NIH Public Access
Author Manuscript
Biometrics. Author manuscript; available in PMC 2013 March 1.

Published in final edited form as:
Biometrics. 2012 March ; 68(1): 1–11. doi:10.1111/j.1541-0420.2011.01654.x.

A Statistical Framework for eQTL Mapping Using RNA-seq Data

Wei Sun[1,2,*]

[1] Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.

[2] Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, U.S.A.

Summary

RNA-seq may replace gene expression microarrays in the near future. Using RNA-seq, the expression of a gene can be estimated using the total number of sequence reads mapped to that gene, known as the Total Read Count (TReC). Traditional eQTL mapping methods, such as linear regression, can be applied to TReC measurements after they are properly normalized. In this paper, we show that eQTL mapping, by directly modeling TReC using discrete distributions, has higher statistical power than the two-step approach: data normalization followed by linear regression. In addition, RNA-seq provides information on allele-specific expression (ASE) that is not available from microarrays. By combining the information from TReC and ASE, we can computationally distinguish *cis*- and *trans*-eQTL and further improve the power of *cis*-eQTL mapping. Both simulation and real data studies confirm the improved power of our new methods. We also discuss the design issues of RNA-seq experiments. Specifically, we show that by combining TReC and ASE measurements, it is possible to minimize cost and retain the statistical power of *cis*-eQTL mapping by reducing sample size while increasing the number of sequence reads per sample. In addition to RNA-seq data, our method can also be employed to study the genetic basis of other types of sequencing data, such as ChIP-seq (chromatin immunoprecipitation followed by DNA sequencing) data. In this paper, we focus on eQTL mapping of a single gene using the association-based method. However, our method establishes a statistical framework for future developments of eQTL mapping methods using RNA-seq data (e.g. linkage-based eQTL mapping), and the joint study of multiple genetic markers and/or multiple genes.

Keywords

Allele-specific Expression (ASE); eQTL; RNA-seq; Total Read Count (TReC)

1. Introduction

Since the first genome-wide study of gene expression quantitative trait locus (eQTL) was published in 2002 (Brem et al., 2002), eQTL mapping has evolved from a novel approach to a standard strategy in many genome-wide studies. It has been shown that eQTL not only provides insight on transcription regulation, but also illuminates the molecular basis of phenotypic outcomes, such as complex diseases (Cookson et al., 2009). High-throughput RNA sequencing, also known as RNA-seq, is becoming a popular technique to measure gene expression abundance (Wang et al., 2009; Mortazavi et al., 2008). Briefly, an RNA sample extracted from single or multiple cells is first converted to a library of cDNA fragments of a few hundred base pairs (bps). Then the sequence segments on one or both

* wsun@bios.unc.edu .

Supplementary Materials The Web Appendix, Tables, and Figures referenced in Sections 2, 3, and 4 are available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

ends of the cDNA fragments are sequenced. One such sequenced segment is called a sequence read and is typically 30bp to 100bp long, depending on the sequencing platform. In a typical RNA-seq experiment, tens of millions of sequence reads are obtained for a sample and the expression of a gene is measured by the number of sequence reads mapped to this gene. RNA-seq offers several advantages over microarrays. For example, RNA-seq is less noisy and has a much larger dynamic range than microarrays, and RNA-seq can identify new transcripts while microarray's detection capability is limited by the probes on the array (Wang et al., 2009).

In terms of eQTL studies, one important advantage of RNA-seq is its ability to measure allele-specific expression (ASE). In diploid individuals, there are two sets of chromosomes: one is the paternal copy and the other is the maternal copy. Therefore each gene has two alleles: the paternal allele and the maternal allele. The transcript abundance of each allele of a gene is referred to as its ASE, which has been used to distinguish *cis*- and *trans*-eQTL (Doss et al., 2005; Ronald et al., 2005). *Cis*-acting regulation is due to DNA variations that directly influence the transcription process in an allele-specific manner. Alternatively, *trans*-acting regulation affects the expression of a gene by modifying the activity or expression of factors that regulate the gene, which leads to the same amount of expression changes for both alleles of the gene (Wittkopp et al., 2004). For example, suppose a DNA polymorphism site $\mathscr{D}$ is located at the promoter of a gene $\mathscr{G}_1$. A mutation of $\mathscr{D}$ disrupts transcription initiation of $\mathscr{G}_1$, and thus $\mathscr{D}$ is a *cis*-eQTL of $\mathscr{G}_1$. The allele of $\mathscr{G}_1$ that harbors the mutated copy of $\mathscr{D}$ has a reduced expression level, while the allele that harbors the normal copy of $\mathscr{D}$ has a normal expression level. Now assume $\mathscr{G}_1$ encodes a transcription factor that regulates the expression of $\mathscr{G}_2$. Then $\mathscr{D}$ is a *trans*-eQTL of $\mathscr{G}_2$. Both alleles of $\mathscr{G}_2$ will have the same amount of expression reduction due to the mutation at $\mathscr{D}$. Gene expression microarrays cannot measure ASE since both alleles of a gene are targeted by the same probe set. Thus *cis/trans*-eQTL cannot be distinguished in traditional eQTL studies using gene expression microarrays. However, because *cis*-eQTLs are often due to DNA polymorphisms near the gene, local and distant eQTLs are often referred to as *cis*-eQTL and *trans*-eQTL, respectively. As Rockman and Kruglyak (2006) have pointed out, "the casual conflation of different usages of *cis* and *trans* has resulted in a significant amount of confusion". With RNA-seq, we can measure ASE using the sequence reads that overlap with at least one heterozygous SNP. Therefore, it is possible to distinguish *cis*-eQTLs from *trans*-eQTLs when we map eQTLs using RNA-seq data. This distinction is important for understanding the mechanism of gene expression regulation, for example, from an evolutionary point of view (Wittkopp et al., 2004).

We propose three statistical methods to map eQTLs using RNA-seq data: the TReC, ASE, and TReCASE methods, together with a statistical test to distinguish between *cis*-eQTL and *trans*-eQTL. The TReC method maps either *cis*- or *trans*-eQTLs by assessing the significant association between the total read count per gene and the SNP genotype via a negative binomial or Poisson regression. The ASE method maps *cis*-eQTLs using allele-specific expression. Specifically, it tests whether the over-expression of one allele of a gene is associated with one allele of a target SNP, while modeling allele-specific read counts using a beta-binomial distribution. Allele-specific reads can only be identified if there are heterozygous SNPs in the coding regions of the gene, and some reads overlap with such heterozygous SNPs. In addition, an association can be tested only if the target SNP is heterozygous. Thus the ASE method can only use part of the sequence reads (those overlapping with at least one heterozygous SNP), and part of the samples (those in which ASE can be measured and the target SNP is heterozygous). The TReC method uses all available samples and all mapable reads, although it is sensitive to possible confounding effects, such as observed/unobserved batch effects. In contrast, although the ASE method only uses part of the data, it is less sensitive to confounding effects. This is because ASE

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

measures the expression of one allele using the other allele as a within-sample control, therefore most confounding effects will be cancelled (Pastinen, 2010). The TReCASE method maps *cis*-eQTLs by combining TReC and ASE data in a likelihood-based framework. Thus TReCASE not only exploits more information than the TReC or ASE method alone, but also is more robust since it is less likely that an unobserved confounding effect will alter TReC and ASE measurements in a consistent manner.

Traditional eQTL mapping methods that were developed for microarray expression data often assume the expression data follow a normal distribution (conditioning on certain covariates including genetic factors), and apply linear regression or equivalent approaches for eQTL mapping (Kendziorski and Wang, 2006). A straightforward application of such linear regression approaches for RNA-seq data requires appropriate normalization of the TReC measurements. For example, Pickrell et al. (2010) normalized TReC measurements of each gene by normal quantile transformation and then applied linear regression for the eQTL mapping. As is shown in simulation and real data studies, our TReC or TReCASE method has significantly higher power than the two-step approach of normalizing TReC data followed by linear regression. In addition, it will be clear from our derivation that by modeling the RNA-seq data using discrete distributions, it is natural to combine TReC and ASE data for joint inference. However, if one transforms TReC and/or ASE measurements into a continuous scale, it is difficult if not impossible to combine them into one model.

Similar to many other new techniques at their emergence, how to design RNA-seq experiments to achieve high power and low cost is of great interest. In this paper, we show that by using our TReCASE method for *cis*-eQTL mapping, it is possible to retain statistical power, while reducing the sample size and increasing the total number of reads per sample. Since the sequencing cost is falling rapidly as the techniques advance, sample collection, especially for human studies, will contribute the major cost of an eQTL experiment. Therefore this strategy of increasing the number of reads per sample, but decreasing the sample size could dramatically reduce the overall cost.

2. Proposed Methods

2.1 *Overview*

The diagram in Figure 1 illustrates an example of a *cis*-eQTL in three individuals. For a hypothetical gene with two exons, we assume there is a SNP (single nucleotide polymorphism) on the first exon, which has two alleles A and T, and we test for the association of this gene's expression with an upstream SNP (target SNP), which has two alleles C and T. To simplify the discussion, we assume there is no read on the exon junction, all the reads in the first exon overlap with the exonic SNP, and the two exons have the same number of reads. Individuals (i) and (ii) have heterozygous genotypes on the exonic SNP, and thus ASE can be measured. Individual (i) has a heterozygous genotype for the target SNP, thus we can test *cis*-eQTL using ASE from that individual (Figure 1(b)). This *cis*-eQTL also manifests its effect through the TReC across the three individuals (Figure 1 (c)). In this example, TReC and ASE eQTLs have the same cause, the expression of the T allele is half of the expression of the C allele.

Another feature illustrated in this diagram is that we need to know the haplotype around the gene for detecting *cis*-eQTL using ASE. For example, in individual (i), given haplotype C-A and T-T, we can assign the ASE to the two alleles of the target SNP. Haplotype information is also needed to combine the ASE across different SNPs in the gene body. The ambiguity of haplotype construction (i.e., phasing) may influence the accuracy of ASE measurements. Usually phasing by statistical methods is not accurate across a long range of genetic distance, thus ASE can hardly be used to identify distant *cis*-eQTLs. However, recent

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

technique advancements show that direct phasing of the whole genome is possible, and may become an integral part of genome-wide analysis in the near future (Fan et al., 2010; Kitzman et al., 2010).

2.2 An Association Model Using Total Read Count (TReC) per Gene

We consider one gene and study the association of its expression with the *j*-th SNP. Let t_i be the total number of reads mapped to this gene in the *i*-th sample, where $1 \le i \le N$, and N is sample size. We model t_i by a Poisson distribution or a negative binomial distribution, depending on whether there is significant over-dispersion for the Poisson distribution, which can be assessed by a score test (Dean, 1992). Let $f_{NB}(; \mu_i, \phi)$ be the density function for a negative binomial distribution with mean μ_i and dispersion parameter ϕ:

$$f_{NB}(t_i;\mu_i,\phi)=\frac{\Gamma(t_i+1/\phi)}{t_i!\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{t_i}. \quad (1)$$

A negative binomial distribution can be considered a generalization of a Poisson distribution to allow for over-dispersion. Specifically, if a random variable X follows a Poisson distribution with mean μ_i, and μ_i follows a gamma distribution, then the resulting distribution for X is a negative binomial distribution. The variance of a negative binomial distribution is $\mu_i+\phi\mu_i^2$, where $\phi\mu_i^2$ is the over-dispersion part of the variance. As $\phi \to 0$, $f_{NB}(t_i; \mu_i, \phi)$ reduces to a Poisson distribution with parameter μ_i.

Denote the major and minor alleles of the *j*-th SNP as A and B, respectively. Let x_{ij}, the genotype of the *j*-th SNP in the *i*-th sample, be the number of B alleles, i.e., x_{ij} = 0, 1, and 2 for genotypes AA, AB and BB, respectively. For either a Poisson or negative binomial regression, we employ a log link function to acknowledge the fact that $\mu_i > 0$:

$$\log(\mu_i)=b_0+b_\kappa\kappa_i+\sum_{u=1}^{p}b_u\eta_{iu}+w\left(b_{x_j},x_{ij}\right), \quad (2)$$

where κ_i is the logarithm of the total number of reads for sample *i*, and $\sum_{u=1}^{p} b_u\eta_{iu}$ controls other confounding effects. In the real data analysis in this paper, following Pickrell et al. (2010), we capture such additional confounding effects by principal component analysis (PCA) of the expression data, and thus η_{iu} is the loading of the *u*-th principal component (PC) in the *i*-th sample. Specifically, PCs are calculated using log-transformed standardized TReCs, where standardization means dividing a TReC by the total number of reads per sample. The coefficients of these confounding effects (e.g., total read counts per sample and PCs) are estimated for each gene separately, since they may influence the expression of different genes by different magnitudes. For example, suppose a batch effect is due to two protocols in RNA amplification. One protocol favors amplification of genes with higher GC content but the other does not. Thus this batch effect may have different influences for genes with different GC contents. Note that PCs from genotype data may also be used as covariates to capture population stratification. We do not include them in our real data analysis since this is a well studied population without apparent population stratification. Gene length is another factor that is often considered in measuring gene expression by RNA-seq. However it is un-relevant to our study since we map eQTL for each gene separately. $w(b_{x_j}, x_{ij})$ models the genetic effect of the *j*-th SNP

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$w\left(b_{x_j}, x_{ij}\right) = \begin{cases} 0 & \text{if } x_{ij}=0 \\ \log\left(1+\exp\left(b_{x_j}\right)\right) - \log(2) & \text{if } x_{ij}=1 \\ b_{x_j} & \text{if } x_{ij}=2 \end{cases}$$

The functional form of $w(b_{x_j}, x_{ij})$ can be derived as follows. First, let

$$\log(\mu_{iAA}) \equiv \log\left(\mu_i | x_{ij}=0\right) = b_0 + \sum_{u=1}^{p} b_u \eta_{iu} + b_\kappa \kappa_i, \tag{3}$$

$$\log(\mu_{iBB}) \equiv \log\left(\mu_i | x_{ij}=2\right) = b_0 + \sum_{u=1}^{p} b_u \eta_{iu} + b_\kappa \kappa_i + b_{x_j}. \tag{4}$$

In other words, we define b_{x_j} as

$$b_{x_j} = \log\left(\frac{\mu_{iBB}}{\mu_{iAA}}\right) = \log\left(\frac{\mu_{BB}}{\mu_{AA}}\right) = \log\left(\frac{2\mu_B}{2\mu_A}\right) = \log\left(\frac{\mu_B}{\mu_A}\right). \tag{5}$$

Thus

$$\log\left(\frac{\mu_{iAB}}{\mu_{iAA}}\right) = \log\left(\frac{\mu_A+\mu_B}{\mu_A+\mu_A}\right) = \log\left(\frac{1+\mu_B/\mu_A}{2}\right) = \log\left\{\frac{1+\exp\left(b_{x_j}\right)}{2}\right\}, \tag{6}$$

and

$$\log(\mu_{iAB}) \equiv \log\left(\mu_i | x_{ij}=1\right) = b_0 + \sum_{u=1}^{p} b_u \eta_{iu} + b_\kappa \kappa_i + \log\left\{\frac{1+\exp\left(b_{x_j}\right)}{2}\right\}. \tag{7}$$

We refer to the model specified by equations (1)-(2) as the total read count (TReC) model. Let $\mathbf{b} = \{b_0, b_1, \ldots, b_p, b_\kappa\}$. In a general form, we write the likelihood of the TReC model of the i-th sample as

$$g_{TR}\left(t_i; \mathbf{b}, b_{x_j}, \phi, \mathbf{X}\right) = I_{NB} f_{NB}(t_i; \mu_i, \phi) + (1 - I_{NB}) f_P(t_i; \mu_i), \tag{8}$$

where $\mathbf{X}$ indicates all the relevant covariates, and I_{NB} is an indicator which equals 1 if a negative binomial distribution is used, and 0 if a Poisson distribution is used. The MLE of the model parameters can be estimated by the following iterative procedure.

Initialization—Fit a null model by a Poisson regression using the confounding variables κ_i and η_{iu} ($1 \le u \le p$), and estimate $\mathbf{b} = \{b_0, b_1, \ldots, b_p, b_\kappa\}$. Then conduct a score test for over-dispersion. If the score test p-value is smaller than a cutoff value, e.g., 0.05, fit a negative binomial regression model and estimate $\mathbf{b} = \{b_0, b_1, \ldots, b_p, b_\kappa\}$ and ϕ. The decision regarding the distribution family is kept for the following iterations. Occasionally, when a SNP has a large effect, we might see that a negative binomial distribution fits the data better under null hypothesis, but a Poisson distribution is not unreasonable under alternative hypothesis. We fit negative binomial models for both the null and alternative models in such cases to facilitate a nested likelihood ratio test, which is more robust and is computationally

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

more efficient. When the sample size is large and computational efficiency is less of a concern, replacing this likelihood ratio test with a Wald test under alternative hypothesis is a reasonable choice.

Iteration

1. Given **b**, estimate b_{x_j} by numerical method, see Supplementary Materials (Section A) for details.
2. Given b_{x_j}, estimate **b** by a Poisson regression with offsets $w(b_{x_j}, x_{ij})$, or estimate **b** and ϕ by a negative binomial regression with offsets $w(b_{x_j}, x_{ij})$. The likelihoods of **b** and ϕ are independent, and hence they can be estimated separately. See Supplementary Materials (Section A) for details of the estimation of ϕ.

Termination—Iterate steps (1) and (2) until estimates of all the parameters converge.

The significance of association can be tested by a likelihood ratio test comparing the null model estimated in the initialization step and the alternative model estimated at the end of the iterations.

2.3 An association model using allele specific expression (ASE)

The measurement of ASE and ASE-based eQTL mapping are two independent steps. We first discuss the former. For a particular individual, suppose the two haplotypes of a gene of interest are known and denote them by h_a and h_b. Then any sequence read that overlaps with at least one heterozygous exonic SNP of this gene can be assigned to either h_a or h_b. The ASE of a haplotype is simply the total number of allele-specific reads mapped to this haplotype. In other words, we merge the allele-specific expression captured by each exonic SNP of this gene using haplotype information. Let n_i be the total number of allele-specific reads mapped to this gene in the i-th sample. If there is no heterozygous SNP within the exon regions of this gene at sample i, then $n_i = 0$ and we do not include this sample in the likelihood. Otherwise, let n_{ih_b} be the number of allele-specific reads mapped to haplotype h_b, and thus $n_{ih_a} = n_i - n_{ih_b}$.

Next let's consider the association between the expression of a gene and SNP j (target SNP). SNP j can be anywhere in the genome and does not need to be within the gene body. We can study the ASE association as long as SNP j is connected with the gene body by a contiguous haplotype. In practice, the haplotype phasing may not be accurate in a long range, thus we may focus on SNPs around the gene body. For example, in the real data study, we examine all the SNPs within the gene body or outside the gene body, but within 200kb of the transcription start or end sites. If SNP j is heterozygous in sample i with genotype AB, let n_{ijB} be the number of AS reads mapped to the same haplotype as allele B. Then $n_{ijB} = n_{ih_b}$ if SNP j's B allele is on haplotype h_b, and $n_{ijB} = n_i - n_{ih_b}$ otherwise. We model n_{ijB} by a beta-binomial distribution, which is an extension of a binomial distribution to allow for possible over-dispersion. Specifically, let n_{ijB} follow a binomial distribution with the number of trials n_i, and the probability of success p_S. If p_S follows a beta distribution with parameters α and β, the resulting distribution for n_{ijB} is a beta-binomial distribution

$$h\left(n_{ij_B}; n_i, \alpha, \beta\right) = \binom{n_i}{n_{ij_B}} \frac{B\left(n_{ij_B} + \alpha, n_i - n_{ij_B} + \beta\right)}{B(\alpha, \beta)}. \tag{9}$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

To connect this ASE model with the TReC model, we adopt a commonly used strategy to parameterize a beta-binomial distribution by $\pi = \alpha/(\alpha + \beta)$ and $\theta = 1/(\alpha + \beta)$ (Griffiths, 1973):

$$h\left(n_{ij_B};n_i,\pi,\theta\right)=\left(\begin{array}{c} n_i \\ n_{ij_B}\end{array}\right)\frac{\prod_{k=0}^{n_{ij_B}-1}(\pi+k\theta)\prod_{k=0}^{n_i-n_{ij_B}-1}(1-\pi+k\theta)}{\prod_{k=1}^{n_i-1}(1+k\theta)}, \tag{10}$$

where π is the expected proportion of AS reads from haplotype H_B and θ is a dispersion parameter. If there is no over-dispersion, then $\theta = 0$ and n_{ijB} follows a binomial distribution. Let π_0 and π_1 be the proportion of AS reads from haplotype H_B under the null and alternative hypotheses, respectively. π_0 is a fixed constant, while π_1 is estimated from the data. Ideally, if the two genome-wide haplotypes of an individual are known, and the sequence reads are mapped to these two haplotypes separately, then there is no mapping bias and $\pi_0 = 0.5$. In practice, the complete haplotypes may be unknown and sequence reads are mapped to the reference genome, which may lead to preferential mapping to the reference alleles of the SNPs. One remedy is to exclude those SNPs with strong mapping bias and then assume $\pi_0 = 0.5$ (Pickrell et al., 2010). We adopt this strategy in our studies.

If SNP j is homozygous in sample i, n_{ijB} equals 0 or n_i, and thus it is not informative for *cis*-eQTL mapping since it does not provide any information regarding the degree of allelic imbalance or the degree of over-dispersion (θ). However, as long as there are heterozygous exonic SNPs in this gene, we still have non-trivial n_{ih_b} (i.e., $0 < n_{ih_b} < n_i$), which is informative for estimation of the dispersion parameter θ. Therefore we have the following likelihood function:

$$\begin{aligned} h\left(n_{ij_B},n_{ih_b};n_i,\pi_1,\pi_0,\theta\right) &= 1 \quad \text{if there is no AS read in sample } i\text{;otherwise} \\ h\left(n_{ij_B},n_{ih_b};n_i,\pi_1,\pi_0,\theta\right) &= \left\{\left(\begin{array}{c} n_i \\ n_{ij_B}\end{array}\right)\frac{\prod_{k=0}^{n_{ij_B}-1}(\pi_1+k\theta)\prod_{k=0}^{n_i-n_{ij_B}-1}(1-\pi_1+k\theta)}{\prod_{k=1}^{n_i-1}(1+k\theta)}\right\}^{\zeta_{ij}} \\ &= \left\{\left(\begin{array}{c} n_i \\ n_{ih_b}\end{array}\right)\frac{\prod_{k=0}^{n_{ih_b}-1}(\pi_0+k\theta)\prod_{k=0}^{n_i-n_{ij_b}-1}(1-\pi_0+k\theta)}{\prod_{k=1}^{n_i-1}(1+k\theta)}\right\}^{1-\zeta_{ij}} \end{aligned} \tag{11}$$

where ζ_{ij} is an indicator, which equals 1 or 0 if SNP j is heterozygous or homozygous in sample i, respectively. Under the null hypothesis that there is no *cis*-association,

$$\begin{aligned} h\left(n_{ih_b};n_i,\pi_0,\theta\right) &= 1 \quad \text{if there is no AS read in sample } i\text{;otherwise} \\ h\left(n_{ih_b};n_i,\pi_0,\theta\right) &= \left(\begin{array}{c} n_i \\ n_{ih_b}\end{array}\right)\frac{\prod_{k=0}^{n_{ih_b}-1}(\pi_0+k\theta)\prod_{k=0}^{n_i-n_{ij_b}-1}(1-\pi_0+k\theta)}{\prod_{k=1}^{n_i-1}(1+k\theta)}. \end{aligned} \tag{12}$$

We refer to the model specified by equations (11) and (12) as the ASE model. The MLE of parameters π_1 and θ can be estimated using a quasi-Newton method (Byrd et al., 1995), see Supplementary Materials (Section B) for details. The significance of *cis*-association can be tested by a likelihood ratio test comparing the null model with a fixed π_0 and an MLE of θ and the alternative model with MLE of π_1 and θ. Note that θ under the null and alternative models are estimated separately.

2.4 Joint study of total read count (TReC) and allele-specific expression (ASE)

We connect the TReC and ASE models by formulating the log odds of observing reads from the same haplotype of allele B or allele A of the target SNP:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$b_{x_j}=\log\left(\frac{\mu_B}{\mu_A}\right)=\log\left(\frac{\pi}{1-\pi_1}\right),\ \text{and thus}\ \pi_1=\frac{\exp\left(b_{x_j}\right)}{1+\exp\left(b_{x_j}\right)}. \tag{13}$$

Therefore we can use the information from both TReC and ASE to estimate b_{x_j}, and we name the combined model the TReCASE model. Let $\mathbf{b} = (b_0, \ldots, b_p, b_\kappa)$ and let $\mathbf{X}$ indicate all the relevant covariates. Based on the above definition of the TReC and ASE models, the likelihood of the TReCASE model across N samples can be written as

$$L\left(\mathbf{b}, b_{x_j}, \theta, \phi | t_i, n_i, n_{i_{j_B}}, \mathbf{X}\right)=\prod_{i=1}^{N} g_{TR}\left(t_i, \mathbf{b}, b_{x_j}, \phi, \mathbf{X}\right)\prod_{i:n_i>0} h\left(n_{i_{j_B}}, n_{ih_b}; n_i, \pi_i, \pi_0, \theta\right). \tag{14}$$

We obtain the MLE of the parameters using the following algorithm.

Initialization—The TReC and ASE models are fitted separately to obtain the initial estimates of $\mathbf{b}$, $\boldsymbol{\phi}$, and θ under the null model.

Iteration

1. Given θ, $\boldsymbol{\phi}$, and $\mathbf{b}$, estimate b_{x_j} by numerical methods (see Supplementary Materials (Section C) for details).
2. Given b_{x_j}, calculate π_1 by equation (13), and then estimate θ given π_1, as shown in the Supplementary Materials (Section B).
3. Given b_{x_j}, estimate $\mathbf{b}$ and $\boldsymbol{\phi}$ by a Poisson or negative binomial regression with offsets 0, log [{1 + exp(b_{x_j})}/2], and b_{x_j}, while the genotype of the SNP is 0, 1, and 2, respectively.

Termination—Iterate steps (1) to (3) until the estimates of all the parameters converge.

Under the null hypothesis, $b_{x_j} = 0$ and $\pi_0 = 0.5$, the likelihoods of TReC and ASE models are independent, and we can obtain the MLE of all the parameters and the corresponding likelihoods, as described in the previous sections. Finally, the significance of the association is tested by a likelihood ratio test comparing the null and alternative models.

2.5 *TReC, ASE or TReCASE?*

We have proposed three methods for eQTL mapping: TReC, ASE, and TReCASE. An immediate question is, which method should be used in practice? If the underlying eQTL is a *cis*-eQTL, then the TReC and ASE methods should give consistent results and the TReCASE method should be used since it combines information from both TReC and ASE measurements, hence it should be more powerful. But if the underlying eQTL is a *trans*-eQTL, ASE provides no information regarding the eQTL and the TReC model alone should be used. Based on the above rationale, we can computationally distinguish a *trans*-eQTL from a *cis*-eQTL by hypothesis testing.

$$H_0\ (cis\text{−eQTL}): b_{x_j}^{(A)}=b_{x_j}^{(T)},\ \text{v.s.}\ H_1\ \ (trans\text{−eQTL}): b_{x_j}^{(A)} \neq b_{x_j}^{(T)},$$

where $b_{x_j}^{(T)}$ and $b_{x_j}^{(A)}$ denote the b_{x_j}'s estimated from the TReC model and ASE model, respectively. Let logLik_M be the log-likelihood of model M. We test the above hypothesis by a likelihood ratio test with test statistic: $-2(\text{logLik}_\text{TReCASE} - \text{logLik}_\text{TReC} - \text{logLik}_\text{ASE})$, where the log-likelihood under H_0 is $\text{logLik}_\text{TReCASE}$ and the log-likelihood under H_1 is

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$\log\text{Lik}_{\text{TReC}} + \log\text{Lik}_{\text{ASE}}$. We use the TReCASE model if H_0 cannot be rejected (i.e., treat the eQTL as a *cis*-eQTL), and the TReC model otherwise (i.e., treat the eQTL as a *trans*-eQTL).

3. Simulation Studies

The power of our methods is affected by multiple factors, such as the effect size and the minor allele frequency (MAF) of the target SNP. We first use simulation studies to compare the power of our methods and an existing approach: normalizing the TReC data by normal quantile transformation followed by linear regression (Pickrell et al., 2010). We simulate t_i, the total number of reads of a gene in the *i*-th sample, by a negative binomial distribution with mean parameter μ_{AA}, μ_{AB}, or μ_{BB} as specified later, and an over-dispersion parameter $\phi = 1.0$. The total number of allele-specific reads, denoted by n_i, is decided by a relation identified from our real data study: $n_i \approx 0.005 \times t_i$. The number of allele-specific reads from one haplotype is simulated by a beta-binomial distribution with mean decided by effect size and over-dispersion parameter $\theta = 0.1$. The values of ϕ and θ are decided by the results from the real data study. For the robustness of likelihood ratio test, the ASE model is only applied to those genes with 5 or more allele-specific reads in at least 5 samples.

In the first simulation setup, we assume sample size $N = 65$, mean values $\mu_{AA} = 500$, $\mu_{AB} = \mu_{AA}f_d$, and $\mu_{BB} = \mu_{AA}(2f_d–1)$, where f_d is the fold change of expression level with one minor allele. When the MAF of the target SNP is small (MAF=0.05, Figure 2 (a)), the TReC, ASE, and linear regression methods have similar power, although TReCASE apparently has higher (almost two-fold) power. When the MAF of the target SNP is moderate (MAF=0.2, Figure 2 (b)), both the TReC and ASE methods have higher power than linear regression. The TReCASE method again has significantly higher power than any other method. The simulation results also show that all methods control type I error at desired levels when the fold change is 1.0.

Suppose we want to improve the power of eQTL mapping given a fixed total number of reads in the experiment. We can either (1) fix the sample size and increase the number of reads per sample, (2) fix the number of reads per sample and increase the sample size, or (3) increase both sample size and the number of reads per sample. We carry out simulations for these three situations. Specifically, while the relation $\mu_{AB} = \mu_{AA}f_d$, and $\mu_{BB} = \mu_{AA}(2f_d - 1)$ remains the same, we consider (1) $\mu_{AA} = 1000$, $N = 65$, (2) $\mu_{AA} = 500$, $N = 130$, and (3) $\mu_{AA} = 650$, $N = 100$. The power of the TReC method increases as the sample size increases (Figure 3 (a)-(b)). However, the power of the ASE and TReCASE methods are similar if we either increase the sample size or the number of reads per sample (Figure 3 (c)-(f)), except for the TReCASE when the MAF is small, where increasing the sample size has a slight advantage over increasing the number of reads per sample (Figure 3 (e)).

Based on the results from Figure 2 and Figure 3, we can conclude that the TReCASE method has better power than any other method we considered. More importantly, using the TReCASE method, the power of *cis*-eQTL mapping can be improved by increasing the number of reads per sample instead of increasing the sample size. Except for simple experimental organisms such as yeast, the sample recruiting process is often very expensive. Thus given a fixed total number of reads per experiment, increasing the number of reads per sample (i.e. increasing read depth) could be much cheaper than increasing the sample size. Therefore application of the TReCASE method could have an important financial impact on designing eQTL experiments using RNA-seq data.

We have derived approximated formulas for the power of the TReC and ASE methods under the assumption that the eQTL effect size is relatively small (see Supplementary Materials (Section D)). Based on our derivation, the test statistics for the TReC and ASE methods are

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

affected by the number of reads through the factors $1/(1/t_i+\phi)$ and $1/(1/n_i+\theta)$, respectively. n_i is often much smaller than t_i. For example, in the real data study presented in the next section, $n_i \approx 0.005 \times t_i$. Therefore if $t_i = 600$, $n_i \approx 3$; and if we double the number of reads per individual, $t_i = 1200$ and $n_i \approx 6$. This only affects the denominator of TReC's test statistic by 1/1200, but affects the denominator of ASE's test statistic by 1/6. Therefore, increasing the read depth can lead to a limited power increase for the TReC model, but a significant power increase for the ASE model.

4. eQTL mapping for HapMap YRI samples

We downloaded the mapped RNA-seq reads of 69 lymphoblastoid cell lines from the Pritchard lab's website (http://eqtl.uchicago.edu/) (Pickrell et al., 2010). These 69 cell lines, which were derived from unrelated individuals from Yoruba in Ibadan (YRI), Nigeria, were part of the samples of the HapMap project (Frazer et al., 2007). Haplotype data were available for 65 of these cell lines, which were downloaded from the HapMap website (Thorisson et al., 2005) (version HapMap3_r2). These 65 samples were used for all the studies which follow. Among the 1,387,466 phased autosome SNPs, we used ~1.1 million (1,131,131) common SNPs with MAF larger than 0.05. The TReC data of 22,032 autosomal genes were downloaded from http://eqtl.uchicago.edu/RNA_Seq_data/results/final_gene_counts.gz. We calculated the ASE of these 22,032 genes using the R function asCount in our R package R/asSeq (see the Supplementary Materials (Section E) for details). The total number of sequence reads per sample varied from 2.7 million to 25.4 million with a median of 10.7 million. The total number of allele-specific reads per sample varied from 12 thousand to 135 thousand, with a median of 58 thousand (Supplementary Table 1). Overall, about 0.5% of the sequence reads were allele-specific reads (Supplementary Figure 1).

Seven confounding factors were included in the TReC model, the total number of reads per sample and six principal components derived from TReC data. No confounding factor was needed for the ASE model, since the ASE from one allele was directly compared with the other allele so that the effects of all the confounding factors in the TReC model were cancelled. Pickrell et al. (2010) have applied linear regression to identify eQTLs in this data using normal quantile normalized TReC data and 3.8 million SNPs in 69 samples. Due to the limitation of haplotype data, we considered eQTLs at 1.1 million phased common SNPs in 65 samples. To make the results from our methods directly comparable with the results from linear regression, we carried out eQTL mapping by linear regression as follows. We transformed the TReC data by normal quantile normalization, regressed out the effects of the seven confounding factors, and then applied normal quantile normalization again to obtain the normalized expression data. Finally we carried out eQTL mapping by linear regression using the normalized expression data together with the 1.1 million phased SNPs in 65 samples. Following Pickrell et al. (2010), for each gene, only local eQTLs within 200kb of the transcription start site were considered.

One gene was often associated with several local SNPs due to linkage disequilibrium among SNPs. To reduce such redundancy, we only considered the most significant local-eQTL for each gene. However, this strategy raised a multiple testing problem since different genes have different numbers of nearby SNPs, and thus the smallest local eQTL p-values across genes were not directly comparable. We corrected this multiple testing problem by calculating a permutation p-value for each gene. Specifically, for each gene we permuted its expression up to 5,000 times. In each permutation, we randomly shu ed the TReC data among individuals followed by a random switch of two ASE measurements per individual. Then the permutation p-value was calculated as the proportion of permutations where we observed more significant p-values than in the non-permuted data.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

As shown in Figure 4 (a), the TReC model clearly had higher power than linear regression, which we believe was due to the TReC model's capability to more accurately model counts with over-dispersion. The TReCASE model had even higher power than the TReC model, and the ASE model had the lowest power. The low power of the latter was mainly the consequence of the limited number of allele-specific reads, due to both the relatively low read depth and incomplete haplotype information. For the robustness of the likelihood ratio test, we employed an ad-hoc rule, to run the ASE model only if at least 5 samples had 5 or more allele-specific reads. With this restriction, the ASE model was only applied to 5,438 (~25%) genes. If there were not enough allele-specific reads to fit the ASE model, then the TReCASE model was degenerated to the TReC model. Therefore the small number of allele-specific reads also limited the power of the TReCASE model.

The results in Figure 4 illustrate the powers of the different methods across a range of permutation p-value cutoffs. In practice, we may choose a permutation p-value cuto by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995; Storey, 2003). We calculate FDR as $E(F\ D)/D$, where D is the number of discoveries at permutation p-value cuto p, and $E(F\ D)$ is the expected number of false discoveries. We estimate $E(F\ D)$ by $D\pi_0 p$, where π_0 is the expected proportion of null hypotheses across all tests (all the 22,032 genes in our case), and this was calculated as 2 times the proportion of genes with permutation p-values ≥ 0.5. By controlling FDR at 10%, linear regression, TReC, ASE, and TReCASE identified 690, 709, 435, and 815 genes with significant local eQTLs, respectively. For FDR 5%, the number of discoveries of linear regression, TReC, ASE, and TReCASE were 447, 483, 295, and 563, respectively.

Figure 4(b) and 4(c) illustrate the results of the ASE model and TReC model for one gene. The p-values for the TReC model were 3.5×10^{-2} and 6.7×10^{-5}, respectively, before and after correcting for confounding factors. This difference underlined the importance of accounting for the confounding effects in the TReC model. The p-value for the ASE model was 5.2×10^{-9}. In contrast, the TReCASE model had a p-value 1.9×10^{-12}, which was much more significant than either the ASE or TReC model. Three examples where eQTLs are missed by linear model, but identified by TReCASE model are presented at Supplementary Figure 2. The general pattern is that TReC model gives slightly more significant p-value than linear model, ASE p-value is comparable or smaller than TReC p-value, and the TReCASE p-value is much smaller than the p-value from TReC, ASE, or linear model.

Finally, we seek to answer an important question: among the local eQTLs, how many were *cis*-eQTL? As shown in Table 1, about 74% to 81% of the local eQTLs were *cis*-eQTLs, based on the hypothesis testing approach described in the Section 2.5 (p-value ≥ 0.05). The proportion of *cis*-eQTLs increased as the p-value cuto became more stringent. This trend was expected, since generally *cis*-eQTL affected the expression more directly than *trans*-eQTL, which often involved some other proteins (Rockman and Kruglyak, 2006), hence *cis*-eQTLs tended to have stronger effects.

5. Discussion

We have developed a statistical framework for eQTL mapping using RNA-seq data. Linear models are often used for eQTL mapping, while gene expression is measured using microarrays (Kendziorski and Wang, 2006). Since RNA-seq measurements are counts of sequence reads, it is natural and as shown in our study, more powerful, to model them by distributions for discrete variables, such as a negative binomial distribution or a beta-binomial distribution. More importantly, we have incorporated allele-specific expression information in our eQTL mapping method, which enables us to computationally distinguish

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

cis- and *trans*-eQTLs, and to further improve the power of *cis*-eQTL mapping. In our real data study, the gain of power is less significant than in our simulation studies; this is mainly due to an insufficient amount of allele-specific reads. One can increase the number of allele-specific reads by increasing the overall read depth and/or the read length. Meanwhile, methods that are particularly designed to enrich allele-specific reads are already available or are being developed (Pastinen, 2010).

When the TReC method identifies an eQTL, if the ASE method detects a consistent association, it is a *cis*-eQTL; if the ASE method detects no association, it is more likely a *trans*-eQTL. Our real data study shows it is very rare that both TReC and ASE models find significant but inconsistent associations (results not shown). However, we do observe situations in which significant associations are identified by the ASE model, but not by the TReC model. One possible explanation is that the allelic imbalance is due to imprinting (parent-of-origin) effects. With proper experimental design, e.g., family trio studies with complete haplotype information from both parents, our method could be further extended to identify such imprinting effects.

In addition to eQTL mapping using RNA-seq data, our method can also be applied to map the genetic loci underlying events captured by ChIP-seq data, for example, transcription factor binding sites (Zheng et al., 2010), open chromatin regions (McDaniell et al., 2010), or DNA methylations (Tycko, 2010). One extra challenge in such studies is that the sites of interest need to be identified from ChIP-seq data before mapping their variations to genetic loci.

In this paper, we have considered a simple eQTL mapping approach by associating each gene with each SNP. Following the work of eQTL mapping using microarray data, our method can be extended to multiple loci mapping (Sun et al., 2010), or simultaneous multiple loci mapping for all genes (Kendziorski et al., 2006; Gelfond et al., 2007; Jia and Xu, 2007; Pan, 2009; Chun and Keles, 2009), which are among our future research interests.

We have implemented our methods, the TReC, ASE, and TReCASE into an R package asSeq, which can be freely downloaded from http://www.bios.unc.edu/~wsun/software.htm. The computationally intensive parts of these methods were implemented by C. Our real data study, involving ~ 22,000 genes and ~ 1.1 million SNPs, and up to 5,000 permutations per gene, took about 100 hours on 64 CPUs with 2.8 GHz Intel EM64T processors. The computational load, though a ordable for most research institutes, is heavy due to the large number of permutations for each gene. We are currently working on possible approximations to improve the computational efficiency.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sun's research is supported in part by NIH RC2 MH089951-01, P50-MH090338-01, and the Gillings Innovative Laboratory in Statistical Genomics at UNC Chapel Hill. I am grateful to two anonymous reviewers and the editors' comments and suggestions, which led to significant improvements in this paper.

References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 1995; 57:289–300.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296:752–755. [PubMed: 11923494]

Byrd R, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 1995; 16:1190–1208.

Chun H, Keles S. Expression Quantitative Trait Loci Mapping With Multivariate Sparse Partial Least Squares Regression. Genetics. 2009; 182:79. [PubMed: 19270271]

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nature Reviews Genetics. 2009; 10:184–194.

Dean C. Testing for overdispersion in Poisson and binomial regression models. Journal of the American Statistical Association. 1992; 87:451–457.

Doss S, Schadt E, Drake T, Lusis A. Cis-acting expression quantitative trait loci in mice. Genome Research. 2005; 15:681. [PubMed: 15837804]

Fan H, Wang J, Potanina A, Quake S. Whole-genome molecular haplotyping of single cells. Nature Biotechnology. 2010; 29:51–57.

Frazer K, Ballinger D, Cox D, Hinds D, Stuve L, Gibbs R, Belmont J, Boudreau A, Hardenbol P, Leal S, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]

Gelfond J, Ibrahim J, Zou F. Proximity model for expression quantitative trait loci (eQTL) detection. Biometrics. 2007; 63:1108–1116. [PubMed: 17425636]

Griffiths D. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics. 1973; 29:637–648. [PubMed: 4785230]

Jia Z, Xu S. Mapping quantitative trait loci for expression abundance. Genetics. 2007; 176:611. [PubMed: 17339210]

Kendziorski C, Chen M, Yuan M, Lan H, Attie A. Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics. 2006; 62:19–27. [PubMed: 16542225]

Kendziorski C, Wang P. A review of statistical methods for expression quantitative trait loci mapping. Mammalian genome. 2006; 17:509–517. [PubMed: 16783633]

Kitzman J, MacKenzie A, Adey A, Hiatt J, Patwardhan R, Sudmant P, Ng S, Alkan C, Qiu R, Eichler E, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nature Biotechnology. 2010; 29:59–63.

McDaniell R, Lee B, Song L, Liu Z, Boyle A, Erdos M, Scott L, Morken M, Kucera K, Battenhouse A, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010; 328:235. [PubMed: 20299549]

Mortazavi A, Williams B, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008; 5:621–628. [PubMed: 18516045]

Pan W. Network-based multiple locus linkage analysis of expression traits. Bioinformatics. 2009; 25:1390. [PubMed: 19336446]

Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nature Reviews Genetics. 2010; 11:533–538.

Pickrell J, Marioni J, Pai A, Degner J, Engelhardt B, Nkadori E, Veyrieras J, Stephens M, Gilad Y, Pritchard J. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

Rockman M, Kruglyak L. Genetics of global gene expression. Nature Reviews Genetics. 2006; 7:862–872.

Ronald J, Brem R, Whittle J, Kruglyak L. Local regulatory variation in Saccharomyces cerevisiae. PLoS Genet. 2005; 1:e25. [PubMed: 16121257]

Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics. 2003; 31:2013–2035.

Sun W, Ibrahim J, Zou F. Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. Genetics. 2010; 185:349. [PubMed: 20157003]

Thorisson G, Smith A, Krishnan L, Stein L. The international HapMap project web site. Genome research. 2005; 15:1592. [PubMed: 16251469]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Tycko B. Allele-specific DNA methylation: beyond imprinting. Human Molecular Genetics. 2010; 1:R11.

Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics. 2009; 10:57–63.

Wittkopp P, Haerum B, Clark A. Evolutionary changes in cis and trans gene regulation. Nature. 2004; 430:85–88. [PubMed: 15229602]

Zheng W, Zhao H, Mancera E, Steinmetz L, Snyder M. Genetic Analysis of Variation in Transcription Factor Binding in Yeast. Nature. 2010; 464:1187. [PubMed: 20237471]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

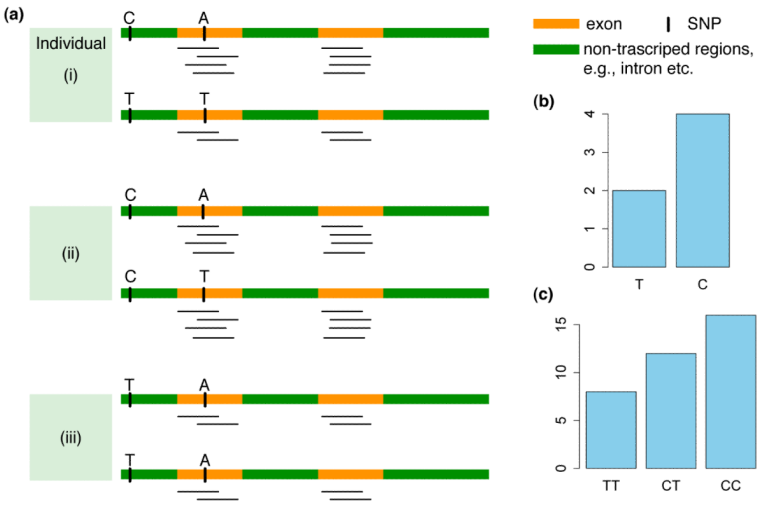

Figure 1.

A diagram to illustrate the RNA-seq count variation of one gene due to an *cis*-eQTL. **(a)** RNA-seq measurements of a hypothetic gene with two exons in three diploid individuals. The target SNP which we test for association has the genotype CT, CC and TT for the three individuals. There is a SNP on the first exon, which has genotype AT, AT, and AA for the three individuals. Allele-specific expression can be measured by those sequence reads that overlap with a heterozygous exonic SNP. Therefore we can measure allele-specific expression for individuals (i) and (ii). However, association testing by ASE is only possible if the target SNP is heterozygous, thus only individual (i) can be used to test for eQTL by ASE **(b)** ASE measurments for individual (i). **(c)** Total Read Count (TReC) measured for the three individuals across the two exons of this gene.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

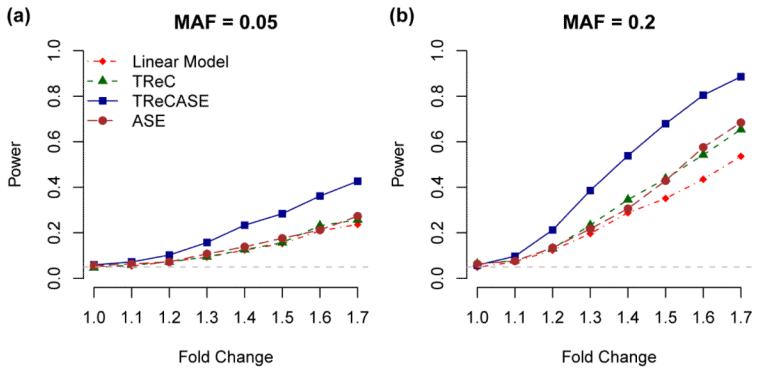

Figure 2.

Comparison of the power of four methods for eQTL mapping when the MAF of the target SNP is 0.05 (a) or 0.2 (b). P-value cutoff of 0.05 is used to call significance and power is calculated as the percentage of simulations where the p-values are smaller than 0.05, among 2,000 simulations. The horizontal dash line at the bottom of each figure corresponds to a power of 0.05. When the fold change is 1.0, all methods' power is approximately 0.05, which indicates that type I errors are controlled at a desired level by all methods.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

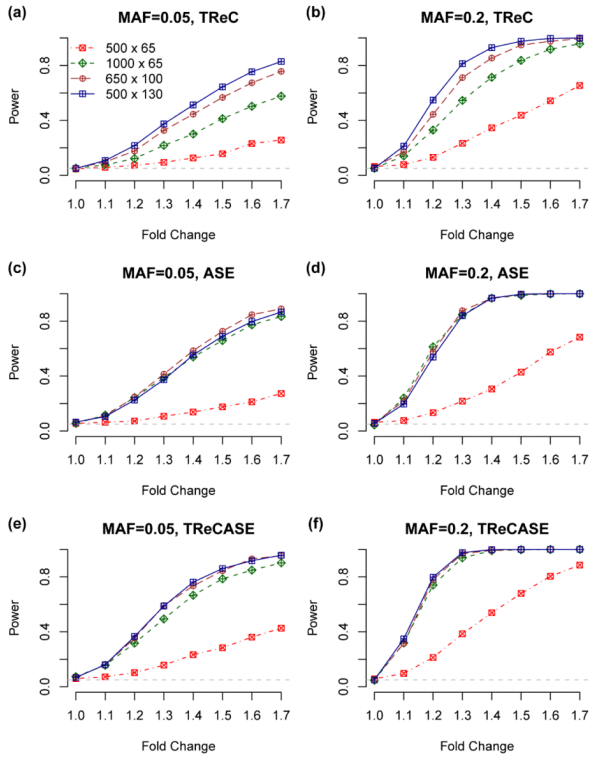

Figure 3.

Comparison of the powers of TReC, ASE, and TReCASE for eQTL mapping. “500 × 65” indicates the baseline situation that μ_{AA} = 500 and sample size N = 65. “1000 × 65”, “650 × 100” , and “500 × 130” indicate three strategies to improve power by increasing the number of reads per sample, increasing the sample size, or both. Similar to Figure 2, a p-value cut-off of 0.05 is used to call significance, and power is calculated as the percentage of simulations where the p-values are smaller than 0.05, among 2,000 simulations. The horizontal dash line at the bottom of each figure corresponds to a power of 0.05.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

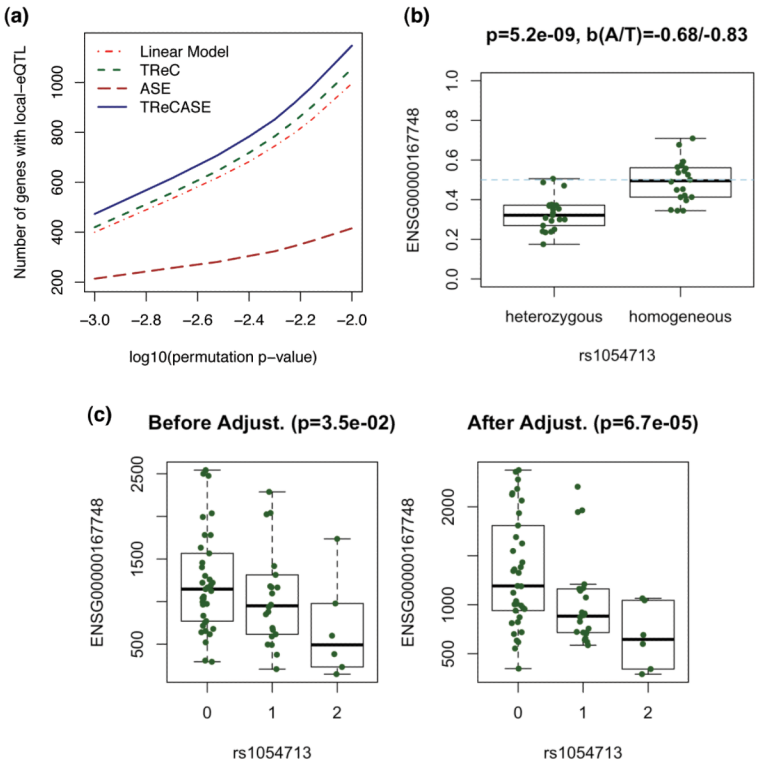

Figure 4.

(a) The number of local-eQTLs identified across permutation p-value thresholds. For each gene, only the most significant local-eQTL is kept and all the other local-eQTLs are discarded. (b) An example of eQTL mapping by the ASE model. b(A/T) indicates the regression coefficient estimates from the ASE model and TReC model, respectively. (c) An example of eQTL mapping by the TReC model. The X-axis is the genotype measured by the number of minor alleles, and the Y-axis is the number of reads per sample. Adjustment means to include seven confounding variables into the TReC model: the total number of reads per sample plus 6 PCs.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

The proportion of cis-eQTL among local eQTL.

p-value threshold for the TReC model	10^{-3}	10^{-4}	10^{-5}	10^{-6}
# of genes passed the TReC p-value threshold	6474	2055	838	412
# of genes with enough AS reads for *cis/trans* test	1093	399	174	83
# of genes with *cis*-eQTL by the *cis/trans* test	809	298	136	67
Proportion of *cis*-eQTL	0.740	0.747	0.782	0.807