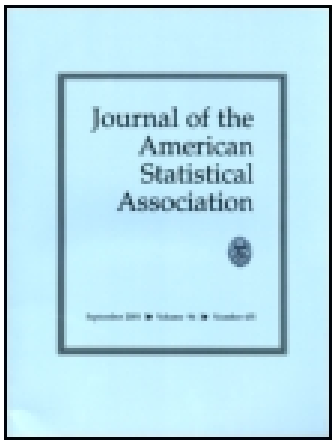




Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

# Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data

Yi-Juan Hu[a], Wei Sun[b], Jung-Ying Tzeng[c] & Charles M. Perou[d]
[a] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322 (Email: )
[b] Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (Email: )
[c] Department of Statistics, North Carolina State University, Raleigh, NC 27695 (Email: )
[d] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599 (Email: )
Accepted author version posted online: 06 May 2015.

To cite this article: Yi-Juan Hu, Wei Sun, Jung-Ying Tzeng & Charles M. Perou (2015): Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data, Journal of the American Statistical Association, DOI: 10.1080/01621459.2015.1038449

To link to this article: http://dx.doi.org/10.1080/01621459.2015.1038449

# Proper Use of Allele-Specific Expression Improves Statistical Power for *cis*-eQTL Mapping with RNA-Seq Data

Yi-Juan HU, Wei SUN, Jung-Ying TZENG, and Charles M. PEROU

Studies of expression quantitative trait loci (eQTLs) offer insight into the molecular mechanisms of loci that were found to be associated with complex diseases and the mechanisms can be classified into *cis*- and *trans*-acting regulation. At present, high-throughput RNA sequencing (RNA-seq) is rapidly replacing expression microarrays to assess gene expression abundance. Unlike microarrays that only measure the total expression of each gene, RNA-seq also provides information on allele-specific expression (ASE), which can be used to distinguish *cis*-eQTLs from *trans*-eQTLs and, more importantly, enhance *cis*-eQTL mapping. However, assessing the *cis*-effect of a candidate eQTL on a gene requires knowledge of the haplotypes connecting the candidate eQTL and the gene, which cannot be inferred with certainty. The existing two-stage approach that first phases the candidate eQTL against the gene and then treats the inferred phase as observed in the association analysis tends to attenuate the estimated *cis*-effect and reduce the power for detecting a *cis*-eQTL. In this article, we provide a maximum-likelihood framework for *cis*-eQTL mapping with RNA-seq data. Our approach integrates the inference of haplotypes and the association analysis into a single stage, and is thus unbiased and statistically powerful. We also develop a pipeline for performing a comprehensive scan of all local eQTLs for all genes in the genome by controlling for false discovery rate, and implement the methods in a computationally efficient software program. The advantages of the proposed methods over the existing ones are demonstrated through realistic simulation studies and an application to empirical breast cancer data from The Cancer Genome Atlas project.

KEY WORDS: ASE; eQTL study; Gene expression; Haplotype; Maximum likelihood.

---

Yi-Juan Hu is Rollins Assistant Professor, Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322 (Email: yijuan.hu@emory.edu). Wei Sun is Associate Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (Email: wsun@bios.unc.edu). Jung-Ying Tzeng is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (Email: jytzeng@ncsu.edu). Charles

M. Perou is May Goldman Shaw Distinguished Professor, Department of Genetics, University of North Carolina, Chapel Hill, NC 27599 (Email: chuck_perou@med.unc.edu). This work was supported by NIH grant R01GM105785 (WS, YJH), P01CA142538-01 (JYT, WS), NCI Breast SPORE program P50CA58223-09A1 (CMP), U24CA143848-05 (CMP), and Breast Cancer Research Foundation (CMP).

## 1. INTRODUCTION

Expression quantitative trait loci (eQTLs) mapping searches across the genome for variants that regulate expression levels of messenger RNAs. Studies of eQTLs can not only discover the genetic basis underlying the variation in gene expression but also decipher molecular mechanisms of loci that were found to be associated with complex diseases by genome-wide association studies (GWAS) (Cookson et al. 2009). At present, high-throughput RNA sequencing (RNA-seq) is rapidly replacing expression microarrays to assess gene expression abundance. To date, The Cancer Genome Atlas (TCGA) project has collected RNA-seq data on >5,000 samples over 17 cancers. These data hold the potential to characterize the functional roles of GWAS hits (Welter et al. 2014), which can later be translated into clinical benefits such as cancer treatments. However, traditional methods for eQTL mapping were developed for microarray expression data (Kendziorski and Wang 2006) and may no longer be appropriate for RNA-seq data.

Compared with microarrays that only measure the total expression of each gene, one important feature of RNA-seq is its ability to measure allele-specific expression (ASE). In a diploid individual, each chromosome consists of a paternal copy and a maternal copy, and each gene comprises a paternal allele and a maternal allele. The expression of each allele of a gene is referred to as its ASE, which has been used to distinguish between *cis*- and *trans*-eQTLs. By definition, a *cis*-eQTL is located on the same chromosome as the target gene (e.g., transcriptional factor binding site of the gene) and the DNA variation on the paternal (or maternal) chromosome only influences the transcription process of the paternal (or maternal) gene allele. By contrast, a *trans*-eQTL can be located anywhere in the genome and influences the transcription of the target gene by modifying the abundance or activity of a protein that regulates the gene; this regulation does not discriminate between the two alleles. Therefore, a *cis*-regulation typically leads to differential ASE between the two alleles of an individual whereas a *trans*-regulation generally results in the same amount of expression change. Such a pattern of ASE can be used to distinguish *cis*-eQTLs from *trans*-eQTLs.

More importantly, the difference in ASE offers another layer of information on top of the total gene expression to enhance *cis*-eQTL mapping.

With RNA-seq data, the ASE can be measured by sequence reads that are unambiguously mapped to each of the two gene alleles. For an individual, the usable reads must overlap with at least one exonic single nucleotide polymorphism (SNP) at which the individual carries a heterozygous genotype. In Figure 1(a), we illustrate the quantification of ASE for a hypothetical gene with two exons and two exonic SNPs. In individual (i), four and two reads harbor alleles A and T, respectively, at the first exonic SNP, and four and two reads harbor alleles A and G, respectively, at the second exonic SNP. Thus, the ASE is estimated to be eight and four for gene alleles containing the haplotypes A-A and T-G, respectively. In individual (ii), only the eight reads overlapping with the first SNP can be counted into the ASE measure. In individual (iii), no read can be counted into ASE.

Using the ASE, *cis*-eQTL mapping assesses whether one allele of a (biallelic) SNP is associated with a higher or lower ASE than the other allele. We use Figure 1(a-b) to illustrate this process. In individual (i), after the haplotypes are inferred as C-A-A and T-T-G and the ASE is quantified as above, alleles C and T of the candidate *cis*-eQTL are linked with the ASE measurements of eight and four, respectively (Figure 1[a]). Because allele C is associated with a higher ASE than allele T, this SNP is likely a *cis*-eQTL (Figure 1[b]). Note that only individuals heterozygous at the candidate eQTL provide information for *cis*-eQTL mapping. The ASE measured as above is only comparable within individuals, with one ASE measurement serving as the control for the other, but not comparable across individuals.

Evidently, the phase information is essential to link the ASE with the alleles of the candidate *cis*-eQTL. Many algorithms (Stephens et al. 2001; Browning and Browning 2009; Li et al. 2010; Delaneau et al. 2012) have been developed to infer the haplotype phase from the genotypes of unrelated individuals. It is often reasonable to assume that the inferred phase is accurate within the gene body, because genes are relatively short (Flicek et al. 2011), and RNA-seq reads that span

over two or more SNPs provide direct information on the phase. However, the phasing accuracy tends to deteriorate when the haplotypes extend to the candidate eQTL, which is often located in an inter-genic region and can be a few hundred kilobases (kb) away from the gene. Incorrectly linking the SNP alleles with the ASE will greatly compromise the power of *cis*-eQTL mapping.

Besides the ASE, the total gene expression, which is measured by the total read count (TReC) (i.e., the total number of reads mapped to the gene), also provides information for *cis*-eQTL mapping. As shown in Figure 1(c), the individual with genotype CC has the highest TReC followed by the individual with CT and then TT, and such a pattern is consistent with the aforementioned observation that allele C is associated with a higher ASE than allele T. Note that the quantification of TReC is not confined to reads that overlap with SNPs showing heterozygous genotypes. Thus, the TReC is generally much greater than the sum of the two ASE measurements and can be compared across individuals.

Traditional methods for eQTL mapping, which were originally developed for microarray expression data, have been applied to the RNA-seq TReC data after normalization (Pickrell et al. 2010). These methods do not exploit the ASE information, ignore the digital nature of the read count data, and cannot distinguish between *cis*- and *trans*-eQTLs. To incorporate both the ASE and TReC data available from RNA-seq for *cis*-eQTL mapping, Sun (2012) proposed a two-stage approach that first phases the candidate eQTL against the target gene and then treats the inferred phase as observed in assessing the *cis*-effect. This approach, pertaining to single imputation in the missing data literature, has two limitations. First, the phasing process disregards the gene expression, which is potentially informative about the missing phase, especially for a true *cis*-eQTL. Second, the downstream association analysis ignores the uncertainties in the inferred phase. Consequently, the two-stage approach tends to attenuate the estimated *cis*-effect and reduce the power for detecting a *cis*-eQTL. A further challenge, which has not been addressed in previous work, remains to develop computationally efficient methods to evaluate the significance of eQTL findings when multiple, correlated SNPs are tested for each gene.

In this article, we propose a maximum-likelihood approach for *cis*-eQTL mapping that incorporates both the ASE and TReC data. Unlike Sun (2012), our approach integrates the phasing of the candidate eQTL and the association analysis into a single stage. Thus, our maximum-likelihood estimators (MLEs) for *cis*-effects are unbiased and statistically efficient in the sense that the corresponding test of association is the most powerful among all valid tests based on the same data and same assumptions. Within the maximum-likelihood framework, we develop a five-step pipeline (Figure 2) for performing the comprehensive analysis of all local eQTLs for all genes in the genome. Here, we restrict the candidate eQTLs to the local SNPs because *cis*-eQTLs are mostly local and the gene body contains little information about the phase of a SNP if the SNP is too distant. The five steps of our pipeline are (1) testing every candidate eQTL for association with the expression of a gene (referred to as a gene-SNP pair) and focusing on the SNP with the minimum *p*-value (referred to as the minimum-*p* SNP) for each gene, (2) assessing the significance of each minimum-*p* SNP, (3) detecting eQTLs among genome-wide minimum-*p* SNPs by controlling for false discovery rate (FDR), (4) distinguishing between *cis*- and *trans*-acting regulations at detected eQTLs, and (5) estimating the effect sizes at detected eQTLs. We implement the proposed methods, including an ultra-fast permutation process for step (2), in a computationally efficient program. We show in realistic simulation studies that the proposed approach is more powerful than the two-stage approach and the use of the TReC data alone. We further demonstrate the usefulness of the new methodology and software program in an application to the breast cancer data from TCGA.



## 2. METHODS

### 2.1 ASE Model

Suppose that each SNP is biallelic with allele values 0 and 1. Thus, each haplotype is a unique sequence of 0s and 1s on one copy of a chromosome. We denote the $\widetilde{K}$ possible haplotypes at

the $\widetilde{M}$ exonic SNPs within a gene by $\widetilde{h}_1, \ldots, \widetilde{h}_{\widetilde{K}}$, and denote the ordered diplotype consisting of haplotypes $\widetilde{h}_m$ and $\widetilde{h}_n$ by $\widetilde{H} = (\widetilde{h}_m, \widetilde{h}_n)$, where $m, n = 1, \ldots, \widetilde{K}$. For the reasons mentioned earlier, we assume that $\widetilde{H} = (\widetilde{h}_m, \widetilde{h}_n)$ is known. For a candidate eQTL, let $G$ be the genotype, which is the number of allele 1, and let $G_1$ be the phase, which is the number of allele 1 on the same chromosomal copy as $\widetilde{h}_m$. The candidate eQTL can be outside or inside the gene body. We focus on the former case, under which $G_1$ is unknown for individuals who are heterozygous at the candidate eQTL; the latter case is trivial as $G_1$ would be known. We use $\widetilde{H}$ to infer $G_1$ by exploiting the linkage disequilibrium (LD) between the exonic SNPs and the candidate eQTL. To this end, we consider the joint distribution of the $M = \widetilde{M} + 1$ SNPs. Let $H = (h_k, h_l)$ be the ordered diplotype at the $M$ SNPs, where $k, l = 1, \ldots, K$ indexing the $K$ possible haplotypes. Write $\pi_k = \Pr(h = h_k)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, where $\boldsymbol{\nu}'$ denotes the transpose of the vector $\boldsymbol{\nu}$. Under Hardy-Weinberg equilibrium (HWE), the frequency of $H = (h_k, h_l)$ is $\pi_k \pi_l$, denoted by $P_{\boldsymbol{\pi}}\{H = (h_k, h_l)\}$. Both $H$ and $G_1$ are unknown. Once $H$ is inferred as $(h_k, h_l)$, $G_1$ is simply the value of $h_k$ in the position of the candidate eQTL.

The ASE data are represented by the number of allele-specific reads mapped to $\widetilde{h}_m$ and the sum of the two allele-specific read counts, denoted by $R_1$ and $R$, respectively. Note that $R_1$ pertains to the same chromosomal copy as $G_1$. Because $R_1$ is not comparable across individuals, we model the conditional probability of $R_1 = r_1$ given $R = r$ by a beta-binomial distribution with a success rate $p_{k,l}$ and a dispersion parameter $\psi$,

$$P_{\beta_A,\psi}(r_1 | r, (h_k, h_l)) = \binom{r}{r_1} \frac{\prod_{j=0}^{r_1-1}(p_{k,l} + j\psi) \prod_{j=0}^{r-r_1-1}(1 - p_{k,l} + j\psi)}{\prod_{j=1}^{r-1}(1 + j\psi)},$$

where

$$p_{k,l} = \begin{cases} \exp(\beta_A)/\{1 + \exp(\beta_A)\} & \text{if } G = 1 \text{ and } G_1 = 1 \\ 1/\{1 + \exp(\beta_A)\} & \text{if } G = 1 \text{ and } G_1 = 0 \\ 0.5 & \text{if } G = 0 \text{ or } 2 \end{cases} \quad (1)$$

with $\beta_A$ characterizing the effect of $G_1$ on $R_1$. The beta-binomial distribution is an extension of the binomial distribution $\text{Bin}(r, p_{k,l})$ to allow for over-dispersion. When there is no allele-specific read, i.e., $r = 0$, we set $P_{\beta_A,\psi}(r_1|r,(h_k,h_l)) = 1$. Note that the covariates (e.g., intercept term, demographical or environmental factors) are ignored, because the comparison is between $R_1$ and $R - R_1$ within an individual so that any covariate effect is canceled out.

Because $H$ (or equivalently, $G_1$) may be missing, the likelihood based on the observed data $(R_{1i}, R_i, G_i, \widetilde{H}_i)$ $(i = 1, \ldots, N)$ involves a summation:

$$L_A(\beta_A, \psi, \boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{H_i \sim (G_i, \widetilde{H}_i)} P_{\beta_A,\psi}(R_{1i}|R_i, H_i) P_{\boldsymbol{\pi}}(H_i), \tag{2}$$

where $H_i \sim (G_i, \widetilde{H}_i)$ denotes the set of $H_i$s that are compatible with $G_i$ and $\widetilde{H}_i$. If $G_i = 1$, such a set consists of two elements, $\{(1, \widetilde{h}_m), (0, \widetilde{h}_n)\}$ and $\{(0, \widetilde{h}_m), (1, \widetilde{h}_n)\}$. If $G_i = 0$ or 2, then $H_i$ is uniquely determined. As noted earlier, only individuals with $G_i = 1$ and $R_i > 0$ provide information about the effect size $\beta_A$. However, all individuals with $R_i > 0$ are used for estimating the over-dispersion parameter $\psi$, and all, including those with $R_i = 0$, are used for estimating the haplotype frequencies in $\boldsymbol{\pi}$. We refer to (2) as the ASE model. The two-stage approach (Sun 2012) can be viewed as a special case of (2) in that the set $H_i \sim (G_i, \widetilde{H}_i)$ only contains the inferred value and $P_{\boldsymbol{\pi}}(H_i)$ factors out.

### 2.2 TReC and TReCASE Models

Denote the TReC of the gene by $T$. We adopt the TReC model of Sun (2012) for assessing the effect of genotype $G$ on $T$. We reformulate his TReC model in this section to conform to the notation in this article. Let $\boldsymbol{X}$ be a set of $p$ covariates including the unit component. The probability of $T = t$ given the genotype $G = g$ and the covariates $\boldsymbol{X} = \boldsymbol{x}$ is formulated through a negative-binomial distribution with mean $\mu_{\boldsymbol{x},g}$ and a dispersion parameter $\phi$:

$$P_{\beta_T,\boldsymbol{\gamma},\phi}(t|\boldsymbol{x}, g) = \frac{\Gamma(t + 1/\phi)}{t!\Gamma(1/\phi)} \left(\frac{1}{1 + \phi\mu_{\boldsymbol{x},g}}\right)^{1/\phi} \left(\frac{\phi\mu_{\boldsymbol{x},g}}{1 + \phi\mu_{\boldsymbol{x},g}}\right)^{t},$$

where $\Gamma(.)$ is the Gamma function. The negative-binomial distribution is a generalization of the Poisson distribution to allow for over-dispersion. The covariate-adjusted effect of $G$ on $T$, characterized by $\beta_{\mathrm{T}}$, is formulated through the regression model

$$\log(\mu_{\boldsymbol{x},g}) = \boldsymbol{\gamma}'\boldsymbol{x} + \mathcal{Z}(g,\beta_{\mathrm{T}}), \tag{3}$$

where

$$\mathcal{Z}(g,\beta_{\mathrm{T}}) = \begin{cases} 0 & \text{if } g = 0 \\ \log\{1+\exp(\beta_{\mathrm{T}})\} - \log 2 & \text{if } g = 1 \\ \beta_{\mathrm{T}} & \text{if } g = 2. \end{cases} \tag{4}$$

The form of $\mathcal{Z}(g,\beta_{\mathrm{T}})$ reflects the additive genetic effect. To see this, we first write $\log(\mu_{\boldsymbol{x},0}) = \boldsymbol{\gamma}'\boldsymbol{x}$ and $\log(\mu_{\boldsymbol{x},2}) = \boldsymbol{\gamma}'\boldsymbol{x} + \beta_{\mathrm{T}}$, from which we obtain $\beta_{\mathrm{T}} = \log(\mu_{\boldsymbol{x},2}/\mu_{\boldsymbol{x},0})$. Then, it follows from the additive effect that $\log(\mu_{\boldsymbol{x},1}) = \log(\mu_{\boldsymbol{x},0} + \mu_{\boldsymbol{x},2}) - \log 2 = \log(\mu_{\boldsymbol{x},0}) + \log(1 + \mu_{\boldsymbol{x},2}/\mu_{\boldsymbol{x},0}) - \log 2$, which satisfies (3) and (4) after replacing $\mu_{\boldsymbol{x},2}/\mu_{\boldsymbol{x},0}$ by $\exp(\beta_{\mathrm{T}})$. The additive model is the most commonly used in eQTL studies. Indeed, a *cis*-effect is additive by definition. Finally, the likelihood based on the data $(T_i, \boldsymbol{X}_i, G_i)$ $(i = 1, \ldots, N)$ takes the form

$$L_{\mathrm{T}}(\beta_{\mathrm{T}}, \boldsymbol{\gamma}, \phi) = \prod_{i=1}^{N} P_{\beta_{\mathrm{T}},\boldsymbol{\gamma},\phi}(T_i|\boldsymbol{X}_i, G_i). \tag{5}$$

When an eQTL *cis*-regulates the gene expression, we have $\beta_{\mathrm{A}} = \beta_{\mathrm{T}}$. To show this, we introduce $\upsilon_1$ and $\upsilon_0$ as the mean ASE corresponding to alleles 1 and 0, respectively, of the candidate eQTL at the baseline of $\boldsymbol{X}$. For individuals with $G = 1$ and $G_1 = 1$, we have $p_{k,l} = \upsilon_1/(\upsilon_1 + \upsilon_0)$ and then $\beta_{\mathrm{A}} = \log\{p_{k,l}/(1 - p_{k,l})\} = \log(\upsilon_1/\upsilon_0)$; for those with $G = 1$ and $G_1 = 0$, we have $p_{k,l} = \upsilon_0/(\upsilon_1 + \upsilon_0)$ and $\beta_{\mathrm{A}} = \log\{(1 - p_{k,l})/p_{k,l}\} = \log(\upsilon_1/\upsilon_0)$. On the other hand, we have $\beta_{\mathrm{T}} = \log(\mu_{\boldsymbol{x},2}/\mu_{\boldsymbol{x},0}) = \log\{2\upsilon_1/(2\upsilon_0)\}$, where the second equation follows from the additivity assumption. Therefore, we obtain $\beta_{\mathrm{A}} = \beta_{\mathrm{T}}$ and denote them by a common effect size $\beta$. To combine the ASE and TReC data for inference on $\beta$, we employ the joint likelihood

$$L_{\mathrm{TA}}(\beta, \boldsymbol{\gamma}, \phi, \psi, \boldsymbol{\pi}) = L_{\mathrm{A}}(\beta, \psi, \boldsymbol{\pi}) L_{\mathrm{T}}(\beta, \boldsymbol{\gamma}, \phi). \tag{6}$$

We refer to (6) as the TReCASE model. Again, the two-stage approach using the combined data is a special case of (6). A *trans*-eQTL implies that $\beta_T \neq 0$ and $\beta_A = 0$. In that case, the TReCASE model is not appropriate because the ASE data provide no information for the association but simply introduce noise. The TReC model alone should be used for *trans*-eQTL mapping.

### 2.3 A Measure of Phasing Accuracy

In theory, the ASE and TReCASE models allow the use of all exonic SNPs in the target gene to predict the phase of a candidate eQTL. In practice, there may be a significant number of exonic SNPs, which would form a large pool of haplotypes and cause some haplotype frequencies too small to be estimated. To balance between the predictive power and the numerical stability, we select a small number of exonic SNPs that provide the best prediction. In this article, we set $\widetilde{M}$ to four following much of the literature (de Bakker et al. 2005; Nicolae 2006; Lin et al. 2008). To guide the selection of exonic SNPs, we develop a measure, called Rsq, to quantify the amount of information in a given set of exonic SNPs for predicting the phase; see Appendix A for details. Rsq takes values between 0 and 1, with 0 and 1 indicating that the set of exonic SNPs provide none and complete information, respectively. For a candidate eQTL, we evaluate all sets of $\widetilde{M}$ exonic SNPs and select the one with the largest value of Rsq.

### 2.4 Testing A Gene-SNP Pair

The genome-wide scan of gene-SNP pairs involves tens of thousands of genes and hundreds or thousands of local candidate eQTLs for each gene. Without any prior knowledge about the *cis* or *trans* mechanism for each of the massive gene-SNP pairs, we apply the TReC and TReCASE models in parallel. The association testing may be performed by the Wald or likelihood-ratio statistic based on the likelihood functions (5) and (6). However, the calculation of these statistics requires solving for the MLEs of all parameters iteratively for each gene-SNP pair, and hence entails a considerable computational burden. By contrast, the score statistics (derived in Appendix

B) are computationally more efficient and numerically more stable, as it only requires solving for the nuisance parameters (i.e., $\psi$, $\boldsymbol{\pi}$, $\boldsymbol{\gamma}$, and $\phi$). With the TReC model, in particular, the nuisance parameters only need to be solved once for all tests pertaining to one gene. Thus, we employ the score statistic of each model to produce a *p*-value for each gene-SNP pair. Both tests have correct type I error even when the additivity assumption for the TReC model or the *cis*-regulation assumption for the TReCASE model does not hold, because under the null hypothesis there does not exist any association between the SNP and the gene expression (either ASE or total expression). Violation of these assumptions only affects the power. When the effect is *cis* (hence additive), the test based on the TReCASE model is the most powerful among all valid tests that use the same information. When the effect is *trans* and additive, the test based on the TReC model is the most powerful. Due to LD, several SNPs can be found to be associated with the expression of a gene by each model. To reduce such redundancy, we focus on the minimum-*p* SNP. Note that the TReC and TReCASE models may yield different minimum-*p* SNPs for the same gene.



### 2.5 Assessing the Significance of the Minimum-*p* SNP

Due to multiple testing, the score-based *p*-value at the minimum-*p* SNP is no longer indicative of the significance level. In addition, such *p*-values are not comparable across genes, as different genes have different numbers of local SNPs. To evaluate the significance of the minimum-*p* SNPs, we propose a permutation process that is tailored to the score statistics and features ultra-fast computation. We permute the dataset by fixing $(T, \boldsymbol{X}, R_1, R)$, shuffling $G$ and $\widetilde{H}$ as a whole among the individuals, and then randomly switching $\widetilde{h}_m$ and $\widetilde{h}_n$ in $\widetilde{H} = (\widetilde{h}_m, \widetilde{h}_n)$ of an individual. Because the nuisance parameters have been estimated without reference to the genetic association in the original dataset, they do not need to be re-estimated in the permuted datasets in which the association is altered. Thus, the analysis of the permuted datasets only involves simple re-evaluation of the cross-products between $(T, \boldsymbol{X}, R_1, R)$ and $(G, \widetilde{H})$ in the score statistics. This is fundamentally different from the traditional permutation process employed by Sun (2012), which repeats exactly

the same (iterative) analysis for the permuted datasets as for the original one. We generate $B$ permuted datasets (e.g., $B = 5,000$) and calculate the permutation $p$-value by taking the proportion of permuted datasets that generate smaller minimum $p$-values than the observed one.

### 2.6 Detecting eQTLs Among Genome-Wide Minimum-$p$ SNPs by FDR Control

We adopt the method of Storey and Tibshirani (Storey and Tibshirani 2003) for estimating FDR. Specifically, for any $p$-value cutoff $p_c$, the FDR is estimated by $p_0 p_c N_{\text{total}}/N_{\text{eQTL}}$, where $N_{\text{total}}$ is the total number of minimum-$p$ SNPs, $N_{\text{eQTL}}$ is the number of those with permutation $p$-values $\leq p_c$, and $p_0$ is the expected proportion of minimum-$p$ SNPs having null effects and is calculated as two times the proportion of minimum-$p$ SNPs with permutation $p$-values $\geq 0.5$. We explore over a fine grid of cutoffs and select the one that generates the desired FDR, e.g., 5% or 10%. Then, the minimum-$p$ SNPs whose permutation $p$-values lower than the final cutoff are declared as eQTLs. When the procedure is applied to the minimum-$p$ SNPs from the TReC and TReCASE models separately with the same FDR, the combined minimum-$p$ SNPs are also controlled for that FDR level.

### 2.7 Distinguishing Between *cis* and *trans* Mechanisms at eQLTs

After an eQTL is identified, the mechanism can be determined by the following *cis*-*trans* test. Let $\beta_{\text{A}} = \beta$ and $\beta_{\text{T}} = \beta + \alpha$. The hypotheses are written as

$$H_0(\textit{cis}\text{-eQTL}) : \alpha = 0 \quad \textit{vs}. \quad H_1(\textit{trans}\text{-eQTL}) : \alpha \neq 0.$$

We employ a score statistic (derived in Appendix B), which is computationally more efficient and numerically more stable than the likelihood-ratio statistic adopted by Sun (2012).

Up to this point, the TReC and TReCASE models work independently and may identify different eQTLs for the same gene. The *cis*-*trans* tests at the two eQTLs may yield the same or different conclusions. If both eQTLs are determined to be *cis*, we only retain the one identified by TReCASE, in consistent with our focus on only one *cis*-eQTL for a gene. Likewise, if both eQTLs are

established to be *trans*, we only retain the one identified by TReC. If one eQTL is determined to be *cis* and another *trans*, we keep both for future validation, as a gene can be subject to different regulations.

### 2.8 Estimating Effect Sizes of eQTLs

For a detected eQTL, the effect sizes in the ASE, TReC and TReCASE models can be estimated by maximizing the likelihood functions (2), (5) and (6), respectively. Note that, although the ASE model on its own is not used in the steps described in Sections 2.4-2.7, its effect size offers a unique insight into the ASE data alone. Because of missing phases, the maximization of (2) and (6) can be carried out by the expectation-maximization (EM) algorithms described in Appendix B. The maximization of (5) can be achieved by a standard Newton-Raphson algorithm. By the classical likelihood theory, the MLEs of the effect sizes are consistent, asymptotically normal and asymptotically efficient. Their variances can be estimated by the Louis formula (Louis 1982), as provided in Appendix B.

## 3. SIMULATION STUDIES

We carried out extensive simulation studies to evaluate the performance of the proposed and two-stage methods in realistic settings. To make a fair comparison between the two approaches, we considered a phasing method for the two-stage approach that uses the same amount of information as the proposed approach. Specifically, the method selects the $\widetilde{M}$ exonic SNPs that yield the maximum Rsq for a candidate eQTL, and infers the diplotype $H_i$ for the $i$th individual by the one that corresponds to the largest $P_{\widetilde{\boldsymbol{\pi}}}(H_i)$ among all $H_i \sim (G_i, \widetilde{H}_i)$, where $\widetilde{\boldsymbol{\pi}}$ is obtained by maximizing (7) in Appendix A. In Supplementary Method S1, we proved that treating the inferred phase as observed in the ASE model yields a valid test of the hypothesis $H_0 : \beta_{\mathrm{A}} = 0$, provided that $\widetilde{H}$ is independent of $R_1$ and $R$. We refer to the proposed maximum-likelihood approach and the two-stage (imputation) approach as MLE and IMP, respectively. As a benchmark, the approach using

the true phase is also included and referred to as TRUE. We set the nominal significance level to be 0.05, which is realistic because the commonly used FDR values of 5% and 10% correspond to the *p*-value cutoff of 0.023 and 0.062 in our analysis of TCGA breast cancer data.

The first study was focused on the effect size estimation and association testing at a *cis*-eQTL, when the four exonic SNPs are fixed. We generated genotype data for various sets of five SNPs according to the LD patterns observed in the TCGA breast cancer data (Supplementary Table S1). For each SNP set, we chose one SNP to be the *cis*-eQTL, leaving the other four to be exonic SNPs. The Rsq values of the exonic SNPs for phasing the *cis*-eQTL range from 0.56 to 0.93, which represent a wide spectrum of phasing accuracy according to the empirical distribution of Rsq values (Figure 6). Given the *cis*-eQTL, the expression data were generated from the association models, whose nuisance parameters were set at typical values observed in the TCGA data. Specifically, the TReC $T$ was simulated from a negative-binomial distribution with $\mu_{X,G} = 900 \times \exp\{0.1X + \mathcal{Z}(G, \beta_T)\}$ and $\phi = 0.2$, where $X$ was a normal random variable with mean zero and variance one. Then, $R$ was obtained as the integer part of $0.034 \times T$, where 0.034 was the average proportion of TReC that were allele-specific, and $R_1$ was simulated from a beta-binomial distribution with $p_{k,l}$ in (1) and $\psi = 0.05$. We generated 79 subjects (unless otherwise specified) for each dataset, which was the sample size of the TCGA data we analyzed. We applied the TReCASE model with the MLE, IMP and TRUE methods, and summarized the results in Table 1. As expected, MLE is virtually unbiased in all cases, and the standard error estimates accurately reflect the variability of the effect size estimates. The type I error rates of MLE are slightly greater than the nominal significance level of 0.05 due to the small sample size of 79. They are much closer to 0.05 when the sample size was increased to 500 (Supplementary Table S2). As predicted by our theory, IMP yielded proper type I error under the null hypothesis. In the presence of association, however, the estimates of $\beta$ produced by IMP are seriously biased towards zero, especially when Rsq is low. Consequently, IMP is uniformly less powerful than MLE; the power difference can be seen more clearly in Figure 3. As Rsq approaches 1, both MLE and IMP are nearly as powerful as TRUE. We

also applied the ASE model to the same data, whose results show similar patterns (Supplementary Table S3 and Supplementary Figure S1).

The second simulation study was designed to assess the validity and power of using the minimum-*p* SNP and the proposed permutation process. Unlike the first study, we generated the genotypes for all the exonic and local SNPs (i.e., within 200 kb of the gene) in various gene regions (Supplementary Table S4) via GWAsimulator (Li and Li 2008), which can mimic the minor allele frequencies (MAF) and LD patterns observed in the TCGA data. We selected a true *cis*-eQTL for each gene, and simulated the expression data in the same manner as in the first study. To significantly speed up the computation, we pruned out redundant SNPs that were in strong LD with others, so that the remaining SNPs have pairwise Pearson correlation coefficients less than 0.9. Note that the true *cis*-eQTL may be pruned out as well. We employed the TReC and TReCASE models; for the TReCASE model, we applied the MLE, IMP and TRUE methods. All the SNPs in and flanking a gene were scanned for association with the gene expression, and in the MLE and IMP methods, at most four exonic SNPs were selected for informing the phase of a candidate eQTL. Table 2 summarizes the characteristics of the minimum-*p* SNPs obtained by various methods. For all genes, the minimum-*p* SNPs yielded by MLE are in stronger LD with the true *cis*-eQTLs than those by IMP and by TReC. We used the proposed permutation process to assess the significance of the minimum-*p* SNPs in all methods. In the absence of any genetic effect, the permutation *p*-values are distributed uniformly between 0 and 1 (Supplementary Figure S2), confirming the validity of the permutation process even for the small sample size. In the present of *cis*-effects, MLE is uniformly more powerful than IMP and TReC (Figure 4), and is as powerful as TRUE for gene *MAP7D1* whose average value of Rsq at the minimum-*p* SNPs is as high as 0.80 (Table 2). When Rsq decreases to 0.4 and below, MLE tends to have lower power than TReC (Supplementary Figure S3). This is not surprising because, in such cases, the uncertainties about the phase overweigh the information contained in the ASE data and incorporating the ASE data even hurt the power from using the TReC data alone.

In the third simulation study, we investigated the MLE and IMP methods in distinguishing between *cis*- and *trans*-eQTLs, using the five-SNP settings in the first study. To examine the type I error of the *cis*-*trans* test, we considered various *cis*-effect sizes. As shown in Figure 5, IMP tends to generate inflated type I error (i.e., declaring excessive eQTLs to be *trans*), especially when Rsq is low and the *cis*-effect sizes are large. This is because IMP generally yields biased estimates for $\beta_A$, thereby creating a difference between $\beta_A$ and $\beta_T$. By contrast, MLE produced reasonable type I error rates in all cases, although a slight inflation can be observed due to the small sample size of 79. We then assessed the power of the *cis*-*trans* test by simulating *trans*-effects, which implies that $\beta_A = 0$. As shown in Supplementary Figure S4, IMP has the same power as TRUE, because the phasing does not matter when $\beta_A = 0$; note that IMP has accurate control of type I error in this particular case. MLE lost power relative to IMP and TRUE due to its enlarged variance to account for phasing uncertainty, but the power loss is very modest. We also assessed the power of the *cis*-*trans* test conditional on $\beta_A \neq 0$, whose alternative hypothesis ($\beta_A \neq 0$ and $\beta_A \neq \beta_T$), although corresponding to neither *cis*- nor *trans*-effect, may exist in reality through other mechanisms. Supplementary Figure S4 displays the power curves when $\beta_A$ was set to 0.5. Now that the phasing does matter, IMP has a shifted power curve compared to MLE and TRUE; the minimum power is not attained at the null of $\beta_T = \beta_A = 0.5$ but somewhere below 0.5.

## 4. APPLICATION TO TCGA DATA

We performed an analysis of local eQTLs (within 200 kb of a gene) for the genome-wide expression in 79 breast cancer (OMIM114480) patients from TCGA, using the RNA-seq data collected from the normal breast tissues only. Although tumor tissues are more relevant to a cancer, studying normal tissues from cancer patients is also important because it provides a baseline for comparison with what happened in tumor tissues. The 79 European females were genotyped at ~1 million SNPs by the Affymetrix 6.0 arrays and imputed against the European population (EUR) in the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) via

MACH (Li et al. 2010), reaching a total of ~36 million SNPs. The TReC and ASE of 19,048 autosomal genes were derived from the RNA-seq data by the `R` packages `asSeq` (Sun 2012) and `isoform` (`http://www.bios.unc.edu/~weisun/software/isoform.htm`). We note that using the imputation-augmented SNP set greatly increased the number of reads that can be counted into the ASE measurements. Supplementary Method S2 has more details on the processing of the RNA-seq data.

We restricted the candidate eQTLs and the exonic SNPs used for phase prediction to common SNPs with MAF $\geq$ 5%. By this criterion, the number of candidate eQTLs per gene varied from 7 to 15,536, with a median of 911, and the number of exonic SNPs per gene varied from 1 to 14,314, with a median of 53. As in the second simulation study, we pre-pruned the SNPs, which resulted in median numbers of 373 (ranging from 1 to 6,898) and 21 (ranging from 1 to 6,282) for candidate eQTLs and exonic SNPs, respectively. The TReC model was applied to all the 19,048 genes, including as covariates the log-transformed total read count per sample and three principal components calculated from the standardized TReC data. For numerical stability, we ran the TReCASE model for a candidate eQTL only if at least five samples who are heterozygous at the candidate eQTL had five or more allele-specific reads. Thus, the TReCASE model was applied to 14,414 genes.

First, the genome-wide gene-SNP pairs were scanned and the minimum-$p$ SNPs were obtained. Figure 6(a) displays the Rsq distributions at the minimum-$p$ SNPs by the TReCASE model with the MLE and IMP methods. Compared to MLE, IMP had a tendency to produce minimum-$p$ SNPs with lower Rsq values. Since the ASE data generally stop providing reliable information on the *cis*-effect when Rsq $\leq$ 0.4, we filtered out the TReCASE results whose minimum-$p$ SNPs had Rsq $\leq$ 0.4; such a quality-control (QC) procedure can also relieve the burden of multiple comparisons. As a result, 71.0% and 66.0% of minimum-$p$ SNPs by MLE and IMP, respectively, passing the QC procedure.

Then, permutation $p$-values for the minimum-$p$ SNPs were calculated. Figure 6(b) shows the

numbers of minimum-*p* SNPs whose permutation *p*-values pass varying cutoffs. Such a plot can be used to compare the power of different methods for identifying eQTLs. It can be seen that the MLE method was uniformly more powerful than IMP at any cutoff. The relatively small power gain was expected because, for most minimum-*p* SNPs, the phases were predicted sufficiently well (Figure 6[a]). In particular, 53.3% and 48.0% of minimum-*p* SNPs by MLE and IMP, respectively, had Rsq $\geq 0.7$. Due to the prevalence of eQTLs in the human genome, we do not expect the permutation *p*-values to follow a uniform distribution; the skewness of the histograms in Supplementary Figure S6 seems reasonable.

Next, we identified a total of 2,486 eQTLs by controlling for an FDR of 5%, and determined their *cis* or *trans* mechanisms using the *cis*-*trans* test at a nominal significance level of 0.01. Here, we chose a relatively stringent nominal significance level because we wish the test to be less sensitive to the difference between $\beta_{\mathrm{A}}$ and $\beta_{\mathrm{T}}$, which should always exist in real data even for *cis*-eQTLs. As shown in Table 3, there were 141 genes that were identified with eQTLs by the TReC model but lack of the ASE data; as a result, the TReCASE model did not apply to these genes and the eQTLs were undetermined for the mechanism. There are 23 eQTLs that were identified only by TReC; they are expected to be *trans*-eQTLs. The reason that 7 of them were determined to be *cis* is that the *cis*-*trans* test may lack power. There are 1,876 eQTLs that were identified only by TReCASE; they are expected to be *cis*-eQTLs. The reason that 195 of them were determined to be *trans* is due to the difference between $\beta_T$ and $\beta_A$ in the real data. The dominant number of eQTLs that were identified only by TReCASE highlights the importance of exploiting the ASE data for *cis*-eQTL mapping. Furthermore, there were 410 genes identified with eQTLs by both models. In particular, the two models identified overlapping *cis*-eQTLs for 179 genes, different *cis*-eQTLs for 154 genes, and one *cis*-eQTL and one *trans*-eQTL for 36 genes. A full list of the 2,486 eQTLs can be found in Supplementary Dataset S1.

At last, we compared the 2,486 eQTLs with the 134 SNPs found to be associated with breast cancer (referred to as trait-associated SNPs [TASs]) and downloaded from the NHLBI GWAS

Catalog (Welter et al. 2014). One eQTL, rs4784227 with alleles C and T, was a TAS. The SNP was reported by Long et al. (2010) to have a highly significant association with breast cancer risk, with an odds ratio of 1.19 per T allele in European Americans. Our analysis identified the SNP to be a *cis*-eQTL for gene *TOX3*, with the effect of allele T consistently estimated to be $-0.77$ by either the ASE, TReC or TReCASE model. Since it has been established that *TOX3* functions as a tumor suppressor in breast cancer cells (Cowper-Sal et al. 2012), our eQTL analysis suggested a putative functional role of rs4784227 on breast cancer risk, which has been confirmed *in vivo* in a breast cancer cell line (Cowper-Sal et al. 2012). Specifically, rs4784227 is located at a binding motif of the transcription factor FOXA1 and allele T of rs4784227 increases the binding affinity of FOXA1, which in turn represses the expression of *TOX3*. This *cis*-eQTL would also be identified using the TReCASE model with IMP, but not using the TReC model alone. In total, 6 eQTLs are in the vicinity (i.e., within 10 kb) of breast cancer TASs (Supplementary Tables S5-S6), whereas only 3 are expected by chance, showing evidence of eQTL enrichment among TASs. Like rs4784227, these eQTLs also hold the potential to elucidate the functional roles of the nearby TASs.



## 5. DISCUSSION

In summary, we developed a powerful and timely tool for *cis*-eQTL mapping with RNA-seq data. We showed that the proposed approach is more powerful than the existing two-stage approach and the use of total gene expression alone. All the proposed methods are implemented in the `R` package XXX, which is publicly available at our website. With the empirical data from the TCGA, we demonstrated the ability of our approach to discover *cis*-eQTLs and to elucidate the functional roles of nearby TASs, and the computational efficiency (e.g., only one day on 50 IBM HS22 machines to analyze the TCGA data) of our software to perform a genome-wide analysis. As functional genomics is rapidly progressing towards unraveling the *cis*-regulatory control of gene expression and many large-scale RNA-seq studies are underway, our analytical tool may play a key part in the coming years.

Our work significantly improves over the one by Sun (2012). First, we do not phase the candidate eQTL as a priori. Our maximum-likelihood methods phase the candidate eQTL and assess the *cis*-effect in an iterative fashion, and thus can remove the bias in the estimated effect size and boost the power in detecting a *cis*-eQTL. Second, we base our methods on score statistics and devise a permutation process that is tailored to the score statistics, thereby reducing the computation time to a mere ~5% of the time required the traditional permutation process. Third, for each SNP-gene pair, Sun (2012) adopted a strategy of first deciding on the *cis* or *trans* mechanism by the *cis-trans* test and then choosing the TReC or TReCASE model based on the test result. This strategy would incur a great computation cost because the computationally intensive *cis-trans* test has to be applied for every possible SNP-gene pair. In addition, it is unnatural to distinguish between the *cis* and *trans* mechanisms before any association is detected.

For the two-stage approach, we have used a small number of exonic SNPs to predict the phase of the candidate eQTL, so that the two-stage approach is compared with our maximum-likelihood approach on equal footing. An alternative is to use a hidden-Markov model (HMM) (Stephens et al. 2001; Browning and Browning 2009; Li et al. 2010; Delaneau et al. 2012), which exploits the LD information over a larger region, and thus may yield more accurate prediction of phase in certain situations. The conclusions regarding the relative merits of the maximum-likelihood approach versus the two-stage approach are expected to hold when an HMM is used.

We have assumed independence between the covariates $\boldsymbol{X}$ and the diplotype $H$, so that the probability $\Pr(\boldsymbol{X}|H)$ factored out from the likelihood (2). This assumption is reasonable if the covariates include demographic variables, such as age and gender, and environmental exposures. If it is believed that $\boldsymbol{X}$ and $H$ are correlated, then we can extend the methodology by adopting a generalized odds-ratio function for $\Pr(\boldsymbol{X}|H)$ (Hu et al. 2010). Likelihoods (2) and (6) will then involve the (potentially infinite-dimensional) distribution function of $\boldsymbol{X}$, which will greatly increase the theoretical and numerical complexity. In addition, assessing the significance of the minimum-$p$ SNP should then rely on the parametric bootstrap, instead of permutation.

We have focused on the minimum-$p$ SNP for each gene. It is possible that other SNPs act independently on the gene expression, but to a lesser extent. To detect such SNPs, we can include the minimum-$p$ SNP as a covariate and search for additional eQTLs conditional on the most significant one. Although it is straightforward to do so with the TReC model, it is less transparent with the ASE model. Not only the regression model (1) should include the minimum-$p$ SNP as a covariate, but also the likelihood (2) should account for the phasing uncertainties of both SNPs.

## APPENDIX A: Rsq Measure of Phasing Accuracy

The development of Rsq follows the rational of Li et al. (2010) in developing a measure of accuracy for imputed SNP genotypes. First, for a candidate eQTL and $\widetilde{M}$ exonic SNPs, we estimate the corresponding $\boldsymbol{\pi}$ by maximizing the likelihood that uses the genotype data only:

$$\widetilde{L}(\boldsymbol{\pi}) = \prod_{i=1}^{N} \sum_{H_i \sim (G_i, \widetilde{H}_i)} P_{\boldsymbol{\pi}}(H_i). \tag{7}$$

Due to the missing $H_i$, the expectation-maximization (EM) algorithm is the method of choice for the maximization. In the E-step, we evaluate

$$\omega_{ikl} = \frac{I\{(h_k, h_l) \sim (G_i, \widetilde{H}_i)\} P_{\boldsymbol{\pi}}(h_k, h_l)}{\sum_{(h_{k^*}, h_{l^*}) \sim (G_i, \widetilde{H}_i)} P_{\boldsymbol{\pi}}(h_{k^*}, h_{l^*})},$$

and in the M-step, we update $\pi_k$ by $N^{-1} \sum_{i=1}^{N} \sum_{l=1}^{K} \omega_{ikl}$, where $I(.)$ is the indicator function, $i = 1, \ldots, N$, and $k, l, k^*, l^* = 1, \ldots, K$. Denote the maximum likelihood estimator (MLE) of $\boldsymbol{\pi}$ by $\widetilde{\boldsymbol{\pi}}$. Then, for an individual with $G = 1$ and $\widetilde{H} = (\widetilde{h}_m, \widetilde{h}_n)$, we can estimate $G_1$ by the expected number of allele 1 on the same chromosome as $\widetilde{h}_m$, i.e., $\widetilde{G}_1 = \widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} / \{\widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} + \widetilde{\pi}_{0,m}\widetilde{\pi}_{1,n}\}$, where the subscript $(1, m)$ represents the haplotype consisting of allele 1 at the candidate eQTL and $\widetilde{h}_m$ at the exonic SNPs. Although $\widetilde{G}_1$ is not a form of phasing due to the continuous nature, it is an unbiased estimator of $G_1$. To quantify the phasing accuracy, a natural measure would be the Pearson correlation coefficient $r^2$ between $\widetilde{G}_1$ and the underlying $G_1$. Since $G_1$ is unknown, we estimate $r^2$ by the ratio of the sample variance of $\widetilde{G}_1$ with what would be expected if $G_1$ was

observed (Li et al. 2010), and refer to the estimated $r^2$ as the Rsq measure. Specifically, Rsq = $\mathrm{Var}(\widetilde{G}_1|G=1)/\mathrm{Var}(G_1|G=1)$. Write $P_1 = \sum_{m,n=1}^{\widetilde{K}} \widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} + \widetilde{\pi}_{0,m}\widetilde{\pi}_{1,n}$. By simple algebra, we show that $\mathrm{E}(\widetilde{G}_1|G=1) = \sum_{m,n=1}^{\widetilde{K}} \widetilde{G}_1 P_{\widetilde{\boldsymbol{\pi}}}(\widetilde{h}_m, \widetilde{h}_n|G=1) = P_1^{-1}\sum_{m,n=1}^{\widetilde{K}} \widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} = 0.5$, and $\mathrm{E}(\widetilde{G}_1^2|G=1) = P_1^{-1}\sum_{m,n=1}^{\widetilde{K}}(\widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} + \widetilde{\pi}_{0,m}\widetilde{\pi}_{1,n})^{-1}(\widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n})^2$. By the symmetry of the two haplotypes, $\mathrm{E}(G_1|G=1) = 0.5$, $\mathrm{E}(G_1^2|G=1) = 0.5$, and thus $\mathrm{Var}(G_1|G=1) = 0.25$. In summary, we develop a measure of phasing accuracy to be

$$\mathrm{Rsq} = -1 + 4P_1^{-1}\sum_{m,n=1}^{\widetilde{K}}(\widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n} + \widetilde{\pi}_{0,m}\widetilde{\pi}_{1,n})^{-1}(\widetilde{\pi}_{1,m}\widetilde{\pi}_{0,n})^2.$$

We note that the calculation of Rsq, though involving an iterative process to solve for $\widetilde{\boldsymbol{\pi}}$, is computationally fast and scalable to the genome-wide scan.

## APPENDIX B: Maximum Likelihood Inference

### *B.1 Inference for ASE model*

We obtain the MLEs of $\beta_{\mathrm{A}}$, $\psi$, and $\boldsymbol{\pi}$ based on the likelihood (2) via the following EM algorithm. The complete-data log-likelihood is

$$l_{\mathrm{A}}(\beta_{\mathrm{A}}, \psi, \boldsymbol{\pi}) = \sum_{i=1}^{N}\sum_{k,l=1}^{K} I\{H_i = (h_k, h_l)\}\log\Big[P_{\beta_{\mathrm{A}},\psi}(R_{1i}|R_i, h_k, h_l)P_{\boldsymbol{\pi}}(h_k, h_l)\Big].$$

In the E-step, we evaluate $\mathrm{E}[I\{H_i = (h_k, h_l)\}|R_{1i}, R_i, G_i, \widetilde{H}_i]$, which can be shown to be

$$\omega_{ikl} = \frac{I\{(h_k, h_l) \sim (G_i, \widetilde{H}_i)\}P_{\beta_{\mathrm{A}},\psi}(R_{1i}|R_i, h_k, h_l)P_{\boldsymbol{\pi}}(h_k, h_l)}{\sum_{(h_{k^*}, h_{l^*})\sim(G_i,\widetilde{H}_i)} P_{\beta_{\mathrm{A}},\psi}(R_{1i}|R_i, h_{k^*}, h_{l^*})P_{\boldsymbol{\pi}}(h_{k^*}, h_{l^*})}. \tag{8}$$

In the M-step, we maximize $l_{\mathrm{A}}(\beta_{\mathrm{A}}, \psi, \boldsymbol{\pi})$ with $I\{H_i = (h_k, h_l)\}$ replaced by $\omega_{ikl}$. Specifically, we update $\pi_k$, $k = 1, \ldots, K$, by

$$\pi_k^{\mathrm{new}} = N^{-1}\sum_{i=1}^{N}\sum_{l=1}^{K}\omega_{ikl}, \tag{9}$$

and update $\beta_A$ and $\psi$ by a one-step Newton-Raphson iteration. Starting with $\beta_A = 0$, $\psi = 0$ and $\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}$, where $\tilde{\boldsymbol{\pi}}$ is the MLE based on the likelihood (S1), we iterate between the E-step and M-step until the change in the observed-data log-likelihood is negligible. Denote the MLEs of $\beta_A$, $\psi$, and $\boldsymbol{\pi}$ by $\hat{\beta}_A$, $\hat{\psi}$, and $\hat{\boldsymbol{\pi}}$.

We can estimate the standard errors of $\hat{\beta}_A$, $\hat{\psi}$, and $\hat{\boldsymbol{\pi}}$ by the Louis formula (Louis 1982). Write $C_{k,l,j} = p_{k,l} + j\psi$, $D_{k,l,j} = 1 - p_{k,l} + j\psi$, and $I_{k,l} = I(h_k = h_l)$, where $p_{k,l}$ was defined in expression (1). For $i = 1, \ldots, N$ and $k, l = 1, \ldots, K$, calculate the vector

$$\dot{\boldsymbol{l}}_{ikl} = \begin{pmatrix} \left(\sum_{j=0}^{R_{1i}-1} C_{k,l,j}^{-1} - \sum_{j=0}^{R_i-R_{1i}-1} D_{k,l,j}^{-1}\right) \partial p_{k,l}/\partial\beta_A \\ \sum_{j=0}^{R_{1i}-1} jC_{k,l,j}^{-1} + \sum_{j=0}^{R_i-R_{1i}-1} jD_{k,l,j}^{-1} - \sum_{j=1}^{R_i-1} j(1+j\psi)^{-1} \\ (I_{k,1} + I_{l,1})/\pi_1 - (I_{k,K} + I_{l,K})/\pi_K \\ \vdots \\ (I_{k,K-1} + I_{l,K-1})/\pi_{K-1} - (I_{k,K} + I_{l,K})/\pi_K \end{pmatrix}.$$

Calculate the block diagonal matrix $\ddot{\boldsymbol{l}}_{ikl}$ with two blocks. The first block is a 2×2 matrix with two diagonal elements,

$$\left(-\sum_{j=0}^{R_{1i}-1} C_{k,l,j}^{-2} - \sum_{j=0}^{R_i-R_{1i}-1} D_{k,l,j}^{-2}\right)(\partial p_{k,l}/\partial\beta_A)^2 + \left(\sum_{j=0}^{R_{1i}-1} C_{k,l,j}^{-1} - \sum_{j=0}^{R_i-R_{1i}-1} D_{k,l,j}^{-1}\right)(\partial^2 p_{k,l}/\partial\beta_A^2)$$

and

$$-\sum_{j=0}^{R_{1i}-1} j^2 C_{k,l,j}^{-2} - \sum_{j=0}^{R_i-R_{1i}-1} j^2 D_{k,l,j}^{-2} + \sum_{j=1}^{R_i-1} j^2(1+j\psi)^{-2},$$

and an off-diagonal element

$$\left(-\sum_{j=0}^{R_{1i}-1} jC_{k,l,j}^{-2} + \sum_{j=0}^{R_i-R_{1i}-1} jD_{k,l,j}^{-2}\right)(\partial p_{k,l}/\partial\beta_A).$$

The second block is a $(K-1)\times(K-1)$ matrix

$$\text{diag}\left\{-(I_{k,1}+I_{l,1})/\pi_1^2,\cdots,-(I_{k,K-1}+I_{l,K-1})/\pi_{K-1}^2\right\}-(I_{k,K}+I_{l,K})/\pi_K^2\boldsymbol{E}_{(K-1)\times(K-1)},$$

where $\boldsymbol{E}_{(K-1)\times(K-1)}$ is a $(K-1)\times(K-1)$ matrix with all elements being 1. For subjects with $R_i=0$, the elements in $\dot{\boldsymbol{l}}_{ikl}$ and $\ddot{\boldsymbol{l}}_{ikl}$ associated with the derivative with respect to $\beta_{\rm A}$ or $\psi$ are zero. The covariance matrix for $(\widehat{\beta}_{\rm A},\widehat{\psi},\widehat{\pi}_1,\ldots,\widehat{\pi}_{K-1})'$ is given by $\boldsymbol{\Sigma}_{\rm A}=\boldsymbol{I}_{\rm A}^{-1}$, where

$$\boldsymbol{I}_{\rm A}=-\sum_{ikl}\omega_{ikl}\ddot{\boldsymbol{l}}_{ikl}-\sum_i\Big[\sum_{kl}\omega_{ikl}\dot{\boldsymbol{l}}_{ikl}\dot{\boldsymbol{l}}_{ikl}'-\Big\{\sum_{kl}\omega_{ikl}\dot{\boldsymbol{l}}_{ikl}\Big\}\Big\{\sum_{kl}\omega_{ikl}\dot{\boldsymbol{l}}_{ikl}\Big\}'\Big],$$

and $\boldsymbol{\nu}'$ denotes the transpose of the vector $\boldsymbol{\nu}$. Note that $\omega_{ikl}$, $\ddot{\boldsymbol{l}}_{ikl}$ and $\dot{\boldsymbol{l}}_{ikl}$ should be evaluated at the MLEs. The standard errors for $\widehat{\beta}_{\rm A}$, $\widehat{\psi}$, and $\widehat{\pi}_1,\ldots,\widehat{\pi}_{K-1}$ are the square-roots of the diagonal elements in $\boldsymbol{\Sigma}_{\rm A}$. The standard error for $\widehat{\pi}_K=1-\sum_{k=1}^{K-1}\widehat{\pi}_k$ is the square root of $(1,\ldots,1)\boldsymbol{B}(1,\ldots,1)'$, where $\boldsymbol{B}$ is the last $(K-1)\times(K-1)$ sub-matrix of $\boldsymbol{\Sigma}_{\rm A}$.

To calculate the score statistic for testing $H_0:\beta_{\rm A}=0$, we first produce the restricted MLE of the nuisance parameter $\boldsymbol{\delta}=(\psi,\pi_1,\ldots,\pi_{K-1})'$, denoted by $\widetilde{\boldsymbol{\delta}}=(\widetilde{\psi},\widetilde{\pi}_1,\ldots,\widetilde{\pi}_{K-1})'$. The $\widetilde{\pi}_1,\ldots,\widetilde{\pi}_{K-1}$ turn out to be the MLEs based on the likelihood (7), and $\widetilde{\psi}$ maximizes $\sum_{i=1}^N\log\left\{P_{\beta_{\rm A}=0,\psi}(R_{1i}|R_i)\right\}$. Then, the score and information functions for $(\beta_{\rm A},\boldsymbol{\delta}')'$, which take the forms $\boldsymbol{U}_{\rm A}=\sum_{ikl}\omega_{ikl}\dot{\boldsymbol{l}}_{ikl}$ and $\boldsymbol{I}_{\rm A}$, respectively, are evaluated at $(0,\widetilde{\boldsymbol{\delta}}')'$. Based on the partition

$$\boldsymbol{U}_{\rm A}=\begin{pmatrix}U_{{\rm A},\beta}\\ \boldsymbol{U}_{{\rm A},\boldsymbol{\delta}}\end{pmatrix}\quad\text{and}\quad\boldsymbol{I}_{\rm A}=\begin{pmatrix}I_{{\rm A},\beta\beta}&\boldsymbol{I}_{{\rm A},\beta\boldsymbol{\delta}}\\ \boldsymbol{I}_{{\rm A},\boldsymbol{\delta}\beta}&\boldsymbol{I}_{{\rm A},\boldsymbol{\delta}\boldsymbol{\delta}}\end{pmatrix},$$

where $\beta$ represents $\beta_{\rm A}$, the score statistic $n^{-1/2}U_{{\rm A},\beta}$ is asymptotically zero-mean normal with a variance that can be consistently estimated by $n^{-1}\left\{I_{{\rm A},\beta\beta}-\boldsymbol{I}_{{\rm A},\beta\boldsymbol{\delta}}\boldsymbol{I}_{{\rm A},\boldsymbol{\delta}\boldsymbol{\delta}}^{-1}\boldsymbol{I}_{{\rm A},\boldsymbol{\delta}\beta}\right\}$.

*B.2 Inference for TReC model*

The MLEs of $\beta_{\rm T}$, $\boldsymbol{\gamma}$, and $\phi$, denoted by $\widehat{\beta}_{\rm T}$, $\widehat{\boldsymbol{\gamma}}$, and $\widehat{\phi}$, based on the likelihood (5) can be obtained via the Newton-Raphson algorithm with the first and negative second derivative of $l_{\rm T}(\beta_{\rm T},\boldsymbol{\gamma},\phi)=$

$\log\{L_{\mathrm{T}}(\beta_{\mathrm{T}}, \boldsymbol{\gamma}, \phi)\}$, denoted by $\boldsymbol{U}_{\mathrm{T}}$ and $\boldsymbol{I}_{\mathrm{T}}$. The covariance matrix for $(\widehat{\beta}_{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}, \widehat{\phi})'$ is given by $\boldsymbol{I}_{\mathrm{T}}^{-1}$, which is evaluated at the MLEs.

To calculate the score statistic for testing $H_0 : \beta_{\mathrm{T}} = 0$, we first produce the restricted MLEs of the nuisance parameter $\boldsymbol{\xi} = (\boldsymbol{\gamma}, \phi)'$, denoted by $\widetilde{\boldsymbol{\xi}} = (\widetilde{\boldsymbol{\gamma}}, \widetilde{\phi})'$. Then, the score and information functions for $(\beta_{\mathrm{T}}, \boldsymbol{\xi}')'$, which take the forms $\boldsymbol{U}_{\mathrm{T}}$ and $\boldsymbol{I}_{\mathrm{T}}$, respectively, are evaluated at $(0, \widetilde{\boldsymbol{\xi}}')'$. Based on the partition

$$\boldsymbol{U}_{\mathrm{T}} = \begin{pmatrix} U_{\mathrm{T},\beta} \\ \boldsymbol{U}_{\mathrm{T},\boldsymbol{\xi}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{I}_{\mathrm{T}} = \begin{pmatrix} I_{\mathrm{T},\beta\beta} & \boldsymbol{I}_{\mathrm{T},\beta\boldsymbol{\xi}} \\ \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\beta} & \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix},$$

where $\beta$ represents $\beta_{\mathrm{T}}$, the score statistic $n^{-1/2} U_{\mathrm{T},\beta}$ is asymptotically zero-mean normal with a variance that can be consistently estimated by $n^{-1}\left\{I_{\mathrm{T},\beta\beta} - \boldsymbol{I}_{\mathrm{T},\beta\boldsymbol{\xi}} \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\beta}\right\}$.

*B.3 Inference for TReCASE model*

Let $\beta = \beta_{\mathrm{A}} = \beta_{\mathrm{T}}$. The MLEs of $\beta$, $\boldsymbol{\delta}$, and $\boldsymbol{\xi}$ based on the likelihood (6) can be obtained via a similar EM algorithm as for the ASE model. The complete-data log-likelihood is $l_{\mathrm{TA}}(\beta, \boldsymbol{\delta}, \boldsymbol{\xi}) = l_{\mathrm{A}}(\beta, \boldsymbol{\delta}) + l_{\mathrm{T}}(\beta, \boldsymbol{\xi})$. In the E-step, we evaluate the $\omega_{ikl}$ in (8). In the M-step, we update $\pi_k$, $k = 1, \ldots, K$, by the $\pi_k^{\mathrm{new}}$ in (9) and update $\beta$, $\psi$, and $\boldsymbol{\xi}$ by a one-step Newton-Raphson iteration. Denote the MLEs of $\beta$, $\boldsymbol{\delta}$, and $\boldsymbol{\xi}$ by $\widehat{\beta}$, $\widehat{\boldsymbol{\delta}}$, and $\widehat{\boldsymbol{\xi}}$. According to their definitions, $\widehat{\boldsymbol{\delta}} \neq (\widehat{\psi}, \widehat{\pi}_1, \ldots, \widehat{\pi}_{K-1})'$ and $\widehat{\boldsymbol{\xi}} \neq (\widehat{\boldsymbol{\gamma}}, \widehat{\phi})'$, as they are MLEs for different models. The covariance matrix for $(\widehat{\beta}, \widehat{\boldsymbol{\delta}}', \widehat{\boldsymbol{\xi}}')'$ is given by $\boldsymbol{\Sigma}_{\mathrm{TA}} = \boldsymbol{I}_{\mathrm{TA}}^{-1}$, where

$$\boldsymbol{I}_{\mathrm{TA}} = \begin{pmatrix} I_{\mathrm{A},\beta\beta} + I_{\mathrm{T},\beta\beta} & \boldsymbol{I}_{\mathrm{A},\beta\boldsymbol{\delta}} & \boldsymbol{I}_{\mathrm{T},\beta\boldsymbol{\xi}} \\ \boldsymbol{I}_{\mathrm{A},\boldsymbol{\delta}\beta} & \boldsymbol{I}_{\mathrm{A},\boldsymbol{\delta}\boldsymbol{\delta}} & \mathbf{0} \\ \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\beta} & \mathbf{0} & \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}.$$

To calculate the score statistic for testing $H_0 : \beta = 0$, we first obtain the restricted MLEs for the nuisance parameter $\boldsymbol{\tau} = (\boldsymbol{\delta}', \boldsymbol{\xi}')'$ to be $\widetilde{\boldsymbol{\tau}} = (\widetilde{\boldsymbol{\delta}}', \widetilde{\boldsymbol{\xi}}')'$, where $\widetilde{\boldsymbol{\delta}}$ and $\widetilde{\boldsymbol{\xi}}$ are restricted MLEs from the ASE and TReC models, respectively. Then, the score and information functions for $(\beta, \boldsymbol{\tau}')'$, which take the forms $\boldsymbol{U}_{\mathrm{TA}} = (U_{\mathrm{A},\beta} + U_{\mathrm{T},\beta}, \boldsymbol{U}_{\mathrm{A},\boldsymbol{\delta}}', \boldsymbol{U}_{\mathrm{T},\boldsymbol{\xi}}')'$ and $\boldsymbol{I}_{\mathrm{TA}}$, respectively, are evaluated at $(0, \widetilde{\boldsymbol{\tau}}')'$. The

score statistic and the variance estimator can be constructed in the same fashion as for the ASE model.

*B.4 Cis-trans test*

Let $\beta_{\mathrm{A}} = \beta$ and $\beta_{\mathrm{T}} = \beta + \alpha$. To calculate the score statistic for testing $H_0 : \alpha = 0$, we first obtain the restricted MLEs for the nuisance parameter $\boldsymbol{\tau}^* = (\beta, \boldsymbol{\tau}')'$ to be $\widetilde{\boldsymbol{\tau}}^* = (\widetilde{\beta}, \widetilde{\boldsymbol{\delta}}', \widetilde{\boldsymbol{\xi}}')'$, where $\widehat{\beta}$, $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\xi}}$ are the MLEs of $\beta$, $\boldsymbol{\delta}$, and $\boldsymbol{\xi}$, respectively, from the TReCASE model. The score and information functions for $(\alpha, (\boldsymbol{\tau}^*)')'$ take the forms

$$\boldsymbol{U}_{cis} = \begin{pmatrix} U_{\mathrm{T},\beta} \\ U_{\mathrm{A},\beta} + U_{\mathrm{T},\beta} \\ U_{\mathrm{A},\boldsymbol{\delta}} \\ U_{\mathrm{T},\boldsymbol{\xi}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{I}_{cis} = \begin{pmatrix} I_{\mathrm{T},\beta\beta} & I_{\mathrm{T},\beta\beta} & \mathbf{0} & \boldsymbol{I}_{\mathrm{T},\beta\boldsymbol{\xi}} \\ I_{\mathrm{T},\beta\beta} & I_{\mathrm{A},\beta\beta} + I_{\mathrm{T},\beta\beta} & \boldsymbol{I}_{\mathrm{A},\beta\boldsymbol{\delta}} & \boldsymbol{I}_{\mathrm{T},\beta\boldsymbol{\xi}} \\ \mathbf{0} & \boldsymbol{I}_{\mathrm{A},\boldsymbol{\delta}\beta} & \boldsymbol{I}_{\mathrm{A},\boldsymbol{\delta}\boldsymbol{\delta}} & \mathbf{0} \\ \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\beta} & \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\beta} & \mathbf{0} & \boldsymbol{I}_{\mathrm{T},\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix},$$

respectively, which are evaluated at $(0, (\widetilde{\boldsymbol{\tau}}^*)')'$. The score statistic and the variance estimator can be constructed in the same fashion as for the ASE model.

## REFERENCES

Browning, B., and Browning, S. (2009), "A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals," *American Journal of Human Genetics*, 84(2), 210–223.

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009), "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, 10(3), 184–194.

Cowper-Sal, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Moore, J. H., Lupien, M. et al. (2012), "Breast Cancer Risk-Associated SNPs Modulate the Affinity of Chromatin for FOXA1 and Alter Gene Expression," *Nature Genetics*, 44(11), 1191–1198.

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005), "Efficiency and power in genetic association studies," *Nature genetics*, 37(11), 1217–1223.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012), "A Linear Complexity Phasing Method for Thousands of Genomes," *Nature Methods*, 9(2), 179–181.

Flicek, P., Amode, M., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. et al. (2011), "Ensembl 2011," *Nucleic Acids Research*, 39(suppl 1), D800.

Hu, Y., Lin, D., and Zeng, D. (2010), "A General Framework for Studying Genetic Effects and Gene–Environment Interactions with Missing Data," *Biostatistics*, 11(4), 583–598.

Kendziorski, C., and Wang, P. (2006), "A Review of Statistical Methods for Expression Quantitative Trait Loci Mapping," *Mammalian Genome*, 17(6), 509–517.

Li, C., and Li, M. (2008), "GWAsimulator: A Rapid Whole-Genome Simulation Program," *Bioinformatics*, 24(1), 140–142.

Li, Y., Willer, C., Ding, J., Scheet, P., and Abecasis, G. (2010), "MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes," *Genetic Epidemiology*, 34(8), 816–834.

Lin, D., Hu, Y., and Huang, B. (2008), "Simple and efficient analysis of disease association with missing genotype data," *Am J Hum Genet*, 82(2), 444–452.

Long, J., Cai, Q., Shu, X.-O., Qu, S., Li, C., Zheng, Y., Gu, K., Wang, W., Xiang, Y.-B., Cheng, J. et al. (2010), "Identification of a Functional Genetic Variant at 16q12.1 for Breast Cancer Risk: Results from the Asia Breast Cancer Consortium," *PLoS Genetics*, 6(6), e1001002.

Louis, T. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society. Series B.*, 44(2), 226–233.

Nicolae, D. L. (2006), "Testing Untyped Alleles (TUNA)applications to genome-wide association studies," *Genetic epidemiology*, 30(8), 718–727.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010), "Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing," *Nature*, 464(7289), 768–772.

Stephens, M., Smith, N., and Donnelly, P. (2001), "A New Statistical Method for Haplotype Reconstruction from Population Data," *American Journal of Human Genetics*, 68(4), 978–989.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *Proceedings of the National Academy of Sciences*, 100(16), 9440–9445.

Sun, W. (2012), "A Statistical Framework for eQTL Mapping Using RNA-seq Data," *Biometrics*, 68(1), 1–11.

The 1000 Genomes Project Consortium (2012), "An Integrated Map of Genetic Variation From 1,092 Human Genomes," *Nature*, 491, 56–65.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. et al. (2014), "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations," *Nucleic Acids Research*, 42(D1), D1001–D1006.

Table 1. Results for the *cis*-effect of a SNP from the TReCASE model in the first simulation study

| | | | | MLE | | | | IMP | | | | TRUE | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Rsq | MAF | $\beta$ | Bias | SE | SEE | PW | Bias | SE | SEE | PW | Bias | SE | SEE | PW |
| A1 | 0.56 | 0.28 | 0.0 | 0.003 | 0.110 | 0.106 | 0.069 | 0.002 | 0.089 | 0.088 | 0.058 | 0.001 | 0.089 | 0.088 | 0.054 |
| | | | 0.3 | −0.001 | 0.103 | 0.101 | 0.829 | −0.095 | 0.094 | 0.090 | 0.618 | 0.002 | 0.089 | 0.088 | 0.929 |
| A2 | 0.64 | 0.39 | 0.0 | 0.001 | 0.092 | 0.090 | 0.059 | 0.001 | 0.080 | 0.079 | 0.053 | 0.001 | 0.080 | 0.079 | 0.055 |
| | | | 0.3 | −0.005 | 0.088 | 0.086 | 0.913 | −0.072 | 0.084 | 0.080 | 0.805 | 0.000 | 0.080 | 0.078 | 0.967 |
| A3 | 0.78 | 0.32 | 0.0 | 0.000 | 0.094 | 0.090 | 0.062 | 0.000 | 0.086 | 0.084 | 0.058 | −0.001 | 0.086 | 0.084 | 0.059 |
| | | | 0.3 | −0.005 | 0.089 | 0.086 | 0.914 | −0.038 | 0.086 | 0.082 | 0.880 | 0.000 | 0.083 | 0.081 | 0.951 |
| A4 | 0.93 | 0.21 | 0.0 | 0.004 | 0.100 | 0.098 | 0.056 | 0.004 | 0.098 | 0.096 | 0.054 | 0.003 | 0.098 | 0.096 | 0.056 |
| | | | 0.3 | −0.001 | 0.100 | 0.098 | 0.867 | −0.011 | 0.099 | 0.097 | 0.853 | 0.001 | 0.098 | 0.097 | 0.881 |

NOTE: MLE, IMP, and TRUE are the maximum-likelihood method, the two-stage method, and the method using the true phase, respectively. A1–A4 denote the four scenarios listed in Supplementary Table S1, with Rsq and MAF indicating the measure of phasing accuracy and the minor allele frequency, respectively, at the candidate eQTL. $\beta$ is the *cis*-effect size. Bias and SE are the bias and standard error of the parameter estimator. SEE is the mean of the standard error estimator. PW is the type I error or power for testing zero parameter value at the 0.05 nominal significance level; the tests are based on the Wald statistics. Each entry is based on 10,000 replicates.



Table 2. Characteristics of minimum-*p* SNPs in the second simulation study

| | *cis*-eQTL | TReC | | MLE | | | IMP | | | TRUE | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Gene | MAF | LD | MAF | LD | MAF | Rsq | LD | MAF | Rsq | LD | MAF |
| *EIF3I* | 0.23 | 0.72 | 0.24 | 0.87 | 0.25 | 0.45 | 0.59 | 0.27 | 0.54 | 0.99 | 0.23 |
| *PIGV* | 0.15 | 0.41 | 0.20 | 0.61 | 0.21 | 0.52 | 0.44 | 0.24 | 0.45 | 0.88 | 0.17 |
| *EIF2C1* | 0.10 | 0.58 | 0.20 | 0.64 | 0.21 | 0.64 | 0.53 | 0.24 | 0.65 | 0.83 | 0.15 |
| *MAP7D1* | 0.11 | 0.38 | 0.22 | 0.71 | 0.16 | 0.80 | 0.65 | 0.17 | 0.73 | 0.78 | 0.15 |

NOTE: *cis*-eQTL is a randomly selected causal SNP for each gene. TReC is the TReC model. MLE, IMP and TRUE are the maximum-likelihood method, the two-stage method, and the method using the true phase, respectively, with the TReCASE model. LD is the average Pearson correlation coefficient between the genotypes of the minimum-*p* SNP and the *cis*-eQTL. MAF is the average minor allele frequency. Rsq is the average phasing accuracy. Each entry is based on $\beta_A = \beta_T = 0.5$ and 1,000 replicates.

Table 3. Genes identified with eQTLs at an FDR of 5% in the TCGA Data

| Test results | | | Genes |
|---|---|---|---|
| TReC | TReCASE (MLE) | *cis-trans* | no. |
| yes | NA | NA | 141 |
| yes | no | *cis* | 7 |
| | | *trans* | 16 |
| no | yes | *cis* | 1681 |
| | | *trans* | 195 |
| yes | yes | *cis* | 179 |
| | | *trans* | 29 |
| | | *cis,cis* | 154 |
| | | *cis,trans* | 36 |
| | | *trans,trans* | 12 |

NOTE: TReC is the TReC model. TReCASE (MLE) is the TReCASE model with the MLE method. Yes and no indicate that the minimum-*p* SNP is declared as significant and nonsignificant, respectively, by FDR control. NA is not applicable. *cis-trans* is the *cis-trans* test. If two different eQTLs within a gene are identified by TReC and TReCASE, their mechanisms are separated by a comma and unordered.

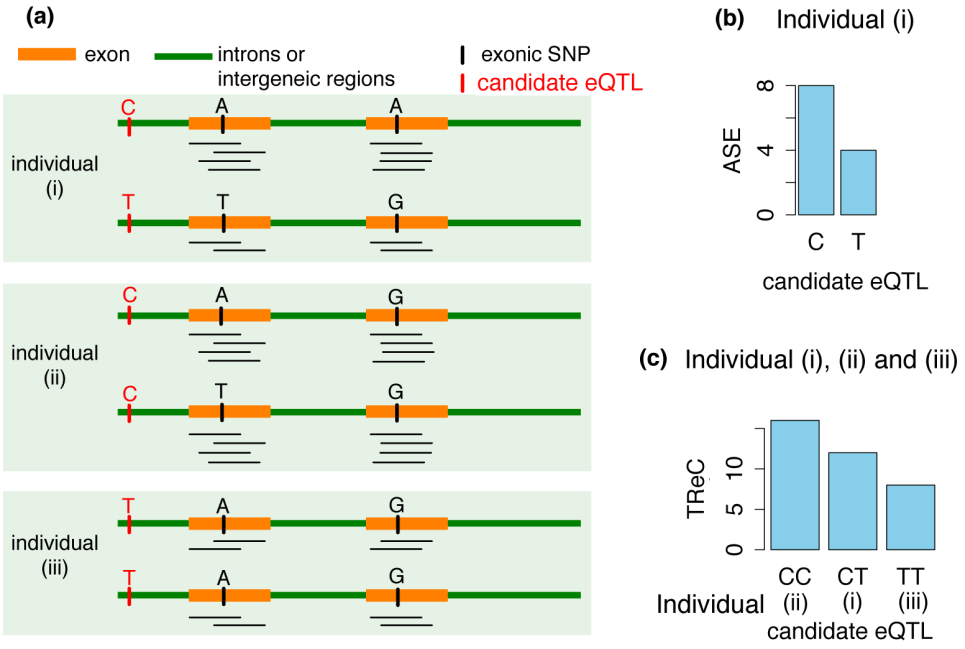


Figure 1. A hypothetical example of *cis*-eQTL mapping, adapted from Sun (2012). (a) RNA-seq measurements of a hypothetic gene with two exons for three diploid individuals. Because individuals (i) and (ii) have at least one exonic SNPs showing a heterozygous genotype, they have non-zero ASE measurements. In addition, only individual (i) has a heterozygous genotype at the candidate eQTL, so association analysis using ASE is only possible within individual (i). (b) The ASE measurements for individual (i). (c) The TReC for the three individuals.

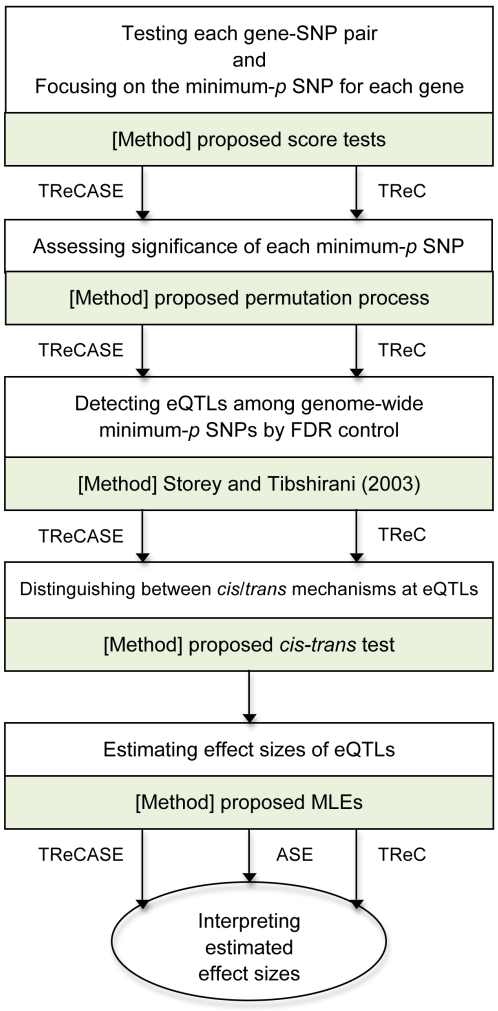


Figure 2. A five-step pipeline for performing a genome-wide analysis of *cis*-eQTLs. TReCASE, TReC and ASE are models that are used in parallel in certain steps.

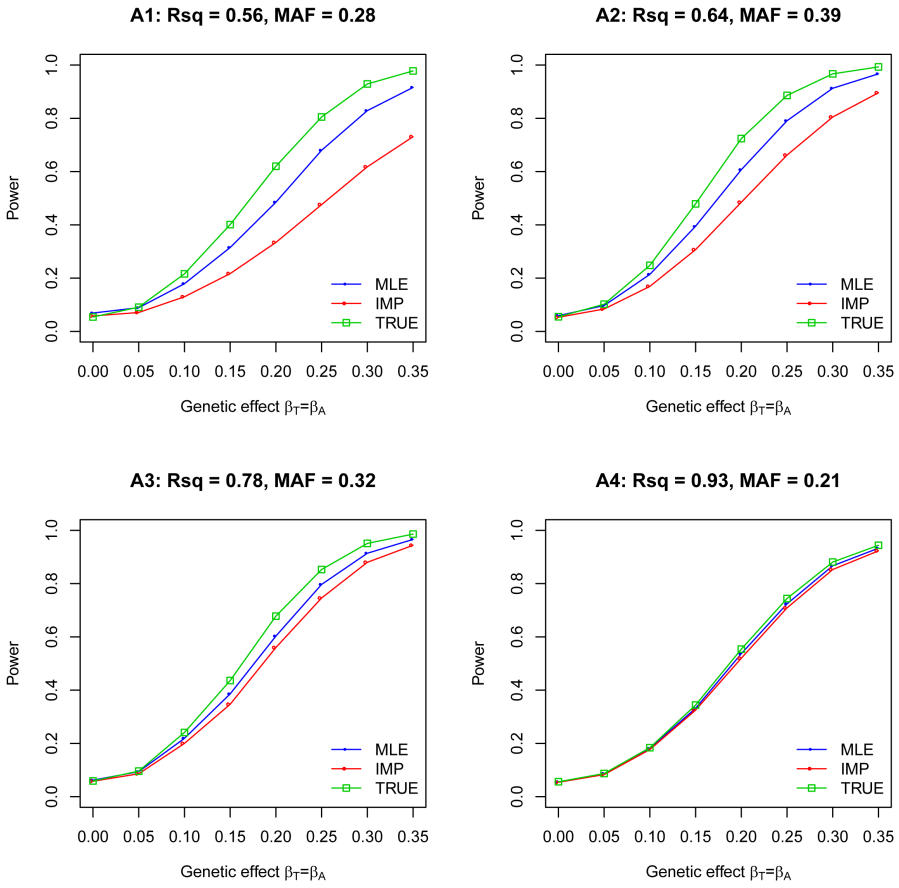


Figure 3. Power of the TReCASE model for testing a *cis*-eQTL in the first simulation study. The nominal significance level is 0.05. Each power estimate is based on 10,000 replicates.

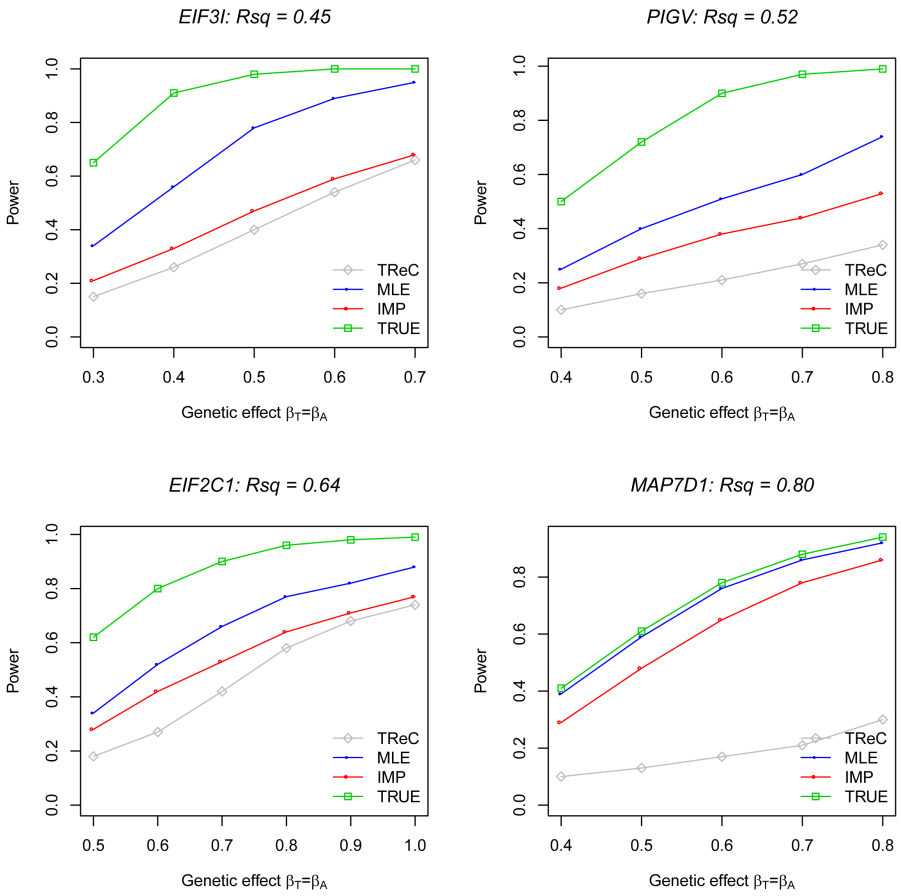


Figure 4. Power of the TReCASE model for detecting a *cis*-eQTL in the second simulation study. Rsq is the average phasing accuracy of the minimum-*p* SNPs yielded by MLE. The nominal significance level is 0.05. Each power estimate is based on 1,000 replicates.

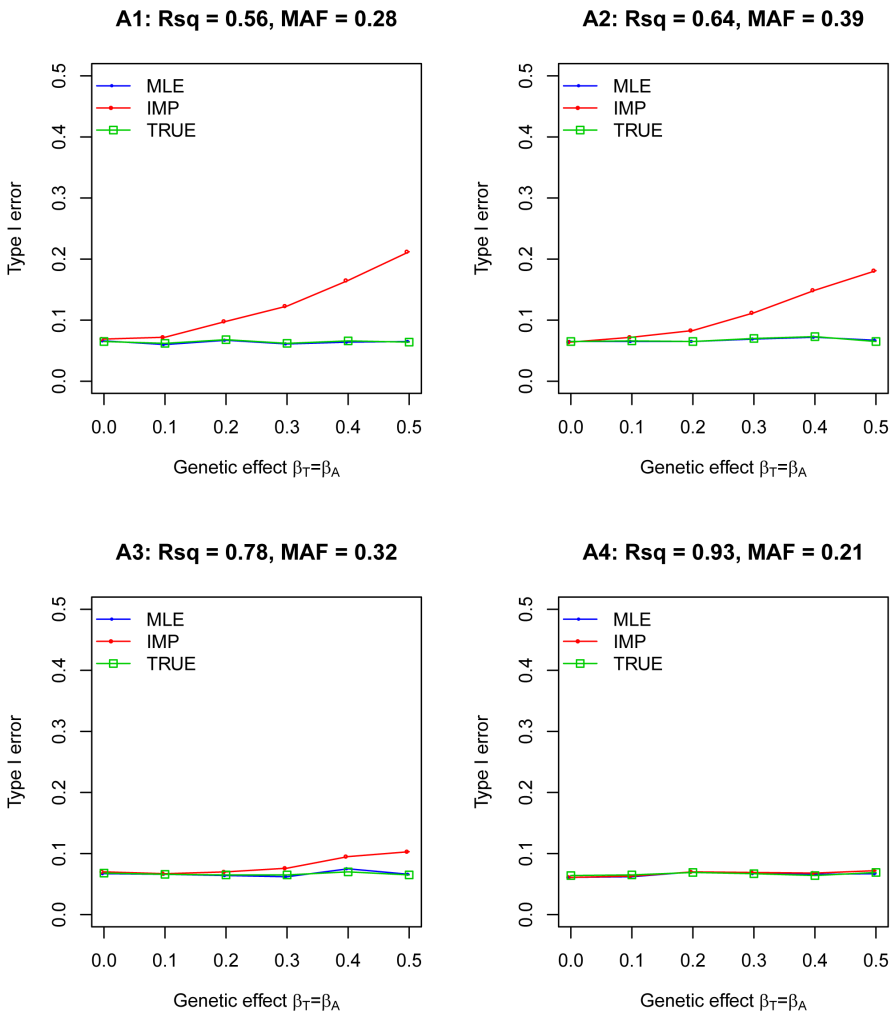


Figure 5. Type I error of the *cis-trans* test in the third simulation study. The nominal significance level is 0.05. Each estimate of type I error rate is based on 10,000 replicates.

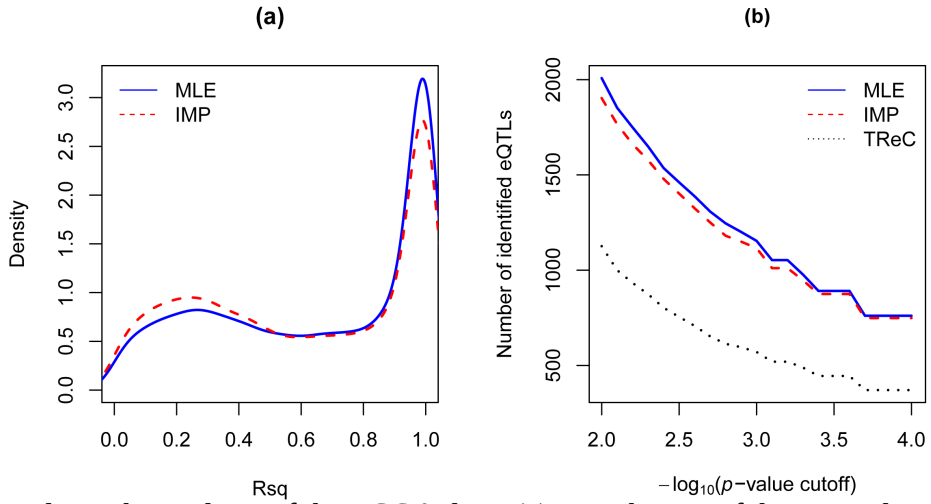


Figure 6. Results in the analysis of the TCGA data. (a) Distribution of the Rsq values at minimum-$p$ SNPs. (b) Numbers of minimum-$p$ SNPs with permutation $p$-values passing varying $p$-value cutoffs at a $-\log_{10}$ scale. MLE and IMP are the maximum-likelihood and two-stage methods, respectively, with the TReASE model. TReC is the TReC model.