Cross-population Meta-analysis of eQTLs: Fine Mapping and Functional Study

Xiaoquan Wen[1*], Francesca Luca[2,3] and Roger Pique-Regi[2,4*]
1 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
2 Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA
3 Department of Obstetrics and Gynecology, Wayne State University, Detroit, MI, USA
4 Department of Clinical and Translational Sciences, Wayne State University, Detroit, MI, USA
∗ **E-mail: xwen@umich.edu, rpique@wayne.edu**

Abstract

Mapping expression quantitative trait loci (eQTLs) has been shown as a powerful tool to uncover the genetic underpinnings of many complex traits at the molecular level. In this paper, we present an integrative analysis approach that leverages eQTL data collected from multiple population groups. In particular, our approach effectively identifies multiple independent *cis*-eQTL signals that are consistently presented across populations, accounting for population heterogeneity in allele frequencies and linkage disequilibrium patterns. Furthermore, by integrating genomic annotations, our analysis framework enables high-resolution functional analysis of eQTLs. We applied our statistical approach to analyze the GEUVADIS data consisting of samples from five population groups. From this analysis, we concluded that i) jointly analysis across population groups greatly improves the power of eQTL discovery and the resolution of fine mapping of causal eQTL. ii) many genes harbor multiple independent eQTLs in their *cis* regions iii) genetic variants that disrupt transcription factor binding are significantly enriched in eQTLs (p-value $= 4.93 \times 10^{-22}$).

Author Summary

Expression quantitative trait loci (eQTLs) are genetic variants associated with gene expression phenotypes. Mapping eQTLs enables studying genetic basis of gene expression variation. In this study, we introduce a statistical framework to analyze genotype-expression data collected from multiple population groups. We show that our approach is particularly effective in identifying potentially multiple independent eQTL signals that are consistently presented across populations in the proximity of a gene. In addition, our analysis framework allows effective integration of genomic annotations into eQTL analysis, which is helpful to dissect the functional basis of eQTLs.

1 Introduction

Expression quantitative trait loci, or eQTLs, are genetic variants that are associated with gene expression levels. Mapping eQTLs can help in dissecting the molecular mechanisms by which genetic variants impact organismal phenotypes. Recent studies ([1–3]) have revealed that there are substantial overlaps between eQTLs and genetic variants identified from genome-wide association studies (GWAS) of disease phenotypes. In addition, eQTL mapping provides a powerful tool for investigating the regulatory machinery in different tissues ([4, 5]) or cellular environments ([6–8]).

In this paper, we *jointly* address three outstanding issues in eQTL mapping. First, due to the high experimental cost, most available eQTL data sets typically have limited sample size. To improve power of eQTL discovery, it becomes necessary to aggregate evidence across multiple data sets. Second, because a gene is typically regulated by many regulatory elements, it is highly likely that there exist multiple independent eQTLs in its proximity (i.e., *cis* region). In this scenario, a multiple SNP analysis is required to uncover all relevant *cis* acting genetic factors involved in the gene regulation process ([9]). Third, the availability of extensive functional annotations ([10–12]) now enables integration of functional genomic information into eQTL analysis, which can be useful to dissect the functional basis of eQTLs . Linking genomic annotations to eQTLs goes beyond genetic association analysis, and helps gain a better understanding of the underlying biological processes. Individually, some of these three issues have been discussed by previous works. For example, [3, 9, 13–16] discussed single SNP analysis of eQTLs jointly from different studies, populations or tissues. But these methods do not naturally extended to multiple SNP analysis. [17–20] examined the enrichment of selected genomic features in *cis*-eQTLs, mostly based on single SNP association results. To the best of our knowledge, there is no existing approach that jointly addresses all three issues in a systematic way.

In this paper, we demonstrate an integrative analysis approach to perform fine mapping and functional study of eQTLs using cross-population samples. Our statistical methods are extended from a Bayesian framework proposed by [14, 15, 21], which has been successfully applied in mapping eQTLs from multiple tissues. We apply our statistical framework to analyze the data from the GEUVADIS project ([20]), where the expression-genotype data are collected from five population groups. In GWAS, trans-ethnic meta-analysis of genetic association data from diverse populations has been shown to be a powerful tool in detecting novel complex trait loci and improving resolution of fine mapping of causal variants by leveraging population heterogeneity in local patterns of linkage disequilibrium (LD) and allele frequencies ([22, 23]). This approach, to the extent of our knowledge, has not been applied to eQTL analysis. Utilizing cross-population expression-genotype data, we are interested in identifying eQTL signals that are *consistently* presented in all populations. Furthermore, we aim to examine whether we have sufficient statistical power to identify multiple independent *cis*-eQTL signals with the available aggregated sample size. Last but not least, we set out to investigate the role of genetic variants that disrupt transcription factor (TF) binding in transcriptional processes. More specifically, we ask the question if such variants more likely to be eQTLs. The three of our main aims are also inter-related. With higher power through sample aggregation in mapping eQTLs, we expect to identify potentially multiple genomic regions that harbor casual eQTLs at a high resolution. And consequently, we anticipate that these efforts improve the statistical power and precision of localization for our functional analysis.

2 Results

2.1 Method Overview

We start by a brief description of our statistical procedure and general strategy for multiple SNP fine mapping analysis in a meta-analytic setting across multiple populations.

2.1.1 Statistical Model and Inference

Consider a genomic region with p SNPs that is interrogated in s different population groups. In each group i, we use a multiple linear regression model to describe the potential genetic associations between the p SNPs and the expression levels of a target gene:

$$\boldsymbol{y}_i = \mu_i \mathbf{1} + \sum_{j=1}^{p} \beta_{i,j} \boldsymbol{g}_{i,j} + \boldsymbol{e}_i, \ \boldsymbol{e}_i \sim \mathrm{N}(\mathbf{0}, \sigma_i^2 \boldsymbol{I}), \ i = 1, ..., s, \tag{1}$$

where the vectors $\boldsymbol{y}_i$ and $\boldsymbol{e}_i$ represent the expression levels and the residual errors in population group i, the parameters μ_i and σ_i^2 denote the population group specific intercept and residual error variance, respectively. The vector $\boldsymbol{g}_{i,j}$ denotes the genotype of SNP j in population group i, and the regression coefficient $\beta_{i,j}$ represents its genetic effect. Across all population groups, the s linear models form a system of simultaneous linear regressions (SSLR, [15]). The problem of mapping eQTLs can be framed as identifying SNPs with non-zero $\beta_{i,j}$ values based on (1). For each SNP j, we define an indicator vector

$$\boldsymbol{\gamma}_j := \Big(\mathbf{1}(\beta_{1,j} \neq 0), \dots, \mathbf{1}(\beta_{s,j} \neq 0)\Big)$$

representing its association status in each of the s population groups. Such indicator is referred to as a "configuration" in the literature of genetic association analysis across multiple subgroups ([14, 15, 24]). For each target gene, our computational procedure is designed to make joint inference with respect to $\boldsymbol{\Gamma} := \{\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p\}$ given observed genotype data, $\boldsymbol{G} := \{\boldsymbol{g}_{1,1}, \dots, \boldsymbol{g}_{1,p}, \dots, \boldsymbol{g}_{s,p}\}$, and phenotype data, $\boldsymbol{Y} := \{\boldsymbol{y}_1, \dots, \boldsymbol{y}_s\}$.

For mapping eQTLs in a meta-analytic setting, we make an important assumption that if a SNP is genuinely associated with a given expression phenotype, its underlying genetic effects are non-zero in *all* population groups. That is, $\boldsymbol{\gamma}_j = \mathbf{1}$, if SNP j is an eQTL, and $\mathbf{0}$ otherwise. This assumption is largely motivated by the biological hypothesis that the regulatory mechanisms behind eQTLs should remain the same across populations.

For each SNP j, we assume an independent prior for $\boldsymbol{\gamma}_j$, which assigns most of the probability mass on $\boldsymbol{\gamma}_j = \mathbf{0}$ and encourages an overall sparse structure of $\boldsymbol{\Gamma}$. This is because most previous studies only identify small numbers of *cis*-eQTLs for any given gene. Particularly in our fine mapping analysis where p is typically in the magnitude of 10^3 to 10^4, we use

$$\Pr\left(\boldsymbol{\gamma}_j = \mathbf{1}\right) = \frac{1}{p}, \tag{2}$$

which implies that we expect a single *cis*-eQTL signal for the target gene *a priori*, because

$$\mathrm{E}\left[\sum_{j=1}^{p} \mathbf{1}(\boldsymbol{\gamma}_j \neq \mathbf{0})\right] = 1.$$

We will further justify this prior specification based on our overall strategy for fine mapping analysis in section 2.1.3, and discuss some other alternative specifications for enrichment analysis in section 2.1.4.

Conditional on SNP j being an eQTL, we model its genetic effects across populations, $\boldsymbol{\beta}_j := (\beta_{1,j}, \dots, \beta_{s,j})$, using a flexible Bayesian prior for meta-analysis proposed in [21]. Briefly, we assume that the effect sizes of an eQTL are correlated while allowing reasonable degree of heterogeneity across population groups (See Method section for details).

Given observed data $(\boldsymbol{Y}, \boldsymbol{G})$ and a specified $\boldsymbol{\Gamma}$ value, a Bayes factor

$$\mathrm{BF}(\boldsymbol{\Gamma}) := \frac{P(\boldsymbol{Y} \,|\, \boldsymbol{G}, \boldsymbol{\Gamma})}{P(\boldsymbol{Y} \,|\, \boldsymbol{G}, \boldsymbol{\Gamma} \equiv 0)},$$

can be analytically approximated with high precision following [15]. Based on this result, we compute the posterior of $\boldsymbol{\Gamma}$ using the Bayes rule, i.e.,

$$\begin{aligned} \Pr(\boldsymbol{\Gamma} \,|\, \boldsymbol{Y}, \boldsymbol{G}) &\propto \Pr(\boldsymbol{\Gamma})\, P(\boldsymbol{Y} \,|\, \boldsymbol{\Gamma}, \boldsymbol{G}) \\ &\propto \left(\prod_j \Pr(\boldsymbol{\gamma}_j)\right) \mathrm{BF}(\boldsymbol{\Gamma}). \end{aligned} \tag{3}$$

We implement an efficient MCMC algorithm to perform the posterior inference and summarize the fine mapping results from the posterior samples. More specifically, we compute the posterior probability for each possible $\boldsymbol{\Gamma}$ by the corresponding frequency in the posterior samples. We will refer to this quantity as "posterior model probability" henceforth. To evaluate the importance of each SNP, we compute a posterior inclusion probability (PIP) for each SNP by marginalizing over all posterior model probabilities. If a SNP is included in posterior models with high frequencies, the corresponding PIP tends to be large.

2.1.2 Dealing with Genetic Data from Multiple Populations

We note that analyzing eQTL data using samples collected from multiple populations is slightly different from the standard settings in meta-analysis of genetic associations. In particular, the allele frequencies of interrogated SNPs and/or patterns of linkage disequilibrium presented in genotype data can be highly heterogeneous in different populations. The proposed multiple SNP fine mapping procedure takes advantage of the unique setting of cross-population samples, and leverages the statistical power for eQTL discovery.

The allele frequency of a SNP is unrelated to its underlying genetic effect with respect to expression levels. However, because of its direct impact on sampling errors, it affects the precision of *estimated* $\beta_{i,j}$ (denoted by $\hat{\beta}_{i,j}$) in model (1). Lower allele frequency typically results in larger uncertainty (e.g., the standard error of $\hat{\beta}_{i,j}$ of a rare SNP is usually larger than that of a common SNP), which implies that rare SNPs tend to be less informative. In the extreme case if a SNP is monomorphic in samples from a particular population, it should be considered completely uninformative for examining genetic associations. As shown by [15], Bayesian procedures, especially in use of Bayes factors, can precisely capture the (non-)informativeness of SNPs in each population group: e.g., including the genetic data of monomorphic SNPs yields identical inference results as discarding such SNPs from corresponding population groups. On the other hand, our adopted Bayesian meta-analysis prior enables aggregating less informative or weak association signals from low frequency SNPs, as long as they are *consistent* across population groups.

In pursuing consistent association signals, our fine mapping procedure takes advantage of potential varying LD patterns across multiple population groups. This is mainly because our method favors identifying eQTLs whose effects are consistent in *all* population groups, whereas SNPs that tag causal variants only in *some* populations (due to population-specific LD structures) are automatically down-weighted. As a consequence, the genomic regions that harbor causal eQTLs can be effectively narrowed down using cross-population data. This advantage becomes even more obvious when performing *multiple* SNP analysis, as all candidate *cis*-SNPs are simultaneously evaluated.

2.1.3 Gene-level Testing Prior to Fine Mapping *cis*-eQTLs

In most of currently available data sets, eQTL discoveries are made only in a proportion of genes. To reduce the computational cost of the fine mapping analysis, we adopt a practical procedure that first screens the genes having at least one *cis*-eQTL (which we will call as "eGenes").

The statistical procedure to identify eGenes across multiple subgroups has been well established in [14]. More specifically in the context of cross-population eQTL mapping, it assumes a similar linear model system as (1), and performs gene-level Bayesian hypothesis testing. Namely, for each gene, we test

$$H_0 : \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_p = \mathbf{0}$$

versus

$$H_1 : \text{some } \boldsymbol{\gamma}_j = \mathbf{1}.$$

The eGenes are identified upon rejections of corresponding null hypothese, and we only follow up the eGenes with the multiple *cis*-eQTL analysis.

The gene-level testing is computationally efficient, it effectively filters out a substantial set of genes that are unlikely to be interesting for fine mapping. Furthermore, this additional procedure justifies our use of prior (2): given that a gene is identified as eGene, expecting a single *cis*-eQTL prior to the fine mapping analysis seems, in average, a slightly conservative assumption.

2.1.4 Integration of Genomic Feature and Enrichment Analysis

In the enrichment analysis, we are interested in identifying some common properties shared by eQTL SNPs genome-wide. Intuitively, in contrast to the fine mapping procedure where each gene is separately processed, the enrichment analysis requires jointly analyzing all gene-SNP pairs.

Our Bayesian framework provides a natural way to integrate genomic features of SNPs into the association analysis. In particular, we use the prior specification of $\Pr(\boldsymbol{\gamma}_j)$ to incorporate the genomic annotations of SNP j with respect to the target gene. More specifically, we assume a general logistic model,

$$\log\left[\frac{\Pr(\boldsymbol{\gamma}_j=\mathbf{1})}{\Pr(\boldsymbol{\gamma}_j=\mathbf{0})}\right]=\alpha_0+\sum_k \alpha_k\delta_{kj}, \tag{4}$$

where δ_{kj} denotes the k-th annotation for SNP j. The regression coefficient α_k represents the strength of association between the k-th genomic annotation and the prior probability of a SNP being an eQTL, and is assumed to be invariant across all gene-SNP pairs. For a binary annotation, a positive α_k implies that an annotated SNP has higher odds to be an eQTL, or equivalently speaking, the annotated SNPs are enriched in eQTLs. For the remaining of this paper, we will refer to α_k's as enrichment parameters.

The inference procedure of the Bayesian model with prior specification (4) is conceptually straightforward, and we give the details in Method section. However, because the estimation of the enrichment parameters requires pooling information across all genes, the computational cost is substantially higher comparing to the gene by gene fine mapping analysis.

To ease the computational burden, we derive an approximate inference procedure focusing on hypothesis testing of α_k's. This procedure first performs separate multiple *cis*-eQTL mapping for *every* gene interrogated (i.e., not just eGenes), using a special case of the prior model (4), i.e.,

$$\log\left[\frac{\Pr(\boldsymbol{\gamma}_j=\mathbf{1})}{\Pr(\boldsymbol{\gamma}_j=\mathbf{0})}\right]=\alpha_0.$$

In particular, we determine α_0 in a similarly conservative way as (2), using the results from gene-level testing. Let g_e and g_s denote the number of eGenes identified and the number of total gene-SNP pairs in the analysis, respectively. We set

$$\Pr(\boldsymbol{\gamma}_j=\mathbf{1})=\kappa=\frac{g_e}{g_s}, \tag{5}$$

which, in a way, assumes that each eGene contains only a single associated gene-SNP pair, and corresponds to $\alpha_0=\log\left(\frac{\kappa}{1-\kappa}\right)$. We then fit a logistic regression model by correlating the resulting PIP of each gene-SNP pair with the genomic annotations of the corresponding SNP. The fitted regression coefficient of each annotation is then regarded as the corresponding α_k estimate. This approximate inference procedure seemingly resembles some of the *post hoc* enrichment analysis approaches that are commonly applied ([7–9, 20]). We provide a statistical justification of this procedure in Method section.

2.2 Analysis of GEUVADIS Data

In this paper, we focused on analyzing the expression and genotype data collected from the GEUVADIS project ([20]). More specifically, the data set consists of RNA-seq data on lymphoblastoid cell line (LCL) samples from five populations: the Yoruba (YRI), CEPH (CEU), Toscani (TSI), British (GBR) and Finns (FIN). In our analysis, we selected 420 samples which were densely genotyped in the 1000 Genomes Phase I data release ([25]) and 11,838 protein coding genes and lincRNAs that are deemed expressed in all five population groups. Throughout, our analysis focused on the SNPs that locate within a 200kb genomic region centered at the transcription start site (TSS) of each gene. In contrast to the original eQTL mapping discussed in [20], we treated each population as a single group and performed *cis*-eQTL analysis jointly across all five groups.

2.2.1 Power Gain in gene-level Meta-Analysis

We started our analysis of the GEUVADIS data by performing gene-level testing jointly across all five population groups.

In total, we identified 6,555 eGenes from 11,838 tested protein coding and lincRNA genes at 5% FDR level. For comparison, the numbers of eGenes identified using each population data alone are given in Table 1. The separate analysis identified no more than 2,100 eGenes in any of the population groups. The union of the eGenes from the separate analysis yielded 3,447 genes, the vast majority of which (except for 60 genes) were included in the discoveries by the meta-analysis. In addition, the meta-analysis identified a total of 3,168 new eGenes.

Examining the set of eGenes uniquely identified in the meta-analysis, we found that they share the following common feature: when performing separate analysis within each population group, the strongest association in each respective *cis*-region only shows modest strength and does not pass the required significance threshold; However, across population groups, the same association signal tends to be highly consistent, and the overall evidence aggregated across population groups becomes quite strong. As a consequence, the joint analysis of all population groups is able to detect such signals. We demonstrate one of such examples in Figure 1 using gene *NME1* (ensemble ID: ENSG00000239672), where the gene-level Bayes factor in the joint analysis is several order of magnitude higher than the corresponding gene-level Bayes factors in each separate population group analysis.

2.2.2 Multiple SNP analysis of eGenes

We followed up the gene-level analysis by performing multi-SNP fine mapping for the set of identified eGenes across all 5 population groups.

One of our primary aims was to identify potential multiple independent *cis*-eQTL signals while accounting for varying LD patterns in different populations. First, we asked how often we can identify multiple *cis*-eQTL signals in eGenes. To this end, for each fine mapped eGene, we computed the expected number of independent *cis*-eQTL signals from the corresponding posterior distributions (Method section). Figure 2 shows the histogram of posterior expected *cis*-eQTL signals in all 6,555 eGenes. It is clear that for the available (accumulated) sample size in the GEUVADIS data, we identified only single eQTL signals for the majority of the eGenes. Nevertheless, for a non-trivial proportion of genes, there are strong evidence that multiple *cis*-acting regulatory genetic variants co-exist and can be confidently identified by our fine mapping procedure. More specifically, there are about 14% of the eGenes (or 7% of all interrogated genes) with the posterior expected number of *cis*-eQTLs ≥ 2.

In the case of gene *LHPP* (Eensemble ID: ENSG00000107902), four independent eQTL signals were confidently identified (Figure 3). Each eQTL signal (except one) is represented by a cluster of highly correlated SNPs in a small genomic region. Due to LD, within each cluster the correlated SNPs tend to have similar PIPs and we could not be certain which SNP is truly driving the association signal. However, the sum of the PIPs within each cluster is very close to 1, indicating almost certainty that an eQTL is located within the region. There were 134 different posterior models examined for gene *LHPP* in the sampling phase of the MCMC run. Interestingly, every model contains exactly 4 SNPs (which results in posterior expected number of *cis*-eQTLs being 4 with variance 0), each from one of the four independent clusters.

As discussed in section 2.1.2, even in the cases when only a single *cis*-eQTL signal can be identified, we observed that the fine mapping procedure takes advantage of varying LD patterns across populations and narrows down the set of candidate causal variants by down-weighting SNPs that tag causal variants only in some populations. For example, in analyzing the data of gene *AGO3* (Ensemble ID: ENSG00000126070) from TSI alone, we identified a strong single eQTL signal within a 144kb region. The region contains 41 SNPs that are in perfect LD and show equal strengths of associations. In other four populations, this particular region is broken into smaller LD blocks and the associations for the 41 SNPs become distinguishable. As a result, when performing multiple-SNP analysis across all populations, we narrowed down the potential causal eQTL into a 1.2kb region, with only 3 candidate SNPs (in high LD in all populations) fully explaining the observed association.

To further quantify the effect of LD filtering, we selected a set of 526 eGenes that highly likely harbor exactly one *cis*-eQTL based on our multiple SNP analysis. More specifically, we selected the genes whose posterior expected number of *cis*-eQTLs = 1 and variance $\leq 1 \times 10^{-4}$. We then ran multiple *cis*-eQTL analysis separately in each of the five populations for each selected gene. For every multiple SNP analysis of each gene, including the meta-analysis, we constructed a 95% credible set that contains the minimum number of SNPs with the sum of PIPs ≥ 0.95, and defined its credible region length as the distance between the the right-most and left-most SNPs in the set. In 486 out of 526 (or 92%) genes, we found that the cross-population meta-analysis yields the smallest credible region length. The median ratio of credible region length by the meta-analysis to the minimum credible region length by the corresponding separate analyses is 0.50 across all 526 genes, indicating that, in average, cross-population meta-analysis can effectively narrow down the genomic regions that harbor *cis*-eQTLs.

Multi-SNP analysis can also be very helpful in explaining some of the extreme heterogeneity of eQTL effects across populations observed in *single* SNP analysis. In particular, we identified a few SNPs that, when analyzed alone, appear to show strong but opposite genetic effects on expression levels in different populations. It seemingly suggests that a particular allele of the variant increases expression levels of the target gene in one population and decreases expression levels in another population. However, going through all such individual examples, we found none of the opposite effect association patterns is supported by our multiple SNP analysis. In mapping *cis*-eQTLs for gene *TTC38* (Ensemble ID: ENSG00000075234), we found a set of tightly linked SNPs displaying opposite directional effects in YRI and the European populations in single SNP analysis. For example, the A allele of SNP rs6008600 shows consistently strong negative effect in the four European populations, whereas in YRI the same allele displays a highly significant positive effect (Figure 5). The multiple SNP fine mapping offered a compelling explanation for this phenomenon: two independent *cis*-eQTL signals were identified in the nearby regions, and interestingly, SNP rs6008600 tags one signal in YRI ($r^2 \approx 0.73$), whereas in the European populations, it is highly correlated (e.g., in CEU $r^2 \approx 1$) with the other signal (Figure 5).

Although our multiple *cis*-eQTL mapping method confidently identified independent association signals accounting for LD, in most of the examples we have examined, it usually could not pinpoint a single causal SNP by fully resolving LD relying only on the association data. This is because highly correlated SNP genotypes are nearly "interchangeable" in our statistical association models, and therefore not identifiable. As a consequence, there are many combinations of SNPs showing equivalence or near equivalence based on observed data. Reporting a single "best" model while ignoring its intrinsic uncertainty can be highly problematic. In the previously mentioned example of gene *LHPP*, we found that a large proportion of the 134 reported posterior models by the MCMC algorithm exhibit very similar likelihood and posterior probabilities, and the maximum posterior model probability is only 0.02. Our Bayesian fine mapping approach summarizes the measure of uncertainties using the posterior probabilities both at model and SNP levels, which can be naturally carried over to potential downstream functional analysis.

2.2.3 Functional analysis of eQTLs

We applied the approximate inference procedure described in section 2.1.4 to perform enrichment analysis using the GEUVADIS data. In particular, we conducted the multiple *cis*-eQTL analysis for all 11,858 genes using the prior (5). Notwithstanding the conceptual difference between priors (2) and (5), the overall results of multiple *cis*-eQTL analysis for identified eGenes are markedly similar. Based on these result, we set out to investigate the relationships between eQTLs and some functional genomic features.

We started with examining the enrichment of eQTL signals with respect to their distances to the TSS of respective target genes. More specifically, we computed the posterior expected numbers of eQTL signals within the non-overlapping 1Kb windows defined by their distances to TSS by summing over the

PIPs of all *cis*-SNPs falling in each respective window. The results are summarized in Figure 6. It is clear that *cis*-eQTLs tend to cluster around TSS, and the decay of eQTL signals away from TSS is fast. In particular, Figure 6 suggests that 50% *cis*-eQTL signals are concentrated within 20kb of TSS. Although this phenomenon is well known ([5, 18]), our results display a much cleaner pattern comparing to the previous reports. This is most likely because we considered multiple *cis*-eQTL signals for a target gene, and our use of PIP accounted for the uncertainty of eQTL calls.

We next asked whether eQTLs are enriched for genetic variants disrupting TF binding. Answering this question helps reveal the underlying regulatory logic between transcription factor binding and gene expression ([17, 18, 26, 27]). To this end, we used the base-pair resolution quantitative annotation of binding variants from the extended CENTIPEDE model ([11, 28]). In brief, this annotation was generated starting from DNase-seq data (for LCLs in the ENCODE project) and seed position weight matrices (PWM) models from TRANSFAC (`http://www.gene-regulation.com/pub/databases.html`) and JASPAR (`http://jaspar.genereg.net/`) databases describing the sequences that the TFs prefer to bind. CENTIPEDE was then used to fit a mixture model that learns for each TF motif:

1. a characteristic footprint shape that evidences active binding
2. a re-calibrated sequence PWM model that is useful to predict the the impact of a genetic variant may have on binding

Each genetic variant from the 1000 Genomes project was annotated as one of the following three mutually exclusive categories:

- SNPs that are not located in any DNaseI footprint region, or in brief, *baseline SNPs*
- SNPs that are in a footprint region but predicted to have little or no impact on TF binding, or in brief, *footprint SNPs*
- SNPs that are in a footprint region and predicted to strongly affect TF binding, or in brief, *binding variants*

About 4% of the total 6.7 million interrogated *cis*-SNPs are annotated as binding variants, and another 4% are annotated as footprint SNPs. Additionally, we controlled for SNP positions with respect to TSS in the model. Our main findings from this analysis are:

- binding variants are 1.49 fold (with 95% confidence interval $[1.38, 1.63]$) more likely than baseline SNPs to be eQTLs, its enrichment in eQTLs is statistically highly significant (p-value $= 4.93 \times 10^{-22}$)
- footprint SNPs are 1.15 fold (with 95% confidence interval $[1.04, 1.27]$) enriched in eQTLs, comparing to baseline SNPs, with the corresponding p-value $= 0.0035$

Our results implies that the sequence variants that potentially disrupt TF binding are more likely to have a functional impact on gene expression. In comparison, footprint SNPs (those predicted to not affect binding) are less likely to affect expression and this is reflected with a relatively lower level of enrichment for eQTLs.

Given the results of the enrichment analysis, the corresponding annotation information can be quantitatively incorporated into the fine mapping analysis. More specifically, we plugged in the point estimates of the enrichment parameters to re-compute the $\Pr(\gamma_j)$ using the logistic model (4) for each gene-SNP pair, and re-ran the MCMC algorithm with the updated priors. As a result, the priors for binding variants were up-weighted comparing to the nearby baseline and footprint SNPs. Because the computation of likelihood function was intact, the PIPs for binding variants were accordingly up-weighted. Using this approach, it becomes plausible to quantitatively distinguish SNPs in perfect LD but belonging to different annotation categories. Take Gene *LY86* (Ensemble ID: ENSG00000112799) as an example (Figure

7). Before incorporating any annotation information, each *cis*-SNP was equally weighted *a priori* and the multiple *cis*-eQTL analysis identified three independent signals, two of which were represented by clusters of highly correlated SNPs. Utilizing the quantitative annotation information, the fine mapping procedure still confidently identified the three original eQTL signals, however two binding variants became the top associated SNPs (ranked according to PIPs) in each respective cluster of SNPs that were indistinguishable in the original analysis (Figure 7).

In the set of 526 genes that we confidently identified as harboring only one single *cis*-eQTL signal, when applying the equal prior without any annotation information, we found that that binding variants and footprint variants are top associated SNPs in 11% and 8% of the genes (or 60 and 43 genes in number), respectively. Using the priors based on quantitative binding annotation and distance to TSS, the percentages of genes with binding variants and footprint SNPs as top associated SNPs increase to 16% and 11% (or 85 and 57 genes in number), respectively.

3 Discussion

In this paper, we have presented a systematic approach to jointly analyze eQTL data from cross-population samples. Our core statistical methods are built on the established Bayesian framework for association analysis of genetic data from heterogeneous groups [14, 15, 21]. For the first time, we have applied them for multiple SNP analysis in a cross-population meta-analytic setting: with a combined sample size ~ 400 in the GEUVADIS data, we have demonstrated that multiple independent *cis*-eQTL signals can indeed be effectively identified in many genes.

The commonly applied strategy for mapping additional association signals is to conduct conditional analysis, which can be viewed as a step-wise variable selection algorithm. One of the notable disadvantages of conditional analysis is that it only reports a single "best" model in the end, and completely ignores its uncertainty. From many of our examples shown in this paper, it is clear that in most cases even we are relatively certain about the number of eQTL signals in a *cis* region, there is typically a great deal of uncertainty in determining the true causal SNPs. As a consequence, the posterior probability of the best model can be unimpressive, and it is potentially dangerous to only use the information from the "best" model in downstream functional analysis, if care is not taken. In comparison, our Bayesian analysis provides much more comprehensive information that fully conveys the uncertainty of the inference result, and the quantified uncertainty information is naturally propagated in our functional analysis.

In this paper, we have employed a two-step procedure that first screens eGenes by performing gene-level hypothesis testing then carries out multiple SNP analysis for identified eGenes. This procedure is analogous to the fine mapping procedure that is commonly used in GWAS, where interesting loci are ranked and selected by single SNP association testing before an in-depth analysis focusing on each flagged high priority locus. We find this procedure not only yields considerable computational savings, but also provides a sound argument to justify our prior specifications for $\Pr(\boldsymbol{\gamma}_j)$. Indeed, this procedure is completely general and can be extended to the fine mapping analysis where subgroups of eQTL data are formed by different tissues ([4,14]) or cellular conditions ([6,8]). Comparing to the meta-analytic setting we have encountered in this paper, the parameter space of $\{\boldsymbol{\gamma}_j\}$ in those applications is more complicated (which includes 2^s potential values, where s is the number of subgroups/tissues). Nevertheless, [14] has provided a principled way to "learn" the priors on possible values that $\boldsymbol{\gamma}_j$ can take by pooling information across genes through a hierarchical model.

Our fine mapping results of eQTLs also demonstrate the benefit of utilizing cross-population samples in genetic association studies. Most importantly, the population heterogeneity of local LD patterns serves as an efficient filter that narrows down the regions harboring casual eQTLs. Nevertheless, varying LD patterns can cause some SNPs to display large degree of heterogeneity across populations in their estimated effect sizes from *single SNP analysis*: in the extreme cases, a SNP may appear to possess strong "population specific" effects. As we acknowledge that genuine population specific eQTLs are certainly interesting phenomena and very much likely exist, we suggest *interpreting* highly heterogeneous eQTL signals from single SNP analysis with caution. It may be necessary to carry out multiple SNP analysis, as we have demonstrated in this paper, to simply rule out the possibility that the seemingly population specific effects are artifacts due to varying LD patterns.

Our Bayesian inference framework naturally incorporates functional annotations in fine mapping eQTLs across population groups. This feature allows us to quantitatively evaluate the enrichment of certain functional feature in eQTLs, and in turn to use the quantified enrichment information to prioritize annotated SNPs for fine mapping analysis. Overall, our model for enrichment testing is similar to what were proposed in [17,18]. However, these previous approaches make simplifying assumption to restrict at most one *cis*-eQTL per gene, such that single SNP association results can be directly used. Our method relaxes this assumption and is fully integrated into our multiple SNP analysis procedure. In addition, our use of CENTIPEDE annotation to examine the relative enrichment of binding variants and footprint SNPs is also novel. Although it is largely expected that binding variants are enriched in eQTLs, it is important to note that the level of enrichment for footprint SNPs is much lower than that for binding variants. Interestingly, this finding seems concurring with the results reported by [28] where the relative enrichment of binding variants vs. footprint SNPs in other cellular and organismal phenotype QTLs is examined.

4 Materials and Methods

The computational methods used in the analysis are implemented in the software packages FM-eQTL (manuscript in prep) and eQTLBMA ([14]). They are freely available at `https://github.com/timflutre/eqtlbma` and `https://github.com/xqwen/fmeqtl`

The fine mapping results of all 6,555 identified eGenes, including scatter plots of PIPs, forest plots of top models, and detailed summaries from separate linear regression analysis are made available on the website `http://www-personal.umich.edu/~xwen/geuvadis/fm_rst_html/`

4.1 Data Pre-processing

The genotype and RNA-seq data were directly downloaded from the GEUVADIS project website. We selected 420 samples whose genotypes are directly measured in the 1000 Genomes project. The samples are evenly distributed in the five population groups, and the detailed breakdown of samples by population is shown in Table 1.

For the RNA-seq data, we used a slightly more stringent threshold than the original analysis to select genes that are expressed in all five populations. Specifically, for each selected gene, we required $> 90\%$ individuals in each population group to have RPKM ≥ 0.1. From the 17,361 Ensemble genes that passed this filter, following the original analysis, we selected a subset of 11,838 genes consisting of annotated protein-coding genes and lincRNAs according to GENCODE ([29]) release 17. We log transformed the RPKM values, and used the pipeline employed in [27,31] to remove the effect of GC content on expression measurements. We then followed the same strategy as described in [20] to remove latent confounding factors using the software package PEER ([30]). However, unlike the original analysis, we ran PEER for each population group separately. In the end, we removed 15, 13, 15, 20 and 20 PEER factors for samples from YRI, CEU, TSI, GBR and FIN, respectively. Finally, the expression levels of each gene were quantile normalized across individuals separately in each population group.

For genotype data, we filtered out SNPs whose sample allele frequencies < 0.03 in the *overall* samples across population groups. Note, we did not apply the allele frequency filter in each population group. The SNPs passing this filter must have sample allele frequencies ≥ 0.03 in at least one population group. In general, as discussed in Method section, the rare SNPs do not impose any statistical or computational problems for our analysis. Nevertheless, removing SNPs that are not likely informative in *any* population group helps improve computational efficiency. Following [17, 31], we defined a 200kb *cis* region for each gene centered at its TSS. In total, the final data set contains 6.7 million gene-SNP pairs.

4.2 Gene-level Analysis of *cis*-eQTLs

For each gene, we tested the null hypothesis which asserts no *cis*-eQTLs. More specifically, we adopted the Bayesian hypothesis testing procedure discussed in [14]. Essentially, [14] assumed a Bayesian model that is mostly similar to (1), except for an additional simplifying assumption, "at most one *cis*-eQTL per gene". Given a gene with p *cis*-sNPs, this additional assumption reduces the possible alternative scenarios into p single SNP association models, for which a gene-level Bayes factor can be easily computed analytically using Bayesian model averaging. In the context of eQTL mapping in multiple tissues, [14] considered all possible configurations, i.e.,$\boldsymbol{\gamma}_j$ values, for each assumed associated SNP, whereas in our analysis we only allowed $\boldsymbol{\gamma}_j \in \{\mathbf{0}, \mathbf{1}\}$ for the reasons discussed in Results section.

Briefly, for the alternative model where the j-th SNP is assumed the lone eQTL, the linear model (1) is simplified to

$$\boldsymbol{y}_i = \mu_i \mathbf{1} + \beta_{i,j} \boldsymbol{g}_{i,j} + \boldsymbol{e}_i, \;\; \boldsymbol{e}_i \sim \mathrm{N}(\mathbf{0}, \sigma_i^2 \boldsymbol{I}), \;\; i = 1, ..., s. \tag{6}$$

We model the correlation of genetic effects, $\beta_{i,j}$'s across population groups through the following prior specification,

$$\begin{aligned} \beta_{i,j} &\sim \mathrm{N}(\bar{\beta}_j, \phi^2), \\ \bar{\beta}_j &\sim \mathrm{N}(0, \omega^2). \end{aligned} \tag{7}$$

Equivalently, the above prior can be represented by a multivariate normal distribution by integrating out the average effect parameter $\bar{\beta}_j$, i.e.,

$$\boldsymbol{\beta}_j = \begin{pmatrix} \beta_{1,j} \\ \vdots \\ \beta_{S,j} \end{pmatrix} \sim \mathrm{N}(\mathbf{0}, \;\boldsymbol{W}), \tag{8}$$

where the variance-covariance matrix $\boldsymbol{W}$ is given by

$$\boldsymbol{W} = \begin{pmatrix} \phi^2 + \omega^2 & \omega^2 & \cdots & \omega^2 \\ \omega^2 & \phi^2 + \omega^2 & \cdots & \omega^2 \\ \vdots & \vdots & \ddots & \vdots \\ \omega^2 & \omega^2 & \cdots & \phi^2 + \omega^2 \end{pmatrix}.$$

The parameter $\phi^2 + \omega^2$ characterizes the overall genetic effects of SNP j, and $\phi^2/(\phi^2 + \omega^2)$ represents the degree of heterogeneity across population groups. Following [14, 15] we considered the values of $\phi^2 + \omega^2$ from a set $E = \{\phi^2 + \omega^2 : 0.1^2, 0.2^2, 0.4^2, 0.8^2, 1.6^2\}$ which covers a wide range of plausible magnitude of genetic effects. We allowed limited degree of heterogeneity by taking $\phi^2/(\phi^2 + \omega^2)$ values from the set $H = \{\phi^2/(\phi^2 + \omega^2) : 0, 0.1\}$ which reflects our prior belief that effects of genuine genetic association should be highly consistent across population groups. Overall, we considered a combination of $|E| \times |H|$ grid for (ϕ^2, ω^2) values for each alternative model. Given this model, a single SNP Bayes factor BF_j can be analytically evaluated following [14, 15], and the corresponding gene-level Bayes factor is obtained by Bayesian model averaging as

$$\mathrm{BF}_{\text{gene}} = \frac{1}{p} \sum_{j=1}^{p} \mathrm{BF}_j. \tag{9}$$

Upon obtaining the gene-level Bayes factors, we used the methods implemented in the software package eQTLBMA ([14]) to select eGenes at FDR 5% levels. eQTLBMA implements two types of FDR control procedures: one is a permutation based procedure which converts a gene-level Bayes factor to a corresponding p-value and control FDR using Storey's procedure; the other procedure is based on the EBF procedure discussed in [32] which directly works with gene-level Bayes factors and avoids any permutations. We found that the latter approach is much more computationally efficient, however slightly conservative. The results reported in Results section are based on the EBF procedure.

4.3 Multiple SNP Analysis of *cis*-eQTLs

In our multiple SNP analysis, we no longer assume "one *cis*-eQTL per gene", and consider the full range of alternative scenarios described by model (1). To make joint inference with respect to $\mathbf{\Gamma} = \{\boldsymbol{\gamma}_i, \dots, \boldsymbol{\gamma}_p\}$, we further specify effect size distribution,

$$\boldsymbol{\beta}_j \,|\, \boldsymbol{\gamma}_j = \mathbf{1}, \phi^2, \omega^2 \ \sim \mathrm{N}(\mathbf{0}, \boldsymbol{W}), \tag{10}$$

where $\boldsymbol{W}$ is constructed in the same way as in the gene-level analysis. In particular, we used the same set of grid values of (ϕ^2, ω^2) in both analyses. Unconditional on $\boldsymbol{\gamma}_j$, the prior on $\boldsymbol{\beta}_j$ is a type of "spike-and-slab", where the "slab" is represented by a mixture of multivariate normal distribution, and the vast majority of the prior probability mass, $1 - \frac{1}{p}$, is assigned to the "spike" (i.e., a point mass at 0).

The described linear model system including the prior specification (i.e., SSLR), is a special case of the general system considered by [15]. Given a specified $\mathbf{\Gamma}$ value, a Bayes factor, $\mathrm{BF}(\mathbf{\Gamma})$, contrasting to the trivial null model, $\mathbf{\Gamma} \equiv \mathbf{0}$, can be analytically approximated by applying the result discussed therein. With this result, the posterior probability, $\Pr(\mathbf{\Gamma} \,|\, \boldsymbol{Y}, \boldsymbol{G})$, can be computed up to an unknown normalizing constant, i.e., $\Pr(\mathbf{\Gamma} \,|\, \boldsymbol{Y}, \boldsymbol{G}) \propto \Pr(\mathbf{\Gamma})\,\mathrm{BF}(\mathbf{\Gamma})$. We then implemented an Metropolis-Hastings algorithm, similar to what is discussed in [15] for multivariate linear regression model (MVLR), to efficiently traverse the space of 2^p possible $\mathbf{\Gamma}$ values. In particular, we implemented a novel proposal distribution that utilizes marginal and conditional analysis results to prioritize SNPs with strong marginal or conditional association signals. In practice, we observed that the resulting MCMC algorithm achieves fast mixing. The details of the algorithm is provided in the supplementary material. In the end, the MCMC algorithm yields a set of $\mathbf{\Gamma}$ samples from the posterior distribution, from which we computed the PIP for each SNP by marginalization.

The posterior expected number of independent *cis*-eQTLs and its variance for each gene was obtained by computing the sample mean and variance of the number of non-zero $\boldsymbol{\gamma}_j$'s in each posterior model. Equivalently, the posterior expected number of *cis*-eQTLs can be computed by the sum of PIPs, i.e.,

$$\mathrm{E}\left[\sum_{j=1}^{p} \mathbf{1}(\boldsymbol{\gamma}_j \neq \mathbf{0}) \,\middle|\, \boldsymbol{Y}, \boldsymbol{G}\right] = \sum_j \Pr(\boldsymbol{\gamma}_j = \mathbf{1} \mid \boldsymbol{Y}, \boldsymbol{G}). \tag{11}$$

For the fine mapping analysis of the GEUVADIS data, the MCMC algorithm was applied for each identified eGene individually. We carried out 25,000 burnin steps and 50,000 repeats for each MCMC run. Taking advantage of parallel processing, we performed multiple-SNP analysis for multiple genes simultaneously in a distributed computing environment, which greatly reduced the overall computational time.

4.4 Enrichment Analysis of *cis*-eQTLs

We specify the prior distribution of $\Pr(\boldsymbol{\gamma}_j)$ by the parametric function (4) to incorporate genomic annotations into the eQTL mapping, whereas the other parts of the Bayesian model remains intact. For the g-th gene and its j-th SNP, we re-write (4) using an equivalent vector form as follows

$$\operatorname{logit}\left[\Pr(\boldsymbol{\gamma}_j^g = 1)\right] = \boldsymbol{\alpha}' \boldsymbol{\delta}_j^g, \tag{12}$$

where the additional super-script for gene emphasizes the annotation data are specific to each gene-SNP pair, and the enrichment parameter $\boldsymbol{\alpha}$ is assumed to be shared across all genes.

Let $\boldsymbol{D}^g$ denote the collection of the annotation data for gene g. For a total of q genes, it follows from the Bayes rule that

$$\Pr\left(\boldsymbol{\alpha}, \boldsymbol{\Gamma}^1, ..., \boldsymbol{\Gamma}^q \,|\, \{\boldsymbol{Y}^1, \boldsymbol{G}^1, \boldsymbol{D}^1\} \ldots \{\boldsymbol{Y}^q, \boldsymbol{G}^q, \boldsymbol{D}^q\}\right) \propto P(\boldsymbol{\alpha}) \prod_{g=1}^{q} \left[\Pr(\boldsymbol{\Gamma}^g \,|\, \boldsymbol{D}^g, \boldsymbol{\alpha})\, \mathrm{BF}(\boldsymbol{\Gamma}^g)\right]. \tag{13}$$

Given a prior specification $P(\boldsymbol{\alpha})$ (e.g. a flat prior), the quantities on the right hand side are individually straightforward to compute (analytically). It is conceptually easy to modify the MCMC algorithm outlined above to jointly sample $(\boldsymbol{\alpha}, \boldsymbol{\Gamma}^1, \ldots, \boldsymbol{\Gamma}^q)$. Nevertheless, due to the extremely high dimensionality of the target space (which is approximately the number of gene-SNP pairs, in the case of the GEUVADIS data ~ 6.7 million), the convergence of the MCMC within a reasonable time frame may be in doubt. Furthermore, the computational resources, especially the memory usage, demanded by the MCMC algorithm may be too high to afford in a practical setting.

Alternatively, we consider an EM algorithm that focuses on finding the MLE of $\boldsymbol{\alpha}$ by treating $(\boldsymbol{\Gamma}^1, ..., \boldsymbol{\Gamma}^q)$ as missing data. The derivation of the EM algorithm is mostly straightforward, we give the relevant details in the supplementary materials. Briefly at $(t+1)$-th iteration, in the Expectation step (E-step), we evaluate $\mathrm{E}\left[\boldsymbol{\gamma}_j^g \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)}\right]$ for each gene-SNP pair given the current estimate $\boldsymbol{\alpha}^{(t)}$. It should be noted that this conditional expectation is exactly the PIP of each gene-SNP pair which can be obtained by running the multiple-SNP analysis for each gene *separately* given the hyperparameter $\boldsymbol{\alpha}^{(t)}$. In the Maximization step (M-step), we maximize $\boldsymbol{\alpha}$ by fitting a logistic regression model relating PIP of each gene-SNP pair to the genomic annotations of the corresponding SNP. Overall, we describe the complete algorithm as an "MCMC-within-EM" algorithm, which is initiated at some arbitrary $\boldsymbol{\alpha}^{(0)}$ value, and iteratively performs multiple *cis*-eQTL mapping using the MCMC algorithm and maximization by fitting logistic regression models until convergence. The main computational benefit of the "MCMC-within-EM" algorithm is that the E-steps involving MCMC runs can be processed in parallel on a distributed computing system, hence the memory requirement is much relaxed.

The approximate enrichment analysis procedure introduced in the Results section can be viewed as a special case of the MCMC-within-EM algorithm with a *single* iteration initiated at $\boldsymbol{\alpha}^{(0)} = (\alpha_0, 0, \ldots, 0)$. In particular, α_0 is computed by (5), and this starting point represents the global null model of no enrichment of any annotations in consideration. This one-step maximization approach is statistically valid: under the null model, the likelihood function is maximized at $\boldsymbol{\alpha}^{(0)}$ with respect to the enrichment parameters, whereas under the alternative, the corresponding parameters should be identified *away from 0* in the maximization step. Analogous to the score test, the approximate procedure is most powerful if the true alternative is close to the null model, i.e., the magnitude of the enrichment is relatively small . Nonetheless in general, the absolute values of the estimated enrichment parameters from this procedure represent a lower bound of the true absolute values of MLEs. The most attractive property of this approximate enrichment procedure is probably its computational cost, as it further simplifies the complete MCMC-within-EM procedure.

Acknowledgments

We thank Matthew Stephens for helpful discussion, and our funding agency for support.

References

1. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS genetics 6: e1000895.

2. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genetics 6: e1000888.

3. Hao K, Bossé Y, Nickle DC, Paré PD, Postma DS, et al. (2012) Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS genetics 8: e1003029.

4. GTEx Consortium (2013) The genotype-tissue expression (gtex) project. Nature Genetics 45: 580–585.

5. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, et al. (2009) Common regulatory variation impacts gene expression in a cell type–dependent manner. Science 325: 1246–1250.

6. Maranville JC, Luca F, Richards AL, Wen X, Witonsky DB, et al. (2011) Interactions between glucocorticoid treatment and cis-regulatory polymorphisms contribute to cellular response phenotypes. PLoS genetics 7: e1002162.

7. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, et al. (2012) Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. Proceedings of the National Academy of Sciences 109: 1204–1209.

8. Raj T, Rothamel K, Mostafavi S, Ye C, Lee MN, et al. (2014) Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 344: 519–523.

9. Brown CD, Mangravite LM, Engelhardt BE (2013) Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs. PLoS Genetics 9: e1003649.

10. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature 447: 799–816.

11. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, et al. (2011) Accurate inference of transcription factor binding from dna sequence and chromatin accessibility data. Genome research 21: 447–455.

12. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, et al. (2012) Integrative annotation of chromatin elements from encode data. Nucleic acids research : gks1284.

13. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, et al. (2013) A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome research 23: 716–726.

14. Flutre T, Wen X, Pritchard J, Stephens M (2013) A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genetics 9: e1003486.

15. Wen X (2014) Bayesian model selection in complex linear systems, as illustrated in genetic association studies. Biometrics 70: 73–83.

16. Sul JH, Han B, Ye C, Choi T, Eskin E (2013) Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS genetics 9: e1003491.

17. Gaffney DJ, Veyrieras JB, Degner JF, Pique-Regi R, Pai AA, et al. (2012) Dissecting the regulatory architecture of gene expression QTLs. Genome Biol 13: R7.

18. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS genetics 4: e1000214.

19. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, et al. (2009) Learning a prior on regulatory potential from eQTL data. PLoS genetics 5: e1000358.

20. Lappalainen T, Sammeth M, Friedländer MR, T Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506–511.

21. Wen X, Stephens M (2014) Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions. The Annals of Applied Statistics 8: 176–203.

22. Morris AP (2011) Transethnic meta-analysis of genomewide association studies. Genetic epidemiology 35: 809–822.

23. Marigorta UM, Navarro A (2013) High trans-ethnic replicability of gwas results implies common causal variants. PLoS genetics 9: e1003566.

24. Li G, Shabalin AA, Rusyn I, Wright FA, Nobel AB (2013) An empirical bayes approach for multiple tissue eQTL analysis. arXiv preprint arXiv:13112948 .

25. 1000 Genomes Project Consortium, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

26. Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends in Genetics 24: 408–415.

27. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390–394.

28. Moyerbrailean GA, Harvey CT, Kalita CA, Wen X, Luca F, et al. (2014) Are all genetic variants in dnase i sensitivity regions functional? bioRxiv : 007559.

29. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) Gencode: the reference human genome annotation for the encode project. Genome research 22: 1760–1774.

30. Stegle O, Parts L, Piipari M, Winn J, Durbin R (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature protocols 7: 500–507.

31. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with rna sequencing. Nature 464: 768–772.

32. Wen X (2013) Robust bayesian FDR control with bayes factors. arXiv preprint arXiv:13113981 .

33. Guan Y, Stephens M, et al. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics 5: 1780–1815.

34. Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70: 849–911.

Supplementary Materials

MCMC Algorithm for Mapping Multiple *cis*-eQTLs

We implement a Metropolis-Hastings algorithm to perform posterior sampling based on equation (3) in the main text. The algorithm is mostly straightforward, to help the Markov chain achieve fast mixing, we implement a novel proposal distribution based on the result of conditional analysis of multiple *cis*-eQTLs.

We propose two simple "local" moves in the MCMC simulations:

1. Change a $\boldsymbol{\gamma}_j$ value for SNP j
2. Swap the values of $\boldsymbol{\gamma}_j$ and $\boldsymbol{\gamma}_k$, for SNPs j and k

where each SNP j is proposed according to a pre-calculated weight w_j. The novelty of the proposal distribution is that we construct the weights w_j's based on the conditional analysis results. More specifically, we start by computing Bayes factors for each *cis*-SNP in a single SNP analysis, and compute a quantity

$$p_j^{(1)} = \frac{\mathrm{BF}_j}{\sum_j^p \mathrm{BF}_j}. \tag{14}$$

(Note that $p_j^{(1)}$ is proportional to the PIP for SNP j assuming only one eQTL in the *cis* region and a uniform prior inclusion probability). We then find the SNP with the maximum $p_j^{(1)}$ value, say SNP k. In the next round, we control for the genotype of SNP k and repeat the single SNP analysis to obtain $p_j^{(2)}$, which mimics the conditional analysis of secondary *cis*-eQTL signals. Note that SNP k and the SNPs in LD will have single SNP Bayes factor close to 1 in this round. We again add the SNP with the maximum $p_j^{(2)}$ value into the control set. We repeat this procedure, with one additional SNP added into the control set in each round, until the maximum single SNP Bayes factor falls below a pre-defined threshold (we use 10 in practice). Suppose the procedure ends in t iterations, we then compute the weight for each SNP using

$$w_j = \sum_{r=1}^{t} \theta_r p_j^{(r)} + \theta_{t+1} \frac{1}{p}, \tag{15}$$

where $\theta_1, ..., \theta_{t+1}$ forms a decreasing geometric series summing up to 1. The trailing $\frac{1}{p}$ term in the weight calculation represents a uniform distribution on candidate *cis*-SNPs.

This particular proposal distribution is an extension of what is used in [33], and should be credited to Matthew Stephens (personal communication). Its theoretical backend is related to *sure-independence screening* proposed by [34] in variable selection context.

Maximum Likelihood Inference of Enrichment Parameters

This section gives the technical details of MCMC-within-EM algorithm. Given the hierarchical model described in the main text, we are interested in performing maximum likelihood inference of enrichment parameter $\boldsymbol{\alpha}$. Treating $\{\boldsymbol{\Gamma}^1, ...\boldsymbol{\Gamma}^q\}$ across all q genes as missing data, the complete data likelihood can be written as

$$P(\{\boldsymbol{Y}^g\}, \{\boldsymbol{\Gamma}^g\} \mid \{\boldsymbol{G}^g\}, \{\boldsymbol{D}^g\}, \boldsymbol{\alpha}) = \prod_{g=1}^{q} P(\boldsymbol{\Gamma}^g \,|\, \boldsymbol{D}^g, \boldsymbol{\alpha}) \cdot \prod_{g=1}^{q} P(\boldsymbol{Y}^g | \, \boldsymbol{\Gamma}^g, \boldsymbol{G}^g). \tag{16}$$

We apply an EM algorithm to find the MLE of $\boldsymbol{\alpha}$. Because vector $\boldsymbol{\gamma}_j^g$ only takes values in $\{\mathbf{0}, \mathbf{1}\}$, using a loose notation, we represent vectors $\mathbf{0}$ and $\mathbf{1}$ with the corresponding binary scalar values. It then follows that

$$P(\boldsymbol{\Gamma}^g \,|\, \boldsymbol{D}^g, \boldsymbol{\alpha}) = \prod_j \left[\left(\frac{\exp(\boldsymbol{\alpha}' \boldsymbol{\delta}_j^g)}{1 + \exp(\boldsymbol{\alpha}' \boldsymbol{\delta}_j^g)} \right)^{\boldsymbol{\gamma}_j^g} \left(\frac{1}{1 + \exp(\boldsymbol{\alpha}' \boldsymbol{\delta}_j^g)} \right)^{1-\boldsymbol{\gamma}_j^g} \right]. \tag{17}$$

The complete data log-likelihood is given by

$$\begin{aligned}\log L(\boldsymbol{\alpha}; \{\boldsymbol{Y}^g\}, \{\boldsymbol{\Gamma}^g\}, \{\boldsymbol{G}^g\}, \{\boldsymbol{D}^g\}) = \sum_{g=1}^{q}\sum_{j=1}^{p_g} \boldsymbol{\gamma}_j^g(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g) - \sum_{g=1}^{q}\sum_{j=1}^{p_g} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g)] \\ + \sum_{g=1}^{q} \log[P(\boldsymbol{Y}^g \,|\, \boldsymbol{\Gamma}^g, \boldsymbol{G}^g)]\end{aligned} \tag{18}$$

The EM algorithm initiates by an arbitrary value of $\boldsymbol{\alpha}$, namely, $\boldsymbol{\alpha}^{(1)}$. In the E-step of t-th iteration, we compute

$$\begin{aligned}\mathrm{E}[\log L(\boldsymbol{\alpha}\,; \{\boldsymbol{Y}^g\}, \{\boldsymbol{\Gamma}^g\}, \{\boldsymbol{G}^g\}, \{\boldsymbol{D}^g\}) \mid \{\boldsymbol{Y}^g\}, \{\boldsymbol{G}^g\}, \boldsymbol{\alpha}^{(t)}] = \sum_{g=1}^{q}\sum_{j=1}^{p_g} \mathrm{E}\left(\boldsymbol{\gamma}_j^g \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)}\right)(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g) \\ - \sum_{g=1}^{q}\sum_{j=1}^{p_g} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g)] + \sum_{g=1}^{q} \mathrm{E}\left(\log[P(\boldsymbol{Y}^g \,|\, \boldsymbol{\Gamma}^g, \boldsymbol{G}^g)] \mid \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)}\right)\end{aligned} \tag{19}$$

Note that the last term does not contain parameter $\boldsymbol{\alpha}$. In the M-step of the t-th iteration, we find

$$\boldsymbol{\alpha}^{(t+1)} = \arg\max_{\boldsymbol{\alpha}} \left(\sum_{g=1}^{q}\sum_{j=1}^{p_g} \mathrm{E}\left(\boldsymbol{\gamma}_j^g \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)}\right)(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g) - \sum_{g=1}^{q}\sum_{j=1}^{p_g} \log[1+\exp(\boldsymbol{\alpha}'\boldsymbol{\delta}_j^g)] \right) \tag{20}$$

The objective function in (20) coincides with the log-likelihood function of a logsitic regression model treating each gene-SNP pair as an independent observation, however with the usual binary response variable replaced by the conditional expectations. By this connection, the maximization step can be carried out by fitting the corresponding modified logistic regression model treating conditional expectations as responses (i.e., via an iterative re-weighted least square algorithm). This also implies that in the E-step, it is only required to compute $\mathrm{E}(\left(\boldsymbol{\gamma}_j^g \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)}\right) = \Pr(\boldsymbol{\gamma}_j^g = 1 \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)})$, i.e., the PIP for each gene-SNP pair, which we obtain from the MCMC sampling.

To summarize, we outline the procedure of the MCMC-within-EM algorithm based on the above derivation as follows

1. At $t = 1$, initiate $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(1)}$

2. Compute prior $\Pr(\boldsymbol{\Gamma}^g \mid \boldsymbol{D}^g, \boldsymbol{\alpha}^{(t)})$, and run MCMC algorithm for multiple *cis*-eQTL analysis for each gene g

3. Compute $\Pr(\boldsymbol{\gamma}_j^g = 1 \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)})$ for each gene-SNP pair from the posterior samples

4. Find $\boldsymbol{\alpha}^{(t+1)}$ by fitting a logistic regression model treating $\Pr(\boldsymbol{\gamma}_j^g = 1 \,|\, \boldsymbol{Y}^g, \boldsymbol{G}^g, \boldsymbol{\alpha}^{(t)})$ as response variable and $\{\boldsymbol{D}^g\}$ as observed covariates

5. Repeat 2 to 4 until convergence

Figure Legends

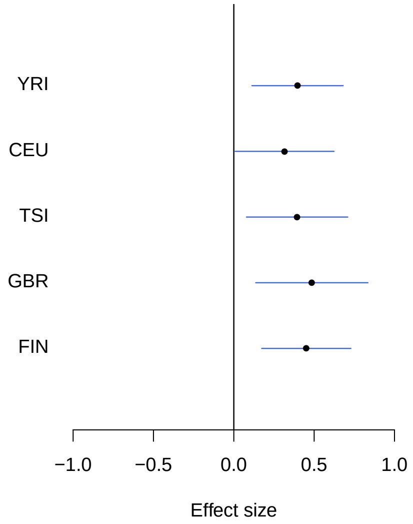


Fig. 1. An example of modest yet consistent eQTL signals across population groups. The forest plot shows the genetic effects of SNP rs7207370 with respect to the expression levels of gene *NME1* (ensemble ID: ENSG00000239672). SNP rs7207370 is one of the top associated *cis*-SNPs in all population groups, yet the strengths of the association signals are modest in all groups (the maximum single SNP Bayes factor, among five groups, is 18.0 in FIN, the corresponding gene-level Bayes factor is 1.8). As a consequence, the gene is not identified as an eGene in any of the separate analyses. Across populations, the SNP exhibit a strongly consistent association pattern. In the cross-population meta-analysis, the gene-level Bayes factor reaches 1.1×10^4 (single SNP Bayes factor for rs7207370 is 9.5×10^5).

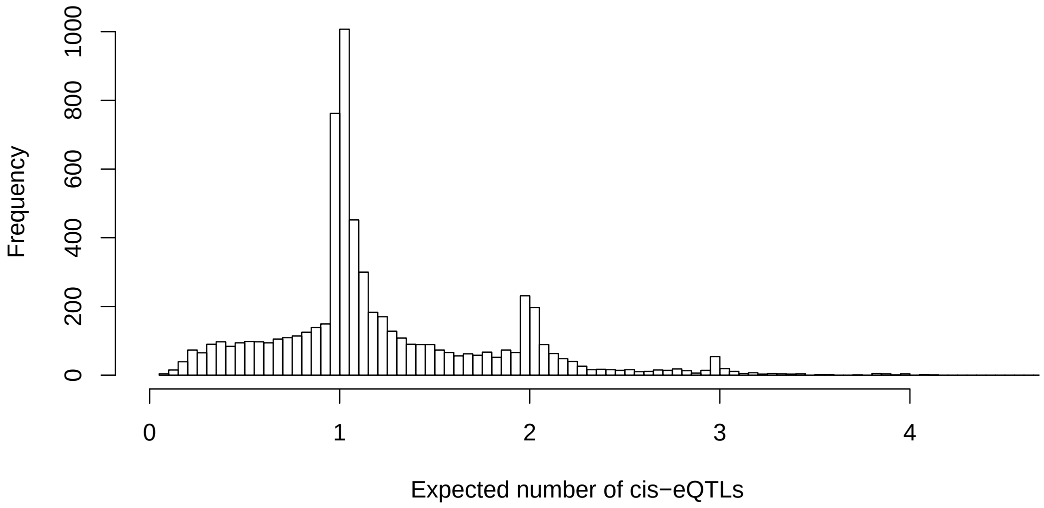


Fig. 2. Histogram of posterior expected number of *cis*-eQTLs in 6,555 identified eGenes. The figure indicates that for most of the eGenes, we identified only single *cis*-eQTLs. However, for a non-trivial proportion of eGenes, multiple independent eQTL signals were identified.

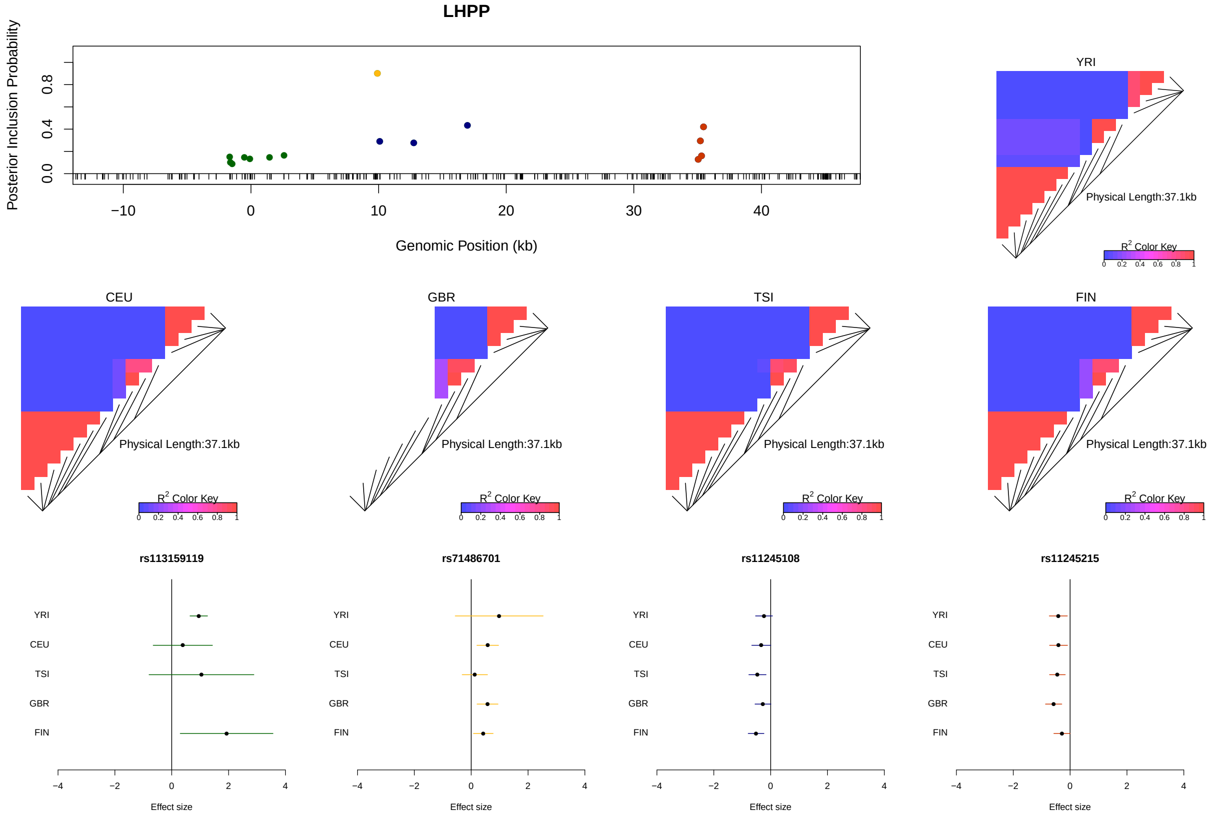


Fig. 3. An example of a gene harboring four independent *cis*-eQTL signals. The top left panel plots the *cis*-SNPs with PIP ≥ 0.02. The locations of the SNPs are with repect to the TSS of gene *LHPP*. The ticks on the x-axis indicate all interrogated *cis*-SNPs in the region. The SNPs with the same color are in high LD and represent the same eQTL signal. In the plot, the sums of the PIPs from the SNPs in the same colors are all ~ 1, indicating that we are confident of the existence of each signal. The heatmaps show the LD patterns in each of the population group. They are qualitatively similar, except that the SNPs representing the first signals are monomorphic in GBR. In the bottom panel, we plot the effect sizes of eQTLs jointly estimated from one of the high posterior probability models. Each of the SNP plotted belongs to a different colored cluster in the PIP plot (as indicated by the color coding of the error bars). The effect sizes and standard errors are estimated from the multiple linear regression models (containing all four SNPs) separately fitted in each population group. All the signals show strong consistency across populations.

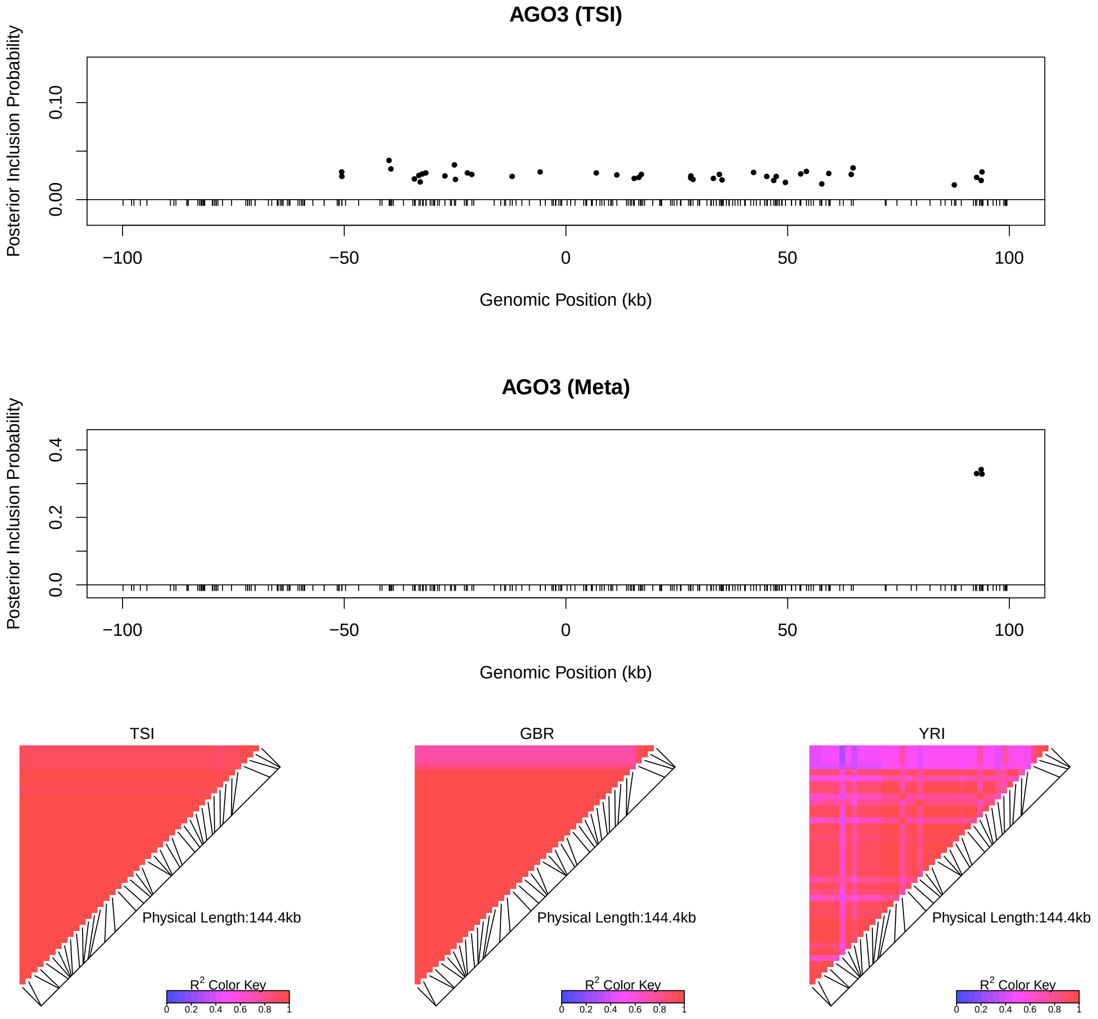
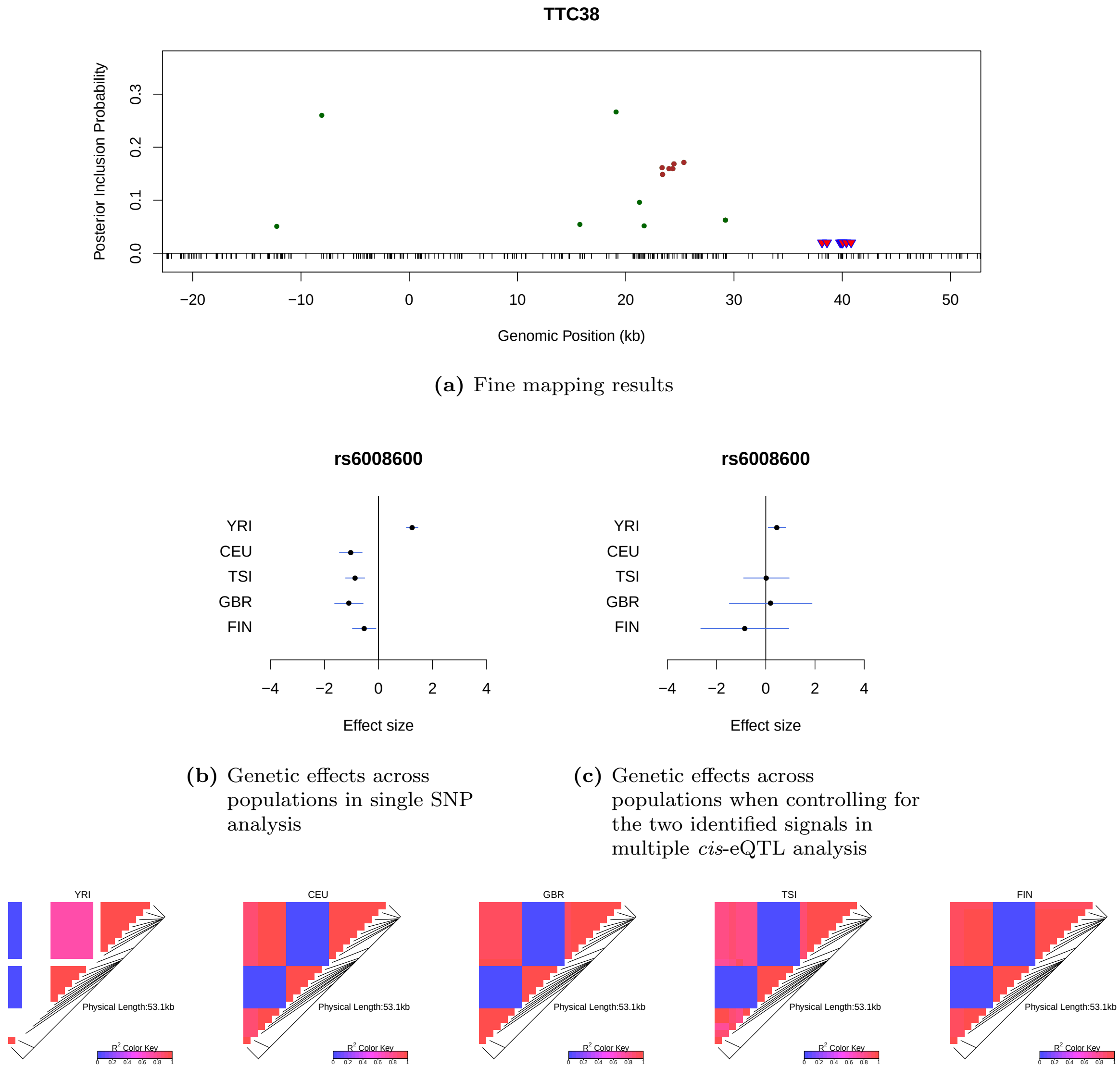


Fig. 4. An example of automatic LD filtering across population groups. The top panel shows the result of multiple *cis*-eQTL analysis for gene *AGO3* using only the data from TSI. The SNPs with PIPs ≥ 0.02 are plotted. All SNPs plotted are in high LD in TSI, and the sum of the PIPs across the genomic region is close to 1. The region spanned by the signals is ~ 140 kb. We repeated the analysis jointly across all five populations, the SNPs with PIPs ≥ 0.02 are plotted in the middle panel. The genomic region harboring the eQTL is narrowed down into a 1.2 kb region enclosed by three SNPs each with PIP ~ 0.33. The bottom panel shows the LD heatmaps between the 41 SNPs plotted in the top panel in TSI, GBR and YRI, respectively. The multiple *cis*-eQTL mapping method takes advantage of the varying LD patterns across populations, and automatically narrows down the region harboring the true causal *cis*-eQTL.

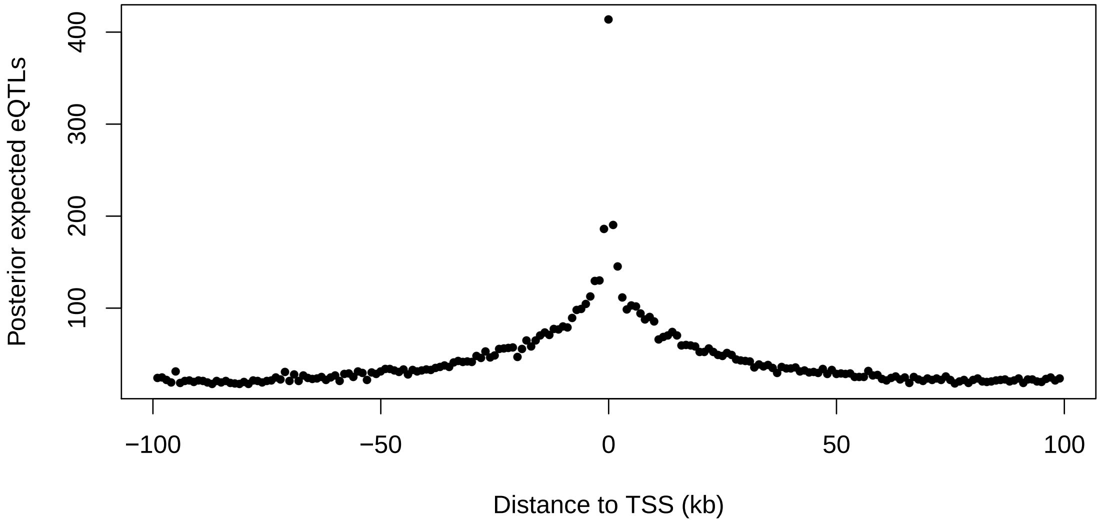


(a) Fine mapping results

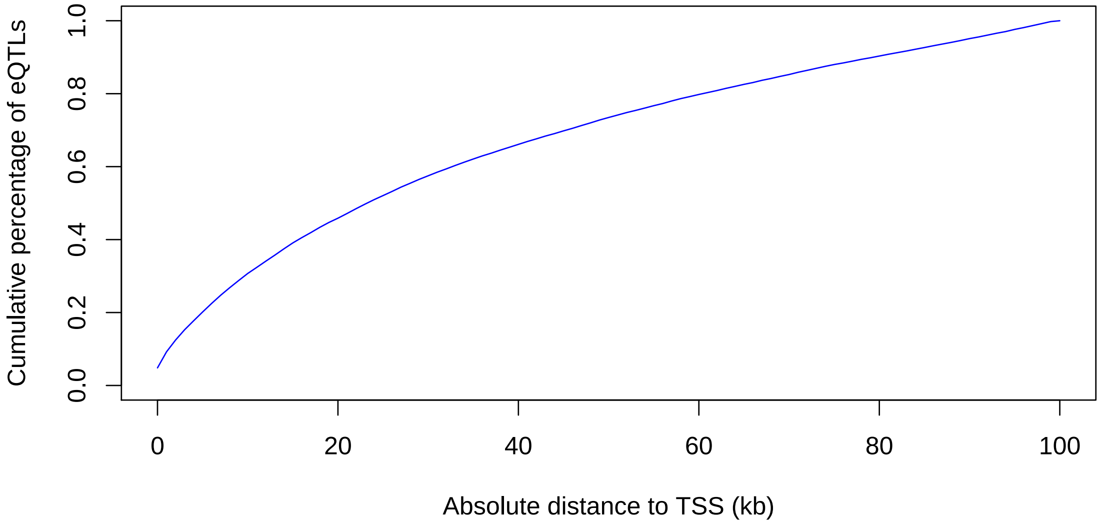
(b) Genetic effects across populations in single SNP analysis

(c) Genetic effects across populations when controlling for the two identified signals in multiple *cis*-eQTL analysis

Fig. 5. Multiple SNP analysis helps explain observed strong heterogeneity in single SNP analysis. SNP rs6006800 and the SNPs in LD (whose genomic locations are indicated by the purple triangels in the top panel) in the *cis*-region of gene *TTC38* shows distinct opposite effects in European and YRI populations when analyzed alone. The middle left panel shows the effect sizes of rs6008600 estimated in the single SNP analysis in the five populations. The top panel shows the multiple *cis*-eQTL analysis result. SNPs with PIPs ≥ 0.02 are plotted. The result suggests that there are two independent signals in the region, with one represented by SNPs colored in green, and the other represented by SNPs colored in brown. The sums of the green and brown SNPs are both very close to 1. SNP rs6008600 and the SNPs in LD all have PIPs ~ 0 in the multiple *cis*-eQTL analysis. The middle right panel shows the effect sizes of rs6006800 estimated from the multiple linear regression models controlling for the two independent signals in each population: the genetic association observed in the single SNP analysis is seemingly "explained away" by the two independent signals identified by the fine mapping analysis. The bottom panel shows LD heatmaps between SNPs highlighted in the top panel (green, brown SNPs and SNPs in LD with rs6006800) in the five populations. Some of the green SNPs are monomorphic in YRI. The opposite effects of rs6006800 is clearly explained by the varying LD patterns: rs6006800 is in high LD with the brown SNPs in YRI, whereas in European populations it tags the green SNPs.

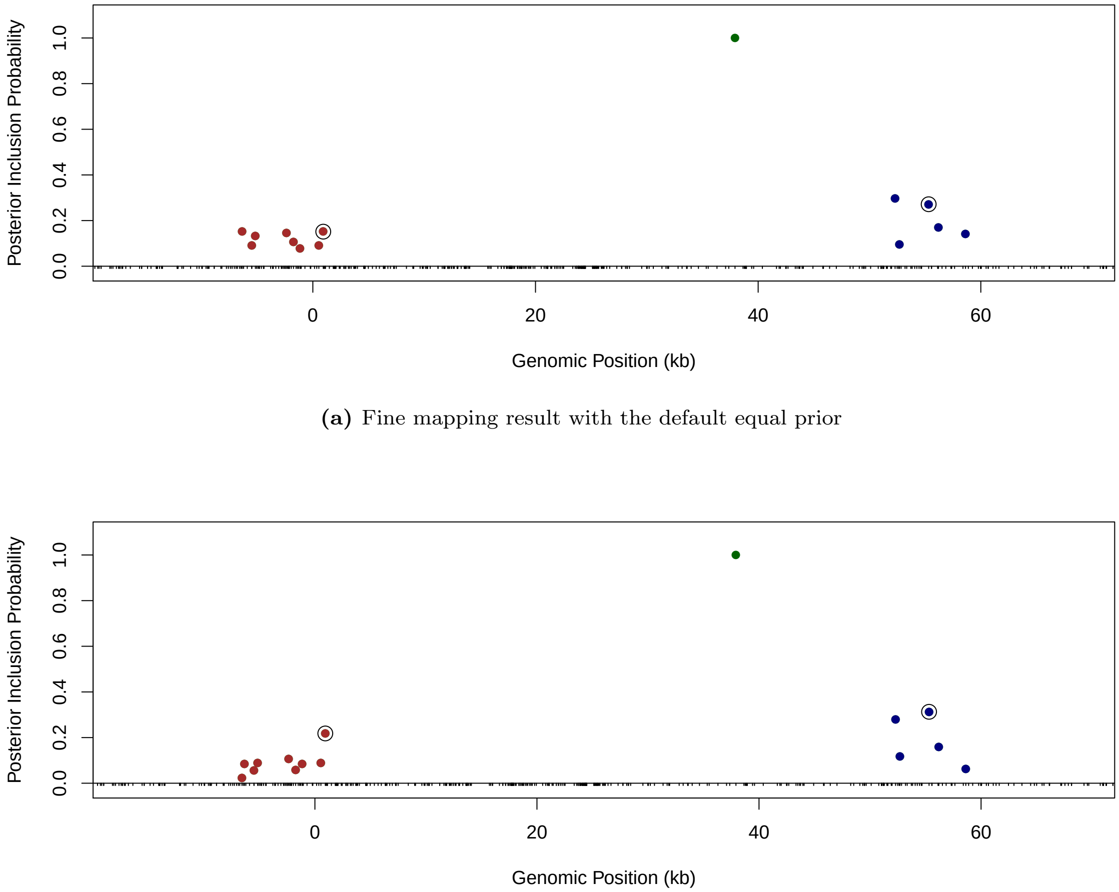


(a) Distribution of posterior expected number of *cis*-eQTLs with respect to distance to TSS



(b) Cumulative percentage of *cis*-eQTLs with respect to distance to TSS

Fig. 6. Enrichment of *cis*-eQTLs with repect to distance to TSS.



(a) Fine mapping result with the default equal prior

(b) Fine mapping result incorporating functional annotations

Fig. 7. Comparison of fine mapping results of gene *LY86* before and after incorporating functional annotations. The top panel is based on the analysis using uniform prior (5). The bottom panel is based on the analysis using prior (4) incorporating SNP distances to TSS and the annotations for binding variants and footprint SNPs. In both panels, SNPs with PIPs ≥ 0.02 are plotted. There are clearly three independent *cis*-eQTLs in the region, represented by different colors of SNPs. SNPs in smae colors are in high LD. The sums of PIPs from SNPs in same colors are all close to 1. The circled points are predicted binding variants. It is clear that binding variants are up-weighted when annotation information is incorporated into the fine mapping analysis.

Tables

Population	Sample Size	Number of eGenes
YRI	77	1042
CEU	78	960
TSI	92	1803
GBR	84	2078
FIN	89	2100
META	-	**6555**

Tab. 1. Comparison of eGene discovery by separate vs. meta analysis. eGenes were declared by rejecting the null hypothesis of no *cis*-eQTLs in the *cis* region at 5% FDR level. The difference of identified eGenes between CEU, YRI and the rest of the European populations in the separate analysis was likely due to the cell line effects (T. Lappalainen, personal communication). Clearly, the meta-analysis improved power of eGene discovery by aggregating samples across population groups.