ANALYSIS OF GENETIC, PARENT OF ORIGIN OR TREATMENT EFFECT ON GENE EXPRESSION USING RNA-SEQ DATA IN HUMAN AND MOUSE

Vasyl Zhabotynsky

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2020

Approved by:

Wei Sun

Fei Zou

Yun Li

Danyu Lin

Fernando Pardo Manuel de Villena

Naim Rashid

ABSTRACT

Vasyl Zhabotynsky: Analysis of Genetic, Parent of Origin or Treatment effect on gene expression using RNA-seq data in Human and Mouse
(Under the direction of Wei Sun and Fei Zou)

RNA sequencing allows us to systematically study allelic imbalance of gene expression, which may be due to genetic factors or genomic imprinting. In order to avoid confounding between genetic and parent-of-origin effects, and to improve the power to detect either effect, we have developed new statistical methods to jointly model both effects. In this dissertation, we consider a situation where modeling and separation of genetic and parent-of-origin effects are more challenging. First, we consider outbred populations such as human. We propose to collect RNA-seq data from children of family trios as well as phased genotype data for each member of those trios. Then we capture the genetic effects by *cis*-acting eQTLs and use the phased genotype data to define parent-of-origin effects. Next we propose a protocol for processing and analysis of RNAseq data with proper integration of total and allele-specific counts. We compare two major methods for final analysis as well as propose an efficient method for estimating permutation p-value. Finally we study for treatment, sex and additive genetic effect the reciprocal inbred crosses (RIX) produced from eight divergent inbred strains.

ACKNOWLEDGMENTS

UNC Department of Genetics

Dr. Paola Giusti and Dr. Terry Magnuson and Dr. Fernando Pardo Manuel de Villena

UNC Department of Biostatistics

Dr. Paul Little

Fred Hutchison Cancer Research Center

Licai Huang

Emory Department of Biostatistics and Bioinformatics

Dr. Yi-Juan Hu

NIEHS, RTI

Dr. Kaoru Inoue

UNC Department of Pharmacology

Dr. Mauro Calabrese

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

High-throughput RNA sequencing, known as RNA-seq, is one of the most popular techniques in the last decade for measuring gene expression abundance. In a typical RNA-seq experiment, for a given sample, tens of millions of sequence reads can be obtained, from which expression of each gene can be quantified as the number of reads mapped to the gene. RNA-seq offers several advantages over microarrays. For example, RNA-seq data are often less noisy with a larger dynamic range than microarray data. In addition, RNA-seq offers a great opportunity for identifying new transcripts while microarray's detection capability is limited by its probes (Mortazavi et al. 2008, Wang et al. 2009). Furthermore, RNA-seq is able to measure allele-specific expression (ASE) not otherwise available from microarray data. In diploid samples, every gene has two alleles: one paternally inherited and one maternally inherited. The transcript abundance of each allele (i.e., the ASE) allows one to dissect *cis*- and *trans*- regulations (Doss et al. 2005, Ronald et al. 2005). A few computational methods have been proposed to estimate genetic effect while combining total and allele-specific read counts (Sun 2012, Sun and Hu 2013, McVicker et al. 2013, Hu et al. 2015, Kumasaka et al. 2016, León-Novelo et al. 2014)). Different strategies have been proposed to estimate parent-of-origin effect, for example, screening using family pedigree (Morcos et al. 2011), or comparing the genotypes of mRNA vs. genomic DNA using genotyping microarrays (Barbaux et al. 2012). A method suggested by (Kumasaka et al. 2016) also allows for indirectly testing potential imprinting effects by searching for the genes that show allelic imbalance

across all samples while the genetic identity of silenced allele varies across individuals.

In studies of haloperidol a typical study involves a few mice of inbred strain, thus not allowing to jointly estimate genetic and treatment effect. Analysis done by (Kim et al. 2018) paper is on the larger side involving 28 C56BL/6j mice. In the setting of the reciprocal cross of two inbred strains, a method directly modeling genetic and parent-of-origin effects utilizing total and allele-specific counts was developed by (Zou et al. 2014). This approach opens new possibilities in studying drug effect on more diverse and at the same time well controlled mice population. For example, it allows to do joint estimation of genetic and treatment effect in total read counts. As was shown in (Zou et al. 2014) to incorporate parent-of-origin effects one also needs to add allele-specific expression, since in total read counts only model has identifiability issue.

1.1 Joint Estimation of Genetic and Parent-of-Origin Effects Using RNA-seq Data From Human Population

1.1.1 TReCASE approach

Total read count (TReC) can be compared across samples of different genotypes to quantify genetic effects on gene expression. Combining RNA-seq data and phased genotype of a diploid genome allows us to estimate allele-specific expression (ASE). Specifically, an RNA-seq read that overlaps with at least one heterozygous SNP can be assigned to one of the two alleles and thus contribute one allele-specific read count. (Sun 2012) developed the first statistical method to combine the data from TReC and ASE to improve the estimation of genetic effects, and the method is referred to as TReCASE. In TReCASE, TReC and allele-specific read counts are modeled by Negative-Binomial and Beta-Binomial distribution, respectively, and both distributions allow for over-dispersion. (Hu et al. 2015) has improved TReCASE method in several aspects. The phasing uncertainty between candidate eQTL and

exonic SNPs was appropriately modeled, the computing efficiency was greatly improved by using a score test, and a more rigorous approach was developed to distinguish *cis*- and *trans*-acting eQTLs.

For a particular gene of interest, let y_i be its total read count in the i-th sample where $i = 1, ..., N$. Let the allele-specific read counts for the two haplotypes be n_{i1} and n_{i2}, and $n_i = n_{i1} + n_{i2}$. If we want to test the association between this gene and the j-th SNP, we need to phase the genotype of SNP j and the gene of interest. For example, in a sample where the genotype of SNP j is AB, if B allele is on the same allele as the second haplotype, $n_{ijB} = n_{i2}$ and otherwise $n_{ijB} = n_i - n_{i2}$. Let $n_{i2'} = n_{ijB}$ if genotype of SNP j is AB, and let $n_{i2'} = n_{i2}$ otherwise. Then the allele-specific read counts can be modeled as:

$$f_{BB}(n_{i2'}; n_i, \pi_i, \varphi) = \binom{n_i}{n_{i2'}} \frac{\prod_{k=0}^{n_{i2'}-1}(\pi_i + k\varphi) \prod_{k=0}^{n_i - n_{i2'}-1}(1 - \pi_i + k\varphi)}{\prod_{k=1}^{n_i-1}(1 + k\varphi)}, \tag{1.1}$$

where over-dispersion φ provides additional flexibility for excessive variance. In the case of $n_i = 0$, we set $f_{BB}(n_{i2'}; n_i, \pi_i, \varphi) = 1$. The above model can be extended to include possible mapping bias.

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \tau_{null}\xi_{ij} + (1 - \xi_{ij})b_0^{(A)}, \tag{1.2}$$

$$\log\left(\frac{\pi_{null}}{1 - \pi_{null}}\right) = \tau_{null}, \tag{1.3}$$

where $\xi_{ij} = 1$ if SNP j is heterozygous for individual i, and 0 otherwise. π_{null} defines a mapping bias. Following similar approaches in earlier works, in this dissertation, we remove SNPs with strong mapping bias and then to assume $\pi_{null} = 0.5$ (or $\tau_{null} = 0$).

The total read counts generally can be modeled by a Negative Binomial distribution with mean value μ_i and dispersion parameter ϕ:

$$f_{NB}(y_i;\mu_i,\phi) = \frac{\Gamma(y_i + 1/\phi)}{y_i!\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i} \tag{1.4}$$

where μ_i is a function of p covariates $\beta_k, k = 1, ..., p$ such as average read depth in the i-th sample, sex, dominant genetic effect, and genetic effect for SNP j, denoted by η_{ij}:

$$\log(\mu_i) = \sum_{k=1}^{p} \beta_k c_{ik} + \eta_{ij}, \tag{1.5}$$

and

$$\eta_{ij} = \begin{cases} 0 & \text{if } g_{ij} = 0(AA) \\ \log\{1+\exp(b_j)\} - \log\{2\} & \text{if } g_{ij} = 1(AB) \\ b_j & \text{if } g_{ij} = 2(BB) \end{cases}$$

By the definition of *cis*-acting regulation, if SNP j is a *cis*-eQTL, $b_0^{(A)} = b_j$. In such case, a joint likelihood can be fit with common parameter b_{0j} shared by the likelihood for total and allele-specific read counts:

$$(\Theta) = \prod_{i=1}^{N} f_{BB}(n_{i2'}, n_i; b_{0j}, \varphi) f_{NB}(y_i; b_{0j}, \phi, \beta_1, ...\beta_p), \tag{1.6}$$

where $\Theta = (b_{0j}, \varphi, \phi, \beta_1, ...\beta_p)$. A likelihood ratio test or score test can be used to assess *cis*-eQTL effect by testing $H_0 : b_{0j} = 0$. In the case of *trans*-eQTL, $b_j \neq b_0^{(A)}$, and thus *cis*- and *trans*-eQTL can be distinguished by a testing $H_0 : b_0^{(A)} = b_j$.

1.1.2 Combined haplotype test (CHT)

Combined haplotype test (CHT) estimates additive genetic *cis*-effects while allowing over-dispersion (van de Geijn et al. 2015). The paper tests whether the genotype of SNP j is associated with read depth and allelic imbalance in a nearby

target region r for each pair $h = \{j, r\}$. Let α_h and β_h be two quantities that measure the gene expression from reference and alternative allele, respectively, so that the expected allelic imbalance in heterozygous case can be written as

$$p_h = \frac{\alpha_h}{\alpha_h + \beta_h}. \tag{1.7}$$

In this model, total read counts are modeled with Beta Negative Binomial distribution that has two over-dispersion parameters, one is specific for a target region, denoted by ϕ_r for region r, and one is specific for an individual, denoted by Ω_i for individual i. In this distribution, the expected number of read counts for individual i and test pair $h = \{j, r\}$ is modeled as

$$\eta_{hi} = \begin{cases} 2\alpha_h T_i & \text{if } g_{ij} = 0 \\ (\alpha_h + \beta_h) T_i & \text{if } g_{ij} = 1 \\ 2\beta_h T_i & \text{if } g_{ij} = 2 \end{cases}$$

where g_{ij} is the genotype of individual i at test SNP j, and T_i is the total number of reads genome-wide for individual i.

$$L(\alpha_h, \beta_h, \Omega_i, \phi_r | D) = \prod_{i=1}^{N} P_{BNB}(y_{ir} | \eta_{hi}, \Omega_i, \phi_r) \tag{1.8}$$

where y_{ir} is number of reads for individual i in target region r.

Allelic imbalance in allele-specific read counts are modeled using Beta-Binomial distribution with separately estimated individual level over-dispersion parameter φ_i. Denote the number of allele-specific read counts from reference allele as n_{1ik} for individual i and target SNP k, and the total number of allele-specific read count as

n_{ik}. Expected fraction of allele-specific reads from reference allele, denoted as p_h, is defined the same way as in equation (1.7), thus producing a likelihood

$$L(\alpha_h, \beta_h | D) = \prod_{i=1}^{N} \prod_{k} \underset{BB}{P}(n_{1ik} | n_{ik}, p_h, \varphi_i) \tag{1.9}$$

Furthermore, this method adjusts for incorrect calls of SNP genotypes by creating a mixture of two Beta-Binomial distributions with H_{ik} - probability that individual i is heterozygous at SNP k.

$$\begin{aligned} \underset{BB-mix}{P}(n_{1ik} | n_{ik}, p_h, \varphi_i, H_{ik}) = H_{ik} \underset{BB}{P}(n_{1ik} | n_{ik}, p_h, \varphi_i) + \\ (1 - H_{ik}) \left[\underset{BB}{P}(Y = n_{1ik} | n_{ik}, p_{err}, \varphi_i) + \underset{BB}{P}(n_{1ik} | n_{ik}, 1 - p_{err}, \varphi_i) \right] \end{aligned} \tag{1.10}$$

with this model for allele-specific counts the joint likelihood is set to

$$(\alpha_h, \beta_h, \phi_r | D) = \prod_{i=1}^{N} \left[\underset{BNB}{P}(y_{ir} | \eta_{hi}, \hat{\Omega}_i, \phi_r) \prod_{k} \underset{BB-mix}{P}(n_{1ik} | n_{ik}, p_h, \hat{\varphi}_i, \hat{H}_{ik}) \right] \tag{1.11}$$

Individual level over-dispersion parameters as well as probability of incorrect heterozygous SNP call for this joint model are estimated separately.

To test for additive genetic *cis*-effect one needs to test a hypothesis $H_0 : \alpha_h = \beta_h$ versus the two sided alternative with LRT.

1.1.3 RASQUAL approach

RASQUAL is another recent method for eQTL mapping while combining total and allele specific read counts (Kumasaka et al. 2016). RASQUAL models total read counts by a Negative Binomial distribution and allele-specific read counts for each SNP within a feature (e.g., a gene) by a Beta-Binomial distribution, while searching and adjusting for potential mapping bias. For a feature, suppose we observe y_i: total

read count in sample i, and for the l-th SNP within the feature, we observe n_{il}: allele-specific count at SNP l, and n_{1il}: the number of alternative read counts at this SNP. We refer to such a SNP within the feature of interest as a feature SNP (fSNP). Let K_i be sample specific offset term reflecting library size and other factors for individual i, and assume it is estimated *a priori*. Assume there is a single *cis*-regulatory SNP (rSNP) with imbalance parameter π. The read count of this feature is modeled by a Negative Binomial distribution with mean value μ_i:

$$\mu_i = \begin{cases} 2(1-\pi)\lambda K_i & \text{if } g_i = 0 \\ \lambda K_i & \text{if } g_i = 1 \\ 2\pi\lambda K_i & \text{if } g_i = 2 \end{cases}$$

where λ is a scale parameter for mean expression, and genotype g_i taking values 0, 1 or 2 based on number of alternative alleles at the rSNP.

For this *cis*-effect, the expected allelic ratio that we measure at available heterozygous fSNPs for a heterozygous or a homozygous rSNP is $\{1-\pi, \pi\}$ or $\{0.5, 0.5\}$, respectively. These allele-specific counts are modeled with Beta-Binomial distribution. Assuming such mean structure, adding δ - the probability that an individual read maps to an incorrect location, π_{null} - reference mapping bias ($\pi_{null} = 0.5$ corresponds to no reference bias) and assuming common over-dispersion parameter ϕ for both Negative Binomial and Beta Binomial distributions, the respective joint likelihood is

$$(\pi, \delta, \pi_{null}, \phi) = \prod_{i=1}^{N} \left[\sum_{g_i} p(g_i) p_{NB}(y_i | g_i; \pi, \lambda_i, \phi) \right. \\ \left. \prod_{l=1}^{L} \left(\sum_{D_{il}} p(D_{il}|g_i) p_{BB}(n_{1il} | n_{il}, D_{il}; \pi, \delta, \pi_{null}, \phi) \right) \right], \quad (1.12)$$

Table 1.1: Diplotype definition for the rSNP and the l-th fSNP in the i-th individual. 0 and 1 indicates the reference allele and alternative allele, respectively. For example, genotype of rSNP is (0,0) means it is homozygous reference alleles. In the definition of diplotype, the order of the two haplotypes can be switched, i.e, h_1/h_2 and h_2/h_1 are the same diplotype. Each diplotype may correspond to multiple combination of allele-specific genotypes. For example, in the 4-th row, if the genotype for rSNP and fSNP are (0,1) and (0,0), the first haplotype is 00, and the second haplotype is 10. If the genotype for rSNP and fSNP are (1,0) and (0,0), the first haplotype is 10, and the second haplotype is 00. Both cases correspond to the diplotype 10/00.

Allele-specific genotype for rSNP and the l-th fSNP: $\{g_i, g_{il}\}$	**Diplotype: $D_{il} = h_1/h_2$, where h_1 and h_2 are two haplotypes**
{(0,0),(0,0)}	00/00
{(0,0),(0,1)} or {(0,0),(1,0)}	00/01
{(0,0),(1,1)}	01/01
{(0,1),(0,0)} or {(1,0),(0,0)}	10/00
{(0,1),(1,0)} or {(1,0),(0,1)}	10/01
{(0,1),(0,1)} or {(1,0),(1,0)}	11/00
{(0,1),(1,1)} or {(1,0),(1,1)}	11/01
{(1,1),(0,0)}	10/10
{(1,1),(1,0)} or {(1,1),(0,1)}	10/11
{(1,1),(1,1)}	11/11

where D_{il} defines the diplotype configuration in individual i between the rSNP and the l-th fSNP (Table 1.1). The genotype and haplotype probabilities $p(g_i)$ and $p(D_{il}|g_i)$ are obtained from SNP phasing and imputation. In addition to common over-dispersion parameter, RASQUAL assumes that neither incorrect mapping δ nor reference mapping bias ϕ influence total read count.

1.1.4 RASQUAL approach to imprinting testing

Without distinguishing paternal and maternal allele, RASQUAL cannot be used to estimate parent-of-origin effect. However, the authors suggest that RASQUAL can be used to search for potential imprinted genes by searching for genes where all samples

show allelic imbalance, while the directions of allelic imbalance cannot be explained by the genetic identities of the alleles. Specifically, they model the unknown parent-origin by a hypothetical rSNP that is heterozygous in all samples. Let g_i be the genotype of this hypothetical rSNP, then

$$p(g_i) = \begin{cases} 0 & g_i = 0 \\ 1 & g_i = 1 \\ 0 & g_i = 2. \end{cases}$$

Assuming the genotype of any feature SNP is independent with the genotype of this hypothetical rSNP, then the probably of diplotype is

$$p(D_{il}|g_i) = \begin{cases} p(g_{il} = 0) & D_{il} = 00/10 \\ 0.5p(g_{il} = 1) & D_{il} = 01/10 \\ 0.5p(g_{il} = 1) & D_{il} = 00/10 \\ p(g_{il} = 2) & D_{il} = 01/11 \\ 0 & \text{otherwise} \end{cases}$$

Authors call a region to be imprinted if the p-value of association with this hypothetical rSNP is lower than the p-values for any QTL, and if estimated effect size π is extreme (> 0.9 or < 0.1).

1.1.5 Approaches in estimating parent-of-origin effect

Morcos et al. (2011) use genotyping arrays to assess allelic imbalance using both RNA and genomic DNA (gDNA) samples to identify potential imprinted genes estimating allelic imbalance fraction coming from one of the alleles at a given SNP

after which if in at least 3 consecutive SNPs average deviation exceeded a 1 SD threshold in at least two samples. Those potential imprinted genes are then screened using family pedigree.

Similarly, Barbaux et al. (2012) used genotyping microarrays to compare mRNA/cDNA vs. genomic DNA to identify new genes presenting mono-allelic expression: such mono-allelic status at the informative SNP was defined as (1) having a heterozygosity on gDNA, (2) apparent homozygosity on the corresponding cDNA in at least two samples (out of five tested) which were further experimentally validated.

Finally, in experimental settings for F_1 cross of highly divergent mouse strains, (Zou et al. 2014) have jointly modeled genetic and parent-of-origin effect. More details on this method are to be described in the next section.

1.2 Modeling Additive, Sex and Treatment Effects in Diverse Recombinant Inbred Cross (RIX)

Typical study of haloperidol effect on mice prior to RIX cross design was done using small sample size of inbred mice. Analysis done by (Kim et al. 2018) paper was done on total expression of 28 C56BL/6j mice. In this experiment haloperidol was studied using similar tissue (striatum) and relatively large sample size - several other recent papers studying haloperidol effect in rodent models are presented in (Kim et al. 2018) Supplementary Table 1 with typical sample size under 10 mice from 6 to 24 mice. The results of previous studies show extreme variability in number of reported significant genes with high variability in number of up-regulated versus down-regulated genes. (Kim et al. 2018) in their study have shown importance of using striatum - even in whole brain tissue authors were able to discover smaller number of treatment effects.

For such experiments quite statistical model design is relatively straightforward due to the fact that no genetic effect can be observed in inbred mice as well as typically

only one sex is chosen for experiment, thus leaving only treatment effect as a main effect of interest. Total expression in (Kim et al. 2018), for example, was fit using Negative-Binomial model as implemented in EdgeR package with gene-wise dispersion. Drug effect was tested using likelihood ratio test.

Incorporating additive and sex effect in more diverse populations such as RIX cross can add both to the stability of estimate and improve identifiability due more diverse population. RNA-seq data collected from F_1 reciprocal crosses of inbred strains allows a straightforward joint modelling of treatment and genetic effects. It also allows attribution of allele-specific reads to one of the parents and, the knowledge of parent's genotype allows to reduce mapping bias at the mapping stage and thus concentrate on estimating of effects of interest. In (Zou et al. 2014), authors define a model for N_1 F_1 mice and N_2 inbred mice - allele-specific counts would be not available for inbred mice since parental crosses are exactly the same. For a given gene total expression y_i is observed, and for each F_1 mouse n_{iB} and n_i represent number of observed allele specific reads from cross B and overall number of allele specific reads (i.e. $n_i = n_{iA} + n_{iB}$ for samples $i = 1...N_1$. Assuming that b_0 represents additive genetic effect and b_{poo} represents parent-of-origin effect, allele-specific reads can be modeled using following Beta-Binomial distribution

$$f_{BB}(n_{iB}; n_i, b_0, b_{poo}, \varphi) = \binom{n_i}{n_{iB}} \frac{\prod_{k=0}^{n_{iB}-1}(\pi_i + k\varphi)\prod_{k=0}^{n_i-n_{iB}-1}(1-\pi_i+k\varphi)}{\prod_{k=1}^{n_i-1}(1+k\phi)}, \qquad (1.13)$$

In this model over-dispersion φ provides additional flexibility for excessive variance and overall model can be represented with a certain π_i that incorporates both genetic and parent-of-origin effects as

$$\log\left(\frac{\pi_{iB}}{1-\pi_{iB}}\right) = b_0^{(A)} + b_{poo}x_i \tag{1.14}$$

$$\tag{1.15}$$

$x_i = 1$ if parent of strain B is father and $x_i = -1$ if parent of strain B is mother.

The total counts y_i generally can be modeled with certain mean μ_i and dispersion parameter ϕ as:

$$f_{NB}(y_i;\mu_i,\phi) = \frac{\Gamma(y_i + 1/\phi)}{y_i!\Gamma(1/\phi)}\left(\frac{1}{1+\phi\mu_i}\right)^{1/\phi}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i} \tag{1.16}$$

where mean can incorporate variety of covariates as well as a link to genetic and parent-of-origin effects η_i

$$\log(\mu_i) = \sum_{k=1}^{p}\beta_k c_{ik} + \eta_i, \tag{1.17}$$

and

$$\eta_i = \begin{cases} 0 & i \in AA \\ \log\{1+\exp(b_0+b_{poo})\} - \log\{1+\exp(b_{poo})\} & i \in AB \\ \log\{1+\exp(b_0-b_{poo})\} - \log\{1+\exp(-b_{poo})\} & i \in BA \\ b_0 & i \in BB \end{cases}$$

The joint model can be written

$$L(\Theta) = \prod_{i=1}^{N_1} f_{BB}(n_{iB}, n_i; b_0, b_{poo}, \varphi) \prod_{i=1}^{N_1+N+2} f_{NB}(y_i; b_0, b_{poo}, \phi, \beta_1, \ldots\beta_p), \tag{1.18}$$

In case of *cis*-effect $b^{(A)} = b_0$ in such case a joint likelihood can be fitted with common

parameter $b_0 = b_0^{(A)}$ for total and allele-specific counts.

The model this way allows to perform multiple tests including a test for genetic *cis*-effect $H_0 : b_0^{(A)} = b_0$ vs two sided alternative.

Similarly to described above extension to allow to fit different genetic and parent-of-origin effects for total and allele specific counts the full model also allows for effects to be different for male and female mice $b_{0F}^{(A)}, b_{0M}^{(A)}$ and $b_{0F}, b_{0M}, b_{pooF}, b_{pooM}$,

The paper discusses the special case of X chromosome, for which the model allows to both adjust for allelic bias on X chromosome in female mice (due to the known partial imprinting called Xce effect) by adding a sample level bias analogous to π_{null} in TReCASE subsection denoted as π_{iXce} in their model defined as chromosome level expression of allele from parent B, as well as adjusting male samples due to the fact that they have only one copy of X chromosome. Setting $\log\left(\frac{\pi_{iXce}}{1-\pi_{iXce}}\right) = \tau_{iXce}$ these adjustments produce a modified mean structure for corresponding crosses in female mice:

$$\log\left(\frac{\pi_{iB}}{1-\pi_{iB}}\right) = \tau_{iXce} + b_{0F}^{(A)} + b_{pooF} x_i \tag{1.19}$$

$$\tag{1.20}$$

$x_i = 1$ if B is paternal and $x_i = -1$ if B is maternal.

and η_i for total read counts to

$$\eta_i = \begin{cases} 0 & i \in AA \\ \log\{1+\exp(\tau_{iXce} + b_{0F} + b_{pooF})\} - \log\{1+\exp(b_{poo})\} + \log\{2\pi_{iXce})\} & i \in AB \\ \log\{1+\exp(\tau_{iXce} + b_{0F} - b_{pooF})\} - \log\{1+\exp(-b_{poo})\} + \log\{2\pi_{iXce})\} & i \in BA \\ b_{0F} & i \in BB \end{cases}$$

while no allele-specific counts for males and since males always have maternally inherited X chromosome which is upregulated (compared to same strain in female mouse) η_i set to

$$\eta_i = \begin{cases} \log\{2\} - \log\{1 + \exp(-b_{pooF})\} & i \in AA, AB \\ b_{0M} - \log\{1 + \exp(-b_{pooF})\} & i \in BB, BA \end{cases}$$

In this model sample level Xce effect is estimated separately by taking a median allele-specific expression among non-fully imprinted X-chromosome genes of that individual.

Additionally, model discusses an issue of identifiability of the parent-of-origin effects in case when only total read counts are available - when no additive effect exists of using total read counts only. In such case if no additive effect exists a parent-of-origin effect defined in model 1.16 is not identifiable. Plugging b_{0F} produces all four variants of η_i to be 0. Thus, for the cases with no allele-specific counts (or very low counts) the only test that can be done is to reduce to a simpler version similar to TReCASE approach discussed at the beginning of the review:

$$\log(\mu_i) = \sum_{k=1}^{p} \beta_k c_{ik} + \beta_{dev} x_i + \eta_i, \tag{1.21}$$

In the equation above we essentially take x_i as it was defined earlier, except now we extend its definition to inbred mice by defining that it is 0 for crosses AA or BB and update

$$\eta_i = \begin{cases} 0 & i \in AA \\ \log\{1 + \exp(b_0)\} - \log\{2\} & i \in AB, BA \\ b_0 & i \in BB \end{cases}$$

One still can test whether there is a deviation from additivity with $H_0 : \beta_{dev} = 0$, however if we observe additive genetic effect $b_0 = 0$ interpretation of deviation from additivity as resulting from parent-of-origin effect becomes spurious as it was mentioned expected values for all four crosses are expected to be the same if $b_0 = 0$. Same remains true for X chromosome.

CHAPTER 2: JOINT ESTIMATION OF GENETIC AND PARENT-OF-ORIGIN EFFECTS UNDER FAMILY TRIO DESIGN

2.1 Method

We assume genetic effects can be captured by a limited number of *cis*-acting eQTLs. Here we define *cis*-eQTLs as those eQTLs that influence allelic-imbalance of gene expression (Sun and Hu 2013). This is a reasonable assumption because most *cis*-acting eQTLs are local eQTLs and because of the limited linkage dis-equilibrium (LD) structure around a gene, the number of independent *cis*-eQTLs of a gene is relatively small. In the following, we assume there is only one *cis*-eQTL to simplify the discussion. Our method can be easily extended to the cases with multiple *cis*-eQTLs.

To help explain the motivation of the model choice and parametrization of our method, we start by a toy example to illustrate how *cis*-eQTL and PoO factors affect both total expression and ASE. Consider the expression of one gene in four individuals, and a *cis*-eQTL of this gene with ordered genotype CC, TT, TC or CT with the first allele listed being inherited from mother and the second from father (Figure 3.1). The genetic effect is that C allele has three times expression of T allele, and the PoO effect is that paternal allele has twice of the expression of maternal allele. It is apparent that there are allelic imbalance in all four individuals. The degree of allelic imbalance depends on both *cis*-eQTL and PoO effects and it is very challenging to discern them without knowing the parent-of-origin of each allele. This toy example also demonstrates that in addition to ASE, both *cis*-eQTL and PoO factors can affect

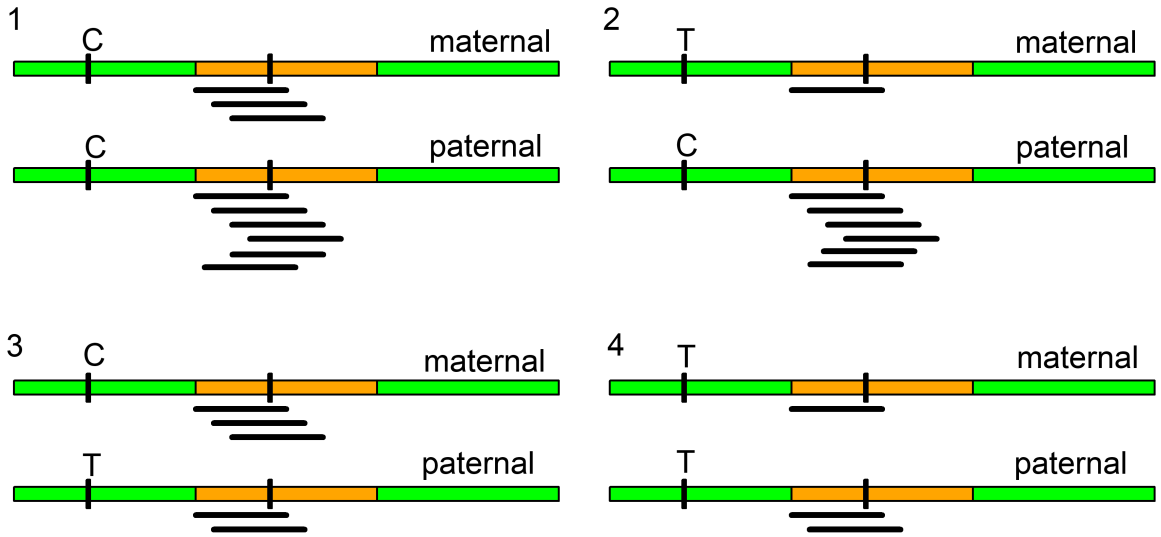

Figure 2.1: Recovering allele-specific signal from RNA-seq data. The diagram illustrates *cis*-eQTL and parent-of-origin effects on gene expression in four individuals (individuals 1-4). The yellow/green boxes indicate exonic/intronic regions, respectively. Assume there is a heterozygous SNP in the exonic region of this hypothetical gene in all four individuals, and thus we can quantify ASE. There is a *cis*-eQTL with C or T allele, and its C allele has three times of expression of T allele. In addition, the paternal copy has twice expression of the maternal copy.

total expression. *Cis*-eQTL's effect on total expression is more apparent since total expression decreases as genotype shift from CC, CT/TC, to TT. PoO modifies the total expression of the two individuals with ordered genotype CT and TC. Without PoO effect, these two individuals should have equal amount of TReC.

2.1.1 Allele specific counts

To simplify the notation, we assume the haplotypes connecting a candidate eQTL and the gene of interest are known. Let n_{i1} and n_{i2} be the allele-specific read counts of the two haplotypes of the gene of interest in the i-th individual (denoted by h_{i1} and h_{i2}), respectively, where $i = 1, ..., N$. Let $n_i = n_{i1} + n_{i2}$. Denote the two alleles of the candidate eQTL as A_1 and A_2, and denote its genotype in the i-th individual as g_i.

We assign different meanings to genotypes A_1A_2 and A_2A_1 such that A_1A_2 means haplotypes h_{i1} and h_{i2} harbor the A_1 and A_2 alleles, respectively, and A_2A_1 means haplotypes h_{i1} and h_{i2} harbor the A_2 and A_1, respectively. We model n_{i1} by a beta-binomial distribution (denoted by f_{BB}):

$$n_{i1} \sim f_{BB}(n_{i1}; n_i, \pi_i, \varphi), \quad \log\left[\pi_i/(1-\pi_i)\right] = b_0 z_i + b_1 x_i, \tag{2.1}$$

where φ is over-dispersion parameter, and z_i and x_i are defined as follows:

$$z_i = \begin{cases} 0 & \text{if } g_i = A_k A_k,\ k = 1 \text{ or } 2 \\ 1 & \text{if } g_i = A_2 A_1 \\ -1 & \text{if } g_i = A_1 A_2; \end{cases}$$

$$x_i = \begin{cases} 1 & \text{if haplotype } h_{i1} \text{ is inherited from the paternal allele,} \\ -1 & \text{if haplotype } h_{i1} \text{ is inherited from the maternal allele.} \end{cases}$$

2.1.2 Total read counts (TReCs)

The TReC of the gene of interest in the i-th individual, denoted by y_i, is modeled by a negative binomial distribution with mean μ_i and over-dispersion parameter ϕ, denoted by $f_{NB}(y_i; \mu_i, \phi)$. We can write the mean structure for negative binomial distribution as:

$$\log(\mu_i) = \gamma_0 + \beta_\kappa \log(\kappa_i) + \sum_{u=1}^{p} \beta_u c_{iu} + \eta_i, \tag{2.2}$$

where β_u, $u = 1, ..., p$, is the regression coefficient for the u-th covariate (e.g., age, gender, batch effects, principal component for population stratification, or surrogate

variables for latent batch effects). η_i is defined as:

$$\eta_i = \begin{cases} 0 & \text{if } g_i = A_1A_1 \\ \log\{1+\exp(b_0+x_ib_1)\} - \log\{1+\exp(x_ib_1)\} & \text{if } g_i = A_1A_2 \text{ or } A_2A_1 \\ b_0 & \text{if } g_i = A_2A_2 \end{cases}$$

Note that the additive genetic and parent-of-origin effects are parametrized by b_0 and b_1, which are the same as the b_0 and b_1 in equation (2.1) for allele-specific read counts.

2.1.3 Joint likelihood

The joint likelihood of total read count (TReC) and ASE is

$$(\Theta) = \prod_{i=1}^{K} f_{BB}(n_{i1}, n_i; b_0, b_1, \varphi) f_{NB}(y_i; b_0, b_1, \phi, \gamma, \beta_\kappa, \beta_1, ...\beta_p), \tag{2.3}$$

where $\Theta = (b_0, b_1, \phi, \varphi, \gamma, \beta_\kappa, \beta_1, ...\beta_p)$. In this model we assume common genetic and parent-of-origin effect for TReC and ASE. We test the genetic and parent-of-origin effects by testing whether b_0 or b_1 equals to 0, respectively, with the likelihood ratio test.

2.1.4 Optimization Algorithm

For a given initial values for non-linear terms (ϕ, φ, b_0, b_1)

Step 1: Optimize linear terms given the initial values of non-linear terms:

$$\beta_{r+1} = \beta_r + (X'W_rX)^{-1}(X'W_rk_r), \text{ diag}(W_r) = \frac{\mu_r}{1+\phi_r^{-1}\mu_r} \quad \text{and} \quad k_r = \frac{y_r - \mu_r}{\mu_r},$$

where W_r is a diagonal matrix.

Step 2: Iteratively estimate b_0 and b_1. Note that the following two steps are

redundant, but improves the robustness of the algorithm.

1. Optimize b_0 and b_1 together using BFGS method.

2. Optimize b_0 and b_1 separately using Brent algorithm.

Step 3: optimize over-dispersion parameter $\log(\phi)$ and $\log(\varphi)$ separately using Brent algorithm. For stability we limit the range of over-dispersion to be between 10^{-4} and 10^4.

Step 4: If likelihood change is larger than a small number ϵ, go to step 1, otherwise finish the estimation process.

2.2 Simulations

We simulated ASE and TReC from the model described by joint likelihood in equation (2.3). We chose the smallest sample size to be 32, which is similar to the dataset size we have for real data analysis. To demonstrate the asymptotic properties, we also simulated data of larger sample sizes of 64, 128, and 256. We set the over-dispersion parameters for beta-binomial (φ) and negative binomial (ϕ) distributions to be 1/4, and 4/3, respectively, which are fairly typical in real data. We further scale the expected counts so that mean total read count was about 250. We set the proportion of reads that are allele-specific to be 10% which is reasonably close to observed value in our real data. Our simulation results show that the estimates of b_0 and b_1 from our method are unbiased. In contrast, if a model is fit with only genetic or parent-of-origin effect while in fact both effects are present, the parameter estimates have significant bias (Figure 2.2).

We also compared the type I error and power of our joint model versus naive fit using *R/glm.nb* function to fit Negative-Binomial model using total read counts only and *R/vglm* function to fit Beta-Binomial model using allele specific counts only. As

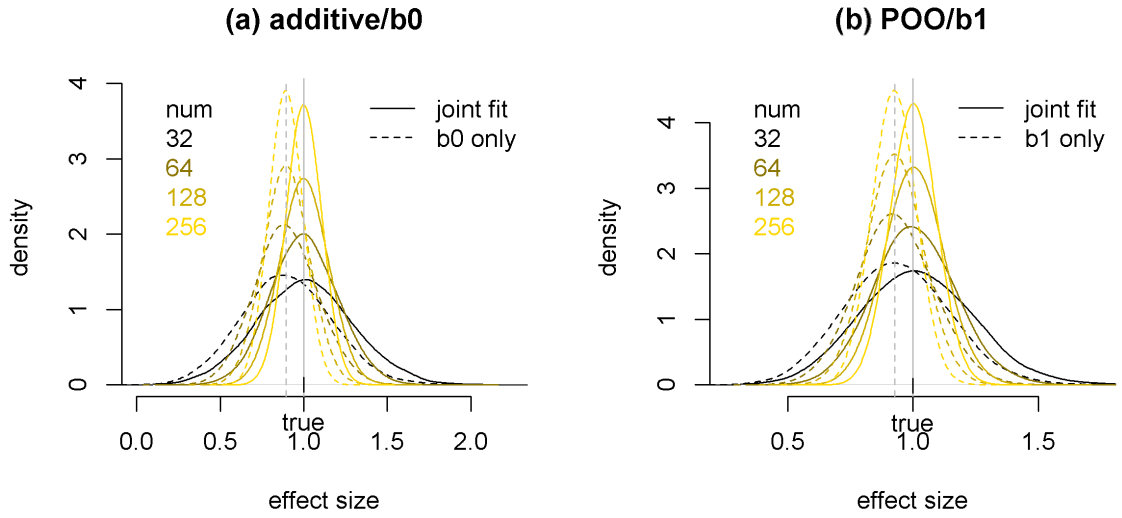

Figure 2.2: Observed bias in model fitting marginal models. Fitting the joint model versus (a) additive genetic effect only or (b) parent-of-origin effect only for various sample sizes. The legend on the left side of each figure shows sample sizes. The solid line shows density estimates obtained from joint estimation of additive and parent-of-origin effect, while dash line shows density estimates when only one effect is fitted (i.e. additive only for left panel and parent-of-origin only for right panel)

shown in Table 2.1, the simple models don't control type I error as well as joint model and they also have lower power to detect either genetic or parent-of-origin effect than the joint model.

The statistical power of the joint model is illustrates in Figure 2.3. Even at sample size of 32, our method has around 80% of power to detect either genetic or parent-of-origin effect at two-fold change, which corresponds to effect size $\log(2) = 0.693$. The power to detect parent-of-origin effect is higher than the power to detect genetic effect. This is because the ASE of all samples can be used to quantify parent-of-origin effect (i.e., comparing paternal vs. maternal allele). In contrast, ASE can be used to quantify genetic effect only if the genotype of the candidate eQTL is heterozygous, so that we can compare the expression of one allele versus the other

Table 2.1: Power Analysis

Parameters		Genetic			Parent-of-Origin		
b_0	b_1	Joint Model	Negative Binomial	Beta Binomial	Joint Model	Negative Binomial	Beta Binomial
0.00	0.00	0.06	0.11	0.07	0.06	0.12	0.07
0.13	0.13	0.09	0.12	0.08	0.11	0.11	0.10
0.25	0.25	0.19	0.14	0.11	0.27	0.11	0.18
0.50	0.50	0.53	0.21	0.21	0.72	0.12	0.46
0.75	0.75	0.83	0.32	0.33	0.96	0.14	0.76
1.50	1.50	1.00	0.76	0.49	1.00	0.39	0.93

allele.

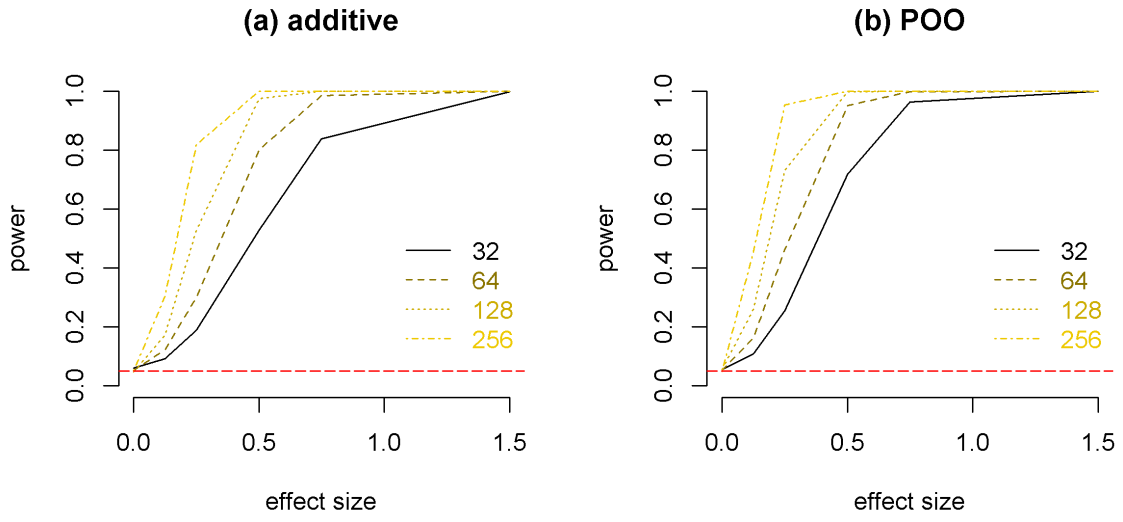

Figure 2.3: The statistical power of joint model. Figure shows the ability of detecting (a) additive genetic effect or (b) parent-of-origin effects for various sample sizes. The horizontal red dash line indicates the p-value cutoff 0.05 to declare statistical significance.

2.2.1 Model misspecification due to only a fraction of individuals having imprinting effect

We also consider a scenario when only a fraction of individuals doesn't have imprinting effect (assuming that they still have the same genetic effect). To do it we

simulate the model as described in method section, but for a randomly chosen fraction we replace b_1 effect with 0. In simulation we included true model, 5%, 10%, 20% and 40% individuals with no imprinting effect. While we simulated the data for sample size 32, 64 and 128, since they show similar results for the future illustrations we chose sample size 32. We do observe that increased fraction of not imprinted individuals leads to larger bias and lower power (Figure 2.4). We also observe that type one error is reasonably controlled in such scenario for all the fractions we used in our simulations. We can see that higher fraction of individuals with no parent of origin effect leads to drop of power with more pronounced drop in parent of origin effect (panel (f)). Additionally, it is clear that the higher is value of the other effect the higher drop of power gets, especially notable for additive genetic effect (panel (e)). At least in part this drop in power is due to the fact that only a fraction of samples have a non-zero effect, and the rest (up to 40%) don't have effect. It could be of interest for the future work to see how much power would improve under a properly specified model for this scenario.

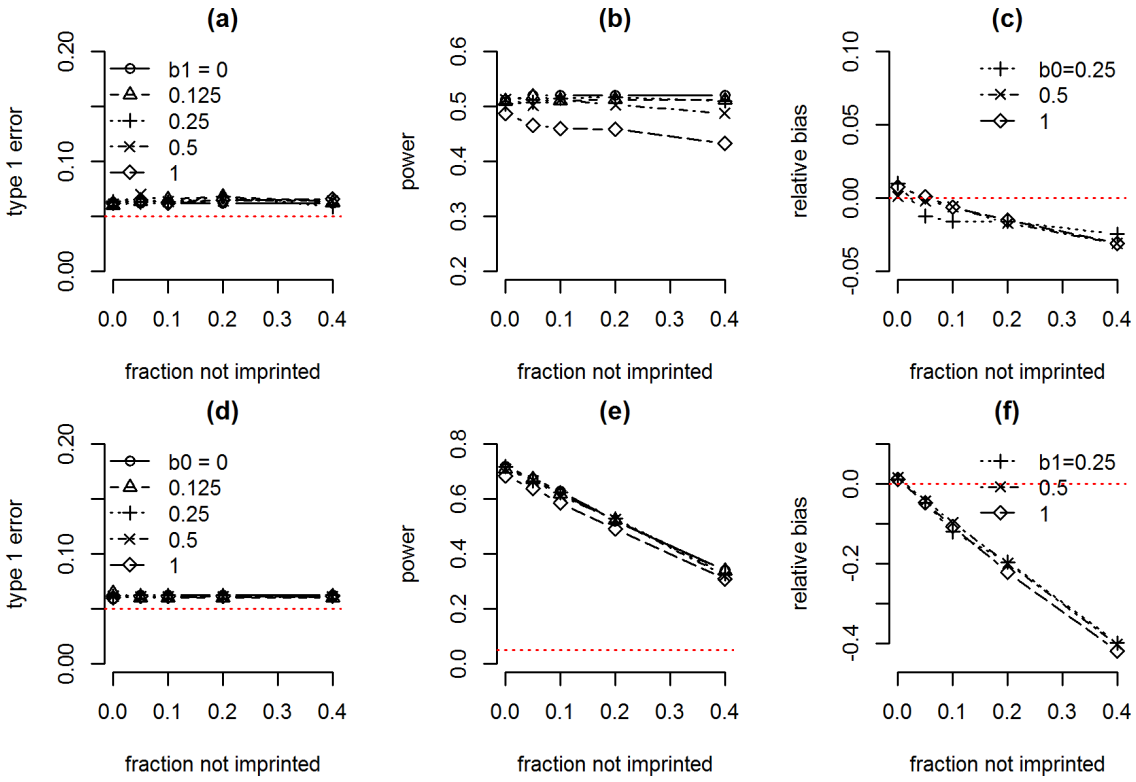

Figure 2.4: Simulation results for 10,000 datasets with variable effect sizes, and fraction imprinting. For readability purposes we choose sample size 32 for presentation. The data simulated with true model perturbed by replacing parent of origin effect (b_1) to 0 for a certain fraction of randomly chosen individuals. Fractions 0, 0.05, 0.10, 0.20 and 0.40 were selected for simulations. Model that assumes that all individuals share common parent of origin effect (b_1) was fitted. Panels (a)-(c) show the type 1 error, power and relative bias introduced to additive genetic effect when it is fit with underlying POO effect of 1 for different fractions of mis-specified parent of origin effects. Panels (d)-(f) show the similar results for parent of origin effect. In panels (b) and (d) we choose effect size of 0.5 since it better illustrates power drop. In panels (c) and (f) we plot a relative bias of a corresponding effect with other effect set to 0. Since it is a relative bias we want to concentrate only on large enough effect sizes to show the trend - we keep only effect sizes 0.25, 0.5 and 1. Even at that you may see that effect size 0.25 gets noisier estimate than 0.5 and 1.

2.2.2 Model mis-specification with perturbation of genotype

We also considered a situation when a fraction of individuals have randomly flipped genotype either due to error or due to the fact that we consider not true eQTL, but the one similar to it. To do it we simulate the model as described in method section, but for a fraction of randomly chosen individuals we replaced genotype by a randomly chosen wrong value. In simulation we included 0%, 5%, 10%, 20% and 40% flipped individuals. Again, we present the results for sample size 32.

We do observe that increased fraction of flipped genotypes leads to larger bias and lower power (Figure 2.5). We also observe that type one error is reasonably controlled in such scenario. We can see that higher fraction of individuals with perturbed genotype leads to higher loss of power. Additionally, it is clear that the higher is value of underlying additive effect the more pronounced is power drop in parent of origin effect due to the incorrect genotypes (panel (f)). Still, main drop due to genotype switch is in genetic effect (panel (e)).

2.2.3 Model mis-specification with perturbation of haplotype

Similarly we modify the main manuscript scheme adding a flip of parental information in 0%, 5%, 10%, 20% and 40% randomly chosen individuals. We performed simulations under several sample sizes (32, 64, 128) samples and present the results for sample size 32.

In this case we also observe that increased fraction of flipped haplotypes leads to larger bias and lower power (Figure 2.6). We also observe that type one error is not controlled as well in parent of origin effect when fraction goes up. The problem is more pronounced with higher underlying genetic effect. This effect is similarly visible in higher sample sizes. We can see that higher fraction of individuals with perturbed haplotype leads to higher loss of power. Additionally, the higher is value of the

incorrect haplotypes impacts parent of origin effect more with higher underlying additive genetic effect (panel (f))

2.2.4 Extended simulations with parameters selected based on real data

While we presented a more controlled simulation scenario in the main text that allows us to evaluate the power for a variety of effect sizes and to observe bias of the short model ignoring genetic or parent of origin effect, we also extended the simulation to better mimic observed dataset to simulate 10,000 genes. Specifically, we aimed to mimic the real dataset in several aspects:

1. We select 10 effect sizes for either genetic (eQTL) or parent of origin effect: 0 and 9 deciles of the observed effect sizes in the real data. For genetic effect these values span the range from 0 to 0.53 and for PoO the effect varies from 0 to 0.31. Among the 10,000 simulated genes, we randomly selected 1,000 genes to assign one of the ten genetic effects, and similarly randomly assign parent of origin effects.

2. To address variability in total read counts, we calculated mean total read count for each expressed gene from our real data analysis, excluded 5% genes with lowest expression and 5% genes with highest expression, and then selected 10 mean read counts using equally spaced quantiles. As result we get mean count for each gene to vary between 30 for the lowest expressed group to 3,588 for the most expressed genes. We randomly assign those read counts so that 1000 genes have one of the 10 read counts.

3. As shown in Figure 2.7, over-dispersion parameter varies with respect to expression level. In addition, over-dispersion distributions were fairly symmetric on log scale. Therefore for each of 10 total read count categories defined in step

2, we estimated mean and variance of both negative binomial and beta-binomial over-dispersion parameters (in log scale) from our dataset, and then randomly generated over-dispersion parameters for each simulated gene by sampling the log over-dispersion from normal distributions with given mean and variance.

4. We also observed that distribution of the fraction of allele-specific counts is notably different for each of 10 total expression categories and in each of them there is notable number of individuals with no allele-specific reads. We created empiric distributions of the fractions of allele-specific counts for each of 10 groups defined in step 2, and for each simulated gene we used the corresponding empiric distribution to simulate fractions of allele-specific counts for each individual.

5. Additionally, we took estimates of the other coefficients we used in our model fit: intercept, library size, and batch effects, and randomly simulated values for each of the simulated gene.

We have also repeated this simulation setup several times with additional a secondary eQTL. Our model only consider the primary eQTL, and thus this is a situation with model mis-specification. We consider the situation when the effect size of the second eQTL is a certain fraction of the effect size of the primary eQTL and its genotype is positively correlated with primary eQTL. The results for both correctly specified model (i.e., with only one eQTL) and several mis-specified models (i.e., with secondary eQTL) are shown in Figure 2.8. The type I error is controlled by both correctly specified model and specified models. Correctly specified model gives unbiased estimates of effect sizes. In the mis-specified models, since we simulated secondary eQTL to be positively correlated with the primary one, we tend to over-estimate eQTL effect and have higher power when the effect of the secondary

eQTL gets bigger. Mis-specification of eQTL effects does not affect our ability to estimate and test the parent of origin effect.

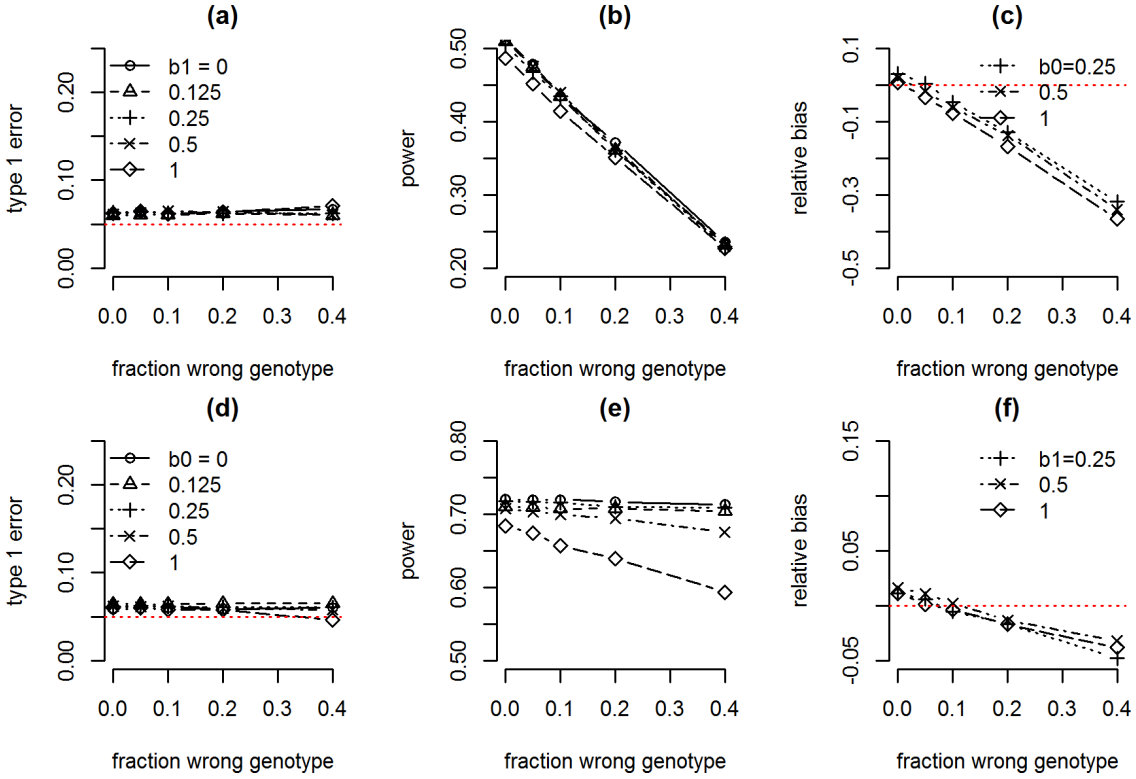

Figure 2.5: Simulation results for 10,000 datasets with variable effect size and fractions of individuals with wrong genotype. For illustration we select sample size 32. The data simulated with true model perturbed genotype replacing it by an opposite for a certain fraction of randomly chosen individuals. Fractions 0, 0.05, 0.10, 0.20 and 0.40 were selected for simulations. Panels (a)-(c) show the type 1 error, power and relative bias introduced to additive genetic effect when it is fit with underlying POO effect of 1 for different fractions of mis-specified parent of origin effects. Panels (d)-(f) show the similar results for parent of origin effect. In panels (b) and (d) we choose effect size of 0.5 since it better illustrates power drop. In panels (c) and (f) we plot a relative bias of a corresponding effect with other effect set to 0. Since it is a relative bias we want to concentrate only on large enough effect sizes to show the trend - we keep only effect sizes 0.25, 0.5 and 1.

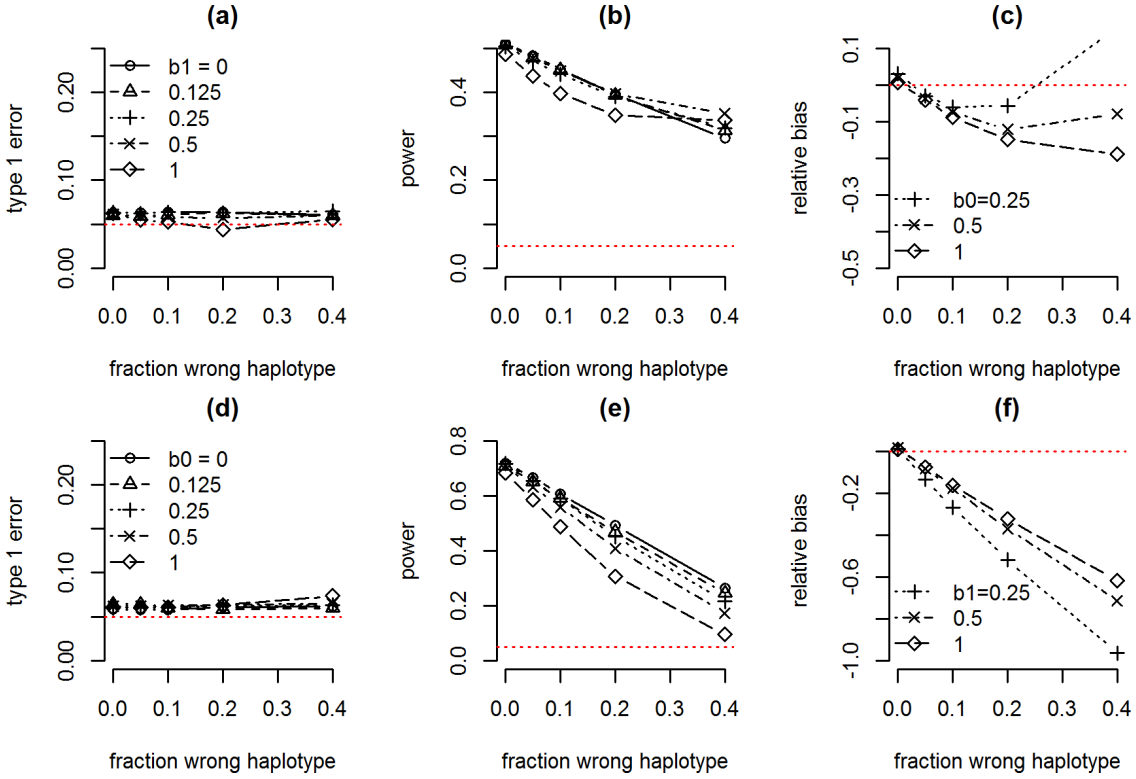

Figure 2.6: Simulation results for 10,000 datasets with variable effect size and fractions of individuals with wrong haplotype. For illustration we select sample size 32. The data simulated with true model perturbed haplotype replacing it by an opposite for a certain fraction of randomly chosen individuals. Fractions 0, 0.05, 0.10, 0.20 and 0.40 were selected for simulations. Panels (a)-(c) show the type 1 error, power and relative bias introduced to additive genetic effect when it is fit with underlying POO effect of 1 for different fractions of mis-specified parent of origin effects. Panels (d)-(f) show the similar results for parent of origin effect. In panels (b) and (d) we choose effect size of 0.5 since it better illustrates power drop. In panels (c) and (f) we plot a relative bias of a corresponding effect with other effect set to 0. Since it is a relative bias we want to concentrate only on large enough effect sizes to show the trend - we keep only effect sizes 0.25, 0.5 and 1.

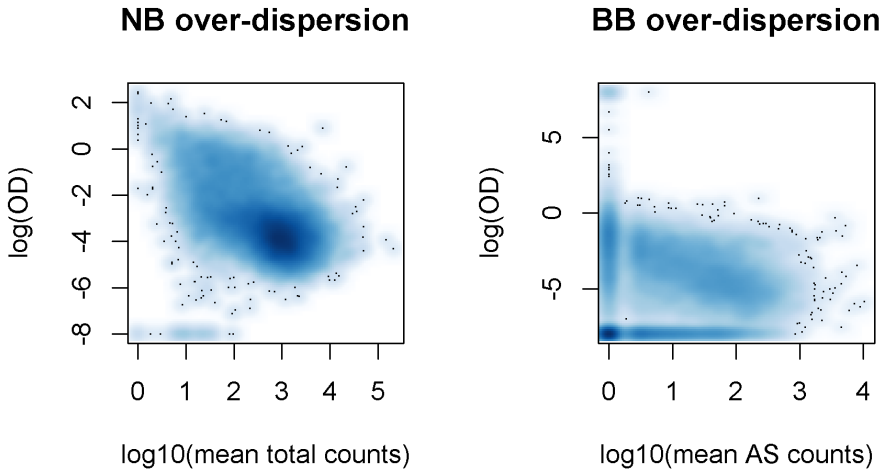

Figure 2.7: The distribution of over-dispersion parameters for total expression φ (left panel) and allele-specific ϕ reads (right panel) plotted versus median expression of a gene, both scales on logarithmic scale. Values for near zero over-dispersion in a group of genes are truncated below.

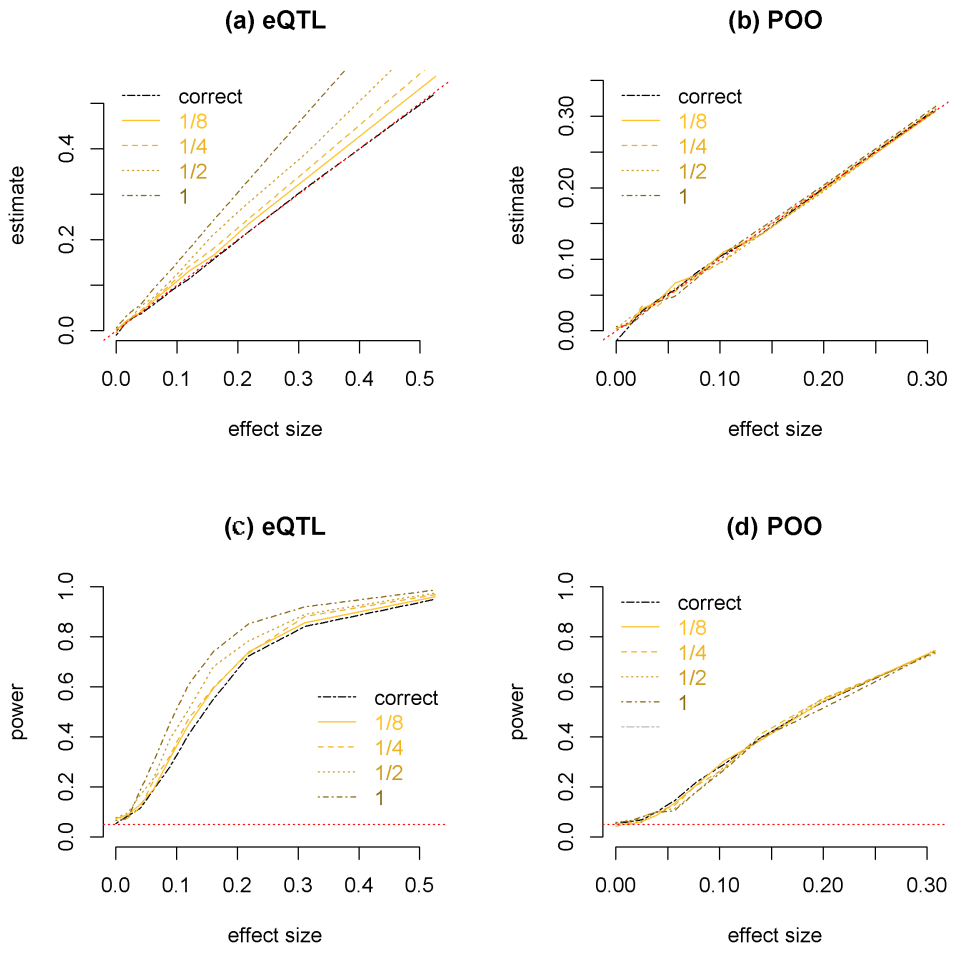

Figure 2.8: Simulation results for 10,000 genes with variable effect sizes, total read counts, and the proportion of allele-specific read counts. There are 5 simulation setups. The "correct" one has one eQTL. In the four mis-specified models, we only model one eQTL, but there is a secondary eQTL. The genotype of the primary and secondary eQTLs have a correlation of 0.5. The effect size of the secondary eQTL is a fraction (1/8, 1/4, 1/2, or 1) of the effect size of the primary eQTL. (a)-(b): Effect size estimation where x-axis is simulated effect size and y-axis is effect size estimate. the red dash line is the diagonal line. (c)-(d): The statistical power to detect genetic effect or parent-of-origin effects. The horizontal red dash line indicates the p-value cutoff 0.05 to declare statistical significance.

2.2.5 Comparison of model-based estimates of standard errors versus the empirically observed ones

Using the aforementioned genome-wide simulations, we demonstrated that the standard error of effect size estimates are consistent with empirically observed ones, with slight under-estimate when sample size is 32 (Table 2.2). The confidence intervals have reasonable coverage, except for sample size 32, where the coverage is slightly lower than expected (Table 2.3).

Table 2.2: Model-based standard errors vs empirical standard errors

eQTL						
sample size	**32**		**64**		**128**	
effect size b_0	**Model**	**Empiric**	**Model**	**Empiric**	**Model**	**Empiric**
0.000	0.120	0.138	0.086	0.097	0.074	0.100
0.019	0.118	0.131	0.086	0.090	0.071	0.076
0.040	0.133	0.150	0.101	0.102	0.070	0.072
0.062	0.123	0.115	0.089	0.093	0.067	0.058
0.089	0.148	0.156	0.105	0.113	0.073	0.067
0.120	0.124	0.121	0.111	0.104	0.077	0.081
0.160	0.119	0.116	0.089	0.090	0.061	0.061
0.218	0.132	0.145	0.102	0.127	0.071	0.062
0.312	0.118	0.137	0.089	0.086	0.067	0.063
0.526	0.123	0.127	0.101	0.112	0.074	0.071
Parent-of-Origin						
sample size	**32**		**64**		**128**	
effect size b_1	**Model**	**Empiric**	**Model**	**Empiric**	**Model**	**Empiric**
0.000	0.142	0.149	0.108	0.120	0.078	0.070
0.012	0.146	0.175	0.110	0.100	0.075	0.069
0.025	0.148	0.164	0.106	0.113	0.077	0.078
0.039	0.146	0.159	0.109	0.115	0.081	0.091
0.056	0.137	0.128	0.101	0.114	0.074	0.074
0.077	0.149	0.159	0.101	0.108	0.074	0.081
0.104	0.151	0.148	0.111	0.122	0.077	0.085
0.140	0.141	0.147	0.101	0.106	0.076	0.073
0.198	0.155	0.142	0.105	0.106	0.078	0.071
0.308	0.142	0.134	0.104	0.114	0.073	0.082

Table 2.3: Model based coverage

b_0	**32**	**64**	**128**	b_1	**32**	**64**	**128**
0.000	92.7%	94.2%	95.0%	0.000	92.4%	94.6%	96.1%
0.019	93.3%	94.4%	93.8%	0.012	93.8%	94.7%	95.7%
0.040	93.7%	95.3%	95.4%	0.025	93.9%	92.8%	94.8%
0.062	95.3%	94.0%	94.0%	0.039	94.1%	93.9%	93.5%
0.089	92.3%	94.6%	95.3%	0.056	93.2%	95.5%	94.5%
0.120	94.2%	93.1%	94.9%	0.077	94.4%	94.7%	95.1%
0.160	93.1%	94.9%	94.1%	0.104	93.9%	94.6%	93.7%
0.218	93.5%	94.3%	96.4%	0.140	93.7%	94.5%	95.5%
0.312	93.9%	94.5%	94.7%	0.198	95.0%	94.9%	95.0%
0.526	92.8%	94.7%	94.1%	0.308	94.1%	94.4%	94.3%
overall	93.5%	94.4%	94.8%	overall	93.9%	94.5%	94.8%

2.2.6 Timing

Model fitting for each gene includes the following steps: first fit TReC only model with b_0 to obtain reasonable initial values, and then fit full TReCASE model estimating b_0 and b_1, as well as two short models described above. The computational time were evaluated by average time per gene.

The computational time of our method scales well with the sample size. The time per gene remains under a second for sample sizes under 200 individuals: see Figure 2.9: the time needed to fit a gene is linearly dependent on sample size and increases for parameters being farther away from zero. For the sample size of around 200 the computation time is about 1 second per gene.

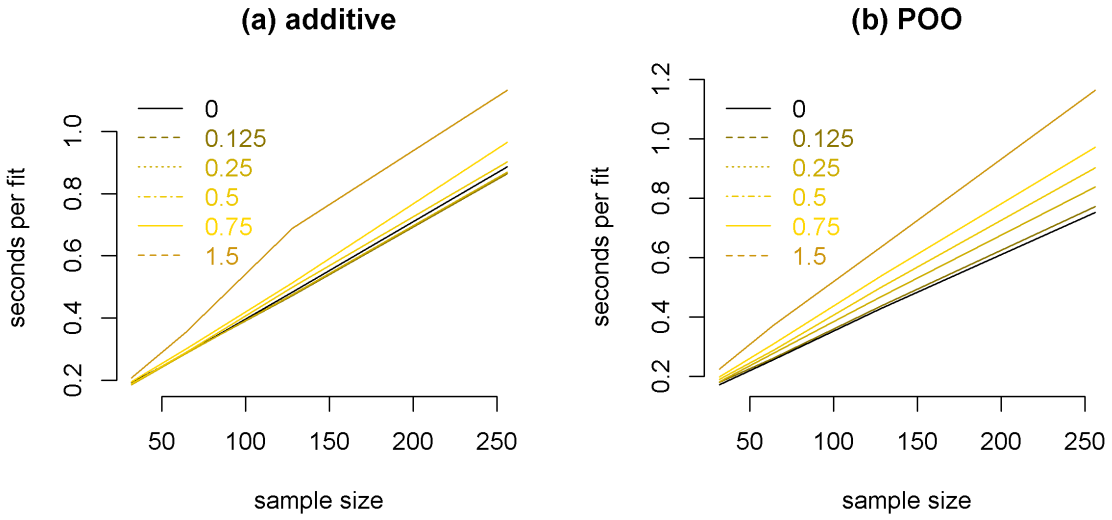

Figure 2.9: Time to fit full and short models in seconds to fit (a) additive genetic and (b) parent-of-origin effects

2.3 Application

2.3.1 Data collection

We collected RNA-seq data from 30 HapMap CEU (Caucasian) samples (15 males + 15 females). All of these samples are children of family trios, where the genotypes/haplotypes of three family members have been reported in the HapMap project, and their lymphoblastoid cell lines are available through Coriell (`http://www.coriell.org/`). For most of these samples, the RNA reads were 150 bp paired-end reads, with an additional smaller run of 75 bp paired-end reads. The median of the total number of reads for these 30 samples is about 20 million. These reads were mapped with Tophat2 using hg38 reference of human genome. HapMap project genotyped about 3.9 million SNPs for these 30 trios. We phased and imputed the genotypes of these 30 trios using shapeit2 (Delaneau et al. 2014) and impute2 (Howie et al. 2012), against 1000 Genome reference panel containing 2,504 individuals with ~82 million SNPs. Finally, based on phased and imputed genotype, we extracted allele-specific reads (i.e., those RNA-seq reads that overlap with at least one heterozygous SNP), and counted the number of allele-specific reads for each haplotype of a gene.

2.3.2 Identification of candidate *cis*-eQTLs

The parents of these 30 HapMap family trios are part of samples included in 1000 Genomes project (1000 Genomes Project Consortium 2012). To improve the power and precision for eQTL mapping, we first identified candidate *cis*-acting eQTLs using the Caucasian samples of 1000 Genome Project. Specifically, we downloaded fastq files for 227 European samples from the Geuvadis consortium (Lappalainen et al. 2013). We mapped all the reads to hg38 reference, and then performed similar process

of phasing and imputation of the genotypes of these 227 samples. Then we calculated total read counts and allele-specific read counts per gene and per sample, and then estimate additive genetic effect using TReCASE method (Hu et al. 2015).

2.3.3 Identification of imprinted genes

Given the candidate *cis*-acting eQTLs identified from 1000G samples, we used our method to estimate genetic (eQTL) effect and imprinting effect for 12,386 genes in the 30 children of the family trios. These genes were selected because they had enough expression in the 30 samples and they had at least one candidate *cis*-eQTL based on eQTL mapping results from the 227 samples of 1000 Genome Project. For the negative binomial model of total read count, we fit our model with additional covariates to capture the effects of read-depth and batches (the RNA-seq data were collected through 3 batches with 10 samples per batch). No additional covariate is needed for the analysis of allele-specific read counts because our model for ASE compared the expression of one allele versus the other allele, and thus the effects of such covariates are canceled. We found 16 genes with significant imprinting effects at q-value cutoff 0.05 (Table 2.4).

Replotting the results of the table 2.4 we can conclude that genes missing 0.05 cutoff generally show lower imprinting disbalance, so notable chunk of them missed 0.05 cutoff due to power issues of a small sample size.

Out of these 16 genes, 10 were known to be imprinted from previous studies and 6 were novel findings. For 14 of these 16 genes, the paternal allele had higher expression than the maternal allele. For all of those 10 known imprinted genes, our estimates of allelic imbalance agree with what were reported before. At a more liberal cutoff of q-value 0.25, we identified 15 additional imprinted genes. After manually examining the expression pattern of these 15 genes, we concluded that 12 of them missed the

Table 2.4: POO genes found: 1 - missed cutoff due to low count and 2 - missed cutoff due to smaller effect size

ID	Name	q-value	Expression	Is Known
ENSG00000269821	KCNQ1OT1	1.2e-08	paternally	yes
ENSG00000204186	ZDBF2	9.6e-08	paternally	yes
ENSG00000167981	ZNF597	2.9e-07	maternally	yes
ENSG00000185513	L3MBTL1	5.1e-06	paternally	yes
ENSG00000177432	NAP1L5	6.6e-06	paternally	yes
ENSG00000242265	PEG10	6.9e-06	paternally	yes
ENSG00000257151	PWAR6	5.8e-05	paternally	no
ENSG00000224078	SNHG14	8.2e-05	paternally	no
ENSG00000130844	ZNF331	1.2e-04	paternally	no
ENSG00000261069	RP11-701H24.4	3.8e-04	paternally	no
ENSG00000225806	RP1-309F20.3	3.8e-04	paternally	no
ENSG00000122390	NAA60	4.3e-04	maternally	yes
ENSG00000128739	SNRPN	3.6e-03	paternally	yes
ENSG00000100138	SNU13	4.4e-03	paternally	no
ENSG00000145945	FAM50B	2.5e-02	paternally	yes
ENSG00000101898	MCTS2P	3.7e-02	paternally	yes
*ENSG*00000279192[1]	PWAR5	8.6e-02	paternally	no
*ENSG*00000174851[2]	YIF1A	9.9e-02	paternally	no
*ENSG*00000182109[1]	RP11-69E11.4	1.3e-01	paternally	no
*ENSG*00000171847[1]	FAM90A1	1.5e-01	maternally	no
*ENSG*00000082781[1]	ITGB5	2e-01	paternally	no
*ENSG*00000178057[1]	NDUFAF3	2e-01	maternally	no
*ENSG*00000254319[1]	RP11-134O21.1	2e-01	paternally	no
*ENSG*00000253633[2]	KB-1980E6.3	2e-01	paternally	no
*ENSG*00000111678[1]	C12orf57	2e-01	maternally	no
*ENSG*00000101160	CTSZ	2e-01	maternally	no
*ENSG*00000054967[2]	RELT	2.1e-01	paternally	no
*ENSG*00000126226[2]	PCID2	2.1e-01	maternally	no
*ENSG*00000135709	KIAA0513	2.1e-01	maternally	no
*ENSG*00000175643	RMI2	2.1e-01	maternally	no
*ENSG*00000262155[1]	RP11-266L9.5	2.2e-01	paternally	no

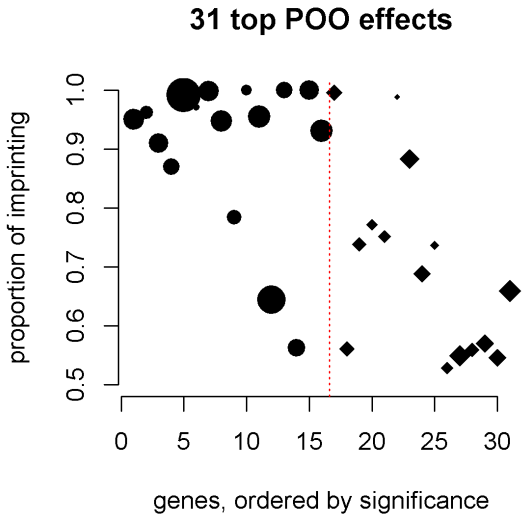

Figure 2.10: A graphical summary of the imprinting effect for the 31 genes that passed q-value 0.25 cutoff X-axis being gene index, and y-axis being the proportion of RNAseq reads from the allele with higher expression. These 31 genes are ordered by q-value and the size of each point reflects the scale of log read counts. Red line (as well as different shape) reflects a q-value 0.05 cutoff. The symbols indicate whether a gene has q-value smaller than 0.05.

q-value cutoff 0.05 due to power - either because parent-of-origin effect was smaller (4 genes) or the number of allele-specific reads was small (8 genes).

We illustrate the read count data of a clearly imprinted gene, ZNF497, which has higher expression on maternal allele (Figure 2.11). The imprinting effect can be observed from both TReC and ASE. We denote the genotype such as the first allele is maternal allele, i.e., genotype AB means A and B are from maternal and paternal allele, respectively. For TReC, the two groups with genotype AA and AB have similar expression because they share the same maternal allele and maternal allele has much higher expression than paternal allele. Similarly, the two groups BA and BB have similar expression. The results are even more evident in allele-specific reads where we observe the proportion reads from paternal allele is far below 50%.

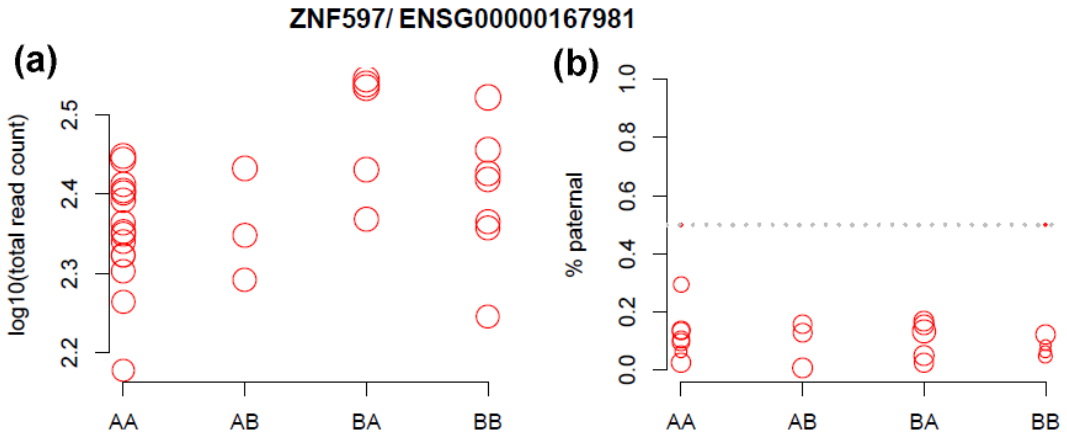

Figure 2.11: ZNF497 - a maternally expressed gene. (a) normalized total read counts ($\log_{10}$ scale), (b) percent of paternally expressed reads. Reads are classified into four categories by their genotype, assuming that first recorded genotype is maternal. Size of circle reflects the scale of log read counts.

We selected known imprinted genes based on the list reported by (Morison et al. 2005) or those genes recorded in the Geneimprint database (Jirtle 2016). There is a total of 90 known imprinted genes, among which 32 were expressed in our dataset and had a candidate eQTL. Of these 32 genes, 10 were found to be significant (q-value < 0.05) by our method. For several genes (such as RB1, KCNQ1, PEG3, and PLAGL1), we observed signal of imprinting, but the signal was too weak to produce significant q-value. In general, though, we observed that even for insignificant results, those with relatively smaller q-value tend to have estimated imprinting direction matched with the reported imprinting direction.

2.3.4 Locations of discovered parent-of-origin effect

For each chromosome, we assess whether the proportion of paternally imprinted genes of this chromosome is different from all the other chromosomes. We performed a test only if there are at least 5 imprinted genes in a chromosome (Table 2.5). These

Table 2.5: POO genes found by chromosome

	q-value < 0.25		q-value < 0.5	
chr	**pat/tot**	**p-value**	**pat/tot**	**p-value**
1	1/1		7/11	0.36
2	1/1		5/11	1.00
3	1/2		2/6	0.68
4	1/1		6/7	0.06
5	0/0		1/8	0.06
6	1/1		5/8	0.49
7	1/1		3/5	0.68
8	2/2		5/8	0.49
9	0/0		1/1	
10	0/0		2/6	0.68
11	3/3		5/10	1.00
12	0/2		1/8	0.06
13	0/1		0/4	
14	0/0		5/10	1.00
15	5/5	0.29	7/8	0.03
16	1/5	0.02	3/11	0.21
17	0/0		5/8	0.49
18	0/0		2/3	
19	1/1		5/11	1.00
20	3/4		5/8	0.49
21	0/0		0/0	
22	1/1		1/3	

formal tests confirm the clusters at chromosome 16 and 15, at imprinting q-value cutoff 0.25 and 0.5, respectively; although if we define imprinted genes at the q-value cutoff 0.5, the p-value of enrichment of imprinted genes for chromosome 15 is not significant after multiple testing correction.

2.3.5 Permutation setup for significant imprinted genes

For the genes with significant parent of origin effect we calculated permutation p-values using up to 100,000 permutations. We seek to permute the data to retain eQTL effect and destroy parent of origin effect. To do this we randomly flip maternal and paternal allele-specific counts. For samples with heterozygous eQTL SNP, e.g., with genotype $A_i^m A_j^p$, $i \neq j$, after flipping the allele-specific counts, we also flip the eQTL allele to be $A_j^m A_i^p$ so that the genetic effect remain unchanged. A naive comparison of the permutation p-value and the LRT p-value may suggest that they have weak correlation (Figure 2.12(a)). However, this is an artifact due to the limited range of permutation p-value. First, since we permute up to 100,000 times, the minimum permutation p-value is 10^{-5}. For some genes, the number of samples with allele-specific reads is small, and thus the total number of possible permutations is limited. For example, we can have at most $2^{10} = 1024$ permutations if there are allele-specific reads in 10 samples. After accounting for such limited range of permutation p-value, we observe strong consistency between LRT p-values and permutation p-values (Figure 2.12(b)).

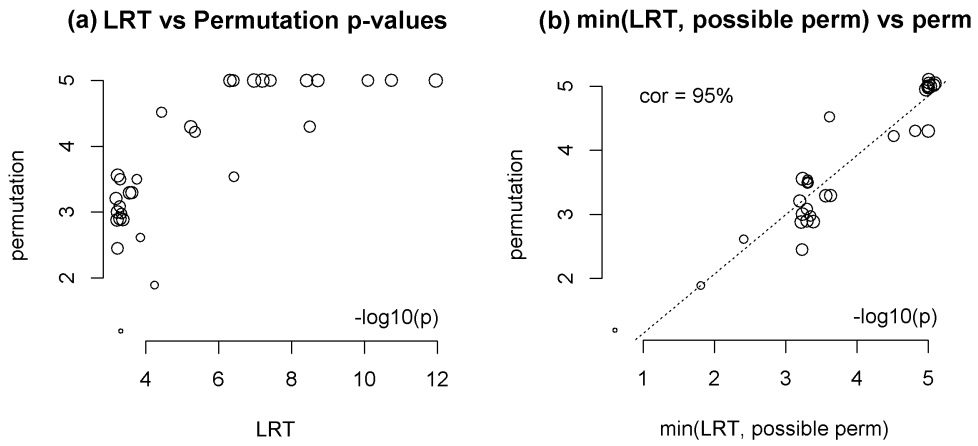

Figure 2.12: Calculating permutation based p-value for 31 significant genes using 100,000 permutations. Panel (a): permutation based p-value vs LRT based p-value, in -log10 scale. (b) Same permutation based p-value vs LRT or one over the number of possible permutations based on number of individuals with allele-specific counts for a given gene. The size of a circle is proportional to number of individuals with allele-specific counts. The dash line is the diagonal line.

2.3.6 Additional study of bias using simpler model and only eQTL and parent of origin effect

We also performed additional study of the bias observed in a Figure 2 of the main text to explore the underlying reason of such bias. We removed all the extra covariates and did a simple modeling of allele-specific counts only using beta-binomial distribution (R/vglm function) or quasi-binomial distribution (R/glm function). The latter one is not useful for testing as it doesn't control type I error under such mis-specification of over-dispersion (as it was shown in (Zou et al. 2014)), but is helpful as a point of comparison. We evaluate the relative bias of parent of origin effect (b_1) if we ignore genetic effect b_0, and illustrate the bias as b_0 increases from 0 to 1 (Figure 2.13 (a)). We can see that increasing size of b_0 is associated with increasing bias of beta-binomial over-dispersion parameter as well as increasing bias of b_1. Therefore, we conjecture that ignoring b_0 leads to over-estimate of over-dispersion parameter, which in turn affects b_1 estimates. In contrast, we didn't see such bias when used a quasi-binomial model (Figure 2.13 (b)), likely due to the fact that it first estimates effect size based on binomial model and then adds an extra step of estimating extra variation, which absorb the model mis-specification.

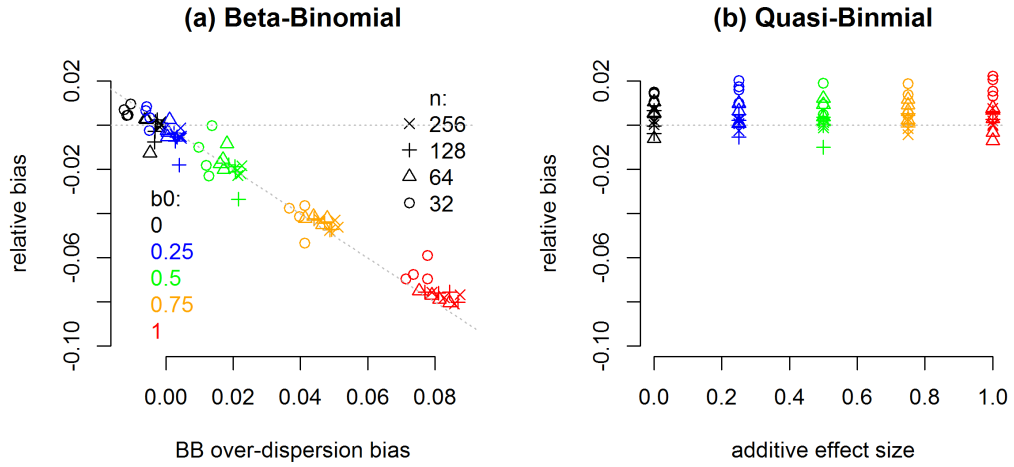

Figure 2.13: Plotting bias vs over-dispersion (a) Bias in effect size estimate (b_1, the parent of origin effect) of a mis-specified beta-binomial model (ignoring genetic effect b_0) is associated with bias in over-dispersion parameter. (b) Mis-specified quasi-binomial model does not lead to bias in effect size estimate (b_1, the parent of origin effect).

2.4 Summary

We developed a systematic approach to jointly estimate *cis*-eQTL and PoO effect in human. Our results recovered about one third of known imprinted genes, and if we excluded genes with low expression and included genes with weaker, but non-contradicting imprinting directions, there was a good overall consistency. None of the genes classified as "predicted imprinting" instead of "imprinting" in Geneimprint database were detected as imprinted genes with high confidence in our results. One possible reason is that most of these genes were selected based on a screening paper (Luedi et al. 2007) and were false positives. We also noted that for these "predicted imprinting" genes, the proportion of genes with predominantly paternal expression is roughly 50%, while this proportion is about 68% (61/90) for known imprinting genes. Several imprinted genes that we found are non-coding RNAs, such as RP11-701H24.4 and RP1-309F20.3, which warrants further studies to elucidate the functional consequence of imprinted non-coding RNAs.

In exonic regions with at least one heterozygous SNPs, one haplotype (e.g., the paternal one) may have more reference alleles than the other haplotype. This may create RNA-seq mapping bias because we map RNA-seq reads to reference genome. As shown by a comprehensive study, mapping bias has minimum effect on eQTL mapping (Panousis et al. 2014). In addition, we have explored the potential impact of mapping bias on estimating parent-of-origin effect and found it does not have any non-ignorable confounding effect.

We have assumed haplotypes around a gene are known or it can be inferred from un-phased genotypes. Phasing uncertainty is less a concern for family trio study because we performed phasing using family information and thus the phasing results are reasonably accurate within a short distance of a gene. For eQTL mapping in unrelated individuals, we employed the eQTL mapping method from (Hu et al. 2015),

which does handle phasing uncertainty.

Another potential issue is that due to genotyping error, some SNP genotypes may be incorrectly labeled as heterozygous. If such mislabel happens for a couple individuals, it can be accommodated by increasing over-dispersion parameter estimate for ASReC. If it happens for more individuals, the effect sizes of TReC and ASReC will be different we will not consider such genes. Such mislabel of genotype may also affect estimation of imprinting effect if it happens in a way consisting with parent-of-origin. However, it is unlikely that such genotyping errors occur frequently in many samples and happen to lead to the same direction of changes (e.g., creating zero-expression maternal allele).

To safeguard our results from such complications, we have used permutation test to evaluate our results. We observed good consistency between our model-based p-values and permutation p-values.

CHAPTER 3: PROTOCOL FOR EQTL MAPPING USING RNA-SEQ DATA AND EVALUATION OF DIFFERENT METHODS

In a diploid genome, most of the genomic loci have two alleles (i.e., the paternal and maternal allele), and thus gene expression can be quantified for each allele based on genetic difference of the two alleles. Most gene expression quantitative trait loci (eQTLs) are *cis*-acting eQTLs Doss et al. (2005), Lagarrigue et al. (2013), McKenzie et al. (2014), Crowley et al. (2015) that lead to allelic imbalance of gene expression, and thus using information from allele-specific expression (ASE) can improve the power of eQTL mapping. A few computational methods have been proposed for eQTL mapping using both total expression and ASE, including TReCASE (Total Read Count + ASE) Sun (2012), Hu et al. (2015), CHT (combined haplotype test) McVicker et al. (2013), and RASQUAL (Robust Allele Specific Quantitation and Quality Control) Kumasaka et al. (2016). TReCASE Sun (2012) was the first method of this kind and was later extended to implement a computationally more efficient score test and to account for phasing errors Hu et al. (2015). CHT was developed in a study with relatively small sample size of $n = 10$. It allows extra over-dispersion in total expression and account for genotyping errors. RASQUAL implemented some elegant strategies to account for sequencing/mapping errors, reference bias, genotyping errors, as well as phasing errors.

We will demonstrate that TReCASE is at least 10 times faster than RASQUAL. Earlier in the TReCASE paper Kumasaka et al. (2016), it has been demonstrated that CHT is at least 10 times slower than RASQUAL and its performance is not

better than TReCASE or RASQUAL. Therefore we will not consider CHT in our evaluations. We will evaluate TReCASE, RASQUAL, and a linear model that ignores ASE by simulations and real data analyzes using eQTL data from 1000 Genomes Project Consortium (2015), Lappalainen et al. (2013) and Genotype-Tissue Expression (GTEx) project Consortium et al. (2017), Aguet et al. (2019). Our analyzes aim to demonstrate the advantages and limitations of each method and provide guidance for their usage. We have also developed a pipeline to prepare ASE for eQTL mapping and a computationally efficient strategy for multiple testing correction. Our method development and results contribute towards a comprehensive computational framework for eQTL mapping using both total expression and ASE.

3.1 Introduction to the TReCASE and RASQUAL approaches with an illustrative example

In this paper we consider two methods that allow integration of allele-specific signal and total expression for association mapping of cis-QTLs: TReCASE and RASQUAL. In addition to this we use the comparison of TReCASE versus MatrixEQTL - a fast alternative that uses only total expression for analysis to both study power increase and precision of the signal detection. Each of these methods models the expression of each gene separately and assumes that genetic effects can be captured by a small number of *cis*-acting eQTLs, defining *cis*-eQTLs as those eQTLs that influence allelic-imbalance of gene expression.

To describe the differences in the way each method approaches the data consider an example provided in Figure 3.1. A *cis* acting SNP is illustrated to the left with T and C alleles for a given individual. The genetic effect is that the C allele has two times expression of T allele. First, one can collect total expression for this individual - 15, repeat this step for other individuals and perform an analysis on total expression

across multiple individuals estimating eQTL effect along with other covariates such as library depth, age, principal components, etc. Both RASQUAL and TReCASE fit this part of the data using Negative-Binomial distribution structure and are nearly identical in this part of the model.

MatrixEQTL, on the other hand, would fit a simple linear model fit on log transformed total expression with eQTL effect and other covariates.

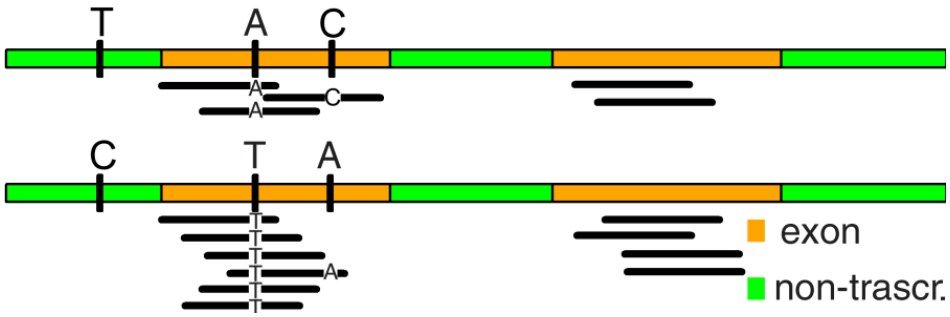

Figure 3.1: This diagram illustrates *cis*-eQTL and several SNPs at which we can collect allele-specific information. There is a *cis*-eQTL with C or T allele, and its C allele has two times of expression of T allele.

While total expression allows to do comparison of expression level between individuals it is clear from the Figure 3.1 that if we have phased SNPs, then for a fraction of reads overlapping with at least one heterozygous SNP we can recover within individual signal: In the figure we see that 6 reads identified with one haplotype and 3 from other haplotype. We will call such counts allele specific expression (ASE). This is an approach used by TReCASE: classifying all reads containing at least one SNP reads in a gene to one of two haplotypes, which further are aggregated on gene level. Reads with conflicting SNP information will be ignored.

RASQUAL also uses allele-specific reads, but collects them at each individual SNP separately. Returning to illustrative Figure 3.1 RASQUAL approach would produce allele-specific counts 6 and 2 for one SNP and for the second SNP they will be 1 and 1. Thus, according to RASQUAL approach, there can be more than one pair of ASE

values: for each SNP that overlaps some reads RASQUAL produces another pair of allele-specific counts. In RASQUAL notation those SNPs would be called fSNP: any within exon SNP such that at least for one of individual the SNP is heterozygous.

Further in analysis in both methods allele-specific reads are fitted assuming Beta-Binomial distribution structure.

The different assumptions lead to several consequences. Two obvious differences are the fact that (a) RASQUAL assumes that allele-specific counts share over-dispersion with total read counts while TReCASE fits Negative-Binomial and Beta-Binomial over-dispersion separately and (b) in RASQUAL some reads can be counted multiple times - once at each SNP. Less obvious, but as we will show critical, difference is that each set of ASE in RASQUAL is considered to be distributed Beta-Binomially - in our example both first SNP count 6 to 2 and second SNP count 1 to 1 are treated as Beta-Binomial counts, while in TReCASE aggregate 6 to 3 ASE count is Beta-Binomial. In nutshell TReCASE considers that Beta-Binomial over-dispersion occurs between samples and that within sample counts are distributed Binomially, while RASQUAL assumes that both within individual fSNP counts and between individual SNP counts come from the same Beta-Binomial distribution.

As an added benefit for RASQUAL choice of collecting the results at a SNP level there is an opportunity to correct some fSNPs. For example, if fSNP is assumed to be homozygous in a particular individual, but is observed to be heterozygous RAQUAL can correct this error. Detailed description of these models can be found in Method Section 3.3 Detailed study of these distinctions can can be found in Simulation Sections 3.4 and 3.5.

The main methodological difference in approach to the data between these two methods is that RASQUAL was designed to be applied to dataset without requiring data filtering while TReCASE relies more heavily on well preprocessed the data.

3.2 Data processing pipeline

Before going into details of data processing steps for processing genotype and gene expression data from each dataset, we first illustrate a general pipeline (Figure 3.2). The steps of obtaining TReC per gene and per sample (the left part of the figure) and phased and imputed genotypes (the right side of the figure) are often standard steps in all the gene expression quantitative trait locus (eQTL) analyses. The middle of the figure shows the steps of obtaining allele-specific reads at gene or SNP level, and it is an extra step to prepare data for eQTL mapping using allele-specific expression (ASE, or allele-specific read count, ASReC).

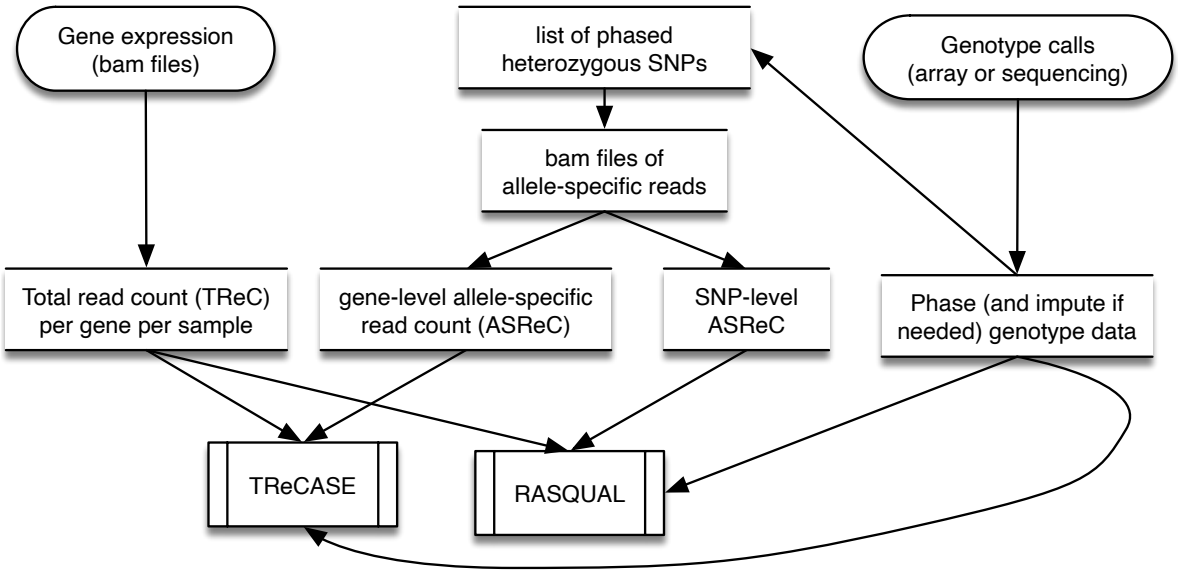

Figure 3.2: Data processing pipeline

3.2.1 The eQTL data from 1000 Genomes Project and Geuvadis Project

The 280 samples of this eQTL dataset are lymphoblastoid cell lines that are part of the samples used in 1000 Genomes Project (1KGP) (1000 Genomes Project Consortium 2015). The genotype data were obtained from SNP arrays and RNA-seq

data were generated by the Geuvadis project (Lappalainen et al. 2013).

Genotype phasing and imputation

SNP genotype data were obtained from the same Geuvadis project. Using SHAPEIT2 (Delaneau et al. 2014) and IMPUTE2 (Howie et al. 2011; 2009), we phase and impute genotypes in the following steps.

1. Convert unphased genotypes to PED/MAP format and ensure that genomic positions are converted to the genome reference that match the reference panel. This coordinate conversion can conveniently be done using liftover tool (Rhead et al. 2009).

2. Run SHAPEIT2 in check mode to get a list of mismatching SNPs. We have observed that there are a notable portion of SNPs labeled as strand mismatch after this step. To fix this problem we flip the SNPs with Plink (Purcell et al. 2007), repeat phasing in a check mode and compare resulting lists. If some of the SNPs present in first error list disappear in the second list they should be flipped and kept, and the rest of the mismatched SNPs can be supplied to SHAPEIT2 at the next step as an exclusion list.

3. Run SHAPEIT2 with exclusion list obtained in previous step.

4. Impute the pre-phased genotype data using IMPUTE2.

5. Using the imputed and phased data from the previous step, we obtain all heterozygous SNPs for each sample. Specifically, T=the main output file of imputation has 3 columns with genotype probabilities per SNP, which represent the probability of observing genotype $G = 0$, 1, or 2 respectively. We selected heterozygous locations (i.e., with high probability of $G = 1$) to output phased genotypes of these locations.

For both phasing and imputation, we used the 1000 Genome reference panel (Howie et al. 2011) (as of summer 2015) containing 2,504 samples with ~82 million SNPs. Effective size of the population was set to the suggested value `-effective-size 20000` and random seed was set to 1234567. Imputation was done by splitting the genome into blocks of 5 Mb (no more than 7 Mb according to the instruction). We also used the same population size option as the one used in phasing step (`-Ne 20,000`), other options used include `-align_by_maf_g` and `-seed 12345`.

The input data are genotypes of ~2.2 million SNPs per sample for 2,123 samples of African or European descent. After the above procedure of phasing and imputation, we end up with genotype data for ~82 million SNPs per sample which corresponds to ~2.2 million heterozygous SNPs per sample heterozygous SNPs per sample. Among all the imputed SNPs, around 6.5 million SNPs have MAF >0.05 and were used as candidate eQTLs for eQTL mapping.

Processing RNA-seq

We downloaded raw RNA-seq data (in fastq format) of 465 samples that are part of the samples for the 1KGP (1000 Genomes Project Consortium 2015). These RNA-seq data were generated by the Geuvadis consortium (Lappalainen et al. 2013), and it is available at
`http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/samples/`. We have used this dataset in a recent work (Zhabotynsky et al. 2019).

The RNA-samples were sequenced by the Illumina HiSeq2000 platform, with paired-end 75-bp reads. We mapped these samples to hg38 reference assembly using TopHat v2 (Trapnell et al. 2009), filtered with the following criteria: ≤ 3 mismatches, read gap length ≤ 3, read edit distance ≤ 3, and read realign edit distance equals to 0.

We filtered RNA-seq reads with average base quality ≥ 30, mapping quality ≥ 20,

and keeping only those uniquely mapped reads, by the `prepareBAM` function from R package `asSeq`. In a typical sample, most of the RNA-seq reads pass these filters.

For each sample, given the list of heterozygous SNPs obtained from phased and imputed genotype data, we filter out allele-specific reads using R function `extractASReads` from `asSeq` package (Sun 2012). This step generates 3 bam files for each sample labeled with *hap*1, *hap*2 and *hapN*, which contain the RNA-seq reads that match haplotype 1, haplotype 2, or with conflicting information. The size of *hapN* file should be much smaller than *hap*1 or *hap*2. Otherwise there may be some systematic mismatch between RNA-seq bam files and the SNP list, e.g., they were generated using different human genome references. Apparent imbalance in *hap*1 and *hap*2 file sizes often suggest some problems in the data preparation steps.

Next we obtained the Total Read Count (TReC) and allele-specific read count (ASReC) per gene using R function `summarizeOverlaps` from `GenomicAlignments` package (Lawrence et al. 2013) using Gencode version 21 (GRCh38). We observed that for some samples, many genes have extreme proportions of reads attributed to one haplotype - this is likely due inconsistency between genotype data and RNA-seq data. We discarded such samples and used the remaining 280 samples for eQTL mapping. The total number of reads mapped to genes in these 280 samples vary from 16 to 82 million with fraction of allele-specific reads from 2.6% to 6.3% (Figure 3.3).

For RASQUAL we produced SNP level allele-specific counts using the same imputed SNPs and mapped bam files as input to `ASEReadCounter` from GATK package (McKenna et al. 2010).

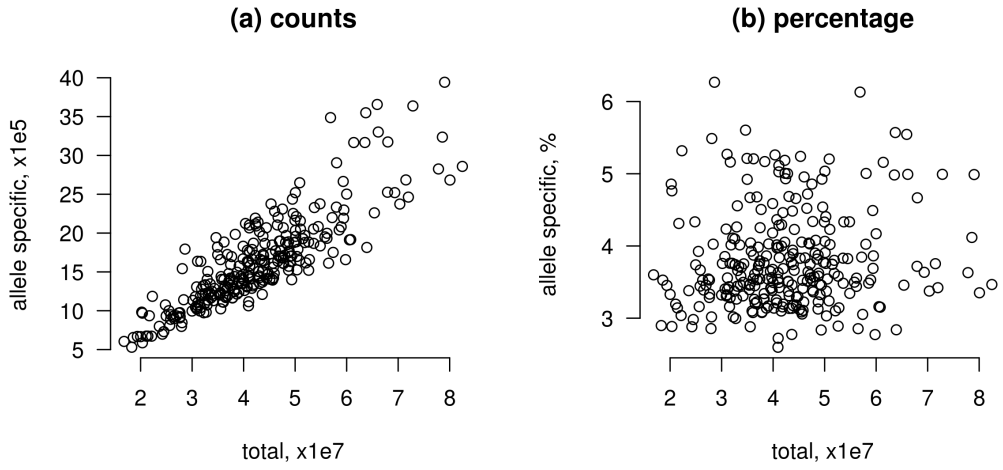

Figure 3.3: Summary of total read counts (x-axis) versus total number of allele-specific reads (panel (a)) or the percentage of reads being allele-specific (panel (b)) across all genes per sample for 1KGP data. Each point indicates one of the 280 samples.

3.2.2 GTEx data

We followed a similar pipeline to 1KPG to analyze Genotype-Tissue Expression (GTEx) data (Consortium et al. 2017). We downloaded mapped RNA-seq data (in SAM format) of 427 whole blood samples (V7). The RNA-samples were sequenced by the Illumina HiSeq2000 platform with paired-end 76-bp reads and were mapped to human genome references hg19. The RNA-seq data is available from dbGaP at `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2`.

We filtered RNA-seq reads with average base quality ≥ 20, mapping quality ≥ 20, and keeping only those uniquely mapped reads, by the `prepareBAM` function from R package `asSeq`.

We downloaded Genotype calls (in VCF format) of 635 samples (release V7, hg19) from dbGap. Indels and multi-allelic SNPs were removed using `bcftools` and variants with missing rate large than 10% were filtered out. We performed phasing on all 635 samples, with the same reference panel and command used when phasing 1KGP data and used the resulting output file to create a list of heterozygous SNPs for each individual. Then we applied the same approaches as for 1KPG data to collect Total Read Count (TReC) and Allele-Specific Read Count (ASReC) per gene and per sample for TReCASE and RASQUAL.

We filtered out genes whose 75 percentile of gene expression is less than 20 or maximum gene expression is more than 25 million. The resulting 16,675 genes were included in the analysis. The total number of reads mapped to genes in whole blood samples vary from 1 to 30 million with fraction of allele-specific reads from 2.3% to 11.73%, with one apparent outlier (Sample ID: YEC3), labeled in red, removed in the following analysis. (Figure 3.4).

Covariates data including genotyping principal components, age, and gender was

downloaded from `https://gtexportal.org/home/datasets`. 354 samples, having RNA-seq data from whole blood, genotype data and covariates data, were included in the following eQTL mapping.

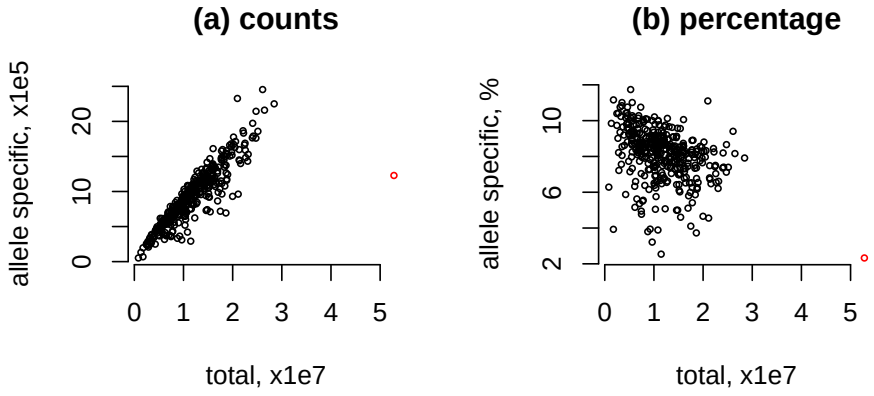

Figure 3.4: Summary of total read counts (x-axis) versus total number of allele-specific reads (panel (a)) or the percentage of reads being allele-specific (panel (b)) across all genes per sample for GTEx data. The red point indicates sample YEC3 has unexpected low proportion of allele-specific reads and it is excluded from further analysis.

More tissue types in GTEx data are available in the V8 release.

We plan to extend our analysis using multiple brain tissues from this release and provide the results for public use.

Table 3.1: GTEx tissues with hundreds of samples

Tissue	Sample size	
	RNA-seq & genotype	RNA-seq
Muscle - Skeletal	706	803
Whole Blood	670	755
Skin - Sun Exposed (Lower leg)	605	701
Artery - Tibial	584	663
Adipose - Subcutaneous	581	663
Thyroid	574	653
Nerve - Tibial	532	619
Skin - Not Sun Exposed (Suprapubic)	517	604
Lung	515	578
Esophagus - Mucosa	497	555
Cells - Cultured fibroblasts	483	504
Adipose - Visceral (Omentum)	469	541
Esophagus - Muscularis	465	515
Breast - Mammary Tissue	396	459
Artery - Aorta	387	432
Heart - Left Ventricle	386	432
Heart - Atrial Appendage	372	429
Colon - Transverse	368	406
Esophagus - Gastroesophageal Junction	330	375
Stomach	324	359
Testis	322	361
Colon - Sigmoid	318	373
Pancreas	305	328
Pituitary	237	283
Adrenal Gland	233	258
Spleen	227	241
Prostate	221	245
Artery - Coronary	213	240
Brain - Cerebellum	209	241
Liver	208	226
Brain - Cortex	205	255
Brain - Nucleus accumbens (basal ganglia)	202	246

3.3 Probability distributions used by TReCASE and RASQUAL

Because we perform eQTL analysis for each gene separately, in the following section we define the probability distribution for one gene and omit the index for gene.

3.3.1 TReCASE definition

Let T_i be the total read count (TReC) for a gene of interest in the i-th sample, with $i = 1, \ldots, n$. In TReCASE model, T_i is assumed to follow a negative binomial distribution with density function defined as:

$$f_{NB}(T_i = t_i; \mu_i, \phi) = \frac{\Gamma(t_i + 1/\phi)}{\Gamma(t_i + 1)\Gamma(1/\phi)} \left(\frac{1/\phi}{1/\phi + \mu_i}\right)^{1/\phi} \left(\frac{\mu_i}{1/\phi + \mu_i}\right)^{t_i}, \tag{3.1}$$

where μ_i is sample-specific mean value and ϕ is an over-dispersion parameter. It can be show that the variance of T_i is

$$\mathrm{Var}(T_i) = \mu_i + \mu_i^2 \phi, \tag{3.2}$$

and in the limiting case when $\phi \to 0$, the negative binomial distribution converges to a Poisson distribution.

Sample-specific mean value μ_i is defined as a function of some covariates:

$$\log(\mu_i) = \beta_0 + \beta_\kappa \log(\kappa_i) + \sum_{u=1}^{p} \beta_u c_{iu} + \eta_i, \tag{3.3}$$

where β_0 is an intercept, β_κ is the coefficient for log-read-depth $\log(\kappa_i)$, β_u, $u = 1, ..., p$, is the regression coefficient for the u-th covariate (e.g., age, gender, batch effects etc.), and η_i is the genetic effect. Given a candidate eQTL with two alleles A and B, let g_i be its genotype, defined as the number of B alleles such that $g_i = 0$, 1, or 2 for genotype of homozygous A allele, heterozygous, or homozygous B allele, respectively.

Then η_i is:

$$\eta_i = \begin{cases} 0 & \text{if } g_i = 0 \\ \log[1 + \exp(b_0)] & \text{if } g_i = 1 \\ b_0 & \text{if } g_i = 2 \end{cases}, \tag{3.4}$$

where b_0 is the genetic effect, defined as log ratio of gene expression for genotype BB vs. AA. The derivation of η_i follows from the assumption that gene expression is additive across the two alleles before log-transformation, see equations (3)-(7) of Sun (2012) for details.

For TReCASE, allele-specific expression (ASE), or allele-specific read counts (ASReCs) are collected for two haplotypes at gene-level, whereas the ASReCs for RASQUAL are measured at SNP level, as described in the next section. For the i-th sample, denote the two ASReCs for haplotypes 1 and 2 by N_{i1} and N_{i2}, so that total ASReC for the i-th sample is $N_i = N_{i1} + N_{i2}$. Note that for each sample, which haplotype is defined as haplotype 1 is arbitrarily decided. The distribution of N_{i1} given N_i can be modeled by a beta-binomial distribution:

$$f_{BB}(N_{i1} = n_{i1}; N_i = n_i, \alpha_i, \beta_i) = \binom{n_i}{n_{i1}} \frac{\Gamma(n_{i1} + \alpha_i)\Gamma(n_i - n_{i1} + \beta_i)}{\Gamma(n_i + \alpha_i + \beta_i)} \frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}, \tag{3.5}$$

where α_i and β_i are sample specific parameters and they are connected with expected proportion of reads of haplotype 1 (denoted by π_i) and over-dispersion (denoted by θ) of this beta-binomial distribution by

$$\pi_i = \frac{\alpha_i}{\alpha_i + \beta_i} \text{ and } \theta = \frac{1}{\alpha_i + \beta_i}. \tag{3.6}$$

The variance of this beta-binomial distribution is

$$\mathrm{Var}(N_{i1}) = n_i\pi_i(1-\pi_i)\frac{1+n_i\theta}{1+\theta},$$

which converges to the variance of binomial distribution when $\theta = 0$.

In an alternative parametrization, the over-dispersion parameter can be rescaled to interval [0,1): $\rho = \theta/(1+\theta)$ (Paul et al. 2005). We will switch to this parametrization in section 3.5.4 to study inflation introduced by splitting reads across several SNPs within a gene.

Let G_i be the ordered genotype of the candidate eQTL, which takes values 0, 1, 2, and 3 for ordered genotype AA, AB, BA, and BB. An ordered genotype is defined based on the order of haplotype 1 followed by haplotype 2. For example, AB indicates that A is on haplotype 1 and B is on haplotype 2. Given G_i, we can model π_i as a function of genetic effect b_0:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \begin{cases} -b_0 & \text{if } G_i = AB \\ b_0 & \text{if } G_i = BA \\ 0 & \text{if } G_i = AA \text{ or } BB \end{cases}. \tag{3.7}$$

The genetic effect b_0 for TReC (equation (3.4)) and ASE (equation (3.7)) are the same and thus it can be estimated by combining the data from TReC and ASE.

3.3.2 RASQUAL definition

To illustrate the difference between TReCASE and RASQUAL, we only consider the basic RASQUAL model without additional features such as capturing sequencing/mapping error or reference allele mapping bias. In addition, we change the notations used by RASQUAL to be consistent with the notations of TReC to facilitate

comparison. We will adopt two terms used in the RASQUAL paper: *cis*-regulatory SNP (**rSNP**) and feature SNP (**fSNP**). An rSNP is a candidate eQTL SNP and a fSNP is a SNP within exonic region of a gene where ASReC can be measured.

RASQUAL also models TReC T_i by a negative binomial distribution. Though instead of a common over-dispersion parameter for all samples, it has a sample-specific over-dispersion parameter ϕ_i:

$$f_{NB}(T_i = t_i; \mu_i, \phi_i) = \frac{\Gamma(t_i + 1/\phi_i)}{\Gamma(t_i + 1)\Gamma(1/\phi_i)} \left(\frac{1/\phi_i}{1/\phi_i + \mu_i}\right)^{1/\phi_i} \left(\frac{\mu_i}{1/\phi_i + \mu_i}\right)^{t_i}. \tag{3.8}$$

The mean and over-dispersion is parametrized by

$$\mu_i = \lambda K_i Q_i \text{ and } \phi_i = \frac{1}{\theta_R K_i Q_i}, \tag{3.9}$$

where λ is a scaling parameter for absolute mean of coverage depth of this gene, K_i is a sample specific offset reflecting library size and other a priori estimated size factors, Q_i is genetic effect defined later, and θ_R is a scaling parameter for over-dispersion. Then the variance of T_i is $\text{Var}(T_i) = \mu_i + \mu_i^2\phi_i = \mu_i(1 + \lambda/\theta_R)$.

Recall that g_i denotes the genotype of a candidate eQTL, and it equals to 0, 1, or 2 for genotype AA, AB/BA, or BB. RASQUAL quantifies genotype effect by

$$Q_i = \begin{cases} 2(1-\pi) & \text{if } g_i = 0, \\ 1 & \text{if } g_i = 1, \\ 2\pi & \text{if } g_i = 2. \end{cases}$$

Then the sample-specific mean value in log scale is

$$\log(\mu_i) = \log(\lambda K_i Q_i) = \log(\lambda) + \log(K_i) + \log(Q_i). \tag{3.10}$$

Comparing the mean value of TReCASE (equation (3.3)) versus RASQUAL (equation (3.10)), it is easy to see that $\log(K_i)$ captures the effects of all the covariates. Although it is not explicitly mentioned in the RASQUAL paper, we expect the scale of K_i to be around 1 because they assume λ captures the absolute mean value μ_i. We can also see the correspondence of genetic effect between TReCASE (b_0) and RASQUAL (π) is $\log[\pi/(1-\pi)] = b_0$.

While there is minor difference between TReCASE and RASQUAL in their TReC model, the major difference is their specifications for the ASE data. In RASQUAL, ASReCs are measured for each feature SNP (fSNP) (with SNP index $l = 1, \ldots, L$), denoted by N_{il0} and N_{il1} (with $N_{il} = N_{il0} + N_{il1}$). Then given N_{il}, N_{il1} is modeled by a beta-binomial distribution

$$f_{BB}(N_{il1} = n_{il1}; N_{il} = n_{il}, \alpha_{il}, \beta_{il}) = \binom{n_{il}}{n_{il1}} \frac{\Gamma(n_{il} - n_{il1} + \alpha_{il})\Gamma(n_{il1} + \beta_{il})}{\Gamma(n_{il} + \alpha_{il} + \beta_{il})} \frac{\Gamma(\alpha_{il} + \beta_{il})}{\Gamma(\alpha_{il})\Gamma(\beta_{il})}. \tag{3.11}$$

Let $0 < h \leq 1/L$ be the relative proportion ASReC contributed by each fSNP, then

$$\alpha_{il} = h\theta_R K_i Q_{il0} \text{ and } \beta_{il} = h\theta_R K_i Q_{il1},$$

where Q_{il0} and Q_{il1} quantify the number of ASReC from haplotype 0 and haplotype 1, and they are defined in Table 3.2. Note that $Q_{il} = Q_{il0} + Q_{il1} = Q_i$, which only depends on the genotype of the *cis*-regulatory SNP (rSNP, or candidate eQTL) but not the genotype of the l-th fSNP.

Then the expected proportion of reads of haplotype 1 and over-dispersion (denoted by ϑ) of this beta-binomial distribution are

$$\pi_i = \frac{\alpha_i}{\alpha_i + \beta_i} = \frac{Q_{il1}}{Q_i} \text{ and } \vartheta = \frac{1}{\alpha_i + \beta_i} = \frac{1}{h\theta_R K_i Q_i}. \tag{3.12}$$

Table 3.2: Relative mean for ASReCs. This table is taken from Supplementary Table 4 of Kumasaka et al. (2016). The first column defines the ordered genotype for rSNP (reference SNP or candidate eQTL) and fSNP (feature SNP where ASReC is measured) where 0 and 1 indicate reference and alternative allele, respectively.

rSNP,fSNP	Q_{il0}	Q_{il1}	$Q_{il} = Q_i$
00,00	$2(1-\pi)$	0	$2(1-\pi)$
00,01 or 00,10	$(1-\pi)$	$(1-\pi)$	$2(1-\pi)$
00,11	0	$2(1-\pi)$	$2(1-\pi)$
01,00 or 10,00	1	0	1
01,10 or 10,01	π	$(1-\pi)$	1
01,01 or 10,10	$(1-\pi)$	π	1
01,11 or 10,11	0	1	1
11,00	2π	0	2π
11,01 or 00,10	π	π	2π
11,11	0	2π	2π

The variance of this beta-binomial distribution is

$$\mathrm{Var}(N_{il1}) = n_{il1}\pi_i(1-\pi_i)\frac{1+n_i\vartheta}{1+\vartheta},$$

which converges to the variance of binomial distribution when $\vartheta = 0$.

In the RASQUAL paper, the authors further set $h = 1$ for the following reasons quoted from page 42 of their supplementary materials:

> "Here the constant h reflecting the proportion of total AS count at each feature SNP is arbitrary ($0 < h \leq 1/L$ for $L > 0$; otherwise $h = 0$). However, in our experience, $h < 1/L$ usually gives worse result in terms of power and fine-mapping than $h \approx 1$. This is partly because the larger the number of feature SNPs L is, the smaller the proportion h each feature SNP accounts for, resulting in overestimation of the dispersion parameter $\hat{\theta}$, resulting in more significant associations in hypothesis testing for features with larger L. To avoid this issue we set $h = 1$ to penalize the

over-dispersion parameter more for large L."

Since h denotes the proportion ASReC from each fSNP, it is very counter-intuitive to set it to be 1. It appears to be an ad-hoc solution to reduce the significance level, especially for the genes with large number of fSNPs. In fact, as shown in this paper, even when $h = 1$, RASQUAL still has inflated type I error, and the degree of inflation increases with the number of fSNPs (Figure 1B in main text). However, the reason is not overestimation of the dispersion parameter $\hat{\theta}$. The estimate of the dispersion parameter is often accurate. It is the mis-specified likelihood model that leads to underestimate of the variance of the eQTL effect size. We will explain in more details in Section D.8.

3.3.3 Definition of RASQUAL-like method: TReCASE-RL

To facilitate more pointed comparison between TReCASE and RASQUAL, in addition to running RASQUAL, we also implemented a modification of TReCASE to adopt two key assumptions made by RASQUAL but not by TReCASE: (1) Equating the over-dispersion parameters of the negative binomial distribution for TReC and the beta-binomial distribution for ASE; (2) Treating ASReC of each fSNP as independent beta-binomial observation. Specifically, for the i-th sample and the l-th SNP, we denote the two ASReCs for haplotypes 1 and 2 by N_{il1} and N_{il2}. The distribution of N_{il1} given $N_{il} = N_{il1} + N_{il2}$ is modeled by a beta-binomial distribution.

$$\begin{aligned} & f_{BB}(N_{il1} = n_{il1}; N_{il} = n_{il}, \alpha_{il}, \beta_{il}) \\ & \quad = \binom{n_{il}}{n_{il1}} \frac{\Gamma(n_{il1} + \alpha_{il})\Gamma(n_{il} - n_{il1} + \beta_{il})}{\Gamma(n_{il} + \alpha_{il} + \beta_{il})} \frac{\Gamma(\alpha_{il} + \beta_{il})}{\Gamma(\alpha_{il})\Gamma(\beta_{il})}, \end{aligned} \tag{3.13}$$

where α_{il} and β_{il} are SNP specific parameters connected with expected proportion of reads of haplotype 1 (denoted by π_i) and over-dispersion (denoted by θ) of this

beta-binomial distribution by

$$\pi_{il} = \frac{\alpha_{il}}{\alpha_{il} + \beta_{il}} \text{ and } \theta = \frac{1}{\alpha_{il} + \beta_{il}}. \tag{3.14}$$

We call the modified TReCASE model as TReCASE-RL, where RL stands for "RASQUAL Like". By comparing TReCASE and TReCASE-RL, we can illustrate the consequence of these two assumptions.

3.3.4 The over-dispersion parameters of the three models

We summarize the over-dispersion parameters of three models: TReCASE, RASQUAL, and TReCASE-RL, in Table 3.3. The over-dispersion parameters of TReCASE are constants across all samples. In contrast, RASQUAL's over-dispersion parameters vary across samples because they depend on K_i and Q_i. We expect both K_i and Q_i vary around the value of 1 and thus we may approximate the over-dispersion parameters of RASQUAL by assuming $K_i = Q_i = 1$. Then we can see the definition of over-dispersion parameters are in reverse scale. For TReCASE, larger over-dispersion parameter (ϕ or θ) means larger over-dispersion, and for RASQUAL, larger over-dispersion parameter (θ_R) means smaller over-dispersion. In our results, when we refer to an over-dispersion, it is the TReCASE over-dispersion.

Table 3.3: Over-dispersion parameters in TReCASE, RASQUAL, and TReCASE-RL models. RASQUAL model modifies the degree of over-dispersion variation with additional offset K_i and relative genetic effect Q_i.

Component	TReCASE	RASQUAL	TReCASE-RL
TReC	ϕ	$1/(\theta_R K_i Q_i) \approx 1/\theta_R$	θ
ASE	θ	$1/(\theta_R K_i Q_i) \approx 1/\theta_R$	θ

3.3.5 Evaluation of the binomial distribution assumption for ASReCs across multiple SNPs within one gene and one sample

TReCASE assumes the summation of ASReCs across multiple SNPs within a gene follows a beta-binomial distribution across samples. This assumption implies that the ASReCs across multiple SNPs within a gene and a sample should follow the same binomial distribution. If these SNP-level ASReCs actually follow a beta-binomial distribution, their summation will not follow a beta-binomial distribution, though it may be approximated by a beta-binomial distribution. Here we evaluate this binomial distribution assumption. It cannot be evaluated for most (gene, sample) pairs because most such cases only have a few heterozygous SNPs with enough coverage. We used RNA-seq data from 30 HapMap trios (NCBI BioProject access number: PRJNA385599) Zhabotynsky et al. (2019) to select a set of genes with allele-specific counts distributed across multiple SNPs. This dataset was used since it has higher read-depth. Specifically, we select those (gene, sample) pairs such that the gene in the sample has at least 6 heterozygous SNPs, with at least 5 overlapping reads per SNP. We ended up with 4,005 (gene, sample) pairs matching these criteria, accounting for less than 1% of all (gene, sample) pairs.

For these 4,005 cases we tested how often the binomial assumption is violated. Deviation from such assumption can be tested using a score statistic developed by Tarone (1979). Since we don't have many SNPs, normal approximation of score statistic cannot be applied. Instead we generated the null distribution of such score statistic using parametric bootstrap. For each gene, we estimated the proportion of reads from one allele, simulated ASReC for each SNP from a binomial

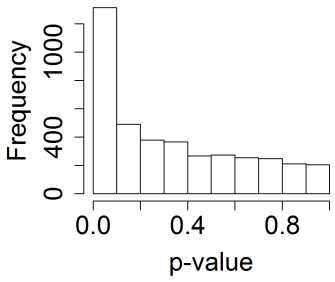

Figure 3.5: The distribution of p-values for testing deviation from binomial distribution across multiple SNPs of the same gene and within the same sample

distribution with this probability and the observed coverage, and recalculated the score statistic. The proportion of times when the observed score statistic is more extreme than the ones from simulations gives us an empiric p-value. Under null we expect a uniform distribution of p-values, however we see an excess of p-values in the category less than 0.1, which transforms to an estimate that the p-value is not uniform for ~23% of the cases (Figure 3.5). Note that this 23% is among those selected 1% of the cases with enough heterozygous SNPs covered by RNA-seq reads, and thus only represent a very small proportion of all the data points we examine in real data analysis.

3.4 Simulation setup

3.4.1 Simulation for TReC

We simulate total read count (TReC) using 4 covariates (including one that can be treated as library depth). Negative binomial over-dispersion parameter is set to be 0.01, 0.1 or 0.5. Genetic effect size b_0 is set to be 0, 0.125, 0.25, or 0.5. We simulate the TReC so that the median across samples is around 100.

3.4.2 Simulation for ASReCs without within-sample over-dispersion

In the basic simulation setup, we simulate ASReCs within a sample by a binomial distribution, and introduce over-dispersion across samples. Specifically, we simulated data by the following steps:

1 We assume 10% of TReC are allele-specific (Figure 3.6(a)) and set the same genetic effect b_0 as for TReC. The beta-binomial over-dispersion is defined as a fraction of negative binomial over-dispersion, with several values: 0, 1/8, 1/4, 1/2, 1 and 2 to cover the fractions observed in the 1KGP data (Figure 3.6(b)). Fraction 0 corresponds to the extreme, but not unlikely scenario when ASReC follow a binomial distribution.

2 After calculating expected proportion of reads coming from allele B for each genotype G, denoted by π_G we generate sample level expected proportions π_i by sampling from a beta distribution with parameters α and β such that $\pi_G = \alpha/(\alpha+\beta)$, and $\theta = 1/(\alpha+\beta)$.

3 Given π_i, we generate ASReC from a binomial distribution.

To simulate SNP level ASReC, we modify step 3 as follows:

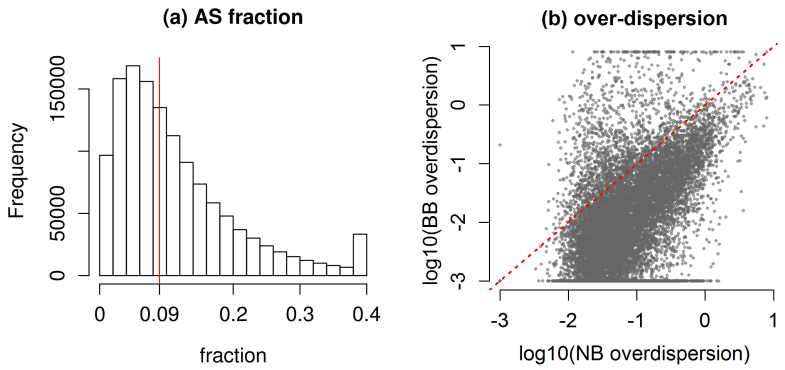

Figure 3.6: Distribution of allele-specific fractions and over-dispersion parameters. (a) The fraction of RNA-seq reads being allele-specific per gene per sample, given the ASReC is larger than 0, and truncated at 0.4. The vertical line indicates median. It is based on 1KGP dataset of 280 samples. (b) The distribution of over-dispersion parameters estimated by TReCASE from the 1KGP dataset of 280 samples. The BB over-dispersion is truncated at 0.001.

3a Uniformly distribute allele-specific reads among 2, 4, or 8 SNPs, and then simulate ASReC on one haplotype by a binomial distribution using the same sample level proportion π_i.

3.4.3 Add within-sample over-dispersion for ASReC data

In order to simulate SNP level ASReC with over-dispersion across multiple SNPs within a sample, we modify step 3a as follows:

3b Uniformly distribute total number of allele-specific reads (that can belong to either haplotype) among 2, 4 or 8 SNPs, and then simulate ASReC for each SNP on one haplotype by a beta-binomial distribution with mean value π_i and an over-dispersion parameter that equals to the between-sample beta-binomial over-dispersion.

This modification allows us to obtain ASReCs with over-dispersion within a sample and at the same time with more similarity within sample than between samples.

3.4.4 To simulate RASQUAL style ASReC data

RASQUAL style SNP level ASReC data have the same over-dispersion across any two SNPs, either two SNPs within a sample or between samples. In other words, there is no extra over-dispersion across samples. To simulate such data, we modify step 2.2 to set the between-sample over-dispersion to be 0. After that we generate all SNP level ASReC from a beta-binomial distribution according to step 2.3b. This is equivalent to generating SNP level proportions π_{il} from sample level proportions π_i with the same over-dispersion θ without discrimination for the SNPs within and between samples.

3.4.5 Simulating the data with genotyping errors

To simulate genotyping errors, we randomly flip a fraction of genotypes from homozygous to heterozygous and from heterozygous to homozygous (fractions 0.05, 0.10 and 0.20 were used). The wrong genotypes have the following consequences.

1. TReCASE only uses heterozygous SNPs to collect ASReCs. In contrast, RASQUAL produces counts for homozygous SNPs too. Consequently whenever we flip a heterozygous SNP to be homozygous, the read counts of this SNP are ignored by TReCASE, while still used by RASQUAL - it would have read counts of both alleles counted, but attach them to an incorrect homozygous genotype status.

2. When a truly homozygous SNP is flipped to be heterozygous, both methods still use such data. They both assume all the reads are from one of the two alleles.

3.5 Simulation Results

3.5.1 Simulation results under RASQUAL assumption

Under RASQUAL assumption, the SNP-level ASReC follows a beta-binomial distribution, such that the similarity of the SNP-level ASReCs is the same for two SNPs within one sample versus two SNPs of two different samples. This is the simulation scenario described in Section C.4. It worth noting that such scenario is not supported by real data. As shown in Figure 3.5, when considering a subset of (gene, sample) pairs with at least 6 heterozygous SNPs and at least 5 overlapping reads per SNP, in 77% of the cases the ASReCs within a sample follow a binomial distribution. We mainly want to use this scenario to demonstrate that RASQUAL does control type I error if its model assumption is correct. In addition, TReCASE model is mis-specified in this scenario, and we demonstrate that TReCASE still has reasonable performance despite its model mis-specification.

As shown in Figure 3.7, both TReCASE and TReCASE-RL control type I error well except that TReCASE-RL is slightly conservative when negative binomial over-dispersion is 2 and beta-binomial over-dispersion is 0.25. The power of the two methods is similar. TReCASE has slightly higher power when the over-dispersion for beta-binomial distribution is small, and TReCASE-RL has slightly higher power when the over-dispersion for beta-binomial is large.

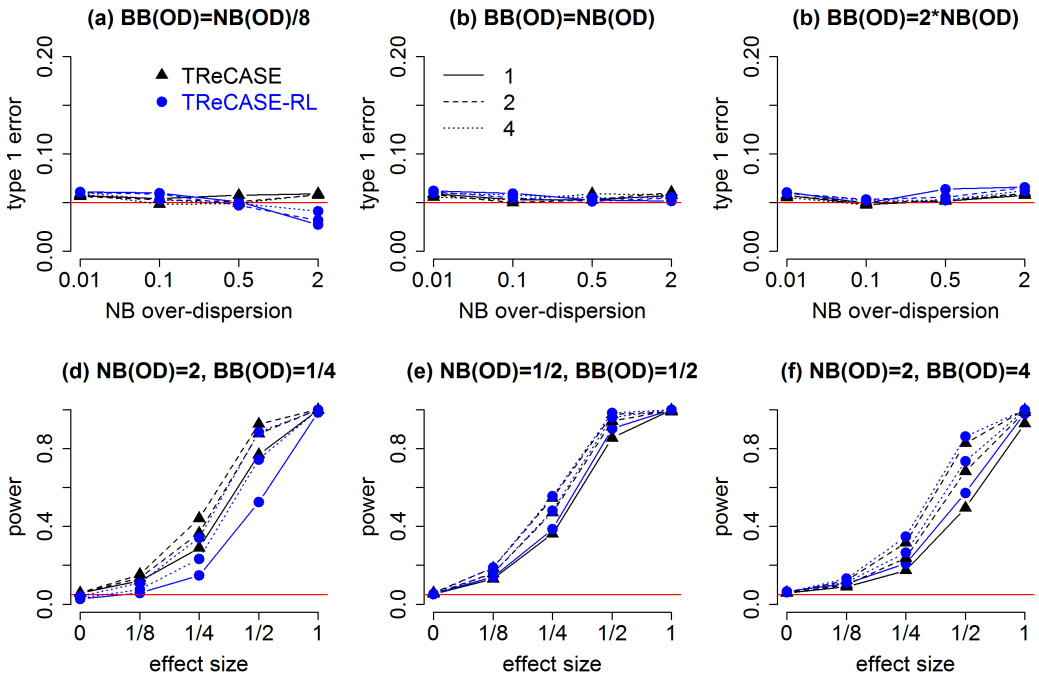

Figure 3.7: Evaluating TReCASE and TReCASE-RL using simulated data under RASQUAL assumption. Under RASQUAL assumption, within sample over-dispersion is the same as between sample over-dispersion. Panels (a)-(c) present type I error and panels (d)-(f) present power. 10,000 genes were simulated for each effect size and over-dispersion profile. The three line types refer to the number of fSNPs per gene. We use sample size 64 in our illustrations.

3.5.2 Simulations given within sample over-dispersion and additional between sample over-dispersion for ASReCs

This is the simulation setting where both models are mis-specified. Following the simulation described in Section C.3, we simulated ASReC data with within sample over-dispersion and extra between-sample over-dispersion. When simulating two SNPs per gene, TReCASE still manages to control type I error, while the RASQUAL style approach has inflated type I error and the inflation increases for larger beta-binomial over-dispersion (Figure 1(C)). In the case of 4 or 8 SNPs per gene, the results are similar. TReCASE still controls type I error, and RASQUAL style approach shows even higher inflation of type I error (results not shown).

3.5.3 Evaluating models for the data simulated under TReCASE assumption

In previous section we show that the majority of the genes do not show evidence of within sample over-dispersion. This is a TReCASE style assumption. In this section, we simulated the data without within sample over-dispersion, and either combined them into one count or split them across two or four SNPs. Our simulation results show that in this scenario, TReCASE controls type I error reasonably well (Figure 3.8 (a)-(c)), while RASQUAL can either produce deflated type I error (Figure 3.8 (a)) or inflated type I error (Figure 3.8 (b)-(c)).

We also evaluated the effect of double counting by randomly selecting 10% of the reads from each SNP and adding them to a neighboring SNP. Double-counting further inflates type I errors in all three simulation settings (Figure 3.8 (d)-(f)).

We observed that RASQUAL is has some additional reasons to produces inflation, not explained by equating total read counts and allele-specific counts over-dispersion and treating each SNP count as independent Beta-Binomial as implemented in

TReCASE-RL version. This is was another motivation to do some comparisons of TReCASE vs TReCASE-RL - this allowed us to more precisely measure assumption violations such as non-equal over-dispersion in total and allele-specific counts and within-gene Binomial distribution.

RASQUAL was still run for reference (as can be seeing in panels h-j), which allows us to show both observed RASQUAL inflation and inflation that we get in RASQUAL-like TReCASE model. In this way we are able to evaluate both inflation observed by RASQUAL and inflation simply due to over-dispersion misspecification, avoiding potential extra discrepancies due to differences in implementation. For the case when we evaluated genotype correction we used RASUQAL itself.

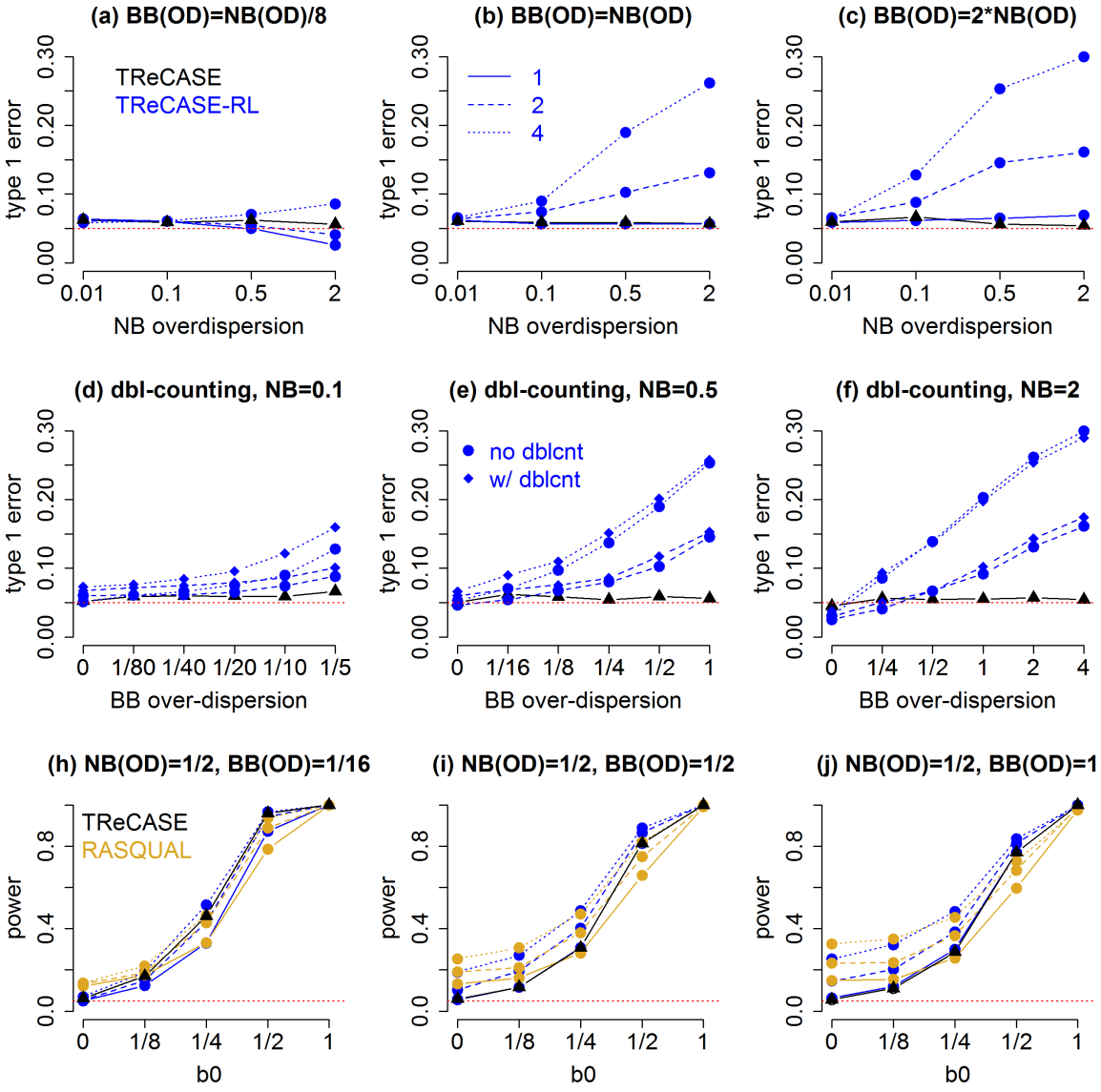

Figure 3.8: Evaluating TReCASE, TReCASE-RL, and RASQUAL models for the data simulated using TReCASE style assumptions. (a)-(c) Evaluation of Type I error across different values of over-dispersion parameters. (d)-(f) Evaluation of type I error given double counting. (h)-(j) Evaluation of power and type I error across eQTL effect sizes. Results are presented for sample size 64.

3.5.4 Evaluation of the inflation of type I error by RASQUAL

In the previous section, we showed applying RASQUAL or TReCASE-RL on simulated RNA-seq data (using both TReC and ASReC), there is inflated type I error. In this section, we study this issue by ignoring TReC and concentrating on ASReCs without within-sample over-dispersion. We set beta-binomial over-dispersion to 0.1 or 0.5 and generate data under null hypothesis of no eQTL effect (proportion of either allele is set to be 0.5). We simulated data for 1,000 genes. For each gene, we first simulated the data assuming there is only one fSNP. Then we split the ASReCs uniformly to multiple fSNPs, while the total number of allele-specific reads is the same. For example, if there are k reads per SNP for 8 SNPs, then there are $2k$ reads per SNP for 4 SNPs, etc. We applied TReCASE-RL on these data. As shown in previous results, there is no inflation of type I error in one SNP scenario. In two SNP scenario, as expected, both likelihood ratio test-statistics (LRT) and -log(p-value) become larger than one SNP scenario (Figure 3.9).

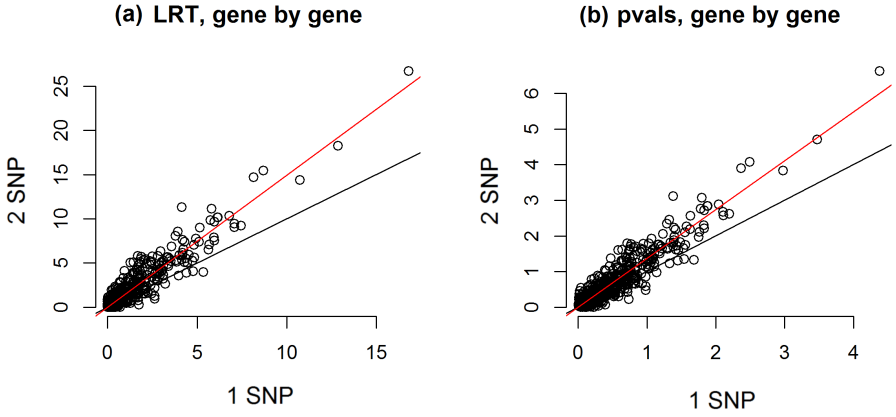

Figure 3.9: Illustration of the type I error inflation by TReCASE-RL. In simulations under H_0: b_0=0, compare likelihood ratio test-statistics (LRT) and -log10(p-values) of two situations: one fSNP or two fSNPs. Each point corresponds to one of 1000 genes. We observe that for the same gene once reads are split into two SNPs we tend to get more significant results.

To understand the underlying causes of type I error inflation, we examine the estimation of eQTL effect and over-dispersion across simulate replicates. We found that after splitting the reads to 2, 4 and 8 fSNPs, the estimation of eQTL effects and over-dispersion are both unbiased (Figure 3.10). As the number of fSNPs increases, the variation of eQTL effect estimates remains similar while the variation of over-dispersion estimates increases. However, due to the mis-specification of likelihood model by RASQUAL, the mode-based standard deviation (sd, sd:mod in Figure 3.10) estimates are smaller than observed sds (sd:obs in Figure 3.10) for both eQTL effects and over-dispersion. The under-estimation of sd for eQTL effects explains the inflation of type I error when we test for eQTL effect. Note that model-based sds does match well with empirical sds when there is only one fSNP. This is expected since there is no model mis-specification when there is only one fSNP.

We derive sd using the Fisher's information matrix derived by Paul et al. (2005), and to be consistent with their work, we quantify over-dispersion by $\rho = \theta/(1+\theta)$, where θ was the over-dispersion parameter defined in Equation (3.1). To make the notation clear, we use legend OD(theta) or OD(rho) in the plots to indicate over-dispersion quantified by θ and ρ, respectively.

Next we examine how the (model-based) sd estimates varies with respect to the number of allele-specific reads. As expected the sd estimates of either eQTL effect and over-dispersion decreases as ASReC increases (Figure 3.11). For eQTL effect, we have observed in Figure 3.10 that the true sds are similar when the number of fSNPs is 1, 2, 4, or 8. Since the sd estimate of 1 SNP scenario is unbiased, the difference of the sd estimates of eQTL effects when the number of fSNPs is 1, 2, 4, or 8 (Figure 3.11) reflects under-estimates of sds due to model mis-specification. Such difference become clearer when we examine relative sd estimates compared with the sd estimate of 1 SNP scenario (Figure 3.11). The standard deviation of over-dispersion parameter is of

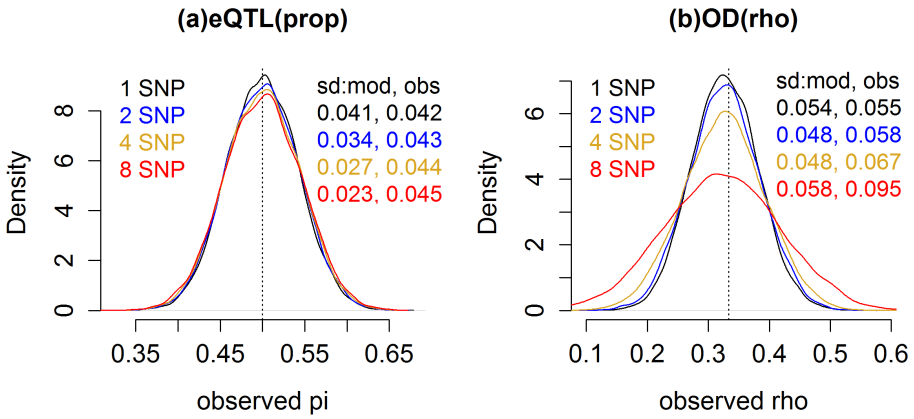

Figure 3.10: Distribution of eQTL effect and over-dispersion estimates. (a) Distribution of eQTL effect estimates in terms of π, the proportion of ASReC from one haplotype. (b) Distribution of over-dispersion estimates in terms of ρ, which is the rescaling of over-dispersion parameter θ to [0, 1) range by $\rho = \theta/(1+\theta)$. In the upper-right corner of each figure, we also list the model-based standard deviation estimate (sd:mod) using Fisher's information matrix (Paul et al. 2005) and empirical standard deviation estimate (sd:obs) across simulation replicates. The data were simulated under null of no genetic effect, and ASReCs were split into 1, 2, 4, and 8 SNPs. Simulation is done for sample size 64 and on average 10 allele-specific reads per sample

less interest in this study. Though we can see that when the number of ASReC is relatively large, the sds of over-dispersion tend to be under-estimated.

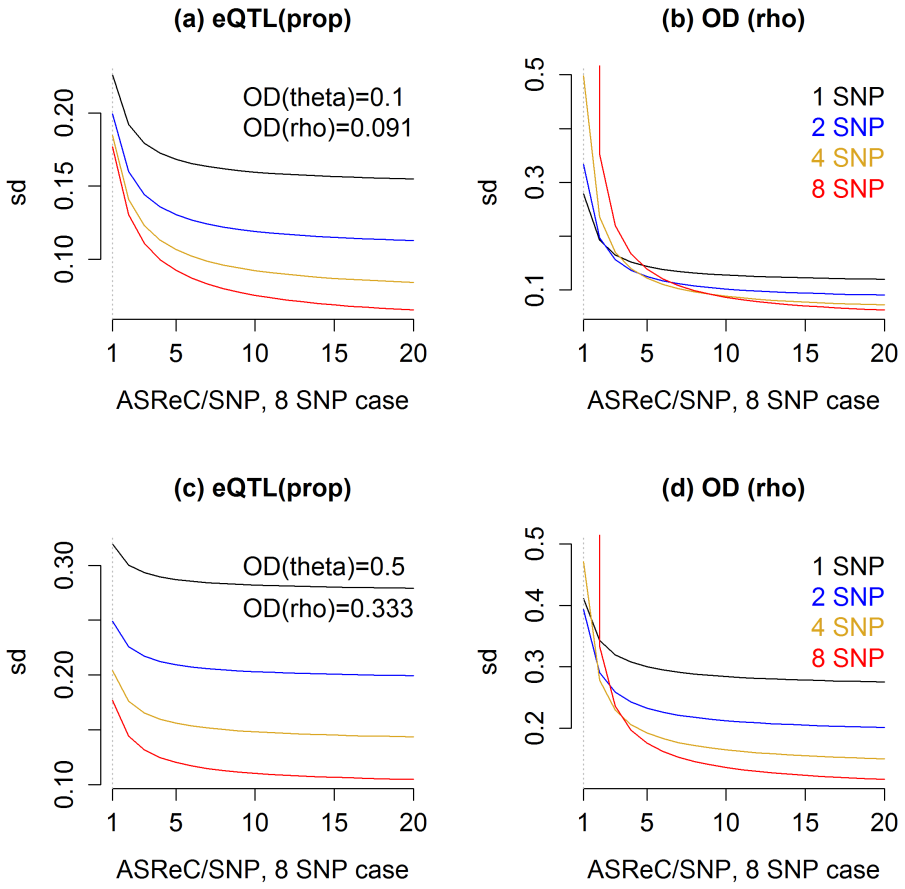

Figure 3.11: Model-based sd estimates by Fisher's information matrix under null. The sd estimates are evaluated for 1, 2, 4, and 8 fSNPs per gene. X-axis is the number of allele-specific reads for each of 8 fSNPs. For example, $x = 5$ means there are 5 reads for each of the 8 SNPs, or 10 reads for each of the 4 SNPs, or 20 reads for each of the 2 SNPs, or 40 reads for one SNP. Panels (a)-(b) present the simulation results for $\phi = 0.1$ and panels (c)-(d) present the simulation results for $\phi = 0.5$

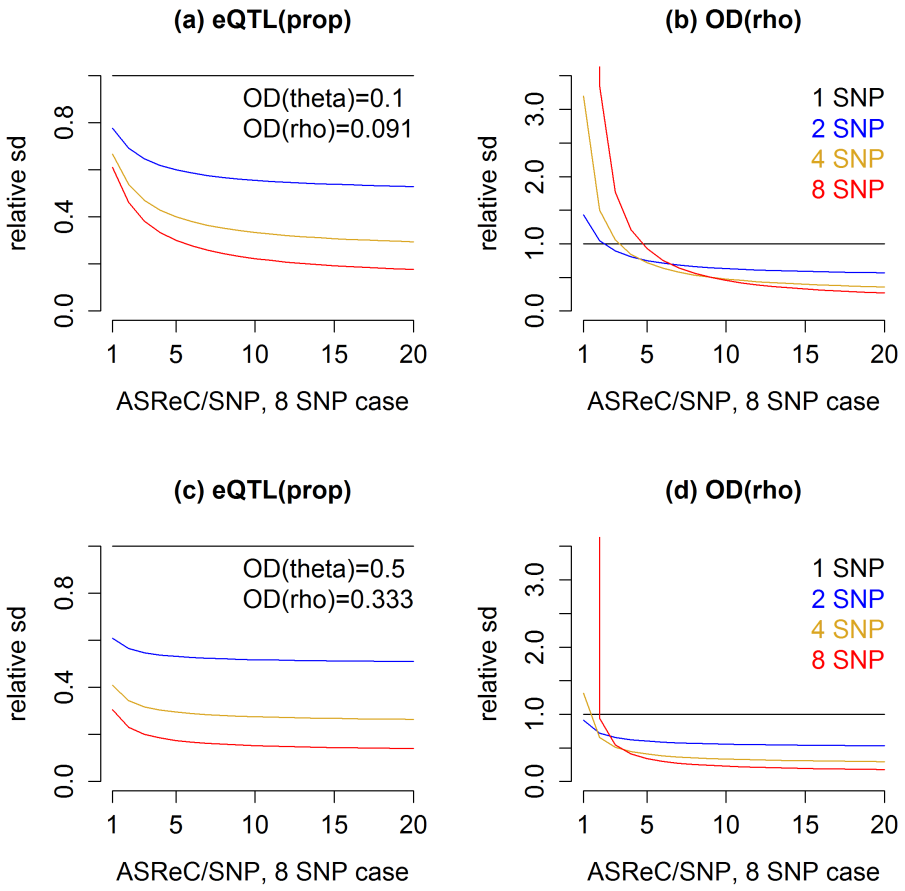

Figure 3.12: Model-based relative sd estimates under null. The same as Figure 3.11, except that the y-axis is the relative sd estimates with respect to the sd estimate for 1 fSNP scenario.

3.5.5 Compare TReCASE and RASQUAL's performance with genotyping errors

Finally, we evaluate the results of TReCASE and RASQUAL when there are genotyping errors using simulated data. We simulated ASReC data without within sample over-dispersion (Section C.2), and then introduced genotyping errors as described in section C.5, where we flipped certain fraction of genotypes from homozygous to heterozygous and from heterozygous to homozygous. When truly heterozygous SNPs are listed as homozygous, they will be discarded by TReCASE but used by RASQUAL, and the latter has a mechanism of correcting the genotype status if it encounters a conflicting SNP. If truly homozygous SNPs are listed as heterozygous, TReCASE will take them at face value while RASQUAL again will try to correct them. Because RASQUAL needs to use multiple SNPs to correct a wrong one, we only consider a scenario when splitting counts to 8 SNPs. Since most genes have less than 8 heterozygous SNPs, these simulation results represent an uncommon situation that favor the genotyping error correction mechanism of RASQUAL.

We consider the simulation setting when the over-dispersion parameters of negative binomial and beta-binomial are the same to match with the assumption made by RASQUAL. We illustrate the type I errors and powers for this simulation setup for a few values of over-dispersion parameters and several sample sizes (Figure 3.13). TReCASE controls type I error in all simulation setups. In contrast, RASQUAL controls type I error only if sample size is small and over-dispersion is small (Figure 3.13 (d)-(f)). The power of TReCASE remains similar as the proportion of genotyping errors increases, with some slight reduction of power when over-dispersion is large and eQTL effect is large. Larger fraction of genotyping errors also reduces the power of RASQUAL in a magnitude slightly larger than for TReCASE (Figure 3.13 (a)-(c)).

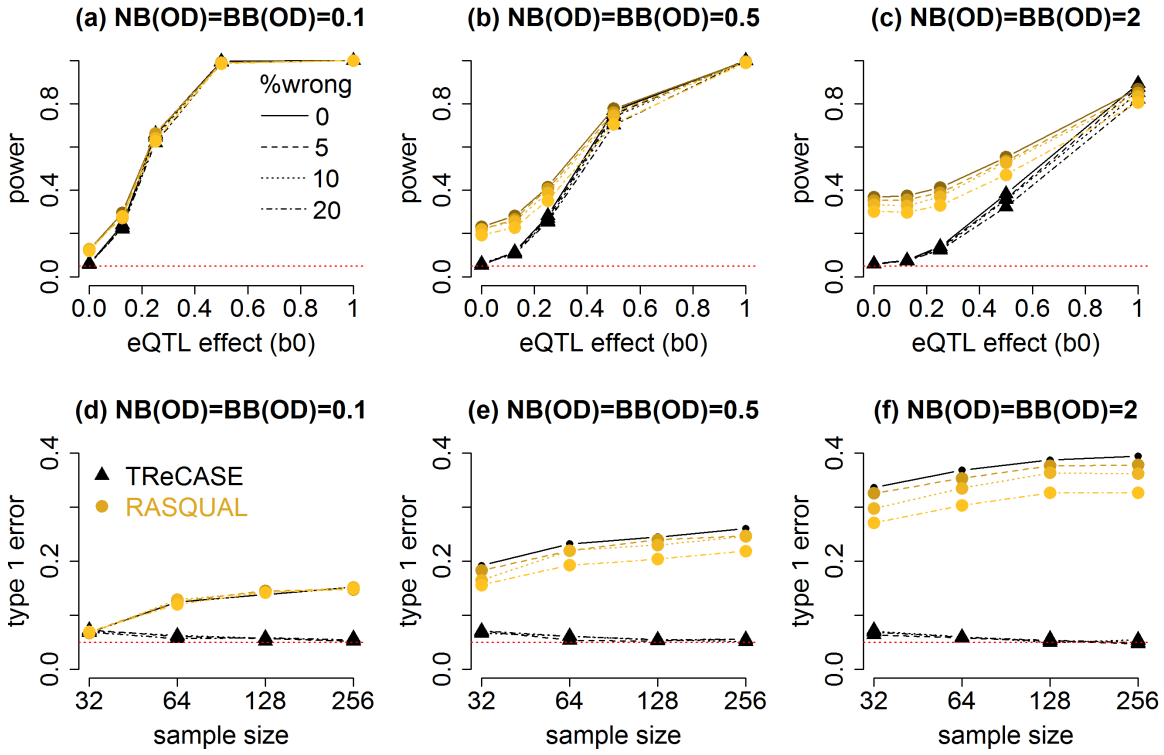

Figure 3.13: Type I errors and powers for a fitting TReCASE and RASQUAL with 0, 5%, 10%, or 20% of all the fSNPs being randomly flipped between homozygous and heterozygous. Sample size 64 was used.

3.6 Summary of observed method performance

Both TReCASE and RASQUAL assume TReC follows a negative binomial distribution that extends Poisson distribution to allow over-dispersion as described in Materials Section 3.3. The negative binomial distribution is widely used to model TReC per gene across multiple samples. We found it is adequate for most of them, though sometime outliers of gene expression data may lead to inflation of type I error. We address this issue by adopting an outlier detection and reduction approach implemented by DESeq2.

Despite using the same distribution family, the model assumption for ASE is very different between RASQUAL and TReCASE. RASQUAL models ASReC for each feature SNP separately. In contrast, TReCASE add up the ASReCs across all the heterozygous SNPs within a gene and model the gene-level ASReCs across samples. An alternative interpretation of the TReCASE model is that it assumes ASReC follow a binomial distribution across multiple SNPs within a sample, and then their summation follows the same binomial distribution. This assumption is reasonable for the majority of the cases (Figure 3.5). In other words, TReCASE assumes the ASReCs across multiple SNPs within a sample are more similar than ASReCs across different samples. In contrast, RSQUAL assumes a constant over-dispersion within or between samples. For example, suppose there are 5 heterozygous SNPs within a gene and there are 100 samples. TReCASE models the gene-level ASReC across the 100 samples by a beta-binomial distribution, while RASQUAL models the 5×100 SNP-level ASReCs by a beta-binomial distribution. In practice, there is often much larger variation across samples than within a sample (Figure 3.5 and Figure 3.6b), due to biological difference across samples. Then RASQUAL's model is mis-specified and it can lead to inflated type I error, which we will discuss in details.

Type I error inflation from RASQUAL's results

We ran both TReCASE and RASQUAL for eQTL mapping using 1KGP data, both before and after permuting SNP genotypes across samples. The same permutation is applied to all the SNPs so that the correlations among the SNPs remain the same and all the eQTL signals are removed. We examined the relation between the percentage of significant findings versus the number of feature SNPs (fSNPs). Using un-permuted data, TReCASE has higher power than RASQUAL for the genes with less than 10 fSNPs and their power become more similar for genes with larger number of fSNPs (Figure 3.14(A)). Using permuted data, all findings should be false positives. We examined the proportion of findings with p-value smaller than 0.05 with respect to the number of fSNPs (Figure 3.14(B)). It is very clear that TReCASE controls type I error well. However, RASQUAL's type I error increases linearly with the number of fSNPs and it is severely inflated for the genes with large number of fSNPs.

Because the major difference between TReCASE and RASQUAL is their models for ASE, we conjecture that the type I error of RASQUAL is due to its assumption of the same beta-binomial distribution within and between samples. Indeed, when this assumption is satisfied, RASQUAL controls type I error (Figure 3.7). When there is within sample and extra between sample over-dispersion, it violates TReCASE's assumption of no over-dispersion within a sample, and also violates RASQUAL's assumption of no additional over-dispersion across samples. In this situation, TReCASE still controls type I error but TReCASE-RL has inflated type I error (Figure 3.14(C)). When there is no over-dispersion within sample, but some over-dispersion across samples, TReCASE-RL still has inflated type I error (Figure 3.14(C)) and the degree of inflation increases with the number of fSNPs.

RASQUAL's approach to collect of the ASReC at the SNP level lead to double-counting certain reads and we seek to quantify the consequence of such double

counting. Finally, comparing all three methods in terms power analysis, RASQUAL has slightly larger type I error inflation than TReCASE-RL, and lower power especially for larger eQTL effect sizes.

3.7 Using MatrixEQTL to perform preliminary screening

MatrixEQTL (Shabalin 2012) is a software package for computationally efficient eQTL mapping using linear model. We first perform normal-quantile transformation for the expression of each gene before running MatrixEQTL, so that a linear model is reasonable. This transformation is a rank-based transformation and thus it can limit the effect of any outliers. We plan to use MatrixEQTL in two ways: for preliminary screening of SNPs before fitting TReCASE model and to estimate the effective number of tests for each gene.

We first compare the number of eQTL findings across p-value cutoffs for all gene-SNP pairs by TReC, TReCASE and MatrixEQTL (Table 3.4 and Table 3.5).

Table 3.4: Gene-SNP p-values by TReC vs MatrixEQTL. Note, we used widely used way to spot influential counts by using a known approach of marking values with Cook's distance bigger than 4/n, where n is sample size. Such value is recommended, for example, in (Hardin et al. 2007). We consider several other candidate cutoffs in the further Section 3.9.6 and confirm that 4/n is more appropriate for our analysis.

TReC / **MatrixEQTL**	(0,1e-6]	(1e-6, 0.001]	(0.001,0.01]	(0.01,0.1]	(0.1,1]
(0,1e-6]	162525	18962	650	115	7
(1e-6,0.001]	28543	359959	77947	7140	898
(0.001,0.01]	619	107229	403969	183961	5440
(0.01,0.1]	278	11369	227985	1562396	530555
(0.1,1]	200	1854	8794	601335	10670619

To screen potential eQTLs, we can first run MatrixEQTL and then select those gene-SNP pairs passing a liberal p-value cutoff, such as 0.01, and only run TReCASE for those selected gene-SNP pairs. Suppose after considering multiple testing

correction, p-value cutoff 10^{-6} is a preliminary cutoff for TReC p-value and we apply MatrixEQTL p-value cutoff 0.1 for screening. We will miss those 200 eQTLs with p-value > 0.1 by MatrixEQTL but p-value < 10^{-6} by TReC. We found it is actually justifiable to "miss" those eQTL findings because they are likely due to some outliers. For example, after refitting TReC model with 6 most expressed samples removed for each gene, 74.3% of these tests weren't significant at 0.01 level and only 10.5% of these cases still had TReC p-value smaller than 10^{-6}.

Table 3.5: Gene-SNP p-values by TReCASE vs MatrixEQTL

TReCASE / **MatrixEQTL**	(0,1e-6]	(1e-6,0.001]	(0.001,0.01]	(0.01,0.1]	(0.1,1]
(0,1e-6]	139821	8808	498	84	8
(1e-6,0.001]	98179	232608	45088	6550	767
(0.001,0.01]	31865	152742	231474	128474	12888
(0.01,0.1]	24655	124162	273335	880042	535208
(0.1,1]	12661	81197	167098	1000798	7834345

3.8 Estimation of permutation p-values

3.8.1 The method for estimation of permutation p-values

When performing eQTL mapping for each gene, we need to scan across a number of SNPs around the gene. The genotypes of these SNPs are often correlated due to linkage disequilibrium. To correct for multiple testing across these local SNPs, we can estimate the permutation p-value of the most significant association. It is computationally infeasible to run TReCASE or RASQUAL on a larger number of permuted datasets. Instead, we seek to estimate a relation between permutation p-value and minimum p-value for each gene separately, while using linear regression for eQTL mapping. This is closely related with the concept of "effective number of tests" since the ratio between permutation p-value and minimum p-value can be

considered as the "effective number of tests". Our model shows that such "effective number of tests" is not a constant for each gene. It varies with the scale of the minimum p-value.

Let $p_{min,i}$ and $p_{perm,i}$ be the minimum p-value for the i-th gene and the corresponding permutation p-value, respectively. Sun et al. (2010) observed that there is an approximate linear relation on log scale:

$$E\left[\log_{10}\left(p_{perm,i}\right)\right] = \beta_0 + \beta_1 \log_{10}\left(p_{min,i}\right). \tag{3.15}$$

We found such a linear model is accurate when the permutation p-value is small. However, when there are relatively larger permutation p-values, e.g., 0.1, a logistic regression has a better fit:

$$logit\left[E\left(p_{perm,i}\right)\right] = \beta_0 + \beta_1 \log_{10}\left(p_{min,i}\right). \tag{3.16}$$

We use the following procedure to produce multiple pairs of minimum p-value and permutation p-value per gene to estimate β_0 and β_1 in the logistic regression.

1. For each gene we create k new datasets using bootstrap with eQTL effect size modified to produce minimum p-value corresponding to permutation p-value in the range from 0.001 to 0.25. In order to approximately achieve a target permutation p-value α, we modify the eQTL effect size so that the minimum p-value is α/E, where E is a preliminary estimate of the effective number of tests by eigenMT tool (Davis et al. 2016). The default value of k is 100. Then the eQTL effect sizes of these 100 datasets are 100 grid points evenly spaced on log scale. We also consider $k = 20$, 50, and 200 in our evaluations and conclude that $k = 100$ is a good balance between accuracy and computational efficiency.

2. Run first 100 permutations. If more than 40% of the points on the grid are below the target 0.001 or more than 30% of the points are above 0.3, do additional adjustment and restart the process (up to 5 trials).

3. Run 1,000 permutations for each bootstrapped dataset, and calculate permutation p-value of the minimum p-value of each dataset.

4. Select only the data-points with observed permutation p-value in the range 0 to 0.25 (for linear model between 0.001 and 0.25).

5. For each gene we fitted a linear regression and a logistic regression. Then the number of independent tests is (permutation p-value) / (minimum p-value) = $\exp(\beta_0)$(minimum p-value)$^{\beta_1 - 1}$.

3.8.2 Evaluation using 1KGP dataset

Running this setup with MatrixEQTL (a linear model approach for eQTL mapping) (Shabalin 2012) on 14,500 genes with 50,100, and 200 grid points take approximately 23, 28, and 42 days for 1000 permutations - the procedure scales quite linearly for practical number of grid points. In contrast, running TReC model once takes at least 18 days, if we only run TReCASE for those gene-SNP pairs with significant association from MatrixEQTL. We summarize total timing required to fit the dataset in Table 3.6.

As a quick alternative one might simply use eigenMT to estimate the effective number of tests, and then obtain the permutation p-value estimate by multiplying minimum p-value with the effective number of tests, and truncating at 1. We evaluate this approach and our methods (linear regression or logistic regression) using the permutation p-values estimated by 10,000 permutations for 14,566 genes as the true permutation p-values. We see that eigenMT tends to be conservative with a large

Table 3.6: Summarizing total time to fit the data using each method. First column gives time to fit the data. Second column - time to estimate permutation p-value using MatrixEQTL using 100 and in parenthesis for the reference smaller grid of 25-200 data points. Third column gives total time to fit the data. Score method calculates permutation p-value automatically, thus it doesn't require calculating estimated permuted p-values and only has total time presented. For TReCASE(score) method 5,000 permutation are done. TReCASE(LRT)* is modification that pre-filters SNP that were not found to be significant after fitting MatrixEQTL (using p-value cutoff 0.01).

method	one run	est. perm.pval	total time
RASQUAL	610	28 (19-42)	638 (629-652)
TReCASE(score)	-		750 (750)
TReCASE(LRT)	53	28 (19-42)	81 (72-95)
TReCASE(LRT)*	18	28 (19-42)	46 (37-60)

number of false negatives, especially at less significant p-values (Table 3.7). This suggests that the eigenMT estimates of the number of tests is too large, particularly so for larger p-values. Overall the results based on linear fit is much more accurate than eigenMT (in terms of smaller number of false positives + false negatives), though it produces unbalanced false positives and false negatives, with more false positives at larger p-value cutoff and more false negatives at smaller p-value cutoffs (Table 3.8). Finally, the logistic regression has the the most accurate estimates with balanced numbers of false positives and false negatives (Table 3.9).

Table 3.7: Permutation p-value estimated by eigenMT

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	5119	8354	5	1088	1093	6207	8359
0.05	4449	9419	4	694	698	5143	9423
0.01	3403	10878	10	275	285	3678	10888
0.005	3094	11253	12	207	219	3301	11265
0.001	2579	11877	25	85	110	2664	11902

In terms of false classifications (Table 3.10), we do not get as much improvement by using more than 100 grid points, particularly for logistic regression (glm) approach.

Table 3.8: Permutation p-value estimates based on gene-by-gene linear regression

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	6200	8262	97	7	104	6207	8359
0.05	5138	9338	85	5	90	5143	9423
0.01	3657	10852	36	21	57	3678	10888
0.005	3270	11241	24	31	55	3301	11265
0.001	2612	11876	26	52	78	2664	11902

Table 3.9: Permutation p-value estimates based on gene-by-gene logistic regression

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	6180	8341	18	27	45	6207	8359
0.05	5109	9398	25	34	59	5143	9423
0.01	3661	10858	30	17	47	3678	10888
0.005	3284	11235	30	17	47	3301	11265
0.001	2636	11863	39	28	67	2664	11902

The performance of linear regression and logistic regression become similar at larger grid points and more significant p-values. Consequently, we suggest using 100 grid points to estimate permutation p-values, though even with 25 grid points we observe large improvement against eigenMT.

Table 3.10: Number of misclassifications of permutation p-value estimates for 25, 50, 100 and 200 grid points.

permutation	eigenMT	lm				glm			
p-value cutoff		25	50	100	200	25	50	100	200
0.1	1093	139	103	10494		66	61	45	42
0.05	698	112	99	90	88	74	53	59	55
0.01	285	65	65	57	57	47	52	47	47
0.05	219	77	59	55	48	59	54	47	51
0.001	110	81	79	78	69	72	59	67	60

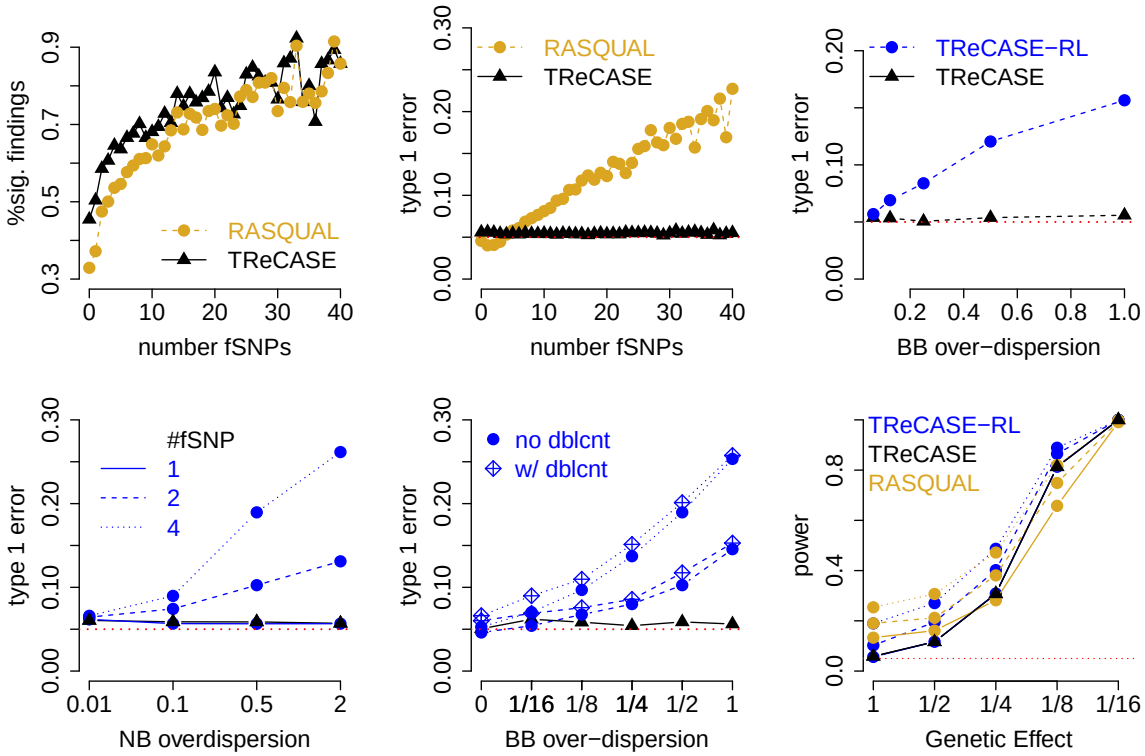

Figure 3.14: Summary of observed method performance: (A) Compare the number of significant findings (q-value < 0.05) between TReCASE and RASQUAL for different number of feature SNPs (fSNPs) using 1KGP data with sample size of 280. (B) The number of significant findings (p-value <0.05) after permuting SNP genotypes, which provides an empirical estimate of type I error. For panels (C)-(F) we run 10,000 replicates per each simulation profile. (C) Evaluation of type I error for TReCASE and TReCASE-RL when there is over-dispersion within a sample and the same amount of extra over-dispersion across samples. We assume there are 2 heterozygous fSNPs per gene and per sample. TReCs were simulated with negative-binomial with over-dispersion 0.5. (D-F) Simulation settings without over-dispersion for ASReC within a sample and with some over-dispersion across samples. We consider the cases where the ASReC is distributed across 1, 2, or 4 fSNPs. (D) type I error when the over-dispersion of negative binomial (NB) and beta-binomial (BB) are the same. (E) Effect of over-counting. We assume 15% double-counting and simulate the data assuming NB over-dispersion to be 0.5. In order to distinguish pure double-counting effect we fit both models RASQUAL like way. (F) power analysis when the over-dispersion of NB and BB are both 0.5.

From the scatter plots of permutation p-value estimates versus "true" permutation p-values estimated by 10,000 permutations, we can see clearly the bias by eigenMT and linear model, as well as the advantage of using larger number of grid points (Figure 3.15).

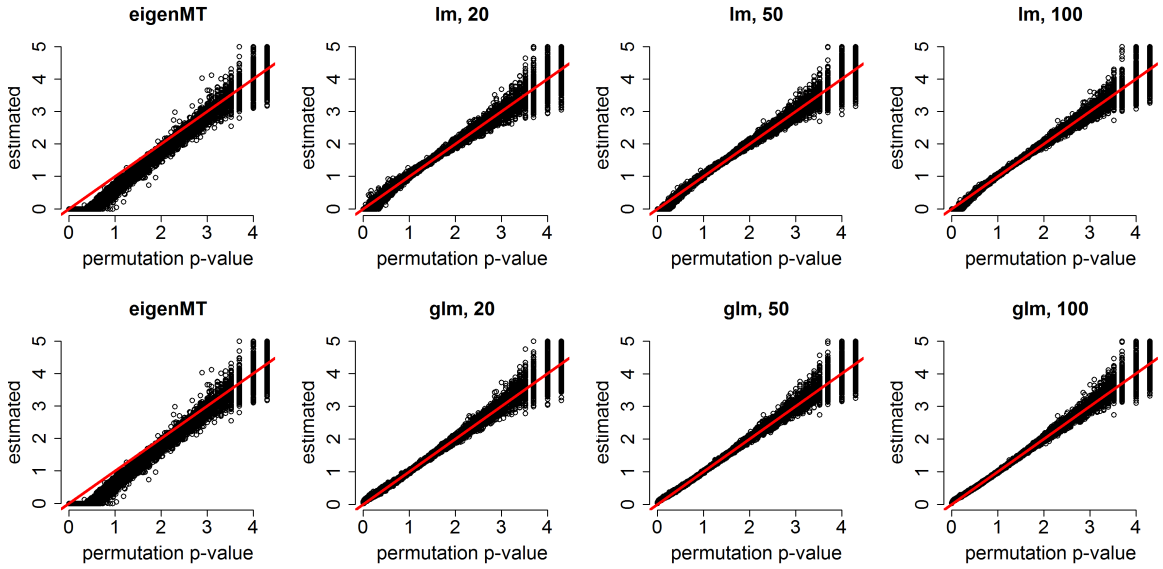

Figure 3.15: Permutation p-value estimation using three methods for 1KGP dataset: eigenMT, linear model (lm), and logistic model (glm). On the x-axis we plot permutation p-values estimated by 10,000 permutations.

3.8.3 Evaluation using GTEx dataset

Evaluation of different methods on GTEx dataset leads to similar results. EigenMT is conservative. Linear regression approach brings significant improvement and logistic regression is still the most stable and accurate one.

Table 3.11: Permutation p-value estimates with eigenMT approach

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	4001	11636	0	1000	1000	5001	11636
0.05	3292	12750	0	595	595	3887	12750
0.01	2285	14148	0	204	204	2489	14148
0.005	2046	14457	3	131	134	2177	14460
0.001	1654	14918	11	54	65	1708	14929

Table 3.12: Permutation p-value estimates with linear regression using 100 grid points.

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	4990	11521	115	11	126	5001	11636
0.05	3882	12658	92	5	97	3887	12750
0.01	2469	14127	21	20	41	2489	14148
0.005	2153	14443	17	24	41	2177	14460
0.001	1664	14908	21	44	65	1708	14929

Table 3.13: Permutation p-value estimates with logistic regression using 100 grid points.

permutation	true		false			number of	
p-value cutoff	pos.	neg.	pos.	neg.	total	pos.	neg.
0.1	4968	11612	24	33	57	5001	11636
0.05	3870	12722	28	17	45	3887	12750
0.01	2471	14124	24	18	42	2489	14148
0.005	2159	14433	27	18	45	2177	14460
0.001	1679	14900	29	29	58	1708	14929

Table 3.14: Number of misclassifications of permutation p-value estimates for 25, 50, 100 and 200 point grid.

permutation p-value cutoff	eigenMT	lm				glm			
		25	50	100	200	25	50	100	200
0.1	1000	184	121	126	106	82	67	57	60
0.05	595	100	108	97	89	55	43	45	44
0.01	204	52	50	41	38	48	45	42	34
0.005	134	65	60	41	48	50	45	45	46
0.001	65	71	62	65	57	55	56	58	54

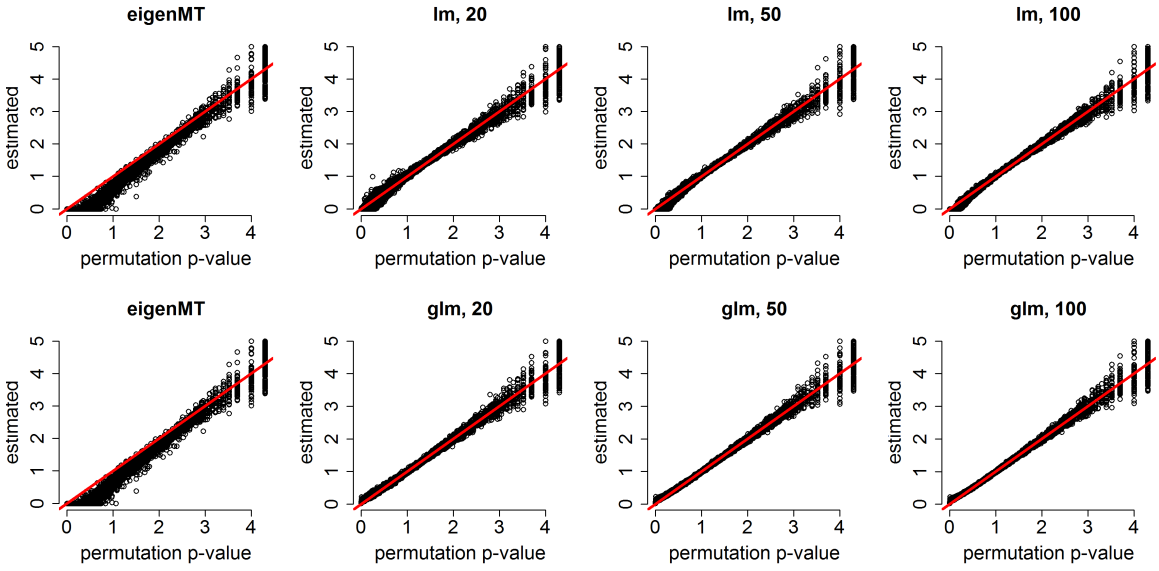

Figure 3.16: Permutation p-value estimation using three methods: eigenMT, linear regression and logistic regression using GTEx dataset. The x-axis are permutation p-values estimated by 10,000 permutations.

3.9 Comparison of MatrixEQTL, TReCASE and RASQUAL using 1KGP dataset

3.9.1 Comparison of computational time

We performed eQTL mapping all the genes for which we had at least 5 samples with at least 5 allele-specific counts. Since RASQUAL is very time consuming, we fitted each gene in parallel. For timing comparability, we did the same for TReCASE. We limited total computational time per gene to be a week, and RASQUAL failed to finish within a week for 9 genes. We summarize the results for the remaining 14,427 genes. The average number of potential eQTL SNPs (rSNPs) per gene was 2,000. The computational time of both TReCASE and RSQUAL increases nearly linearly with sample sizes, and TReCASE is more than 10 times faster than RASQUAL (Figure 3.17(a)). For the full 1KGP with sample size of 280, RASQUAL took 610 days and TReCASE using likelihood ratio test (LRT) took 53 days (Table 3.6).

We have developed a modified version of TReCASE to perform testing using score test (Hu et al. 2015). This TReCASE (score) method is computationally more efficient to perform permutations since some elements of the score test statistic can be calculated only once and used for many permutations. For 280 samples in the 1KGP dataset, TReCASE (score) method took 750 days for 5,000 permutations (Table 3.6), and thus it is doable within a week using 100+ computing jobs. However, one limitation of score test is that the p-values become less stable when sample size is relatively small, such as n=50 or 100. While the permutation p-values are still accurate, one should be cautious when using the p-values with small sample sizes.

The relative computational time of RASQUAL versus TReCASE increases with respect to the number of fSNPs (Figure 3.17(b)) and the number of rSNPs (Figure 3.17(c)). When genotype data were obtained by whole genome sequencing

(e.g., GTEx dataset), there are about 10 times of fSNPs per gene (Figure 3.18) than the genotype data obtained from imputation (e.g., 1KPG dataset). Therefore, as a consequence, RASQUAL can be 100 times slower than TReCASE.

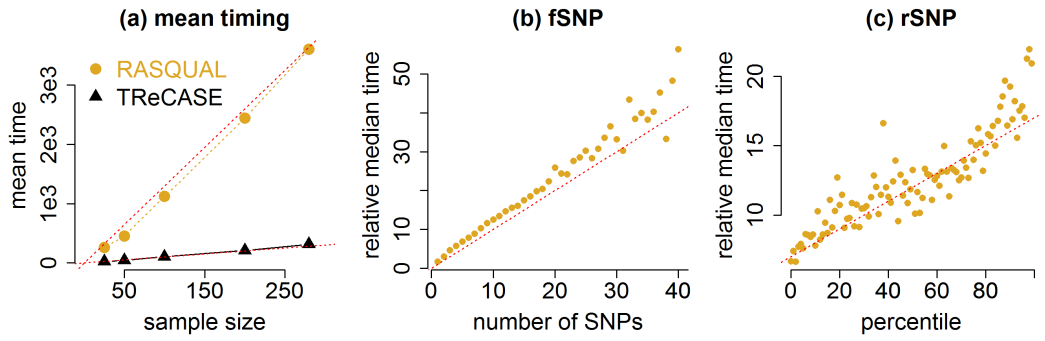

Figure 3.17: TReCASE vs RASQUAL timing to fit a gene: (a) The mean time (seconds) for eQTL mapping per gene by sample size - dotted lines $y = x$ and $y = 13x$ are added for reference. (b-c) The relative median time for eQTL mapping per gene using RASQUAL versus TReCASE with respect to the number of fSNPs (with $y = x$ line added for reference) or the number of rSNPs (with line $y = 7 + 0.1x$ added for the reference).

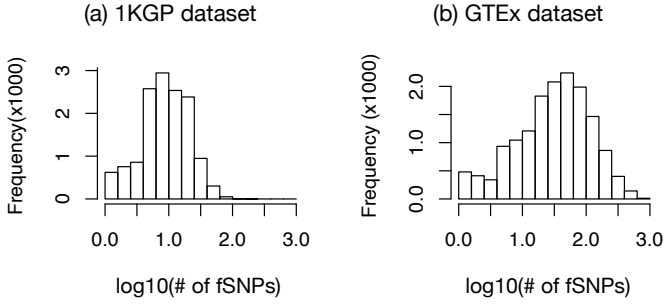

Figure 3.18: Number of fSNPs per gene. The distribution of the number of fSNPs per gene for 1KGP dataset (a) and GTEx dataset (b), respectively.

3.9.2 Choose a permutation p-value cutoff to control FDR

Given the permutation p-value for each gene, we can choose a permutation p-value cutoff to control FDR. We estimated fraction of genes in null distribution by doubling fraction of genes with permutation p-value above 0.5 after which Storey q-value is calculated. Because a larger proportion of genes has significant eQTLs, for a FDR cutoff, the corresponding permutation p-value can be even larger than the FDR (Table 3.15). For example, to control FDR at 0.05, the p-value cutoffs are larger than 0.05. Here we use FDR cutoff 0.01 to choose permutation p-value cutoffs for the three methods.

Table 3.15: Permutation p-value cutoffs for different FDR cutoffs, using 1KGP dataset with sample size 280.

FDR	TReCASE	RASQUAL	MatrixEQTL
0.001	0.001	0.0009	0.0005
0.01	0.014	0.011	0.009
0.05	0.092	0.071	0.074
0.1	0.211	0.161	0.189
0.2	0.474	0.362	0.477
0.25	0.616	0.470	0.631

It is interesting to check how the number of discoveries varies with sample size. We down-sampled the 1KGP dataset to sample sizes from 35 to 140, and calculated the number of significant discoveries by FDR 0.01 (Table 3.16). It is clear that TReCASE discovery a much larger number of eQTLs than MatrixEQTL, a linear model-based approach.

Table 3.16: Number of significant genes by method at FDR 0.01.

sample size	MatrixEQTL	TReC	TReCASE
35	0	224	454
70	266	481	1119
140	1498	1865	3583
280	4501	5038	7447

3.9.3 Results of MatrixEQTL and TReCASE show consistent patterns

Comparing p-values versus MatrixEQTL we observe that their correlations grow with sample size (Table 3.17)

Table 3.17: Correlations of TReC and TReCASE p-values vs MatrixEQTL p-values on -log10 scale

samples	TReC	TReCASE
35	0.72	0.67
70	0.89	0.77
140	0.95	0.85
280	0.97	0.87

3.9.4 Compare the results of RASQUAL vs. TReCASE

We classified the fitted genes by their significance according to each of two methods (Table 3.18). Although the results of the two methods are consistent for the majority of the genes, there is some notable discrepancy for a subset of genes. We study potential sources for discrepancies in the following subsections.

Table 3.18: Number of genes passing corresponding cutoff of q-values applied on permuted p-values.

TReCASE / **RASQUAL**	$[0.01, 1]$	$[10^{-3}, 0.01)$	$[10^{-4}, 10^{-3})$	$[10^{-5}, 10^{-4})$	< 1e-5	total
$[0.01, 1]$	6747	1103	303	151	153	8457
$[10^{-3}, 0.01)$	429	420	231	109	130	1319
$[10^{-3}, 10^{-4})$	195	189	134	121	173	812
$[10^{-4}, 10^{-5})$	103	92	92	94	177	558
< 1e-5	377	246	199	254	2344	3420
total	7851	2050	959	729	2977	14566

Additionally we notice that RASQUAL estimates one over-dispersion that is shared by its negative binomial and beta-binomial components and TReCASE estimates the over-dispersion parameters for these two components separately. We observe a quite clear pattern that the over-dispersion from RASQUAL is very similar to the

over-dispersion of TReCASE negative binomial component, but they are often much larger than the over-dispersion of TReCASE beta binomial component (Figure 3.19).

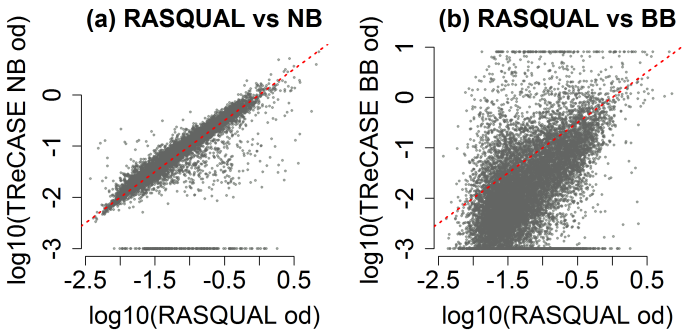

Figure 3.19: Comparing an estimate of RASQUAL over-dispersion versus observed: (a) TReCASE negative-binomial over-dispersion estimates and (b) TReCASE beta-binomial over-dispersion estimates. We trimmed over-dispersion values from TReCASE output to $[-3, 1]$ range.

3.9.5 Discrepancy of the results between TReCASE and RASQUAL

For each gene, we took the smallest p-value across multiple rSNPs by TReCASE and RASQUAL, truncated them at 10^{-15}, and refer them as TReCASE p-value and RASQUAL p-value, respectively. We sought to explore the discrepancies of TReCASE and RASQUAL p-values by a linear regression with

$$y = \log_{10}(\text{TReCASE p-value}) - \log_{10}(\text{RASQUAL p-value})$$

as the response variable and 16 covariates (Tables 3.19-3.20):

- alternative allele frequency
- $\log_{10}$ p-value of χ^2 test for Hardy Weinberg equilibrium
- estimated mapping error (Delta), which is an output of RASQUAL

- reference allele bias (Phi Bias), which is an output of RASQUAL
- the number of feature SNPs per gene, centered to median.
- the number of rSNPs per gene, log scale.
- over-dispersion from total read counts (OD_{NB}) estimated by TReCASE, in log scale and centered.
- over-dispersion from allele-specific counts (OD_{BB}) estimated by TReCASE, in log scale and centered.
- the total allele-specific counts by RASQUAL, in log scale and centered.
- the total allele-specific counts by TReCASE, in log scale and centered.
- interaction of the previous two counts. Two methods approach differently to count allele-specific reads. TReCASE count them at gene level while RASQUAL count them SNP by SNP. Therefore if one read overlap with two heterozygous SNPs, it will be counted twice. The potential degree of over-counting is illustrated at Figure 3.21.
- interactions of three covariates: RASQUAL allele-specific counts, the number of fSNPs, and beta binomial over-dispersion. Based on our simulations and study of information matrix we believe that p-value inflation of RASQUAL is caused by its model of ASReC. These three covariates are all important for the ASReC model.
- median p-value of RASQUAL across all rSNPs of a gene, using permuted genotype. This quantity measures the magnitude of RASQUAL type I error for this gene.

Table 3.19: Type 1 (sequential) and Type 3 (added last) ANOVAs for linear regression analysis of $\log_{10}$(TReCASE p-value) - $\log_{10}$(RASQUAL p-value). The direction (Dir.) indicates whether RASQUAL (R) or TReCASE(T) has smaller p-value.

Parameter	Dir.	Type 1 R^2	P-val	Type 3 R^2	P-val	Marg. R^2
OD_{BB}	R	6.7	6e-244	8.2	3e-294	6.7
$ASReC_R$	R	2.7	8e-101	1	2e-39	0.6
n-fSNP	R	1.9	1e-71	1.6	4e-63	2.7
$ASReC_R$:n-fSNP	R	0.5	2e-19	0.7	4e-26	0
$ASReC_R$:OD_{BB}	R	5	1e-185	3.2	1e-118	1.3
n-fSNP:OD_{BB}	R	0.8	6e-31	1	1e-37	0.4
$ASReC_R$:n-fSNP:OD_{BB}	R	0.2	4e-8	0.2	2e-10	0.1
n-rSNP	T	0	0.73	0	0.52	1
OD_{NB}	T	0	0.009	0	0.004	0.8
$ASReC_T$	T	0.4	1e-15	0.4	3e-16	0.1
$ASReC_R$:$ASReC_T$	R	0.3	6e-13	0.2	1e-10	0
AF	T	0	0.004	0.1	0.002	0
HWE χ^2	T	0.2	5e-9	0.2	1e-8	0.3
Mapping error	T	0	0.09	0	0.24	1.1
Ref. Allel Bias	R	0.3	2e-13	0.3	2e-13	2.1
Med(perm-p)	T	0	0.97	0	0.97	0.4

All covariates were normalized to have standard deviation of 1.

This linear model explains 15% of the variance of y. More variance of y can be explained by this model if we only consider a subset of genes with more discrepant p-values. For example, for a subset of genes passing a cutoff of $|y| \geq 5$ or $|y| \geq 10$, this linear model explains 42% or 57% of the variance of y, respectively. We note that the more discrepant set of genes we select, the higher fraction of genes with smaller p-values by RASQUAL (Figure 3.20 (a)), suggesting stronger discrepancies are more likely due to the inflation of Type I error by RASQUAL.

We observe that three factors, beta-binomial over-dispersion (OD_{BB}), the number of fSNPs (n-fSNP), and RASQUAL style ASReC ($ASReC_R$), along with interactions of these terms have strongest associations with the discrepancy of the two methods (Table 3.20). Larger values of these three factors are all associated with smaller RASQUAL p-value. This is consistent with our findings that the beta-binomial

Table 3.20: Linear regression of y = $\log_{10}$(TReCASE p-value) - $\log_{10}$(RASQUAL p-value) versus a set of potential factors. OD_{BB} and OD_{NB} indicate $\log_{10}$ over-dispersion for beta-binomial and negative binomial, respectively. n-fSNP and n-rSNP indicate the number of feature SNPs and regulatory SNPs, respectively. Subscript $_R$ and $_T$ indicate RASQUAL and TReCASE, respectively.

Parameter	Est.	SE	P-val	Marg.Est	Marg P-val
intercept	0.18	0.14	0.2	0.31	1.6e-43
OD_{BB}	1.04	0.03	3e-294	0.68	3e-213
$ASReC_R$	0.81	0.06	2e-39	0.20	4e-19
n-fSNP	0.44	0.03	4e-63	0.43	4e-84
$ASReC_R$:n-fSNP	0.24	0.023	4e-26	0.013	0.51
$ASReC_R$:OD_{BB}	0.52	0.022	1e-118	0.26	7e-42
n-fSNP:OD_{BB}	0.29	0.022	1e-37	0.13	2e-12
$ASReC_R$:n-fSNP:OD_{BB}	0.11	0.017	2e-10	-0.04	0.0027
n-rSNP	-0.014	0.022	0.52	0.26	6e-32
OD_{NB}	-0.07	0.024	0.004	0.24	1e-26
$ASReC_T$	-0.46	0.056	3e-16	0.084	2e-4
$ASReC_R$:$ASReC_T$	0.12	0.019	1e-10	0.037	0.048
AF	-0.061	0.02	0.002	-0.044	0.046
HWE χ^2	-0.12	0.02	1e-8	-0.16	3e-12
Mapping Error	-0.028	0.024	0.24	0.28	2e-36
Ref. Allel Bias	0.19	0.026	2e-13	0.38	6e-65
Med(perm p)	-0.001	0.034	0.97	-0.16	5e-13

component of RASQUAL treats multiple fSNPs within a sample as independently distributed, which causes larger inflation of type I error. In addition, more discrepant genes also tend to have weaker correlations of over-dispersion estimates between TReCASE and RASQUAL (Figure 3.20 (c)).

Among other covariates, the following relations are notable. Significant associations with reference allele bias and Hardy-Weinberg disequilibrium suggest the advantage of RASQUAL to model these factors. Smaller TReCASE p-values are associated with the case when we observed relatively higher TReCASE style allele-specific counts ($ASReC_T$), which suggests that the ASReC by RASQUAL and TReCASE are different, most likely due to double counting by RASQUAL (Figure 3.21).

Median RASQUAL p-value using permuted data (Med(perm p)) is very significant

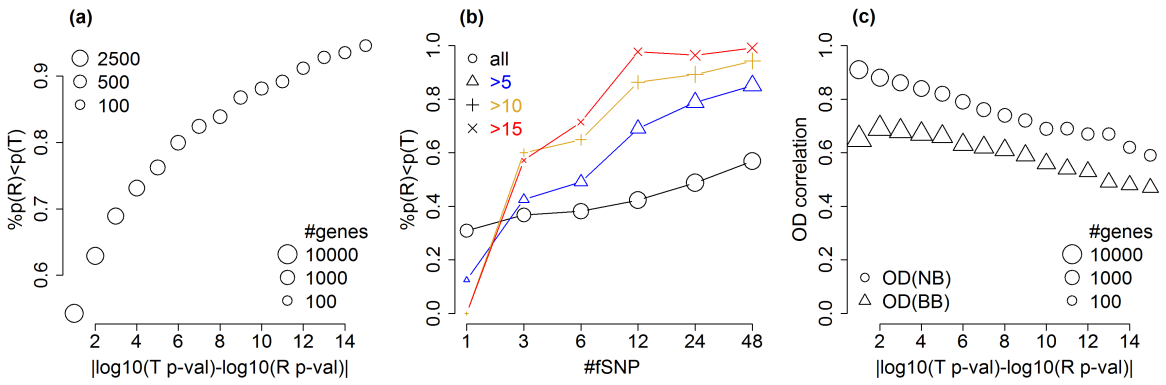

Figure 3.20: Comparison of TReCASE and RASQUAL results using 1KGP dataset. We use "T p-val" and "R p-val" as abbreviations of "TReCASE p-value" and "RASQUAL p-value", respectively. "%p(R) < p(T)" denotes the proportion of genes with RASQUAL p-value smaller than TReCASE p-value. (a) Genes with more discrepant p-values tend to smaller RASQUAL p-values. (b) The genes with larger number of fSNPs also tend to have smaller RASQUAL p-values. Different point symbols indicate the genes with the absolute value of the difference between $\log_{10}$(TReCASE p-value) and $\log_{10}$(RASQUAL p-value) is larger than certain threshold. (c) When there are larger discrepancies of p-values, the over-dispersion estimates by RASQUAL are less similar to either negative binomial (NB) or beta-binomial (BB) over-dispersion estimates by TReCASE.

in marginal model, but becomes much less significant in the joint model. This is expected because inflation of type I error is also associated with other factors in the join model (see Section 3.5.4 for more details). In both cases, smaller RASQUAL p-value using permuted data are associated with smaller RASQUAL p-values.

We further examine the discrepancy of significant findings by RASQUAL and TReCASE with respect to the number of fSNPs. We classified the genes to be significant or not at several FDR cutoffs and plotted them versus the number of fSNPs. The fraction of significant findings of both methods generally grows with respect to the number of fSNPs (Figure 3.22(a) and (d)), which is expected since the number of allele-specific reads would also be higher with more fSNPs. However, this fraction grows quicker for RASQUAL than TReCASE, and it increases regardless the

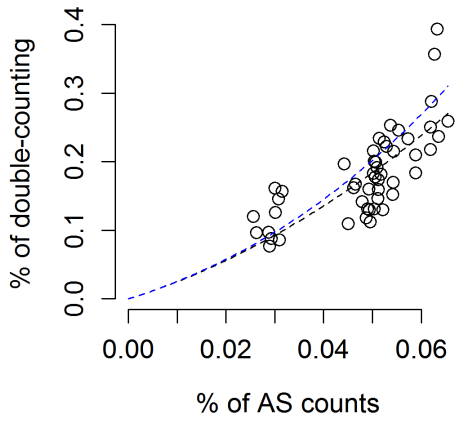

Figure 3.21: Estimating double-count in real data. We again used 30 samples from PRJNA385599. For each sample we counted number of allele-specific reads using our TReCASE procedure and produced fraction with respect to total number of reads (x) and will plot on x scale. In addition for the reads overlapping several heterozygous SNPs we counted such read several time - once for each SNP (define this number as z) Z is inflated with overcounting. We quantify this excess of counts by defining $y = z/x - 1$ and plotting them on the y axis. 10 of the samples in this dataset were measured with both 150bp reads and shorter 75bp reads. They are plotted separately with 10 points around 3% allele-specific counts representing summary for shorter reads.

significance level of TReCASE (Figure 3.22(b) and (c)). This suggests that the association between the number of fSNPs and RASQUAL p-values may not depend on the actual strength of eQTL signals and thus implies inflated type I error. In contrast, conditioning on being significant or insignificant using RASQUAL method we see much weaker association between the number of fSNPs and the fraction of significant TReCASE findings (Figure 3.22(e) and (f)). In fact, given significant RASQUAL results, the number of significant TReCASE findings has slight decrease as number of fSNPs increases (Figure 3.22(e)). This is likely because RASQUAL tends to find higher fraction of false positive when the number of fSNPs is large.

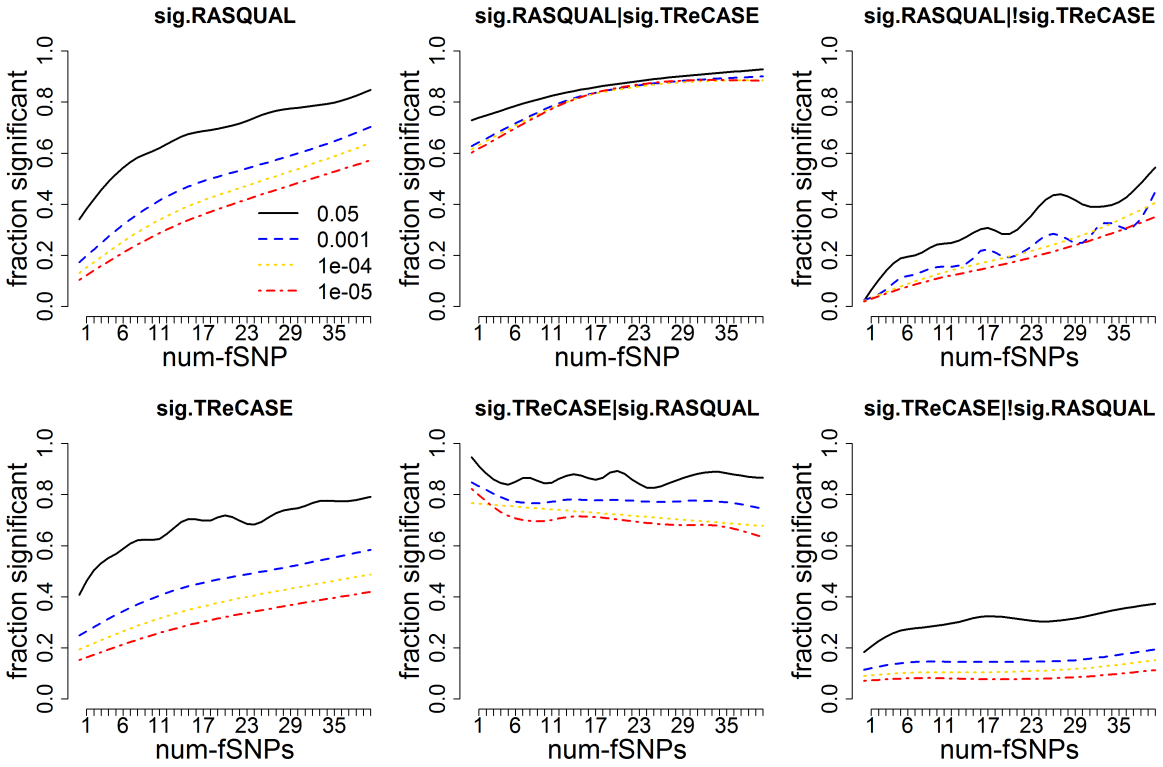

Figure 3.22: Method discrepancy conditioned on significance status of each method. We classify the genes into significant or not significant category using FDR cutoffs presented in the legend: 0.05, 1e-3, 1e-4 and 1e-5 plotted vs number of fSNPs. The curve is obtained using a spline. Panels (a) and (d) consider overall dependency of fraction of genes found to be significant plotted versus number of fSNPs. Panel (b) considers proportion of genes passing a cutoff in RASQUAL model for all the genes passing cutoff for TReCASE. Panel (e) does it other way around - fraction of significant genes found by TReCASE among the genes significant in RASQUAL. Panels (c) and (f) provide similar curves for fraction of genes found to be significant by one of the methods, given that they weren't found to be significant by the other method.

3.9.6 eQTL mapping using permuted genotype data

In this subsection, we evaluate potential type I error inflation of TReCASE and RASQUAL using permuted genotype data. We only include the potential *cis*-acting eQTLs in our evaluation because ASE is only informative for *cis*-eQTL mapping. To identify *cis*-acting eQTLs, we test whether the eQTL effects estimated by TReC and ASE are the same by a cis-trans test (Sun 2012), and consider the cases with cis-trans test p-value > 0.05 as the potential *cis*-acting eQTLs. For TReCASE method, we consider the standard TReCASE using likelihood ratio test (LRT) (Sun 2012) as well as another version using score test (Hu et al. 2015).

From the distribution of all the eQTL p-values, it is clear that RASQUAL has severe inflated type I error (Figure 3.23(c,f)). This is consistent with our analysis of likelihood model (Section 3.5.4), simulation results, and comparison of the results on 1KGP dataset between RASQUAL and TReCASE (Section 3.9.5). At this large sample size of 280, the p-value distribution from TReCASE (score) is slightly deviated from uniform distribution (Figure 3.23(a,d)), though such deviation becomes larger for smaller sample size of 100 (Figure 3.24(a,d)). Therefore when using TReCASE (score) method for small sample size, we recommend using permutation p-values rather than the p-values from asymptotic distribution.

We observed that TReCASE also has slight inflated type I error. This may be due to model mis-specification for some genes, either because the distribution assumption is not accurate or missing some covariates (e.g., genetic effect are missing after permuting genotype data). Such inflation of type I error can be removed after trimming outlier values using an approach implemented in DESeq2 Love et al. (2014). An observation is defined as an outlier if it Cook's distance is larger than a thresholds, and found the threshold of $4/n$ effectively removes the inflation of type I error (Figure 3.25).

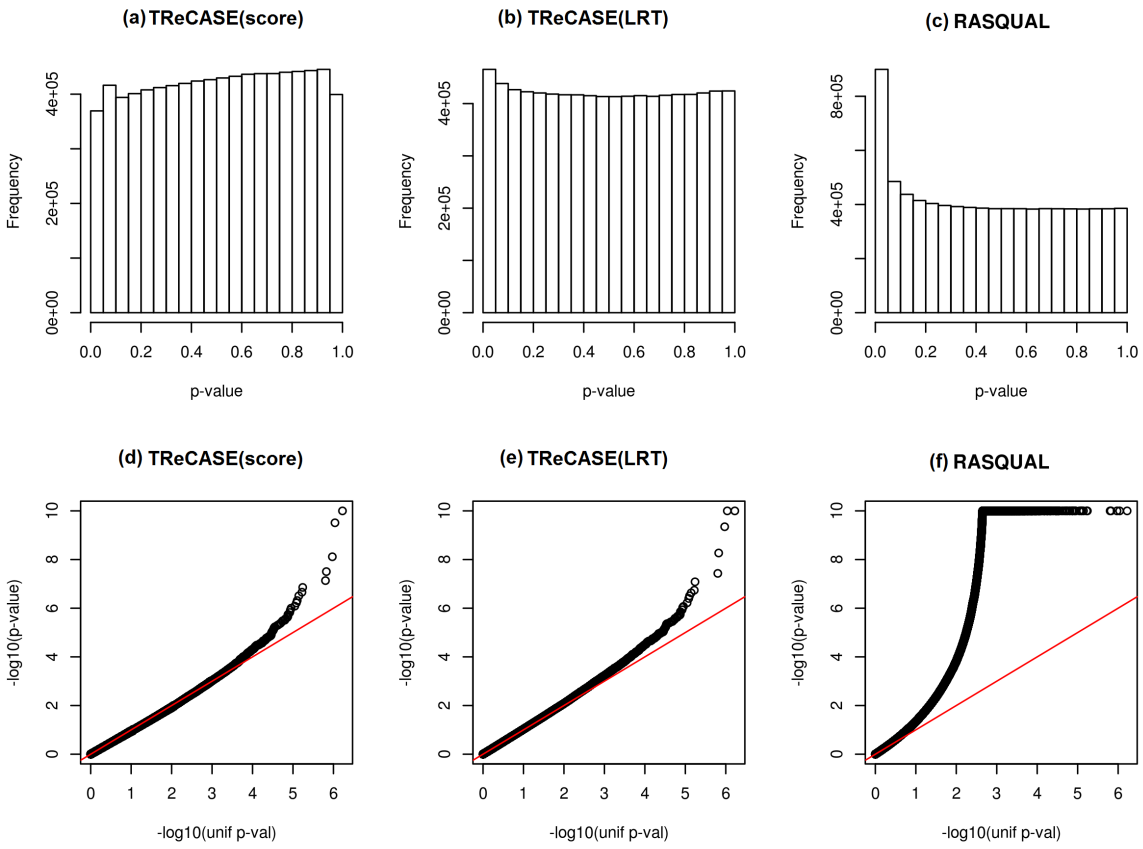

Figure 3.23: The distributions ((a)-(c)) and QQ-plots ((d)-(f)) of eQTL p-values using permuted genotypes by three methods: TReCASE (LRT), TReCASE(Score) and RASQUAL, using 1KGP dataset with sample size 280.

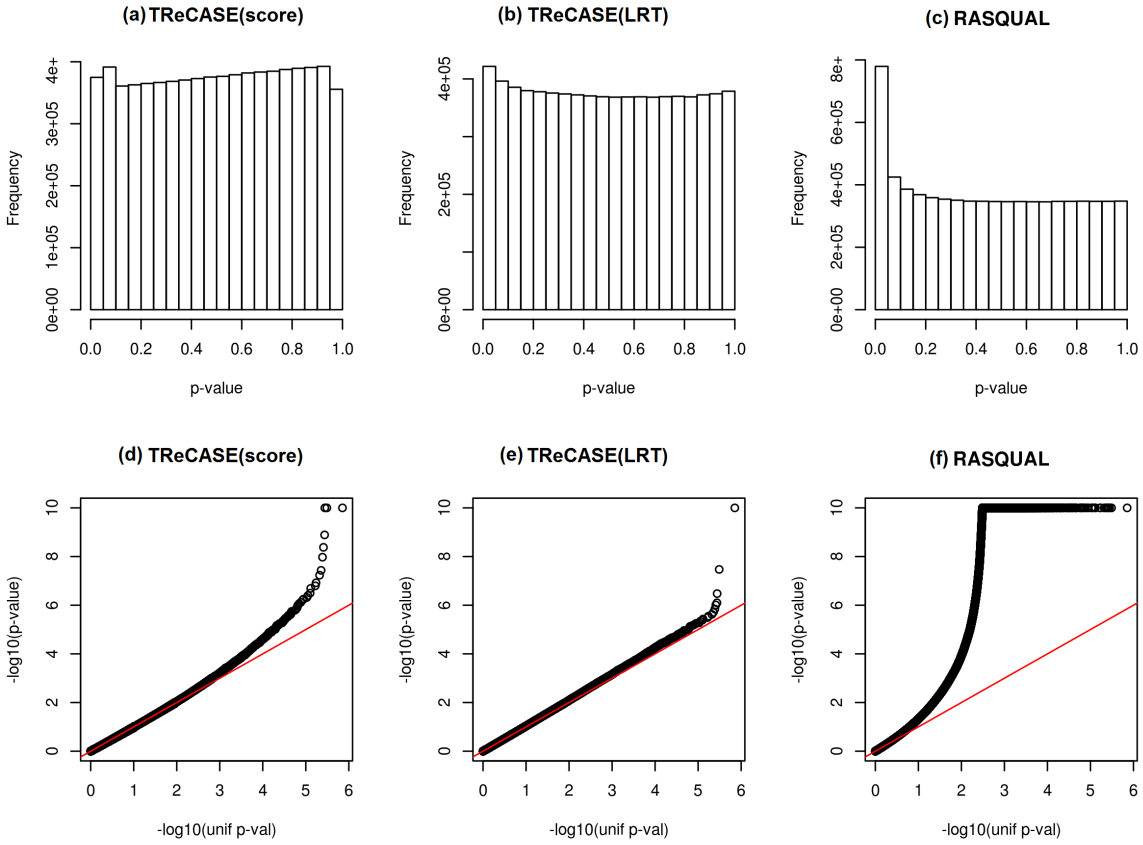

Figure 3.24: The distributions ((a)-(c)) and QQ-plots ((d)-(f)) of eQTL p-values using permuted genotypes by three methods: TReCASE (LRT), TReCASE(Score) and RASQUAL, using 1KGP dataset with sample size 100.

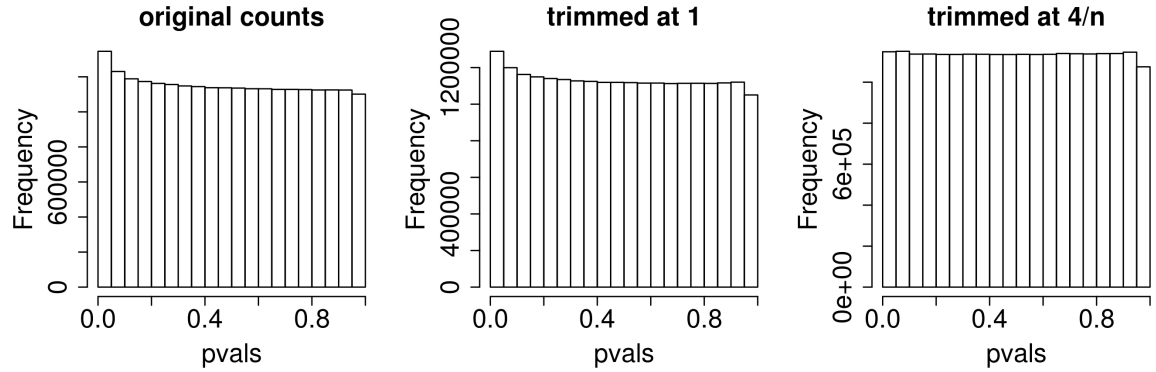

Figure 3.25: Fitting the data with permuted genotypes using TReC model after trimming counts with significant Cook's distances

3.9.7 eQTL positions with respect to transcription start and transcription end

We examine the locations of eQTLs. For each gene, we chose the SNP with the smallest p-value as its eSNP and further select gene-eSNP pairs using different permutation p-value cutoffs. Then for those selected gene-eSNP pairs, we ask where those eSNPs are located with respect to the locations of their associated genes. We found that eSNPs are enriched at Transcription Starting Sites (TSS) or the end of the transcript. We observe strong enrichment on TSS for genes in both positive and negative strand. (Figure 3.26).

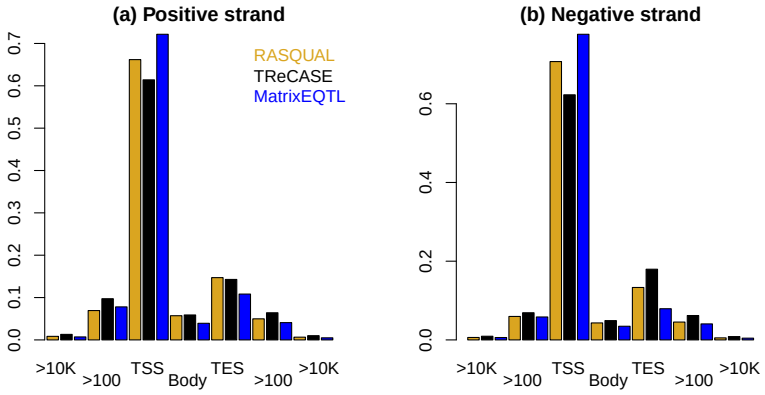

Figure 3.26: Distance from most significant SNP to transcription starting site (Using FDR cutoff 0.01 applied to permutation p-values): (a) Positive strand and (b) Negative strand. For each of three methods genes are classified into 7 categories with respect to gene-body: those that are more than 10K bases from transcription starting site (TSS), those within 10K bases from TSS, but more than 100 bases from TSS, 100 bases around TSS, within body genes, within 100 bases around transcription end site (TES), 100 to 10K bases from TSS and more than 10K from TSS plotted in this order. To adjust for the fact that each category had different width we normalized counts to adjust for interval length.

3.10 Compare the eQTLs identified by TReCASE versus MatrixEQTL using both 1KGP and GTEx data

We compare the number of findings by TReCASE and MatrixEQTL at different q-value thresholds using the results of 1KGP dataset (Table 3.21). MatrixEQTL was able to produced 3,356 genes passing $q = 0.01$ cutoff versus 6,715 found by TReCASE. Applying eigenMT to MatrixEQTL results would make the results even more conservative declaring only 2,773 genes to be significant at q-value 0.01. This decrease of power is more visible after applying FDR correction, since for more conservative eigenMT method estimated fraction of null genes π_0 is much higher: while for TReCASE permutation p-values it is estimated to be 23.5%, for MatrixEQTL permutation p-value estimate of π_0 is 41.5% and for eigenMT - 94.2%.

Table 3.21: Number of genes passing corresponding cutoff of q-values applied on permuted p-values in 1000 Genomes dataset. TReCASE vs MatrixEQTL

MatrixEQTL / **TReCASE**	$[0.01, 1]$	$[10^{-3}, 0.01)$	$[10^{-4}, 10^{-3})$	$[10^{-5}, 10^{-4})$	< 1e-5	total
$[0.01, 1]$	7574	168	38	15	56	7851
$[10^{-3}, 0.01)$	1709	224	60	30	27	2050
$[10^{-4}, 10^{-5})$	655	147	73	29	55	959
$[10^{-5}, 10^{-4})$	417	103	72	46	91	729
< 1e-5	855	299	239	223	1361	2977
total	11210	941	482	343	1590	14566

GTEx dataset shows more dramatic gain from MatrixEQTL to TReCASE Table 3.22 with only 1,878 genes passing $q = 0.01$ cutoff fpr MatrixEQTL versus 7,850 found by TReCASE. Applying eigenMT to MatrixEQTL results would make the results even more conservative declaring only 1,662 genes to be significant at the given level. We do observe similar level of estimate of π_0 for TReCASE - 21.2% while higher than in previous case fractions for either MatrixEQTL permuted p-values or eigenMT corrected p-values - 57.4% and 100%

Table 3.22: Number of genes passing corresponding cutoff of q-values applied on permuted p-values in GTEx. TReCASE vs MatrixEQTL

MatrixEQTL / **TReCASE**	$[0.01, 1]$	$[10^{-3}, 0.01)$	$[10^{-4}, 10^{-3})$	$[10^{-5}, 10^{-4})$	< 1e-5	total
$[0.01, 1]$	8721	36	11	2	7	8777
$[10^{-3}, 0.01)$	2067	71	26	11	9	2184
$[10^{-4}, 10^{-5})$	927	78	28	15	11	1059
$[10^{-5}, 10^{-4})$	543	58	30	20	28	679
< 1e-5	2491	264	212	152	809	3928
total	14749	507	307	200	864	16627

Next we study whether the eSNPs (the most significant eSNPs per gene) found by MatrixEQTL and TReCASE are located in different genomic regions, in terms of the 18 chromatin states classification provided by Roadmap Epigenomic Consortium (Kundaje et al. 2015). We consider the results of the two methods are concordant if their eSNPs of the same gene are within 10kb. Considering the SNPs significant at permutation p-value $\alpha = 0.01$ level and contrasting them to the genes non-significant (at $\alpha = 0.1$ level). Then each gene can be assigned to one of 8 categories based on 3 factors, whether the eSNPs found by MatrixEQTL and TReCASE are concordant, and whether the eQTL association is significant for each method. We observed the distribution of these 8 groups has large difference for a few chromatin states (Figure 3.27). The eQTLs identified by both method or by TReCASE only are less likely located in the Weak Repressed PolyComb (ReprPCwk) or Quiescent/Low (Quies) regions. We observed similar patterns in the results from GTEx dataset (Figure 3.28).

Alternatively we considered a different classification from the same project in which certain DNase enriched regions were classified as promoter or enhancer. For these genes we estimated the probability of the SNP falls into a promoter or enhancer region based on the significance level by each method and distance from gene to SNP (Table 3.23). We observe that in both datasets stronger TReCASE p-value is

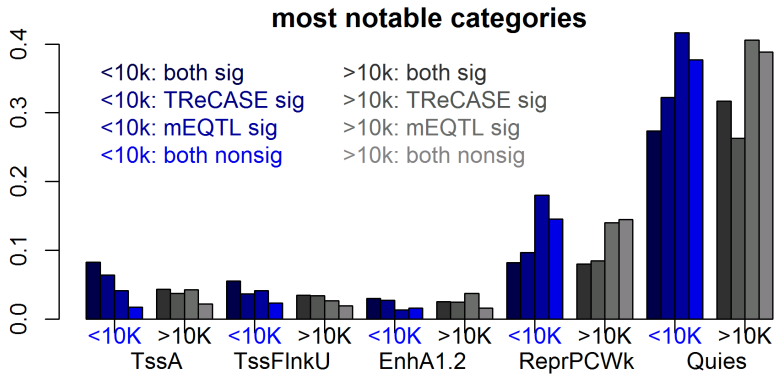

Figure 3.27: Distribution of eQTLs (from 1KGP dataset) in different chromatin states. For each of the 8 categories, we calculated the proportion of eQTLs located in each of the 18 chromatin states. Only the 5 chromatin states with larger difference across the 8 categories are shown. The groups we consider are: (1) TssA - Active TSS, (2) TssFlnkU - Flanking TSS Upstream, (3) EnhA1 - Active Enhancer 1 and Active Enhancer 2, (4) ReprPCwk - Weak Repressed PolyComb and (5) Quies - Quiescent/Low

associated with higher chance to be located within enhancer or particularly promoter while the signal for MatrixEQTL is weaker to non-existent. This observation suggests that incorporation of ASReC not only increases power, but also improves precision.

Table 3.23: Promoter or enhancer status by significance and two method concordance. Results of the logistic model fit with p-values and distance as predictors. P-values are on negative $\log_{10}$ scale, distance is on $\log_{10}$ scale with 1 added. The distance from the gene to a SNP within this gene is considered to be 0.

	1000 Genomes				GTEx			
	promoter		enhancer		promoter		enhancer	
	coef	p-val	coef	p-val	coef	p-val	coef	p-val
Intercept	-2.5	0	-3.21	8e-290	-3.7	4e-254	-3.6	8e-253
p_{mEQTL}	0.01	0.37	-0.02	0.07	-5e-5	0.99	-0.02	0.07
$p_{TReCASE}$	0.05	8e-5	0.02	0.001	0.011	0.03	0.01	0.0003
distance	-0.06	1e-15	-0.005	0.58	-0.014	0.18	0.003	0.76

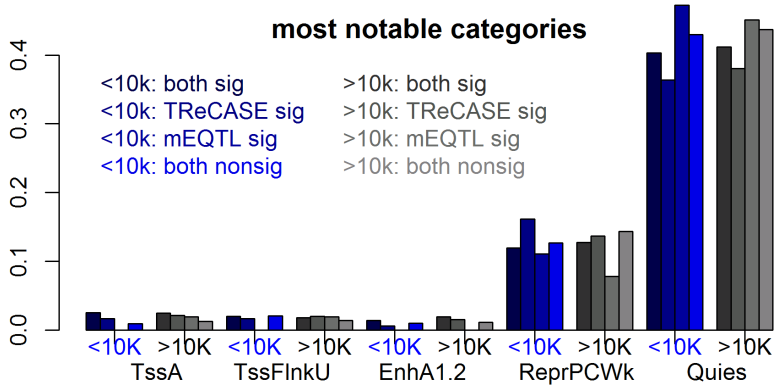

Figure 3.28: The distribution of eQTLs (from GTEx dataset) in different chromatin states. Figure uses the same categories as in previous figure.

3.11 Analysis of brain tissues using version 8 GTEx data release

We applied our pipeline for multiple tissues from v8 GTEx data release and updated the results for whole blood tissue.

We obtained access to 76bp paired RNA-seq reads (in BAM format) mapped to hg38 reference of 670 whole blood samples (V8), and for 5 different brain tissues with counts listed in the Table 3.24. The RNA-seq data is available from at The NHGRI AnVIL `https://anvil.terra.bio`.

We filtered RNA-seq reads keeping only those uniquely mapped reads limiting only to proper pairs, by the `scanBamFlag` function from R package `Rsamtools`.

Phased genotype calls (in VCF format) from whole genome sequencing of 838 samples (release V8, hg38) were obtained from NHGRI AnVIL. Based on this genotype dataset, we created a list of heterozygous SNPs for each individual. Then we applied the same approaches as for 1KPG data to collect Total Read Count (TReC) and Allele-Specific Read Count (ASReC) per gene and per sample for TReCASE and RASQUAL.

For analysis we used only the genes with at least 20% of samples having at least 10 total read counts. For allele-specific counts we added an extra check for genes with evidence of conflicting SNP information: Alelle-specific counts for the whole gene were removed in the following scenarios: (a) more then 5% of individuals had reads had conflicting parental information, (b) 1% of individuals had conflicting parental information and fraction of individuals with extreme allele-specific proportion (<0.10 or >0.90) exceeded 20% of the samples, we also removed all individual allele-specific counts if this individual had at least 10% conflicting information allele-specific reads. In the tissues we analyzed only a small fraction of genes had such conflicting information: allele-specific information was lost for 19-38 genes.

Covariates data including 5 genotyping principal components, gender, PCR (an

indicator whether sequencing protocol for the sample was PCR-based or PCR-free), platform (a two factor variable for Illumina HiSeq 2000 and Illumina HiSeq X) and PEER factors (for sample size N: $150 \leq N < 250$ - 30 factors were used, for $250 \leq N < 350$ - 35 factors, and for $N \geq 350$ - 60 factors). These covariates were downloaded from `https://gtexportal.org/home/datasets`. To adjust for difference in number of reads per sample we included library depth - log number of reads. We also considered shorter model with PEER factors excluded.

Significant discoveries at q-value= 0.01 are provided in Table 3.24

Table 3.24: GTEX version 8 brain analysis results. Full model, including all the covariates used in GTEx data analysis: library depth, PCR/PCR-free flag, platform, sex, 5 principal components and PEER factors and short model excluding PEER factors. We presented the results for TReCASE, MatrixEQTL with p-values corrected using our estimated permutation p-value scheme and MatrixEQTL results with p-values corrected using EigenMT scheme. We applied q-value 0.01 **cutoff to these corrected p-values**

Tissue	N.sam.	N.genes	TReCASE	MatrixEQTL	EigenMT
Full model					
Caudate bg	194	21205	10671	3619	3189
Cerebellar	175	21581	11732	5181	4598
Cortex	205	21137	10527	5599	5024
Frontal Cortex	175	21016	10219	3210	2771
Nucleus abg	202	21395	10483	3720	3281
Short model					
Caudate bg	194	21205	6876	1979	1681
Cerebellar	175	21581	7813	3130	2699
Cortex	205	21137	7146	2920	2563
Frontal Cortex	175	21016	6418	1860	1596
Nucleus abg	202	21395	7025	2109	1824

We can confirm previously noted results: TReCASE is much more powerful both in full and short model compared to simpler total read counts only based analysis. Using EigenMT is conservative in each of these datasets as well. Including PEER factors notably increases power which is consistent with GTEx results.

This analysis can easily be extended to other GTEx tissues.

CHAPTER 4: STUDYING ADDITIVE, SEX AND TREATMENT EFFECTS IN DIVERSE RECOMBINANT INBRED CROSS (RIX) RNA-SEQ DATA

4.1 Introduction

Schizophrenia is a chronic brain disorder that affects about 1% of the population worldwide, and it is associated with substantial loss in life expectancy and personal costs. Haloperidol is the first generation antipsychotic treatment of choice. It is known to have significant side effects often affecting the patient behavior which leads to frequent discontinuation of a treatment after relatively short term use.

Additionally, there is large inter-individual variation in both side-effects and significant heterogeneity in therapeutic response to antipsychotics with variety of literature suggesting a role of genetic variation (Lerer et al. 2005, Lieberman et al. 2005, Patsopoulos et al. 2005, Bakker et al. 2006).

Consequently, a better understanding of haloperidol effects on organism could lead to a safer and more efficient use of existing drugs as well as lead to ideas about future drug development.

Mouse model is appropriate in this study not only because mice are practical replacement of human subjects, but also because it is well known that mice often develop side-effects similar to human side-effects of haloperidol, including variety of motoric disorders such as jaw tremors, tongue protrusions, and vacuous chewing movements (Tomiyama et al. 2001, Turrone et al. 2002, Crowley et al. 2012; 2014). This suggests that haloperidol effects on nervous system of mice and on nervous

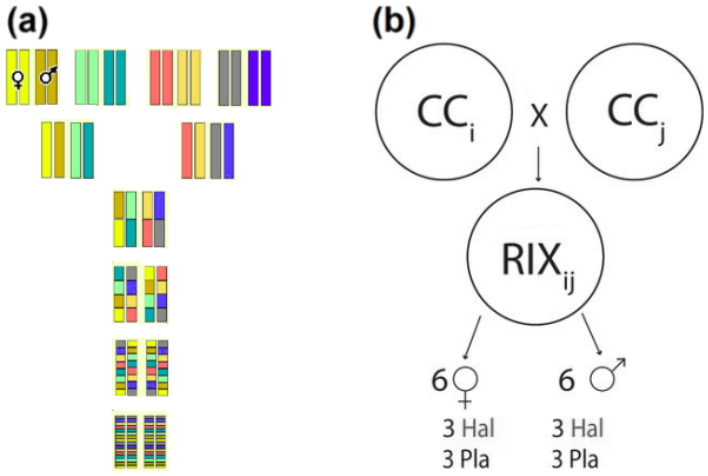

Figure 4.1: Experiment design. (a) Derivation of Recombinant Inbred (RI) strain, (b) RIX cross production and haloperidol exposure.

system of humans are similar in at least some important regards.

A more recent study by Kim et al. (2018) using RNA sequencing of striatal tissue from C57BL6/J mice chronically treated with haloperidol, observed an overlap between the genetic variation underlying the pathophysiology of schizophrenia and the molecular effects of haloperidol confirming a potential of mouse model.

Strong strain-by-treatment interactions were observed for various phenotypes in another recent study by Giusti-Rodríguez et al. (2019) using genetically diverse mice, which points towards the need to evaluate haloperidol effects in diverse populations.

We intend to extend the analysis by studying treatment effects of haloperidol along with additive genetic and sex effects in a diverse population of crosses between recombinant inbred mouse strains. For these purposes in this study we also use use genetically diverse Collaborative Cross (CC) recombinant inbred inter-cross (RIX) mice (Threadgill et al. 2002, Churchill et al. 2004, Consortium et al. 2012).

The CC recombinant inbred (RI) lines were produced as illustrated in panel (a) of Figure 4.1: eight founders ordered randomly, first went through funnel breeding and then went through inbreeding process. Since recombination of eight founders happens

independently, each RI at the end of multiple generations of inbreeding is nearly inbred and genetically diverse. At the same time most of the locations of genome can be traced back to one of the founders which makes an analysis using RIX crosses both diverse and precise.

4.2 Data collection and processing

We used 24 recombinant inbred (RI) strains to generate 22 F_1 hybrids (RIX) as outlined in panel (b) of Figure 4.1. We target to have 6 treated and 6 control mice per cross balanced between male and female mice leading to 3 mice per treatment-sex combination.

Due to variety of issues at the mice production stage we got 232 mice, with number of mice per cross summarized in Table 4.1. Most of the crosses included 12 animals, which is close to the original 6 male and 6 female design; although some crosses had fewer animals (one cross was composed of only 2 male and 2 female animals).

We obtained striatum tissue for these 232 mice, and collected 100bp stranded single-end RNA reads. These reads were mapped using Tophat2 to the appropriate cross pseudogenomes modified from mm10 reference. The data was processed at a lane level with each sample being split into up to 6 lanes. Quality control had shown significant issues with quality for a notable fraction of lanes, particularly in terms of duplication level, fraction of mapped reads and, after summarizing reads at a gene level, fraction of mapped reads among the reads that were mapped to an exon.

After several rounds of quality control we dropped the lanes with duplication level greater than 40%, the fraction of mapped reads less than 75% or the fraction of reads mapped to exons (among mapped reads) less than 65% (for details see Appendix C). Such filtering reduced number of lanes from 1,904 to 1,658, which, after collapsing the data to the sample level, left us with 198 remaining unique samples. We also noted

Table 4.1: Initial number of mice per cross and number of mice after QC Filtering

cross	all samples		filtered	
	Drug	**Placebo**	**Drug**	**Placebo**
13140x3015	6	6	5	5
15156x1566	6	5	4	3
1566x8002	4	6	4	6
16188x3252	6	5	6	5
16211x13140	6	6	5	6
16211x559	5	6	4	4
16211x8043	4	3	4	3
16441x8005	6	6	6	6
3015x15156	6	6	5	5
3154x16211	5	6	3	6
3252x3154	6	6	3	4
5119x13067	6	6	5	6
5489x16188	2	2	1	2
559x8031	5	6	3	5
8005x16188	5	4	5	4
8005x8024	5	5	5	4
8008x8016	6	3	6	2
8016x8034	4	6	3	5
8026x8042	6	5	5	4
8042x8008	6	5	6	5
8042x8043	6	7	6	7
867x3252	6	5	4	3

some mismatches in labeling and used Y chromosome data to fix several sex mismatches. After discarding 4 more mice for which we could not confidently recover cross information and one more mouse that looked as a clear outlier after principal component analysis, performed on the normalized gene expression, we ended up with the final set of 193 mice from 22 crosses available for further analysis. In this analysis the typical cross still had about 4.5 mice per treatment-cross combination with one cross having just 1 treated and 2 placebo mice.

4.3 Total Read Count Model

4.3.1 Modeling autosomes and X chromosome

For a gene of interest on autosomes, for each RIX we can infer founder status to be one of eight original inbred strains. Given F - number of founders for the gene (up to 8), we denote the genotype of each gene from a RIX as A_iA_j where A_i, A_j denote a particular founder allele ($i, j = 1...F$) - with the first allele (A_i) coming from mother and the second allele (A_j) coming from father. For more than a third of the genes F is as high as 8, but for the rest of the genes it is lower: in the dataset of interest we observed F as low as 3 with overall distribution provided in Table 4.2.

For the gene of interest, denote the total number of reads from samples as y_s, with $s = 1, 2, ..., N$. We modeled y_s as Negative-Binomial distribution with mean μ_s and over-dispersion parameter ϕ.

We considered following covariates to be included in Equation 4.1: library depth, treatment, founder information and principal components with corresponding variables encoded as κ_s, sex_s, trt_s, $fnd_{1,s}$, ... $fnd_{F,s}$, $PC_{1,s}$,... $PC_{P,s}$, respectively. We used the first P principal components calculated from the normalized gene expression as covariates.

We assume founder effects to be additive for the purpose of this analysis, so for a

RIX cross with founders A_iA_j $(i \neq j)$ covariates $fnd_{i,s}$ and $fnd_{j,s}$ equal to 1 and for a cross with A_iA_i covariate $fnd_{i,s}$ equals to 2. Consequently, for a given sample we can have one or two non-zero values of fnd_f, s with all of them adding up to 2. To avoid over-parametrization we remove the last founder category and treat it as a reference.

$$y_s \sim f_{NB}(y_s; \mu_s, \phi), \text{ for } s = 1, 2, ..., N, \text{with} \quad (4.1)$$

$$\log(\mu_s) = \beta_0 + \beta_\kappa \times \kappa_s + \beta_{sex} \times sex_s + \beta_{trt} \times trt_s + \sum_{f=1}^{F-1} \beta_f \times fnd_{f,s} + \sum_{k=1}^{P} \beta_{PCk,s} \times PC_{k,s}$$

We test treatment, sex or additive effects with likelihood ratio test $H_0 : \beta_{trt} = 0$, $H_0 : \beta_{sex} = 0$ or $H_0 : \beta_1 = ...\beta_{F-1} = 0$, respectively.

Modeling X chromosome

To model X chromosome we modify the defined above autosomal model by accounting for the fact that in male mice only maternal chromosome is present. Thus female mice are treated similarly to autosomal subsection model and male mice, any cross with founders A_iA_j are treated as A_iA_i setting $fnd_{i,s}$ to 2 and leaving all the other founder variables to be 0.

4.4 Analysis

In this analysis we compared RIX results with those from Kim et al. (2018), because that study analyzed haloperidol effect on C57BL/6J mice using similar tissue (striatum) with a relatively large sample size of 28 mice.

We fitted the model to 13,523 genes with mean expression of at least 20 counts. 80% of the genes had at least 7 founders present and over 98% had at least 6 founders, but in a few cases number of founders could go to as little as 3 founders, as presented in Table 4.2.

Table 4.2: Distribution of number of founders in the analyzed genes

founders	**3**	**4**	**5**	**6**	**7**	**8**
genes	1	24	210	2439	6157	4692

Table 4.3: Number of significant results by number of PCs

PC	**Additive**	**Sex**	**Treatment**
5	10578	230	104
10	10661	330	196
15	9822	378	340
20	9075	531	372
25	8258	459	467
27	7542	297	514
30	7029	284	427

For the main analysis using all the mice jointly and incorporating the founder information to capture additive genetic effect along with sex and treatment effects, we considered using different number of principal components. We observed that using fewer principal components to fit RIX dataset, lead to smaller number of significant treatment or sex discoveries (at a q-value cutoff 0.10). We also observed that the number of discoveries of *cis*-acting eQTLs in this analysis is consistent with previous studies such as Crowley et al. (2015). Increasing the number of principal components lead to higher number of significant treatment and sex effects at a cost of reducing number of significant additive genetic effects as can be seen in Table 4.3. To finalize the number of principal components in our analysis we applied several methods suggested in the literature on selection of number of principal components (summarized in Appendix C Table 7). Using the consensus results of more numerically stable Kaiser, Parallel analysis and Optimal Coordinates methods, we ended up selecting 27 principal components for autosomal analysis and 20 for X chromosome analysis as shown in Figure 4.2.

To evaluate our method performance with various number of principal components included in the model, we checked for the overlap with Kim et al. (2018) dataset as a

comparison. We observe that using 27 principal components (a consensus choice of number of principal components) produces the best overlap as shown in Table 4.4.

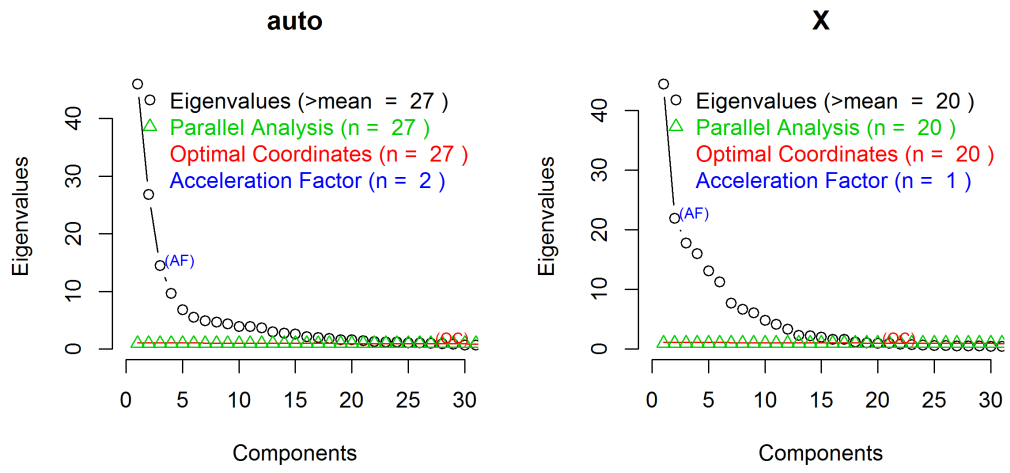

Figure 4.2: Final subset of methods for PC selection: Kaiser method - selecting PC's with eigenvalues bigger than 1, Parallel analysis - a sample based adaptation of the population based Kaiser rule and Optimal Coordinates - an extrapolation of the preceding eigenvalue by a regression line between the eigenvalue coordinates and the last eigenvalue coordinates.

The overlap of the genes found to be significant in RIX dataset and in Kim et al. (2018) dataset is higher than one would expect to observe by chance. To assess robustness of RIX results we included the results with lower number of principal components included to the model as presented in Table 4.4. We observed that using different number of PCs we still observe higher overlap and we also can see that 27 principal components provide the highest number of discoveries as well as the best overlap.

Overall, in Kim et al. (2018) dataset 78 (36 down/42 up-regulated or 54% of genes being up-regulated) genes were found to be significant at q-value cutoff 0.05 and 1,510 (729 down/781 up-regulated or 52% of genes being up-regulated) at q-value cutoff 0.10. While imbalance between up-regulated and down-regulated genes in Kim et al.

Table 4.4: Between dataset overlap for expressed genes. RIX dataset fitted with 15, 20 and 27 principal components compared to Kim et al. (2018) results. At several q-value cutoffs we counted number of genes having significant treatment effect in reference Kim et al. (2018) dataset (Ref), number of genes having significant treatment effect in RIX dataset, number of genes declared significant by both methods, and provided an estimate of excess number of significant genes under assumption that both lists produced randomly. The between method overlap is calculated using 12,589 genes that were expressed in both datasets. Note, this excludes several genes that were found to be significant in RIX dataset, but were not tested in Kim et al. (2018) dataset and several genes that were found significant in Kim et al. (2018) dataset, but were not tested in RIX dataset.

		15 PC			20 PC			27 PC		
q-val	**Ref.**	**RIX**	**Both**	**Exc.**	**RIX**	**Both**	**Exc.**	**RIX**	**Both**	**Exc.**
0.05	56	326	15	13.5	361	15	13.4	492	17	14.8
0.1	1264	541	129	74.7	578	123	65.0	774	164	86.3
0.15	3947	748	301	66.5	780	308	63.4	1013	395	77.4
0.2	5734	977	477	32.0	1028	504	35.8	1311	634	36.9
0.25	7116	1252	748	40.3	1349	803	40.5	1608	928	19.1

(2018) dataset is not statistically significant, we noticed that RIX dataset has the same direction of imbalance - more up-regulated than down-regulated genes, and in RIX dataset this imbalance is statistically significant irrespectively whether we select significant genes at 0.05 and 0.10 q-value cutoffs with various number of principal components in the model. For the model with 27 principal components, we got 198 down versus 316 up-regulated genes at 0.05 q-value cutoff, which corresponds to 61% of significant genes being up-regulated. Applying two-sided binomial test we get a highly significant p-value $2e-7$, rejecting a hypothesis that percentage of up-regulated genes is 50%. At 0.10 q-value cutoff we observed 342 down to 473 up-regulated genes or 58% of significant genes being up-regulated and corresponding p-value $5e-6$.

This result holds even if we include as little as 15 PCs: we still observed more up-regulated than down-regulated genes - 309 versus 260 genes which constitutes 54% of significant genes being up-regulated. This percentage of up-regulated genes is still

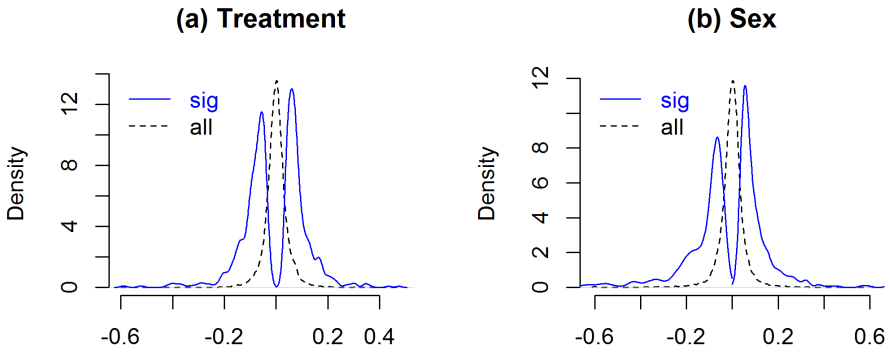

Figure 4.3: Effect size of the genes found to be significant (a) treatment and (b) sex effects. (a) Treatment effect - comparing haloperidol versus placebo, (b) Sex effect - male expression vs female expression

significantly different from 50% (two sided binomial test yields p-value 0.044). At a stronger q-value cutoff 0.05 we observed 56% of significant genes being up-regulated which again is significantly different from 50% (p-value 0.023). Thus, we conclude that RIX dataset strongly indicates that haloperidol tends to up-regulate genes.

Using only the genes that were declared significant in both datasets (at q-value cutoff 0.10) we considered a final cross-check by looking at the direction of the haloperidol effect. Figure 4.4 shows that both in autosomes and in X chromosome RIX and Kim et al. (2018) datasets are generally in agreement regarding the direction of the haloperidol effect.

4.5 Pathway categories discussion

We performed a pathway analysis of genes targeting categories from GO database. We separately considered the genes that were either down-regulated or up-regulated upon haloperidol treatment using GOrilla (http://cbl-gorilla.cs.technion.ac.il); for genes with a q-value < 0.10 and q-value < 0.05). Using GOrilla we were able to

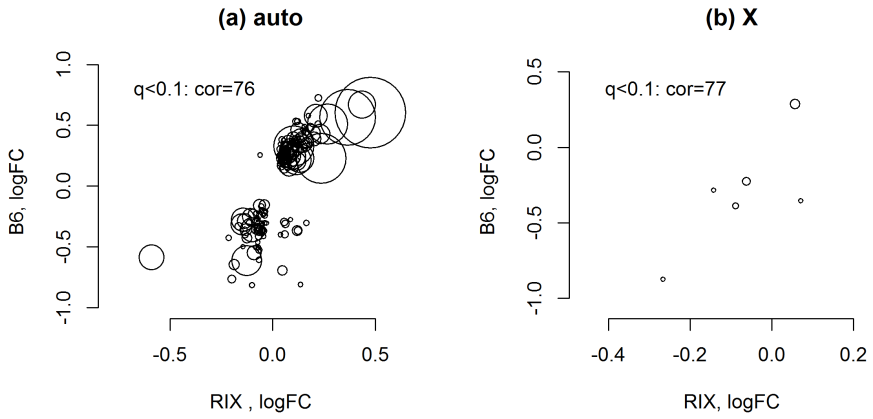

Figure 4.4: Between dataset direction consistency (a) autosomes, (b) X chromosome using genes with haloperidol effect significant at q-value cutoff 0.10 **in both datasets. Figure is based on log fold change estimates for 129 autosomal and 3 X chromosome genes with circle size proportional to corresponding** $-log10(q-value)$ **in RIX dataset.**

examine processes, functions, and components associated with the up or down-regulated genes. In this analysis we see mostly up-regulated categories: none of down-regulated categories was found to be significant at FDR 0.05 level with couple significant at FDR 0.10 level. At the same time for up-regulated categories we observed 43 component categories (Table 4.5) and 10 process categories (Table 4.6). No GO Function terms passed FDR correction for either up or down-regulated genes.

For up-regulated genes (q-value < 0.05), the top terms for GO Component included synapse part, neuron part, plasma membrane, vesicles, neuronal cell body, and neuron projection, suggesting that haloperidol may be altering neuronal morphology and/or density. The top terms associated with GO Process are related to cell secretion, transmembrane transporter activity, ion transport, and synaptic plasticity, which suggests that haloperidol is altering synaptic plasticity and cell signaling, via alterations in channel expression, localization, or modulation.

We observed similar pattern with up-regulated genes creating more significant

Table 4.5: Significant GO Component categories (all up-regulated)

GO Term	Description	FDR	Enr.	Genes
GO:0044456	synapse part	1.62e-9	2.27	82
GO:0097458	neuron part	2.93e-9	1.89	115
GO:0005886	plasma membr.	5.42e-8	1.58	161
GO:0016020	membr.	1.70e-6	1.31	261
GO:0031410	cytoplasmic vesicle	3.82e-5	1.76	86
GO:0097708	intracellular vesicle	3.52e-5	1.75	86
GO:0099503	secretory vesicle	3.85e-5	2.74	33
GO:0043025	neuronal cell body	3.85e-5	2.32	44
GO:0043005	neuron projection	4.17e-5	1.85	72
GO:0044433	cytoplasmic vesicle part	5.84e-5	2.41	39
GO:0031982	vesicle	5.39e-5	1.7	88
GO:0120025	plasma membr. bounded cell proj.	9.58e-5	1.66	91
GO:0045202	synapse	9.18e-5	1.85	66
GO:0042995	cell projection	1.00e-4	1.59	102
GO:0044297	cell body	1.88e-4	2.1	46
GO:0120038	plasma membr. bounded cell proj.p	8.12e-4	1.64	77
GO:0044463	cell projection part	7.64e-4	1.64	77
GO:0030133	transport vesicle	7.80e-4	2.81	23
GO:0070382	exocytic vesicle	7.45e-4	3.08	20
GO:0044425	membr. part	1.37e-3	1.28	193
GO:0030425	dendrite	1.47e-3	2.12	36
GO:0008021	synaptic vesicle	1.57e-3	3.12	18
GO:0044459	plasma membr. part	1.79e-3	1.5	95
GO:0098793	presynapse	1.86e-3	2.62	23
GO:0098590	plasma membr. region	3.10e-3	1.75	53
GO:0097060	synaptic membr.	3.49e-3	2.26	28
GO:1990761	growth cone lamellipodium	3.65e-3	26.66	3
GO:0098563	intr. comp. of syn. vesicle membr.	5.43e-3	4.37	10
GO:0030141	secretory granule	8.77e-3	2.68	18
GO:0042734	presynaptic membr.	1.00e-2	4.04	10
GO:0033267	axon part	1.18e-2	2.02	30
GO:0030054	cell junction	1.20e-2	1.62	56
GO:0030659	cytoplasmic vesicle membr.	1.23e-2	2.77	16
GO:0012506	vesicle membr.	1.72e-2	2.59	17
GO:0005938	cell cortex	2.05e-2	2.74	15
GO:0044306	neuron projection terminus	2.90e-2	2.89	13
GO:0005576	extracellular region	3.07e-2	1.65	45
GO:0030139	endocytic vesicle	3.22e-2	2.84	13
GO:0098831	presynaptic active zone cyto. comp.	3.40e-2	6.66	5
GO:0005905	clathrin-coated pit	3.95e-2	3.95	8
GO:0098805	whole membr.	4.06e-2	1.77	34
GO:0030285	int. comp. of syn. vesicle membr.	4.46e-2	4.34	7

Table 4.6: Significant GO Process categories (all up-regulated)

GO Term	Description	FDR	Enr.	Genes
GO:0032879	reg. of localization	2.47e-3	1.55	123
GO:0065008	reg. of biological quality	4.19e-3	1.46	140
GO:0022898	reg. of transmembr. transp. act.	3.66e-2	2.89	22
GO:0003008	system process	3.06e-2	1.82	56
GO:1903530	reg. of secr. by cell	2.52e-02	2.02	43
GO:0051049	reg. of transp.	2.40e-2	1.59	84
GO:0032409	reg. of transp. act.	2.91e-2	2.78	22
GO:0032412	reg. of ion transmembr. transp. act.	2.63e-2	2.86	21
GO:0051046	reg. of secretion	2.54e-2	1.95	44
GO:0043269	reg. of ion transport	2.70e-2	2.02	40

pathways than down-regulated genes with DAVID (https://david.ncifcrf.gov) Functional Annotation Analysis. Though DAVID was less powerful, it also provided an additional useful feature of clustering categories together. We observed that the most significant of these cluster categories were often haloperidol related (such as Sprouty2) and, even for the clusters with less significant enrichment, we observed very consistent results: if a cluster was enriched with up-regulated genes, it was either completely absent or severely depleted when the similar analysis was applied to down-regulated genes and if a cluster was enriched with down-regulated genes, it would be depleted with up-regulated genes (for details see Appendix C Table 8). We consider such consistent enrichment to be an additional way to confirm consistency of the genes discovered in this study, including the categories that weren't found to be significant at an individual level in DAVID analysis.

4.5.1 Locations of discovered effects

Illustration of the locations of the genes with significant (at q-value 0.10) effects doesn't show obvious spatial patterns (Figure 4.5). The genes with the most significant additive effect include Rpl18-ps1 on chromosome 1, Atp6v0c on chromosome 17 and Scg5 on chromosome 2. For sex effect top genes are Akr1e1 on

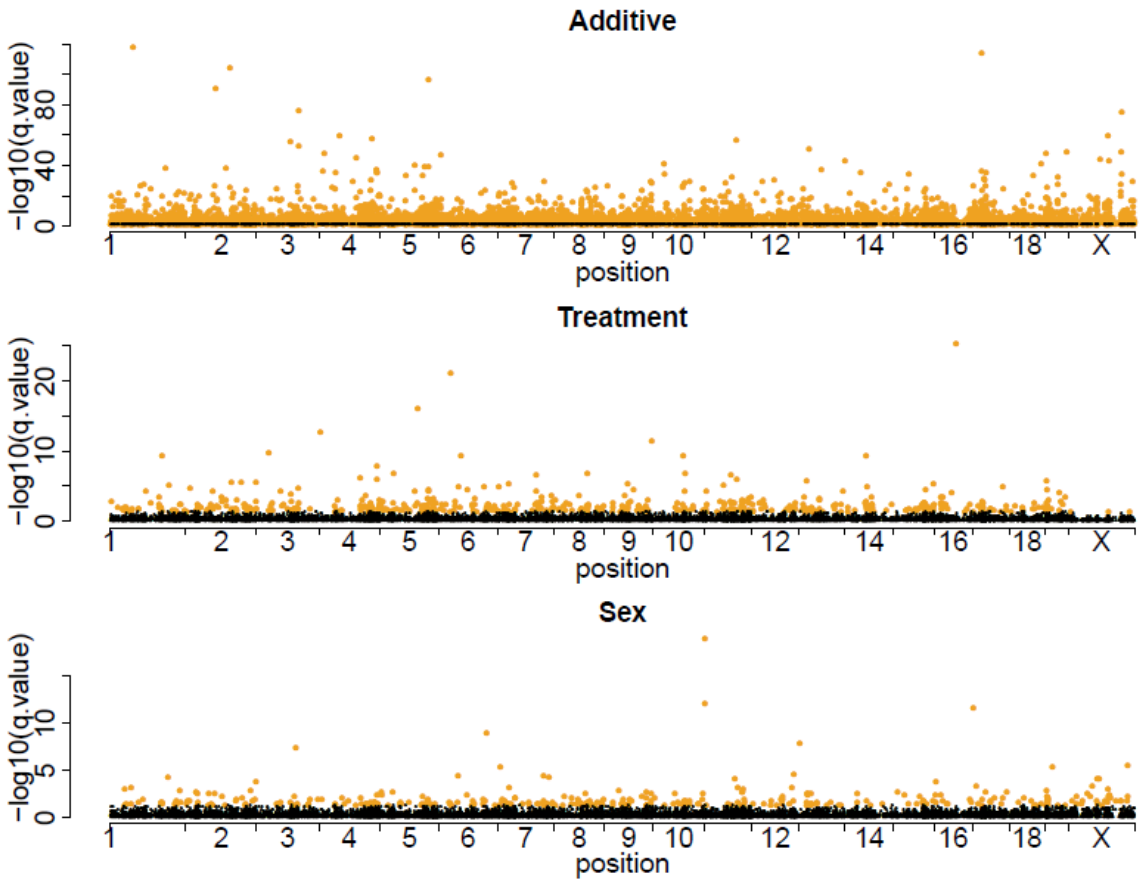

Figure 4.5: Positions of discovered effects. Golden dots represent genes with q-values significant at the level of 0.01, while black represent non-significant genes.

chromosome 13, Ptgfrn on chromosome 3, A2m on chromosome 6 and Pisd-ps1 on chromosome 11. The two genes with the most significant haloperidol effect are Strip2 on chromosome 6 and Tomm70 on chromosome 16

4.6 Conclusions

In conclusion we show that while using RIX is a more complicated task, it is also quite productive: we were able to confirm some of the previous results: Kim et al. (2018) in terms of haloperidol effect and Crowley et al. (2015) in terms of additive genetic effect. We also got more power and were able to test for imbalance between

haloperidol up-regulated and down-regulated genes, for which Kim et al. (2018) dataset had shown direction concordant to RIX, but was not statistically significant due to lower power to discover genes significantly affected by treatment. In addition, this analysis identified significant haloperidol effect for 522 genes at q-value 0.05 (827 at q-value 0.10) unobserved in previous Kim et al. (2018) dataset.

Attempt to analyze haloperidol effect for each cross separately proved to be extremely unstable to outliers and didn't replicate previous results.

RIX dataset also produced more consistent results in pathway analysis showing more significant results in up-regulated pathway categories. Additional look at the top terms associated with GO Process suggests that haloperidol is altering synaptic plasticity and cell signaling, via alterations in channel expression, localization, or modulation.

APPENDIX A: TECHNICAL DETAILS FOR CHAPTER 2

A.1 eQTL consistency with respect to GEUVADIS dataset

Overall consistency of 30 trios additive effects vs E-GEUV-1 dataset (Figure 9) which is likely to be due to the problems with SNP genotypes for those samples.

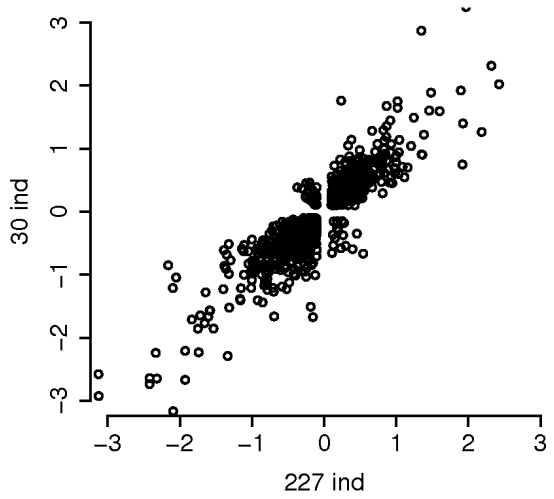

Figure 6: Consistency of additive effects in 30 trios vs 227 samples from E-GEUV-1 dataset

Note, we do observe that either using smaller dataset or larger dataset (for example, in our two data sets with very different sample sizes > 200 vs. 32), we observe similar distribution of distances from a SNP to transcription start of a gene (Figure 7).

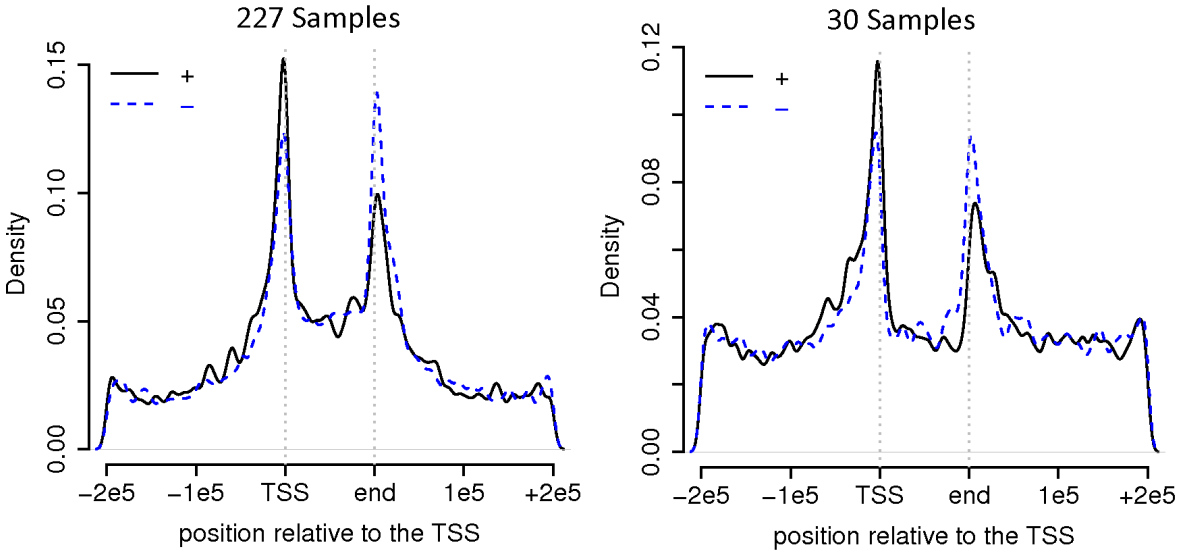

Figure 7: Distance to transcription start site: left - 227 samples from E-GEUV-1 dataset, right - 30 children from trios dataset. Genes on positive or negative strands were plotted separately

A.2 Additional simulations

A.2.1 Checking whether proposed algorithm converges to the proper maximum

One of the concerns was whether our algorithm converges to the proper place. We considered two additional checks whether algorithm converges to the maximum. To do this we took one of the profiles of original simulations (with additive genetic effect and parent of origin effects values 1) and added extra layer of refitting:

First, since in our algorithm we fit 8 parameters, grid search around local maximum would be either too consuming or too imprecise; we considered a following alternative simulation:

1. Perform initial run according to our initial scheme. This run gives us a good impression of variability of the parameters for a given sample size and we

compare these maximum points to the refits produced in the next steps.

2. Estimate standard deviation for each of the parameters typical for such sample size (Table 7). This would give us a reasonable distribution to sample initial values around local maximum.

3. Randomly sample 1000 points around local optimum using MLE from the first step and standard deviations from the second step from normal distribution and refit the model from those starting points. If they lead to a different location also check likelihood if it is comparable. With this step we avoid running the 8-dimensional grid around local optimum and can see whether we see any signs of optimizing at the wrong place.

For 10,000 datasets 1000 re-sampling for each all the samples converged in the vicinity of original fit (Table 8). We also don't see potential source for bias - on average deviation of parameter estimates from refitted iterations are located very close to the fits from original scheme.

Table 7: Standard deviations for all 8 parameters in an initial simulation used to select initial values

	ϕ	φ	b_0	b_1	γ_0	β_1	β_2	β_κ
sd	0.26	0.47	0.31	0.31	0.34	0.738	0.33	0.17

Second approach to evaluate how likely we were to end up in some local maximum was to choose initial values in a less informed fashion than initially suggested and to see whether we would get a different and better result. For each of 10,000 simulations we fitted 100 more models with different initial values selected from a multivariate normal distribution around mean **0** and with standard deviations (Table 7) produced to mimic observed standard deviations for each parameter for such sample size multiplied by 3.

Table 8: Differences of refitted likelihoods, b_0's and b_1's from initial likelihood fit, b_0 and b_1

deviation	effect size	min	1%	50%	99%	max
log-lik	-	-1.9e-02	-5.1e-04	-6.1e-11	3.0e-04	1.5e-02
b_0	1.0	-4.7e-02	-6.1e-03	1.4e-08	5.2e-03	3.5e-02
b_1	1.0	-9.7e-03	-7.7e-04	1.5e-08	9.8e-04	7.8e-03
$\log(\phi)$	-0.3	-1.8e-02	-2.1e-03	-3.9e-09	1.5-03	1.8e-02
$\log(\varphi)$	-0.3	-3.1e-02	-3.3e-03	-1.2e-07	2.4e-03	3.5e-02
γ_0	3.5	-3.5e-02	-5.0e-03	-7.2e-09	5.9e-03	4.7e-02
β_1	2.1	-4.7e-02	-2.4e-03	1.5e-11	2.4e-03	3.7e-02
β_2	0.05	-1.8e-02	-1.1e-03	3.4e-12	1.1e-03	1.5e-02
β_κ	0.5	-6.9e-03	-6.8e-04	-7.6e-10	6.4e-04	1.3e-02

As result of those 100 reruns for 10,000 simulated datasets we've observed 353 cases when one of 100 runs deviated from the consensus location (our initial scheme always landed at the same location as consensus). Each of deviating runs also had inferior likelihood which suggests that poorly chosen initials may reduce stability of algorithm convergence. However, with reasonably selected initial values we didn't observer any cases when algorithm didn't converge to inferior location.

A.3 Additional real data analysis results

A.3.1 Effect locations

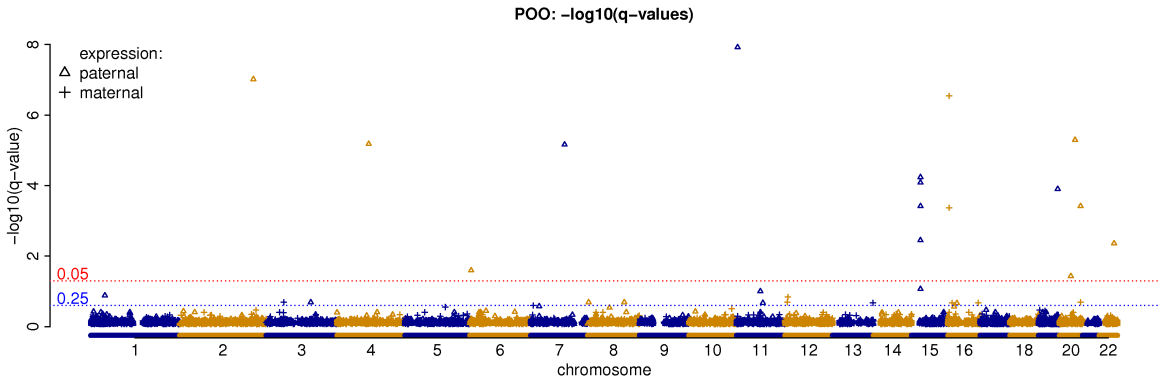

Figure 8: Positions of discovered parent-of-origin effects.

APPENDIX B: TECHNICAL DETAILS FOR CHAPTER 3

B.1 Additional data preparation details

Creating the lists of heterozygous SNPs for each sample

We obtained the list of heterozygous SNPs for each sample from the results of imputation. The main output file (other files have additional extensions to the name as `_haps`, `_allele_probs` etc) has 3 columns with genotype probabilities per SNP, which represent the probability of observing genotype $G = 0$, 1, or 2 respectively. We selected heterozygous locations (i.e. with high probability of $G = 1$) to output phased genotypes of these locations.

Removing abnormal samples in 1000 Genomes dataset

Once we got a list of heterozygous SNPs, we extracted allele-specific reads for two haplotypes using R function `extractASReads` from R package `asSeq` (Sun 2012). After processing E-GEUV-1 dataset we observed that some samples had abnormal number of genes having majority of reads classified as one or the other haplotype (Figure 9) which is likely to be due to the problems with SNP genotypes for those samples.

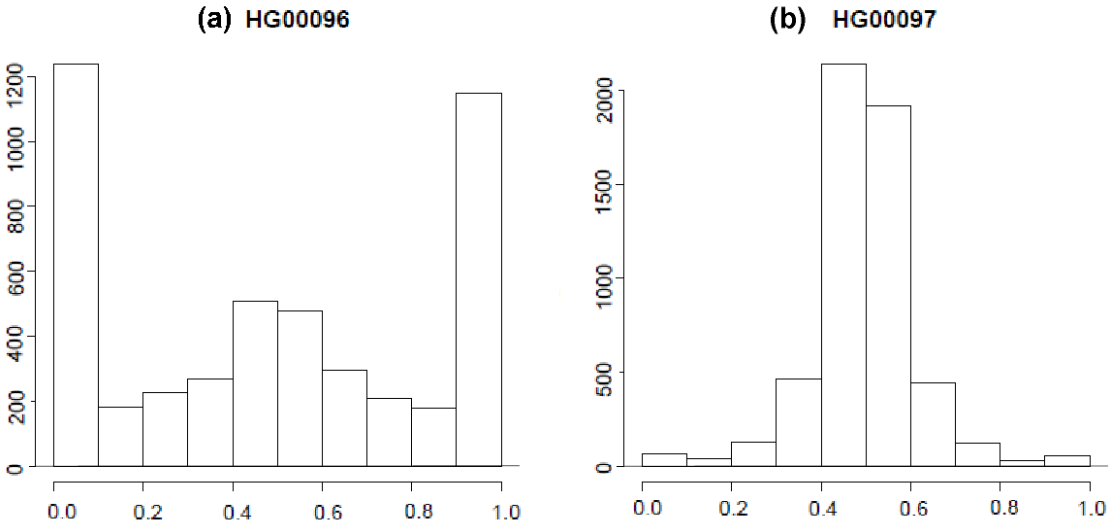

Figure 9: Illustration of a bad vs a good sample: (a) a sample with too many genes ending up having majority of reads from one of haplotypes (b) a sample with reasonable distribution of gene level allele specific counts

B.1.1 RASUQAL inflation in GTEx data

After permutation of whole blood in GTEx data we see inflation of a similar style to what we've observed in 1000 Genome dataset dependent on number of fSNPs as can be seen in Figure 10

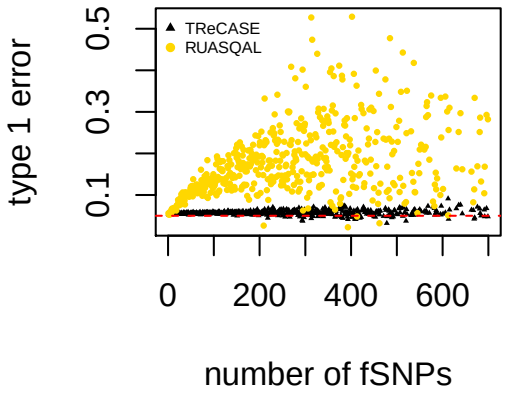

Figure 10: RASQUAL inflation for permuted GTEx dataset. Figure illustrates both generally higher number of fSNPs and inflation of RASQUAL depending on number of fSNPs

B.1.2 Additional cross-method comparisons

Comparing the methods vs full sample TReCASE results we classified genes to be significant at q-value 0.01 and compare both fraction of recovered results and fraction of false positives (Table 9)

Table 9: Comparing to TReCASE results

	Power			FDR		
N.samp	**MatrixEQTL**	**TReC**	**TReCASE**	**MatrixEQTL**	**TReC**	**TReCASE**
35	0	0.02	0.05	-	0.26	0.19
70	0.03	0.06	0.14	0.14	0.09	0.08
140	0.19	0.24	0.45	0.06	0.05	0.06
280	0.58	0.67	1	0.05	0	0

Using TReC fit of 280 samples as gold standard (Table 10)

Table 10: Comparing to TReC results

	Power			FDR		
N.samp	**MatrixEQTL**	**TReC**	**TReCASE**	**MatrixEQTL**	**TReC**	**TReCASE**
35	0	0.03	0.07	-	0.36	0.28
70	0.04	0.08	0.18	0.15	0.11	0.17
140	0.27	0.34	0.54	0.08	0.08	0.23
280	0.81	1	1	0.09	0	0.33

Using MatrixEQTL fit of 280 samples as gold standard (Table 11)

Table 11: Comparing to TReC results

	Power			FDR		
N.samp	**MatrixEQTL**	**TReC**	**TReCASE**	**MatrixEQTL**	**TReC**	**TReCASE**
35	0	0.03	0.07	-	0.37	0.28
70	0.05	0.09	0.2	0.09	0.13	0.2
140	0.31	0.37	0.57	0.07	0.12	0.28
280	1	0.91	0.95	0	0.19	0.42

APPENDIX C: TECHNICAL DETAILS FOR CHAPTER 4

C.1 Additional information on quality control and filtering

C.1.1 Initial QC filtering

We only considered samples that passed 3 cutoffs: filtering by duplication (at most 40% duplication), percentage of mapped reads (at most 25% reads not mapping) and percentage of mapped reads being mapped to a gene (at most 35% not being mapped to a gene) 11.

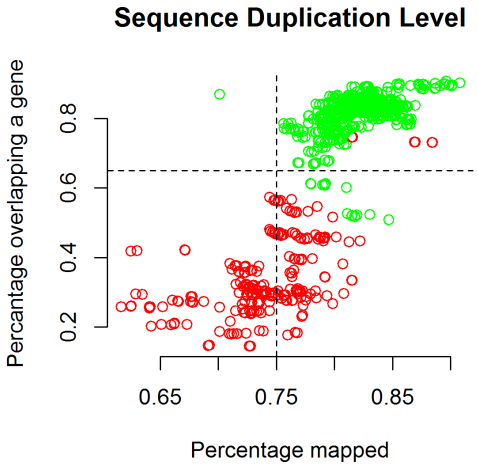

Figure 11: Quality control filters: top-right corner with green color was deemed adequate to proceed

These criteria lead to filtering out 34 samples (CEGS001, CEGS002, CEGS003, CEGS004, CEGS005, CEGS006, CEGS007, CEGS008, CEGS009, CEGS010, CEGS011, CEGS012, CEGS013, CEGS014, CEGS015, CEGS016, CEGS017, CEGS018, CEGS021, CEGS022, CEGS023, CEGS026, CEGS028, CEGS031, CEGS034, CEGS035, CEGS038, CEGS040, CEGS055, CEGS067, CEGS070, CEGS071, CEGS235, CEGS280).

C.1.2 Extra filtering

At the cross recovery step we filtered out four more samples CEGS072, CEGS178, CEGS274 and CEGS253.

Finally we removed sample CEGS057 as an obvious outlier which can be seen in Figure 12. This was a sample for which lane level quality control results were on the boundary - several of the lanes were removed at in the first QC step and the lanes that remained were close to the cutoffs we defined as exclusion criteria in the first QC step. This gives another confirmation that 34 samples removed at first QC step were unreliable.

C.1.3 Principal Component Analysis

Principal Component selection is a procedure known to have many potential issues. As a part of the analysis we performed multiple methods of selection in order to get an idea about the range of the number of principal components that would be achieved by multiple methods. We found nFactors (Raiche 2010) package to be very useful in this respect as it combines quite a few of the methods known in the literature, including Bartlett Test (Bartlett 1950), Lawley Test (Lawley 1956), Anderson Test (Anderson 1963), Kaiser rule (Kaiser 1960), the Parallel Analysis (PA) (Horn 1965), and the Scree test (Cattell 1966) as well as Gorsuch scree test (Gorsuch and Nelson 1981) and Bentler Test (Bentler and Yuan 1998). Finally, this package also adds two more measures - acceleration factor (AC) - a numeric solution to the elbow of a scree plot and optimal coordinates (OC) giving an extrapolation of the preceding eigenvalue by a regression line between the eigenvalue coordinates and the last eigenvalue coordinates.

We found that these methods can produce extremely different results with more parsimonious methods being acceleration factor (2 principal components) and

Table 12: Summary of PC methods. Number of principal components selected by each of the methods.

chr.	AC	Gorsuch	Bentler	Bartlett	Lawley	Anderson	Kaiser	PA	OC
auto	2	3	50	12-21	12-22	18-24	27	27	27
X	1	3	188	18-19	18-19	20-33	20	20	20

Gorsuch scree test (3 principal components). Bentler Test was on the other side of the range suggesting using 50 principal components.

Some of the classic methods such as Bartlett, Lawley and Anderson tests tended to be sensitive to number of eigen-values supplied to the procedure producing suggested numbers of principal components in 12-21 range for Bartlett test, 12-22 for Lawley test and 18-24 for Anderson test.

Finally, Kaiser rule, Parallel analysis and Optimal Coordinates tended to agree the most suggesting using 27 principal components for autosomes.

We observed similar pattern when we checked X-chromosome. Overall results can be summarized in a Table 12

C.1.4 Additional Pathway Analysis

We have looked at the top pathway clusters produced by DAVID Analysis applied to the gene lists from discussed datasets. All of the pathway clusters in RIX dataset were consistently either enriched among up-regulated genes or enriched among down-regulated genes as presented in Table 13: we didn't observe among presented categories any to be enriched both among up and down-regulated at the same time and we generally saw higher enrichment for pointed (only up-regulated or only down-regulated) subset of the genes then in overall list of genes. For the reference we applied the same procedure to 729 down and 781 up-regulated genes from Kim et al. (2018). In RIX dataset we saw more categories enriched among up-regulated genes 18 versus 8 among down-regulated - which is concordant with overall results, other

Table 13: Up or down-regulated pathway clusters in RIX and comparable groups in Kim et al. (2018) dataset. Searching for pathways among down-regulated genes, presented first, and up-regulated genes presented second. For reference we always provide overall enrichment score and in parenthesis up-regulated/down-regulated scores. If in Kim et al. (2018) we observed similar pathway we provide it in the Ref column. Dash represents that pathway cluster was not present at all.

Cluster	RIX	Ref.
Sprouty/SRA	-/2.13	-/2.15
EVH1/WH1	-/1.54	1.62./2.15
Rotamase/isomerase act	-/1.45	-/-
Neuropeptide	-/1.43	-/-
Microtubule	0.04/1.19	0.42/0.76
RGS	-/1.18	-/-
protein folding	-/1.17	-/0.35
Tubulin	-/1.17	-/-
Secreted/Glycoprotein	3.9/0.47	1.75/2.22
Calmodulin-binding	2.51/-	0.37/-
Membrane	2.36/0.15	0.02/2.22
Synapse	2.05/0.3	0.34/5.8
Sodium:neurotransmitter symporter	1.94/-	-/-
VWFC	1.82/-	-/-
growth cone; dendritic spine	1.82/-	-/-
Dilated cardiomyopathy	1.53/-	0.42/1.16
Exocytosis	1.43/-	-/-
Ras-GEF	1.39/-	0.32/-
cGMP	1.29/-	-/-
TSP1	1.26/-	-/0.08
protein kinase	1.19/-	0.15/1.65
ECM-receptor interaction 6	1.12/-	-/-
Prenylation	1.1/-	-/-
coronary vasculature development	1.07/-	-/-

Table 14: Significant categories in RIX using DAVID

category	description	FDR	Enrich.	Genes
Glycoprotein	.	0.0025	1.59	83
Signal	.	0.001	1.60	78
glycosylation site	N-linked (GlcNAc...)	0.020	1.58	104
Disulfide bond	.	0.033	1.55	54
signal peptide	.	0.061	1.59	62
Secreted	.	0.031	1.90	32
GO:0005576	extracellular region	0.060	1.84	34
Calmodulin-binding	.	0.030	3.84	11
Membrane	.	0.0037	1.29	150
GO:0016020	membrane	0.067	1.21	152
GO:0045202	synapse	0.030	2.27	25
GO:1990761	growth cone lamellipodium	0.091	40.40	3

pathway analysis results and Kim et al. (2018) results. Top significant categories from this analysis are presented in Table 14. The only individually significant categories were found among up-regulated genes.

Several enriched clusters were found in both RIX and Kim et al. (2018) dataset and matched the direction (such as Sprouty and EVH1), however some of them didn't match (Membrane, Synapse, VWFC, Dilated cardiomyopathy, protein kinase), didn't show consistent enrichment (Glycoprotein) or were absent (Rotamase, Neuropeptide, RGS, protein folding, Tubulin, dendritic spine, TSP1, Calmodulin-binding, Microtubule). Also there are some categories that were found in Kim et al. (2018) dataset that are not present in our analysis (particularly Zink-finger, Potassium channel, C2, circadium rhythm, HECT, secretion, ubiquitin protein, BTB, GABAergic synapse, Pleckstrin homology).

Literature search shows a lot of connections to schizophrenia related research for categories found by RIX dataset. For example "Decreased expression of Sprouty2 in the dorsolateral prefrontal cortex in schizophrenia and bipolar disorder: a correlation with BDNF expression" by Pillai (2008) and "Pleckstrin homology domain containing 6 protein (PLEKHA6) polymorphisms are associated with psychopathology and

response to treatment in schizophrenic patients" by Spellmann et al. (2014).

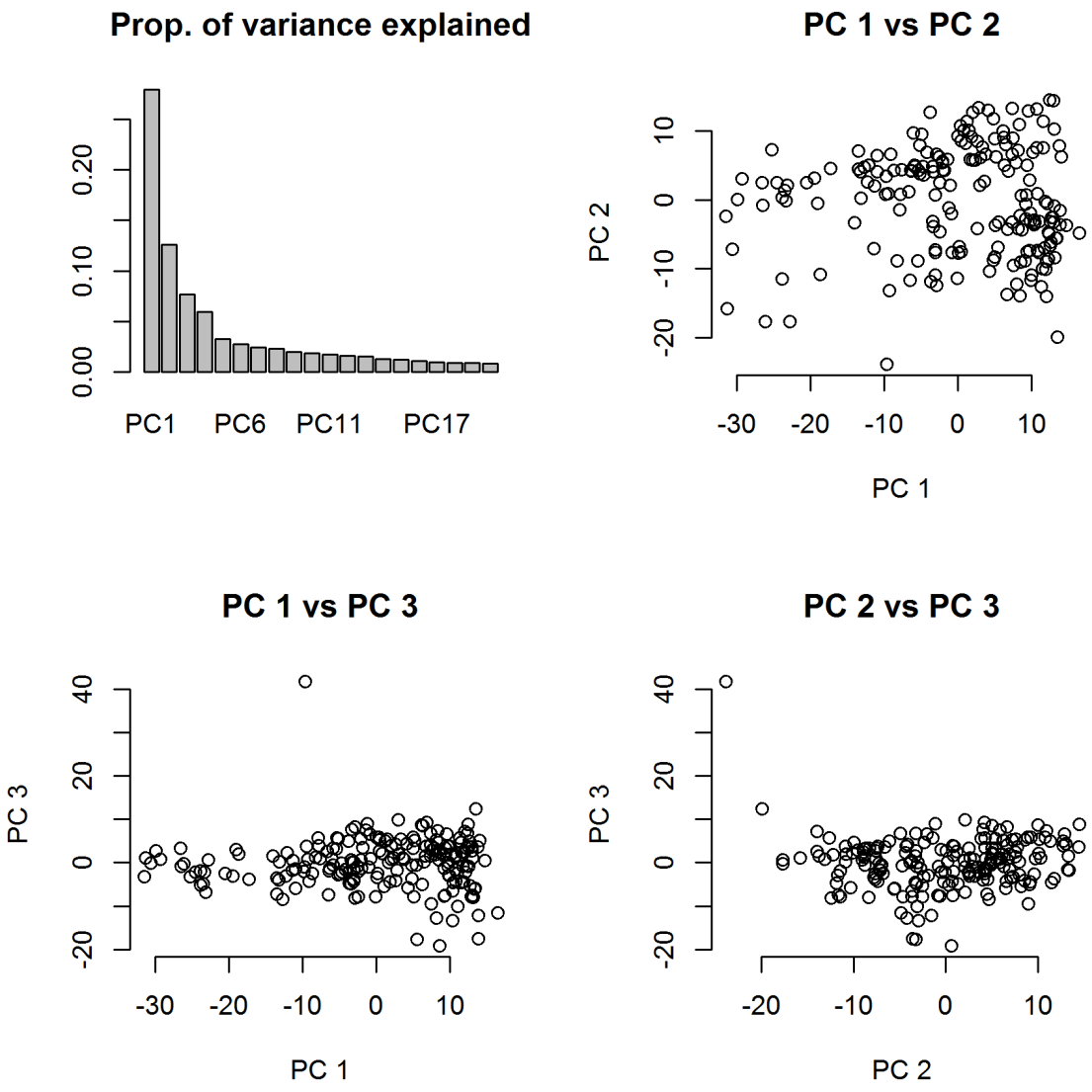

Figure 12: PC outlier: final sample to be removed

BIBLIOGRAPHY

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

1000 Genomes Project Consortium (2015), "A global reference for human genetic variation," *Nature*, 526, 68.

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Parsana, P. E., Flynn, E., Fresard, L., Gaamzon, E. R., Hamel, A. R., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., Park, Y., Saha, A., Segrć, A. V., Strober, B. J., Wen, X., Wucher, V., Das, S., Garrido-Martín, D., Gay, N. R., Handsaker, R. E., Hoffman, P. J., Kashin, S., Kwong, A., Li, X., MacArthur, D., Rouhana, J. M., Stephens, M., Todres, E., Viñuela, A., Wang, G., Zou, Y., , Brown, C. D., Cox, N., Dermitzakis, E., Engelhardt, B. E., Getz, G., Guigo, R., Montgomery, S. B., Stranger, B. E., Im, H. K., Battle, A., Ardlie, K. G., and Lappalainen, T. (2019), "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *bioRxiv*.

Anderson, T. W. (1963), "Asymptotic theory for principal component analysis," *The Annals of Mathematical Statistics*, 34, 122–148.

Bakker, P. R., van Harten, P. N., and van Os, J. (2006), "Antipsychotic-induced tardive dyskinesia and the Ser9Gly polymorphism in the DRD3 gene: a meta analysis," *Schizophrenia research*, 83, 185–192.

Barbaux, S., Gascoin-Lachambre, G., Buffat, C., Monnier, P., Mondon, F., Tonanny, M.-B., Pinard, A., Auer, J., Bessières, B., Barlier, A., et al. (2012), "A genome-wide approach reveals novel imprinted genes expressed in the human placenta," *Epigenetics*, 7, 1079–1090.

Bartlett, M. S. (1950), "Tests of significance in factor analysis," *British Journal of statistical psychology*, 3, 77–85.

Bentler, P. M. and Yuan, K.-H. (1998), "Tests for linear trend in the smallest eigenvalues of the correlation matrix," *Psychometrika*, 63, 131–144.

Cattell, R. B. (1966), "The scree test for the number of factors," *Multivariate behavioral research*, 1, 245–276.

Churchill, G. A., Airey, D. C., Allayee, H., Angel, J. M., Attie, A. D., Beatty, J., Beavis, W. D., Belknap, J. K., Bennett, B., Berrettini, W., et al. (2004), "The Collaborative Cross, a community resource for the genetic analysis of complex traits," *Nature genetics*, 36, 1133.

Consortium, C. C. et al. (2012), "The genome architecture of the Collaborative Cross mouse genetic reference population," *Genetics*, 190, 389–401.

Consortium, G. et al. (2017), "Genetic effects on gene expression across human tissues," *Nature*, 550, 204.

Crowley, J. J., Kim, Y., Lenarcic, A. B., Quackenbush, C. R., Barrick, C. J., Adkins, D. E., Shaw, G. S., Miller, D. R., de Villena, F. P.-M., Sullivan, P. F., et al. (2014), "Genetics of adverse reactions to haloperidol in a mouse diallel: a drug–placebo experiment and Bayesian causal analysis," *Genetics*, 196, 321–347.

Crowley, J. J., Kim, Y., Szatkiewicz, J. P., Pratt, A. L., Quackenbush, C. R., Adkins, D. E., van den Oord, E., Bogue, M. A., Yang, H., Wang, W., et al. (2012), "Genome-wide association mapping of loci for antipsychotic-induced extrapyramidal symptoms in mice," *Mammalian genome*, 23, 322–335.

Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., et al. (2015), "Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance," *Nature genetics*, 47, 353–360.

Davis, J. R., Fresard, L., Knowles, D. A., Pala, M., Bustamante, C. D., Battle, A., and Montgomery, S. B. (2016), "An efficient multiple-testing adjustment for eQTL studies that accounts for linkage disequilibrium between variants," *The American Journal of Human Genetics*, 98, 216–224.

Delaneau, O., Marchini, J., Consortium, . G. P., et al. (2014), "Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel," *Nature communications*, 5.

Doss, S., Schadt, E. E., Drake, T. A., and Lusis, A. J. (2005), "Cis-acting expression quantitative trait loci in mice," *Genome research*, 15, 681–691.

Giusti-Rodríguez, P., Xenakis, J., Crowley, J. J., Nonneman, R. J., DeCristo, D. M., Ryan, A., Quackenbush, C. R., Miller, D. R., Shaw, G. D., Zhabotynsky, V., et al. (2019), "Antipsychotic behavioral phenotypes in the mouse Collaborative Cross recombinant inbred inter-crosses (RIX)," *BioRxiv*, 761353.

Gorsuch, R. and Nelson, J. (1981), "CNG scree test: an objective procedure for determining the number of factors," in *annual meeting of the Society for Multivariate Experimental Psychology*.

Hardin, J., Hardin, J., Hilbe, J., and Hilbe, J. (2007), *Generalized Linear Models and Extensions, Second Edition*, A Stata Press publication, Taylor & Francis.

Horn, J. L. (1965), "A rationale and test for the number of factors in factor analysis," *Psychometrika*, 30, 179–185.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012), "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing," *Nature genetics*, 44, 955–959.

Howie, B., Marchini, J., and Stephens, M. (2011), "Genotype imputation with thousands of genomes," *G3: Genes, Genomes, Genetics*, 1, 457–470.

Howie, B. N., Donnelly, P., and Marchini, J. (2009), "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies," *PLoS Genet*, 5, e1000529.

Hu, Y.-J., Sun, W., Tzeng, J.-Y., and Perou, C. M. (2015), "Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data," *Journal of the American Statistical Association*, 110, 962–974.

Jirtle, R. L. (2016), "Gene imprint, Imprinted Gene Database (Internet)," Available at `http://www.geneimprint.com/site/genes-by-species`.

Kaiser, H. F. (1960), "The application of electronic computers to factor analysis," *Educational and psychological measurement*, 20, 141–151.

Kim, Y., Giusti-Rodriguez, P., Crowley, J. J., Bryois, J., Nonneman, R. J., Ryan, A. K., Quackenbush, C. R., Iglesias-Ussel, M. D., Lee, P. H., Sun, W., et al. (2018), "Comparative genomic evidence for the involvement of schizophrenia risk genes in antipsychotic effects," *Molecular psychiatry*, 23, 708.

Kumasaka, N., Knights, A. J., and Gaffney, D. J. (2016), "Fine-mapping cellular QTLs with RASQUAL and ATAC-seq," *Nature genetics*, 48, 206–213.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015), "Integrative analysis of 111 reference human epigenomes," *Nature*, 518, 317.

Lagarrigue, S., Martin, L., Hormozdiari, F., Roux, P.-F., Pan, C., Van Nas, A., Demeure, O., Cantor, R., Ghazalpour, A., Eskin, E., et al. (2013), "Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage," *Genetics*, 195, 1157–1166.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013), "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, 501, 506–511.

Lawley, D. (1956), "Tests of significance for the latent roots of covariance and correlation matrices," *biometrika*, 43, 128–136.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013), "Software for computing and annotating genomic ranges," *PLoS Comput Biol*, 9, e1003118.

León-Novelo, L. G., McIntyre, L. M., Fear, J. M., and Graze, R. M. (2014), "A flexible Bayesian method for detecting allelic imbalance in RNA-seq data," *BMC genomics*, 15, 920.

Lerer, B., Segman, R. H., Tan, E.-C., Basile, V. S., Cavallaro, R., Aschauer, H. N., Strous, R., Chong, S.-A., Heresco-Levy, U., Verga, M., et al. (2005), "Combined analysis of 635 patients confirms an age-related association of the serotonin 2A receptor gene with tardive dyskinesia and specificity for the non-orofacial subtype," *International Journal of Neuropsychopharmacology*, 8, 411–425.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., et al. (2005), "Effectiveness of antipsychotic drugs in patients with chronic schizophrenia," *New England journal of medicine*, 353, 1209–1223.

Love, M. I., Huber, W., and Anders, S. (2014), "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome biology*, 15, 550.

Luedi, P. P., Dietrich, F. S., Weidman, J. R., Bosko, J. M., Jirtle, R. L., and Hartemink, A. J. (2007), "Computational and experimental identification of novel human imprinted genes," *Genome research*, 17, 1723–1730.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010), "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome research*, 20, 1297–1303.

McKenzie, M., Henders, A. K., Caracella, A., Wray, N. R., and Powell, J. E. (2014), "Overlap of expression quantitative trait loci (eQTL) in human brain and blood," *BMC medical genomics*, 7, 31.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013), "Identification of genetic variants that affect histone modifications in human cells," *Science*, 342, 747–749.

Morcos, L., Ge, B., Koka, V., Lam, K. C., Pokholok, D. K., Gunderson, K. L., Montpetit, A., Verlaan, D. J., and Pastinen, T. (2011), "Genome-wide assessment of imprinted expression in human cells," *Genome biology*, 12, 1.

Morison, I. M., Ramsay, J. P., and Spencer, H. G. (2005), "A census of mammalian imprinting," *TRENDS in Genetics*, 21, 457–465, `http://www.otago.ac.nz/IGC`.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008), "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature methods*, 5, 621–628.

Panousis, N. I., Gutierrez-Arcelus, M., Dermitzakis, E. T., and Lappalainen, T. (2014), "Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies," *Genome biology*, 15, 467.

Patsopoulos, N. A., Ntzani, E. E., Zintzaras, E., and Ioannidis, J. P. (2005), "CYP2D6 polymorphisms and the risk of tardive dyskinesia in schizophrenia: a meta-analysis," *Pharmacogenetics and genomics*, 15, 151–158.

Paul, S. R., Balasooriya, U., and Banerjee, T. (2005), "Fisher Information Matrix of the Dirichlet-multinomial Distribution," *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47, 230–236.

Pillai, A. (2008), "Decreased expression of Sprouty2 in the dorsolateral prefrontal cortex in schizophrenia and bipolar disorder: a correlation with BDNF expression," *PLoS One*, 3, e1784.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. (2007), "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, 81, 559–575.

Raiche, G. (2010), "an R package for parallel analysis and non graphical solutions to the Cattell scree test," .

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., et al. (2009), "The UCSC genome browser database: update 2010," *Nucleic acids research*, gkp939.

Ronald, J., Brem, R. B., Whittle, J., and Kruglyak, L. (2005), "Local regulatory variation in Saccharomyces cerevisiae," *PLoS Genet*, 1, e25.

Shabalin, A. A. (2012), "Matrix eQTL: ultra fast eQTL analysis via large matrix operations," *Bioinformatics*, 28, 1353–1358.

Spellmann, I., Rujescu, D., Musil, R., Giegling, I., Genius, J., Zill, P., Dehning, S., Cerovecki, A., Seemüller, F., Schennach, R., et al. (2014), "Pleckstrin homology domain containing 6 protein (PLEKHA6) polymorphisms are associated with psychopathology and response to treatment in schizophrenic patients," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 51, 190–195.

Sun, W. (2012), "A statistical framework for eQTL mapping using RNA-seq data," *Biometrics*, 68, 1–11.

Sun, W. and Hu, Y. (2013), "eQTL mapping using RNA-seq data," *Statistics in biosciences*, 5, 198–219.

Sun, W., Wright, F. A., et al. (2010), "A geometric interpretation of the permutation p-value and its application in eQTL studies," *The Annals of Applied Statistics*, 4, 1014–1033.

Tarone, R. E. (1979), "Testing the goodness of fit of the binomial distribution," *Biometrika*, 66, 585–590.

Threadgill, D. W., Hunter, K. W., and Williams, R. W. (2002), "Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort," *Mammalian genome*, 13, 175–178.

Tomiyama, K., McNamara, F. N., Clifford, J. J., Kinsella, A., Koshikawa, N., and Waddington, J. L. (2001), "Topographical assessment and pharmacological characterization of orofacial movements in mice: dopamine D1-like vs. D2-like receptor regulation," *European journal of pharmacology*, 418, 47–54.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009), "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, 25, 1105–1111.

Turrone, P., Remington, G., and Nobrega, J. N. (2002), "The vacuous chewing movement (VCM) model of tardive dyskinesia revisited: is there a relationship to dopamine D2 receptor occupancy?" *Neuroscience & Biobehavioral Reviews*, 26, 361–380.

van de Geijn, B., McVicker, G., Gilad, Y., and Pritchard, J. K. (2015), "WASP: allele-specific software for robust molecular quantitative trait locus discovery," *Nature methods*, 12, 1061–1063.

Wang, Z., Gerstein, M., and Snyder, M. (2009), "RNA-Seq: a revolutionary tool for transcriptomics," *Nature reviews genetics*, 10, 57–63.

Zhabotynsky, V., Inoue, K., Magnuson, T., Calabrese, J. M., and Sun, W. (2019), "A Statistical Method for Joint Estimation of Cis-eQTLs and Parent-of-Origin Effects under Family Trio Design," *Biometrics*.

Zou, F., Sun, W., Crowley, J. J., Zhabotynsky, V., Sullivan, P. F., and de Villena, F. P.-M. (2014), "A novel statistical approach for jointly analyzing RNA-Seq data from F1 reciprocal crosses and inbred lines," *Genetics*, 197, 389–399.