THE UNIVERSITY OF CHICAGO

EQUAL LOCAL LEVELS: A GLOBAL TESTING APPROACH WITH APPLICATION TO TRANS EQTL DETECTION

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
TAKINTAYO AKINBIYI

CHICAGO, ILLINOIS
DECEMBER 2020

For my wife Hilary and my future children

This was the hardest task I've had in my life. I thought I couldn't do it, but I'm glad I finished.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I must thank Mary Sara McPeek and Mark Abney for their tireless patience with me. The Department of Statistics has been beyond kind and patient. I left the program and returned to open arms and welcoming hands. No one has been anything but helpful.

I would also like to thank my wife Hilary and my parents, and my wife's parents. Without their support we would not have been able to do this.

ABSTRACT

In order to detect trans eQTLs in a given tissue type, it is common to perform an association test between each pair consisting of an expression level for a gene and a genetic variant that is "trans" for that gene, for D different genes and M different genetic variants, where D could be on the order of thousands or tens of thousands and M could be on the order of hundreds of thousands or millions. Then, a multiple testing correction, e.g., Bonferroni or FDR, for $M * D$ hypothesis tests would commonly be imposed, in order to try to identify significant (gene, trans eQTL) pairs. Maintaining correct type 1 error without sacrificing power becomes particularly challenging in this context.

For trans eQTLs, we might expect that association signals for individual (gene, trans eQTL) pairs would commonly be of only moderate or weak size, which could make the standard approach hopelessly under-powered. However, if a trans eQTL has moderate or weak association with multiple expression traits, then it might be possible to detect the trans eQTL using a global test, in which, for each SNP, we test the global null hypothesis that the SNP is not associated with the expression levels of any genes for which it is trans, against the alternative hypothesis that is it associated with the expression level for at least one of the genes for which it is trans. In the global testing approach, the number of tests performed is M, so the final multiple testing correction is comparable to that needed for an ordinary genome-wide association study. This work focuses on development of a global test that will be powerful for trans eQTL detection.

We propose a global testing approach based on equal local levels (ELL), where our approach allows for dependence among the tests. The ELL test statistic is based on Z-scores from tests of association for individual (gene, SNP) pairs. This allows the method to be applied in situations when only summary statistics are available. Another key feature of the method is that it is computationally fast so that it is feasible for genome-wide analysis. We verify the

type 1 error of the method and compare its power to other available methods in simulations. We find a huge power improvement over other available approaches for a wide range of values for the number of associated expression traits for a given trans eQTL. We apply ELL to perform trans eQTL mapping in the LG/J x SM/J advanced intercross line of mice, which is a multi-generational outbred population dataset consisting of 523,028 SNPs and expression levels of 15,561 genes recorded for the hippocampus, pre-frontal cortex and striatum.

CHAPTER 1

INTRODUCTION

1.1 eQTL Studies

Expression quantitative trait locus (eQTL) mapping is an approach to identify sites on the genome that are associated with transcription of particular genes in a given tissue type, with the goal of understanding the impact of polymorphisms at these sites. In eQTL mapping studies, gene expression levels are viewed as quantitative traits, and association is tested between these traits and SNPs or other genetic variants for which genotypes are available. A common observation, from a variety of both linkage and association eQTL studies, is that numerous genes have eQTLs nearby, which are called cis eQTLs, and are likely in cis regulatory elements [8]. Furthermore, several eQTL-mapping studies [8, 25, 19] have shown that disease-predisposing variants identified in genome-wide association studies are often cis eQTLs for nearby genes. As a result, eQTL mapping has become a widely used tool for identifying genetic variants that affect gene regulation. A few recent studies have also identified trans eQTLs [25], which are eQTLs that are associated with expression traits for one or more genes that are not nearby in the genome. Because the targets of trans eQTLs could potentially be anywhere in the genome, not just among nearby genes, trans eQTLs are expected to be much harder to detect, assuming that their effect sizes are not any larger than those for cis eQTLs. [8, 19]

1.2 Notation and Framework for Trans eQTL Mapping

We consider the problem of detecting trans eQTLs in the following scenario. Suppose that, for each of n subjects, we observe $\tilde{D}$ transcript levels in a given tissue type as well as genotypes for M genomewide SNPs. For any given SNP, say SNP i, we define a subset of the traits, $\mathcal{D}_i$, for which SNP i is not a cis SNP, e.g., $\mathcal{D}_i$ might consist of all traits whose genes

lie at least a certain distance from SNP i, or whose genes lie on a different chromosome from SNP i, depending on how we choose to define "cis SNP" in a given study. Then we will identify SNP i as a potential trans eQTL if it is significantly associated with at least one trait in $\mathcal{D}_i$.

For notational simplicity, we suppose that $|\mathcal{D}_i| = D$ for all i. One approach to trans eQTL detection is to perform $M * D$ hypothesis tests, one for each trans SNP - trait pair, and then apply a multiple comparisons correction, e.g., Bonferroni or FDR, for $M * D$ hypothesis tests, in order to try to identify significant trans SNP - trait pairs. A major drawback of this approach is that unless individual trans eQTL - trait association signals tend to be extremely large, the power of this approach will be very low because of the extremely stringent multiple testing correction that must be applied [19] (for $M * D$ tests instead of the usual M tests that would need to be corrected for in an ordinary GWAS).

One plausible scenario for trans eQTL mapping is that while individual trans eQTL - trait association signals might commonly be of only moderate or weak size, if a trans eQTL has moderate or weak association with multiple traits, then it might be possible to detect the trans eQTL using a global test. Therefore, we propose to perform M global hypothesis tests, one for each SNP, where for SNP i, the null hypothesis for its global test is H_0 : SNP i is not associated with any trait in $\mathcal{D}_i$. This will be followed by correction for M tests, in the same way as is usually done for a GWAS. The focus of this work is development of a global test for each SNP that will be powerful for trans eQTL mapping.

One notable feature of the trans eQTL mapping problem is that D may be quite large (thousands or tens of thousands), and for a given trans eQTL, the number of associated traits might be only a tiny fraction of D, with the effect of the trans eQTL being of only moderate or weak size for each associated trait, the so-called "rare and weak scenario" [7] for

global hypothesis testing. In this setting, when considering the D p-values for a given trans eQTL (corresponding to the association tests of the trans eQTL with each of the D traits), the smallest p-value is likely to correspond to a true null hypothesis, with the p-values for the true alternative hypotheses being less extreme. As a result, focusing attention on only the minimum of the D p-values would not be expected to be a very powerful approach to global hypothesis testing in some plausible settings for trans eQTL mapping.

1.3 Previous Work

1.3.1 Trans-eQTL Mapping Literature

Before we begin our description of the ELL method we first review past approaches to addressing the multiple comparisons problem in trans eQTL mapping.

One basic approach, common in model organisms [5, 11], is to first consider each expression trait separately, and, for each trait, perform a permutation test to obtain genomewide p-values for association of genetic variants with the trait. To correct for testing multiple traits, one could then apply a multiple corrections [5] procedure such as Bonferroni or FDR [2]. Alternatively, Bonferroni or FDR could be applied to all $M * D$ hypothesis tests, corresponding to all possible (expression trait, SNP) pairs in the data set, in order to discover individual (expression trait, SNP) pairs. A drawback of all such approaches is that the high multiple comparisons penalty can result in a lack of power to detect trans eQTLs, as detailed in the previous subsection.

One approach to avoiding the high multiple comparisons penalty is to find ways to make either M, the number of genetic variants tested, or D, the number of expression traits tested, smaller. For example, one could make M smaller by restricting the set of genetic variants tested to be only known transcription factors [27]. Another approach GBAT [15], involves

making M smaller by forming leave-one-out cross-validated BLUP predictions of a given gene's expression levels using SNPs cis to that gene, and then treating the BLUP predictions as if they were genetic variants and mapping them against expression levels of traits to which they are trans. Approaches to making D smaller include clustering similar expression traits or applying a dimension reduction method such as principal components to the expression traits[12]. On the one hand, incorporating biological information to effectively make M or D smaller has the potential to increase power if the investigator guesses right about what biological information is most relevant, but it also has the potential to miss important signals and is limited by the initial choices made in the dimension reduction. Furthermore, approaches that are based on, say, linear combinations of SNPs or of expression traits may yield results that are difficult to interpret biologically.

The cross-phenotype meta-analysis (CPMA) [13, 4] test is a global test for a single SNP with a specified set of expression traits. The global null hypothesis is that the SNP is not associated with any expression trait in the set. The idea behind the test is that if the expression traits in the set were independent, then the set of -$\log_e$ p-values for the association tests between the SNP and each trait would be i.i.d. Exp(1) random variables under the global null hypothesis. To calculate the CPMA test statistc, one models the set of -$\log_e$ p-values as i.i.d. Exp(λ) under the alternative model and sets the CPMA statistic to be the likelihood ratio chi-squared test statistic for testing $H_0 : \lambda = 0$ vs. $H_A : \lambda > 0$ in the i.i.d. case. In the case when the expression traits are correlated, the test statistic remains the same, and a Monte Carlo procedure is used to assess significance instead of the usual χ^2_1 distribution. [4]

1.3.2 Other Relevant Literature

Various methods have been proposed for combining Z scores or p-values from D individual hypothesis tests in order to test (or to compute a p-value for) the global null hypothesis

that all D individual null hypotheses are true vs. the alternative that at least one is false. For many of these methods, the assumption is that the D p-values (or Z scores) for the individual hypothesis tests are i.i.d. Unif(0,1) (or N(0,1)) random variables under the global null hypothesis. In trans eQTL mapping, it is expected that most of the individual null hypotheses will be true even under the global alternative that at least one individual null hypothesis is false, i.e., association signals are moderately "sparse" or "rare". In that scenario, particularly relevant examples of global tests that combine Z scores or p-values include the Berk-Jones statistic [3], Simes' method [20], higher criticism [6, 7], and methods based on local levels [9, 10].

There are also extensions of some of these methods to dependent tests. For example, Barnett et al. [1] developed generalized higher criticism (GHC) a version of higher criticism that allows dependence among the Z scores under the null hypothesis. They applied it to test for SNP-set effects on a trait, i.e., the set of test statistics for which a global test is performed is the set of association test statistics between a given trait and each of a set of SNPs. However, even in the case of independence, the higher criticism test has been shown to under-perform relative to what would be expected based on asymptotic theory.[9] (For example, not until D is of the order of 10^{69} do the asymptotic optimality results seem to hold. [9]) As an alternative approach, Sun and Lin [22] developed generalized Berk-Jones (GBJ), a version of the Berk-Jones test that allows dependence among the Z scores under the null hypothesis, which they also applied to test for the association between a SNP-set and outcome. Software is available for both GHC and GBJ, and we used these for trans eQTL mapping in order to compare to our ELL method. However, in 3.5, we show that the software for both GHC and GBJ is not computationally feasible to run in the settings we consider.

CHAPTER 2

MODELS & DETECTION METHODS

For notational simplicity we first consider the global hypothesis test for just one of the M SNPs, say SNP 1. Assume we have Z-scores $Z_1, ..., Z_D$ with corresponding p-values $\pi_d = 2\Phi(-|Z_d|),\ d = 1, ..., D$ that are ordered as $\pi_{(1)} \leq \cdots \leq \pi_{(D)}$. Each Z_d corresponds to a test statistic for the individual null hypothesis of no association between the SNP and gene transcript d. We describe in subsection (2.4) how we calculate Z_d specifically for the trans eQTL mapping problem. The global null hypothesis is that the SNP is not associated with any trait, and in that case, each Z_d is assumed to be marginally $N(0, 1)$ distributed. In practice there is often correlation between gene transcript levels (see [30] and the references therein). This propagates to correlation between the test statistics Z_d, $1 \leq d \leq D$, above. Let $\Omega = V_0(\boldsymbol{Z})$ be the variance matrix for the vector $\boldsymbol{Z} = (Z_1, ..., Z_D)$ under the global null hypothesis, where Ω is in fact a correlation matrix, because the Z_d's each have variance 1 under the global null hypothesis. Under certain modeling assumptions, Ω can be shown to be the same across SNPs, and this, as well as estimation of Ω, is discussed below in section 2.4. Under the global null hypothesis, we assume $\boldsymbol{Z}$ has the distribution $MVN\,(0,\ \Omega)$.

Significant inflation of some of the components of $\boldsymbol{Z}$ would be viewed as evidence against the global null. Equivalently, such evidence would be manifest by smaller values of some of the components of $(\pi_{(1)}, ..., \pi_{(D)})$ than would be expected under the null. On the one hand, if trans eQTL signals are only of moderate or weak size, then, e.g., $\pi_{(1)}$ or $\pi_{(2)}$ might actually represent null tests, and the true alternatives could be represented by smaller than expected $\pi_{(d)}$ for values of d that are perhaps of small to moderate size. On the other hand, a finding that $\pi_{(d)}$ is smaller than expected for only large d, e.g., d close to D, would be difficult to interpret and might not seem to be compelling evidence for the presence of a trans eQTL. Therefore, we base our ELL test only on a subset, $\mathcal{D}$, of the pvalues, where $\mathcal{D}$ is of the form $\{\pi_{(1)}, \ldots, \pi_{(|\mathcal{D}|)}\}$, where $|\mathcal{D}| < D$. For example $|\mathcal{D}|$ could be $0.2D$ or $0.5D$, meaning that we

would only consider the smallest 20% or 50% of the p-values. (Our default choice in what follows is $|\mathcal{D}| = 0.2D$.) This inspires a class of hypothesis tests for the global null hypothesis parameterized by a set $0 < h_1 \leq \cdots \leq h_{|\mathcal{D}|} < 1$ defined by the rejection rule: reject the global null if $\pi_{(d)} < h_d$ for at least one $d \in \mathcal{D}$. The type 1 error of such a test would be α such that

$$P_0\left[\forall\ d \in \mathcal{D},\ \pi_{(d)} \geq h_d\right] = 1 - \alpha, \tag{2.1}$$

where the zero subscript in $P_0(\cdot)$ indicates the global null hypothesis is true. We define $\eta_d(h) \equiv P_0\left[\pi_{(d)} < h\right]$. In the case when $\Omega = I$, $\eta_d(h_d)$ is referred to as the dth local level [10] of the rejection region defined by $h_1, ..., h_{|\mathcal{D}|}$, and we adopt this term also for the case when $\Omega \neq I$. for the case when $\Omega = I$, Berk and Jones [3] first proposed and demonstrated the efficiency of choosing the h_d such that

$$\forall\ d \in \mathcal{D}, \quad \eta_d(h_d) = \eta \equiv \eta(\alpha) \tag{2.2}$$

where $\eta(\alpha)$ is chosen so that

$$P_0\left[\forall\ d \in \mathcal{D},\ \pi_{(d)} \geq h_d\right] = 1 - \alpha.$$

(Note, however, that this is not the same as the so-called "Berk-Jones" statistic [3, 10].) Equation (2.2) corresponds to having equal local levels (ELL), and we refer to tests satisfying (2.2) as ELL tests. For the case $\Omega = I$, ELL tests have been advocated by Gonscharuk, Landwehr and Finner [9, 10] because they have similar asymptotic properties to those of higher criticism [6, 7], but better finite-sample properties.

The p-value for the ELL test would be the largest α for which the ELL test would not reject, i.e., $\max\left\{\alpha\ :\ \forall\ d \in \mathcal{D},\ \pi_{(d)} \geq h_d\right\}$. We can rewrite this in simpler terms by noting

that, because η_d is a strictly monotone function when the underlying Z_d's have a continuous distribution,

$$\begin{aligned} \forall\ d \in \mathcal{D},\ \pi_{(d)} \geq h_d &\iff \forall\ d \in \mathcal{D},\ \eta_d\left(\pi_{(d)}\right) \geq \eta_d\left(h_d\right) = \eta(\alpha) \\ &\iff \min_{d \in \mathcal{D}} \eta_d\left(\pi_{(d)}\right) \geq \eta(\alpha) \end{aligned} \tag{2.3}$$

From (2.3) we would like to define the ELL statistic as $T = \min_{d \in \mathcal{D}} \eta_d\left(\pi_{(d)}\right)$ and reject the global null hypothesis at level α if $T < \eta(\alpha)$ where $\eta(\alpha)$ is chosen so that the type 1 error is α. However this requires determining $\eta(\alpha)$ as well as the corresponding h_d values.

For the special case when $\Omega = I$, under the global null hypothesis, the p-values $\pi_1, ..., \pi_D$ are i.i.d. Uniform(0,1), and each $\pi_{(d)}$ is distributed as Beta($d\ ,\ D + 1 - d$). Thus the η_d function is the cumulative distribution function (CDF) for a Beta($d\ ,\ D + 1 - d$) random variable. Weine and McPeek [24] derived exact recursive formulas for $\eta(\alpha)$ for the case $\Omega = I$. However, when Ω is not the identity matrix, then these analytical formulas no longer hold.

2.1 ELL Test Statistic

While we do not have analytical formulas for $\eta(\cdot)$ or the $\eta_d(\cdot)$, we can find approximations for these functions. Let $S(c) = \sum_{d=1}^{D} \mathcal{I}\left\{|Z_d| > c\right\}$ where $\mathcal{I}\{\cdot\}$ is the indicator function, and let Φ be the CDF of N(0,1). For $0 < h < 1$,

$$\begin{aligned} P_0\left[\pi_{(d)} < h\right] &= P_0\left[|Z|_{(D+1-d)} > -\Phi^{-1}(h/2)\right] \\ &= P_0\left[\left(\sum_{k=1}^{D} \mathcal{I}\left\{|Z_k| > -\Phi^{-1}(h/2)\right\}\right) \geq d\right] \\ &= P_0\left[S\left(-\Phi^{-1}(h/2)\right) \geq d\right] \end{aligned} \tag{2.4}$$

If $\Omega = I$, then for $c \geq 0$, $S(c)$ has the null distribution of a Binomial$(D, 2\Phi(-c))$ random variable, and

$$P_0\left[\pi_{(d)} < h\right] = 1 - \sum_{j=0}^{d-1} \binom{D}{j} h^j (1-h)^{D-j}, \quad h \in [0,1]$$

which is the CDF of a Beta$(d, D+1-d)$.

When $\Omega \neq I$, the null mean of $S(c)$ will remain the same, $E_0[S(c)] = 2D\Phi(-c)$, however the null variance is

$$\begin{aligned} V_0(S(c)) &= V_0\left(\sum_k \mathcal{I}\left\{|Z_k| > c\right\}\right) \\ &= 2D\Phi(c)(1-2\Phi(c)) + 2 * \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \operatorname{cov}\left(\mathcal{I}\left\{|Z_i| > c\right\}, \mathcal{I}\left\{|Z_j| > c\right\}\right) \\ &> 2D\Phi(c)(1-2\Phi(c)) \end{aligned}$$

Where the last inequality follows from the positive correlation of the magnitudes of any correlated, mean-zero, bivariate Gaussians.

Following the method proposed by [1] (see also [22]), we approximate the distribution of $S(c)$ with a beta-binomial distribution $BB(D, \lambda, \gamma)$ where we choose λ and γ to match the mean and variance of $S(c)$. The beta-binomial is the distribution formed from the binomial distribution when the success probability is drawn from a beta distribution. If X is drawn from a BB$(D,\ \lambda,\ \gamma)$, where $\lambda > 0$, $\gamma > 0$, and D is a positive integer, then X has the

following properties:

$$f_{D,\lambda,\gamma}(x) \equiv P(X=x) = \binom{D}{x} \frac{\prod_{k=0}^{x-1}(\lambda+\gamma k)\prod_{k=0}^{D-x-1}(1-\lambda+\gamma k)}{\prod_{k=0}^{D-1}(1+\gamma k)} \tag{2.5}$$

$$\lambda = E(X)/D$$

$$\frac{\gamma}{1+\gamma} = \frac{V(X)-D\lambda(1-\lambda)}{D(D-1)\lambda(1-\lambda)},$$

where we follow the convention that

$$\prod_{k=0}^{a} c_k = 1 \text{ for } a < 0.$$

To approximate the distribution of $S(c)$ we choose λ and γ by the method of moments. Barnett et al. [1] derived the following:

$$\begin{aligned} E_0\left[S\left(c\right)\right] &= 2D\Phi(-c) \\ V_0\left[S\left(c\right)\right] &= D\left[2\overline{\Phi}(c) - 4\overline{\Phi}^2(c)\right] + 4D(D-1)\phi^2(c)\sum_{i=1}^{\infty}\mathcal{H}_{2i-1}^2(c)\rho(2i)/(2i)! \end{aligned} \tag{2.6}$$

Where $\mathcal{H}_i$ are the Hermite polynomials (see [16]), $\rho(i) = \frac{2}{D(D-1)}\sum_{1\le k<l\le D}\left(\Omega_{kl}\right)^i$, where Ω_{kl} is the (k,l)th entry of Ω, and ϕ and $\overline{\Phi}$ are the density and survivor functions of a standard normal, respectively. Note that (2.5) and (2.6) imply that once λ is set by the method of moments, knowing λ determines c which determines $V_0\left[S\left(c\right)\right]$ and thus determines γ. So we will drop the writing of γ (and D) below to simplify notation, and define $f_\lambda(x)$ to be the beta-binomial density with parameters (D,λ,γ), where λ and γ are related by the equations $\lambda = 2\Phi(-c)$ and

$$\frac{\gamma}{1+\gamma} = \frac{V_0[S(c)] - D\lambda(1-\lambda)}{D(D-1)\lambda(1-\lambda)},$$

with $V_0[S(c)]$ given in 2.6.

Applying this to (2.4)

$$\begin{aligned}
\eta_d(h) &= P_0\left[\pi_{(d)} < h\right] \\
&= P_0\left[S\left(-\Phi^{-1}\left(h/2\right)\right) \geq d\right] \\
&= 1 - \sum_{k=0}^{d-1} P_0\left[S\left(-\Phi^{-1}\left(h/2\right)\right) = k\right] \\
&\approx 1 - \sum_{k=0}^{d-1} f_h\left(k\right) \\
&\equiv \hat{\eta}_d(h). \qquad (2.7)
\end{aligned}$$

Using this approximation we define the ELL statistic:

$$T_{ELL} \equiv \min_{d \in \mathcal{D}} \hat{\eta}_d\left(\pi_{(d)}\right). \qquad (2.8)$$

To assess significance of the ELL statistic T_{ELL}, we consider two different approaches, (1) Monte Carlo simulation and (2) application of a Markov approximation [22] to the joint distribution of $(S(c_1), \ldots, S(c_k))$ for any finite k and $c_1 < \ldots < c_k$.

2.2 Assessment of Significance of ELL by Monte Carlo

Suppose we have an estimate $\hat{\Omega}$ for Ω. (In section (2.4) below describe how we obtain $\hat{\Omega}$.) We simulate R i.i.d vectors $\tilde{Z}^{(j)} \sim \ MVN_D\left(0\ ,\ \hat{\Omega}\right),\ j = 1, .., .R$, where R is very large. For each $\tilde{Z}^{(j)}$ we calculate the ELL statistic $T^{(j)}$. Since η is just the inverse CDF of T_{ELL} under the global null we can estimate it as the inverse empirical CDF of T_{ELL} using the

data set $T^{(1)}, ..., T^{(R)}$:

$$\hat{\eta}(\alpha) \equiv \max \left\{ x \ : \ 1 - \alpha \leq \frac{1}{R} \sum_{j=1}^{R} \mathcal{I} \left(T^{(j)} \geq x \right) \right\}, \text{ for } \alpha \in (0, 1). \tag{2.9}$$

The p-value of an observed T_{ELL} is the largest α such that the global null would not be rejected. We call this method ELL.

2.3 Assessment of Significance of ELL by a Markov Chain Approximation

We also consider trying to find an analytical estimate for $\hat{\eta}(\alpha)$. Sun and Lin [22] introduced a technique for analytically approximating a calculation of the form

$$Pr\{\forall d = 1, ..., D, \ |Z|_{(d)} \leq b_d \mid \mathbf{Z} \sim MVN(0, \ \Omega)\} \tag{2.10}$$

where $\mathbf{Z}$ is the $D \times 1$ vector whose dth element is Z_d. Note that (2.10) can be re-written as

$$Pr\{\forall d = 1, .., D, \ S(b_d) \leq D - d \mid \mathbf{Z} \sim MVN(0, \ \Omega)\} \tag{2.11}$$

Sun and Lin [22] propose a recursion to solve (2.11). Define

$$q_{d,a} = Pr \left\{ S(b_d) = a, \bigcap_{k=1}^{d-1} \{S(b_k) \leq d - k\} \mid \ \mathbf{Z} \sim MVN(0, \ \Omega) \right\}. \tag{2.12}$$

Then (2.11) is just $q_{D,0}$. Sun and Lin [22] make the assumption that conditional on $S(b_d)$ the event $S(b_{d+1}) = a$ is approximately independent of $S(b_k)$ for $k < d$, and they approximate the conditional distribution of $S(b_{d+1})$ given $S(b_d)$ with an extended Beta-Binomial distribution, using the method of moments. They then derive a simple recursion for solving

$q_{d,a}$.

We considered using the same approach to find an approximate analytical p-value for the ELL statistic. To do this, note that we can rewrite

$$Pr\left\{T_{ELL} \geq t\right\} = Pr\left\{\forall d = 1, ..., D,\ |Z|_{(d)} \leq -\Phi^{-1}\left(\hat{\eta}_d^{-1}(t)/2\right)\right\}. \tag{2.13}$$

We can then use (2.12) with $b_d = -\Phi^{-1}\left(\hat{\eta}_d^{-1}(t)/2\right)$. We call this method ELL-Recursion.

Pre-computation For The ELL Test

Our ELL method lends itself to a pre-computation to reduce computation time when it will be applied to a large number of SNPs and a large number of traits. Suppose we observe M SNPs along with the D traits. Then, from Equation (2.7), calculating M *ELL* statistics involves $M \times \sum_{j=1}^{|\mathcal{D}|} j = M(|\mathcal{D}|+1)(|\mathcal{D}|)/2 \approx \frac{M|\mathcal{D}|^2}{2}$ evaluations of a beta-binomial probability mass function $f_h(k)$.

However, we can note from (2.5) that for a given positive integer k_1, the computation of $f_h(k_1)$ involves the simultaneous computation of all the terms needed to construct $f_h(k_2)$ for every integer k_2 satisfying $0 \leq k_2 \leq k \leq k_1$. This is particularly clear if we look at $\log\left[f_h(k)\right]$. Each $\log\left[f_h(k_1)\right]$ involves computing various sums where, for each of these sums, we can just take a corresponding partial sum to form $\log\left[f_h(k_2)\right]$ for $0 \leq k_2 \leq k_1$.

This motivates a computationally efficient strategy for approximating the function $\hat{\eta}_d$ by a step function $\tilde{\eta}_d$. We choose a set of pre-computation points, $\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$, where $0 < h_1 < \cdots < h_{|\mathcal{H}|} < 1$, and evaluate each $\hat{\eta}_d$ only on the points in $\mathcal{H}$. Making use of the approach in the previous paragraph also requires an upper and lower bound on the k for evaluating $f_h(k)$. Our strategy is to choose an η^L and η^R, which we describe below, and for

each $h \in \mathcal{H}$, we evaluate $\hat{\eta}_d(h)$ only for $d \in \mathcal{D}$ such that $\eta^L < \hat{\eta}_d(h) < \eta^R$. For each $d \in \mathcal{D}$, we use a binary search to find

$$\begin{aligned} \hat{\eta}_d^{-1}\left(\eta^L\right) &= \max\left\{h \in \mathcal{H} \ : \ \hat{\eta}_d(h) \leq \eta^L\right\} \\ \hat{\eta}_d^{-1}\left(\eta^R\right) &= \min\left\{h \in \mathcal{H} \ : \ \hat{\eta}_d(h) \geq \eta^R\right\} \end{aligned} \tag{2.14}$$

For $d \in \mathcal{D}$ but $h \notin \mathcal{H}$ we use the following approximation:

$$\tilde{\eta}_d(h) = \begin{cases} \eta^L & \mathrm{h} \leq \hat{\eta}_d^{-1}\left(\eta^L\right) \\ \hat{\eta}_d\left(\min\left\{x \in \mathcal{H} : \ x \geq h\right\}\right), & \hat{\eta}_d^{-1}\left(\eta^L\right) < h < \hat{\eta}_d^{-1}\left(\eta^L\right) \\ \eta^R & h > \hat{\eta}_d^{-1}\left(\eta^R\right) \end{cases} \tag{2.15}$$

We will need to choose values of η^R and η^L. For a fixed resolution (i.e., interpoint distance) of the points in $\mathcal{H}$, the larger the range $\eta_R - \eta_L$, the better the above approximation, however, the more computations and thus the slower the pre-computation. In practice we have found $\eta^L = 10^{-12}$ and $\eta^R = 1 - 10^{-12}$ to work well. However, in general one can use the following upper bound as guidance,

$$\begin{aligned} P_0\left[\forall \ d \in \mathcal{D}, \ \eta_d\left(\pi_{(d)}\right) \in \left[\eta^L, \ \eta^R\right]\right] &\geq 1 - \sum_{d \in \mathcal{D}} P_0\left[\eta_d\left(\pi_{(d)}\right) \notin \left[\eta^L, \ \eta^R\right]\right] \\ &= 1 - \sum_{d \in \mathcal{D}}\left(1 + \eta^L - \eta^R\right) \\ &\geq \delta \mathcal{I}\left(\eta^R - \eta^L \geq 1 - \frac{1-\delta}{|\mathcal{D}|}\right), \end{aligned}$$

for every $0 < \delta < 1$. In our Python code, we typically take $|\mathcal{H}| = 1000|\mathcal{D}|$. A naive means of assigning the values of $\mathcal{H}$ would be to create an evenly spaced grid between $\hat{\eta}_1^{-1}(\eta^L)$ and $\hat{\eta}_{|\mathcal{D}|}^{-1}(\eta^R)$. Unfortunately, this tends to not give enough resolution for small values of h. As a result, power would suffer as it relies on the accuracy of $\hat{\eta}_d$ particularly for small h. Instead

we choose $\mathcal{H}$ to be a geometric sequence:

$$h_d = \eta_{|\mathcal{D}|}^{-1}\left(\eta^R\right)\left(\frac{\eta_1^{-1}\left(\eta^L\right)}{\eta_{|\mathcal{D}|}^{-1}\left(\eta^R\right)}\right)^{\frac{|\mathcal{H}|-d}{|\mathcal{H}|-1}}. \tag{2.16}$$

This choice results in a set $\mathcal{H}$ that behaves better in terms of power.

2.4 Models For Trans-EQTL Mapping; Estimation of Ω

From the introduction to the ELL method in the beginning of this chapter, we can see that the ELL method lends itself to a broad range of applications in which one has a set of correlated test statistics that each have a N(0,1) distribution under their respective null hypotheses, provided that one is willing to assume a joint multivariate normal distribution under the global null hypothesis that all of the individual null hypotheses are true.

Here, we are concerned primarily with the application to trans-eQTL mapping, and in that context, we now describe a joint model for all the expression traits as well as marginal models that are used to calculate individual Z scores for testing association between a SNP and an expression trait. To jointly model the expression traits, we consider a linear model of the form

$$Y = X\beta + G\delta + \epsilon \tag{2.17}$$

where $Y_{N\times D}$ is the matrix of transcript levels; $X_{N\times c}$ is a matrix of covariates which is always assumed to include a column of ones (corresponding to an intercept term), and might also include covariates such as age, sex, batch effects, and ancestry-informative vectors (e.g., the first few principal components from the genetic relatedness matrix (GRM)); $\beta_{c\times D}$ are fixed effects of the covariates; $G_{N\times M}$ is a matrix of the genotypes of the N individuals in the study; $\delta_{M\times D}$ are the effects of the genotypes on the expression traits, either random or fixed,

where the (m, d)th element of δ is the effect of SNP m on gene d; $\epsilon_{N\times D}$ is the error matrix, where $\text{vec}(\epsilon) \sim N(0, V_\epsilon \otimes I_N)$, I_N is the $N \times N$ identity matrix, and V_ϵ is a $D \times D$ variance matrix. The model in equation (2.17) is somewhat vague, in that we have not yet specified which of the entries of δ are fixed and which are random. Below, we consider various more specific versions of this model (and sub-models of it) in the context of hypothesis testing.

To test whether a given SNP, say SNP m, is a trans-eQTL, we consider a model in which the effects of SNP m on the gene expression levels are considered separately from the effects of all the other SNPs, and the latter are aggregated into a random effect:

$$Y = X\beta + G_m\delta_m^T + [U\zeta^{1/2} + \epsilon(I_D - \zeta)^{1/2}]\eta^{1/2}, \tag{2.18}$$

where G_m is the vector of genotypes for SNP m; δ_m is a column vector and may be fixed or random; ζ is a $D \times D$ diagonal matrix of heritabilities; and I_D is a $D \times D$ identity matrix. The matrix $U_{N\times D}$ is the additive polygenic effect comprising the effect of the genotypes aside from the mth SNP; $\text{vec}(U) \sim MVN(0, \rho_u \otimes K)$, where K is the GRM with SNP m removed, and ρ_u is a $D \times D$ correlation matrix; $\text{vec}(\epsilon) \sim MVN(0, \rho_e \otimes I_N)$ where ρ_e is a $D \times D$ correlation matrix; and η is a $D \times D$ diagonal matrix with dth entry σ_d^2 equal to the residual variance for trait d.

To test whether SNP m is a trans-eQTL, one possible approach would be to model δ_m as a vector of random effects, $\delta_m \sim \text{MVN}(0, \sigma_\delta^2 I)$. Then, in principle, one could test the global null hypothesis $H_0 : \ \sigma_\delta^2 = 0$ by, for example, a score test. In order to perform the score test, it would be necessary to obtain the maximum likelihood estimates of the nuisance parameters under the null hypothesis, where the nuisance parameters are the $c \times D$ β values, the D diagonal elements of η, the D heritabilities, the $D(D-1)/2$ off-diagonal entries of ρ_u, and the $D(D-1)/2$ off-diagonal entries of ρ_e. Currently, software exists [29] that makes

this estimation computationally feasible for D of approximately 5 or so, but trans-eQTL mapping would commonly involves D in the thousands or tens of thousands, making this approach currently impractical. Furthermore, one would intuitively expect the score test to have high power when the signal is pervasive, rather than rare or sparse, i.e., when each trans eQTL is associated with a large proportion of the expression traits, even if the effect sizes are small. If a given eQTL is associated with a relatively small fraction of the expression traits, then other methods such as ELL would be expected to have higher power than the score test.

Instead, we suppose that a test for genetic association has been performed for each SNP-trait pair, resulting in a $M \times D$ matrix Z whose (m,d)th element, $Z_{m,d}$ is the Z-score for association of SNP m and trait d. For example, these might be summary statistics from a GWAS. To obtain $Z_{m,d}$, we assume that the following model has been used to fit the data:

$$Y_d = X\beta_d + G_m\delta_{m,d} + \sigma_d(\sqrt{\zeta_d}U_d + \sqrt{1-\zeta_d}\epsilon_d) \tag{2.19}$$

Here Y_d is a vector of length N containing the gene expression data for the dth gene; U_d and ϵ_d are independent with $U_d \sim \text{MVN}_N(0, K)$ and $\epsilon_d \sim \text{MVN}_N(0, I)$; $0 \leq \zeta_d \leq 1$ and $\sigma_d > 0$ are scalars. We can write Y_d in terms of its conditional distribution:

$$Y_d|X, G \sim \text{MVN}\left(X\beta_d + G_m\delta_{m,d}\ ,\ \sigma_d^2(\zeta_d K + (1-\zeta_d)\, I)\right), \tag{2.20}$$

which can be written $\text{MVN}\left(X\beta_d + G_m\delta_{m,d}, \sigma_d^2\Sigma_d\right)$, where we define $\Sigma_d = \zeta_d K + (1-\zeta_d)\, I$. We perform a score test of $H_0 : \delta_{m,d} = 0$ in model (2.20), which has test statistic:

$$Z_{m,d} = \frac{G_m^T P_d Y_d}{\sqrt{\hat{\sigma}_d^2 G_m^T P_d G_m}}, \tag{2.21}$$

where

$$P_d = \hat{\Sigma}_d^{-1} - \hat{\Sigma}_d^{-1}X(X^T\hat{\Sigma}_d^{-1}X)^{-1}X^T\hat{\Sigma}_d^{-1}, \tag{2.22}$$

$\hat{\Sigma}_d$ is the matrix Σ_d evaluated at $\hat{\zeta}_d$, which is the maximum likelihood estimate of ζ_d in (2.20) under the null hypothesis, i.e., when $\delta_{m,d} = 0$, and $\hat{\sigma}_d^2$ is an estimate of σ_d^2 under the null hypothesis. Under the null hypothesis, $Z_{m,d} \sim N(0,1)$, asymptotically. We evaluate $\mathrm{Cov}_0(Z_{m1,d1}, Z_{m2,d2}|X,G)$ under the global null hypothesis, assuming the model of equation (2.18), with $\delta_m = 0$. In that case,

$$\mathrm{Cov}_0(Z_{m1,d1}, Z_{m2,d2}|X,G) \approx \frac{G_{m1}^T P_{d1} A P_{d2} G_{m2}}{\sqrt{(G_{m1}^T P_{d1} G_{m1})(G_{m2}^T P_{d2} G_{m2})}}, \tag{2.23}$$

where

$$A = \sqrt{\zeta_{d1}\zeta_{d2}}\rho_u[d1,d2]K + \sqrt{(1-\zeta_{d1})(1-\zeta_{d2})}\rho_e[d1,d2]I, \tag{2.24}$$

with $\rho_u[d1,d2]$ the $(d1,d2)$th element of ρ_u and $\rho_e[d1,d2]$ the $(d1,d2)$th element of ρ_e. In either of the following 2 special cases, (i) $\zeta_d = \zeta$ for all $1 \le d \le D$ and $\rho_u = \rho_e$, or (ii) $\zeta_d = 0$ for all $1 \le d \le D$, we have $\Sigma_d = \Sigma$ for all d, so in large samples, $P_d \approx P$ for all d and equation (2.23) simplifies to

$$\mathrm{Cov}_0(Z_{m1,d1}, Z_{m2,d2}|X,G) \approx \frac{\rho_e[d1,d2] G_{m1}^T P G_{m2}}{\sqrt{(G_{m1}^T P G_{m1})(G_{m2}^T P G_{m2})}}, \tag{2.25}$$

which results in

$$\mathrm{Var}_0(\mathrm{vec}(Z)|X,G) = \Omega \otimes R, \tag{2.26}$$

where, in special cases (i) and (ii), $\Omega = \rho_e$ and $R : M \times M$ has (m_1, m_2)th entry

$$R_{m1,m2} = \frac{G_{m1}^T P G_{m2}}{\sqrt{(G_{m1}^T P G_{m1})(G_{m2}^T P G_{m2})}}, \tag{2.27}$$

which can be viewed as a version of the correlation between SNPs $m1$ and $m2$.

Motivated by the above heuristic, in our trans-eQTL analysis, we make the general modeling assumption that each row of Z has the same correlation matrix, which we call Ω. In special

cases (i) and (ii) above, we have Ω equal to the trait correlation matrix. However, in the more general case, the modeling assumption is only an approximation to the truth, and we no longer necessarily have Ω equal to the trait correlation matrix. Therefore, we propose to estimate Ω from the matrix Z rather than from the matrix Y. Furthermore, this approach of estimating Ω from Z has the benefit that it can also be applied in the case when only summary statistics (which presumably include Z) are available instead of individual-level data.

Suppose that, from a data set, we have access to summary statistics $Z_1, ..., Z_M$ where each Z_m, $1 \leq m \leq M$ is a vector of association test statistics between SNP m and a set of expression levels for genes trans to m. Let $\mathcal{D}_m$ be the set of expression traits for genes trans to SNP m. Then Z_m is a vector of length $|\mathcal{D}_m|$ for $1 \leq m \leq M$. Suppose that for each SNP, we wish to test the global null hypothesis that the given SNP, say SNP m, has no association with any of the expression traits in $\mathcal{D}_m$. We estimate Ω by $\hat{\Omega}$, whose (i,j)th element, $\hat{\Omega}_{ij}$, is the sample correlation coefficient between the Z scores for the ith and j expression traits, calculated across all SNPs that are trans to both traits. If we define $Q_{ij} = \{1 \leq m \leq M : \{i,j\} \subset \mathcal{D}_m\}$, then

$$\hat{\Omega}_{ij} = \frac{\sum_{m \in Q_{ij}} (Z_{m,i} - \bar{Z}_{\cdot i})(Z_{m,j} - \bar{Z}_{\cdot j})}{\sqrt{\sum_{m' \in Q_{ij}} (Z_{m'i} - \bar{Z}_{\cdot i})^2 \sum_{m'' \in Q_{ij}} (Z_{m''j} - \bar{Z}_{\cdot j})^2}},$$

for $1 \leq i \leq D$, $1 \leq j \leq D$, where $\bar{Z}_{\cdot i} = |Q_{ij}|^{-1} \sum_{m \in Q_{ij}} Z_{m,i}$, for $1 \leq i \leq D$.

2.5 Identifying Which Expression Traits are Associated with a Significant SNP

As the ELL method is a test of a global null hypothesis, in practice for those SNPS for which we reject the global null hypothesis, we need a methodology to identify the particular gene

expression levels we believe the SNP is associated with.

We use a methodology derived from [18]. Given a set of M SNPs to be considered for eQTL mapping and a genome wide cutoff, we calculate ELL p-values for each. Let m (possibly 0) be the number of SNPs whose ELL pvalue is below the genome wide cutoff. For each SNP for which the ELL p-value reaches the genomewide significance threshold, find the associated expression traits as follows: take the set of marginal association p-values for that SNP and apply FDR with target false discovery rate $0.05 * c$, where c is an adjustment factor that is the same for every SNP and is $c =$ (SNPs discovered)/(SNPs tested)$=m/M$.

CHAPTER 3

SIMULATION STUDIES

We perform simulation studies to evaluate the ELL and ELL-Recursion methods and to compare them to other methods. We consider type 1 error, power and computation time. Our simulations are in the context of correlated (i.e. $\Omega \neq I$) association statistics across gene transcript levels.

3.1 Other Test Statistics Considered in Simulations

Every test statistic we consider in the simulations can be formulated as a global test of the null hypothesis that a given SNP, m, is not associated with any of a set of D expression traits. Furthermore, each test statistic is a function of the Z scores, $Z_{m1}, \ldots, X_{mD}$, computed as described in equation (2.21). Below, we drop the subscript m from the Z scores, for notational simplicity, to obtain $Z_1, \ldots, X_D$. We let $\pi_d = 2\Phi(-|Z_d|)$, be the dth p-value, for $1 \leq d \leq D$, and we define $\pi_{(1)} \leq \ldots \leq \pi_{(D)}$ to be the order statistics of the p-values.

3.1.1 Sum of Z-Squared

This test statistic is $T_{sumZ^2} = \sum_{d=1}^{D} Z_d^2$. In the case when the traits are uncorrelated, the resulting test is approximately equivalent to the variance component score test of $H_0 : \sigma_\delta^2 = 0$ for the model (2.18) where $\delta_m \sim N(0, \sigma_\delta^2 I)$. In simulations, we evaluate significance for this test statistic by Monte Carlo.

3.1.2 Minimum P-Value

The minimum p-value [21] test statistic is $T_{minP} = \pi_{(1)} = \min_{1 \leq d \leq D} \pi_d$. In simulations, rather than evaluating significance by applying a conservative Bonferroni correction, we

instead use Monte Carlo.

3.1.3 FDR

Another approach would be to use the Benjamini-Hochberg (BH) false discovery rate (FDR) procedure to discover SNP-trait pairs at some false discovery rate α^*. In order to compare FDR to ELL and the other tests, we apply FDR to the D SNP-trait pairs corresponding to each SNP, and then we "reject" if we make any discoveries at FDR level α^* and do not reject otherwise. This corresponds to a test statistic of $T_{FDR} = \min_d \frac{\pi_{(d)}}{d/D},\ d = 1, \ldots, D$. This is the same global test statistic used in stage 1 of the testing procedure of Peterson et al. [18]. Because the BH procedure guarantees type 1 error α^*, we could choose to set α^* equal to the global test level α that we are using for the other tests. However, the BH procedure tends to be conservative, so to avoid disadvantaging FDR in the comparison, we choose a less conservative α^*, i.e., $\alpha^* > \alpha$, where α^* is chosen by Monte Carlo so that the empirical type 1 error of our FDR procedure is α.

3.1.4 CPMA

The idea behind the CPMA test [13, 4] is that if the expression traits in the set were independent, then the set of -log p-values for the association tests between the SNP and each trait would be i.i.d. Exp(1) random variables under the global null hypothesis. To calculate the CPMA test statistc, one models the set of -log p-values as i.i.d. Exp(λ) under the alternative model and sets the CPMA statistic to be the likelihood ratio chi-squared test statistic for testing $H_0 : \lambda = 0$ vs. $H_A : \lambda > 0$ in the i.i.d. case. In the case when the expression traits are correlated, the test statistic remains the same, namely

$$T_{CPMA} = -2 \cdot \frac{\log\left\{\prod_{d=1}^{D} f_{\mathrm{Exp}(1)}(-\log \pi_d)\right\}}{\log\left\{\prod_{d=1}^{D} f_{\mathrm{Exp}(\hat{\lambda})}(-\log \pi_d)\right\}}$$

where $f_{\text{Exp}(\lambda)}$ is the pdf of an Exponential(λ). The authors [13] propose using a Monte Carlo scheme to simulate the null distribution of the CPMA statistic under a given estimate of Ω allowing for calculation of p-values of the statistic. However, code is provided by the authors only for the calculation of the statistic itself and not for the p-value calculation. We assess significance using our own Monte Carlo procedure, though we use their code for calculation of the statistic.

For Sum of Z-Squared, minP, FDR, and CPMA we use Monte Carlo to simulate the test statistic's null distribution. For simplicity we illustrate the Monte Carlo method for CPMA, as the other three are analogous. We simulate replicates $\tilde{Z}_r$, $r = 1, ..., R = 10^6$ iid from $N(0, \hat{\Omega})$ with corresponding p-value vectors $\tilde{\pi}_r$. Given a vector Z with corresponding p-value vector π we calculate $T_{CPMA}(\pi)$ and estimate its global p-value by comparing it to $\{T_{CPMA}(\tilde{\pi}_r),\, r = 1, ..., R\}$.

3.1.5 Generalized Higher Criticism

The higher criticism [6, 7] statistic, first suggested by Tukey [23], is based on the order statistics for the p-values. The idea is to find the order statistic that is most extreme, relative to what would be expected for that order statistic under the global null hypothesis. Similarly to what we proposed for the ELL statistic, only a certain fraction of the smallest p-values are considered in constructing the higher criticism statistic. Donoho and Jin [7] consider this fraction to be a tuning parameter with a default value of 50%, and this value is used in the equation below, i.e., only the smallest $D/2$ p-values are considered.

$$HC = \max_{d \in \{1,...,D/2\}} \sqrt{D} \frac{d/D - \pi_{(d)}}{\sqrt{\pi_{(d)}(1 - \pi_{(d)})}}.$$

Barnett et al. [1] extend the higher criticism statistic to incorporate dependence, i.e. to the case $\Omega \neq I$, and they call the resulting test "generalized higher criticism" (GHC) with test

statistic

$$\text{GHC} = \max_{d \in \{1,\ldots,D/2\}} \left\{ \left. \frac{d - D\pi_{(d)}}{\sqrt{\widehat{\text{var}}(S(t))}} \right|_{t=|Z|_{(D+1-d)}} \right\}.$$

It is possible to analytically derive an asymptotic null distribution of the GHC. However, the rate of convergence is slow. Barnett et al. [1] instead propose a method for calculating approximate p-values for GHC using a Beta Binomial approximation to $S(t)$ similar to that used in (2.5). In our simulations we use the software provided by the authors in the GHC R package for calculating the GHC statistic and deriving corresponding p-values.

3.1.6 Generalized Berk-Jones

Berk and Jones [3] introduced an order statistic based test — referred to as the Berk-Jones test — that is an asymptotic approximation to equal local levels for the case when $\Omega = I$.

$$BJ = \max_{1 \le d \le D/2} \left\{ \pi_{(d)} \log\left(\frac{\pi_{(d)}}{d/D}\right) + (1 - \pi_{(d)}) \log\left(\frac{1 - \pi_{(d)}}{1 - d/D}\right) \right\}$$

Sun and Lin [22] showed that the BJ statistic can be rewritten as a likelihood ratio

$$\max_{1 \le d \le D/2} \log \left\{ \frac{\Pr\{S(|Z|_{(D-d+1)}) = d \mid E(Z) = \widehat{\mu}_d \mathbf{1},\ \text{cov}(Z) = \mathbf{I}\}}{\Pr\{S(|Z|_{(D-d+1)}) = d \mid E(Z) = \mathbf{0},\ cov(Z) = \mathbf{I}\}} \right\} 1\left\{ \pi_{(d)} < \frac{d}{D} \right\}$$

where Z is assumed to be multivariate normal and where $\hat{\mu}_d > 0$ solves the equation: $d/D = 1 - \{\Phi(|Z|_{(D-d+1)} - \widehat{\mu}_d) - \Phi(-|Z|_{(D-d+1)} - \widehat{\mu}_d)\}$. This leads to a natural extension to the case $\Omega \neq I$, resulting in a test they call "Generalized Berk-Jones" (GBJ), with test statistic

$$GBJ = \max_{1 \le d \le D/2} \log \left\{ \frac{\Pr\{S(|Z|_{(D-d+1)}) = d \mid E(Z) = \widehat{\mu}_d \mathbf{1},\ \text{cov}(Z) = \mathbf{\Omega}\}}{\Pr\{S(|Z|_{(D-d+1)}) = d \mid E(Z) = \mathbf{0},\ cov(Z) = \mathbf{\Omega}\}} \right\} 1\left\{ \pi_{(d)} < \frac{d}{D} \right\}.$$

In order to approximate the probabilities in the numerator and denominator, Sun and Lin use the extended beta binomial distribution (where the extended beta binomial differs from

the beta binomial in that it allows under-dispersion as well as over-dispersion) with parameters fit by the method of moments. They use a Markov Chain approximation for p-value calculation, as described in section 2.3. The authors give evidence that the GBJ statistic has advantages over higher criticism when the number of $Z_1, ..., Z_D$ with non-zero mean is small. In our simulations we use the GBJ R package code provided by the authors for calculation of GBJ statistics and the corresponding p-values.

3.2 Methodology

For our simulations we use $N = 1200$ individuals, $D = 1200$ or 10,000 traits, and $c = 1$, corresponding to an intercept term and no covariates in the linear model. SNPs are simulated as bi-allelic, and simulated values of a SNP are in the set $\{0, 1, 2\}$ corresponding to whether the individual has 0, 1 or 2 copies of the reference allele. We simulate genotypes using gene dropping based on 60 pedigrees of size 20, each with the same structure, where the pedigree of size 20 is a random subset from the pedigree of the AIL mice used in the data application in section (4). In the AIL mouse data, the founders of the pedigree are members of inbred lines, so the minor allele frequency of every SNP in the dataset is .5 in the founders, so we use this feature in our gene dropping simulation also.

We use the software Limix [14] to fit linear mixed models (LMMs). Suppose we have a length N Gaussian vector Y of expression levels of a given trait, an $N \times M$ matrix G where each row represents one subject's genotype for M SNPs. For a given (SNP, trait) pair, Limix uses the univariate model

$$Y = \mu + G_i\beta + \sigma(\zeta U + (1 - \zeta)\epsilon) \tag{3.1}$$

where G_i is the ith column of G; $-\infty < \mu < \infty$, $-\infty < \beta < \infty$ $\sigma \geq 0$, and $\zeta \in [0, 1]$ (the heritability) are unknown scalar parameters; $U \sim N(0,\ K)$ and $\epsilon \sim N(0,\ I)$. For each

$i = 1, ..., M$, Limix calculates the score test statistic, where the maximum likelihood estimate of ζ is obtained under null model in which $\beta = 0$ and his held constant across i with only σ being re-fit for each i. The $\hat{\beta}/se(\hat{\beta})$ estimated by Limix for a given SNP has a Student's t_{n-1} distribution under the null hypothesis of $\beta = 0$, so we convert it to the standard normal value of the same percentile to get a Z score. We run Limix on each (SNP, trait) pair and concatenate the resulting Z scores to get a $M \times D$ matrix.

We also use Limix to compute the matrix K from a matrix G using the formula

$$K = \frac{1}{M} \sum_{i=1}^{M} (G_i - g_i) (G_i - g_i)^T / s_i^2 \tag{3.2}$$

where g_i and s_i are the column mean and standard deviation of G_i respectively. Recent studies have shown that including candidate variants in the subset used for estimating the GRM can lead to loss of power [26]. Therefore, in our data analysis in chapter 4, we use the leave-one-chromosome-out (LOCO) method [26], and in our simulations, we do not include any tested SNPs in the GRM.

3.3 Type 1 Error Validation

For validating the type 1 error of ELL and the other above-mentioned statistics we generate a matrix Y according to (2.18) but with $\delta_m = 0$. In one set of simulations we set D=1200 and and in another, D =10,000. To maintain realistic parameter values we set ρ_u to the $D \times D$ empirical correlation matrix for the first D columns and first 104 rows of the 208×14461 matrix of observed AIL mouse gene expression levels from chapter 4. ρ_e is similarly formed but using the bottom 104 rows. The diagonals of the matrix ζ are drawn iid from a Beta(1, 10) distribution — this approximately matches the empiric distribution found in [17]. 100,000 SNPs are simulated, 50,000 for estimating K and the rest for testing

association. η, the diagonal matrix of trait variances, is set to be I_D, the $D \times D$ identity matrix. Limix was run as above to generate an $M \times D$ matrix Z, where $M = 50,000$. $\hat{\Omega}$ was estimated as the $D \times D$ empirical correlation matrix of the columns of Z. For each of the $M = 50,000$ tested SNPs, we calculated a test statistic for each method considered.

3.3.1 Type 1 Error Validation Results

First, we found that the GHC and GBJ testing methods were far too slow for us to be able to complete type 1 error (or power) simulations for them in a reasonable amount of time in the two settings we consider, namely, $D = 1200$ and D=10,000. This can be seen in section (3.5) in which we benchmark the methods based on computation time. For example, on a single processor, the GBJ method took over 15 hours per SNP to analyze data in which D=2,000. GHC was even slower. Therefore, we were not able to obtain type 1 error or power results for these methods.

For the remaining 6 methods, ELL, ELL-Recursion, sumZ^2, minP, FDR and CPMA, for the case when D=1200, Table (3.1) shows empirical Type 1 error for each method at levels .01, .001 and .0001. From the Table, we can see that ELL-Recursion had type 1 error that differed significantly from the nominal in both cases, while the other 5 methods, which all had significance assessed based on Monte Carlo, all had type 1 error not significantly different from the nominal level. Similar conclusions can be drawn from Figure (3.1), which shows QQ plots of the p-values from each method on the the 50,000 SNPs, with simultaneous 95% confidence bands included for reference.

Based on these type 1 error results for $D = 1200$, we did not consider further the Markov chain approximation to assess significance for ELL, i.e., ELL-Recursion. From here on we only consider the Monte Carlo p-value method (ELL). In particular, for the type 1 error

α	ELL	CPMA	sumZ^2	FDR	minP
10^{-4}	0.00012 (2e-05)	0.00011 (2e-05)	0.0001 (2e-05)	0.00012 (1e-05)	0.00012 (1e-05)
10^{-3}	0.00111 (7e-05)	0.00091 (6e-05)	0.00092 (5e-05)	0.00108 (5e-05)	0.00109 (4e-05)
10^{-2}	0.00983 (0.00014)	0.00941 (0.00012)	0.00942 (0.00013)	0.01018 (0.00017)	0.01024 (0.00016)

Table 3.1: Empirical Type 1 error for 6 different methods based on 50,000 SNPs averaged over 10 simulations. Standard error in parenthesis. Bold denotes empirical type 1 error that is significantly different from the nominal level ($p < .05$).

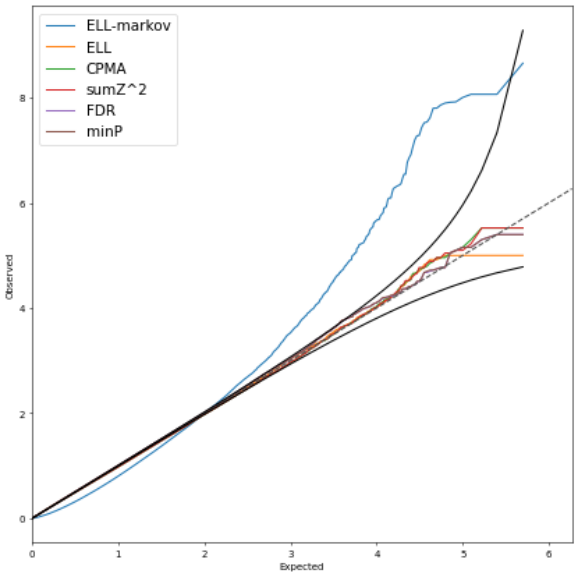

Figure 3.1: Combined QQ Plots of p-values for each of 6 methods for a set of 50,000 SNPs, where $D = 1{,}200$ expression traits over 10 simulations. For each method and each simulation, the 10,000 ordered p-values are plotted versus their expected values under the null hypothesis. The dashed line represents the expectation, and the solid lines are 95% simultaneous confidence bands under the null hypothesis.

simulation for D =10,000 we only use ELL and not ELL-Recursion.

The case when $D = 10{,}000$ is potentially of interest, because in that case D is much larger than the sample size, which is only $N = 1200$ (while the number of SNPs is M=50,000). As can be seen from Table (3.2) and Figure (3.2), the 5 methods that use the Monte Carlo assessment of significance all have type 1 error not significantly different from the nominal level, in agreement with the results for $D = 1200$.

α	ELL	CPMA	sumZ^2	FDR	minP
0.0001	0.0001 (1e-05)	0.0001 (1e-05)	0.0001 (1e-05)	0.0001 (0.0)	0.0001 (0.0)
0.001	0.00094 (4e-05)	0.00087 (5e-05)	0.00087 (5e-05)	0.001 (5e-05)	0.001 (4e-05)
0.01	0.01005 (0.00018)	0.00923 (0.00019)	0.00922 (0.00017)	0.01069 (0.00015)	0.0107 (0.00015)

Table 3.2: Empirical Type 1 error for 5 different methods based on 10,000 SNPs averaged over 10 simulations. No values in the table were significantly different from the corresponding nominal level ($p > .05$ in each case).

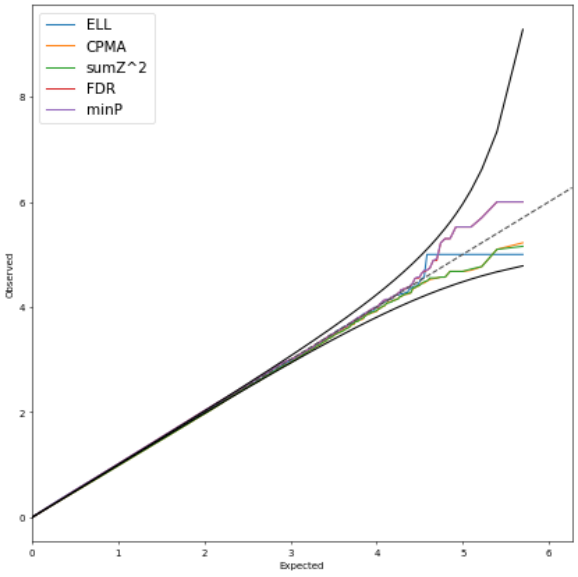

Figure 3.2: Combined QQ Plot of p-values for each of 5 methods for a set of 50,000 SNPs, where D = 10,000 expression traits over 10 simulations. For each method and each simulation, the 10,000 ordered p-values (each corresponding to a SNP) are plotted versus their expected values under the null hypothesis. The dashed line represents the expectation, and the solid lines are 95% simultaneous confidence bands under the null hypothesis.

3.4 Power analysis

Under the alternative hypothesis, we assume that a given trans eQTL is associated with a set of traits $\mathcal{D}^*$, where we fix $|\mathcal{D}^*|$ to some pre-specified value a. Given $|\mathcal{D}^*| = a$, the a traits in $\mathcal{D}^*$ are chosen at random, independently for each SNP. The effect of the trans eQTL on each trait in $\mathcal{D}^*$ is assumed to be $\pm c_a/\sqrt{2f(1-f)}$, where c_a is a positive constant that depends on the choice of a, f is the minor allele frequency (MAF) of the the SNP, and the sign of the effect is chosen at random.

To save time in the simulations, we actually first generate a data set under the null hypothesis, and then, for each simulated trans eQTL, we start from the null data set and add only the effects of that one trans eQTL to the traits in $\mathcal{D}^*$, and then analyze the resulting data set. Thus, each trans eQTL is effectively analyzed in a data set in which it is the only trans eQTL. For each null data set, we simulated and analyze 1,000 trans eQTLs this way. To ensure reliability, we repeat the entire process 10 times, so a total of 10,000 trans eQTLs are simulated and analyzed for each simulation setting. We try several different sizes of $|\mathcal{D}^*|$: 1, 2, 4, 10, 50, 150, 500. For each of the ten replicates, we calculate power under each choice of $|\mathcal{D}^*|$.

3.4.1 Power Results

Table (3.3) and figure (3.3) show the results of the the power simulation for D=1,200. Each row corresponds to a particular value of $|D^*|$, the number of associated traits for each trans eQTL. The left-most column of the table gives the effect size values, c. Not surprisingly, when only 1 expression trait is associated, minP is the most powerful. For 2 associated traits, FDR becomes more powerful, and then already by 4 associated traits, ELL is far more powerful than any other method and remains so until the number of associated traits reaches 500, which represents 42% of all the traits being associated with the trans eQTL. At

c	$\lvert\mathcal{D}^*\rvert$	ELL	CPMA	sumZ^2	FDR	minP
3.14	1	0.019(0.0009)	0.011(0.0008)	0.011(0.0008)	0.623(0.0241)	**0.762**(0.013)
3.125	2	0.581(0.0449)	0.0120.0008)	0.012(0.0007)	**1.000**(0)	0.662(0.0222)
2.89	4	**0.789**(0.0617)	0.012(0.0008)	0.013(0.0009)	0.024(0.0016)	0.012(0.0009)
2.568	10	**0.688**(0.0634)	0.015(0.0009)	0.015(0.001)	0.012(0.0007)	0.012(0.0008)
2	50	**0.756**(0.0421)	0.030(0.0015)	0.031(0.0014)	0.012(0.0007)	0.011(0.0007)
1.53	150	**0.746**(0.0212)	0.061(0.0035)	0.056(0.0030)	0.011(0.0008)	0.011(0.0009)
1.3	500	0.476(0.1227)	**0.588**(0.0203)	0.428(0.0178)	0.010(0.0009)	0.010(0.0009)

Table 3.3: Empirical power with standard error in parentheses for each of 5 methods for $D = 1200$ traits and various choices of the number of associated traits for each trans eQTL ($|\mathcal{D}^*|$). Each simulation study is replicated 10 times with 1,000 trans eQTLs per replicate, for a total of 10,000 simulated trans eQTLs. Cell contents are the average power at 0.01 level over 10 experimental replicates. S.E. in parentheses calculated as the standard deviation divided by the square root of 10. c is the effect size. Note that power is only comparable across methods within one row of the table, not between different rows of the table.

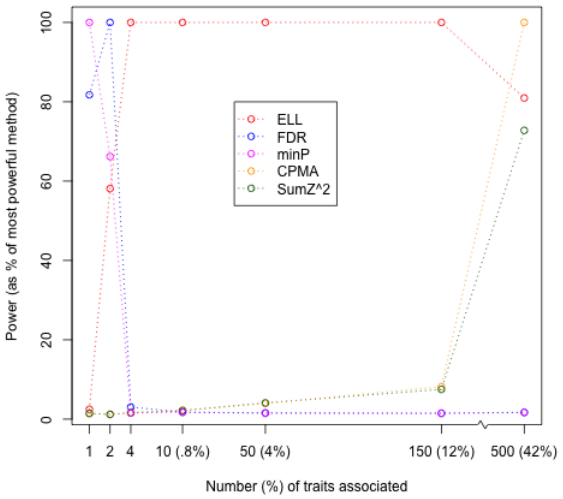

Figure 3.3: Empirical Power as a percentage of the maximum observed power vs. the number of traits associated with each trans eQTL, for each of 5 methods, for 1200 traits and various choices of the number of associated traits for each trans eQTL. Each simulation study is replicated 10 times with 1,000 trans eQTLs per replicate, for a total of 10,000 simulated trans eQTLs. Empirical power is assessed at level .01

c	$\lvert\mathcal{D}^*\rvert$	ELL	CPMA	sumZ^2	FDR	minP
3.433	1	0.0124(0.001)	0.0108(0.0007)	0.0111(0.0007)	0.4877(0.0171)	**0.6784**(0.0084)
3.357	2	0.015(0.0013)	0.011(0.0008)	0.0112(0.0008)	**0.8006**(0.0141)	0.0122(0.0017)
3.263	4	0.0362(0.0038)	0.0112(0.0008)	0.0114(0.0008)	**0.8272**(0.0166)	0.0118(0.0017)
3.076	10	**0.3474**(0.0782)	0.0115(0.0008)	0.0116(0.0008)	0.0151(0.0016)	0.0118(0.0017)
2.634	50	**0.3901**(0.0794)	0.0135(0.001)	0.014(0.001)	0.0119(0.0017)	0.0118(0.0017)
2.299	150	**0.4882**(0.0648)	0.0178(0.0012)	0.0188(0.0014)	0.0117(0.0017)	0.0116(0.0017)
1.899	500	**0.8429**(0.0288)	0.0345(0.0015)	0.036(0.0016)	0.0114(0.0016)	0.0114(0.0017)

Table 3.4: Empirical power with standard error in parentheses for each of 5 methods for D = 10,000 traits and various choices of the number of associated traits for each trans eQTL ($|\mathcal{D}^*|$). Each simulation study is replicated 10 times with 1,000 trans eQTLs per replicate, for a total of 10,000 simulated trans eQTLs. Cell contents are the average power at 0.01 level over 10 experimental replicates. S.E. in parentheses calculated as the standard deviation divided by the square root of 10. c is the effect size. Note that power is only comparable across methods within one row of the table, not between different rows of the table.

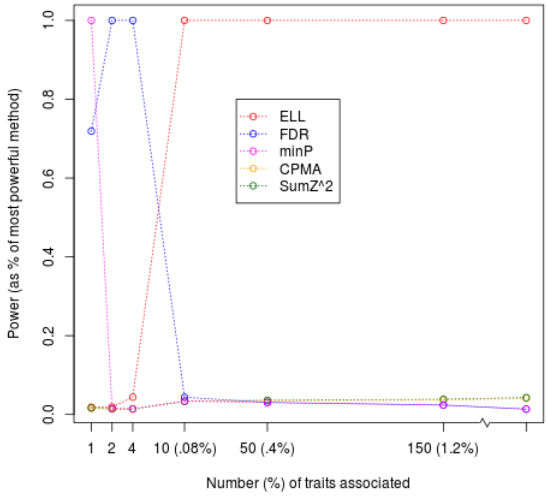

Figure 3.4: Empirical Power as a percentage of the maximum observed power vs. the number of traits associated with each trans eQTL, for each of 5 methods, for 10,000 traits and various choices of the number of associated traits for each trans eQTL. Each simulation study is replicated 10 times with 1,000 trans eQTLs per replicate, for a total of 10,000 simulated trans eQTLs. Empirical power is assessed at level .01

that point, ELL, sumZ^2 and CPMA are all similarly powerful, with CPMA being the most powerful. It is interesting to note that we have chosen to use ELL with only the top 20% of p-values considered, so that might have reduced its power somewhat when 42% of the traits are associated. The results indicate that ELL dominates the majority of the range of $|D^*|$ providing compelling evidence of the high power of ELL over a large range of scenarios.

The fact that minP and FDR tend to have similar power across scenarios and that CPMA and sumZ^2 tend to have similar power across scenarios should not be surprising. The former pair are quite similar in construction and inherently focused on the smallest p-values. CPMA and sumZ^2 are both based on sums of the association statistic, CPMA uses $\hat{\lambda} = \frac{1}{D}\sum_{i=d}^{D}\log(\pi_d)$, while sum$Z^2$ uses $\sum_{d=1}^{D} Z_d^2$. The use of summation is inherently more sensitive to higher density effects that are not lost in the respective averaging.

Table (3.3) and figure (3.3) show the results of the the power simulation for D=10,000. These results mimic the ones for $D = 1,200$. The ELL dominates except for the scenarios with a very small number of very strong non-null signals. We conjecture that if we were to consider $|D^*|$ much larger than 500 for the $D = 10,000$ case, we would eventually see that CPMA and sumZ^2 would come to have higher power compared to the other methods.

3.5 Computation Time

A key requirement for a method to be useful for eQTL detection is reasonable computation times. It is common for eQTL detection studies to consider thousands or tends of thousands of genes and hundreds of thousands of SNPs [11, 25, 8]. We compared the computation time of the ELL method to those of the GBJ and GHC. In our scenario, we applied each method on three vectors Z_1, Z_2 and Z_3 each of length 2000. They were all drawn from the same MVN distribution with mean 0 and correlation matrix taken from the empirical correlation matrix for the gene transcript levels of 2000 genes in the AIL data in chapter 4. We run the

methods assuming a fixed Ω.

For GBJ and GHC, Table (3.5) gives the average time per SNP to calculate a p-value. The ELL has an additional upfront pre-compute time which is the time to compute the function in equation (2.15) for all $h \in \mathcal{H}$ and $d \in \mathcal{D}$. This pre-compute time, given in column 2 of the table, is the same no matter how many SNPs are analyzed, and the additional time per SNP to compute a p-value is given in column 3. For each entry in the 3rd column, the time per SNP is the average over three replicates. We can see that for D=2,000 expression traits, GBJ takes over 15 hours per SNP to obtain a p-value, and GHC is even slower. In contrast, the ELL is much faster and can be run in a reasonable amount of time, even for larger numbers of traits and SNPs.

Method	Pre-compute Time (minutes)	Time Per SNP (minutes)
ELL	17.12	0.003
GBJ	0	937.52
GHC	0	1781.93

Table 3.5: Computation Times for ELL, GBJ, GHC using $D = 2000$. Cell values are time in minutes. The ELL has an additional upfront pre-compute time which is the same no matter how many SNPs are analyzed (column 2), and the additional time to compute a p-value is given in column 3. For each entry in the 3rd column, the time per SNP is the average over three replicates.

As a second experiment we compare the computation time of the ELL method against the other methods used in our simulations: FDR, minP, sumZ^2, CPMA. We did this using two different values for D:1200 and 10,000. In each case we drew $Z_i \sim MVN(0, \Omega)$ iid $i = 1, ..., 10^6$ for a fixed Ω and calculated the average time per SNP and the computation time for calculating p-values for the 10^6 Z vectors.

Method	Pre-compute Time (minutes)	Average Compute Time (minutes)
ELL	14.1	40.30
CPMA	0	2210.3
minP	0	37.5
sumZ^2	0	250.5
FDR	0	44.4

Table 3.6: Average time for each of ELL, CPMA, minP, sumZ^2, and FDR to be computed on 10^6 snps with $D = 10,000$. Pre-compute time refers to the time to complete the pre-computation for the ELL method. 10 experimental repetitions, each with 10^6 snps, were performed and total time to score the 10^6 snps averaged over the repetitions.

CHAPTER 4

APPLICATION TO AIL DATA

We apply the ELL method to map eQTLs on data from an Advanced Intercross Line (AIL) of mice [11]. The LG/J x SM/J advanced intercross line of mice (LG x SM AIL) is a multi-generational outbred population. High minor allele frequencies, a simple genetic background, and the fully sequenced LG and SM genomes make it a powerful population for genome-wide association studies.

The dataset contains originally 1067 mice of which 208 have gene expression data. The transcript levels of 15,071 genes were measured in the hippocampus region of each mouse, and each mouse was genotyped at 523,027 locations. Sex and batch number for each mouse were also recorded. The SNPs occur on each of the 19 autosomal chromosomes, and measurements for gene expression traits are present for all 19 chromosomes.

Imputation was performed on the data set by the authors [11] for missing observations. SNPs with very low MAF were removed. The preprocessing used by [11] included finding an eigen decomposition of the correlation matrix of the quantile-normalized expression traits, and then regressing 71 of the top 100 eigenvectors (referred to as principal components or "PCs") out of the expression data (where 29 of the top 100 PC's were excluded because they were associated with at least one SNP with nominal p-value $<$ 8.68e-6), but not regressing these PCs out of the SNP data. While regressing out a large number of latent factors might be reasonable for the task of identifying cis-eQTLs, it could be undesirable for identifying trans-eQTLs, because trans signal for trans-eQTLs regulating multiple genes could be expected to manifest as a latent factor, which would then be removed from the data.[4] Indeed, we found that this pre-processing resulted in candidate eQTLs being associated with only a handful of genes. Among the other concerns we had about the pre-processing, one was that with only 208 observations, when 71 covariates are regressed out of only the trait,

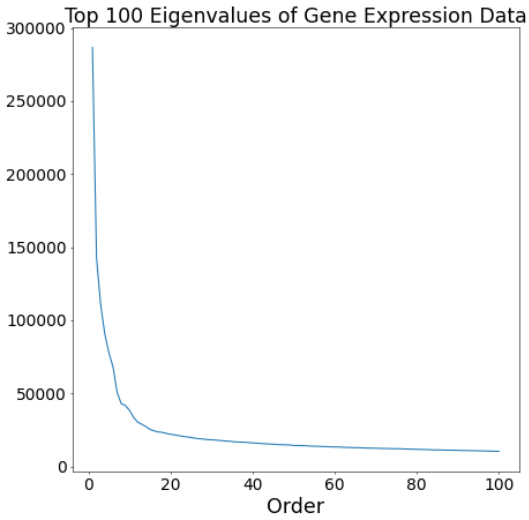

Figure 4.1: Eigenvalues from the processed AIL gene expression data

but not the predictor (SNP), prior to association mapping, that could result in inflation of p-values. As a result of these concerns and a few others, we decided to pre-process the data somewhat differently from [11]. We quantile normalized the gene expression data, and then regressed out sex, and batch. We then standardized the column variance of the residual. We then found the top 100 eigenvalues of the correlation matrix of the result, they are plotted in Figure (4.1). Based on this plot we chose to remove the first 10 PCs from the above residuals. The result of this step is used as our response matrix subsequently. We also removed sex, batch and the same 10 PCs from each SNP and used these SNP residuals in place of the SNPs in all subsequent analyses. The main reason that we removed all these covariates (sex, batch and 10 PCs) from both traits and SNPs as pre-processing, rather than including them in the LMM is that inclusion of covariates in the LMM slows down the LMM computations considerably, and with almost 8 billion LMMs to fit (one for each SNP-trait pair) it was more expedient to remove the covariates as a pre-processing step.

For each chromosome, we use GEMMA [28] to construct a "leave one chromosome out" (LOCO) estimate of the GRM, which constructs the GRM using SNPs from all other chro-

mosomes. For each gene, we use GEMMA to run a univariate LMM for that gene's expression levels against each SNP:

$$Y = \beta_0 + \beta_1 SNP + \sigma(\zeta U + (1-\zeta)\epsilon) \tag{4.1}$$

where Y is the gene's expression levels (actually residuals, as described above), SNP is the SNP genotype for each mouse (actually residuals, as described above), $U \sim N(0, K)$ where K is the LOCO GRM for the chromosome containing the given SNP, $\epsilon(0, I)$ and $\zeta \in [0, 1]$.

GEMMA estimates all parameters using maximum likelihood. We estimate ζ once per gene expression trait using the null model $\beta_1 = 0$ and then the same estimated ζ is used for each SNP. For each gene-SNP pair we record the statistic

$$Z = \Phi^{-1}\left[t_{204}\left(\frac{\hat{\beta}}{se(\hat{\beta})}\right)\right]$$

where Φ is the cdf of a standard normal and t_{204} the cdf of a student's t distribution with 204 degrees of freedom. This last step is done to normalize the test statistic.

For each SNP, we calculated an ELL pvalue and compared it to a "naive" p-value formed for each SNP by taking the minimum marginal p-value for association between it and each gene and multiplying that minimum by the number of genes considered as an adjustment. This approach mimics the strategy used in [11]. The authors there proposed a genome wide cutoff of 8.06×10^{-6} and we re-use that threshold in our analysis as well.

4.0.1 AIL Analysis Results

Four sets of analyses were performed and figures for each set are given below. Firstly, we produced Manhattan plots for SNPs on each of the 19 autosomal chromosomes. In these plots

we compare the Naive method and the ELL-pvalue. Figures 4.2 to 4.3 give the results. The p-values reported by ELL have a lower bound of 10^{-7}, which is determined by the number of Monte Carlo replicates used to assess significance, and in this data set the significant SNPs identified by ELL tend to have reported p-values right up against that boundary. Thus, if, for example, there is a nominal p-value between a SNP and a trait of 10^{-60}, then the naive p-value will obviously be much smaller than the ELL p-value. However, in that case, no fancy statistical method is needed to identify that eQTL. The proposed contribution of ELL would be for the cases when there is a trans-eQTL for which no single p-value is sufficiently small that it would be significant after correction for multiple comparison. From the Manhattan plots, it is clear that the ELL method tends to identify many additional candidate SNPs that do not have any single p-value that meets the genome-wide cut-off. To investigate this further, we apply the method of [18] to identify the expression traits associated with each significant SNP. Figures 4.4 to 4.5 show, for each significant SNP, the number of associated genes identified by the procedure of [18]. Interestingly, there are several trans-eQTLs identified by ELL that seem to each be associated with a large number of traits, where these trans-eQTLs cannot be identified by the naive method, because no single p-value for that trans-eQTL with any gene is sufficiently small.

Thirdly, We examined the overlap between the traits associated with different candidate eQTLs identified by the ELL method. Figures 4.6 to 4.10 show the results. Unsurprisingly, candidate eQTLs near to each other tend to have high overlap in which traits they are associated with. The overlap score between two candidate eQTLs is the ratio of the number of traits they are both associated with over the smaller of the number of traits each is associated with. We plotted the number of traits each candidate eQTL is associated with in black. Sharp differences in the number of traits two proximal eQTLs are associated with often coincides with the traits associated with one being a super set of that of the other.
Lastly, we examined the relationship difference between ELL and the Naive method in terms

of the distribution of the number of traits snps that were significant by each method were associated with. Figure 4.11 shows the results. For both methods we used the [18] method to determine the number of associated traits even if a SNP was significant by the Naive Method but not by ELL. It is clear that the SNPs detected by ELL tend to be associated with a larger number of traits. This combined with the results of the manhattan plots - where most regions found to be significant by the Naive Method were also significant by ELL - indicates that the ELL discovers additional candidate eQTLs that have a perhaps weaker association with a larger number of traits.

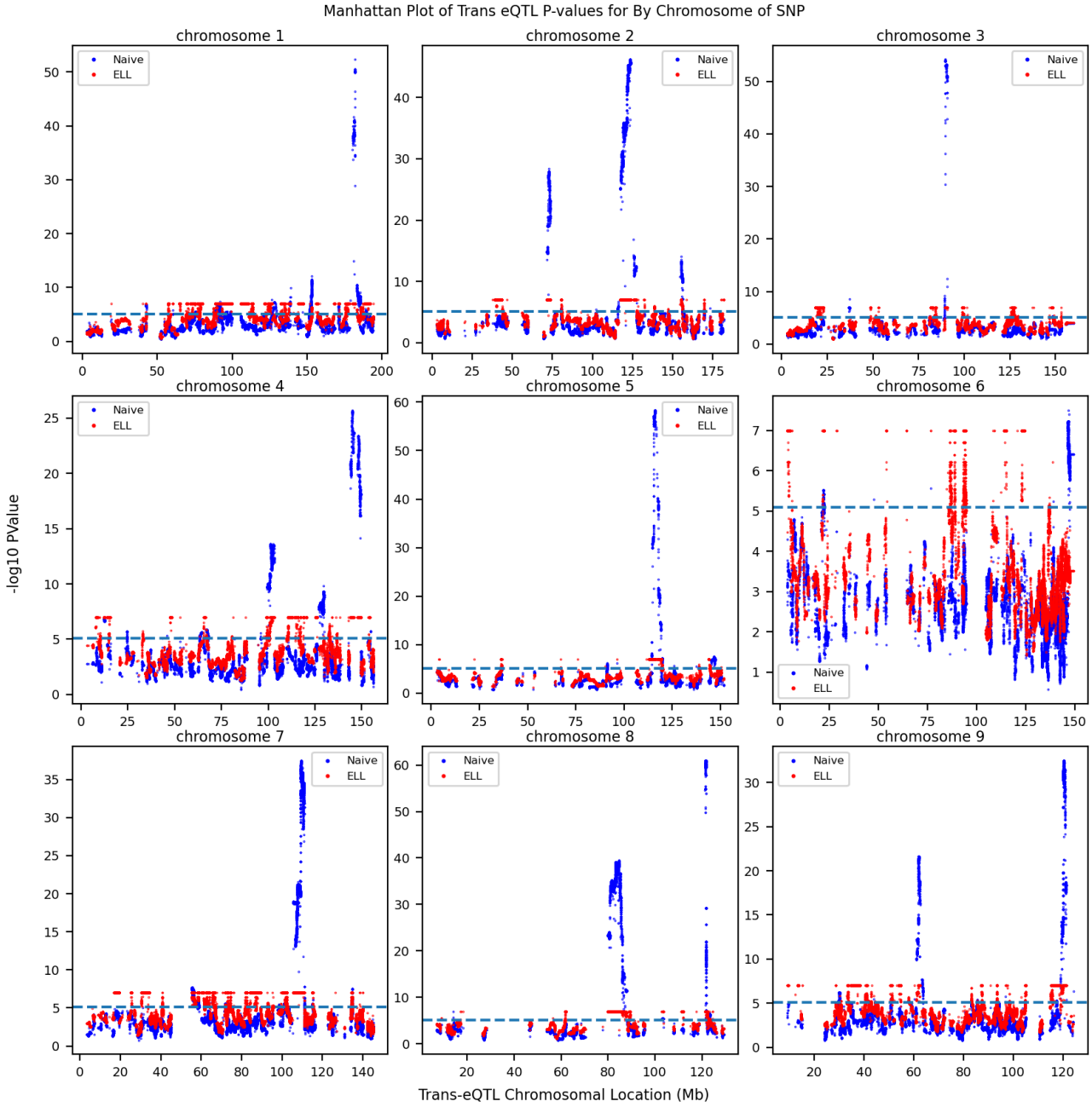

Figure 4.2: Manhattan plots for SNPs on the first 9 chromosomes of the AIL dataset. 15,071 gene expression levels analyzed. For each SNP location, the blue points are the minimum p-value from LMM of each trans-gene against that SNP adjusted for the number of genes. Red is the ELL p-value for the SNP. Horizontal dashed line is genome wide cutoff from [11] Points are jittered.

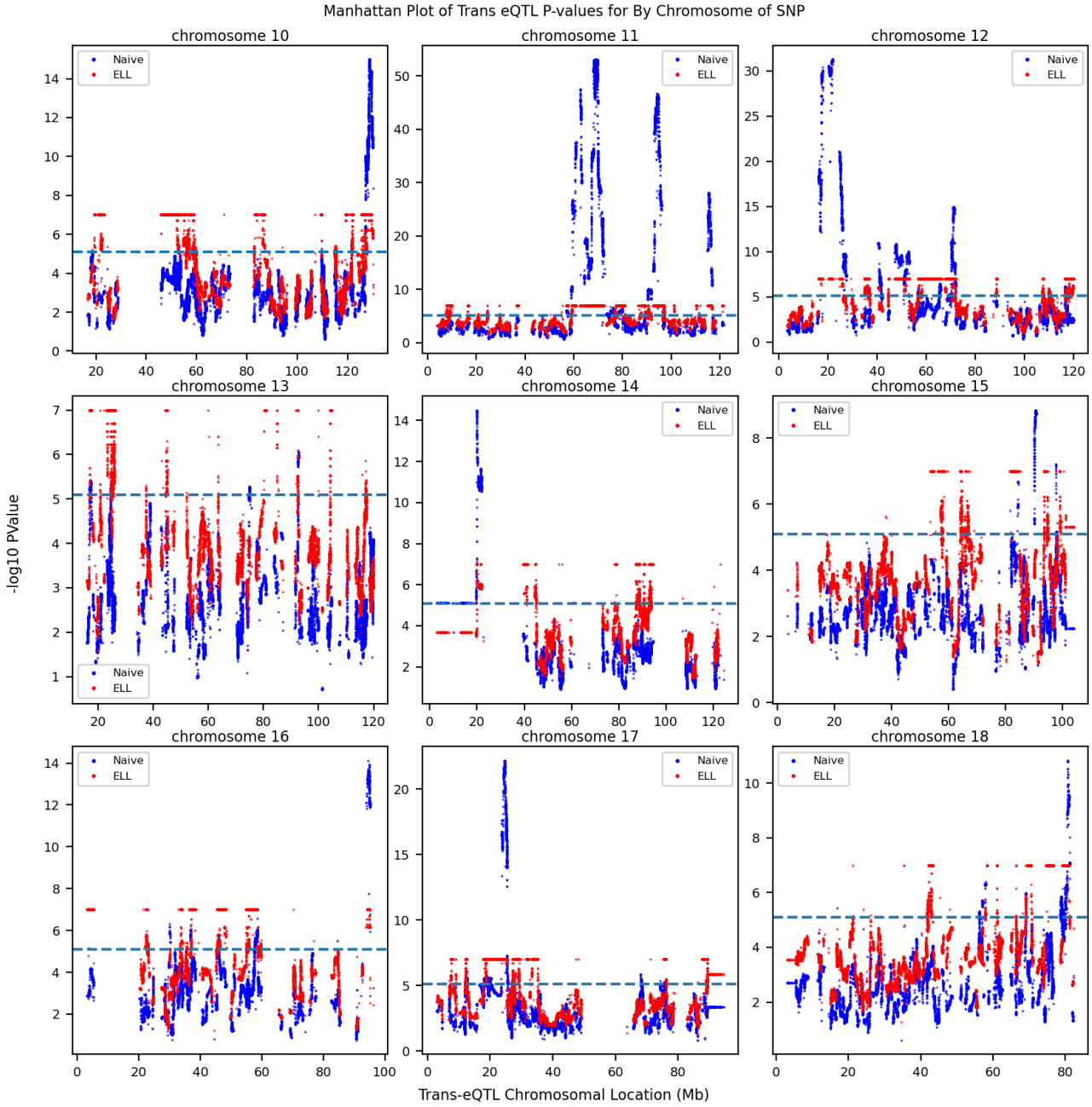

Figure 4.3: Manhattan plots for SNPs on chromosomes 10-18 of the AIL dataset. 15,071 gene expression levels analyzed. For each SNP location, the blue points are the minimum p-value from LMM of each trans-gene against that SNP adjusted for the number of genes. Red is the ELL p-value for the SNP. Horizontal dashed line is genome wide cutoff from [11] Points are jittered.

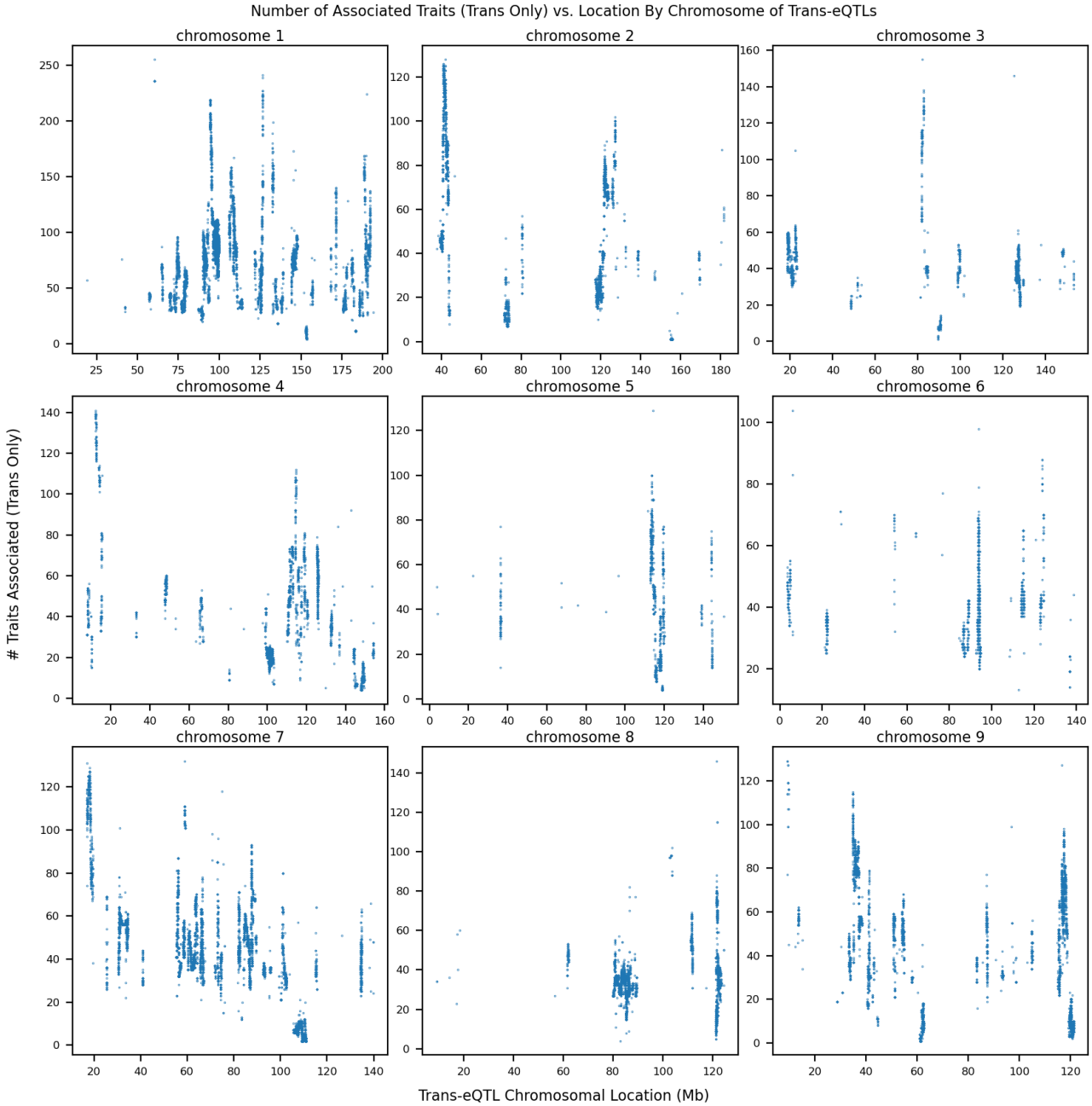

Figure 4.4: Number of expression traits associated with each significant trans-eQTL vs. trans-eQTL location for signficant trans-eQTLs located on the first 9 chromosomes of the AIL dataset. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered.

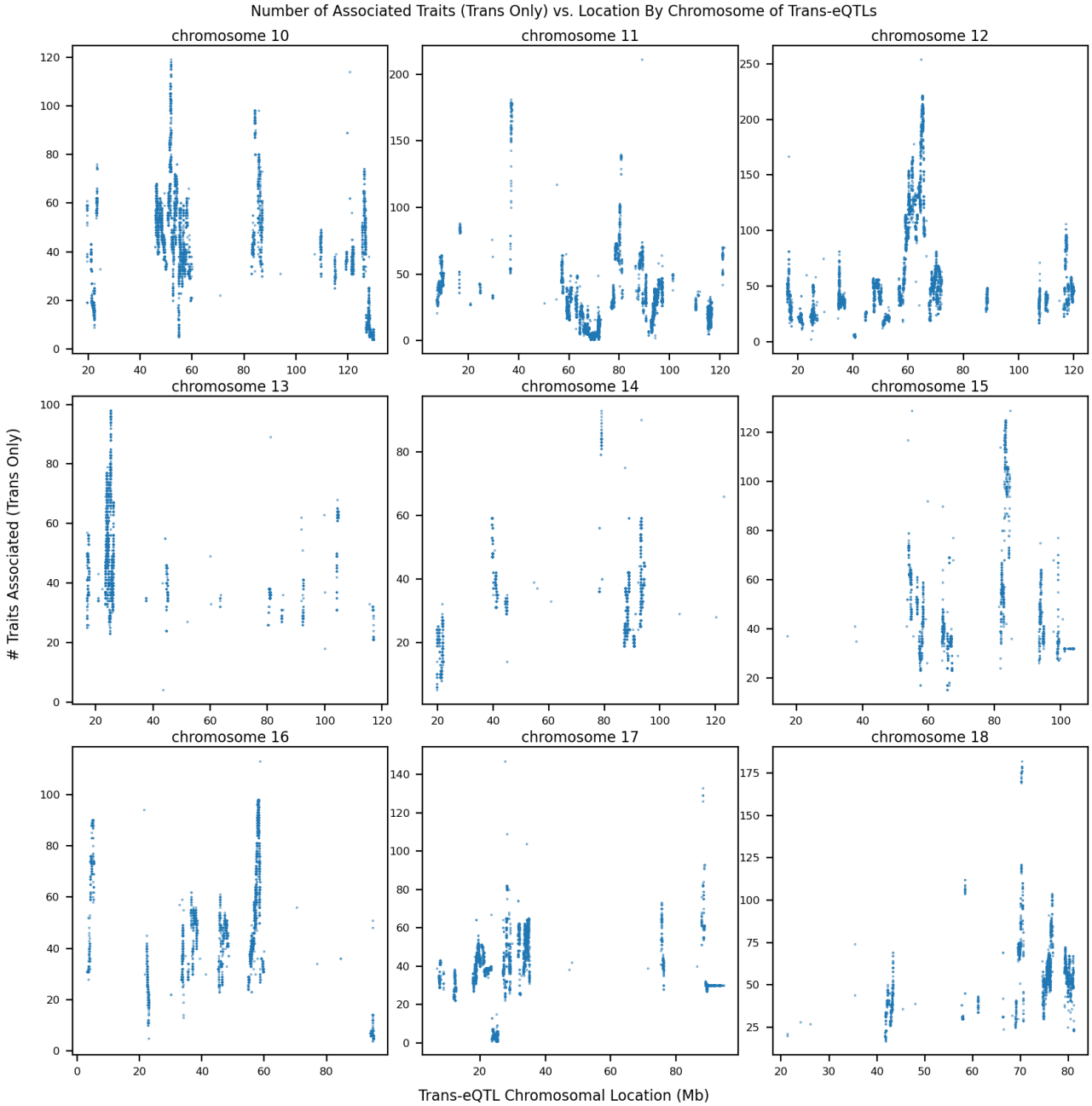

Figure 4.5: Number of expression traits associated with each significant trans-eQTL vs. trans-eQTL location for signficant trans-eQTLs located on chromosomes 10-18 of the AIL dataset. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered.

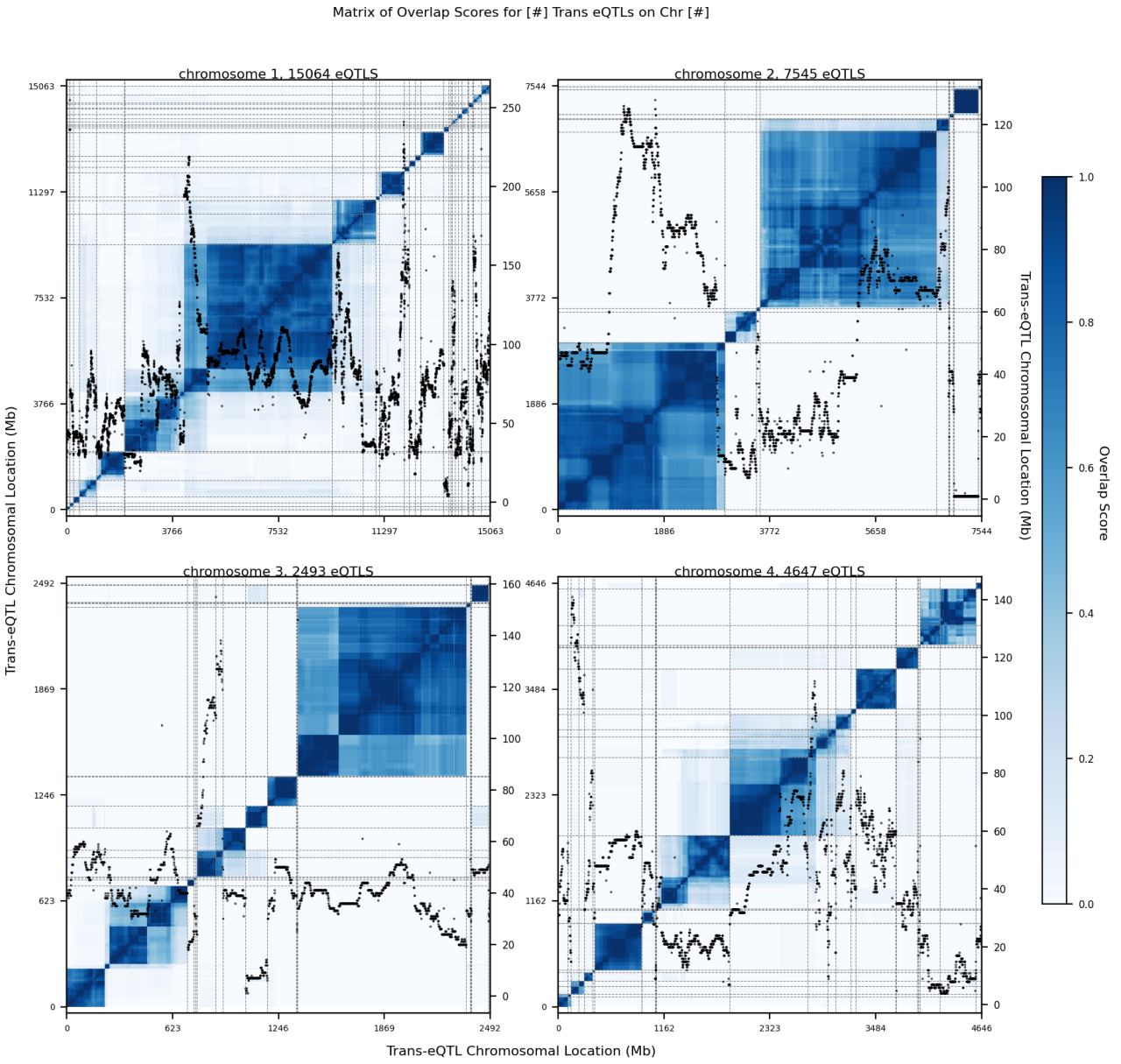

Figure 4.6: Chromosome wide overlap of eQTLs for eQTLs on the first 4 chromosomes. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered. Number of traits each SNP is associated with in black.

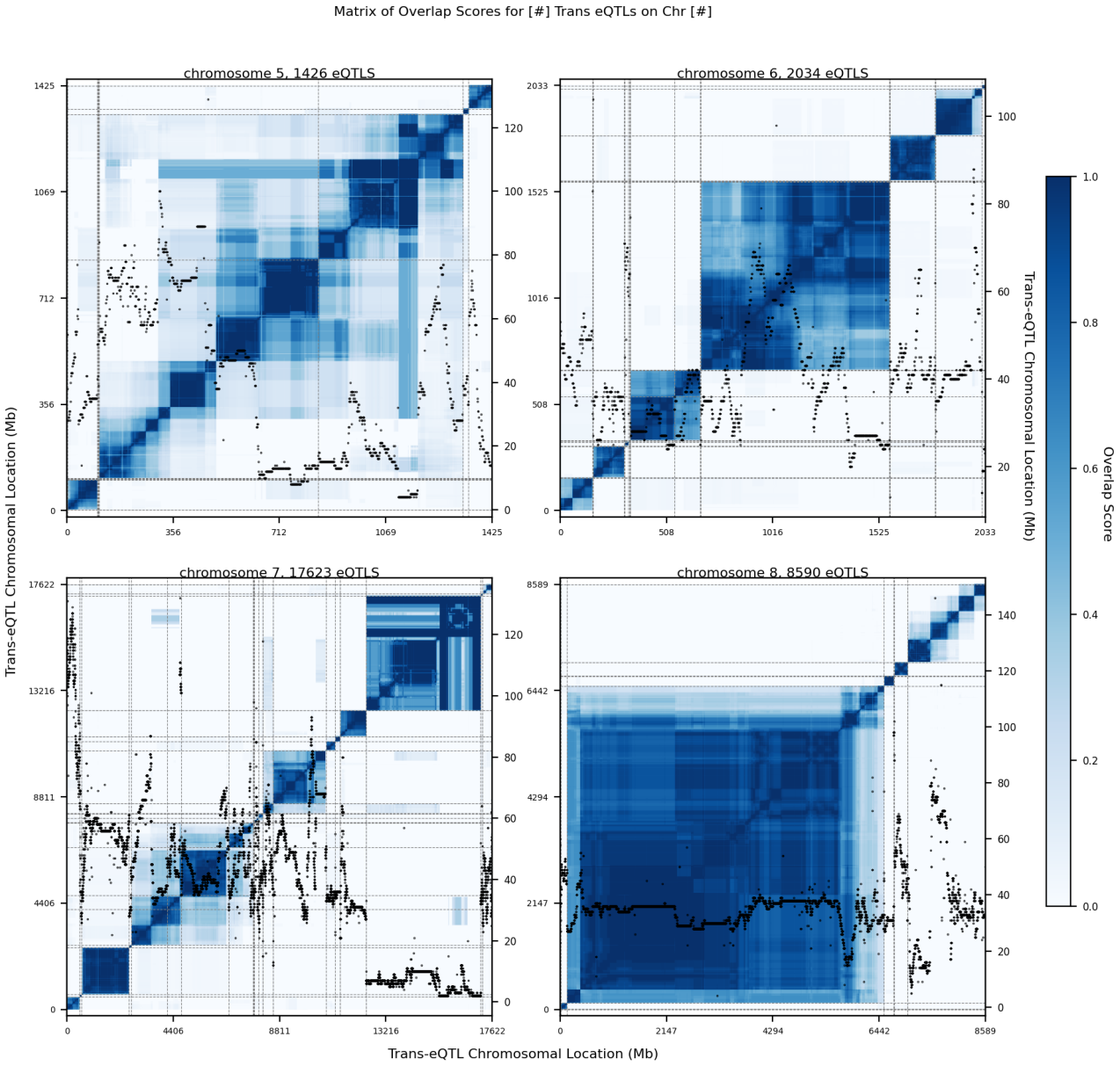

Figure 4.7: Chromosome wide overlap of eQTLs for eQTLs on the chromosomes 4-7. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered. Number of traits each SNP is associated with in black.

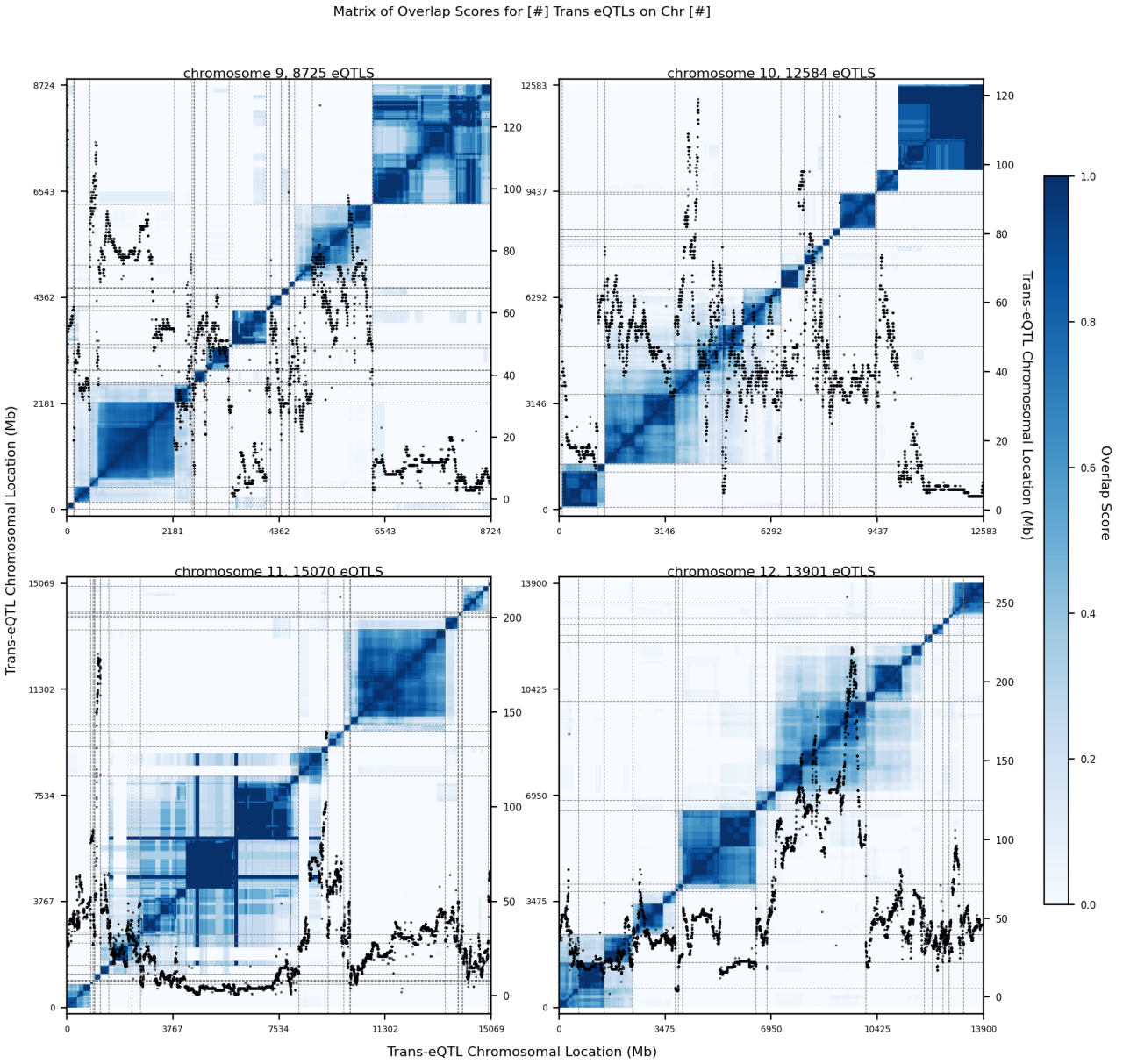

Figure 4.8: Chromosome wide overlap of eQTLs for eQTLs on the chromosomes 8-11. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered. Number of traits each SNP is associated with in black.

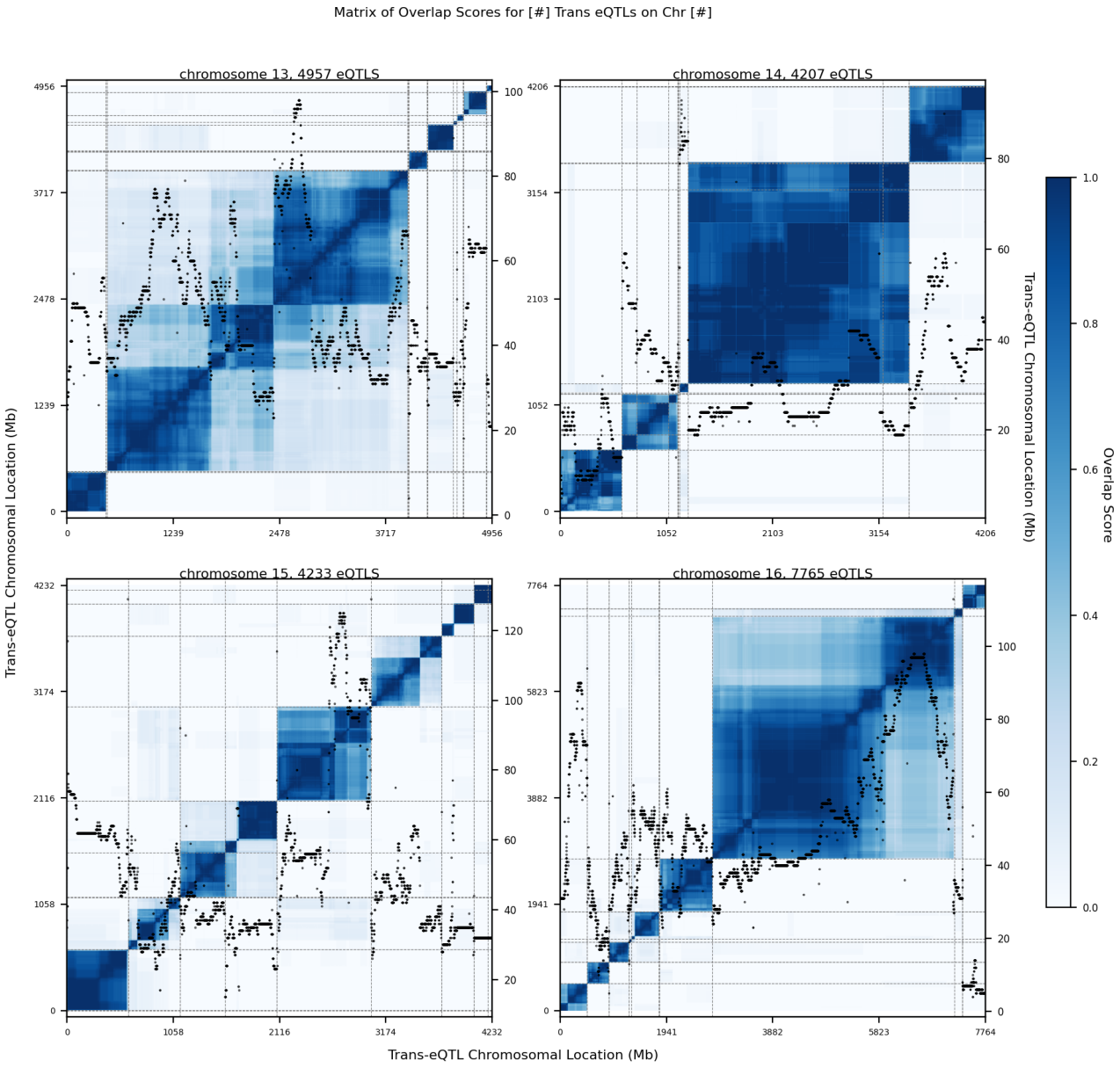

Figure 4.9: Chromosome wide overlap of eQTLs for eQTLs on the chromosomes 12-15. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered. Number of traits each SNP is associated with in black.

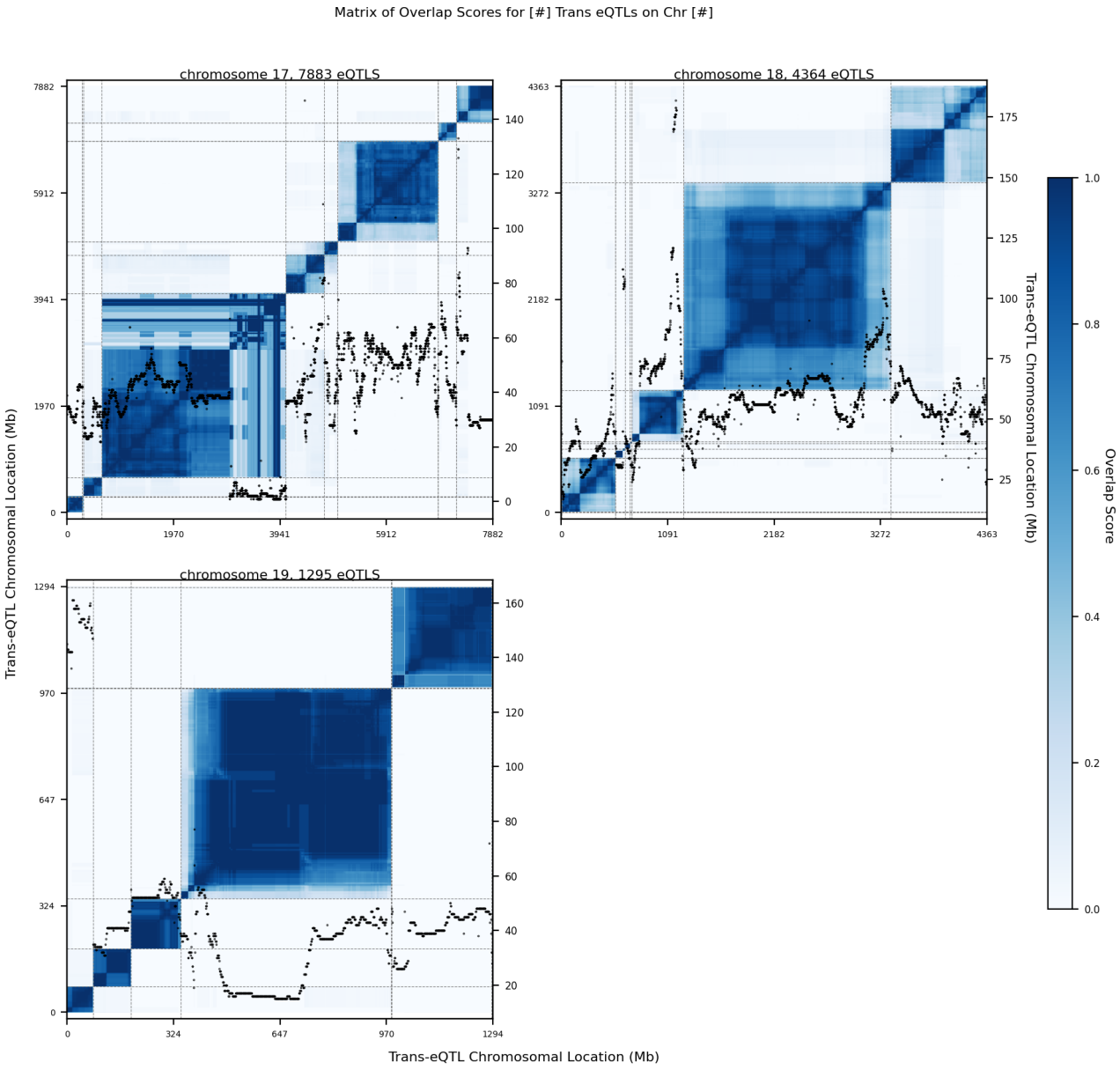

Figure 4.10: Chromosome wide overlap of eQTLs for eQTLs on the chromosomes 16-18. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered. Number of traits each SNP is associated with in black.

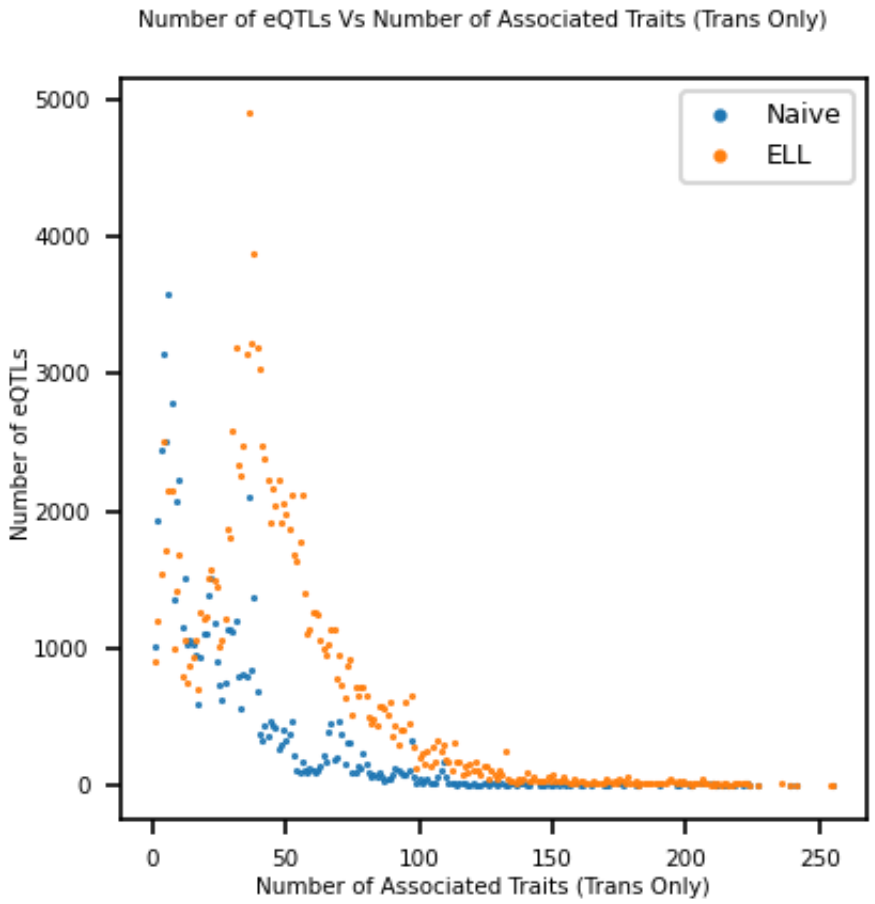

Figure 4.11: Number of eQTLs found by either ELL or Naive for each different number of associated trans traits. 15,071 gene expression levels analyzed. Only traits on a different chromosome from the trans-eQTL are considered.

CHAPTER 5

DISCUSSION

We have developed a novel global testing approach, ELL, with applications to trans-eQTL detection. Through extensive simulation studies, we have demonstrated the type 1 error validity of ELL and the power improvement it can provide over other methods.

In particular, ELL is shown to dominate a significant region of sparsity/density of true signals between very sparse but strong signals, where FDR and minP dominate, and very dense but weak signals where CPMA and sumZ^2 dominate.

A novelty of the method among trans-eQTL mapping methods is the incorporation of $\hat{\Omega}$ directly into the statistic. This provides for higher power while controlling type 1 error. We explored two different strategies for calculating p-values for the ELL statistic: an analytical approximation and a empirical approximation. The former was shown to not have correct type 1 error and thus we have adopted the latter strategy. Further work may be conducted to evaluate potential alternative analytical approximations.

The method has been demonstrated to have quite low computation time, particularly compared to other methods which attempt to incorporate Ω into the statistic directly. We have also introduced a highly accurate yet extremely fast method of pre-computing the ELL statistic.

Through application to the AIL mouse dataset [11] we showed that ELL can be run quickly on large, real data sets and can produce improved power compared to other methods. Further work will expand upon this analysis.

While ELL is demonstrated here in the context of trans-eQTL detection, it is a general

method with broad applications. No conditions are placed on the correlation matrix Ω, it works with small samples - as long as there is enough data to estimate Ω - and relies only on having summary association test statistics.

5.0.1 Limitations

We are assuming that the correlation structure between association statistics for the same SNP but different genes, remains the same from one SNP to the next. In contexts in which this is not a good approximation, the method might not perform well. Future work could evaluate whether the estimation of Omega can be improved by taking into account the correlation among SNPs, rather than just taking the product-moment estimator.

Our pre-computation strategy is necessarily more memory intensive than computing one statistic at a time. It requires holding the matrix of pre-computed values in memory. This requires a computer with large memory.

5.0.2 Future Work

We would like to evaluate alternatives to the BetaBinomial approximation used. Perhaps using maximum likelihood instead of the method of moments would be better.

In this work, particularly in the simulations, we focussed on controlling the type 1 error. However, it could be possible to use quite analogous methodology but to choose a cutoff per SNP that controls the false discovery rate.

In our initial analyses of the AIL dataset, we found that, as expected, a pre-processing approach that removes a huge number of latent factors could severely limit the ability to detect interesting trans-eQTL signal. After changing the pre-processing approach, we had

much more substantial trans-eQTL results (though of course the final verdict is still out on the results until they can be replicated in other data sets). Further study of how best to pre-processs data in a way that removes artifacts while still preserving tran-eQTL signal could be important and useful.

REFERENCES

[1] Ian Barnett, Rajarshi Mukherjee, and Xihong Lin. The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, 112, 06 2016.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[3] Robert H. Berk and Douglas H. Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(1):47–59, 1 1979.

[4] Boel Brynedal, JunMyung Choi, Towfique Raj, Robert Bjornson, Barbara E. Stranger, Benjamin M. Neale, Benjamin F. Voight, and Chris Cotsapas. Large-scale trans-eqtls affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *American Journal of Human Genetics*, 100:581–591, 2017.

[5] O. Carlborg, D. J. De Koning, K. F. Manly, E. Chesler, R. W. Williams, and C. S. Haley. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, 21:2383–2393, 2005.

[6] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962–994, 2004.

[7] David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science*, 30:1–25, 2015.

[8] Yoav Gilad, Scott A. Rifkin, and Jonathan K. Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics : TIG*, 24 8:408–15, 2008.

[9] V. Gontscharuk, S. Landwehr, and H. Finner. The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biometrical Journal*, 57:159–180, 2015.

[10] Veronika Gontscharuk, Sandra Landwehr, and Helmut Finner. Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli*, 22(3):1331–1363, 08 2016.

[11] Natalia M. Gonzales, Jungkyun Seo, Ana I. Hernandez Cordero, Celine L. St. Pierre, Jennifer S. Gregory, Margaret G. Distler, Mark Abney, Stefan Canzar, Arimantas Lionikas, and Abraham A. Palmer. Genome wide association analysis in a mouse advanced intercross line. *Nature Communications*, 9(1):5162, 12 2018.

[12] Hong Lan, Jonathan P. Stoehr, Samuel T. Nadler, Kathryn L. Schueler, Brian S. Yandell, and Alan D. Attie. Dimension reduction for mapping mrna abundance as quantitative traits. *Genetics*, 164:1607–1614, 2003.

[13] Xiaoyin Li and Xiaofeng Zhu. Cross-phenotype association analysis using summary statistics from gwas. *Methods in molecular biology (Clifton, N.J.)*, 1666:455–467, 2017. 28980259[pmid].

[14] Christoph Lippert, Franceso Paolo Casale, Barbara Rakitsch, and Oliver Stegle. Limix: genetic analysis of multiple traits. *bioRxiv*, 2014.

[15] Xuanyao Liu, Joel A Mefford, Andrew Dahl, Meena Subramaniam, Alexis Battle, Alkes L Price, and Noah Zaitlen. Gbat: a gene-based association method for robust trans-gene regulation detection. *bioRxiv*, 2018.

[16] Keith Patarroyo. A digression on hermite polynomials, 01 2019.

[17] C. Peterson, S. Service, A. Jasinska, F. Gao, I. Zelaya, T. Teshiba, C. Bearden, V. Reus, G. Macaya, C. López-Jaramillo, M. Bogomolov, Y. Benjamini, E. Eskin, G. Coppola, N. Freimer, and C. Sabatti. Characterization of expression quantitative trait loci in pedigrees from Colombia and Costa Rica ascertained for bipolar disorder. *PLoS Genet*, 12:e1006046, 2016.

[18] Christine B. Peterson, Marina Bogomolov, Yoav Benjamini, and Chiara Sabatti. Many phenotypes with many false discoveries: error controlling strategies for multitrait association studies. *Genetic Epidemiology*, 40:45–56, 2016.

[19] Matthew V. Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, 11 2006.

[20] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 1986.

[21] Jae Hoon Sul, Towfique Raj, Simone De Jong, Paul I.W. de Bakker, Soumya Raychaudhuri, Roel Ophoff, Barbara Stranger, Eleazar Eskin, and Buhm Han. Accurate and fast multiple-testing correction in eqtl studies. *American journal of human genetics*, 96, 05 2015.

[22] Ryan Sun and Xihong Lin. Genetic variant set-based tests using the generalized berk–jones statistic with application to a genome-wide association study of breast cancer. *Journal of the American Statistical Association*, 0(0):1–13, 2019.

[23] J. W. Tukey. T13 n: The higher criticism. course notes, statistics 411, princeton. 1976.

[24] E Weine and M S McPeek. Efficient calculation of alternative agnostic testing bands for qq-plots.

[25] Harm-Jan Westra, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, Benjamin P. Fairfax, Katharina Schramm, Joseph E. Powell, Alexandra Zhernakova, Daria V. Zhernakova, Jan H. Veldink, Leonard H. Van den Berg, Juha Karjalainen, Sebo Withoff, André G. Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A. C. 't Hoen, Eva Reinmaa,

Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B. Singleton, Dena G. Hernandez, Michael A. Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A. Gharib, Daniel A. Enquobahrie, Thomas Lumley, Grant W. Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C. Jansen, Peter M. Visscher, Julian C. Knight, Bruce M. Psaty, Samuli Ripatti, Alexander Teumer, Timothy M. Frayling, Andres Metspalu, Joyce B. J. van Meurs, and Lude Franke. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 10 2013.

[26] Jian Yang, Noah A. Zaitlen, Michael E. Goddard, Peter M. Visscher, and Alkes L. Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100–106, 2 2014.

[27] Gael Yvert, Rachel B. Brem, Jacqueline Whittle, Joshua M. Akey, Eric Foss, Erin N. Smith, Rachel Mackelprang, and Leonid Kruglyak. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, 35:57–64, 2003.

[28] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44:821–824, 2012.

[29] X. Zhou and M. Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nat Methods*, 11:407–409, 2014.

[30] Xiaofeng Zhu, Tao Feng, Bamidele Tayo, Jingjing Liang, J Young, Nora Franceschini, Jennifer Smith, Lisa Yanek, Yan Sun, Todd Edwards, Wanwan Chen, Mike Nalls, Ervin Fox, Michele Sale, Erwin Bottinger, Charles Rotimi, Yongmei Liu, Barbara McKnight, Kiang Liu, and Susan Redline. Meta-analysis of correlated traits via summary statistics from gwass with an application in hypertension. *American journal of human genetics*, 96, 12 2014.