35 3.17e-87
chr19	46272974	46282617	DMPK	ENSG00000104936.13	Brain_Frontal_Cortex_BA9	q13.2	8.80e-26
chr19	45542297	45574214	CLASRP [a]	ENSG00000104859.10	Heart_Atrial_Appendage Brain_Caudate_basal_ganglia Whole_Blood Adipose_Subcutaneous	q13.2	2.45e-37 1.12e-25 1.38e-21 1.13e-07
chr19	44847405	44871377	ZNF112 [a]	ENSG00000062370.12	Breast_Mammary_Tissue	q13.2	6.23e-23
chr19	45971252	45978437	FOSB [a]	ENSG00000125740.9	Adipose_Subcutaneous	q13.2	5.59e-18

chr19	45251092	45263301	BCL3 [a]	ENSG00000069399.8	Lung Colon_Transverse Small_Intestine_Terminal_Ileum Brain_Anterior_cingulate_cortex_BA24	q13.2	7.94e-22 4e-13 1.41e-22 2.45e-68
chr19	44886458	44898703	ZNF285 [a]	ENSG00000267508.1	Heart_Left_Ventricle	q13.2	3.38e-10
chr19	46110251	46148887	EML2	ENSG00000125746.10	Spleen	q13.2	1.25e-10
chr19	1597170	1602425	UQCR11 [a]	ENSG00000127540.7	Brain_Cortex	p13.3	1.70e-09
chr19	44416780	44439430	ZNF45	ENSG00000124459.7	Lung	q13.2	5.86e-49
chr19	45574757	45579846	ZNF296 [a]	ENSG00000170684.4	Esophagus_Mucosa Brain_Cortex Colon_Transverse	q13.2	1.57e-11 2.21e-09 1.11e-49
chr19	45385283	45394133	CTB-129P6.4	ENSG00000267282.1	Ovary	q13.2	7.13e-24
chr19	44576296	44576444	ZNF284 [a]	ENSG00000186026.6	Brain_Cortex Artery_Aorta	q13.2	7.98e-21 1.61e-25
chr2	128644129	128644759	RP11-395A13.2	ENSG00000272667.1	Brain_Anterior_cingulate_cortex_BA24	q14.3	2.29e-09
chr19	45281125	45303891	CBLC	ENSG00000142273.6	Lung Colon_Transverse Whole_Blood Adipose_Subcutaneous	q13.2	7.96e-09 7.23e-29 7.61e-16 2.68e-22
chr7	142180514	142181016	TRBV6-5	ENSG00000211721.2	Breast_Mammary_Tissue	q34	3.40e-11
chr19	44971857	44977444	ZNF285B	ENSG00000176761.6	Lung Brain_Cortex Colon_Sigmoid	q13.2	4.08e-26 1.32e-16 6.61e-09
chr14	93170151	93215047	LGMN	ENSG00000100600.10	Artery_Aorta	q32.12	9.45e-08

chr16	70514470	70557468	COG4	ENSG00000103051.14	Heart_Atrial_Appendage	q22.3	1.35e-07
chr19	46009836	46030241	VASP [a]	ENSG00000125753.9	Brain_Hippocampus Colon_Transverse Adipose_Visceral_Omentum Breast_Mammary_Tissue	q13.2	3.32e-08 2.52e-72 7.31e-46 1.73e-21
chr19	44501047	44506988	RP11-15A1.7 [a]	ENSG00000266921.1	Heart_Atrial_Appendage Cells_EBV-transformed_lymphocytes Brain_Hippocampus Esophagus_Muscularis Brain_Anterior_cingulate_cortex_BA24	q13.2	1.35e-07 2.93e-14 4.4e-12 3.12e-10 7.5e-28
chr7	143104905	143220542	EPHA1-AS1	ENSG00000229153.1	Lung Spleen Whole_Blood Adipose_Visceral_Omentum Adipose_Subcutaneous	q35	3.12e-08 1.21e-10 3.93e-10 1.12e-09 9.81e-10
chr19	44530177	44537265	ZNF222 [a]	ENSG00000159885.9	Brain_Hippocampus Small_Intestine_Terminal_Ileum Breast_Mammary_Tissue	q13.2	1.91e-07 1.99e-19 1.03e-08
chr19	44455374	44471861	ZNF221 [a]	ENSG00000159905.10	Esophagus_Mucosa Whole_Blood	q13.2	1.24e-34 6.33e-12
chr19	46196670	46207247	QPCTL [a]	ENSG00000011478.7	Cells_EBV-transformed_lymphocytes Brain_Cortex Ovary Heart_Left_Ventricle Adipose_Visceral_Omentum Brain_Anterior_cingulate_cortex_BA24	q13.2	7.72e-19 8.95e-13 1.47e-07 4.44e-08 2.4e-19 2.38e-31
chr7	143548467	143599291	TCAF1	ENSG00000198420.5	Stomach	q35	3.55e-09

chr19	45457847	45496599	CLPTM1	ENSG00000104853.11	Ovary	q13.2	9.86e-19
chr19	45583164	45591957	GEMIN7 [a]	ENSG00000142252.6	Brain_Hippocampus	q13.2	7.47e-16
chr19	46292516	46296060	DMWD	ENSG00000185800.7	Esophagus_Mucosa Brain_Cortex	q13.2	6.55e-11 1.13e-23
chr19	44794233	44795064	CTC-512J12.7	ENSG00000270679.1	Cells_EBV-transformed_lymphocytes	q13.2	1.57e-09
chr11	86748885	87034800	TMEM135 [a]	ENSG00000166575.12	Lung	q14.2	1.79e-08
chr19	46030684	46088100	OPA3	ENSG00000125741.4	Esophagus_Muscularis Small_Intestine_Terminal_Ileum	q13.2	5.7e-10 2.26e-12
chr19	45312327	45324671	BCAM [a]	ENSG00000187244.6	Brain_Frontal_Cortex_BA9 Brain_Cortex Artery_Aorta	q13.2	1.07e-25 2.58e-21 3.14e-47
chr2	128383571	128384423	RP11-286H15.1	ENSG00000272789.1	Brain_Frontal_Cortex_BA9	q14.3	3.41e-08
chr19	45741889	45748628	AC006126.4 [a]	ENSG00000267045.1	Colon_Transverse Heart_Left_Ventricle Artery_Aorta	q13.2	4.36e-15 4.05e-09 6.61e-19
chr11	86013252	86056969	C11orf73	ENSG00000149196.11	Pancreas	q14.2	1.78e-09
chr19	45579767	45593649	CTB-179K24.3	ENSG00000267348.2	Esophagus_Mucosa Small_Intestine_Terminal_Ileum	q13.2	6.16e-13 4.44e-10
chr19	45916691	45982086	ERCC1	ENSG00000012061.11	Colon_Sigmoid	q13.2	6.22e-08
chr2	127805602	127864931	BIN1	ENSG00000136717.10	Esophagus_Mucosa Cells_EBV-transformed_lymphocytes Pancreas Esophagus_Muscularis Heart_Left_Ventricle	q14.3	1.48e-07 4.67e-18 5.66e-08 9.42e-08 5.71e-08
chr1	4619071	4619564	SNRPD2 [a]	ENSG0000012574	Adipose_Visceral_Omentum	q13.2	1.72e-15

9	1	7		3.6	Small_Intestine_Terminal_Ileum		5.93e-12

The table shows AD-associated genes and corresponding loci identified by T-GEN
[a] Genes replicated on the GWAX data by T-GEN

Supplementary Figures and Table

S1 Fig. More SNPs with functional potentials were idenfied by T-GEN under the ChromHMM-25 model.

The figure compares the percentages of SNPs in gene expression imputation models having any of active ChromHMM25 annotated states for three different methods (elnt, vb and vb.logit) using gene expression and genotype data from GTEx in 26 tissues. Each dotted line represents the mean values of percengates of identified SNPs having active ChromHMM25 annotated states for each method. The detail of each method is as descriped in the caption of **Fig 2**.

S2 Fig. Correlation of imputed expression for genes located in the same loci

Correlations of imputed gene expression levels in whole blood for genes located in the same loci are compared across five methods and also for the total gene expression level from GTEx (indicated by "true" in the figure). Values located above red diamonds indicate the mean levels for boxplots.

S3 Fig. Comparison of R^2 between observed and imputed gene expression levels across five methods.

Using 5-fold cross validation in the GTEx data, R^2 between imputed and observed expression levels was calculated in elastic net (elnt) and vb.logit models. Dotted lines indicate the mean values of R^2 for each method across all tissues. The mean values of vb.annot and T-GEN models are very close to each other, which lead to the overlaid green and blue dotted lines.

S4 Fig. R^2 between observed and imputed gene expression across five methods for the CommonMind data.

Squared Pearson correlation (R^2) between observed gene expression in the CommonMind dataset and predicted gene expression based on genotypes in the CommonMind dataset are shown for all 5 methods. T-GEN has the lowest values of R^2 among all five methods. Y axis is truncated at the value of 2 times the third quarters of each boxplot for visualization.

S5 Fig. More loci were identified as trait-associated by T-GEN across 208 traits from the LD Hub.

Applied to 208 traits from LD Hub, significantly trait-associated genes were identified in 26 tissues (p-values threshold: 0.05 divided by the number of gene-tissue pairs). Those identified associated-genes were further grouped into pre-defined cytobands. Each boxplot represents the distribution of the number differences between that identified from our tissue-specific analysis and that identified from the four other methods. Y axis is truncated at the value of 3 times the third quarters of each boxplot for visualization.

S1 Table. Active states in ChromHMM-15 and ChromHMM-25 model
The table shows 11/21 active states in ChromHMM-15 and 25 models

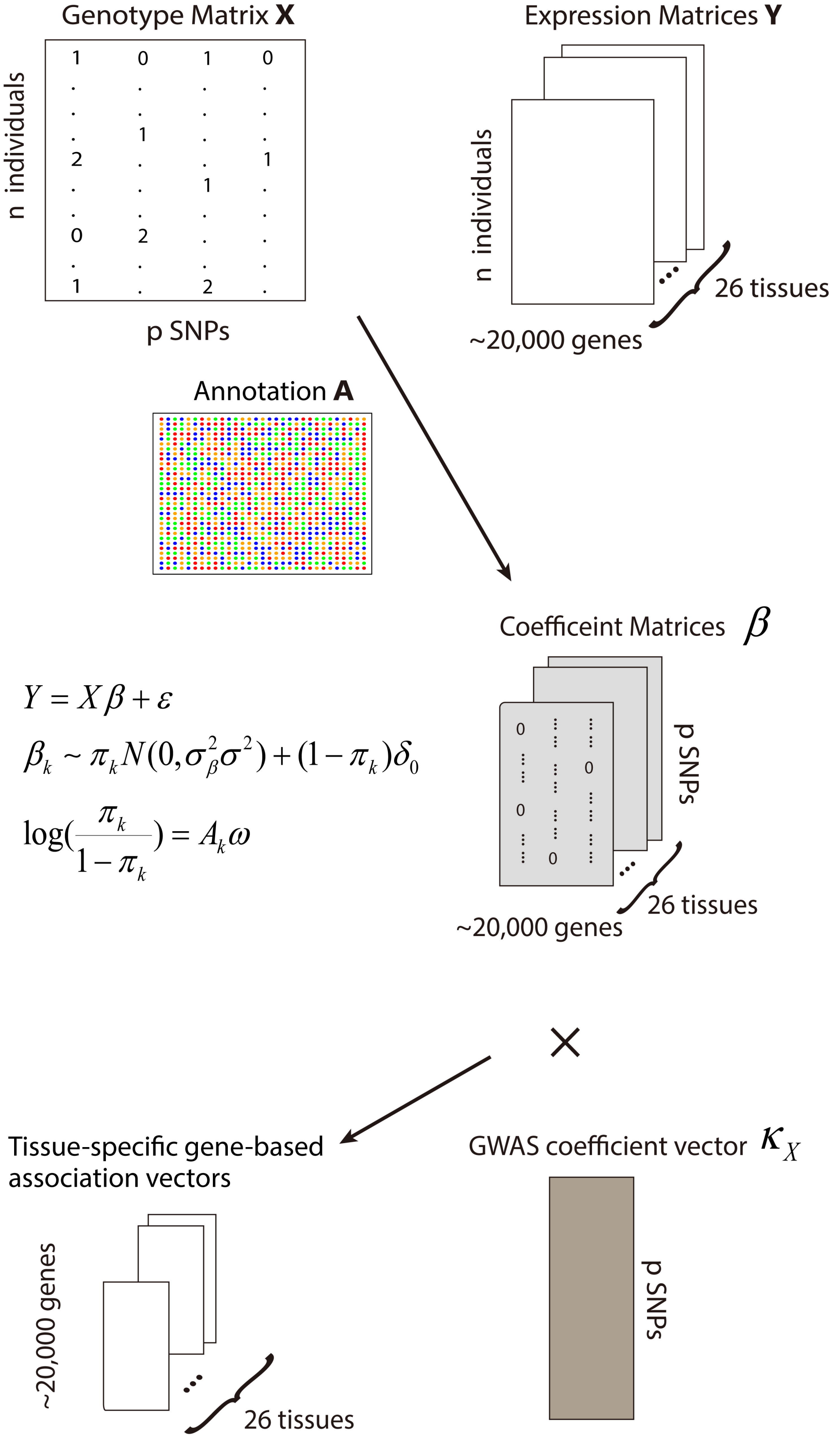
S2 Table 93 lipid traits in LD Hub
The table shows the 93 lipid traits included in the LD Hub database

S3 Table. Numbers of singifcant genes in the most relevent tissue across traits in LDhub
The table shows the number of significant gens in the tissues with most-enriched heritability across different traits

S4 Table. eQTLs for novel genes identified in AD by T-GEN
The table shows eQTLs identified for novel AD-associated genes identified by T-GEN

S5 Table. Significant associated-genes identified in GWAX data by T-GEN
The table shows significantly-associated genes in the external GWAS data.

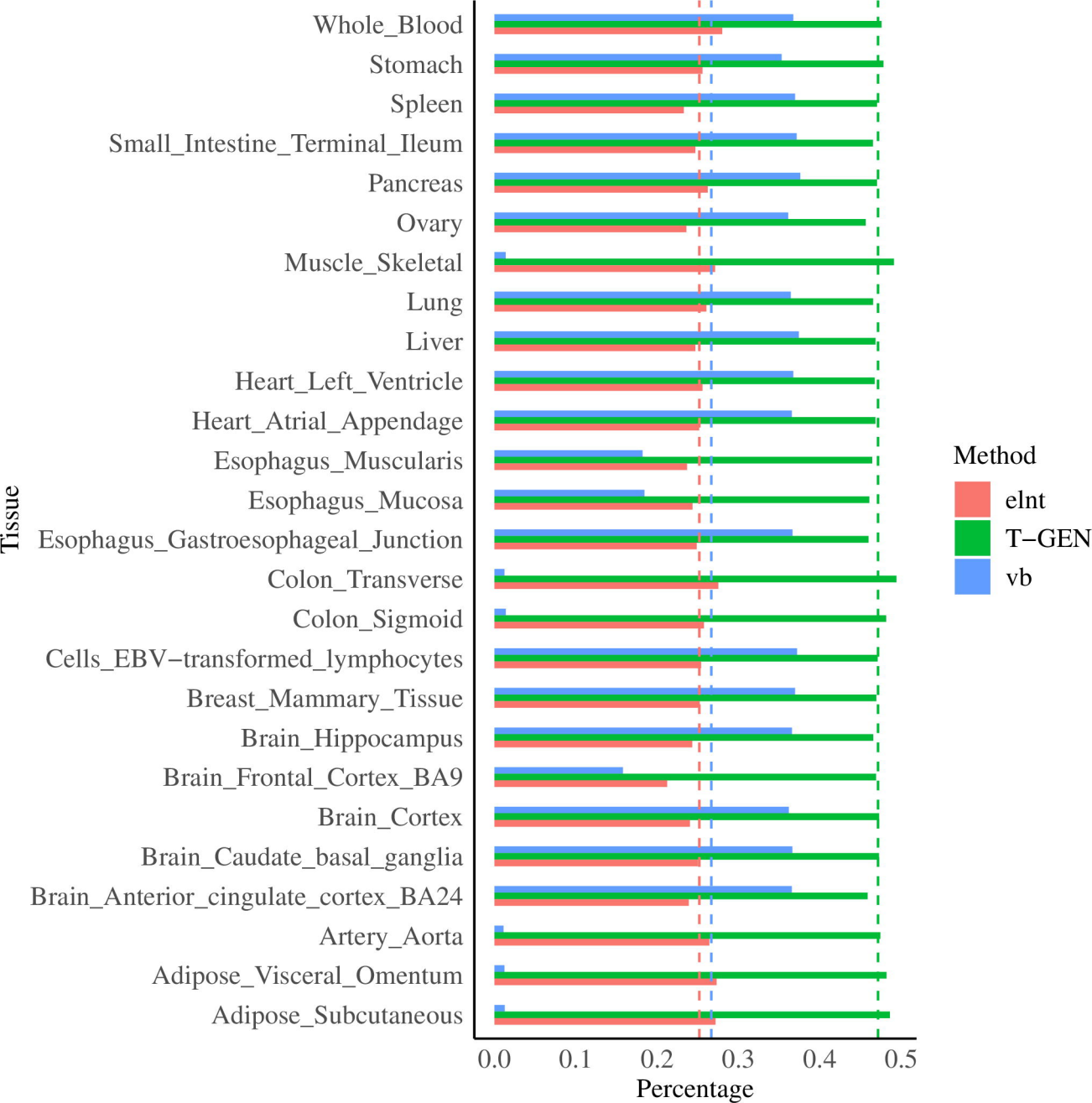
S6 Table. Numbers of Roadmap annotation categories for each cell types and corresponding tissues
The table shows the numbers of Roadmap annotation categories (like DNA methylation, H3K4me3 and H3K9ac) in Roadmap cell types and corresponding 26 Roadmap tissues

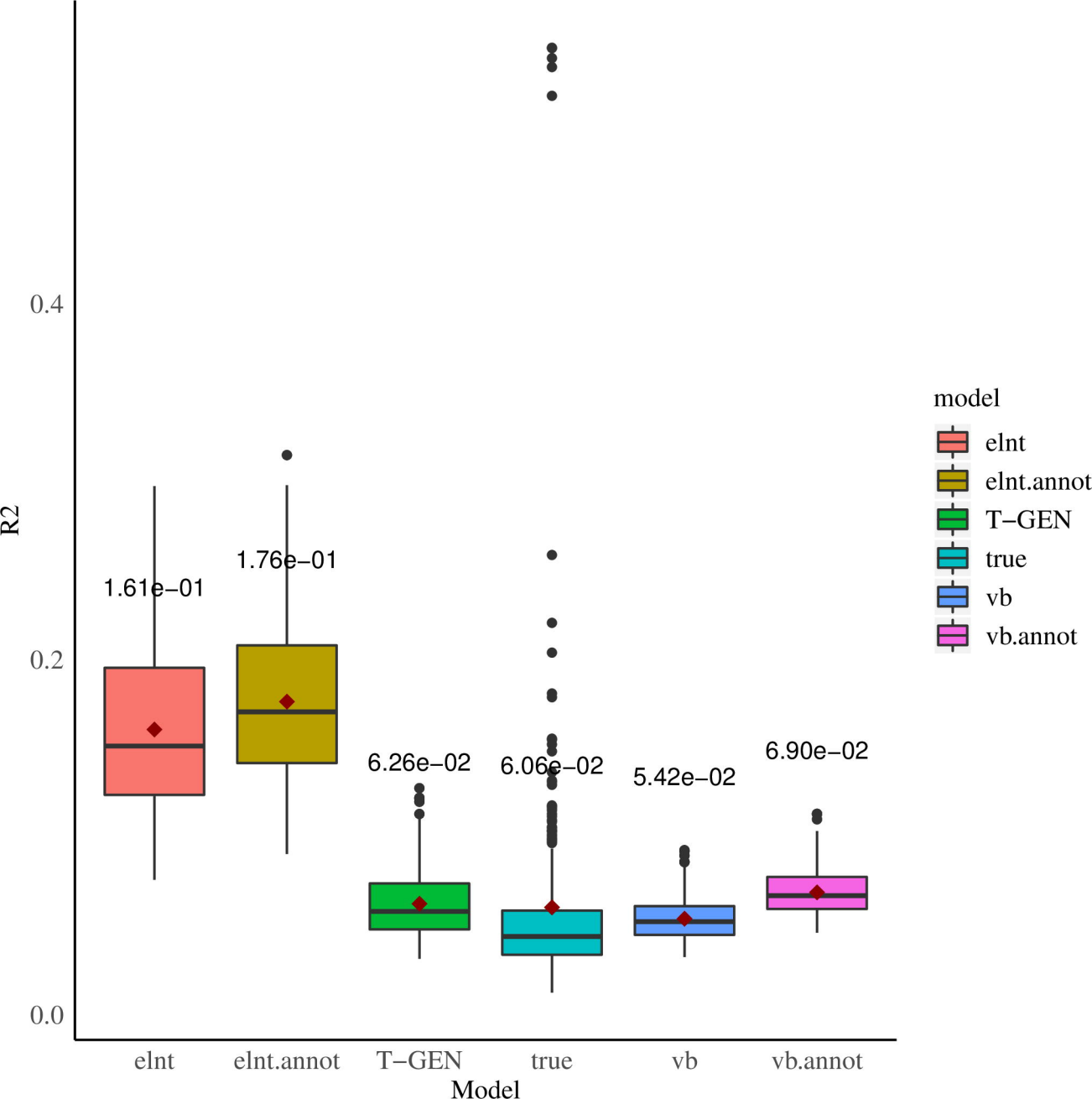


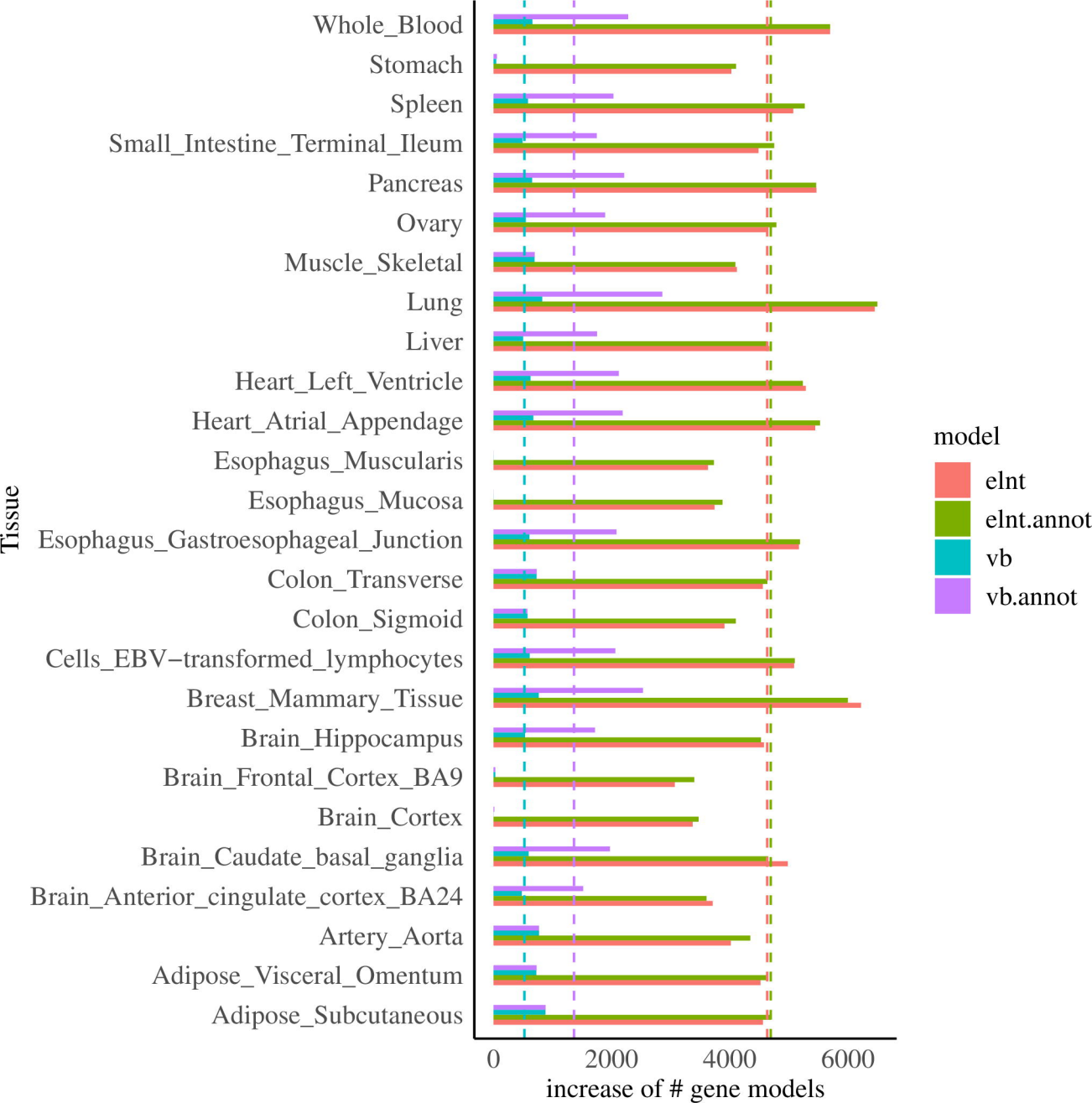
$$Y = X\beta + \varepsilon$$

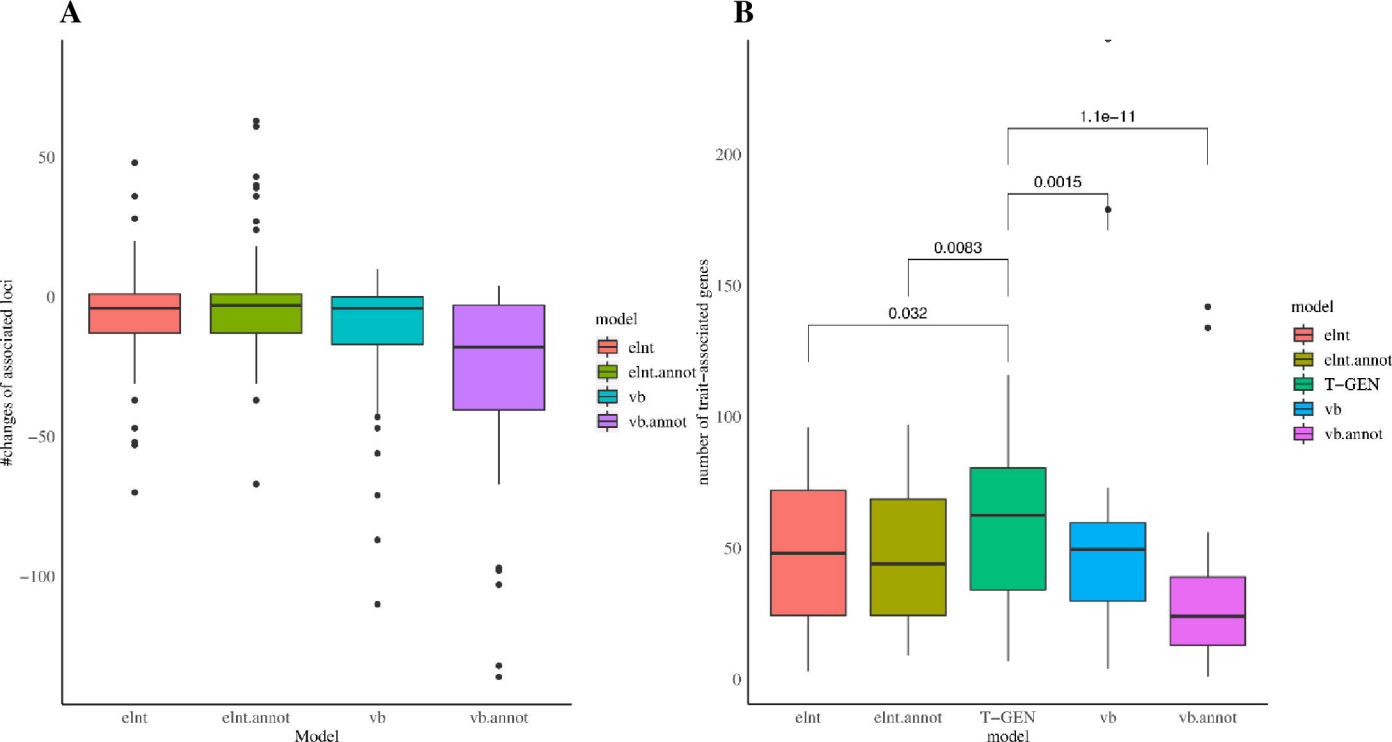
$$\beta_k \sim \pi_k N(0, \sigma_\beta^2 \sigma^2) + (1 - \pi_k)\delta_0$$

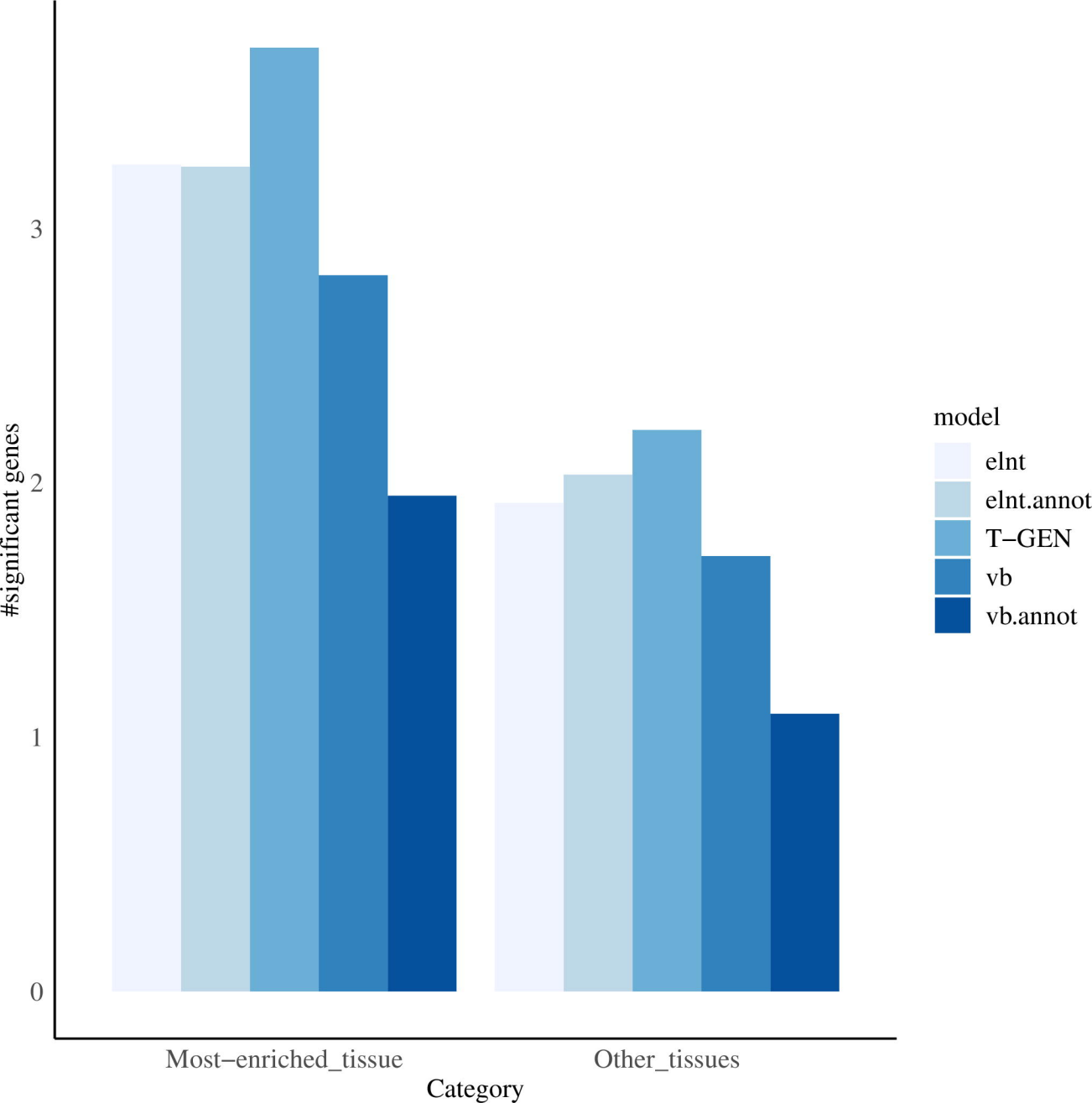
$$\log\left(\frac{\pi_k}{1 - \pi_k}\right) = A_k \omega$$

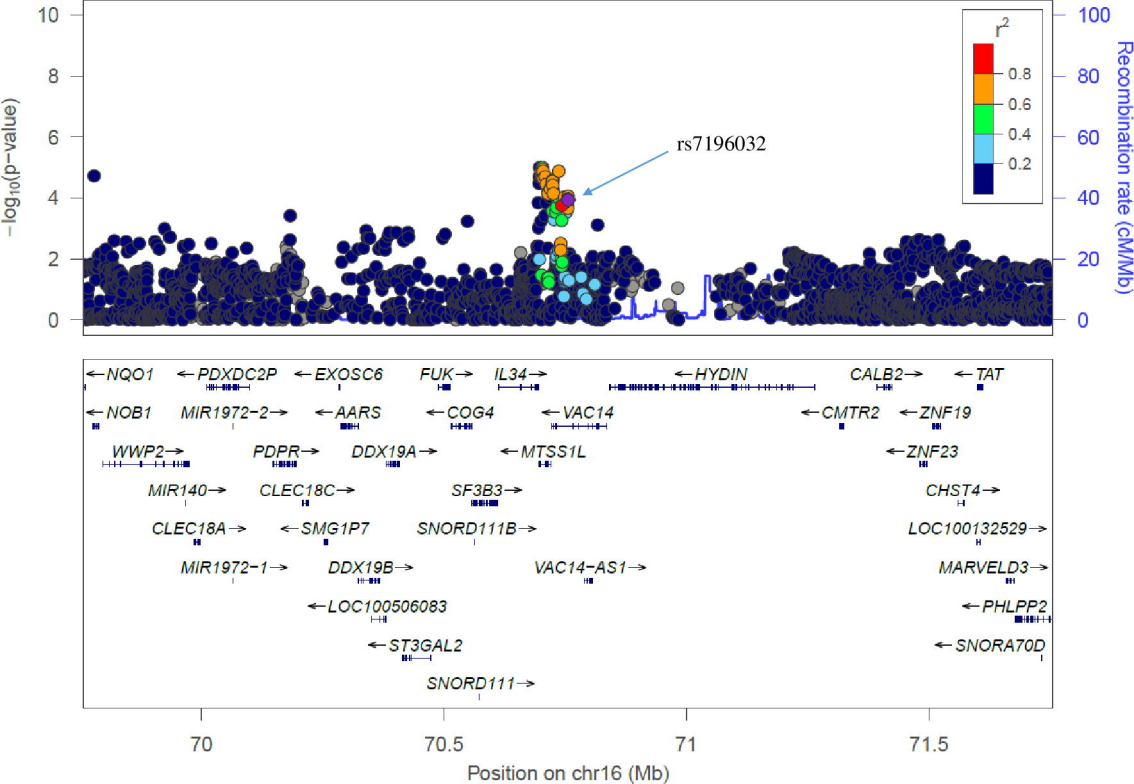

Whole_Blood
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Pancreas
Ovary
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV−transformed_lymphocytes
Breast_Mammary_Tissue
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Artery_Aorta
Adipose_Visceral_Omentum
Adipose_Subcutaneous
Tissue
0.0
0.1
0.2
0.3
0.4
0.5
Percentage
Method
elnt
T−GEN
vb

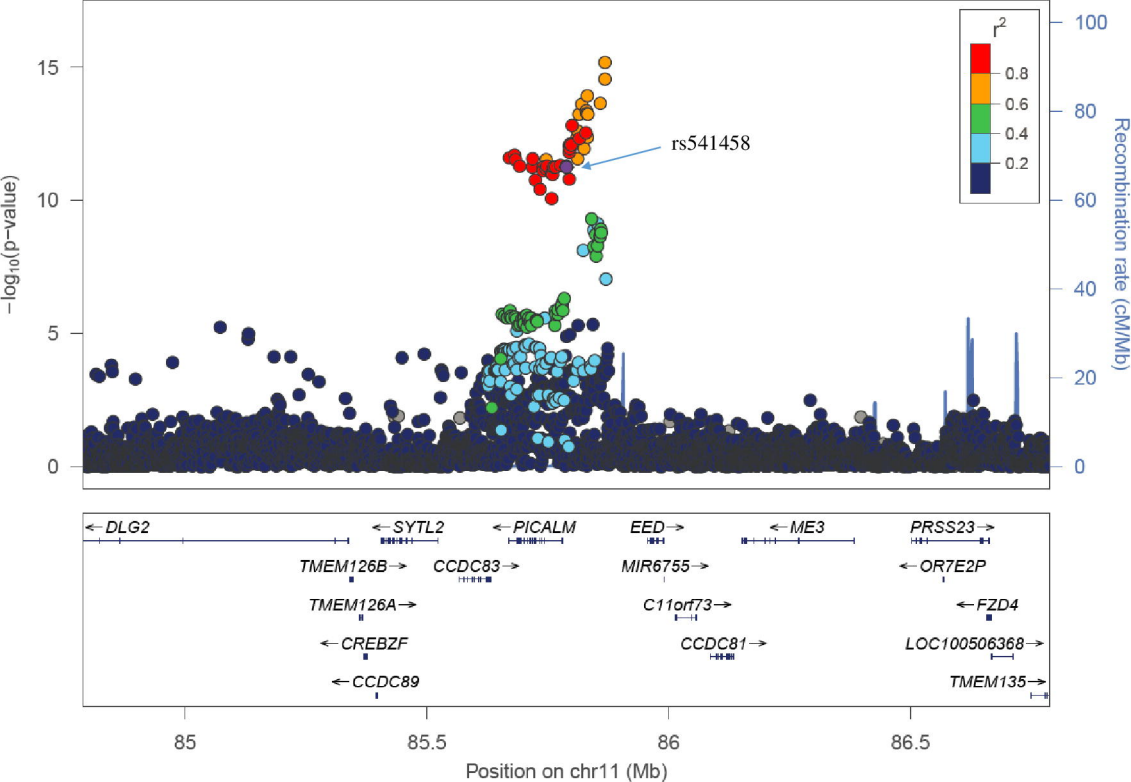

R2
0.0
0.2
0.4
1.61e−01
1.76e−01
6.26e−02
6.06e−02
5.42e−02
6.90e−02
elnt
elnt.annot
T−GEN
true
vb
vb.annot
Model
model
elnt
elnt.annot
T−GEN
true
vb
vb.annot

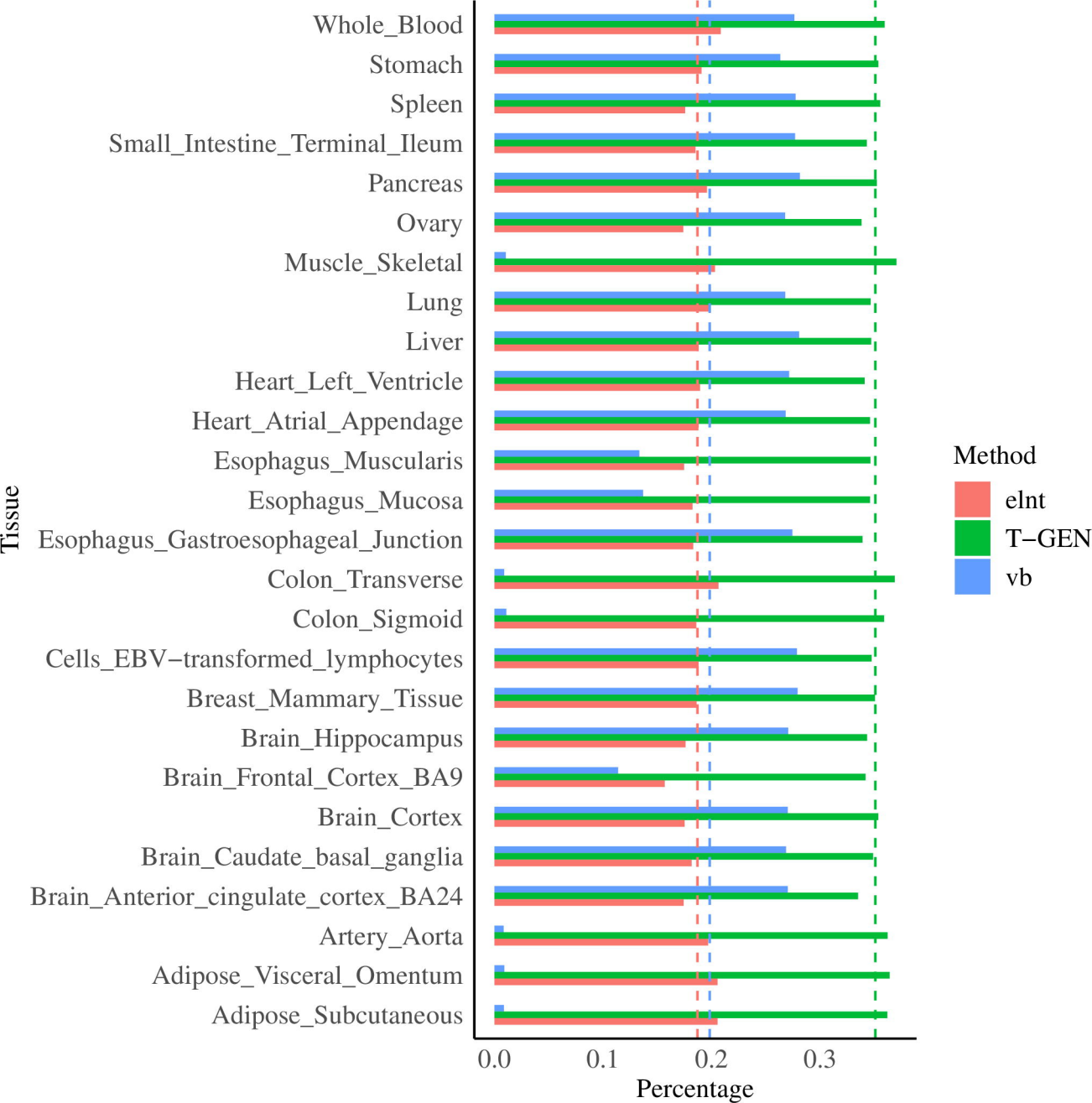

Tissue
Whole_Blood
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Pancreas
Ovary
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV−transformed_lymphocytes
Breast_Mammary_Tissue
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Artery_Aorta
Adipose_Visceral_Omentum
Adipose_Subcutaneous
0
2000
4000
6000
increase of # gene models
model
elnt
elnt.annot
vb
vb.annot

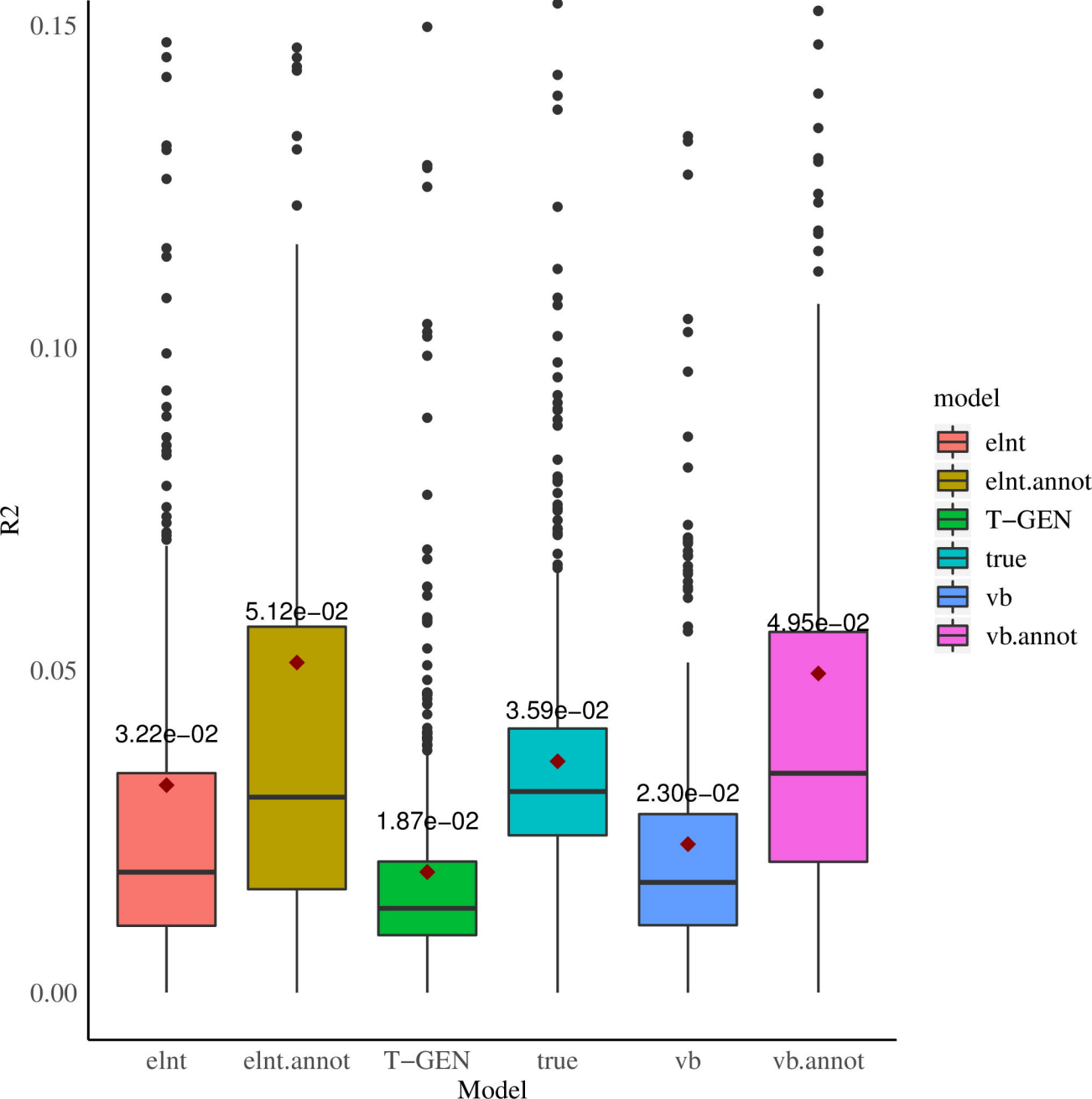

A
50
0
−50
−100
#changes of associated loci
elnt
elnt.annot
vb
vb.annot
Model
model
elnt
elnt.annot
vb
vb.annot
B
200
150
100
50
0
number of trait-associated genes
1.1e−11
0.0015
0.0083
0.032
elnt
elnt.annot
T−GEN
vb
vb.annot
model
model
elnt
elnt.annot
T−GEN
vb
vb.annot

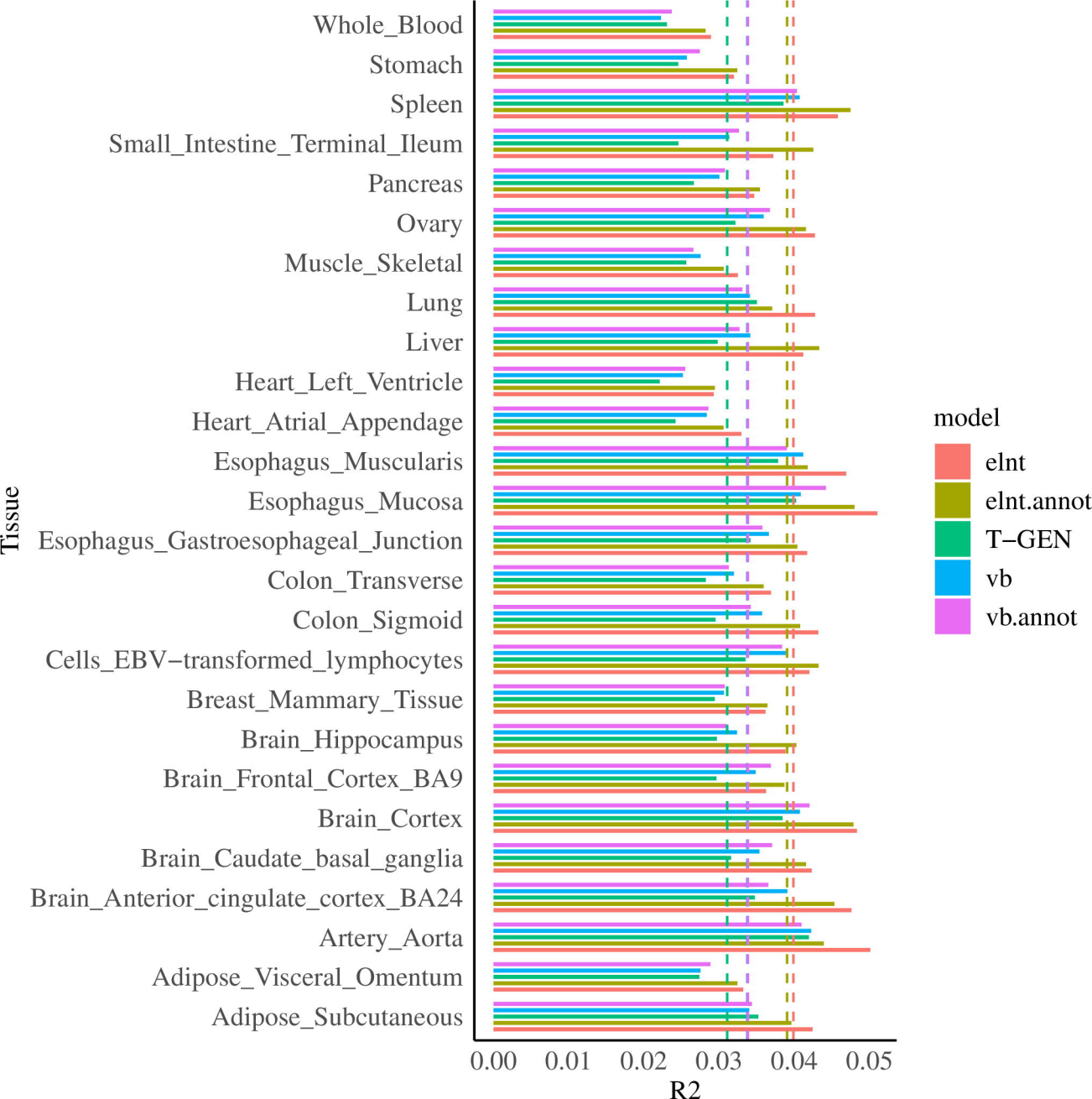

#significant genes
0
1
2
3
Most−enriched_tissue
Other_tissues
Category
model
elnt
elnt.annot
T−GEN
vb
vb.annot

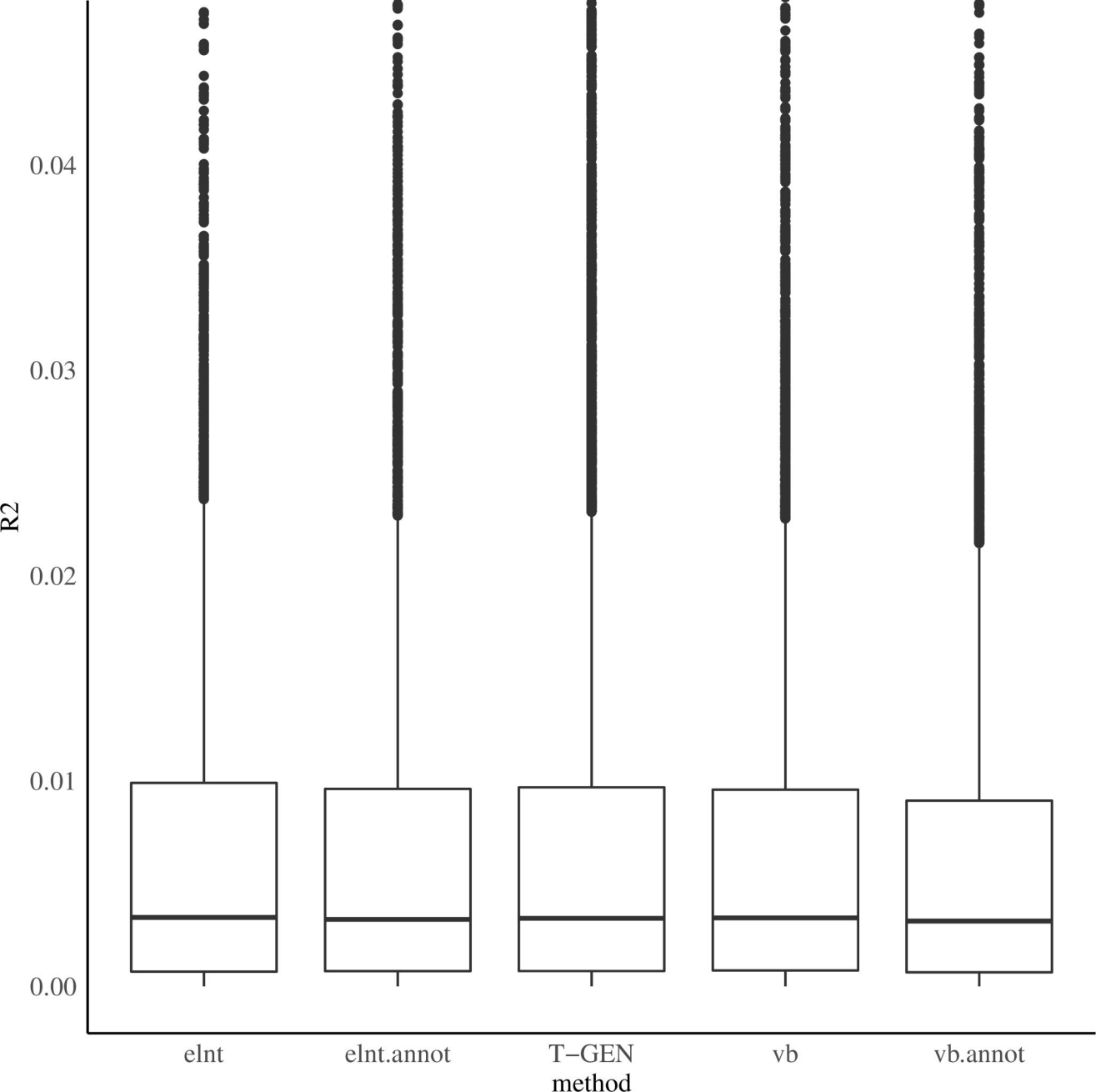

rs7196032
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
NQO1
PDXDC2P
EXOSC6
FUK
IL34
HYDIN
CALB2
TAT
NOB1
MIR1972−2
AARS
COG4
VAC14
CMTR2
ZNF19
WWP2
PDPR
DDX19A
MTSS1L
ZNF23
MIR140
CLEC18C
SF3B3
CHST4
CLEC18A
SMG1P7
SNORD111B
LOC100132529
MIR1972−1
DDX19B
VAC14−AS1
MARVELD3
LOC100506083
PHLPP2
ST3GAL2
SNORA70D
SNORD111
Position on chr16 (Mb)

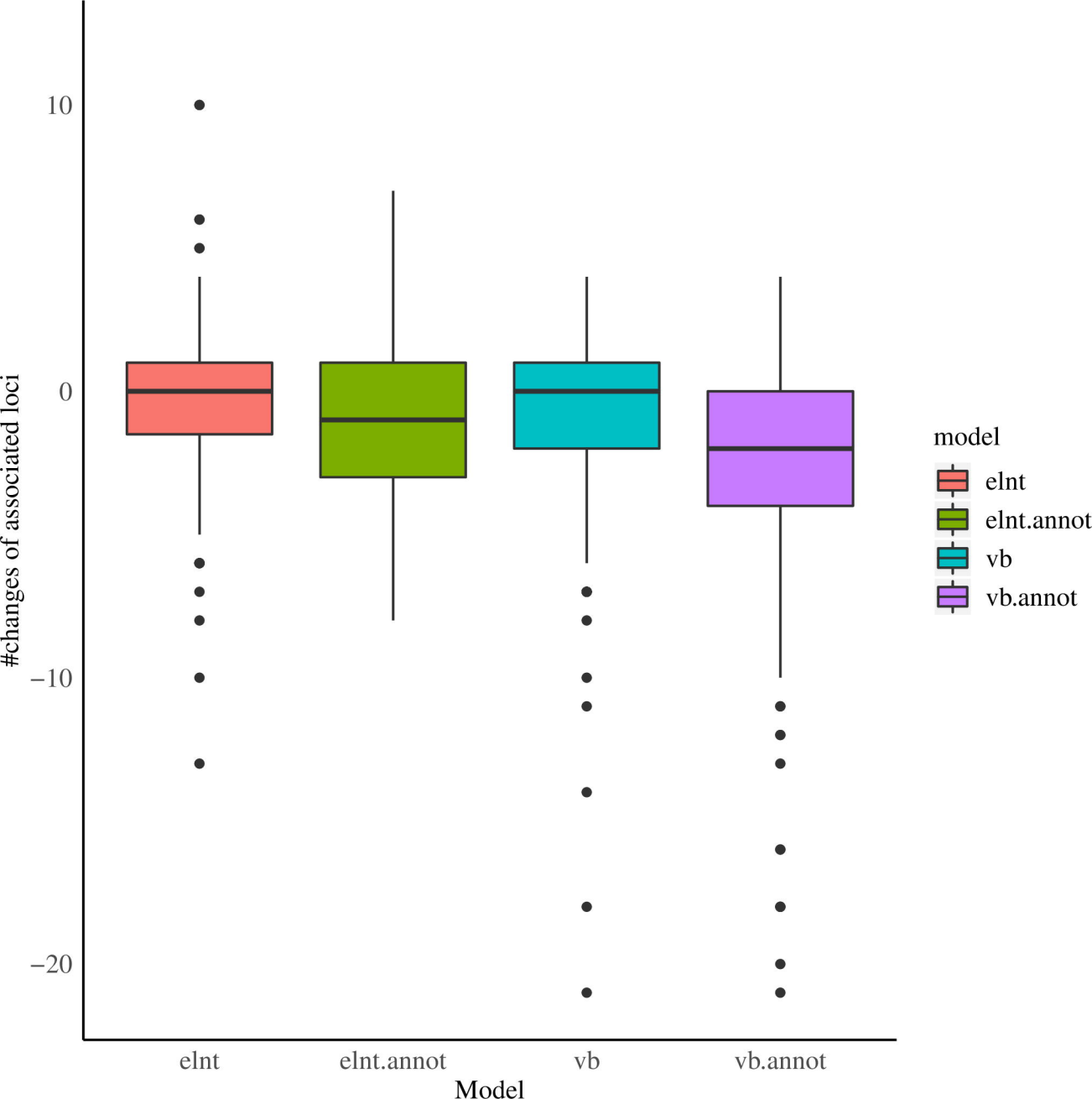

rs541458
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
DLG2
SYTL2
PICALM
EED
ME3
PRSS23
TMEM126B
CCDC83
MIR6755
OR7E2P
TMEM126A
C11orf73
FZD4
CREBZF
CCDC81
LOC100506368
CCDC89
TMEM135
Position on chr11 (Mb)


Whole_Blood
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Pancreas
Ovary
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV−transformed_lymphocytes
Breast_Mammary_Tissue
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Artery_Aorta
Adipose_Visceral_Omentum
Adipose_Subcutaneous
Tissue
0.0
0.1
0.2
0.3
Percentage
Method
elnt
T−GEN
vb


0.15
0.10
0.05
0.00
R2
3.22e−02
5.12e−02
1.87e−02
3.59e−02
2.30e−02
4.95e−02
elnt
elnt.annot
T−GEN
true
vb
vb.annot
Model
model
elnt
elnt.annot
T−GEN
true
vb
vb.annot


Whole_Blood
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Pancreas
Ovary
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV−transformed_lymphocytes
Breast_Mammary_Tissue
Brain_Hippocampus
Brain_Frontal_Cortex_BA9
Brain_Cortex
Brain_Caudate_basal_ganglia
Brain_Anterior_cingulate_cortex_BA24
Artery_Aorta
Adipose_Visceral_Omentum
Adipose_Subcutaneous
Tissue
0.00
0.01
0.02
0.03
0.04
0.05
R2
model
elnt
elnt.annot
T−GEN
vb
vb.annot


0.04
0.03
0.02
0.01
0.00
R2
elnt
elnt.annot
T−GEN
vb
vb.annot
method


10
0
−10
−20
#changes of associated loci
elnt
elnt.annot
vb
vb.annot
Model
model
elnt
elnt.annot
vb
vb.annot