Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations

Matthew Weiser*, Sayan Mukherjee§, Terrence S. Furey†

*Curriculum in Bioinformatics and Computational Biology, †Departments of Genetics and Biology, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599 USA, §Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, North Carolina 27708 USA
SHORT RUNNING TITLE
Mapping distal eQTL using gene dependency structure

KEY WORDS
eQTL; gene expression; gene networks; genetical genomics

CORRESPONDING AUTHOR
Terrence S. Furey
120 Mason Farm Road CB#7240
Genetic Medicine Building
University of North Carolina
Chapel Hill, NC 27599
919-966-7033
tsfurey@email.unc.edu

FUNDING
This study was supported by NIH training grant 5T32GM067553-09, as well as the University Cancer Research Fund (UCRF) from UNC-Chapel Hill, and NIH Grants U01-CA157703 and R01-CA166447.

COMPETING INTERESTS
The authors have declared that no competing interests exist.
ABSTRACT

Mapping expression quantitative trait loci (eQTL) has identified genetic variants associated with transcription rates, and has provided insight for genotype-phenotype associations obtained from genome-wide association studies (GWAS). Traditional eQTL mapping methods present significant challenges for multiple testing burden, resulting in a limited ability to detect eQTL that reside distal to the affected gene. To overcome this, we developed a novel eQTL testing approach, NetLIFT, which performs eQTL testing based on the pairwise conditional dependencies between genes’ expression levels. When applied to existing data from yeast segregants, NetLIFT replicated most previously-identified distal eQTL, and identified 46% more genes with distal effects compared to local effects. In liver data from mouse lines derived through the Collaborative Cross project, NetLIFT detected 5,744 genes with local eQTL while 3,322 genes had distal eQTL. This analysis revealed founder of origin effects for a subset of local eQTL that may contribute to previously described phenotypic differences in metabolic traits. In human lymphoblastoid cell lines, NetLIFT was able to detect 1,274 transcripts with distal eQTL that had not been reported in previous studies, while 2,483 transcripts with local eQTL were identified. In all species, we found no enrichment for transcription factors facilitating eQTL associations; instead, we find that most trans-acting factors were annotated for metabolic function, suggesting that genetic variation may indirectly regulate multi-gene pathways by targeting key components of feedback processes within regulatory networks. Furthermore, the unique genetic history of each population appears to influence the detection of genes with local and distal eQTL.
INTRODUCTION

Gene expression is highly heritable, indicating a strong genetic component (Cheung et al. 2003; Schadt et al. 2003). Expression quantitative trait loci (eQTL) mapping strives to uncover the underlying genetic architecture of transcriptional regulation. An important concept in dissecting complex regulatory processes is to identify both local and distal variants that are associated with gene expression. Local eQTL are largely thought to regulate proximal genes by affecting the activity of regulatory elements that directly influence transcription rates, such as through alterations in genomic sequence that affect binding affinities of regulatory factors. In contrast, distal eQTL map to genomic locations far from the affected gene, possibly on different chromosomes, and likely act initially on the expression or function of some nearby, intermediate gene that then affects the associated target gene in *trans*. Notably, in genetically diverse populations such as humans, the reported effect sizes and significance levels for distal associations are weaker than for local eQTL (Brem et al. 2002; West et al. 2007; Doss et al. 2005). This is likely attributable to the greater noise inherent in indirect effects that occur within the context of a protein-protein interaction network.

Initial eQTL discovery analyses performed association tests for all pairs of genomic variants and genes (Holloway et al. 2011; Mehta et al. 2012; Alberts et al. 2011), leading to challenges in both sensitivity and interpretation. Although recent methods have greatly reduced the computational burden for this approach (Shabalin 2012), the reduced statistical power due to multiple testing correction still present significant problems, especially in detecting distal eQTL. Using this technique, the reported frequency of distal effects has varied from 2% to 75% of all detected eQTL (Mehta et al. 2012; Göring et al. 2007; Yvert et al. 2003), and it remains unclear whether this is attributable to differences in regulatory architecture or statistical power. Indeed, in several recent eQTL analyses using human data, distal eQTL mapping was either not performed or not reported (Pickrell et al. 2010; Lappalainen et al. 2013), likely due to the inability to detect any distal eQTL whatsoever. Additionally,
inferring the direction of effect of distal associations that result from protein interactions is difficult when dealing with gene expression data that is often noisy and highly correlated.

To detect distal eQTL with greater power, some recently-developed methods assume an underlying regulatory architecture in which the local regulation of an intermediate gene leads to widespread expression variation in a large set of target genes (Kompass and Witte 2011; Bottolo et al. 2011; Duarte and Zeng 2011; Rotival et al. 2011). Modules of target genes are defined by factor analysis or gene-gene correlation statistics, and association testing is performed between genotypes and summary statistics of each module. In this setting, strong associations are thought to represent master regulators that exert broad, but potentially weak, effects in the regulatory network. These approaches reduce the multiple testing burden, as thousands of genes are replaced by a few dozen modules; however, there remain several drawbacks. First, if the regulatory activity of a trans-acting factor (TAF) affects only a handful of target genes, the initial clustering approach may not identify the small gene module. Secondly, the intermediate genes regulating the expression of gene modules are often not identified. Finally, expression for individual genes belonging to a module do not always correlate with the eQTL associated with the module, raising doubts about the validity of the results (Kompass and Witte 2011).

Others have developed methods focused on addressing interpretability and directionality of associations using randomization of genetic variables (Chen, Emmert-Streib, and Storey 2007) and causal model selection tests (Neto et al. 2013) as a foundation for statistical inference. In these methods, conditional dependence between expression of genes and/or latent variables is used to probabilistically determine whether the association between the genetic variant and target gene is causal. In this study, we present a novel eQTL detection method: “Network-based, Large-scale Identification of distal eQTL” (NetLIFT), which, rather than performing causal model selection or
randomization, uses pairwise partial correlations derived from gene expression data to restrict distal association testing, thereby reducing the multiple testing burden and highlighting candidate regulatory genes. In this framework, statistically significant local associations are first identified, and then local eQTL variants are tested for distal associations only for genes whose expression values show evidence of direct effects. We show that NetLIFT identifies individual SNP-gene distal associations with greater power than traditional pairwise eQTL testing, scales well to large data sets, and provides interpretability regarding the mechanism of association by highlighting potential trans-acting factors. In simulation studies, NetLIFT better identified distal eQTL, especially those with small numbers of target genes, when compared with a traditional all-SNPs-vs-all-genes approach, a module-based approach (Independent Components Analysis, adapted from (Rotival et al. 2011)), and a method designed to identify causal associations using randomization of genotype data (Chen, Emmert-Streib, and Storey 2007). Applying NetLIFT to a data set consisting of 112 yeast segregants (Brem and Kruglyak 2005), we recapitulated previously reported distal associations and putative regulators, while discovering several additional eQTL with plausible biological mechanisms of association. In mouse livers, we discovered founder of origin effects for a subset of local eQTL that drive differential expression of target genes in a subspecies-of-origin-specific manner, suggesting a possible role for these loci in transcriptomic and phenotypic differences between strains. Using data from human lymphoblast cell lines (Pickrell et al. 2010), we identified over one thousand distal associations not previously reported. We note that individuals from each of these three populations (yeast, mice, human) have unique genetic histories, and our analysis suggests that this influences the number and type of eQTL detected in each study.
MATERIALS AND METHODS

Description of the NetLIFT Model

The analysis workflow for the NetLIFT model is outlined in Figure 1, and was designed to parallel our understanding of the mechanism of trans regulatory effects. That is, if SNP $s_i$ affects the transcription of gene $g$ in trans, we expect that $s_i$ first directly affects the transcription level of an intermediate gene $g_i$, and that the transcription rate of $g_i$ directly or indirectly affects the transcription rate of $g_j$. There are three main steps to the NetLIFT algorithm:

Step 1: Identify local eQTL:

Local association tests are performed for all variants that lie within an a priori defined window of each gene (Figure 1a). Allele counts are regressed on the genes expression values, using a univariate, additive linear model. Since some genes contain many more variants than others, we control the false positive rate in local testing by retaining only associations that meet a Bonferroni-corrected significance cutoff of 0.05. Significant associations represent variants that may have a direct effect on the transcription rate of nearby genes, likely by altering activity of cis regulatory elements.

Step 2: Estimate pairwise partial correlations for all genes:

Pairwise partial correlations are estimated for all gene pairs (Figure 1b) to identify genes with expression level dependencies. The distribution of connections for gene networks has been shown to follow a power-law distribution (Barabási and Oltvai 2004; Jeong et al. 2000; Yook, Oltvai, and Barabási 2004; Lorenz et al. 2011) with an overall small numbers of edges. Therefore, we estimate the partial correlation matrix $G$ using a method that enforces sparsity on the entries of $G$ via L1 regularization, and which has been shown to accurately identify network hubs (Peng et al. 2009; Allen et al. 2012).
Briefly, this method performs joint sparse regression on all \( p \) variables (genes) simultaneously, by minimizing the penalized loss function:

\[
L = \frac{1}{2} \left( \sum_{i=1}^{p} \left\| \mathbf{g}_i - \sum_{j \neq i} \rho_{ij} \sqrt{\frac{\sigma_{jj}}{\sigma_{ii}}} \mathbf{g}_j \right\|^2 \right) + \lambda \sum_{1 \leq i < j \leq p} |\rho_{ij}|
\]

where \( \mathbf{g}_i \) and \( \mathbf{g}_j \) are the expression vectors for genes \( i \) and \( j \), \( \rho_{ij} \) denotes the partial correlation between genes \( i \) and \( j \), and \( \sigma_{ii} \) and \( \sigma_{jj} \) are the \( i \)th and \( j \)th diagonal entries of the inverse covariance matrix. The L1 penalty \( \lambda \) controls the sparsity of the network, and was optimized by minimizing the BIC criterion outlined in (Peng et al. 2009).

For \( p \) genes, the resulting \( p \times p \) matrix \( \mathbf{G} \) consists of entries \( G_{ij} \) that represent the correlation between expression vectors \( \mathbf{g}_i \) and \( \mathbf{g}_j \), conditioned on the expression of all other genes' expression:

\[
G_{ij} = \text{corr}(g_i, g_j | g_k, k \neq i, j)
\]

\( \mathbf{G} \) can be interpreted as an undirected network, where each node represents a gene, and an edge is drawn between two nodes if and only if the corresponding entry in the matrix \( \mathbf{G} \) is nonzero.

**Step 3: Distal eQTL testing:**

Distal eQTLs are called by integrating the results from these two steps (Figure 1c). For each variant \( s \) that shows significant association to a local gene \( g_i \), we test \( s \) for association with distal genes \( g_j \) that are nearby \( g_i \) in the partial correlation network defined by \( \mathbf{G} \). Since the edges of \( \mathbf{G} \) only account for direct relationships between two genes, we exploit the network structure to search for second-degree (downstream) regulatory effects as well. Specifically, we require two conditions for \( s \) to be tested for a distal effect on \( g_j \):

i) \( s \) must be strongly associated with expression of the putative trans-acting factor (TAF), \( g_i \); and
ii) genes $g_i$ and $g_j$ must be separated in the partial correlation network by no more than two edges, i.e. either $G_{ij} \neq 0$, or there exists a third gene $g_k$ such that $G_{ik} \neq 0$ and $G_{kj} \neq 0$. Additionally, we incorporate a threshold whereby two-degree genes are tested only if the association between $s_i$ and the intermediate gene $g_k$ meets a user-defined significance level (we selected $p < 0.2$ for this cutoff in all analyses presented here).

Although longer-range interaction effects could be considered by testing genes at increased distances within the network, doing so would exponentially increase the number of tests performed at each distance cutoff. We sought to balance this tradeoff by limiting the edge distance to two.

If a locally-affected gene contains many significantly associated variants, only the variant with the strongest local association is tested with distal genes. Furthermore, we impose directionality in the ambiguous case where two directly connected genes both have local eQTL, by only recording the direction with the strongest distal association. We note that since $G$ is a symmetric matrix representing an undirected network of correlated genes, we make no assumption regarding the direction of potential gene-gene effects, and therefore no assumption about how variant-to-gene effects may propagate through the network. Instead, we use the network structure only to select which variant-gene pairs to test for associations. Although significant associations do not provide conclusive evidence of trans associations, we expect that many of the distal eQTL will be acting in trans, potentially through the putative TAF identified by our method.

We note that the correlation-based network structure used to guide the distal association tests will likely lead to correlations among test statistics. The Benjamini-Yekutieli (BY) FDR correction holds rigorously under general dependence of test statistics (Benjamini and Yekutieli 2001); however, this correction is generally considered to be overly-conservative. Instead, we use the standard Benjamini-Hochberg FDR (Benjamini and Hochberg 1995), which in simulation studies was shown to perform
ICA Method

The Independent Components Analysis (ICA) methodology was adopted from (Rotival et al. 2011) and applied to the simulated data for comparison with NetLIFT. ICA identifies a predefined number of hidden variables (“independent components”) by factoring the gene expression data matrix, $X$, into a product of two matrices: $X \sim SA$. Each column of matrix $S$ corresponds to an independent component or factor, and the $i$-th element of a column is the “activation” level of the $i$-th gene in that factor. These factors are meant to model some latent or underlying biological process. The $k$-th row of matrix $A$ reflects the amount of activation of the $k$-th independent component across all individuals, $A_{ij}$ is activation on the $j$-th individual for component $i$. Rows of $A$ serve as the response vector when testing SNPs in a linear model. We used the fastICA function implemented in the R programming language to factor the expression data. This algorithm minimizes the statistical dependencies between the columns of $S$, so that each column of $S$ defines groups of co-expressed genes. Since the method requires an a priori-defined number of components to use in factorization, we set this parameter to 14; the number of modules in each simulated expression data set. To assign individual genes to components, we used the fdrtool function, which models a column’s scores as a mixture of null and alternative distributions. Each entry of the column is assigned an FDR corresponding to the likelihood of belonging to the null. For each component (column of $S$), a corresponding component-set was defined for genes with FDR < 0.05. Association tests were performed by regressing allele counts on rows of $A$, which represent the activation of each component across individuals. SNP-component associations with Benjamini-Hochberg corrected FDR < 0.05 were considered significant. For each association between a true comparison with the BY correction in the case of general dependency, and in particular for two sided t-statistics (Romano, Shaikh, and Wolf 2008).
local eQTL and a component, we defined the number of true positives to be the number of component-set genes which were downstream of the locally-affected driver gene. False positives were defined as any other gene assigned to that component-set.

**Trigger Method**

The Trigger method is described in (Chen, Emmert-Streib, and Storey 2007). This method aims to infer causality of a genetic variant on expression of a gene by treating genetic variants as randomized variables, and leveraging the causality equivalence theorem to identify the direction of effect. Briefly, let: \( s_i \) bet the genetic variant to be tested for association, and let \( g_i \) be a nearby gene. Trigger first tests for association between \( s_i \) and \( g_i \) (graphically: \( s_i \rightarrow g_i \)) using a standard likelihood ratio test. This gives \( \Pr(s_i \rightarrow g_i) \). If the probability of a local association exceeds a defined threshold, the variant is then considered for distal association testing. A similar likelihood test is used for defining the probability of linkage between \( s_i \) and \( g_j \), for all other genes \( g_j \), under the condition that \( s_i \rightarrow g_i \) (denoted \( \Pr(s_i \rightarrow g_i | s_i \rightarrow g_j) \)). Finally, we test whether \( s_i \) and \( g_j \), are *independent*, given the expression of \( g_i \): \( \Pr(s_i \perp g_j | g_i | s_i \rightarrow g_i \text{ and } s_i \rightarrow g_j) \). The causality equivalence theorem can be used to show that:

\[
\Pr(s_i \rightarrow g_i \rightarrow g_j) = \Pr(s_i \rightarrow g_i) \times \Pr(s_i \rightarrow g_j | s_i \rightarrow g_i) \times \Pr(s_i \perp g_j | g_i | s_i \rightarrow g_i \text{ and } s_i \rightarrow g_j),
\]

so multiplying the probability estimates yields an estimate for direct effect of \( s_i \) on \( g_j \). We use the R package “trigger” for implementation of this algorithm.

**Data Simulation Procedure**

A total of ten gene expression data sets were simulated, each with 500 genes and 250 samples. For each set of 500 genes, a network gene structure consisting of 14 disconnected gene modules of varying numbers of genes was imposed. Sizes of gene modules in each data set were as follows: 100 (x2), 50 (x2), 10 (x10), leaving 100 genes that were independent of any module. Module topologies
are depicted in Figure S1. For each module, the hub gene’s expression values for 250 samples were simulated first, by drawing from a standard normal distribution. Each successive downstream gene’s expression was modeled as a linear combination of the upstream gene plus random error, using an effect size of ±1, and a random error drawn from a standard normal distribution, represented as follows:

\[ g_{ds} = \beta g_{us} + \epsilon \]

where \( g_{ds} \) and \( g_{us} \) represent expression of the downstream and upstream genes, respectively, and \( \epsilon \sim N(0,1) \). Genes directly downstream of either the hub gene or a highly connected gene (defined as a gene with degree greater than 20) were chosen to have effect sizes of 1, while all other effect sizes were assigned randomly as -1 or 1 with probability 0.3 and 0.7, respectively.

Next, for each gene, the total number of SNPs for that gene was drawn from a gamma(4,0.2) distribution and rounded to the next highest integer. Minor allele frequencies for each SNP were drawn from a uniform(0.05, 0.5) distribution; from these, diploid genotype frequencies encoded 0, 1, 2 were derived under the assumption of Hardy-Weinberg equilibrium.

For each module, a single gene, not necessarily the hub gene, was chosen to have a local eQTL effect. Since the network topology is undirected, local eQTL effects on non-hub driver genes may lead to spurious distal associations in the analysis. In order to investigate the sensitivity and specificity of the method under these potentially confounding circumstances, we assigned local eQTL effects to hub genes in some modules, and to genes downstream of the hub in others. Furthermore, thirty percent of the 100 independent genes were assigned at random to have local eQTL effects. If a gene was not chosen to have an eQTL, genotypes were assigned randomly to the 250 samples. For genes chosen to have an eQTL, the direction of effect was chosen to be positive or negative with probability 0.7 and 0.3, respectively. Genotype labels were assigned using a genetic algorithm that
sought to maximize the effect size under the condition that the significance of association lie within a certain range (here, between 5e-05 and 1e-08). In cases where the eQTL was assigned to the hub gene, all genes in the module were considered as distal targets; however, to model cases where confounding associations may occur between the eQTL SNP and genes “upstream” of the locally-affected gene, we also assigned eQTL effects to non-hub genes.

The retrospective allele assignment allowed the specification of desired eQTL effect sizes and significance levels without the need to explicitly consider the pairwise correlations between genes when performing the genotype simulation. This procedure was carried out for 10 simulated data sets. Each data set consisted of gene expression networks for the same module topologies, and each module’s expression was characterized by an identical underlying genetic architecture. We defined true distal associations as those genes downstream of the locally-associated gene in the expression topology. Working code and a representative simulated data set is available for download at: http://fureylab.web.unc.edu/software/netlift/.

**Yeast Data**

Gene expression and genotype data, described previously (Brem and Kruglyak 2005) were obtained from R Brem. 112 yeast segregants were mated from parent strains BY4716 and RM11-1a and grown in culture. Strains were genotyped at 2,957 markers and expression measurements were assayed for 6,216 ORFs. Genes with no available annotation information were removed, leaving a total of 5,647 genes for analysis.

**Mouse Liver Data**

Gene expression data was previously assayed on the Affymetrix Mouse Gene 1.0 ST array, and was obtained from GEO (accession number GSE22297) (Aylor et al. 2011). Expression values were
normalized using the “rma-sketch” option in the Affymetrix Power Tools package. Probes containing SNPs were masked in the normalization procedure. Probesets that were expressed at a level above 6 on a log2 normalized scale in at least 87.5% of mice were retained, leaving a total of 9,377 probesets for further analysis. Genotypes for 181,752 markers from the “A” test array for the Mouse Diversity Array were obtained from D. Aylor.

**Human Lymphoblastoid Cell Line Data**

Gene expression data and HapMap phase 2 and 3 genotypes were obtained from [http://eqtl.uchicago.edu](http://eqtl.uchicago.edu). Normalization and processing were performed as described previously (Pickrell et al. 2010). Additionally, the top 25% of transcripts ranked by expression level were retained for further analysis, based on median expression level of the pre-quantile normalized data across all 69 individuals, leaving 9,810 transcripts that were retained for analysis.
RESULTS

Simulation Analysis

To assess the sensitivity and specificity of NetLIFT for identifying distal eQTLs, we applied the method to ten simulated data sets consisting of paired expression and genotype data (see Methods).

For comparison, we also tested three previously described eQTL detection methods: Independent Component Analysis (ICA), Trigger, and an All-vs-All pairwise testing approach (AvA) (Figure S2). The ICA method is primarily suited to identify eQTL that drive the expression of large numbers of distal genes; however, we note that the number of desired components must be defined according to some empirical criteria, and no specific intermediate gene is pinpointed as the trans-acting factor responsible for large scale variations. Therefore, this method does not identify local eQTLs.

We first compared the network structures inferred by NetLIFT’s partial correlation analysis to the true simulated regulatory architecture. We found that NetLIFT estimates the gene-gene partial correlation structure with high sensitivity, but note that as module connectivity increases, specificity decreases (Table S1, Figure S3). However, since the network structure is used primarily to determine which SNP-gene tests to perform, the main effect of false network edges is a slight increase in testing burden. As a result, we were willing to tolerate a reduction in network accuracy so long as the sensitivity remained high.

For detection of local eQTL effects, NetLIFT, Trigger, and AvA both identified true positives with 100% success (FDR < 0.05, Table S2). The local eQTL false positive rate for NetLIFT was identical to AvA under this FDR; setting a stricter FDR cutoff of 0.001 resulted in only one false positive for both methods. Additionally, we observed a large number of false positive local eQTL for Trigger, likely due to a lenient default thresholding criterion in the local eQTL testing step. Since we are particularly
interested in this method’s ability to detect distal eQTL, and since distal eQTL identification is conditional on local linkages for this method, we chose to retain the permissive threshold and focus primarily on results for distal associations.

Intra-module distal eQTL were predicted using each method simultaneously considering all genes and SNPs from all simulated modules. For each module, the true set of distal effects was defined as all SNP-gene associations between the module eQTL and genes downstream of the locally-affected gene. Thus, for modules where the eQTL acted on the hub gene, all combinations of the local eQTL SNP with non-hub genes were considered “true positives.” For modules with eQTL acting on non-hub genes, the true positives were defined as the eQTL-gene pairs in which the associated genes were downstream of the locally-affected, driver gene. False positives were defined as eQTL-gene associations where the associated gene was not downstream of the locally-affected gene. Figure 2 details the performance of each of the four methods.

In this case, NetLIFT identified true distal associations at a higher rate for all module topologies (overall 77.9% detection rate), at the cost of a slightly elevated false positive rate. These false positives were mostly due to eQTL SNPs being linked distally to genes that were in the same module, but that were not downstream of the locally-affected gene. Since our network estimation step cannot infer directionality of expression effects, these false associations reflect our inability to distinguish true functional associations from those that are due to confounding gene expression correlations present in the data. However, we note that the estimation of direct gene-gene effects and subsequent testing procedure prevents many upstream genes from being tested against the eQTL SNP, reducing the overall burden of these false associations. Moreover, in a rank based test performed on FDR values, true positives were found to have higher significance values than the false positives (\(p = 4.92\times 10^{-96}\)), again suggesting that the false positive count is strongly dependent on the FDR threshold chosen.
The AvA approach performed poorly, as most true associations were lost after correcting for multiple hypothesis testing. ICA performed well in large module settings, but poorly for small modules, suggesting that this approach is underpowered for detecting small co-regulated gene modules under the influence of a common variant. Trigger performed better than an AvA approach, though in general identified fewer than 12% of true distal associations. NetLIFT was the only method to consistently identify distal effects in all network topologies.

We next evaluated NetLIFT’s performance in detecting “hotspot” eQTL loci, where a hotspot is defined as a locus that is associated with more transcripts than is expected by chance. To derive a FWER for each locus, we used the procedure described in (Breitling et al. 2008), which permutes genotypes among samples but preserves the correlation structure present in the gene expression data. Performing association testing with the permuted genotype data sets yields a distribution of the expected maximum number of linkages under the null hypothesis of no eQTL associations. When restricting to a FWER of 0.05, NetLIFT identified the eQTL for all hub-based gene modules as hotspots in 10/10 simulated data sets, while the AvA approach identified these eQTL as hotspots only 20-60% of the time, and with many fewer linkages (Table S3).

To investigate whether a larger simulated data set affected the sensitivity and/or specificity of our method, we generated and analyzed an additional simulated data set consisting of 2,000 genes. We observed that the overall fraction of true and false positives remained similar in this analysis (data not shown). These simulation results indicate that in addition to scaling well to large data sets, NetLIFT may discover distal eQTL that are not readily identifiable with existing detection methods.
Analysis of 112 yeast segregants

We applied NetLIFT to previously analyzed paired genotype/gene expression data for 112 haploid yeast segregants (Brem and Kruglyak 2005). After filtering for genes with available annotation, 5,647 genes and 2,956 variants were retained for analysis. Variants within 10kb of the gene’s transcribed region were considered “local,” and all other linkages were denoted as distal eQTL. At an FDR of 0.05, we identified a total of 1,124 (19.9%) and 1,642 (29.1%) genes with local and distal eQTL effects, respectively (Figure S4). Local and distal effects were observed to have a similar effect size and level of significance (Table S4). The large effect sizes for distal eQTL are in line with previously reported results, and are likely attributable to the extreme diversity between the two strains of yeast.

A GO analysis using all 143 genes identified as intermediate trans-acting factors (TAFs) for at least 10 downstream targets revealed enrichments for a wide range of functions, with top hits reserved for metabolic function and transport (Table 1). This corroborates previous findings where putative regulators located near hotspots were not found to be enriched for transcription factors; instead, evidence suggests that many trans regulators exert widespread transcriptional effects by mediating levels of key metabolites or regulating post-translational processes (Yvert et al. 2003; Litvin et al. 2009). A comprehensive list of all putative regulators is provided in Table S5.

For most previously identified hotspots, NetLIFT correctly identified biologically validated regulators (Table 2). Several predicted novel regulators with more than 15 target genes were also found, many involved in metabolic and biosynthetic processes. In some cases, we provide regulatory evidence for novel drivers not identified previously for detected hotspots; furthermore, our results suggest that there may be numerous secondary drivers within previously identified hotspot regions, indicating that local association signals arising from two or more distinct loci may influence a similar set of distal target genes. One example is the hotspots on chromosome 2 where target genes are enriched for
ribosome biogenesis and ncRNA processing (Table 2). Previous results implicated AMN1 and MAK5 as trans-acting factors for subsets of the target genes; however, patterns of linkage to distinct regions within this locus suggest that additional regulators lie on chromosome 2 (Brem et al. 2002). In addition to AMN1, NetLIFT implicated at least seven new candidate regulators on chromosome 2 – TBS1, ARA1, YSW1, TOS1, UMP1, NPL4, and YBR197C – that were strongly linked with local eQTL (p < 1.0e-05) and were associated with highly overlapping sets of distally-associated genes (Figure S5). Notably, we fail to identify MAK5, as this putative regulator was shown to contain a loss of function mutation which has no effect on transcription (Brem et al. 2002). By definition, distal effects arising from amino acid substitutions affecting protein function of the trans-acting factor will be undetectable using NetLIFT, as we specifically seek to identify distal effects that arise from local, cis-regulatory effects.

Given the strong enrichment for ribosome function among target genes linking to the chromosome 2 loci, we hypothesized that causal variants would significantly affect growth rates via widespread differential transcription originating from direct up-/down- local regulation of the candidate TAF. To investigate this, we used segregants’ gene expression profiles to predict relative growth rate, using previously described methods (Airoldi et al. 2009). We then tested each of the candidate regulators’ distal eQTLs for association with the growth rate phenotype. After correction for multiple testing, we found that nearly all of the underlying variants attained significance at FDR < 0.05. We propose that differential expression of the putative regulators influences growth rate by perturbing common, growth-related pathways in trans.

We found numerous loci linking to small sets of target genes that are functionally related, as might be expected from the simulation results. TEC1, a transcription factor that targets filamentation genes, was found to have a significantly associated local variant that was distally linked to 16 genes enriched
for pseudohyphal growth annotation (p=1.03e-03). Additionally, for 5 of these 16 genes (31.2%), the YEASTRACT database shows direct evidence of TEC1 DNA binding and transcriptional regulation (Teixeira et al. 2014). Of the 25 genes that mapped to the lead variant (defined as the variant with strongest local effect on TEC1) in an all versus all test, only 4 (16%) showed direct evidence of TEC1 binding and regulation, suggesting that NetLIFT is better able to identify biologically relevant associations.

We identify several putative regulators that are metabolic enzymes and whose target gene sets are enriched for metabolic and biosynthesis annotations. For example, a locus on chromosome 2 that acts as a local eQTL for LYS2 was distally associated with 167 target genes enriched for the GO term “lysine biosynthetic process via aminoadipic acid” (p=1.27e-07). LYS2 catalyzes the reduction of alpha-aminoadipate to alpha-aminoadipate semialdehyde (αAASA), the fifth step in the lysine biosynthesis pathway. Downstream of this reaction, glutamate-forming saccharopne dehydrogenase, which consists of the structural determinant LYS9 and the regulatory product LYS14, converts αAASA to saccharopine. LYS9 loss of function increases intracellular levels of αAASA, which induces the regulatory activity of Lys14p and results in the up-regulation of several genes in the pathway, including LYS1, LYS9, LYS2, LYS4, LYS20, and LYS21 (Becker et al. 1998). In a previous experiment, a mutant strain with loss of function for both LYS2 and LYS9 was shown to have decreased intracellular αAASA and lower levels of transcriptional activation of pathway genes, relative to the LYS9 single mutant (RAMOS, DUBOIS, and PIERARD 1988; Feller et al. 1999). We hypothesize that strains harboring the genomic variant associated with decreased transcription of LYS2 will have a similar reduction of intracellular αAASA concentration, and thus a decreased potential for transcriptional activation of Lys14p. Of the previously mentioned lysine biosynthesis genes that are targeted by Lys14p, we find four linked distally to the putative eQTL (LYS1, LYS9, LYS20, and LYS21). We note that the direction of effect between the eQTL and the downstream
genes reflects what we expect under the proposed mechanism (Figure S6). Among the set of transcriptional targets are four additional genes whose promoters contain the \emph{Lys14p} binding motif, TCCRNYGGA, one of which, \emph{LYS12}, is involved in lysine biosynthesis and has a directional expression pattern matching the other \emph{Lys14p} targets (Figure S6).

\textbf{Analysis of 156 partially inbred mouse lines}

To test how well NetLIFT scales to larger data sets, and for organisms with more complex mechanisms of gene regulation, we analyzed paired genotype and liver gene expression data from 156 partially inbred mice originating from eight founder mice (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HILtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ), part of the Collaborative Cross (CC) project (Churchill et al. 2004; Collaborative Cross Consortium 2012). Founder strains of the CC were chosen to provide a high level of genetic diversity, and represent three subspecies of origin: \emph{Mus mus domesticus}, \emph{Mus mus castaneus}, and \emph{Mus mus musculus}. Wild-derived WSB/EiJ and classical inbred strains A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HILtJ have a genetic background comprised mostly of the \emph{Mus mus domesticus} subspecies, while the wild-derived CAST/EiJ and PWK/PhJ founder strains are primarily representative of the \emph{Mus mus castaneus} and \emph{Mus mus musculus} subspecies, respectively (Churchill et al. 2004; Collaborative Cross Consortium 2012).

We filtered for probe sets expressed above background levels and retained 9,377 genes for analysis. PCA analysis revealed no batch effects in the data (Figure S7). Genotypes for the same mice were available for 171,761 markers. In a previous analysis, a total of 6,182 eQTL were discovered for 5,733 genes at a 5\% genome-wide threshold; 75\% of eQTL were within 10cM of the affected gene (Aylor et al. 2011).
For eQTL testing, we defined local effects as those where variants were within 1Mb of the affected gene, based on the marker-to-gene distances for linkages reported previously for these data (Aylor et al. 2011). We detected a total of 5,744 genes (61%) with a local eQTL, and 3,322 (35%) with at least one distal eQTL (FDR < 0.05). Of the genes with a distal eQTL, 1,102 (12%) were linked to one SNP, 574 (6%) were linked to two SNPs, 400 (4%) were linked to three SNPs, and 1,246 (13%) were linked to four or more SNPs.

We next investigated patterns of large-scale effects on the regulatory architecture that are attributable to founder and/or subspecies of origin. For the 293 genes with a local eQTL that was linked to at least 5 genes on different chromosomes, genes inherited from a PWK genetic background showed more extreme expression variation than genes inherited from the other founder strains (Figure S8). Mice from the CC have been shown to be phenotypically diverse for various immune related phenotypes (Phillippi et al. 2013; Ferris et al. 2013), body weight (Philip et al. 2011), and behavior (Philip et al. 2011), with variance for some traits exceeding that observed in the founder strains (Philip et al. 2011). One plausible reason for this is that epistatic interactions between alleles inherited from distinct subspecies (castaneus, domesticus, and musculus) may severely mis-regulate gene expression and homeostasis. To investigate whether allele inheritance from different subspecies of origin led to more extreme expression for particular combinations of locally-acting eQTL alleles and target genes, we mapped both eQTL SNPs and target genes to their subspecies of origin. Since alleles inherited from PWK mice appeared to be driving extreme expression variation in locally-affected genes, we reduced the locally-affected set of genes to a subset of 61 genes for which the Mus musculus musculus-derived PWK allele explained at least half of the overall genetic effect on expression (Figure 5A). We observed that for these SNPs, expression of distally-linked genes showed differential variation based on the combinatorial genetic backgrounds of the locally-associated variant and target gene (Figure 5B).
These transcriptomic differences may in turn affect phenotype. Body weight for wild-derived founder strains (CAST/EiJ, PWK/PhJ, WSB/EiJ) used in the Collaborative Cross is lower than in classical laboratory strains (Aylor et al. 2011). A GO analysis performed for the 142 distal genes linking to the PWK-driven eQTL revealed annotation for various terms related to metabolism and lipid processes (Table 3). This enrichment suggests a possible role for the candidate trans acting factors in regulating weight, via a broad but subtle effect on gene expression.

**Analysis of 69 human individuals**

RNA-seq data from lymphoblastoid cell lines and HapMap genotype data for 69 Nigerian individuals were recently interrogated for eQTLs (Pickrell et al. 2010). For NetLIFT analysis, expression data was corrected for GC content and batch, and was normalized as described previously. We selected 9,810 Ensembl transcripts in the top quartile based on median expression level for further analysis. Genotype data for the same individuals, consisting of 9.5 million SNPs, were obtained from HapMap phase 2 and 3, release 27.

Using a local regulatory window of 200kb, similar to the original analysis (Pickrell et al. 2010), we identified 2,483 transcripts (25.3%) with a local eQTL effect (FDR < 0.10). Of the 929 transcripts previously identified as having local associations at the same FDR, we replicated 538. The remainder not found consisted of transcripts that we removed from the data set due to low median expression level, with the exception of 3 transcripts that were not identified in our analysis. In addition, we identified 1,945 novel local associations, likely attributable to greater power resulting from testing only the most highly expressed quartile of transcripts.
NetLIFT identified 1,274 transcripts (13.0%) with at least one distal eQTL (FDR < 0.10, Figure S9). None were reported in the previous analysis (Pickrell et al. 2010). A traditional all SNPs-vs-all genes testing approach on this filtered set of genes and variants yielded only 5 significant distal associations at this FDR, indicating that our method is better powered for detecting these associations. A GO analysis for the 64 candidate regulators that were linked to at least 3 transcripts (FDR < 0.1) again suggested enrichment for metabolic and biosynthetic processes (Table 4).
DISCUSSION

Genome Wide Association Studies (GWAS) have so far identified thousands of quantitative trait loci associated with hundreds of complex traits (Hindorff et al. 2009). However, the success of GWAS has been tempered by a lack of understanding of the mechanism of association for many variants. eQTL studies have shown excellent promise in highlighting potential biological mechanisms of SNP-phenotype associations, and prioritizing particular variants for follow up studies (Mehta et al. 2012). Furthermore, the correlation between significance levels of SNP-phenotype associations and eQTL associations may help to identify tissue types that play a key role in disease etiology (Kang et al. 2012). Recently, gene-gene interaction evidence has been incorporated in the GWAS setting to identify epistatic effects on phenotype (Ma, Clark, and Keinan 2013), suggesting that correlation based testing may increase power to detect associated variants. We described here a novel method, NetLIFT, that addresses the problems of computational burden and power in traditional eQTL testing, by reducing the search space and using conditional dependencies between genes’ expression to prioritize variant-gene testing. The reduced multiple testing correction penalty under our algorithm allows detection of weaker eQTL effects that are missed by currently available methods. Furthermore, our results provide immediate interpretability of the mechanism of association, by highlighting potential regulatory genes that mediate discovered distal effects. We note that in the current implementation of our code, runtime and memory usage increases nonlinearly as the number of genes increases, and that the the major bottleneck in runtime is the estimation of the partial correlation matrix. Therefore, when the number of genes exceeds 10,000, users may wish to filter gene expression data sets by most highly expressed or most variable genes.

Importantly, we showed through simulations that NetLIFT can identify instances where distal eQTL only affect a small number of genes, not just the large hub genes found by other methods. Additionally, candidate regulators that are putatively affected in cis by the causal variant can be
identified, highlighting potential mechanisms of association. We note that since our method seeks to identify distal effects that arise via alterations in the expression level of trans-acting factors located nearby the eQTL, we are unable to detect associations mediated by a loss-of-function coding variant in the trans-acting factor.

We demonstrated the ability of NetLIFT to identify distal eQTL in three very different data sets. In yeast segregants, we replicated numerous distal eQTL reported previously, as well as the biologically validated regulators for many of the associations. Additionally, we identified several novel biologically plausible distal associations. In inbred lines from genetically diverse founder mice, we detected an interesting pattern of eQTL effects driven by PWK-derived alleles, which may provide clues as molecular underpinnings of downstream phenotypes such as reduced mouse size in the wild-type derived PWK mice. Lastly, in a set of 69 human individuals, NetLIFT was able to find over 1,200 gene transcripts with significant distal eQTL due to its increased power, whereas previously only 5 had been identified.

Intuitively, one might think that the best candidates for asserting regulatory influence on distal genes would be transcription factors that directly participate in controlling gene transcription rates. In accordance with previous results, though, we found no enrichment for transcription factor annotation among genes implicated by our method as trans-acting factors; instead, we find that many of these genes play a role in metabolic and biosynthesis pathways. This suggests that more commonly, the regulation of key genes in these pathways play a role in feedforward or feedback processes that then affect transcription rates of downstream target genes within the same pathway. These indirect effects are more subtle than the direct effects associated with local eQTL, but they can have significant effects on phenotypes, such as growth rates (seen in yeast) and size (seen in mouse).
Our results also highlight an often unaddressed topic in complex trait mapping; namely, that eQTL discovery and interpretability of mapping results is significantly influenced by the genetic and genomic diversity within the sample population. The two yeast strains from which the analyzed segregants were derived were extremely diverse, with an estimated sequence divergence of 0.5-1%. This, and overall genome complexity, likely contributed to many distal effects being found to be as strong as local effects, enabling their easier detection. Genetic incompatibilities between progenitors can result in atypical patterns of linkage disequilibrium, which present challenges in identifying causal versus linked markers. In an inbred mouse model, we were able to identify numerous distal linkages where expression variation in the distally-affected genes appears to be driven by differences in the genetic background at the local and distal loci. However, the resolution of the eQTL mapping is ultimately restricted by the randomization of the genome that is mediated by recombination events. On the other hand, human studies typically involve genetically diverse individuals, whose genomes are randomized to a greater extent. Thus a model organism may allow for accurate eQTL mapping at the expense of precision, whereas in human populations we expect to identify eQTL with precision, but reduced accuracy.
ACKNOWLEDGEMENTS

The authors wish to thank Dr. Rachel Brem for providing genotype and eQTL data for yeast strains, as well as Dr. Steve Haase, Dr. David Aylor, and Dr. Fernando Pardo-Manuel de Villena for their contributions in interpreting eQTL results from yeast and pre-Collaborative Cross mice.
REFERENCES

Airoldi, Edoardo M, Curtis Huttenhower, David Gresham, Charles Lu, Amy A Caudy, Maitreya J Dunham, James R Broach, David Botstein, and Olga G Troyanskaya. 2009. “Predicting Cellular Growth from Gene Expression Signatures.” Edited by Andrey Rzhetsky. PLoS Computational Biology 5 (1). Public Library of Science: e1000257. doi:10.1371/journal.pcbi.1000257.

Alberts, Rudi, Hairong Chen, Claudia Pommerenke, August B Smit, Sabine Spijker, Robert W Williams, Robert Geffers, Dunja Bruder, and Klaus Schughart. 2011. “Expression QTL Mapping in Regulatory and Helper T Cells from the BXD Family of Strains Reveals Novel Cell-Specific Genes, Gene-Gene Interactions and Candidate Genes for Auto-Immune Disease.” BMC Genomics 12 (1): 610. doi:10.1186/1471-2164-12-610.

Allen, Jeffrey D, Yang Xie, Min Chen, Luc Girard, and Guanghua Xiao. 2012. “Comparing Statistical Methods for Constructing Large Scale Gene Networks.” Edited by Petter Holme. PloS One 7 (1). Public Library of Science: e29348. doi:10.1371/journal.pone.0029348.

Aylor, David L, William Valdar, Wendy Foulds-Mathes, Ryan J Buus, Ricardo A Verdugo, Ralph S Baric, Martin T Ferris, et al. 2011. “Genetic Analysis of Complex Traits in the Emerging Collaborative Cross.” Genome Research 21 (8): 1213–22. doi:10.1101/gr.111310.110.

Barabási, Albert-László, and Zoltán N Oltvai. 2004. “Network Biology: Understanding the Cell’s Functional Organization.” Nature Reviews. Genetics 5 (2): 101–13. doi:10.1038/nrg1272.

Becker, Benedicte, Andre Feller, Mohamed El Alami, Evelyne Dubois, and Andre Pierard. 1998. “A Nonameric Core Sequence Is Required Upstream of the LYS Genes of Saccharomyces Cerevisiae for Lys14p-Mediated Activation and Apparent Repression by Lysine.” Molecular Microbiology 29 (1): 151–63. doi:10.1046/j.1365-2958.1998.00916.x.

Benjamini, Yoav, and Yosef Hochberg. 1995. “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. doi:10.2307/2346101.

Benjamini, Yoav, and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing under Dependency.” The Annals of Statistics 29 (4): 1165–88. http://projecteuclid.org/euclid.aos/1013699998.

Bottolo, Leonardo, Enrico Petretto, Stefan Blankenberg, François Cambien, Stuart A Cook, Laurence Tiret, and Sylvia Richardson. 2011. “Bayesian Detection of Expression Quantitative Trait Loci Hot Spots.” Genetics 189 (4): 1449–59. doi:10.1534/genetics.111.131425.

Breitling, Rainer, Yang Li, Bruno M Tesson, Jingyuan Fu, Chunlei Wu, Tim Wiltshire, Alice Gerrits, et al. 2008. “Genetical Genomics: Spotlight on QTL Hotspots.” Edited by Gary A. Churchill. PLoS Genetics 4 (10). Public Library of Science: e1000232. doi:10.1371/journal.pgen.1000232.

Brem, Rachel B, and Leonid Kruglyak. 2005. “The Landscape of Genetic Complexity across 5,700 Gene Expression Traits in Yeast.” Proceedings of the National Academy of Sciences of the United States of America 102 (5): 1572–77. doi:10.1073/pnas.0408709102.
Brem, Rachel B, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. 2002. “Genetic Dissection of Transcriptional Regulation in Budding Yeast.” *Science (New York, N.Y.)* 296 (5568): 752–55. doi:10.1126/science.1069516.

Chen, Lin S, Frank Emmert-Streib, and John D Storey. 2007. “Harnessing Naturally Randomized Transcription to Infer Regulatory Relationships among Genes.” *Genome Biology* 8 (10): R219. doi:10.1186/gb-2007-8-10-r219.

Cheung, Vivian G, Laura K Conlin, Teresa M Weber, Melissa Arcaro, Kuang-Yu Jen, Michael Morley, and Richard S Spielman. 2003. “Natural Variation in Human Gene Expression Assessed in Lymphoblastoid Cells.” *Nature Genetics* 33 (3). Nature Publishing Group: 422–25. doi:10.1038/ng1094.

Churchill, Gary A, David C Airey, Hooman Allayee, Joe M Angel, Alan D Attie, Jackson Beatty, William D Beavis, et al. 2004. “The Collaborative Cross, a Community Resource for the Genetic Analysis of Complex Traits.” *Nature Genetics* 36 (11): 1133–37. doi:10.1038/ng1104-1133.

Collaborative Cross Consortium. 2012. “The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population.” *Genetics* 190 (2): 389–401. doi:10.1534/genetics.111.132639.

Doss, Sudheer, Eric E Schadt, Thomas A Drake, and Aldons J Lusis. 2005. “Cis-Acting Expression Quantitative Trait Loci in Mice.” *Genome Research* 15 (5): 681–91. doi:10.1101/gr.3216905.

Duarte, Christine W, and Zhao-Bang Zeng. 2011. “High-Confidence Discovery of Genetic Network Regulators in Expression Quantitative Trait Loci Data.” *Genetics* 187 (3): 955–64. doi:10.1534/genetics.110.124685.

Feller, Andre, Fernando Ramos, Andre Pierard, and Evelyne Dubois. 1999. "In Saccharomyces Cerevisiae, Feedback Inhibition of Homocitrate Synthase Isoenzymes by Lysine Modulates the Activation of LYS Gene Expression by Lys14p." *European Journal of Biochemistry* 261 (1): 163–70. doi:10.1046/j.1432-1327.1999.00262.x.

Ferris, Martin T, David L Aylor, Daniel Bottomly, Alan C Whitmore, Lauri D Aicher, Timothy A Bell, Birgit Bradel-Tretheway, et al. 2013. "Modeling Host Genetic Regulation of Influenza Pathogenesis in the Collaborative Cross." *PLoS Pathogens* 9 (2): e1003196. doi:10.1371/journal.ppat.1003196.

Gat-Viks, Irit, Renana Meller, Martin Kupiec, and Ron Shamir. 2010. “Understanding Gene Sequence Variation in the Context of Transcription Regulation in Yeast.” *PLoS Genetics* 6 (1): e1000800. doi:10.1371/journal.pgen.1000800.

Göring, Harald H H, Joanne E Curran, Matthew P Johnson, Thomas D Dyer, Jac Charlesworth, Shelley A Cole, Jeremy B M Jowett, et al. 2007. “Discovery of Expression QTLs Using Large-Scale Transcriptional Profiling in Human Lymphocytes.” *Nature Genetics* 39 (10): 1208–16. doi:10.1038/ng2119.

Hindorff, Lucia A, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. 2009. “Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): 9362–67. doi:10.1073/pnas.0903103106.
Holloway, Beth, Stanley Luck, Mary Beatty, J-Antoni Rafalski, and Bailin Li. 2011. “Genome-Wide Expression Quantitative Trait Loci (eQTL) Analysis in Maize.” *BMC Genomics* 12 (January): 336. doi:10.1186/1471-2164-12-336.

Jeong, H, B Tombor, R Albert, Z N Oltvai, and A L Barabási. 2000. “The Large-Scale Organization of Metabolic Networks.” *Nature* 407 (6804): 651–54. doi:10.1038/35036627.

Kang, Hyunsoek Peter, Alex A Morgan, Rong Chen, Eric E Schadt, and Atul J Butte. 2012. “Coanalysis of GWAS with eQTLs Reveals Disease-Tissue Associations.” *AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science* 2012 (January): 35–41. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3392070&tool=pmcentrez&rendertype=abstract.

Kompass, Kenneth S, and John S Witte. 2011. “Co-Regulatory Expression Quantitative Trait Loci Mapping: Method and Application to Endometrial Cancer.” *BMC Medical Genomics* 4 (January): 6. doi:10.1186/1755-8794-4-6.

Lappalainen, Tuuli, Michael Sammeth, Marc R. Friedländer, Peter A. C. ‘t Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, et al. 2013. “Transcriptome and Genome Sequencing Uncovers Functional Variation in Humans.” *Nature* 501 (7468). Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 506–11. doi:10.1038/nature12531.

Litvin, Oren, Helen C Causton, Bo-Juen Chen, and Dana Pe’er. 2009. “Modularity and Interactions in the Genetics of Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (16): 6441–46. doi:10.1073/pnas.0810208106.

Lorenz, W Walter, Rob Alba, Yuan-Sheng Yu, John M Bordeaux, Marta Simões, and Jeffrey F D Dean. 2011. “Microarray Analysis and Scale-Free Gene Networks Identify Candidate Regulators in Drought-Stressed Roots of Loblolly Pine (P. Taeda L.).” *BMC Genomics* 12 (1): 264. doi:10.1186/1471-2164-12-264.

Ma, Li, Andrew G Clark, and Alon Keinan. 2013. “Gene-Based Testing of Interactions in Association Studies of Quantitative Traits.” Edited by Scott M. Williams. *PLoS Genetics* 9 (2). Public Library of Science: e1003321. doi:10.1371/journal.pgen.1003321.

Mehta, Divya, Katharina Heim, Christian Herder, Maren Carstensen, Gertrud Eckstein, Claudia Schurrmann, Georg Homuth, et al. 2012. “Impact of Common Regulatory Single-Nucleotide Variants on Gene Expression Profiles in Whole Blood.” *European Journal of Human Genetics : EJHG*, June. Macmillan Publishers Limited. doi:10.1038/ejhg.2012.106.

Neto, Elias Chaibub, Aimee T Broman, Mark P Keller, Alan D Attie, Bin Zhang, Jun Zhu, and Brian S Yandell. 2013. “Modeling Causality for Pairs of Phenotypes in System Genetics.” *Genetics* 193 (3): 1003–13. doi:10.1534/genetics.112.147124.

Peng, Jie, Pei Wang, Nengfeng Zhou, and Ji Zhu. 2009. “Partial Correlation Estimation by Joint Sparse Regression Models.” *Journal of the American Statistical Association* 104 (486): 735–46. doi:10.1198/jasa.2009.0126.
Philip, Vivek M, Greta Sokoloff, Cheryl L Ackert-Bicknell, Martin Striz, Lisa Branstetter, Melissa A Beckmann, Jason S Spence, et al. 2011. “Genetic Analysis in the Collaborative Cross Breeding Population.” *Genome Research* 21 (8): 1223–38. doi:10.1101/gr.113886.110.

Phillippi, J, Y Xie, D R Miller, T A Bell, Z Zhang, A B Lenarcic, D L Aylor, et al. 2013. “Using the Emerging Collaborative Cross to Probe the Immune System.” *Genes and Immunity*, November. doi:10.1038/gene.2013.59.

Pickrell, Joseph K, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. 2010. “Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing.” *Nature* 464 (7289). Macmillan Publishers Limited. All rights reserved: 768–72. doi:10.1038/nature08872.

RAMOS, Fernando, Evelyne DUBOIS, and Andre PIERARD. 1988. “Control of Enzyme Synthesis in the Lysine Biosynthetic Pathway of Saccharomyces Cerevisiae. Evidence for a Regulatory Role of Gene LYS14.” *European Journal of Biochemistry* 171 (1-2): 171–76. doi:10.1111/j.1432-1033.1988.tb13773.x.

Romano, Joseph P., Azeem M. Shaikh, and Michael Wolf. 2008. “Control of the False Discovery Rate under Dependence Using the Bootstrap and Subsampling.” *TEST* 17 (3): 417–42. doi:10.1007/s11749-008-0126-6.

Rotival, Maxime, Tanja Zeller, Philipp S Wild, Seraya Maouche, Silke Szymczak, Arne Schillert, Raphaelle Castagné, et al. 2011. “Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans.” Edited by Gregory S. Barsh. *PLoS Genetics* 7 (12). Public Library of Science: e1002367. doi:10.1371/journal.pgen.1002367.

Schadt, Eric E, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, et al. 2003. “Genetics of Gene Expression Surveyed in Maize, Mouse and Man.” *Nature* 422 (6929). Nature Publishing Group: 297–302. doi:10.1038/nature01434.

Shabalin, Andrey A. 2012. “Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations.” *Bioinformatics (Oxford, England)* 28 (10): 1353–58. doi:10.1093/bioinformatics/bts163.

Smith, Erin N, and Leonid Kruglyak. 2008. “Gene-Environment Interaction in Yeast Gene Expression.” Edited by Trudy Mackay. *PLoS Biology* 6 (4). Public Library of Science: e83. doi:10.1371/journal.pbio.0060083.

Teixeira, Miguel Cacho, Pedro Tiago Monteiro, Joana Fernandes Guerreiro, Joana Pinho Gonçalves, Nuno Pereira Mira, Sandra Costa dos Santos, Tânia Rodrigues Cabrito, et al. 2014. “The YEASTRACT Database: An Upgraded Information System for the Analysis of Gene and Genomic Transcription Regulation in Saccharomyces Cerevisiae.” *Nucleic Acids Research* 42 (Database issue): D161–6. doi:10.1093/nar/gkt1015.

West, Marilyn A L, Kyunga Kim, Daniel J Kliebenstein, Hans van Leeuwen, Richard W Michelmore, R W Doerge, and Dina A St Clair. 2007. “Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis.” *Genetics* 175 (3): 1441–50. doi:10.1534/genetics.106.064972.
Yook, Soon-Hyung, Zoltán N Oltvai, and Albert-László Barabási. 2004. “Functional and Topological Characterization of Protein Interaction Networks.” *Proteomics* 4 (4): 928–42. doi:10.1002/pmic.200300636.

Yvert, Gaël, Rachel B Brem, Jacqueline Whittle, Joshua M Akey, Eric Foss, Erin N Smith, Rachel Mackelprang, and Leonid Kruglyak. 2003. “Trans-Acting Regulatory Variation in *Saccharomyces Cerevisiae* and the Role of Transcription Factors.” *Nature Genetics* 35 (1): 57–64. doi:10.1038/ng1222.

Zhu, Jun, Pavel Sova, Qiuwei Xu, Kenneth M Dombek, Ethan Y Xu, Heather Vu, Zhidong Tu, Rachel B Brem, Roger E Bumgarner, and Eric E Schadt. 2012. “Stitching Together Multiple Data Dimensions Reveals Interacting Metabolomic and Transcriptomic Networks That Modulate Cell Regulation.” Edited by Andre Levchenko. *PLoS Biology* 10 (4). Public Library of Science: e1001301. doi:10.1371/journal.pbio.1001301.

Zhu, Jun, Bin Zhang, Erin N Smith, Becky Drees, Rachel B Brem, Leonid Kruglyak, Roger E Bumgarner, and Eric E Schadt. 2008. “Integrating Large-Scale Functional Genomic Data to Dissect the Complexity of Yeast Regulatory Networks.” *Nature Genetics* 40 (7): 854–61. doi:10.1038/ng.167.
Figure 1. Schematic of the NetLIFT method. Top: genotypes for 'm' markers ($s_1, s_2, \ldots, s_m$) and 'p' genes ($g_1, g_2, \ldots, g_p$) are assayed for the same 'n' individuals ($a_1, a_2, \ldots, a_n$). Markers and genes which map to the same locus are color coded. Local eQTL mapping is performed for markers and nearby genes using an a priori defined genomic distance for local effects (A), yielding a local eQTL effect matrix (significant marker-gene associations depicted in green). A sparse partial correlation matrix is inferred from the expression data, representing a network of gene-gene interactions (B). Finally, significantly associated local eQTL markers are tested for distal eQTL effects on genes near the locally-affected gene in the interaction network (C).
**Figure 2. Number of detected distal associations, by module topology/method.** Topology of each network module is depicted in header. Black nodes depict genes with an assigned local eQTL effect, and red nodes represent “true” distally-associated genes. Total number of “true” distal associations given in parentheses. Each cell value reports the mean and standard deviation of TP / FP, over the ten simulated data sets. Cells are colored according to fraction of true positives discovered. Rightmost column (bottom row) reports the number of false positive distal associations where the locally-regulated gene and target gene belonged to disjoint modules.
Figure 3. Distal eQTL associations in pre-Collaborative Cross mice. X axis gives the genomic coordinates of marker SNPs; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.05, for markers that were at least 1Mb from the associated gene.
Figure 4. Expression variability for PWK-driven \textit{trans}-acting factors and target genes, in pre-Collaborative Cross mice. Top: Distribution of absolute expression deviation from median, for putative \textit{trans}-acting factors with a PWK-driven local eQTL, grouped by founder strain genetic background at the eQTL locus. Only putative \textit{trans}-acting factors that were linked to at least 5 target genes on a different chromosome were considered. Bottom: Expression distribution for target genes of PWK-driven eQTL loci, stratified by subspecies of origin allele (\textit{castaneus/domesticus/musculus}) at both the local and distal loci. Each boxplot represents the expression deviation for all target genes, for each possible combination of local/distal alleles.
Figure S1. Simulated gene module topologies. Each module’s expression effects were simulated by first generating the hub gene’s expression; each successive downstream gene’s expression values were simulated using the upstream gene’s expression as a baseline (dependencies indicated by arrows). For each module, a single local eQTL effect was simulated for a SNP assigned to either the hub gene (black), or to a gene downstream of the hub (red), but not both.
**Figure S2.** Illustration of eQTL detection methods. SNP is depicted as a red node; genes depicted in green. **A** performs a standard regression significance test for all pairs of SNPs and genes. **B** seeks to identify distal associations that are mediated by a locally associated variant-gene pair (local associations depicted with blue arrows). Genes downstream of the inferred direction of gene-gene effects (represented by green arrows) should be associated with the variant (true distal associations = solid black arrows), while genes upstream of gene effects will not show association (dashed black arrows). **C** first factors expression data into Independent Components, then performs association tests between allele frequency and the activation levels of components across samples. **D** first performs local linkage tests for a SNP and nearby genes (blue arrows). For significant linkages (solid blue arrow), distal eQTL tests are performed for all genes in the network which are one- or two-edges removed from the locally affected gene (black arrows).
**Figure S3.** Partial correlation structure from network detection step, for representative 100 gene, 50 gene, and 10 gene modules. True positive correlations depicted with green edges, false negatives correlations in red, false positives in gray.
**Figure S4.** Local and distal eQTL linkages in yeast. X axis shows the genomic coordinates of marker variants; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.05.
**Figure S5.** Pairwise overlap of target gene sets enriched for ribosomal annotation. Cell [i,j] shows the target gene overlap for between proposed regulators i, j.
**Figure S6.** eQTL effects for LYS2 local regulatory variant and downstream genes. The allele associated with lower LYS2 expression (“0”) is associated with lower expression of known Lys14p targets LYS2, LYS1, LYS9, LYS20, and LYS21. The same allele also associates with higher expression of three non-LYS genes containing Lys14p binding motifs (DYS1, TOP2, DAD2), and the Lys14p motif-containing LYS12.
Figure S7. PCA analysis for pre-CC mice. Top two principal components for gene expression data in 156 pre-CC mice.
Figure S8. Expression variability by founder strain, for locally-regulated genes with at least 5 distal targets. Gene expression values were binned according to the genetic background of the locally-affected gene. Violin plot shows the level of variation compared to the overall sample expression medians, for each of the eight founder strains.
**Figure S9.** Local and distal eQTL linkages in human lymphoblastoid cell lines. X axis shows the genomic coordinates of SNPs; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.1.
Table 1. GO ANNOTATION ENRICHMENT FOR CANDIDATE REGULATORS IN YEAST. GO analysis was performed for genes with >= 10 distal associations; top 20 enrichment terms reported in right column.

| Pvalue   | Term                                                |
|----------|-----------------------------------------------------|
| 2.00E-06 | asparagine catabolic process                        |
| 5.89E-06 | cellular response to nitrogen starvation            |
| 5.89E-06 | cellular response to nitrogen levels                |
| 4.66E-05 | asparagine metabolic process                        |
| 4.90E-05 | glutamine family amino acid catabolic process       |
| 0.000172 | aspartate family amino acid catabolic process       |
| 0.001328 | cellular response to nutrient levels                |
| 0.001784 | response to nutrient levels                         |
| 0.001784 | cellular response to extracellular stimulus         |
| 0.001784 | cellular response to external stimulus              |
| 0.002359 | response to external stimulus                       |
| 0.002359 | response to extracellular stimulus                  |
| 0.003704 | cellular amino acid catabolic process               |
| 0.003936 | developmental process involved in reproduction      |
| 0.004111 | cellular response to starvation                     |
| 0.005043 | response to starvation                              |
| 0.005191 | amino acid transmembrane transport                 |
| 0.005905 | carbon catabolite regulation of transcription from RNA polymerase II promoter |
| 0.005931 | copper ion transport                                |
| 0.007164 | viral reproduction                                  |
Table 2. DISTAL REGULATORY LOCI AND CANDIDATE REGULATORS IDENTIFIED IN YEAST. First column indicates eQTL identified by: previous methods and NetLIFT (**); NetLIFT only (**); previous methods only (*). Third and fourth columns list candidate regulators implicated by NetLIFT, previous methods, respectively. Fifth column gives the number of genes linked to the locus by NetLIFT. Top GO enrichment for linked transcripts listed in sixth column. For eQTL on chromosome 2 that were linked to genes with ncRNA and ribosomal annotation, association testing was performed for the marker and growth rate phenotype (far right column).

| Method | eQTL Position | TAF | Previously Predicted Regulators | # Targets | GO Annotation Enrichment | GO pVal | FDR - Growth Rate Association |
|--------|---------------|-----|---------------------------------|-----------|--------------------------|---------|-------------------------------|
| ***    | chrII:376668  | TAT1| TRM7(Gat-Viks et al. 2010)      | 265       | cytoplasmic translation  | 9.63E-37| NA                            |
| ***    | chrII:555596  | AMN1| AMN1(Gat-Viks et al. 2010; Yvert et al. 2003), MAK5(Yvert et al. 2003) | 307       | ribosomal biogenesis     | 2.90E-12| 0.0036                        |
| ***    | chrII:697894  | GPX2| None(Gat-Viks et al. 2010; Yvert et al. 2003) | 205       | ncRNA processing         | 1.53E-17| 0.012                         |
| ***    | chrIII:92127  | LEU2| LEU2(Gat-Viks et al. 2010; Yvert et al. 2003; Zhu et al. 2008; Zhu et al. 2012) | 113       | organic acid biosynthetic process | 4.05E-25| NA                            |
| ***    | chrIII:105042 | ILV6| ILV6(Zhu et al. 2008; Zhu et al. 2012) | 93        | organic acid biosynthetic process | 2.45E-22| NA                            |
| ***    | chrIII:201116 | MATALPHA1 | MATALPHA1(Gat-Viks et al. 2010; Yvert et al. 2003; Smith and Kruglyak 2008; Zhu et al. 2008) | 40        | response to pheromone      | 1.78E-08| NA                            |
| ***    | chrV:117056   | URA3| URA3(Gat-Viks et al. 2010; Yvert et al. 2003; Zhu et al. 2008; Zhu et al. 2012) | 28        | ‘de novo’ UMP biosynthetic process | 8.66E-09| NA                            |
| ***    | chrVIII:111682| GPA1| GPA1(Gat-Viks et al. 2010; Yvert et al. 2003; Litvin et al. 2009; Smith and Kruglyak 2008; Zhu et al. 2008) | 29        | conjugation                | 1.14E-15| NA                            |
| ***    | chrXII:659357 | HAP1| HAP1(Gat-Viks et al. 2010; Yvert et al. 2003; Litvin et al. 2009; Smith and Kruglyak 2008; Zhu et al. 2008) | 29        | steroid metabolic process  | 3.80E-09| NA                            |
| ***    | chrXII:1067121| YLR464W | YRF1-4(Gat-Viks et al. 2010), YRF1-5(Gat-Viks et al. 2010), YLR464(Gat-Viks et al. 2010) | 15        | telomere maintenance via recombination | 1.81E-05| NA                            |
| **chr**  | **Position** | **Gene** | **Reference(s)** | **Fisher Test Value** | **Log10(P Value)** | **Metabolic Function** |
|----------|--------------|----------|------------------|----------------------|--------------------|----------------------|
| chrIV:371953 | NAM9 | MKT1 (Zhu et al. 2008), SAL1 (Zhu et al. 2008) | 25 | mitochondrial translation | 1.55E-21 | NA |
| chrXV:174364 | PHM7 | PHM7 (Zhu et al. 2008; Zhu et al. 2012), IRA2 (Litvin et al. 2009; Smith and Kruglyak 2008) | 107 | cellular ketone metabolic process | 8.89E-08 | NA |
| chrXV:382531 | CRS5 | CAT5 (Gat-Viks et al. 2010; Yvert et al. 2003) | 11 | cellular respiration | 3.77E-05 | NA |
| chrI:11638 | SEO1 | NA | 17 | monocarboxylic acid metabolic process | 1.11E-06 | NA |
| chrII:376872 | NRG2 | NA | 32 | asparagine catabolic process | 1.85E-06 | NA |
| chrII:401568 | TEC1 | NA | 16 | pseudohyphal growth | 1.03E-03 | NA |
| chrII:477206 | LYS2 | NA | 167 | lysine biosynthetic process via aminoadipic acid | 1.27E-07 | NA |
| chrIV:96259 | HEM3 | NA | 21 | cytokinesis | 5.47E-04 | NA |
| chrIV:1149761 | FCF1 | NA | 18 | endonucleolytic cleavage involved in rRNA processing | 4.02E-04 | NA |
| chrV:420595 | LCP5 | NA | 102 | ncRNA metabolic process | 1.90E-13 | NA |
| chrV:504714 | YER160C | NA | 19 | DNA integration | 6.65E-24 | NA |
| chrVII:402871 | PRM8 | NA | 23 | cellular zinc ion homeostasis | 5.72E-06 | NA |
| chrVII:916675 | ZPR1 | NA | 27 | ribosome biogenesis | 2.56E-05 | NA |
| chrX:24739 | REE1 | NA | 18 | formate metabolic process | 3.32E-08 | NA |
| chrX:262593 | SIP4 | NA | 17 | mitochondrial outer membrane translocase complex assembly | 2.03E-04 | NA |
| chrXII:126934 | PUF3 | NA | 22 | transposition, RNA-mediated | 1.01E-06 | NA |
| chrXII:468981 | ASP3-1 | NA | 50 | oxidation-reduction process | 7.84E-07 | NA |
| chrXII:956366 | PUN1 | NA | 64 | beta-alanine metabolic process | 1.29E-04 | NA |
| chrXII:28694 | PHO84 | NA | 32 | negative regulation of catalytic activity | 5.17E-05 | NA |
| chrXVI:523450 | SWI1 | NA | 40 | regulation of DNA metabolic process | 2.63E-04 | NA |

*** chrXIII:149075 | NA | SMA2 (Zhu et al. 2008) | NA | NA | NA | NA |

** chrXIII:149075 | SMA2 (Zhu et al. 2008) | NA | NA | NA | NA | NA |
Table 3. GO ENRICHMENTS FOR DISTAL GENES LINKING TO PWK-DRIVER eQTL IN PRE-COLLABORATIVE CROSS MICE. GO analysis was performed for the pooled set of genes that linked to a PWK founder-driven eQTL with at least 5 distal effects; top 20 GO enrichments are reported in right column.

| Pvalue          | Term                                                |
|-----------------|-----------------------------------------------------|
| 0.00116742      | malate metabolic process                            |
| 0.00192771      | progesterone metabolic process                      |
| 0.00192771      | negative regulation of nitric oxide biosynthetic process |
| 0.002854168     | organic acid metabolic process                      |
| 0.003725601     | carboxylic acid metabolic process                   |
| 0.004640455     | small molecule metabolic process                    |
| 0.00524957      | positive regulation of heart contraction             |
| 0.005659446     | lipid transport                                     |
| 0.005687178     | oxoacid metabolic process                           |
| 0.006687313     | phagocytosis, engulfment                            |
| 0.006687313     | complement activation, alternative pathway           |
| 0.007432993     | steroid metabolic process                            |
| 0.008282274     | protein targeting to plasma membrane                |
| 0.009233885     | monocarboxylic acid metabolic process               |
| 0.010029798     | regulation of the force of heart contraction         |
| 0.010029798     | C21-steroid hormone metabolic process               |
| 0.010037416     | cellular response to lipid                          |
| 0.011642566     | lipid localization                                  |
| 0.011925326     | natural killer cell differentiation                 |
| 0.011925326     | membrane invagination                               |
Table 4. GO TERM ENRICHMENT FOR PUTATIVE TRANS-ACTING FACTORS IN HUMAN LBCs. GO analysis was performed for the set of putative trans-acting factors linked to >= 3 distal genes; enrichments at significance p < 0.01 reported in right column.

| Pvalue     | Term                                               |
|------------|----------------------------------------------------|
| 8.27E-05   | folic acid metabolic process                       |
| 0.000759   | folic acid-containing compound metabolic process   |
| 0.001212   | one-carbon metabolic process                       |
| 0.001766   | pteridine-containing compound metabolic process    |
| 0.00537    | histidine biosynthetic process                     |
| 0.00537    | glycyl-tRNA aminoacylation                         |
| 0.00537    | histidine metabolic process                        |
| 0.00537    | regulation of hippo signaling cascade              |
| 0.00537    | imidazole-containing compound metabolic process    |