Aim 1: Discovering trans eQTLs from summary statistics

Significance. The approach proposed here provides a novel and powerful method for discovering trans eQTLs. It is computationally fast, requiring only summary statistics, and has greater power over a wide range of models than other trans eQTL mapping approaches. Identifying trans eQTLs is biologically significant because it will lead to a deeper understanding of the regulatory circuits and mechanisms of gene expression. Although many cis eQTLs have been found, relatively little is known about trans regulation in humans. This method will enable much more effective discovery of trans eQTLs than has previously been possible.

Context. Trans eQTLs should be pervasive in the human genome, yet relatively few have been discovered. Model organism studies, particularly in yeast [18], have shown that there can be many such eQTLs in the genome, each of which may regulate from several up to hundreds of genes [14]. Studies in humans, however, where experimental conditions cannot be as carefully controlled, have yielded relatively few trans eQTLs even with a large meta-analysis [11]. Although one hypothesis might be that there are relatively few trans eQTLs in humans, similarity of biology across organisms suggests this to be unlikely. A more probable reason for the dearth in discovered trans eQTLs is that the statistical methods used are not very effective at finding them

Finding trans eQTLs presents challenges that can be distinctly different from those for finding genetic associations for other types of traits. First, trans eQTL mapping has a much higher multiple testing burden than that for cis eQTL mapping or for a typical GWAS, because in trans eQTL mapping doing an association test for every SNP-gene pair entails approximately $10^{10} - 10^{11}$ tests. If trans effects were sufficiently strong, it might be possible to overcome this burden. Unfortunately, trans effects seem likely to be no stronger than cis eQTL effects [14]. Clearly, then, the strategy of many single SNP-gene association tests will be dramatically underpowered.

A common approach to avoid the heavy multiple testing burden is to simply test a small subset of preselected SNPs, with as large a sample size as possible [11]. This has the obvious drawback that only trans eQTLs that are already suspected of being present will be discovered. Furthermore, increasing the sample size is often not possible with a limited budget. Meta-analysis can help, but does not address the underlying problem of using an inefficient approach. False discovery rate (FDR) [19] based error control, being more liberal than Bonferroni correction, allows for more discoveries [11,20] but still suffers from being underpowered because it does not alter the underlying test methodology.

Innovation. A solution to the problem of underpowered methods arises by taking advantage of the characteristic of trans eQTLs to regulate multiple genes. In this case, much power can be regained by shifting away from many individual null hypotheses of SNP-gene association to a test of a single global null hypothesis of a SNP to many genes. When the global null hypothesis is rejected we conclude that the SNP is associated with some subset of the genes. This strategy has a precedent in genetics. For instance, heritability analyses detect an overall genetic effect without identifying which variants are likely the causative ones, or rare variant tests will group together many SNPs that individually have low power but collectively may be detectable. This paradigm is likely to be a powerful one for mapping trans eQTLs, and a very recent paper [21] uses this approach. Their method is likely to be well powered when a trans eQTL affects hundreds of genes, or more, but may have low power if fewer genes are targeted by the trans eQTL. Our proposed method adopts this paradigm, thus avoiding a large multiple testing burden, and has greater power than other methods over a wide range of models.

In addition, to reject the null hypothesis of a SNP not being a trans eQTL, a trans eQTL mapping method should, ideally, identify which transcripts are liely to be the targets of the eQTL. Methods that simply test the global null of a SNP's genotype not being associated with any transcript do not necessarliy select candidate targets, e.g. the CPMA method [21]. Our method does identify likely target transcripts.

Our proposed method is novel; we are not aware of similar methods in the genetic literature, and is based on the statistical concept of "local levels" [22], a recent innovation developed in the context of goodness of fit testing. The application of appropriate local levels to the global null hypothesis enables a new testing approach with possible avenues for future work to develop even more powerful tests.

Methods. The method we propose requires only summary statistics and is justified by an underlying model for trans eQTLs. The key aspect of this model is that a single SNP, if it is a trans eQTL, will cause variation in M gene expression levels, where $M > 0$ and may typically range from a few to hundreds. The total number of measured transcripts P, on the other hand will normally be on the order of 10^4. We view the underlying, generative model as a multivariate linear model with a single predictor variable for the SNP with P outcomes and P effect size parameters, only M of which are non-zero. When the data are analyzed, the SNP is tested against each trait separately in a linear model or linear mixed model framework resulting in P p-values. Because the M effects are small, Bonferroni correction results in power that is only nominally larger than the type-1 error rate. That is, in our model we assume thests where the alternative model is

true to be rare among all tests and to have a "weak" effect. Specifically, consider having P test statistics $Z_i, 1 \leq i \leq P$ with the null hypothesis $Z_i \sim N(0,1)$ against the mixture alternative $Z_i \sim (1-\epsilon)N(0,1) + \epsilon N(\mu,1)$, where $N(\cdot,\cdot)$ denotes a normal density. In the RW setting ϵ is "small" (in the range $1/P^{1/2}$ to $1/P$) and μ is weak enough that it does not noticeably affect the bulk distribution of p-values. In particular, the largest values of Z_i that are drawn from $N(\mu,1)$ will generally be smaller than the largest Z_i that are drawn from $N(0,1)$ (because there are very many more drawn from $N(0,1)$).

This is a highly challenging testing regime and an approach such as FDR will be underpowered because it implicitly assumes high enrichment of true alternative tests among the smallest p-values. The "higher criticism" (HC) test has asymptotic optimality properties in this setting [23, 24], but the asymptotic approximations have poor accuracy until the sample size P is of the order of 10^{69} [25]. We propose a new method, the G-Null test for testing the global null hypothesis that a single SNP is not associated with any transcript. We show in the **Preliminary results** that the G-Null test has, for realistic sample sizes, more power than HC and other trans eQTL mapping approaches.

Our approach is based on local levels [22]. Given the ordered p-values $\pi_{(1)} < \pi_{(2)} < \cdots < \pi_{(P)}$, the local level α_i is defined as $\alpha_i = \Pr(\pi_{(i)} \leq h_i), i = 1, \ldots, P$ for some threshold value h_i, assuming the global null is true. A general level-α test of the global null requires $\Pr(\pi_{(1)} > h_1, \pi_{(2)} > h_1, \ldots, \pi_{(P)} > h_P) = 1 - \alpha$. In other words, we construct the level α test by selecting thresholds $h_i, i = 1, \ldots, P$ such that the ordered uniform random variables are all greater than their respective thresholds $1 - \alpha$ of the time. We reject the global null if at least on of the p-values is small enough. What, then, should the h_i be? Asymptotically, the HC statistic is optimal and implicitly selects h_i such that all local levels α_i are equal, although this no longer holds true with finite samples [25]. Selecting equal local levels, then, is a reasonable criterion for choosing the h_i. That is we need an α_{loc} such that $\alpha_{\text{loc}} = \Pr(\pi_{(i)} \leq h_i)$ for every i, while maintaining the criterion for a level-α test. To find α_{loc} first note that under the null $\pi_{(i)} \sim \text{Beta}(i, P+1-i)$, whose CDF we call F_i, i.e. $F_i(h_i) = \alpha_{\text{loc}}$. We propose obtaining α_{loc} with the following algorithm: (1) Simulate T sorted data sets of $U \sim$ unif(0,1) random variables, each of size P, (2) set $\alpha_*^{(k)}$ to an initial value ($k = 1$), (3) obtain values $h_i = F_i^{-1}(\alpha_*^{(k)})$, (4) Compute R, the fraction of data sets where $U_{(i)} > h_i$ for every i, (5) obtain $\alpha_*^{(k+1)}$: if $R < 1 - \alpha$, decrease $\alpha_*^{(k)}$; else if $R > 1 - \alpha$, increase $\alpha_*^{(k)}$; else $R = 1 - \alpha$, set $\alpha_{\text{loc}} = \alpha_*^{(k)}$ and stop, (6) with the new $\alpha_*^{(k+1)}$ goto step (3). Since $0 < \alpha_{\text{loc}} < \alpha$, a binary search quickly converges on the correct α_{loc}.

To perform the G-Null test with α_{loc}, from the observed and ordered p-values $\pi_{(i)}$ compute $l_i = F_i(\pi_{(i)})$ for every i and reject the global null if there are any $l_i < \alpha_{\text{loc}}$. We call the l_i the "l-value," and it is intuitively a "local p-value." Note that though the p-values are ordered $\pi_{(1)} < \cdots < \pi_{(P)}$ the corresponding l-values $l_1, \ldots, l_P$, in general, are not. Instead, the ith l-value is a measure of the enrichment of p-values smaller than the ith smallest p-value. In the setting where effect sizes are weak and rare among the tests, the smallest p-values are often from tests where the null hypohtesis is true. Nevertheless, tests where the alternative is true will tend to have small p-values even if they are not the smallest. The l-values tend to be smallest at more intermediate to small p-values, where there have been a sufficient number of tests where the alternative model is true.

We can also leverage the property that the l-values are indicative of the accumulation of tests under the alternative to propose an approach to identify a candidate set of transcripts to be the target of the eQTL. Among l-values $l_i, i = 1, \ldots, P$ define $G = \min_j l_j^*$ where $l_j^* \in \{l_i | l_i < \alpha_{\text{loc}}\}$, then the candidate set of tests are those tests whose p-values $\pi_k \leq \pi_{(G)}$. That is, when tests are ordered by p-value, candidate tests are those that come before, or are equal to, the first test whose l-value is less than α_{loc}.

Preliminary results. We show now that the G-Null test has more power than current methods for mapping trans eQTLs. We undertook simulations in a simplified setting and compared power of our approaches with the power of the recently proposed CPMA test [21] and the more commonly used FDR approach. The CPMA statistic tests the null hypothesis that the p-values, from tests of association of a single SNP with many transcripts, are uniformly distributed by fitting an exponential distribution to the log of the p-values and performing a likelihood ratio test. In our FDR implementation we use the Benjamini-Hochberg [19] procedure and set the FDR to 0.05.

	M=80 $R^2 = 0.0091$	M=40 $R^2 = 0.013$	M=20 $R^2 = 0.018$
HC	0.56	0.68	**0.78**
Score	0.83	0.53	0.36
CPMA	0.74	0.43	0.30
FDR	0.36	0.49	0.58
G-Null	**0.87**	**0.77**	**0.77**

Table 1: Power of the methods under different trans eQTL models

We simulate a single SNP genotype for 500 individuals with a minor allele frequency of 0.3 and 10^4 quantitative transcripts. For simplicity we simulate all individuals and transcripts

as independent. We examine four different models with a varying number M of the transcripts associated with the genotype with the effect size parameterized by R^2. For each SNP-transcript pair we perform linear regression from which we obtain p-values. These are input to the methods that compute the statistics as described above. We obtain the null distribution of the statistics by performing 500 replicates where none of the transcripts are associated with the SNP. For each of the models we perform 200 replicates and test the global null hypothesis at level $\alpha = 0.05$.

The power of the methods is shown in Table 1. Across all models tested G-Null has the highest power. The FDR approach suffers in all scenarios because the effect sizes are weak enough that the tests where there is an association are unable to pass the Benjamini-Hochberg threshold. The score test (see **Proposed plan**) and CPMA have power closer to G-Null when there are a large number of non-null effects, even when those effects are very small. In contrast, when the number of non-nulls drops, and the signal becomes sparse, these test do poorly compared to HC and G-Null, the latter of which maintains high power across all settings.

We also looked at the enrichment of true alternative models in our candidate set, as defined above. For the model where 80, 40 and 20 tests out of 10^4 were under the alternative, the true discover rates (the fraction of candidates that were true) were 0.61, 0.69 and 0.78, respectively. Note that all models are in the "non-estimable" range [23,24]. That is, even asymptotically the numbers of type 1 and type 2 errors do not converge to zero. This is indicative of the extreme difficulty of reliably picking our the few alternative tests from the large number of null tests.

Proposed plan. We propose to develop a complete software implementation of the methods described above that will be capable of analyzing realistic data sets from humans as well as other organisms. We will verify the validity of the statistical methods via simulation studies and compare the power of our approach with other methods for trans eQTL mapping. In addition, we will apply our method to map trans eQTLs in the GTEx and GEUVADIS data sets and compare the results to other methods and to known trans eQTLs.

We will first create a software implementation of the G-Null method. In our preliminary work we have a simplified implementation that assumes independence between the tests, but in order to be useful for real data the method will have to incorporate the correlation between transcript p-values. As input, our software package will take a $P \times R$ matrix of p-values for P transcripts by R SNPs and output (1) a list of those SNPs that rejected the global null hypothesis and at what significance level, and (2) for each trans eQTL a list of candidate transcripts that are targets of the eQTL.

In the case where all tests are independent, there exists an analytic solution to the asymptotic value of α_{loc}. In expression data, however, transcript levels and SNPs are often not independent, resulting in correlations between p-values. In this case, we must obrain an empiric null distribution which can then be used to find the proper α_{loc}Although permutations of the raw data could provide this distribution, we obtain dramatic copmutational savings by directly working with the summary statistics. From the input $P \times R$ p-value matrix we first convert to p-values to z-scores and compute the $P \times P$ z-score sample covariance matrix ($R \gg P$ so this matrix is positive definite). An empiric null data set can then be generated directly from a multivariate normal distribution from which one can easily, and quickly, generate large sample sizes. From the empiric null z-scores, we obtain the empiric null distribution of minimum l-values. The p-value of each G-Null test of the real data is obrained by the proportion of empiric null minimum l-values that are smaller than a SNP's minimum l-value. Note that finding the minimum l-value for a single G-Null test is extremely fast, thousands can easily be done in approximately one minute. Computing an empiric null distribution and performing a genomewide scan, once multi-core parallelism is incorporated, can be done in several hours.

Following completion of the G-Null software, we will also implement a score statistic based test that requires only summary statistics. The test follows from a proportional odds generalized linear mixed model with the genotype as teh outcome and the transcript effects as random. The score statistic is of the form $Q = \mathbf{g}^t\mathbf{X}^t\mathbf{W}\mathbf{X}\mathbf{g}$, where $\mathbf{g}$ is a vector of genotypes, $\mathbf{X}$ is a matrix of transcript levels and $\mathbf{W}$ is a weight matrix. This can be rewritten to use only summary statistics $Q = \sum_i w_i z_i^2$ where w_i are the weights and z_i the z-score for the ith transcript with the SNP. Although the G-Null test has the highest power over a wide range of realistic trans eQTL models, the score test has relatively high power in the case where a single SNP affects several hundred transcripts, each very weakly. Our simulations show power similar to or higher than CPMA. Including this test in the software will give users flexibility in their search for trans eQTLs.

Simulation studies will be done to check the validity of the methods and to compare the power to other approaches under a variety of alternative models. To make simulations realistic, they will be based on real transcript correlation structure from the GTEx and GEUVADIS studies. Our simulations will have sample sizes in the range of 500–2000 individuals and 10^4 transcripts with varying genetic architectures. Transcripts will have from 0–400 eQTLs with total heritabilities from 0–0.9. We will compare the power of our method with CPMA, FDR and any other statistical approaches we find in the literature. We will also check the efficiency of our method for identifying candidate target

transcripts. We will evaluate AUC of the ROC curves as well as the TDR and FNDR for the proposed approach for setting the identification threshold.

Following the simulation studies we will analyze both the GTEx and GEUVADIS data sets. We will follow the recommended protocols for QC and population stratification correction. We will use linear regression to compute p-values for SNP-transcript pairs across select tissues. From these p-values we will use our method to find trans eQTLs that replicate in both data sets with consistency in the tissues, target genes and direction of effect.

Possible problems and alternative approaches. Computationally, our methods as mentioned above, is very fast and can scale to completing analysis of genomewide size studies in a matter of hours. This is possible because our method (A) requires only summary statistics, alleviating the need to run new analyses on the raw data; and (B) is highly parallelizable, allowing the software to easily take advantage of the multi-core architecture of modern processors. In order for this scaling to be successful, our code will have to use RAM efficiently. We will accomplish this by using memory mapping techniques, available in R packages such as BigMemory, to allow the code to efficiently load only the necessary portion of the data needed at any time.

Benefits of the proposed method. Our method has substantially more power to discover trans eQTLs than other currently used approaches. Even as sample sizes get larger this translates into a greater number of discoveries. Furthermore, unlike methods such as CPMA, our method identifies a candidate set of target transcripts which are enriched for true discoveries. These characteristics together allow for a far more effective use of challenging to obtain data, resulting in a deeper understanding of the trans regulatory landscape of the genome.

Aim 2: Removing unknown covariates while preserving trans genetic variation in expression data

Significance. Genetic regulatory mechanisms are critical for many biological processes, from gene co-expression to the orchestration of developmental processes, yet understanding the architecture of these mechanisms has proven challenging. A basic question in describing this architecture is simply to determine to what degree a gene's expression level is controlled by cis versus trans variation. Accurately assessing this question in gene expression data requires estimating and removing unknown sources of non-genetic variation in the data, while taking care not to eliminate genetic variation. The method proposed here takes a novel approach that preserves the heritability in the data while allowing the removal of unknown, non-genetic covariates.

Context. Although many transcript levels are highly heritable, some studies have argued that most of this heritability is due to trans loci [12,14,15], while more recent findings instead suggest that cis heritability is the majority [9,13]. One important difference in the recent work is the use of methods for removing hidden covariates from the expression data. Though these methods are recognized as playing an important role in removing technical variation, and thus increasing the power to find cis eQTLs [26,27], they also run the risk of removing trans genetic effects as well. The most recent approaches [27] aim to preserve trans eQTL effects, but new work shows that these sorts of data treatments can negatively bias the estimated effect of trans eQTLs on gene expression and instead suggest eschewing methods for removing hidden variation when studying trans effects [21,28].

Although not estimating hidden sources of variation has the advantage of preserving trans effects, and estimates of trans heritability specifically, it has the consequence of leaving undesirable sources of variation in the data. Noisier data masks genetic signal which also results in negatively biased estimates of heritability. A preferable approach is to use a method that allows one to find the hidden, non-genetic sources of variation that need to be removed while preserving the genetic signal that exists genomewide. There are a variety of approaches designed to estimate hidden sources of variation in multivariate data in general [29–31] and gene expression data in particular [26,27,32,33]. In these methods information to estimate hidden variates derives from correlations in the data that are assumed to arise from non-genetic (e.g., technical) sources. If these are the dominant sources of the correlation and variation in the data, the methods readily pick them out. However, estimates of heritability also rely on the presence of correlation in the data, and the genetic correlations are not distinguishable from the technical correlations, based solely on the expression data. The discovered variates, then, are a mixture of the genetic and non-genetic. When these discovered variates are regressed out of the data, both non-genetic and genetic effects are removed. As a consequence, heritability estimates can be significantly reduced. One approach [27] attempts to remedy this by using a subset of SNPs as potential sources of major trans signal. Unfortunately, this relies on knowing the correct set of potential trans eQTLs and only adjusts for strong trans effects and not the large number of smaller trans effects that exist throughout the genome. Our goal, then, is to develop a method that preserves the genetic effect, in general, and trans heritability in particular, while simultaneously estimating the hidden sources of non-genetic variation.