Leveraging functional annotation to identify genes associated with complex diseases

Wei Liu[1], Mo Li[2], Wenfeng Zhang[2], Geyu Zhou[1], Xing Wu[3], Jiawei Wang[1], Qiongshi Lu[4,5,6], Hongyu Zhao[1,2,7*]

[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 Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA 06510

[4] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI, USA 53706

[5] Department of Statistics, University of Wisconsin-Madison, WI, USA 53706

[6] Center for Demography of Health and Aging, University of Wisconsin-Madison, WI, USA 53706

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

* 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 transcriptome-wide association study (TWAS) methods have been proposed using gene expression as a mediating trait linking genetic variations and diseases. These methods first predict 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 may not be functional in the corresponding tissue, due to linkage disequilibrium (LD) and the correlation of gene expression between tissues. Here, we introduce a new method called T-GEN (**T**ranscriptome-mediated identification of disease-associated **G**ens with **E**pigenetic a**N**notation) to identify disease-associated genes leveraging epigenetic information. Through prioritizing SNPs with tissue-specific epigenetic annotation, T-GEN can better identify SNPs that are both statistically predictive and biologically functional. We found that a significantly higher percentage (an increase of 18.7% to 47.2%) of eQTLs identified by T-GEN are inferred to be functional by ChromHMM and more are deleterious based on their Combined Annotation Dependent Depletion (CADD) scores. Applying T-GEN to 207 complex traits, we were able to identify more trait-associated genes (ranging from 7.7 % to 102%) than those from existing methods. Among the identified genes associated with these traits, T-GEN can better identify genes with high (>0.99) pLI scores compared to other methods. When T-GEN was applied to late-onset Alzheimer's disease, we identified 96 genes located at 15 loci, including two novel loci not implicated in previous GWAS. We further replicated 50 genes in an independent GWAS, including one of the two novel loci.

Author summary

TWAS-like methods have been widely applied to understand disease etiology using eQTL data and GWAS results. However, it is still challenging to discriminate the true disease-associated genes from those in strong LD with true genes, which is largely due to the misidentification of eQTLs. Here we introduce a novel statistical method named T-GEN to identify disease-associated genes considering epigenetic information. Compared to current TWAS methods, T-GEN can not only identify eQTLs with higher CADD scores and function potentials in gene-expression imputation models, but also identify more disease-associated genes across 207 traits and more genes with high (>0.99) pLI scores. Applying T-GEN in late-onset Alzheimer's disease identified 96 genes at 15 loci with two novel loci. Among 96 identified genes, 50 genes were further replicated in an independent GWAS.

Introduction

Genome-Wide Association Studies (GWAS) have been very successful in identifying single nucleotide polymorphisms (SNPs) associated with human diseases [1]. 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 to link genes with identified SNPs and provide insights for downstream analysis [2–6]. PrediXcan [7] and similar methods [8–12] have been developed for utilizing transcriptomic data, such as those from GTEx [13], to interpret identified GWAS non-coding signals and to identify additional disease associated genes. These methods first impute (i.e. predict) gene expression levels from SNP genotypes and then identify disease-associated genes by associating phenotypes with predicted expression levels. At the SNP level, SNPs used in the gene expression imputation models are selected through statistical correlation between these SNPs' genotypes and gene expression levels. Since SNPs in the same LD block are correlated, it is hard to differentiate regulatory SNPs from others statistically, which may lead to incorrect identifications of genes with regulatory SNPs that are in strong LD with true trait/disease genes.

To more accurately identify regulatory eQTLs that play functional roles, we assume that SNPs with active epigenetic annotations are more likely to regulate tissue-specific gene expression. As for available tissue-specific epigenetic data, we consider epigenetic marks that are known hallmarks for DNA regions with important functions, such as H3K4me1 signals that are often associated with enhancers [14]. We note that these epigenetic marks have been used to infer regulatory regions and prioritize eQTLs in some published studies [15–18]. Reported GWAS hits are enriched in regions with active epigenetic signals [19] and they can help fine-map true GWAS hits with functional impacts [20]. Based on these previous findings, we use epigenetic signals to select SNPs among all candidate cis-SNPs when modeling the relationship between gene expression levels and SNP genotypes. We have developed a new method, called T-GEN (**T**ranscriptome-mediated identification of disease-associated **G**ens with **E**pigenetic a**N**notation), that leverages tissue-specific epigenetic information to identify

disease-associated genes. By prioritizing the SNPs likely having regulatory roles (through the use of epigenetic marks) when building gene expression imputation models, T-GEN identified SNPs with higher deleterious effects and higher function potential, Further application of T-GEN identified more trait-associated genes in 207 traits, compared to other gene expression imputation models. More specifically, in late-onset Alzheimer's disease (AD), T-GEN identified the largest number of genes with novel loci, indicating the importance of cholesterol transportation, neuron activity and mitochondrial dysfunction in late-onset AD.

.

Results

Method overview

Similar to previous methods [8,21,22], we use two linear models to study gene-level genetic effects on traits mediated by gene expression regulation. Firstly, individual-level genotype and gene expression data are used to build gene expression imputation models for each tissue. One novel feature of our method is the integration of epigenetic data from the Roadmap Epigenomics Project [23] to prioritize regulatory SNPs in gene expression imputation. As a result, SNPs located in regions with active epigenetic marks are more likely to be selected, consistent with the enrichment of epigenetic marks (e.g. H3K4me3 and DNase-I hypersensitivity) in regulatory DNA regions for gene expression [24]. After obtaining tissue-specific gene expression imputation models, 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**.

By utilizing a spike-and-slab prior, SNPs regulating gene expression levels were selected considering their epigenetic signals through the following model:

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

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

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

where Y is the vector of gene expression values in a given tissue for a gene, X is the genotype matrix for the candidate cis-SNPs for the gene, β is the tissue-specific effect vector of genotypes on expression level, and ϵ denotes random noise. For a cis-SNP k, its effect β_k follows a mixture prior of a normal distribution and a point mass around 0. The probability π_k of the SNP being an eQTL of its nearby gene is linked to its epigenetic annotation matrix A_k via a logit model, with ω being the annotation coefficient vector. We use a variational Bayesian method [25] to estimate the coefficient vector β. More details are shown in the Methods section.

For comparison, we also consider four other gene expression imputation methods, including elastic net (elnt) [7,26] used in both PrediXcan and FUSION, linear models with spike-and-slab priors solved by variational Bayes (vb) [27,28], and both elastic net

and linear models with spike-and-slab priors applied only to SNPs having active epigenetic signals (elnt.annot and vb.annot).

In general, we can describe the relationship between the imputed gene expression level and selected SNPs in the form of $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ denotes 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, which was also used for gene-trait testing in many TWAS methods[8,21,22]:

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

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

T-GEN prioritizes biologically functional SNPs in gene expression imputation

To study whether our method can better prioritize functional SNPs in gene expression imputation models, we evaluated the functional states of the identified eQTLs using the ChromHMM-annotated SNP status [29]. T-GEN identified higher percentage eQTLs (47%) having one of the 11 active states (out of 15 states) (**S1 Table**), with an 87% increase over elnt models (25%) and 77% increase over vb (27%) models (**Fig 2**). Overall, these results demonstrate that our method can better identify SNPs with functional potential to regulate gene expression.

We note that T-GEN utilizes the epigenetic information in SNP selections, and the same information is also used in ChromHMM models. Therefore, we expect to select more SNPs annotated as functionally active in ChromHMM models. To further evaluate the

functional potential of the SNPs selected by T-GEN, we considered the CADD scores of the identified SNPs across all five methods. T-GEN-identified eQTLs have higher CADD scores (3.36 on average, representing a 0.9% increase compared to elnt and vb, $p<2e-16$, Wilcoxon rank sum test) and a higher percentage (0.34%) of functionally deleterious SNPs (larger than 20, representing a 2.3% and 2.9% increase to elnt and vb respectively). The net CADD score improvement is not substantial, which may be partially explained by the purifying selections undergone by cis-eQTLs [30,31] and their consequent low deleteriousness. These results indicate the statistically significant higher functional potential of identified eQTLs by T-GEN than those identified by other models except direct SNP filtering in elnt.annot and vb.annot, which is expected.

More genes can be effectively imputed by T-GEN

The number of genes that can assessed in transcriptome-wide association analysis is affected by the quality of expression imputation. For all five imputation methods, gene expression imputation models were filtered based on the significance of imputation models with an FDR cutoff of 0.05. Therefore, significant trait-gene associations can only be detected for genes with significant imputation models.

With the same FDR cutoff, T-GEN had an increase in the range of 2.8% to 55.3% (3,807 in whole blood) compared with the elastic net methods in 25 of the 26 tissues. There was a smaller increase or even slight decrease in some tissues with a smaller sample size like brain cortex (2.8%, 269, n=136) and brain frontal cortex BA9 (-0.55%, -50, n=118). 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-eQTL. Compared to elastic net methods (elnt and elnt.annot), variational Bayes (both vb and T-GEN) methods can better impute genes, which can be partly attributed to its improved variable selection performance [32].

More genes and higher percentage of functionally conserved genes identified in 207 traits

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 207 traits (**S2 table**) from LD Hub. After Bonferroni adjustment, significant trait-associated genes were identified in each tissue. T-GEN showed a 25% (9.6 genes, compared with elnt.annot) to 175% (30.4 genes, compared with vb.annot) increase in the average number of significant trait-associated genes across 207 traits (**Fig 4**). After aggregating identified genes into pre-defined cytogenetic bands, we observed a similar pattern in the numbers of identified trait-associated loci across 207 traits. T-GEN identified the largest number of associated loci (14.3 loci on average) across all five methods (**S1 Fig**), showing an 11% (1 locus, compared with elnt) to 86% increase (5 loci, compared with vb.annot).

To investigate the functional potential of trait-associated genes identified by each method, we further compared the enrichment pattern of associated genes having pLI scores larger than 0.99 for each method. Among all identified genes, some genes are identified as trait-associated in multiple traits. Grouping identified genes into three categories based on the number of associated traits, we found that T-GEN identified the largest percentage of trait-associated genes having higher pLI (>0.99) in gene groups with fewer than 5 associated traits (**Fig 5**). This may indicate that T-GEN is more likely to identify conserved genes specific to 1-2 diseases compared to other methods ($p=0.009$ to elnt, $p=0.002$ to vb). Across all categories, T-GEN also showed the strongest enrichment signal (fold change: 1.15, binomial test p value: 0.038) compared to all the other four methods, which didn't show significant enrichment pattern (**S3 table**). Although pLI score is an indirect measure of gene importance in human traits and more associated with fitness, it does provide an important way to prioritize genes whose heterozygous mutations can be phenotypically harmful [33,34].

To assess the ability of identifying associated genes in the tissue most relevant to traits, which is defined as the tissue with the highest heritability enrichment estimated by LDSC [35], we compared the numbers of identified genes in the most relevant tissues

(**S4 Table**). Using LDSC and annotation from GenoSkyline-Plus [36], we identified the tissue with the highest heritability enrichment for each trait. We compared the number of significantly associated genes identified in heritability-enriched tissues across five methods (**Fig 6**). In the most-enriched tissues, T-GEN identified the largest numbers of significantly associated genes. When comparing the ratios of the number of genes identified in the most-enriched tissue and those in the other tissues across 151 lipid-associated traits (**S2 Fig**), T-GEN showed a significant increase compared to the vb (Wilcoxon rank sum test, p=0.048) and vb.annot models (p=3.5e-3).These results suggest that T-GEN can better identify disease genes in trait-relevant tissues.

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

T-GEN identifies novel genes for late-onset AD

To further investigate the performance of our method in identifying trait-associated genes in detail, we analyzed the biological functions of genes associated with AD (N=74,046) that were only identified by our method. Considering the total number of tissue-gene pairs (258,039), 96 significantly associated genes at 15 loci (**S5 Table, Fig 7**) were identified by T-GEN, with five loci identified in the brain tissues (caudate basal ganglia, anterior cingulate cortex BA24, hippocampus, cortex, and frontal cortex BA9) and four loci identified in the whole blood. Thirteen out of the 15 loci have been implicated in AD GWAS[37]. In comparison (**S3 Fig**), 79 genes at 10 loci were identified by elastic net models, 81 genes at eight loci by elnt.annot models, 61 genes at three loci by vb.annot models, and 81 genes at 11 loci by vb method. Not only the number of associated genes by T-GEN is the largest, the heritability enrichment in signal-contributing eQTLs for associated genes by T-GEN is also the highest (**S6 table**).

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* [37]. *SLC24A4* and *RIN3* are two potential genes contributing to the GWAS signal of this locus in previous GWAS studies, whose functions are not yet understood. *LGMN* encodes protein AEP that cleaves inhibitor 2 of PP2A and can trigger tau pathology in AD brain [38], which makes *LGMN* a potential signal gene at this locus. Another locus is 16p22.3, where *COG4* was identified (p=1.35e-7). Two of the identified eQTLs (**S7 Table**) of COG4 are potential GWAS hits (rs7196032, p=1.1e-4 and rs7192890: p=3.5e-3) (**S4 Fig**). *COG4* encodes a protein involved in the structure and function of the Golgi apparatus. Recent research has shown that defects in the Golgi complex are associated with AD and Parkinson's disease by affecting the functions of Rab-GTPase and SNAREs [39]. The third 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 [37] with *CD2AP* reported as the locus signal genes affecting amyloid precursor protein (APP) metabolism and production of amyloid-beta [40]. While 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* located on 16q22.3 [41], which was also the gene only identified by our method and encodes a protein that may serve as an antioxidant strategy target for AD [42].

Although other loci identified by our method have been implicated in GWAS, including *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), we identified additional, potentially functionally-impactful genes at these loci. At 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 a potential gene in late-onset AD via gene-gene interaction analysis [43]. At 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 AD etiology [44].

Two out of 15 loci identified by our method are novel compared with published GWAS loci [37,45–47]. 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 and is discussed above. The associated gene identified at the other novel locus, *TMEM135* (p=1.80e-8) is located 1MB (**Fig 8**) downstream of the identified GWAS locus (the *PICALM* locus), and two of the identified eQTLs of *TMEM135* by our method were located at 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 [48]. Increased cholesterol levels in brain tissue are associated with accumulated AβPP [46] suggesting the potential role of *LXR* and *TMEM135* in the etiology of AD.

To validate the T-GEN results, we tried to replicate our findings in an independent GWAS for AD using inferred phenotypes based on family history (GWAX, N=114,564) [49]. Among the 96 genes identified by our method using the GWAS of clinical AD, 50

(52%) genes were successfully replicated in this independent GWAS study ($p < 5.2e-4$) including *TMEM135* in the novel loci. For the other four methods, only 9.9% (8/81, vb) to 41% (33/81, elnt.annot) genes were replicated in the GWAX data. Pathway enrichment analysis via Enrichr [50] identified apoptosis-related network (p=2e-3), statin pathway (p=9e-3), and ApoE-related inflammation and atherosclerosis (p=0.04). Statin pathway was again implicated (p=0.067) in the GWAX data despite a lack of statistical significance. The high replication rate of the results in an external study further confirmed the power of T-GEN.

Discussion

In this paper, we have introduced a new method called T-GEN, which leverages epigenetic signals to improve gene expression imputation and identify trait-associated genes. Different from previous methods, T-GEN uses data from GTEx and Roadmap Epigenomics Project to prioritize SNPs with active epigenetic annotations for gene expression imputation. We found that T-GEN models were more likely to include SNPs with functional potential and more genes can be effectively imputed compared to other methods. Applied to more than 200 traits, T-GEN identified more genes/loci associated with these traits. T-GEN is especially more likely to identify genes with potential function importance (high pLI scores).

When applied to AD GWAS, T-GEN identified the largest number of associated genes (96 genes in 15 loci) compared with four other methods. We found novel association signals at previously identified loci including *LGMN* in the *SLC24A4/RIN3* locus, *ADRA1A* at the *PTK2B* locus and *OSBP* at 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*), cholesterol transportation (indicated by *TMEM135* and *OSBP*), and neuron activity (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, T-GEN showed improved performance in prioritizing functional SNPs and identifying disease-associated genes. Limited by the lack of individual-level epigenetic data and small sample sizes, T-GEN also can be further improved in different aspects, such as the low accuracy of predicting gene expression levels, which may be further improved, for example by simultaneously using non-parameteric Bayesian model [51] and considering epigenetic annotation. Nevertheless, the relationship between imputation accuracy and the power of TWAS-like methods, and how GWAS signals and

eQTL signals contribute to final results are rather complex and worth further investigation (supplementary discussion). The eQTLs used for expression imputation in these methods also need further experimental validations. Utilizing epigenetic annotations in the reference panel from Roadmap also limits the potential 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-eQTL [52–54]. Integrating trans-eQTL into gene expression imputations would help to identify more trait-associated genes and also co-regulatory peripheral genes and core genes [55]. Methods like T-GEN, using gene expression as a mediated trait to study genetic effects and to identify disease genes, can only provide evidence for associations between genes and traits/diseases. The causal relations between identified genes and traits should be further validated by functional analysis in future research. Also, results from TWAS studies are hard to validate in silico and a benchmarking dataset would help in the comparison of different TWAS-like methods. Considering multiple TWAS methods when identifying disease genes may aid in controlling false discoveries.

By using individual-level data with larger sample sizes and matched individual-level epigenetic annotation would further improve power to identify disease-associated genes under our modeling framework. Apart from functional analysis, gene-level fine mapping or Mendelian randomization (MR) would also help in discriminating causal relations from associations between genes and traits.

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 variable selection model to select SNPs:

$$\mathrm{Y} = \mathrm{X}\beta + \epsilon,$$

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

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

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

$$\gamma_{\mathrm{k}} \sim Bernoulli(\pi_k),$$

where Y is the $\mathrm{n} \times 1$ gene expression level from the GTEx [56] database (v6p) (n denotes the sample size in this tissue), X denotes the $\mathrm{n} \times \mathrm{p}$ centered genotype matrix for cis-SNPs for this gene**,** n is the sample size, and p is the number of SNPs. In analysis, Y denotes the gene expression in a specific tissue while X denotes the genotype matrix for SNPs located within 1MB from the upstream or downstream of the gene. We use the $\mathrm{p} \times 1$ vector to represent the effects of those cis-SNPs on the gene expression level while ϵ is the $\mathrm{n} \times 1$ 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 n 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 that the effect size vector comes from a mixture of normal distribution and a point mass at zero. γ_{k} 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. $\gamma_{\mathrm{k}} = 1$), we assume that its effect size follows a normal distribution with 0 mean and variance $\sigma_\beta^2 \sigma^2$.

To more accurately identify eQTL with potential biological functions, we integrate the epigenetic annotation to prioritize 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:

$$\mathrm{logit}(\pi_{\mathrm{k}}) = \mathrm{A}_{\mathrm{k}} \omega,$$

$$\omega \sim \mathrm{N}(0, \eta^{-1} \mathrm{I}),$$

where A_{k} is the epigenetic signal $1 \times \mathrm{m}$ vector for the SNP k and the $\mathrm{m} \times 1$ vector is the epigenetic signal effect vector where m is the number of epigenetic signals considered, such as DNA methylation and histone marker status. We assume that the variance of the coefficients is η^{-1} .

Prior assumptions on hyper parameters

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

$$\sigma^2 \sim IG(a, b),$$

$$\sigma_\beta^2 \sim IG(c, d),$$

$$\eta \sim \mathrm{Gamma}(a_0, b_0),$$

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

Variational Bayes inference

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

$$\mathrm{PPS(k)} = \sum_\theta P(\gamma_k = 1|X, Y, \theta)P(\theta|X, Y)$$

which is basically the weighted average of the posterior probability of this SNP being effective while the weight is the model likelihood, similar to that used in varbvs.

To address the computational challenge in fitting our model to high-dimensional genotype data, we applied the variational Bayes method, which is an alternative of the Markov Chain Monte Carlo (MCMC) method for statistical inference with less computation time and lower computation burden. 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 $\theta = \{\sigma^2, \sigma_\beta^2, \eta\}$, 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:

$$\alpha_k = P(\gamma_k = 1|X, Y, \theta).$$

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:

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

More details about fitting the model can be found in the **supplementary text**.

Model training and evaluation

Gene expression level imputation models for 26 tissues were trained using the RNA-seq and genotype data from the GTEx(v6p) project 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 the 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. More specific details of preprocessing expression data from GTEx were described in our previous publication [22].

We further 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 (R^2) was used. In our method, with the number of epigenetic categories existing in Roadmap varying across different tissues (**S8 Table**), all available epigenetic annotation categories were used for each tissue. The continuous annotation values were used for training models. For elastic net models (elnt), as what previous studies did [7], the parameter α was set to be 0.5 as in PrediXcan and the optimal λ was selected via the function cv.glmnet provided in the 'glmnet' package [57]. For elastic net models with epigenetic signals direct filtering (elnt.annot), we first removed cis-SNPs without positive values for H3K4me1 [58], H3K4me3 [59] or H3K9ac [60] signals, which have been reported as associated with gene expression regulation. After filtering, the remaining cis-SNPs were 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 signed 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 levels in an external brain expression data set of the 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 $\mathrm{Y} = \mathrm{X}\beta + \epsilon$ while the expression-trait relation can be modelled by $\mathrm{T} = \mu + \mathrm{Y}\kappa_{\mathrm{Y}} + \tau$,

where μ is the intercept and τ is the error term in the model. The basic idea of this test is similar to what is used in UTMOST and PrediXcan.

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

$$\mathrm{Z} = \frac{\hat{\kappa}_Y}{se(\hat{\kappa}_Y)}$$

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

$$\widehat{\kappa_Y} = \frac{cov(Y,T)}{var(Y)} = \frac{\hat{\beta}^T cov(X,T)}{var(Y)} = \frac{\hat{\beta}^T var(X)\hat{B}}{var(Y)}$$

where $\hat{\beta}$ is the vector of estimated SNP effects on gene expression levels, $\hat{B}$ denotes the vector of GWAS SNP-level effect sizes for those identified effective SNPs in the gene expression imputation model. Besides, $\mathrm{var(X)}$ is a diagonal matrix whose elements are the genotype variances for each SNP and var(Y) is the variance of the imputed expression levels for this gene in the tissue.

To calculate the test statistic z score, we also need to derive the stand error of the estimated κ_Y:

$$\mathrm{se}(\widehat{\kappa_Y}) = \sqrt{var(\widehat{\kappa_Y})} = \sqrt{\frac{var(\tau)}{var(Y) \times n_{GWAS}}}$$

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

$$\mathrm{R}^2_{T,Y} = \hat{\kappa}_Y^2 \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_T^2}$$

$$\mathrm{R}^2_{T,X} = \hat{\kappa}_X^2 \frac{\hat{\sigma}_X^2}{\hat{\sigma}_T^2}$$

where $\hat{\kappa}_X$ is the estimated SNP-level effect size presented in the GWAS result and $\hat{\sigma}_X^2$ is the estimated SNP variance diagonal matrix. Therefore, the standard error of $\hat{\kappa}_Y$ is $\sqrt{\frac{se(\hat{\kappa}_X)\hat{\sigma}_X^2(1-\mathrm{R}^2_{T,Y})}{(1-\mathrm{R}^2_{T,X})\hat{\sigma}_Y^2}}$.

Then, the Z score is $\mathrm{Z} \approx \hat{\beta}^T \Lambda \hat{Z}_X$, where $\hat{\beta}$ is the SNP coefficient vector in the gene expression imputation model, Λ 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 $\hat{Z}_X$ is the Z 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 207 traits (not based on UK biobank data) were downloaded from the LD Hub website (http://ldsc.broadinstitute.org/gwashare/) by March of 2018, 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 expression.

Acknowledgements

Supported in part by NIH grants R01 GM122078 and P30 AG021342, and NSF grant DMS 1902903.

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. 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.
Q.L. advised on the manuscript.
H.Z. advised on statistical and genetic issues.
All authors read and approved the manuscript.

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 epigenetic 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 get the gene-level association statistics for each disease in each tissue.

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

This figure compares the percentages of SNPs in gene expression imputation models having active ChromHMM15 annotated states across three 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, and the "T-GEN" model was 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. T-GEN can build more gene expression imputation models than other methods across all tissues.

The left panel shows the ratios of gene model numbers (FDR < 0.05) in each method over that in elastic net models. The right panel indicates the difference in the number of genes models between that from each method and that from elastic net. Different colors of each tissue indicate their sample sizes, from upper to lower: [401,501), [301, 401), [201, 301), [101, 201).

Fig 4. More genes were identified as trait-associated by T-GEN across 207 traits from the LD Hub.

Applied to 207 traits from LD Hub, significant 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.

Fig 5. Higher pecentage of significant genes by T-GEN have pLI scores larger than 0.99 for genes identified in leass than 5 traits

Considering the number of traits that each identified gene is associated with, all significant trait-associated genes were groupped into three categories. The bar plot shows the percentage of genes identified by each method having larger pLI scores (>0.99) in each category. Error bars indicate the standard error calculated using bootstraping (120 traits each time, for 20 times).

Fig 6. More genes were identified by T-GEN as trait-associated in tissues most enriched for genetics 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 207 traits in the LD Hub.

Fig 7. Gene-level manhattan plot of T-GEN results in IGAP data

The plot shows the gene-level association with LOAD attained from T-GEN. Several significant genes are indicated in the figure.

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 eQTL of *TMEM135* identified by T-GEN, this SNP also has the strongest GWAS signal in the published AD GWAS study.

Supplementary Figures and Tables

S1 Fig. More loci were identified as trait-associated by T-GEN across 207 traits from the LD Hub.

Applied to 207 traits from the LD Hub, significant 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 those identified from our tissue-specific analysis and those identified from the four other methods. Y axis is truncated at the value of 3 times the third quarters of each boxplot for visualization.

S2 Fig. The ratio of number of genes identified in trait-relavant tissues over that identified in non-associated tissues.

Each boxplot shows the distribution of the ratios across 207 traits in the LD Hub.

S3 Fig. The Venn diagram of genes identified as associated with LOAD in all five methods

Twenty nine genes were identified by all five methods. 62 genes were shared between T-GEN and elastic-net methods, which may indicate the consistency of gene findings in T-GEN and other TWAS methods based on elastic-net models.

S4 Fig. 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*.

S5 Fig. Comparison of R^2 between observed and imputed gene expression levels across five methods.

Using 5-fold cross validation in the GTEx data, $\mathrm{R2}$ between imputed and observed expression levels was calculated in elastic net (elnt) and vb.logit models. Dotted lines indicate the mean values of $\mathrm{R2}$ 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.

S6 Fig. R^2 between observed and imputed gene expression across five methods for the CommonMind data.

Squared values of Pearson correlation (R^2) between observed gene expression in the CommonMind dataset and predicted gene expression based on five different methods. T-GEN showed 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.

S7 Fig. **The percentages of identified trait-associated genes in 207 traits having pLI>0.99 and their relationship with their imputation accuracy**

For each method, all trait-associated gene-tissue models are classified into 20 bins based on their imputation accuracy (R^2). For each bin, the percentage of genes having pLI>0.99 is calculated. The linear equation indicates the linear relationship between the mean R^2 and the percentage and the p value indicate the significance level of the

association.

S8 Fig. The percentage of genes identified as trait-associated in any of 207 traits and its relationship with the corresponding imputation accuracy

For each method, all gene-tissue models (not just trait-associated ones) are classified into 20 bins based on their imputation accuracy (R^2). For each bin, the percentage of gene-tissue models identified as trait-associated (bars) and the mean level of imputation accuracy for trait-associated ones (squares with crosses) are calculated. The linear equation indicates the linear relationship between the mean imputation accuracy and the percentage of trait-associated gene-tissue models. The p value indicates the significance level.

S1 Table. Active states in ChromHMM-15 model

This table shows 11/21 active states in the ChromHMM-15 model

S2 Table. 207 traits from LD Hub

207 traits from the LD Hub considered in our paper and also their corresponding GWAS studies

S3 Table. Enrichment pattern of genes with pLI > 0.99 in identified trait-associated genes

This table shows the enrichment analysis results for all five methods, the p values were obtained using binomial test.

S4 Table. Numbers of singifcant genes in the most relevent tissue across traits in LDhub

This table shows the number of significant gens in the tissues with most-enriched heritability across different traits

S5 Table. T-GEN results in LOAD case study

This table shows T-GEN association results in LOAD, genes with green background in the table are those replicated in GWAX data

S6 Table. Heritability enrichment results in LOAD

eQTLs contributing to identified LOAD-associated genes were identified as an annotation used in LD score regression. Heritability enrichment analysis was further conducted. The table shows the results for all five methods

S7 Table. eQTL for novel genes identified in AD by T-GEN

This table shows eQTLs for novel AD-associated genes identified by T-GEN

S8 Table. Numbers of Roadmap annotation categories for each cell types and corresponding tissues

This table shows the numbers of roadmap annotation categories (like DNA methylation, H3K4me3 and H3K9ac) in Roadmap cell types and corresponding 26 Roadmap tissues

References

1. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577: 179–189. doi:10.1038/s41586-019-1879-7
2. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011;89: 82–93. doi:https://doi.org/10.1016/j.ajhg.2011.05.029
3. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, et al. Insights into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data. Am J Hum Genet. 2010;86: 860–871. doi:https://doi.org/10.1016/j.ajhg.2010.04.014
4. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;99: 1245–1260. doi:https://doi.org/10.1016/j.ajhg.2016.10.003
5. Joehanes R, Zhang X, Huan T, Yao C, Ying S, Nguyen QT, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 2017;18: 16. doi:10.1186/s13059-016-1142-6
6. Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, et al. Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. Am J Hum Genet. 2018;102: 1169–1184. doi:https://doi.org/10.1016/j.ajhg.2018.04.011
7. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47: 1091–1098. doi:10.1038/ng.3367
8. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016/02/08. 2016;48: 245–252. doi:10.1038/ng.3506
9. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLOS Genet. 2017;13: e1006646. Available: https://doi.org/10.1371/journal.pgen.1006646
10. Bhutani K, Sarkar A, Park Y, Kellis M, Schork NJ. Modeling prediction error improves power of transcriptome-wide association studies. bioRxiv. 2017; 108316. doi:10.1101/108316
11. Xu Z, Wu C, Wei P, Pan W. A Powerful Framework for Integrating eQTL and GWAS Summary Data. Genetics. 2017/09/11. 2017;207: 893–902.

doi:10.1534/genetics.117.300270

12. Yang Y, Shi X, Jiao Y, Huang J, Chen M, Zhou X, et al. CoMM-S2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz880

13. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank. 2015;13: 311–319. doi:10.1089/bio.2015.0032

14. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009;459: 108–112. doi:10.1038/nature07829

15. Berger SL. Histone modifications in transcriptional regulation. Curr Opin Genet Dev. 2002;12: 142–148. doi:https://doi.org/10.1016/S0959-437X(02)00279-4

16. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol. 2011;12: R15. doi:10.1186/gb-2011-12-2-r15

17. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Res. 2011;40: 553–568. doi:10.1093/nar/gkr752

18. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13: R53. doi:10.1186/gb-2012-13-9-r53

19. Consortium TEP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57. Available: https://doi.org/10.1038/nature11247

20. Spisák S, Lawrenson K, Fu Y, Csabai I, Cottman RT, Seo J-H, et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med. 2015;21: 1357. Available: https://doi.org/10.1038/nm.3975

21. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9: 1825. doi:10.1038/s41467-018-03621-1

22. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet. 2019;51: 568–576. doi:10.1038/s41588-019-0345-7

23. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Roadmap for regulation. Nature. 2015;518: 314–316. doi:10.1038/518314a
24. Li B, Carey M, Workman JL. The Role of Chromatin during Transcription. Cell. 2007;128: 707–719. doi:https://doi.org/10.1016/j.cell.2007.01.015
25. Carbonetto P, Stephens M. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. Bayesian Anal. 2012;7: 73–108. doi:10.1214/12-BA703
26. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. J R Stat Soc Ser B (Statistical Methodol. 2005;67: 301–320. Available: http://www.jstor.org/stable/3647580
27. Carbonetto P, Zhou X, Stephens M. varbvs: Fast Variable Selection for Large-scale Regression. arXiv Prepr arXiv170906597. 2017.
28. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLOS Genet. 2013;9: e1003264. Available: https://doi.org/10.1371/journal.pgen.1003264
29. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9: 215–216. doi:10.1038/nmeth.1906
30. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2013/10/03. 2014;24: 14–24. doi:10.1101/gr.155192.113
31. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proc Natl Acad Sci. 2015;112: 15390 LP – 15395. doi:10.1073/pnas.1503027112
32. Ray K, Szabo B. Variational Bayes for high-dimensional linear regression with sparse priors. 2019; 1–40. Available: http://arxiv.org/abs/1904.07150
33. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. PLOS Genet. 2013;9: e1003709. Available: https://doi.org/10.1371/journal.pgen.1003709
34. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017;49: 806. Available: https://doi.org/10.1038/ng.3831
35. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47: 1228. Available: https://doi.org/10.1038/ng.3404

36. Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, Hu Y, et al. Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLOS Genet. 2017;13: e1006933. Available: https://doi.org/10.1371/journal.pgen.1006933

37. Lambert J-C, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45: 1452–1458. doi:10.1038/ng.2802

38. Basurto-Islas G, Grundke-Iqbal I, Tung YC, Liu F, Iqbal K. Activation of Asparaginyl Endopeptidase Leads to Tau Hyperphosphorylation in Alzheimer Disease. J Biol Chem . 2013;288: 17495–17507. doi:10.1074/jbc.M112.446070

39. Climer LK, Dobretsov M, Lupashin V. Defects in the COG complex and COG-related trafficking regulators affect neuronal Golgi function . Frontiers in Neuroscience . 2015. p. 405. Available: https://www.frontiersin.org/article/10.3389/fnins.2015.00405

40. Qing-Qing Tao Zhi-Ying Wu Y-CC. The role of CD2AP in the Pathogenesis of Alzheimer's Disease. Aging and disease. pp. 901–907. Available: http://www.aginganddisease.org

41. Amlie-Wolf A, Tang M, Mlynarski EE, Kuksa PP, Valladares O, Katanic Z, et al. INFERNO: inferring the molecular mechanisms of noncoding genetic variants. Nucleic Acids Res. 2018;46: 8740–8753. doi:10.1093/nar/gky686

42. Wadsworth TL, Bishop JA, Pappu AS, Woltjer RL, Quinn JF. Evaluation of coenzyme Q as an antioxidant strategy for Alzheimer's disease. J Alzheimer's Dis. 2008;14: 225–234.

43. Meda SA, Koran MEI, Pryweller JR, Vega JN, Thornton-Wells TA. Genetic interactions associated with 12-month atrophy in hippocampus and entorhinal cortex in Alzheimer's Disease Neuroimaging Initiative. Neurobiol Aging. 2013;34: 1518.e9-1518.e18. doi:https://doi.org/10.1016/j.neurobiolaging.2012.09.020

44. Jaeger CUP and S. Functional Role of Lipoprotein Receptors in Alzheimers Disease. Current Alzheimer Research. 2008. pp. 15–25. doi:http://dx.doi.org/10.2174/156720508783884675

45. Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KL, Hunt P, et al. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. Hum Mol Genet. 2009;18: 4650–4661. doi:10.1093/hmg/ddp413

46. Hong C, Tontonoz P. Liver X receptors in lipid metabolism: opportunities for drug discovery. Nat Rev Drug Discov. 2014;13: 433–444. doi:10.1038/nrd4280

47. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al.

GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017. doi:10.1093/database/bax028

48. Renquist BJ, Madanayake TW, Hennebold JD, Ghimire S, Geisler CE, Xu Y, et al. TMEM135 is an LXR-inducible regulator of peroxisomal metabolism. bioRxiv. 2019; 334979. doi:10.1101/334979

49. Liu JZ, Erlich Y, Pickrell JK. Case–control association mapping by proxy using family history of disease. Nat Genet. 2017;49: 325–331. doi:10.1038/ng.3766

50. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14: 128. doi:10.1186/1471-2105-14-128

51. Nagpal S, Meng X, Epstein MP, Tsoi LC, Patrick M, Gibson G, et al. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am J Hum Genet. 2019;105: 258–266. doi:https://doi.org/10.1016/j.ajhg.2019.05.018

52. Price AL, Patterson N, Hancks DC, Myers S, Reich D, Cheung VG, et al. Effects of cis and trans Genetic Ancestry on Gene Expression in African Americans. PLOS Genet. 2008;4: e1000294. Available: https://doi.org/10.1371/journal.pgen.1000294

53. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44: 1084–1089. doi:10.1038/ng.2394

54. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, et al. Functional Architectures of Local and Distal Regulation of Gene Expression in Multiple Human Tissues. Am J Hum Genet. 2017;100: 605–616. doi:https://doi.org/10.1016/j.ajhg.2017.03.002

55. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell. 2019;177: 1022-1034.e6. doi:https://doi.org/10.1016/j.cell.2019.04.014

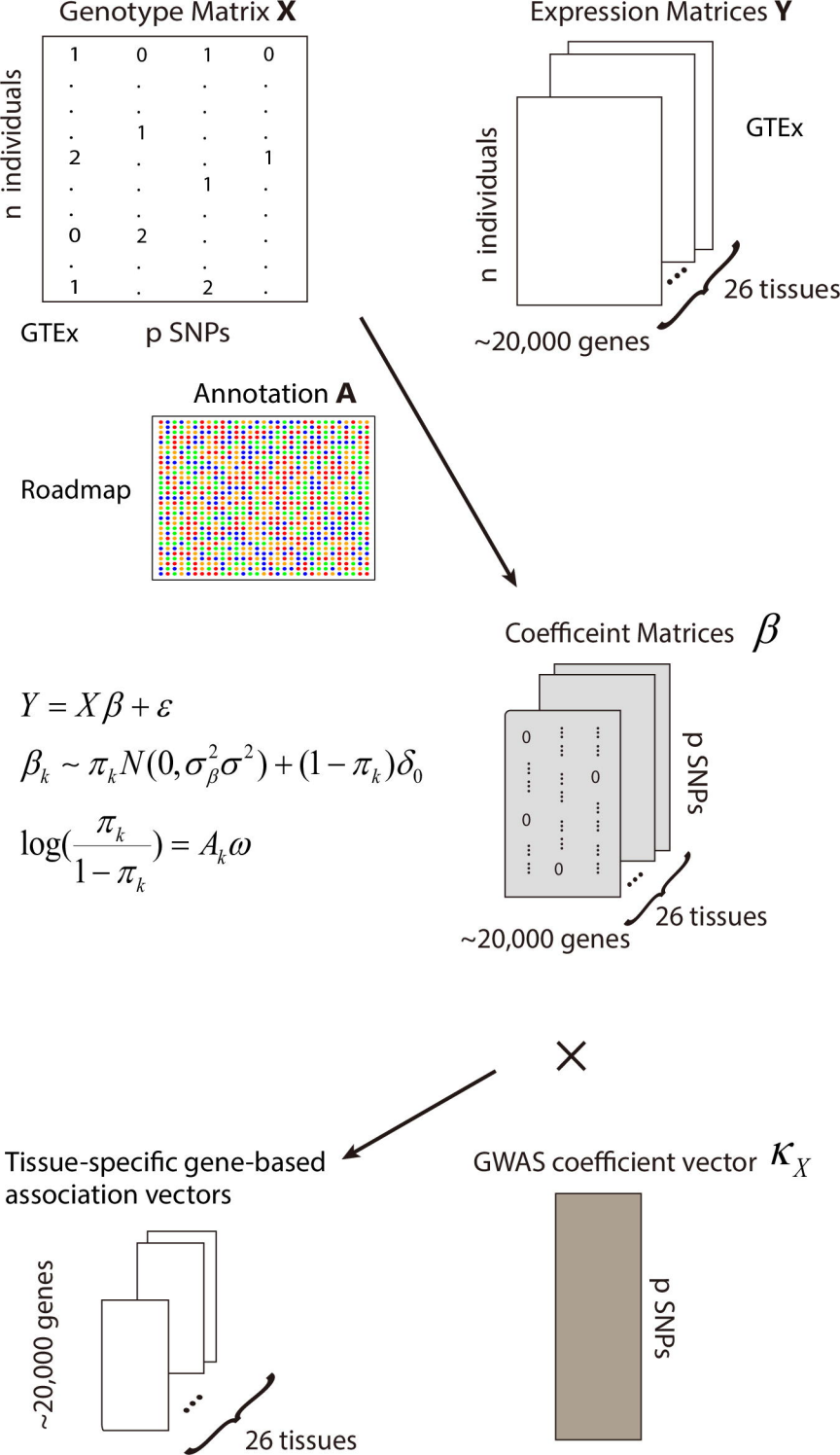
56. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015;348: 648 LP – 660. doi:10.1126/science.1262110

57. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33: 1–22. Available: https://www.ncbi.nlm.nih.gov/pubmed/20808728

58. Cheng J, Blum R, Bowman C, Hu D, Shilatifard A, Shen S, et al. A Role for H3K4 Monomethylation in Gene Repression and Partitioning of Chromatin Readers. Mol Cell. 2014;53: 979–992. doi:https://doi.org/10.1016/j.molcel.2014.02.032

59. Liang G, Lin JCY, Wei V, Yoo C, Cheng JC, Nguyen CT, et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A. 2004/05/03. 2004;101: 7357–7362. doi:10.1073/pnas.0401866101

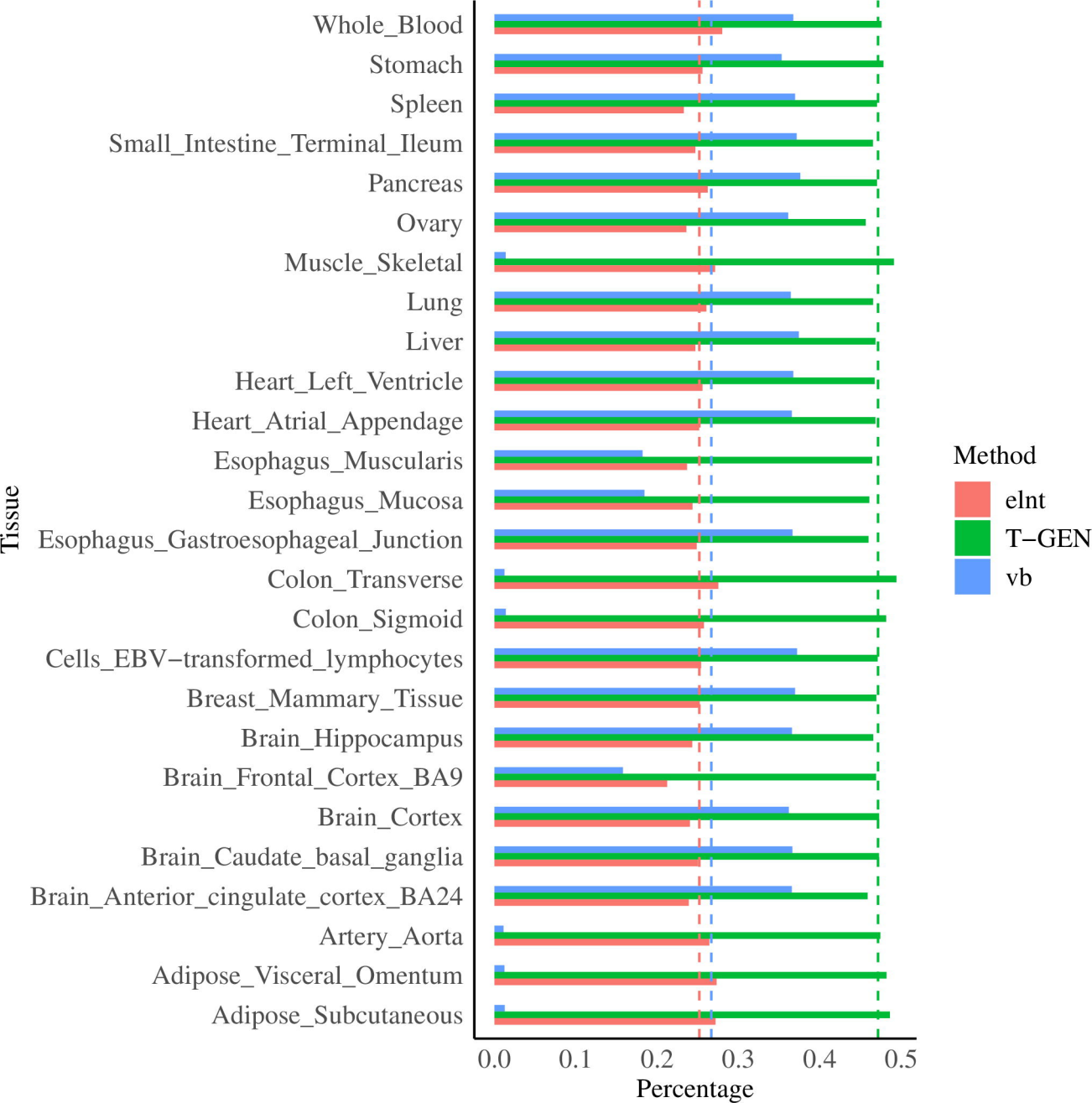
60. Zhou J, Wang X, He K, Charron J-BF, Elling AA, Deng XW. Genome-wide profiling of histone H3 lysine 9 acetylation and dimethylation in Arabidopsis reveals correlation between multiple histone marks and gene expression. Plant Mol Biol. 2010;72: 585–595. doi:10.1007/s11103-009-9594-7



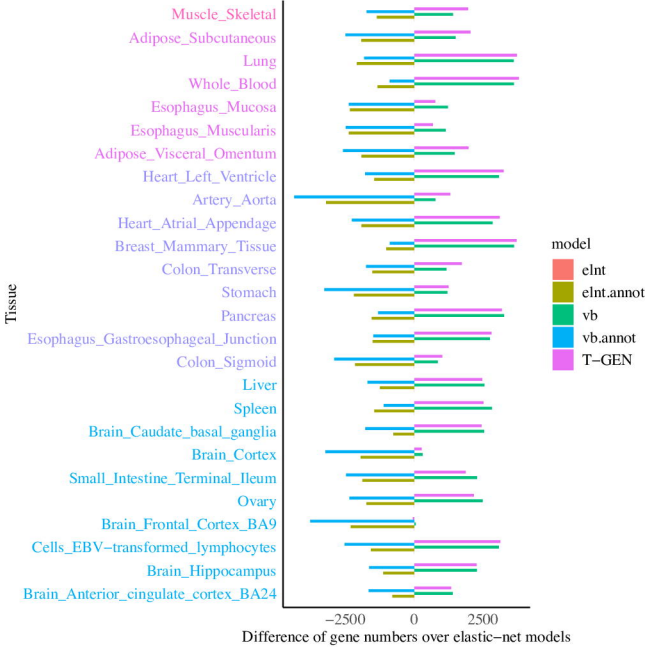
$$Y = X\beta + \varepsilon$$

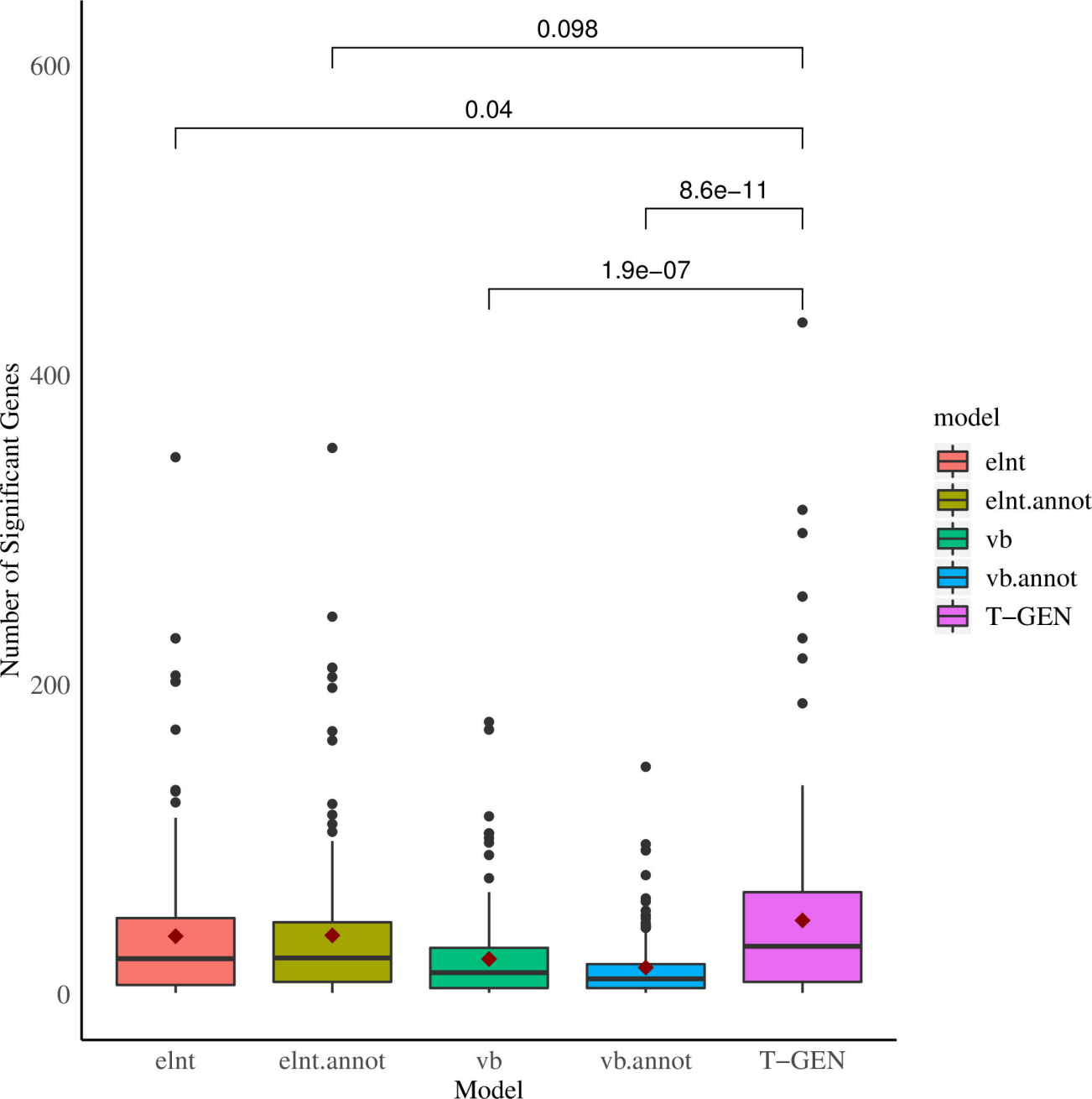
$$\beta_k \sim \pi_k N(0, \sigma_\beta^2 \sigma^2) + (1 - \pi_k)\delta_0$$

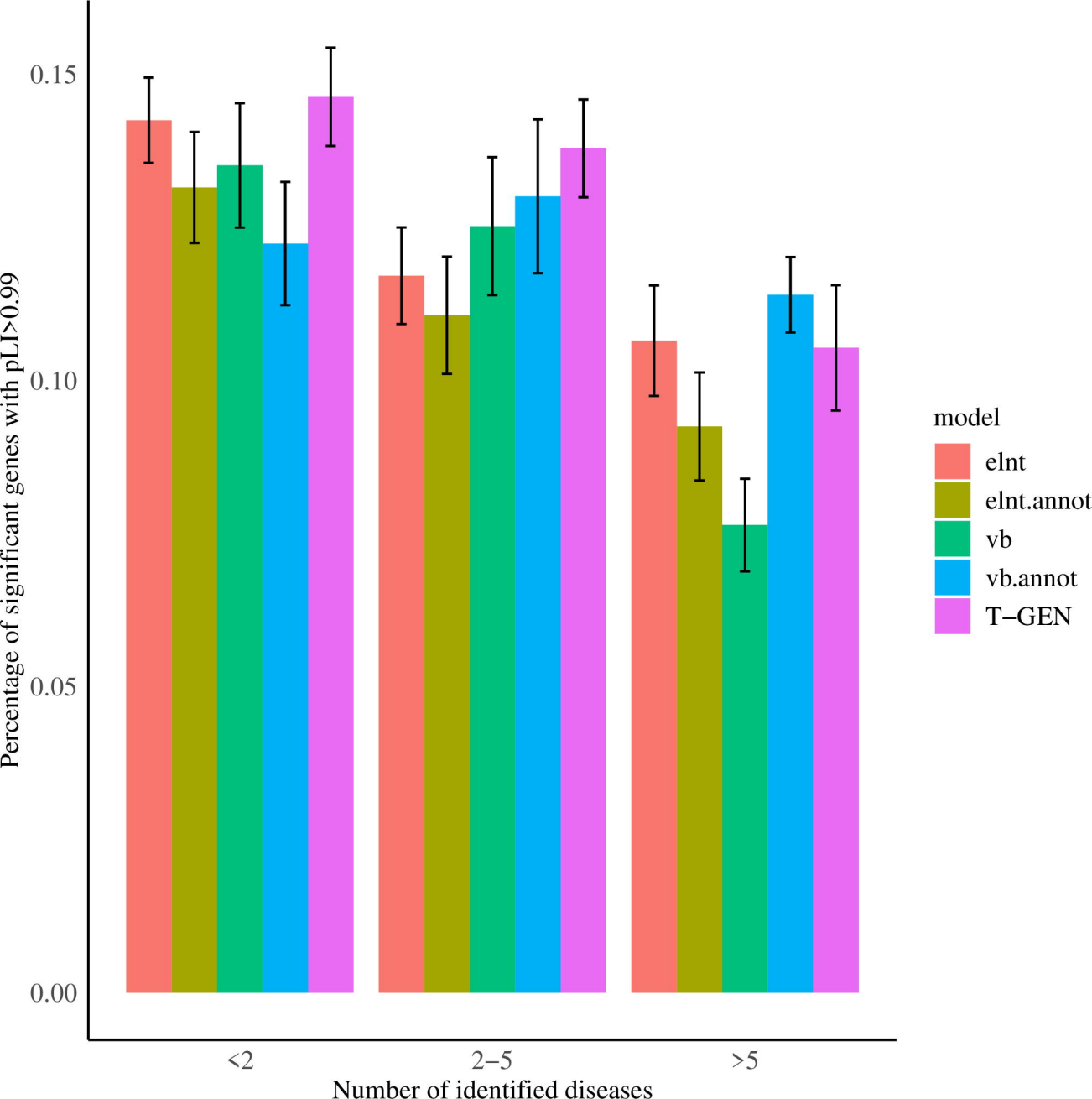
$$\log\left(\frac{\pi_k}{1 - \pi_k}\right) = A_k \omega$$

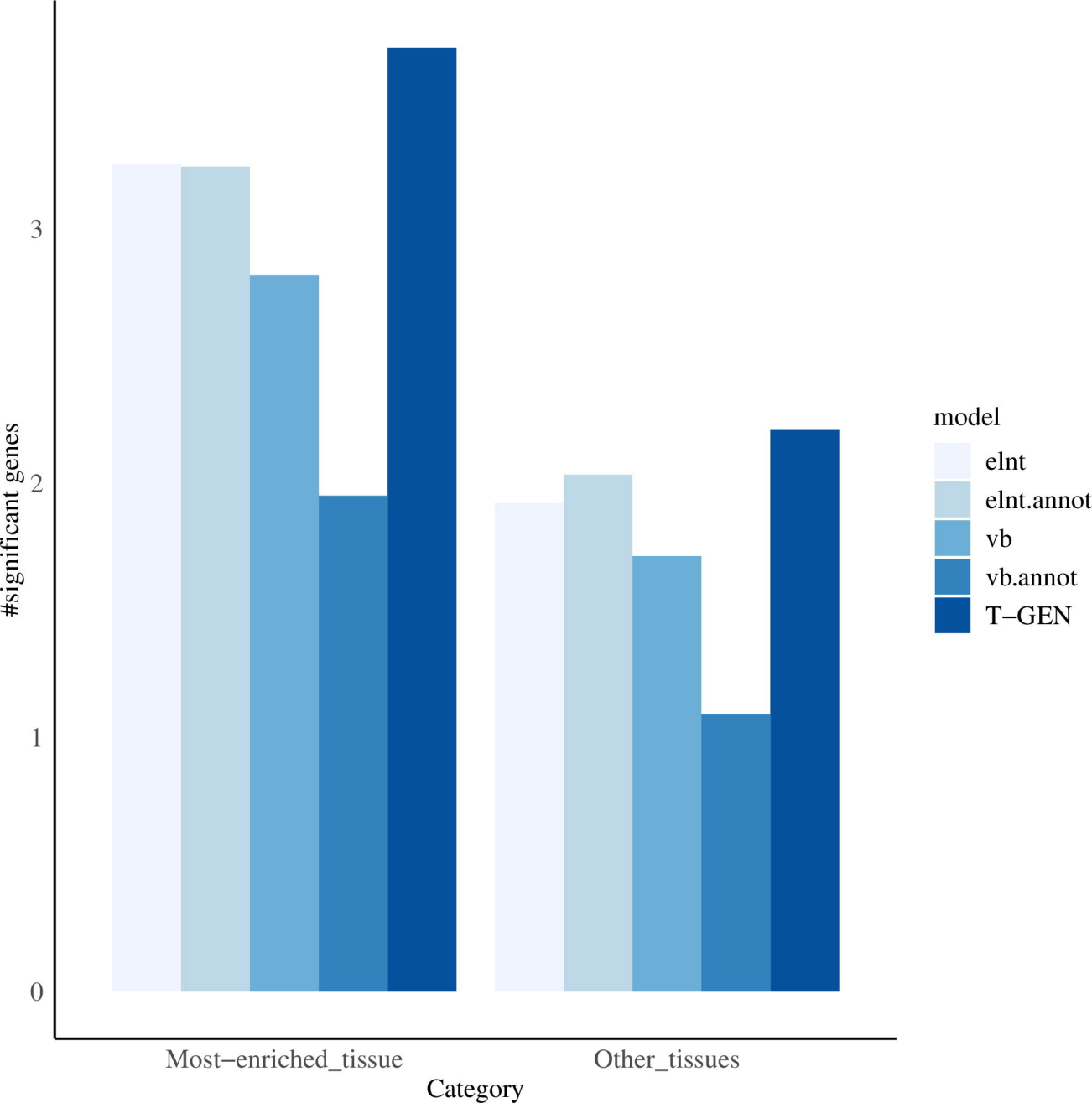

Whole_Blood
Stomach
Spleen
Small_Intestine_Terminal_Ileum
Pancreas
Ovary
Muscle_Skeletal
Lung
Liver
Heart_Left_Ventricle
Heart_Atrial_Appendage
Esophagus_Muscularis
Esophagus_Mucosa
Esophagus_Gastroesophageal_Junction
Colon_Transverse
Colon_Sigmoid
Cells_EBV−transformed_lymphocytes
Breast_Mammary_Tissue
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

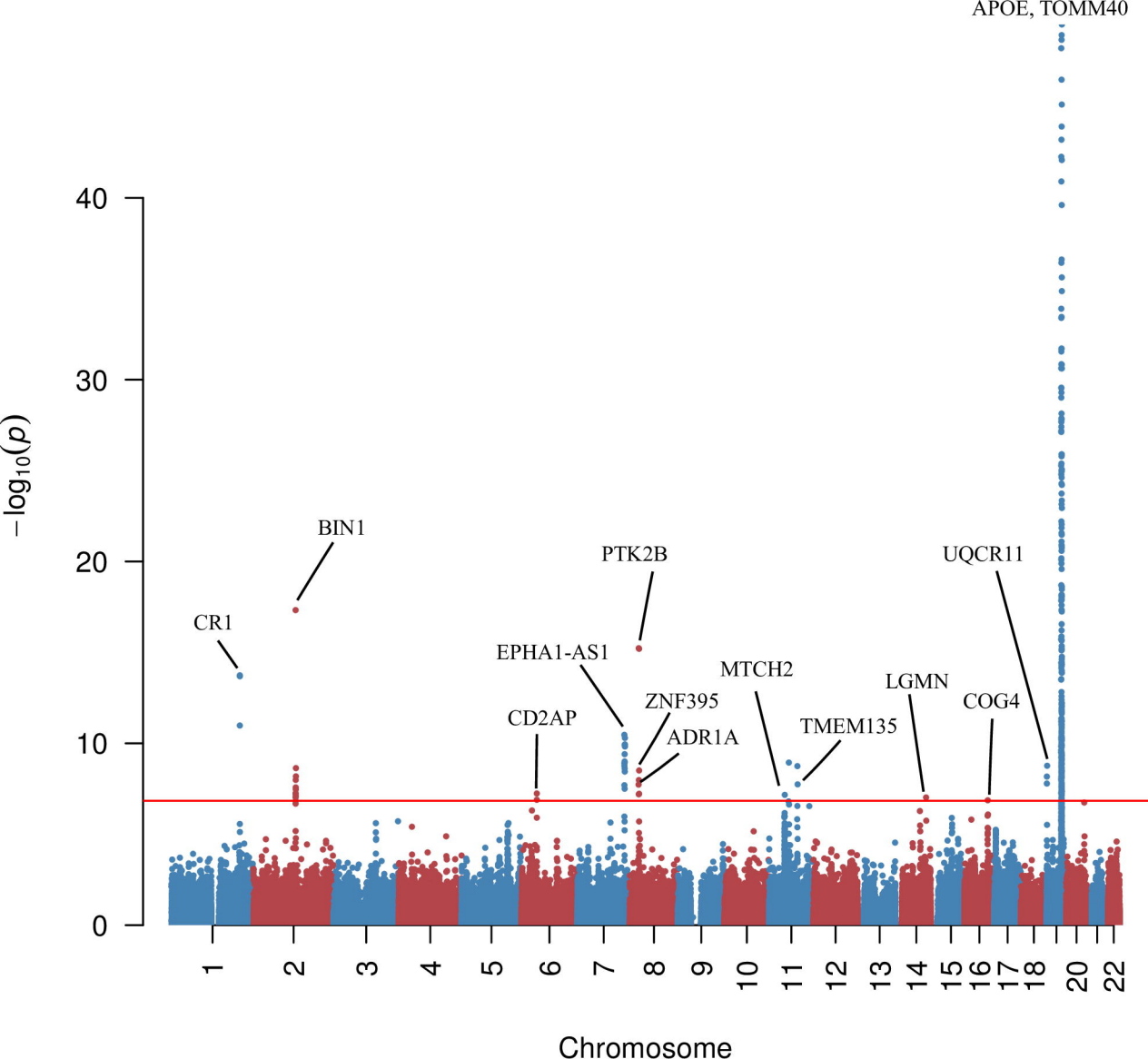
Muscle_Skeletal
Adipose_Subcutaneous
Lung
Whole_Blood
Esophagus_Mucosa
Esophagus_Muscularis
Adipose_Visceral_Omentum
Heart_Left_Ventricle
Artery_Aorta
Heart_Atrial_Appendage
Breast_Mammary_Tissue
Colon_Transverse
Stomach
Pancreas
Esophagus_Gastroesophageal_Junction
Colon_Sigmoid
Liver
Spleen
Brain_Caudate_basal_ganglia
Brain_Cortex
Small_Intestine_Terminal_Ileum
Ovary
Brain_Frontal_Cortex_BA9
Cells_EBV−transformed_lymphocytes
Brain_Hippocampus
Brain_Anterior_cingulate_cortex_BA24
Tissue
−0.25
0.00
0.25
0.50
ratio of gene numbers over elastic−net models

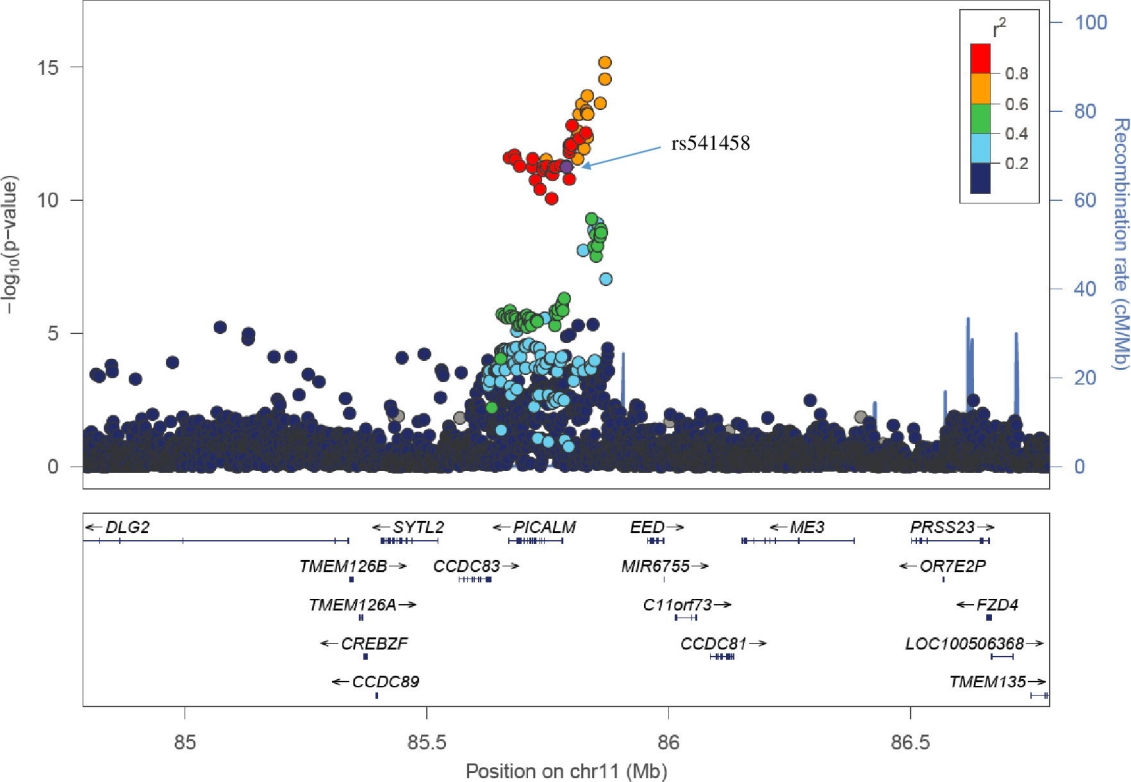

Muscle_Skeletal
Adipose_Subcutaneous
Lung
Whole_Blood
Esophagus_Mucosa
Esophagus_Muscularis
Adipose_Visceral_Omentum
Heart_Left_Ventricle
Artery_Aorta
Heart_Atrial_Appendage
Breast_Mammary_Tissue
Colon_Transverse
Stomach
Pancreas
Esophagus_Gastroesophageal_Junction
Colon_Sigmoid
Liver
Spleen
Brain_Caudate_basal_ganglia
Brain_Cortex
Small_Intestine_Terminal_Ileum
Ovary
Brain_Frontal_Cortex_BA9
Cells_EBV−transformed_lymphocytes
Brain_Hippocampus
Brain_Anterior_cingulate_cortex_BA24
Tissue
−2500
0
2500
Difference of gene numbers over elastic−net models
model
elnt
elnt.annot
vb
vb.annot
T−GEN


0.098
0.04
8.6e−11
1.9e−07
600
400
200
0
Number of Significant Genes
elnt
elnt.annot
vb
vb.annot
T−GEN
Model
model
elnt
elnt.annot
vb
vb.annot
T−GEN


0.15
0.10
0.05
0.00
Percentage of significant genes with pLI>0.99
<2
2–5
>5
Number of identified diseases
model
elnt
elnt.annot
vb
vb.annot
T–GEN


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


APOE, TOMM40
40
30
20
10
0
$-\log_{10}(p)$
BIN1
CR1
PTK2B
EPHA1-AS1
CD2AP
ZNF395
ADR1A
MTCH2
TMEM135
LGMN
COG4
UQCR11
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
22
Chromosome


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)