Novel Distal eQTL Analysis Demonstrates Effect of Population Genetic Architecture on Detecting and Interpreting Associations

Matthew Weiser,* Sayan Mukherjee,† and Terrence S. Furey‡,[1]
*Curriculum in Bioinformatics and Computational Biology, ‡Departments of Genetics and Biology, Carolina Center for Genome Sciences, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, and †Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, North Carolina 27708

ABSTRACT Mapping expression quantitative trait loci (eQTL) has identified genetic variants associated with transcription rates and has provided insight into genotype–phenotype associations obtained from genome-wide association studies (GWAS). Traditional eQTL mapping methods present significant challenges for the 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, "**net**work-based, **l**arge-scale **i**dentification o**f** dis**t**al eQTL" (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 5744 genes with local eQTL while 3322 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 1274 transcripts with distal eQTL that had not been reported in previous studies, while 2483 transcripts with local eQTL were identified. In all species, we found no enrichment for transcription factors facilitating eQTL associations; instead, we found that most *trans*-acting factors were annotated for metabolic function, suggesting that genetic variation may indirectly regulate multigene 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.

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; Doss *et al.* 2005; West *et al.* 2007). 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 (Alberts *et al.* 2011; Holloway *et al.* 2011; Mehta *et al.* 2012), 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 presents significant problems, especially in detecting

doi: 10.1534/genetics.114.167791
Manuscript received July 29, 2014; accepted for publication September 9, 2014; published Early Online September 16, 2014.
Available freely online through the author-supported open access option.
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167791/-/DC1.
[1]Corresponding author: 120 Mason Farm Rd., CB 7240, Genetic Medicine Bldg., University of North Carolina, Chapel Hill, NC 27599. E-mail: tsfurey@email.unc.edu

distal eQTL. Using this technique, the reported frequency of distal effects has varied from 2% to 75% of all detected eQTL (Yvert *et al.* 2003; Göring *et al.* 2007; Mehta *et al.* 2012), 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 are 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 (Bottolo *et al.* 2011; Duarte and Zeng 2011; Kompass and Witte 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, several drawbacks remain. 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. Second, the intermediate genes regulating the expression of gene modules are often not identified. Finally, expression for individual genes belonging to a module does 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 *et al.* 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 the target gene is causal. In this study, we present a novel eQTL detection method, "**net**work-based, **l**arge-scale **i**dentification o**f** dis**t**al 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 *et al.* 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 >1000 distal associations not previously reported. We note that individuals from each of these three populations (yeast, mice, and humans) have unique genetic histories, and our analysis suggests that this influences the number and type of eQTL detected in each study.

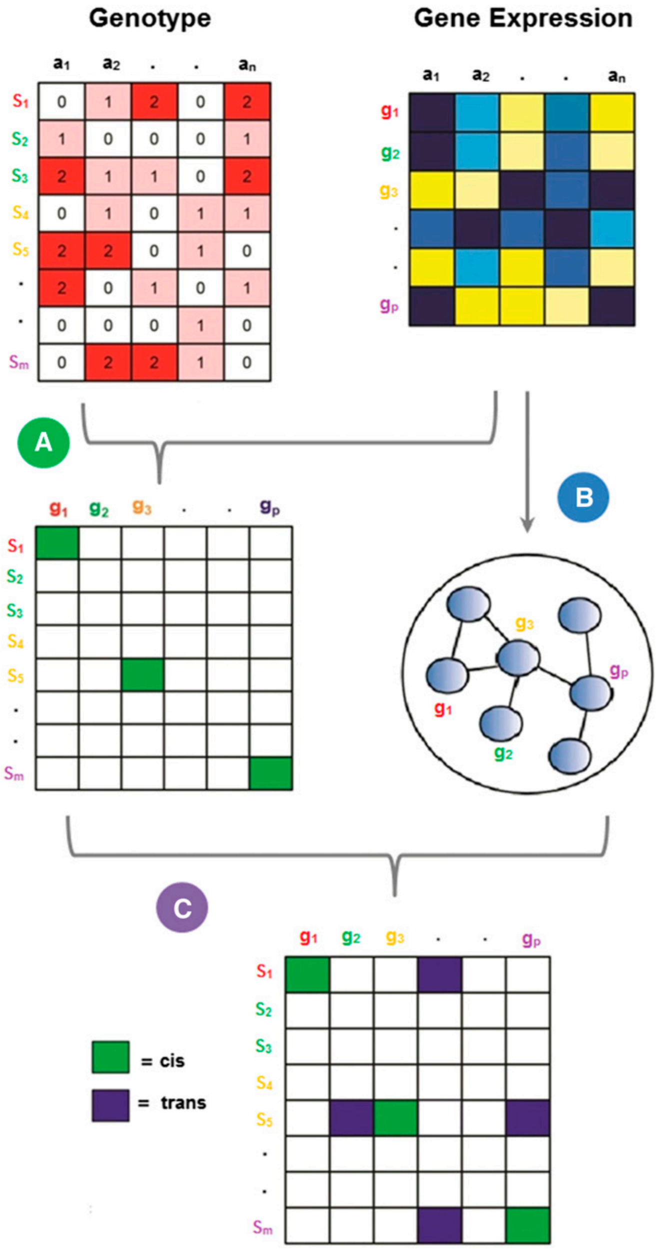
Materials and Methods

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_j 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 gene's 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



	a_1	a_2	.	.	a_n
S_1	0	1	2	0	2
S_2	1	0	0	0	1
S_3	2	1	1	0	2
S_4	0	1	0	1	1
S_5	2	2	0	1	0
.	2	0	1	0	1
.	0	0	0	1	0
S_m	0	2	2	1	0

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 that 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).

(Jeong *et al.* 2000; Barabási and Oltvai 2004; Yook *et al.* 2004; Lorenz *et al.* 2011) with an overall small number 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 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

$$L = \frac{1}{2}\left(\sum_{i=1}^{p}\left\|\mathbf{g_i} - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\mathbf{g_j}\right\|^2\right) + \lambda\sum_{1\leq i<j\leq p}\left|\rho^{ij}\right|,$$

where $\mathbf{g_i}$ and $\mathbf{g_j}$ are the expression vectors for genes *i* and *j*, ρ^{ij} denotes the partial correlation between genes *i* and *j*, and σ^{ii} and σ^{jj} are the *i*th and *j*th diagonal entries of the inverse covariance matrix. The L1 penalty λ controls the sparsity of the network and was optimized by minimizing the Bayesian information criterion outlined in Peng *et al.* (2009).

For *p* genes, the resulting $p \times p$ matrix **G** consists of entries $G_{i,j}$ that represent the correlation between expression vectors $\mathbf{g_i}$ and $\mathbf{g_j}$, conditioned on the expression of all other genes' expression:

$$G_{ij} = \operatorname{corr}\left(g_i, g_j \middle| g_k, k\neq i,j\right).$$

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 **G** is nonzero.

Step 3: Distal eQTL testing: Distal eQTL are called by integrating the results from these two steps (Figure 1C). For each variant s_i that shows significant association to a local gene g_i, we test s_i for association with distal genes g_j that are nearby g_i in the partial correlation network defined by **G**. Since the edges of **G** account only 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_i to be tested for a distal effect on g_j:

1. s_i must be strongly associated with expression of the putative TAF, g_i.
2. Genes g_i and g_j must be separated in the partial correlation network by no more than two edges; *i.e.*, either $G_{i,j} \neq 0$ or there exists a third gene g_k such that $G_{i,k} \neq 0$ and $G_{k,j} \neq 0$. 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 trade-off 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 recording only 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) false discovery rate (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 comparably with the BY correction in the case of general dependency and in particular for two-sided t statistics (Romano *et al.* 2008).

Independent components analysis 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: $\mathbf{X} \sim \mathbf{SA}$. Each column of matrix **S** corresponds to an independent component or factor, and the ith element of a column is the "activation" level of the ith gene in that factor. These factors are meant to model some latent or underlying biological process. The kth row of matrix **A** reflects the amount of activation of the kth independent component across all individuals, and A_{ij} is activation on the jth 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 coexpressed 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 $\text{FDR} < 0.05$.

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 $\text{FDR} < 0.05$ were considered significant. For each association between a true local eQTL and a component, we defined the number of true positives to be the number of component-set genes that 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 *et al.* (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 be 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 $\Pr(s_i \rightarrow g_j \mid s_i \rightarrow g_i)$]. Finally, we test whether s_i and g_j are *independent*, given the expression of g_i: $\Pr(s_i \perp g_j \mid g_i \mid s_i \rightarrow g_i \text{ and } s_i \rightarrow g_j)$. 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 10 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 (×2), 50 (×2), and 10 (×10), leaving 100 genes that were independent of any module. Module topologies are depicted in Supporting Information, 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

$$\mathbf{g_{ds}} = \beta \mathbf{g_{us}} + \boldsymbol{\varepsilon},$$

where $\mathbf{g_{ds}}$ and $\mathbf{g_{us}}$ represent expression of the downstream and upstream genes, respectively, and $\varepsilon \sim N(0,1)$. Genes directly downstream of either the hub gene or a highly connected gene (defined as a gene with degree >20) were chosen to have effect sizes of 1, while all other effect sizes were assigned randomly as −1 or 1 with probabilities 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, and 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 nonhub driver genes may lead to spurious distal associations in the analysis. 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, 30% 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 probabilities 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 5*e*-05 and 1*e*-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 nonhub 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 are 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 (Buck Institute, Novato, CA). A total of 112 yeast segregants were mated from parent strains BY4716 and RM11-1a and grown in culture. Strains were genotyped at 2957 markers and expression measurements were assayed for 6216 ORFs. Genes with no available annotation information were removed, leaving a total of 5647 genes for analysis.

Mouse liver data

Gene expression data were previously assayed on the Affymetrix Mouse Gene 1.0 ST array and were obtained from GEO (accession no. 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. Probe sets that were expressed at a level >6 on a log2 normalized scale in at least 87.5% of mice were retained, leaving a total of 9377 probe sets for further analysis. Genotypes for 181,752 markers from the "A" test array for the Mouse Diversity Array were obtained from D. Aylor (North Carolina State University, Raleigh, NC).

Human lymphoblastoid cell line data

Gene expression data and HapMap phase 2 and 3 genotypes were obtained from 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 prequantile normalized data across all 69 individuals, leaving 9810 transcripts that were retained for analysis.

Results

Simulation analysis

To assess the sensitivity and specificity of NetLIFT for identifying distal eQTL, we applied the method to 10 simulated data sets consisting of paired expression and genotype data (see *Materials and Methods*).

For comparison, we also tested three previously described eQTL detection methods: 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 eQTL.

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 as long as the sensitivity remained high.

For detection of local eQTL effects, NetLIFT, Trigger, and AvA all identified true positives with 100% success (FDR $<$ 0.05, Table S2). The local eQTL false positive rate for NetLIFT was identical to that for 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.

Intramodule 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 nonhub genes were considered "true positives." For modules with eQTL acting on nonhub 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 the subsequent testing procedure prevent 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 ($P = 4.92e\text{-}96$), 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 coregulated gene modules under the influence of a common variant. Trigger performed better than the AvA approach, although in general identified $<$12% of true distal associations. NetLIFT was the only method to consistently identify distal effects in all network topologies.

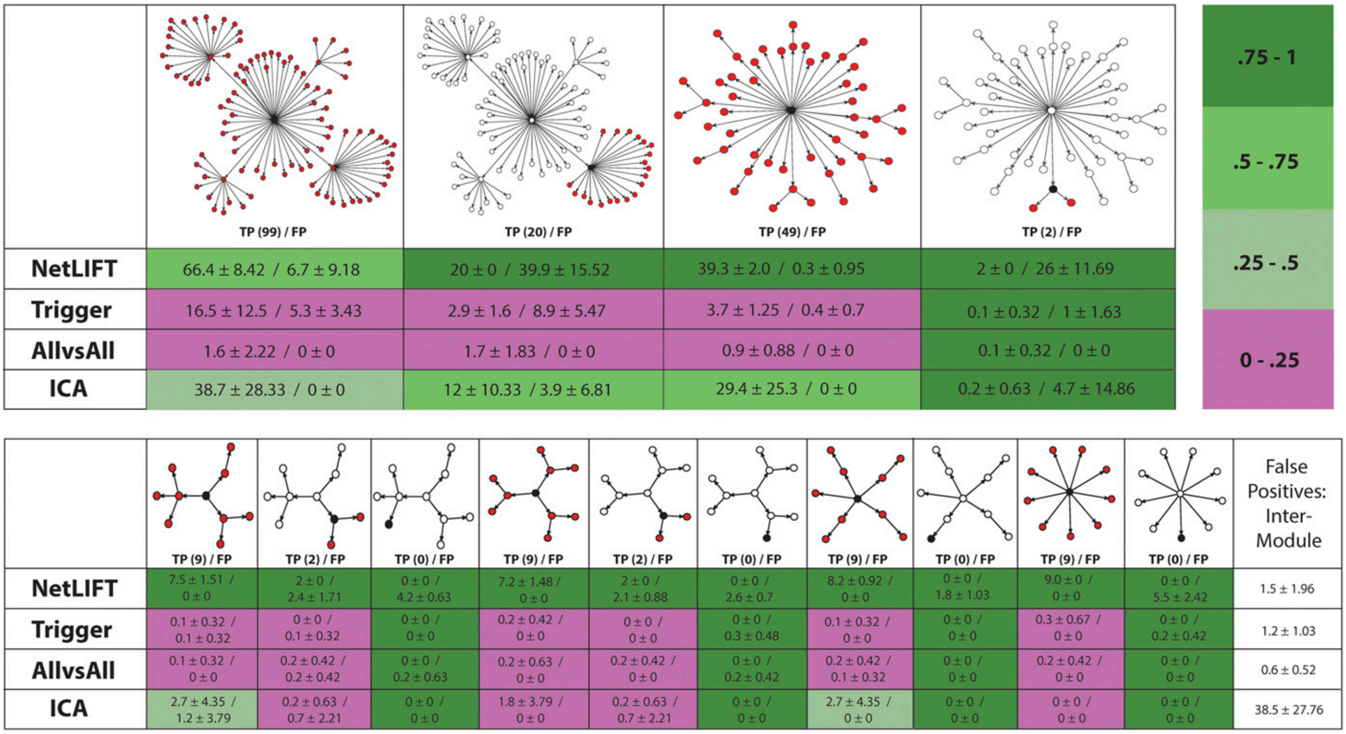
We next evaluated NetLIFT's performance in detecting "hotspot" eQTL loci, where a hotspot is defined as a locus that is associated with more transcripts than are expected by chance. To derive a family-wise error rate (FWER) for each locus, we used the procedure described in Breitling *et al.* (2008), 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 2000 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, 5647 genes and 2956 variants were retained for analysis. Variants within 10 kb 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 1124 (19.9%) and 1642 (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 Gene Ontology (GO) analysis using all 143 genes identified as intermediate 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.



	TP (99) / FP	TP (20) / FP	TP (49) / FP	TP (2) / FP
NetLIFT	66.4 ± 8.42 / 6.7 ± 9.18	20 ± 0 / 39.9 ± 15.52	39.3 ± 2.0 / 0.3 ± 0.95	2 ± 0 / 26 ± 11.69
Trigger	16.5 ± 12.5 / 5.3 ± 3.43	2.9 ± 1.6 / 8.9 ± 5.47	3.7 ± 1.25 / 0.4 ± 0.7	0.1 ± 0.32 / 1 ± 1.63
AllvsAll	1.6 ± 2.22 / 0 ± 0	1.7 ± 1.83 / 0 ± 0	0.9 ± 0.88 / 0 ± 0	0.1 ± 0.32 / 0 ± 0
ICA	38.7 ± 28.33 / 0 ± 0	12 ± 10.33 /3.9 ± 6.81	29.4 ± 25.3 / 0 ± 0	0.2 ± 0.63 / 4.7 ± 14.86

	TP (9) / FP	TP (2) / FP	TP (0) / FP	TP (9) / FP	TP (2) / FP	TP (0) / FP	TP (9) / FP	TP (0) / FP	TP (9) / FP	TP (0) / FP	False Positives: Inter-Module
NetLIFT	7.5 ± 1.51 / 0 ± 0	2 ± 0 / 2.4 ± 1.71	0 ± 0 / 4.2 ± 0.63	7.2 ± 1.48 / 0 ± 0	2 ± 0 / 2.1 ± 0.88	0 ± 0 / 2.6 ± 0.7	8.2 ± 0.92 / 0 ± 0	0 ± 0 / 1.8 ± 1.03	9.0 ± 0 / 0 ± 0	0 ± 0 / 5.5 ± 2.42	1.5 ± 1.96
Trigger	0.1 ± 0.32 / 0.1 ± 0.32	0 ± 0 / 0.1 ± 0.32	0 ± 0 / 0 ± 0	0.2 ± 0.42 / 0 ± 0	0 ± 0 / 0 ± 0	0 ± 0 / 0.3 ± 0.48	0.1 ± 0.32 / 0 ± 0	0 ± 0 / 0 ± 0	0.3 ± 0.67 / 0 ± 0	0 ± 0 / 0.2 ± 0.42	1.2 ± 1.03
AllvsAll	0.1 ± 0.32 / 0 ± 0	0.2 ± 0.42 / 0.2 ± 0.42	0 ± 0 / 0.2 ± 0.63	0.2 ± 0.63 / 0 ± 0	0.2 ± 0.42 / 0 ± 0	0 ± 0 / 0.2 ± 0.42	0.2 ± 0.42 / 0.1 ± 0.32	0 ± 0 / 0 ± 0	0.2 ± 0.42 / 0 ± 0	0 ± 0 / 0 ± 0	0.6 ± 0.52
ICA	2.7 ± 4.35 / 1.2 ± 3.79	0.2 ± 0.63 / 0.7 ± 2.21	0 ± 0 / 0 ± 0	1.8 ± 3.79 / 0 ± 0	0.2 ± 0.63 / 0.7 ± 2.21	0 ± 0 / 0 ± 0	2.7 ± 4.35 / 0 ± 0	0 ± 0 / 0 ± 0	0 ± 0 / 0 ± 0	0 ± 0 / 0 ± 0	38.5 ± 27.76

Figure 2 Number of detected distal associations, by module topology and method. Topology of each network module is depicted at the top of each section. Black nodes depict genes with an assigned local eQTL effect, and red nodes represent "true" distally associated genes. The total number of true distal associations is given in parentheses. Each cell value reports the mean and standard deviation of true positives and false positives, over the 10 simulated data sets. Cells are colored according to fraction of true positives discovered. The rightmost column (bottom row) reports the number of false positive distal associations where the locally regulated gene and the target gene belonged to disjoint modules.

For most previously identified hotspots, NetLIFT correctly identified biologically validated regulators (Table 2). Several predicted novel regulators with >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 noncoding RNA (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 ($P < 1.0e\text{-}05$) and were associated with highly overlapping sets of distally associated genes (Figure S5). Notably, we failed to identify *MAK5*, as this putative regulator was shown to contain a loss-of-function mutation that 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 eQTL 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 ($P = 1.03e\text{-}03$). Additionally, for 5 of these 16 genes (31.2%), the YEASTRACT database

Table 1 GO annotation enrichment for candidate regulators in yeast

P-value	Term
2.00*E*-06	Asparagine catabolic process
5.89*E*-06	Cellular response to nitrogen starvation
5.89*E*-06	Cellular response to nitrogen levels
4.66*E*-05	Asparagine metabolic process
4.90*E*-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

GO analysis was performed for genes with ≥10 distal associations; the top 20 enrichment terms are reported in the right column.

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-*vs.*-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" ($P = 1.27e\text{-}07$). *LYS2* catalyzes the reduction of α-aminoadipate to α-aminoadipate semialdehyde (αAASA), the fifth step in the lysine biosynthesis pathway. Downstream of this reaction, glutamate-forming saccharopine 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 upregulation 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 *et al.* 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). Within the set of transcriptional targets are four additional genes whose promoters contain the *Lys14p* binding motif, TCCRNYGGA, one of which, *LYS12*, is involved in lysine biosynthesis and has a directional expression pattern matching the other *Lys14p* targets (Figure S6).

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 8 founder mice (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ), part of the Collaborative Cross (CC) project (Churchill *et al.* 2004; Collaborative Cross Consortium 2012) (Figure 3). Founder strains of the CC were chosen to provide a high level of genetic diversity and represent three subspecies of origin: *Mus mus domesticus*, *M. m. castaneus*, and *M. m. musculus*. Wild-derived WSB/EiJ and classical inbred strains A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, and NZO/HlLtJ have a genetic background composed mostly of the *M. m. domesticus* subspecies, while the wild-derived CAST/EiJ and PWK/PhJ founder strains are primarily representative of the *M. m. castaneus* and *M. m. musculus* subspecies, respectively (Churchill *et al.* 2004; Collaborative Cross Consortium 2012).

We filtered for probe sets expressed above background levels and retained 9377 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 6182 eQTL were discovered

Table 2 Distal regulatory loci and candidate regulaotrs identified in yeast

Method	eQTL position	TAF	Previously predicted regulators	No. targets	GO annotation enrichment	GO *P*- value	FDR– growth rate association
[a]	ChrII:376668	TAT1	TRM7 (Gat-Viks *et al.* 2010)	265	Cytoplasmic translation	9.63*E*-37	NA
[a]	ChrII: 555596	AMN1	AMN1 (Yvert *et al.* 2003; Gat-Viks *et al.* 2010), MAK5 (Yvert *et al.* 2003)	307	Ribosome biogenesis	2.90*E*-12	0.0036
[a]	ChrII: 697894	GPX2	None (Yvert *et al.* 2003; Gat-Viks *et al.* 2010)	205	ncRNA processing	1.53*E*-17	0.012
[a]	ChrIII: 92127	LEU2	LEU2 (Yvert *et al.* 2003; Zhu *et al.* 2008, 2012; Gat-Viks *et al.* 2010)	113	Organic acid biosynthetic process	4.05*E*-25	NA
[a]	ChrIII: 105042	ILV6	ILV6 (Zhu *et al.* 2008, 2012)	93	Organic acid biosynthetic process	2.45*E*-22	NA
[a]	ChrIII: 201116	MATALPHA1	MATALPHA1 (Yvert *et al.* 2003; Smith and Kruglyak 2008; Zhu *et al.* 2008; Gat-Viks *et al.* 2010)	40	Response to pheromone	1.78*E*-08	NA
[a]	ChrV: 117056	URA3	URA3 (Yvert *et al.* 2003; Zhu *et al.* 2008, 2012; Gat-Viks *et al.* 2010)	28	*De novo* UMP biosynthetic process	8.66*E*-09	NA
[a]	ChrVIII: 111682	GPA1	GPA1 (Yvert *et al.* 2003; Smith and Kruglyak 2008; Zhu *et al.* 2008; Litvin *et al.* 2009; Gat-Viks *et al.* 2010)	29	Conjugation	1.*14E*-15	NA
[a]	ChrXII: 659357	HAP1	HAP1 (Yvert *et al.* 2003; Smith and Kruglyak 2008; Zhu *et al.* 2008; Litvin *et al.* 2009; Gat-Viks *et al.* 2010)	29	Steroid metabolic process	3.80*E*-09	NA
[a]	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	1.81*E*-05	NA
[a]	ChrXIV: 371953	NAM9	MKT1 (Zhu *et al.* 2008), SAL1 (Zhu *et al.* 2008)	25	Mitochondrial translation	1.55*E*-21	NA
[a]	ChrXV: 174364	PHM7	PHM7 (Zhu *et al.* 2008, 2012), IRA2 (Smith and Kruglyak 2008; Litvin *et al.* 2009)	107	Cellular ketone metabolic process	8.89*E*-08	NA
[a]	ChrXV: 382531	CRS5	CAT5 (Yvert *et al.* 2003; Gat-Viks *et al.* 2010)	11	Cellular respiration	3.77*E*-05	NA
[b]	ChrI: 11638	SEO1	NA	17	Monocarboxylic acid metabolic process	1.11*E*-06	NA
[b]	ChrII: 376872	NRG2	NA	32	Asparagine catabolic process	1.85*E*-06	NA
[b]	ChrII: 401568	TEC1	NA	16	Pseudohyphal growth	1.03*E*-03	NA
[b]	ChrII: 477206	LYS2	NA	167	Lysine biosynthetic process via aminoadipic acid	1.27*E*-07	NA
[b]	ChrIV: 96259	HEM3	NA	21	Cytokinesis	5.47*E*-04	NA
[b]	ChrIV: 1149761	FCF1	NA	18	Endonucleolytic cleavage involved in rRNA processing	4.02*E*-04	NA
[b]	ChrV: 420595	LCP5	NA	102	ncRNA metabolic process	1.90*E*-13	NA
[b]	ChrV: 504714	YER160C	NA	19	DNA integration	6.65*E*-24	NA
[b]	ChrVII: 402871	PRM8	NA	23	Cellular zinc ion homeostasis	5.72*E*-06	NA
[b]	ChrVII:916675	ZPR1	NA	27	Ribosome biogenesis	2.56*E*-05	NA
[b]	ChrIX: 33795	YIL166C	NA	30	Oligopeptide transport	2.22*E*-03	NA
[b]	ChrIX: 141014	RPI1	NA	21	L-asparagine biosynthetic process	1.34*E*-05	NA
[b]	ChrX: 24739	REE1	NA	18	Formate metabolic process	3.32*E*-08	NA
[b]	ChrX: 262593	SIP4	NA	17	Mitochondrial outer membrane translocase complex assembly	2.03*E*-04	NA
[b]	ChrXII: 126934	PUF3	NA	22	Transposition, RNA mediated	1.01*E*-06	NA
[b]	ChrXII: 468981	ASP3-1	NA	50	Oxidation-reduction process	7.84*E*-07	NA

(continued)

Table 2, *continued*

Method	eQTL position	TAF	Previously predicted regulators	No. targets	GO annotation enrichment	GO *P*- value	FDR– growth rate association
[b]	ChrXII: 956366	PUN1	NA	64	β-alanine metabolic process	1.29*E*-04	NA
[b]	ChrXIII: 28694	PHO84	NA	32	Negative regulation of catalytic activity	5.17*E*-05	NA
[b]	ChrXVI: 523450	SWI1	NA	40	Regulation of DNA metabolic process	2.63*E*-04	NA
[c]	ChrXIII: 149075	NA	SMA2 (Zhu *et al.* 2008)	NA	NA	NA	NA

The third and fourth columns list candidate regulators implicated by NetLIFT and previous methods, respectively. The fifth column gives the number of genes linked to the locus by NetLIFT. Top GO enrichment for linked transcripts is listed in the 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). Chr, chromosome.

[a] eQTL identified by previous methods and NetLIFT.

[b] eQTL identified by NetLIFT only.

[c] eQTL identified by previous methods only.

for 5733 genes at a 5% genome-wide threshold; 75% of eQTL were within 10 cM of the affected gene (Aylor *et al.* 2011).

For eQTL testing, we defined local effects as those where variants were within 1 Mb 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 5744 genes (61%) with a local eQTL and 3322 (35%) with at least one distal eQTL (FDR < 0.05). Of the genes with a distal eQTL, 1102 (12%) were linked to one SNP, 574 (6%) were linked to two SNPs, 400 (4%) were linked to three SNPs, and 1246 (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 (Ferris *et al.* 2013; Phillippi *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 misregulate 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 *M. m. musculus*-derived PWK allele explained at least half of the overall genetic effect on expression (Figure 4, top). 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 the target gene (Figure 4, bottom).

These transcriptomic differences may in turn affect phenotype. Body weight for wild-derived founder strains (CAST/EiJ, PWK/PhJ, and 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

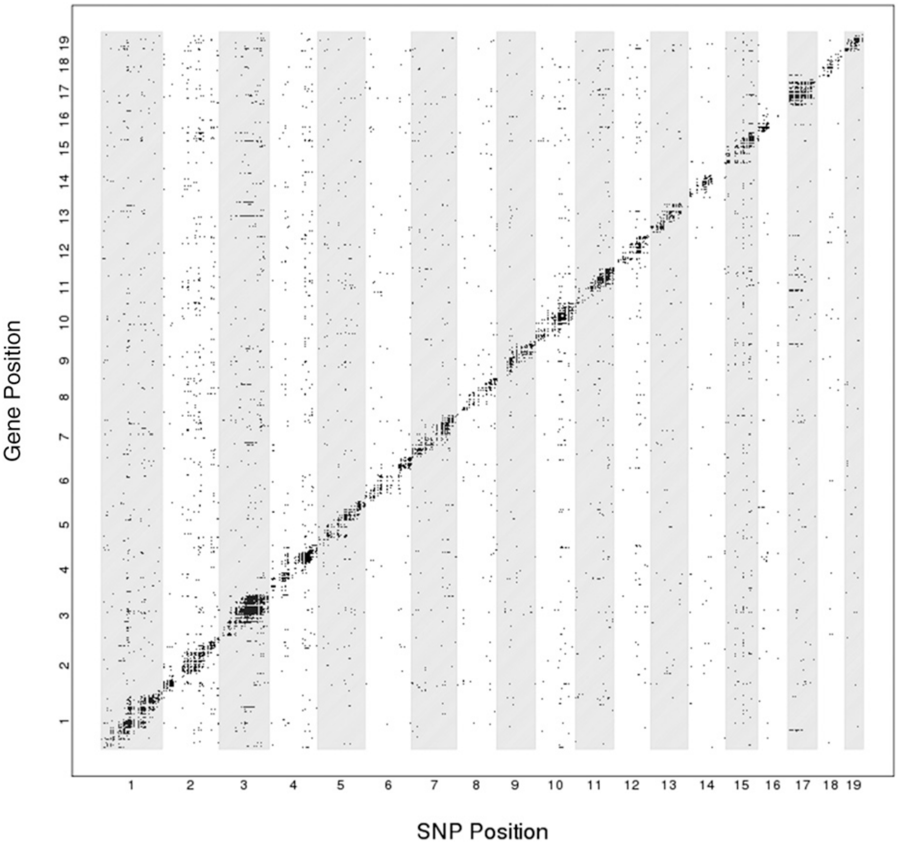


Figure 3 Distal eQTL associations in pre-Collaborative Cross mice. The *x*-axis gives the genomic coordinates of marker SNPs; the *y*-axis represents gene position. Each dot represents a significant marker–gene association at FDR $<$ 0.05, for markers that were at least 1 Mb from the associated gene.

interrogated for eQTL (Pickrell *et al.* 2010). For NetLIFT analysis, expression data were corrected for GC content and batch and were normalized as described previously. We selected 9810 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 phases 2 and 3, release 27.

Using a local regulatory window of 200 kb, similar to the original analysis (Pickrell *et al.* 2010), we identified 2483 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 1945 novel local associations, likely attributable to greater power resulting from testing only the most highly expressed quartile of transcripts.

NetLIFT identified 1274 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 five 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 *et al.* 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-

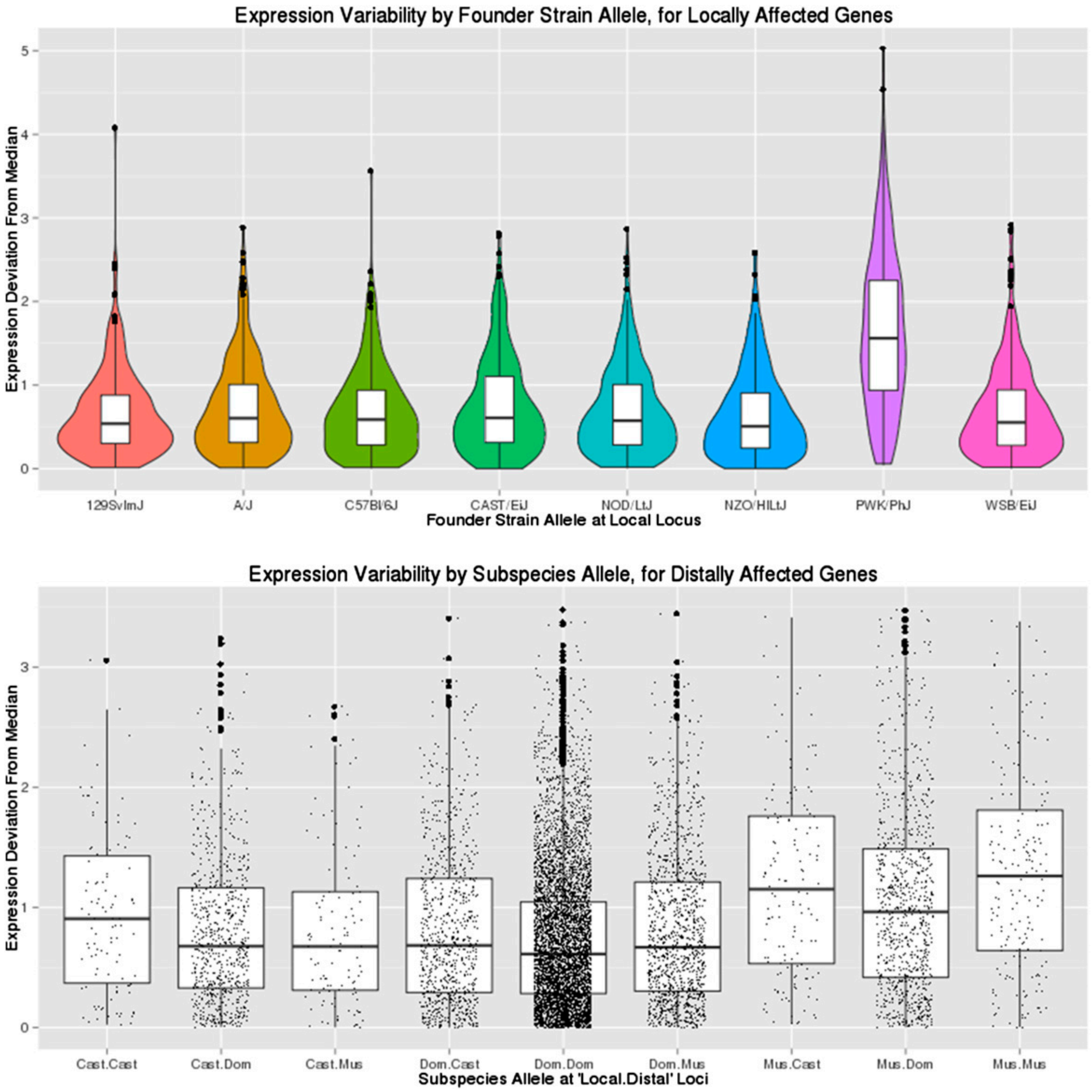


Figure 4 Expression variability for PWK-driven *trans*-acting factors and target genes, in pre-Collaborative Cross mice. (Top) Distribution of absolute expression deviation from median, for putative *trans*-acting factors with a PWK-driven local eQTL, grouped by founder strain genetic background at the eQTL locus. Only putative *trans*-acting factors that were linked to at least five 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 (*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.

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 increase nonlinearly as the number of genes increases and 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 affect only 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

Table 3 GO enrichments for distal genes linking to PWK-driver eQTL in pre-Collaborative Cross mice

P-value	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

GO analysis was performed for the pooled set of genes that linked to a PWK founder-driven eQTL with at least five distal effects; the top 20 GO enrichments are reported in the right column.

Table 4 GO term enrichment for putative trans-acting factors in human LBCs

P-value	Term
8.27*E*-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

GO analysis was performed for the set of putative *trans*-acting factors linked to three or more distal genes; enrichments at significance $P < 0.01$ are reported in the right column.

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. Finally, in a set of 69 human individuals, NetLIFT was able to find >1200 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, however, 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 plays 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 mice).

Our results also highlight an often unaddressed topic in complex trait mapping, namely, that eQTL discovery and interpretability of mapping results are 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 *vs.* 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.

Acknowledgments

The authors thank Rachel Brem for providing genotype and eQTL data for yeast strains and Steve Haase, David Aylor, and Fernando Pardo-Manuel de Villena for their contributions in interpreting eQTL results from yeast and pre-Collaborative Cross mice. The authors declare that no competing interests exist. This study was supported by National Institutes of Health (NIH) training grant 5T32GM067553-09, by the University Cancer Research Fund from the University of North Carolina at Chapel Hill, and by NIH grants U01-CA157703 and R01-CA166447.

Literature Cited

Airoldi, E. M., C. Huttenhower, D. Gresham, C. Lu, A. A. Caudy *et al.*, 2009 Predicting cellular growth from gene expression signatures. PLoS Comput. Biol. 5(1): e1000257.

Alberts, R., H. Chen, and C. Pommerenke, A. B. Smit, S. Spijker *et al.*, 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 autoimmune disease. BMC Genomics 12(1): 610.

Allen, J. D., Y. Xie, M. Chen, L. Girard, and G. Xiao, 2012 Comparing statistical methods for constructing large scale gene networks. PloS ONE 7(1): e29348.

Aylor, D. L., W. Valdar, and W. Foulds-Mathes, R. J. Buus, R. A. Verdugo *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 21(8): 1213–1222.

Barabási, A.-L, and Z. N. Oltvai, 2004 Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5(2): 101–113.

Becker, B., A. Feller, M. El Alami, E. Dubois, and A. 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. Mol. Microbiol. 29(1): 151–163.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57(1): 289–300.

Benjamini, Y., and D. Yekutieli, 2001 The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29(4): 1165–1188.

Bottolo, L., E. Petretto, S. Blankenberg, F. Cambien, S. A. Stuart *et al.*, 2011 Bayesian detection of expression quantitative trait loci hot spots. Genetics 189: 1449–1459.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, C. Wu *et al.*, 2008 Genetical genomics: spotlight on QTL hotspots. PLoS Genet. 4(10): e1000232.

Brem, R. B., and L. Kruglyak, 2005 The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci. USA 102(5): 1572–1577.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568): 752–755.

Chen, L. S., and F. Emmert-Streib, and J. D. Storey, 2007 Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol. 8(10): R219.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K.-Y. Jen *et al.*, 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. Nat. Genet. 33(3): 422–425.

Churchill, G. A., D. C. Airey, H. Allayee, J. M. Angel, A. D. Attie *et al.*, 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36(11): 1133–1137.

Collaborative Cross Consortium, 2012 The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190: 389–401.

Doss, S., E. E. Schadt, T. A. Drake, and A. J. Lusis., 2005 Cis-acting expression quantitative trait loci in mice. Genome Res. 15(5): 681–691.

Duarte, C. W., and Z.-B. Zeng, 2011 High-confidence discovery of genetic network regulators in expression quantitative trait loci data. Genetics 187: 955–964.

Feller, A., F. Ramos, A. Pierard, and E. Dubois, 1999 In Saccharomyces cerevisae, feedback inhibition of homocitrate synthase isoenzymes by lysine modulates the activation of LYS gene expression by Lys14p. Eur. J. Biochem. 261(1): 163–170.

Ferris, M. T., D. L. Aylor, D. Bottomly, A. C. Whitmore, L. D. Aicher *et al.*, 2013 Modeling host genetic regulation of influenza pathogenesis in the Collaborative Cross. PLoS Pathog. 9(2): e1003196.

Gat-Viks, I., R. Meller, M. Kupiec, and R. Shamir, 2010 Understanding gene sequence variation in the context of transcription regulation in yeast. PLoS Genet. 6(1): e1000800.

Göring, H. H., J. E. Curran, M P Johnson, T. D. Dyer, J. Charlesworth *et al.*, 2007 Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat. Genet. 39(10): 1208–1216.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106(23): 9362–9367.

Holloway, B., S. Luck, and M. Beatty, J.-A. Rafalski, and B. Li, 2011 Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics 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): 651–654.

Kang, H. P., A. A. Morgan, R. Chen, E. E. Schadt, and A. J. Butte, 2012 Coanalysis of GWAS with eQTLs reveals disease-tissue associations. AMIA Jt. Summits Transl. Sci. Proc. 2012: 35–41.

Kompass, K. S., and J. S. Witte, 2011 Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer. BMC Med. Genomics 4: 6.

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. A. C. 't Hoen, J. Monlong *et al.*, 2013 Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468): 506–511.

Litvin, O., H. C. Causton, B.-J. Chen, and D. Pe'er, 2009 Modularity and interactions in the genetics of gene expression. Proc. Natl. Acad. Sci. USA 106(16): 6441–6446.

Lorenz, W. W., R. Alba, Y.-S. Yu, J. M. Bordeaux, M. Simões *et al.*, 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.

Ma, L., A. G. Clark, and A. Keinan, 2013 Gene-based testing of interactions in association studies of quantitative traits. PLoS Genet. 9(2): e1003321.

Mehta, D., K. Heim, C. Herder, M. Carstensen, G. Eckstein *et al.*, 2012 Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. Eur. J. Hum. Genet. 20: 995–998.

Neto, E. C., A. T. Broman, M. P. Keller, A. D. Attie, B. Zhang *et al.*, 2013 Modeling causality for pairs of phenotypes in system genetics. Genetics 193: 1003–1013.

Peng, J., P. Wang, N. Zhou, and J. Zhu, 2009 Partial correlation estimation by joint sparse regression models. J. Am. Stat. Assoc. 104(486): 735–746.

Philip, V. M., and G. Sokoloff, C. L. Ackert-Bicknell, M. Striz, L. Branstetter *et al.*, 2011 Genetic analysis in the Collaborative Cross breeding population. Genome Res. 21(8): 1223–1238.

Phillippi, J., Y. Xie, D. R. Miller, T. A. Bell, Z. Zhang *et al.*, 2013 Using the emerging Collaborative Cross to probe the immune system. Genes Immun. 15: 38–46.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt *et al.*, 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464 (7289): 768–772.

Ramos, F., E. Dubois, and A. Pierard, 1988 Control of enzyme synthesis in the lysine biosynthetic pathway of Saccharomyces cerevisiae. Evidence for a regulatory role of gene LYS14. Eur. J. Biochem. 171(1–2): 171–176.

Romano, J. P., A. M. Shaikh, and M. Wolf, 2008 Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 17(3): 417–442.

Rotival, M., T. Zeller, P. S. Wild, S. Maouche, S. Szymczak *et al.*, 2011 Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. PLoS Genet. 7(12): e1002367.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.*, 2003 Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929): 297–302.

Shabalin, A. A., 2012 Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28(10): 1353–1358.

Smith, E. N., and L. Kruglyak, 2008 Gene-environment interaction in yeast gene expression. PLoS Biol. 6(4): e83.

Teixeira, M. C., P. T. Monteiro, J. F. Guerreiro, J. P. Gonçalves, N. P. Mira *et al.*, 2014 The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 42(Database issue): D161–D166.

West, M. A. L., K. Kim, D. J. Kliebenstein, H. van Leeuwen, R. W. Michelmore *et al.*, 2007 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 175: 1441–1450.

Yook, S.-H., Z. N. Oltvai, and A.-L. Barabási, 2004 Functional and topological characterization of protein interaction networks. Proteomics 4(4): 928–942.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss *et al.*, 2003 Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet. 35(1): 57–64.

Zhu, J., and B. Zhang, E. N. Smith, B. Drees, R. B. Brem *et al.*, 2008 Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat. Genet. 40(7): 854–861.

Zhu, J., P. Sova, Q. Xu, K. M. Dombek, E. Y. Xu *et al.*, 2012 Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. PLoS Biol. 10(4): e1001301.

Communicating editor: G. A. Churchill

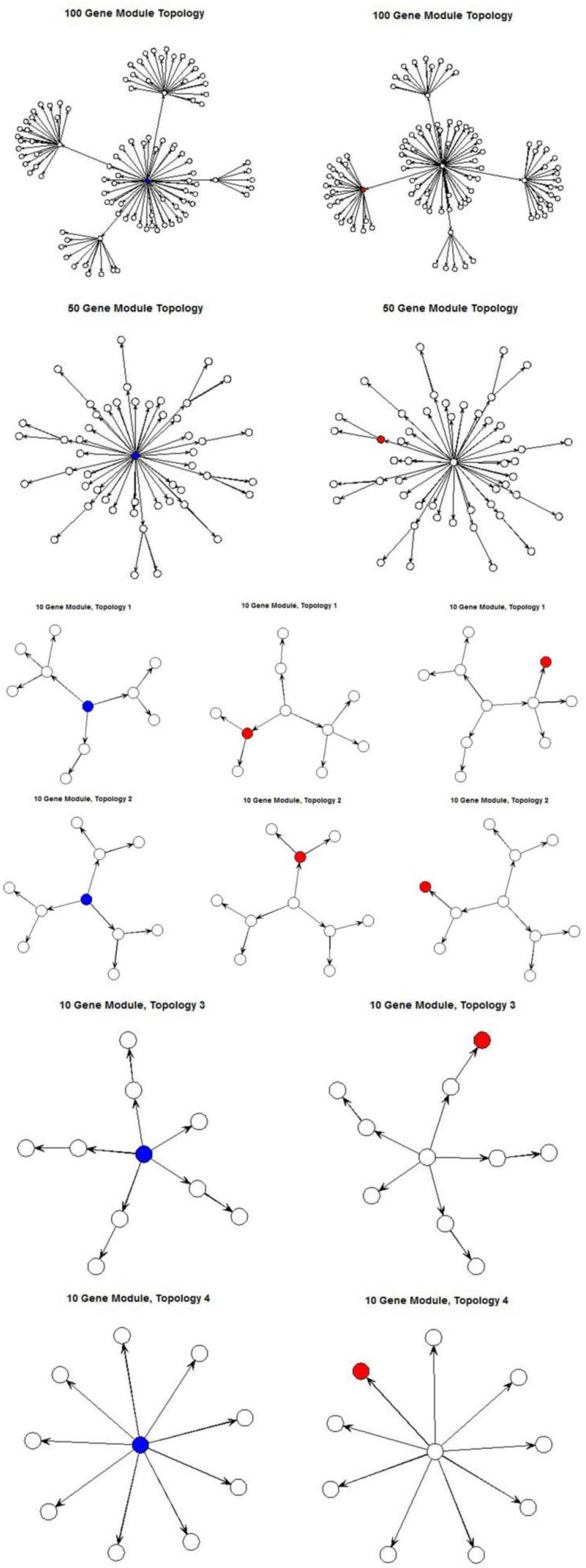
GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167791/-/DC1

Novel Distal eQTL Analysis Demonstrates Effect of Population Genetic Architecture on Detecting and Interpreting Associations

Matthew Weiser, Sayan Mukherjee, and Terrence S. Furey

DOI: 10.1534/genetics.114.167791



100 Gene Module Topology

100 Gene Module Topology

50 Gene Module Topology

50 Gene Module Topology

10 Gene Module, Topology 1

10 Gene Module, Topology 1

10 Gene Module, Topology 1

10 Gene Module, Topology 2

10 Gene Module, Topology 2

10 Gene Module, Topology 2

10 Gene Module, Topology 3

10 Gene Module, Topology 3

10 Gene Module, Topology 4

10 Gene Module, Topology 4

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.

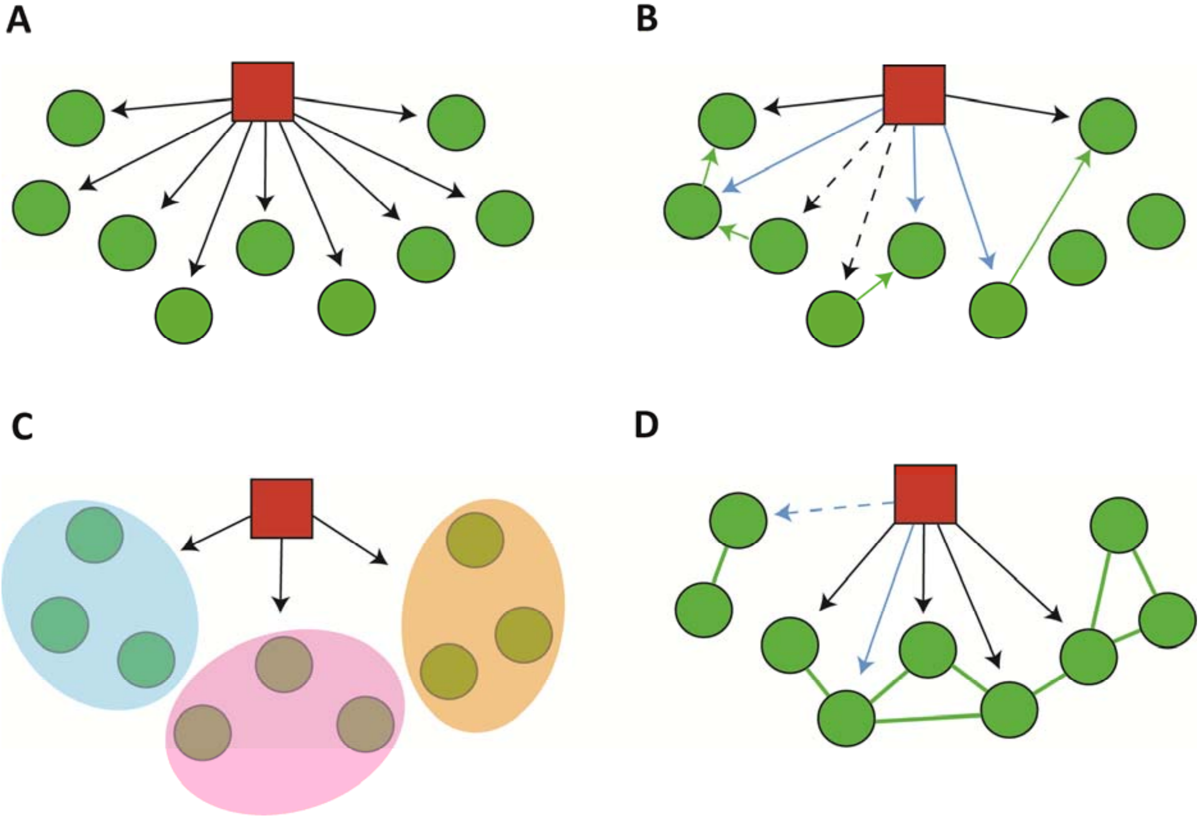


Figure S2 Illustration of eQTL detection methods. SNP is depicted as a red node; genes depicted in green. All vs All (A) performs a standard regression significance test for all pairs of SNPs and genes. Trigger (B) seeks to identify distal associations that are mediated by a locally associated variant-gene 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). Independent Components Analysis (C) first factors expression data into Independent Components, then performs association tests between allele frequency and the activation levels of components across samples. NetLIFT (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).

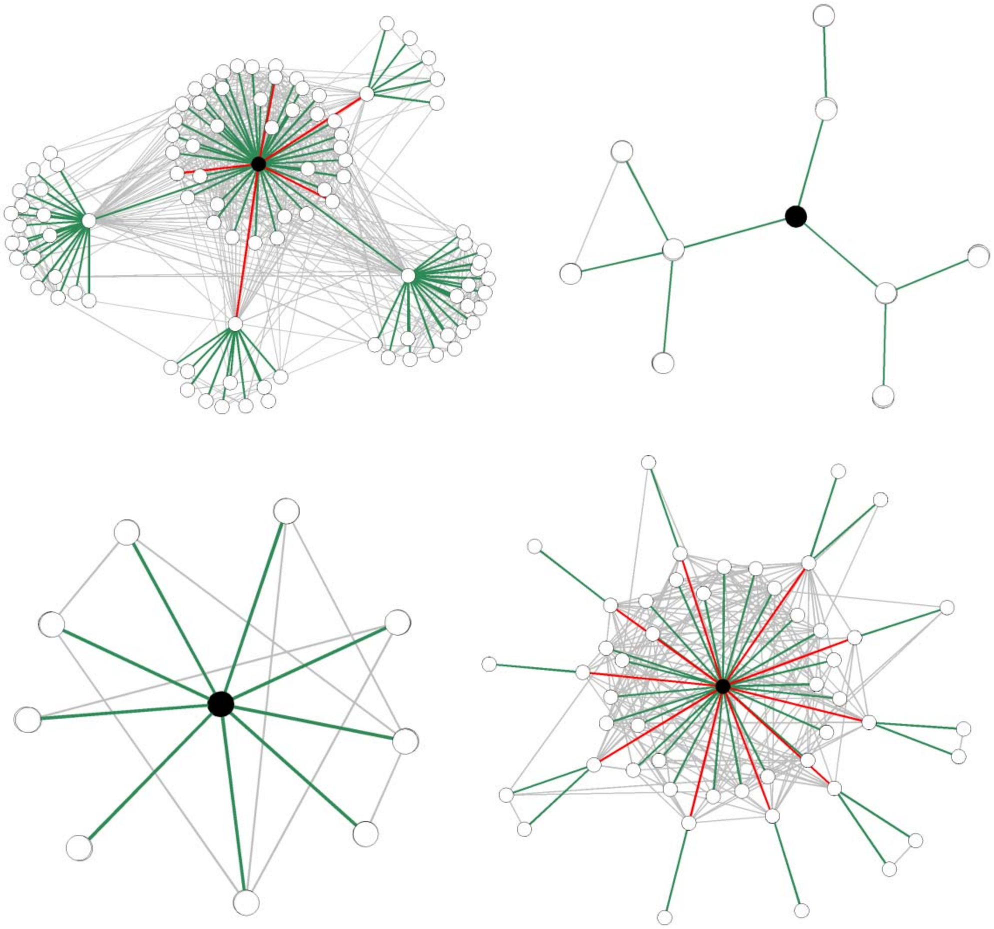


Figure S3 Partial correlation structure from network detection step, for representative 100 gene, 50 gene, and 10 gene modules. True positive correlations depicted with green edges, false negatives correlations in red, false positives in gray.

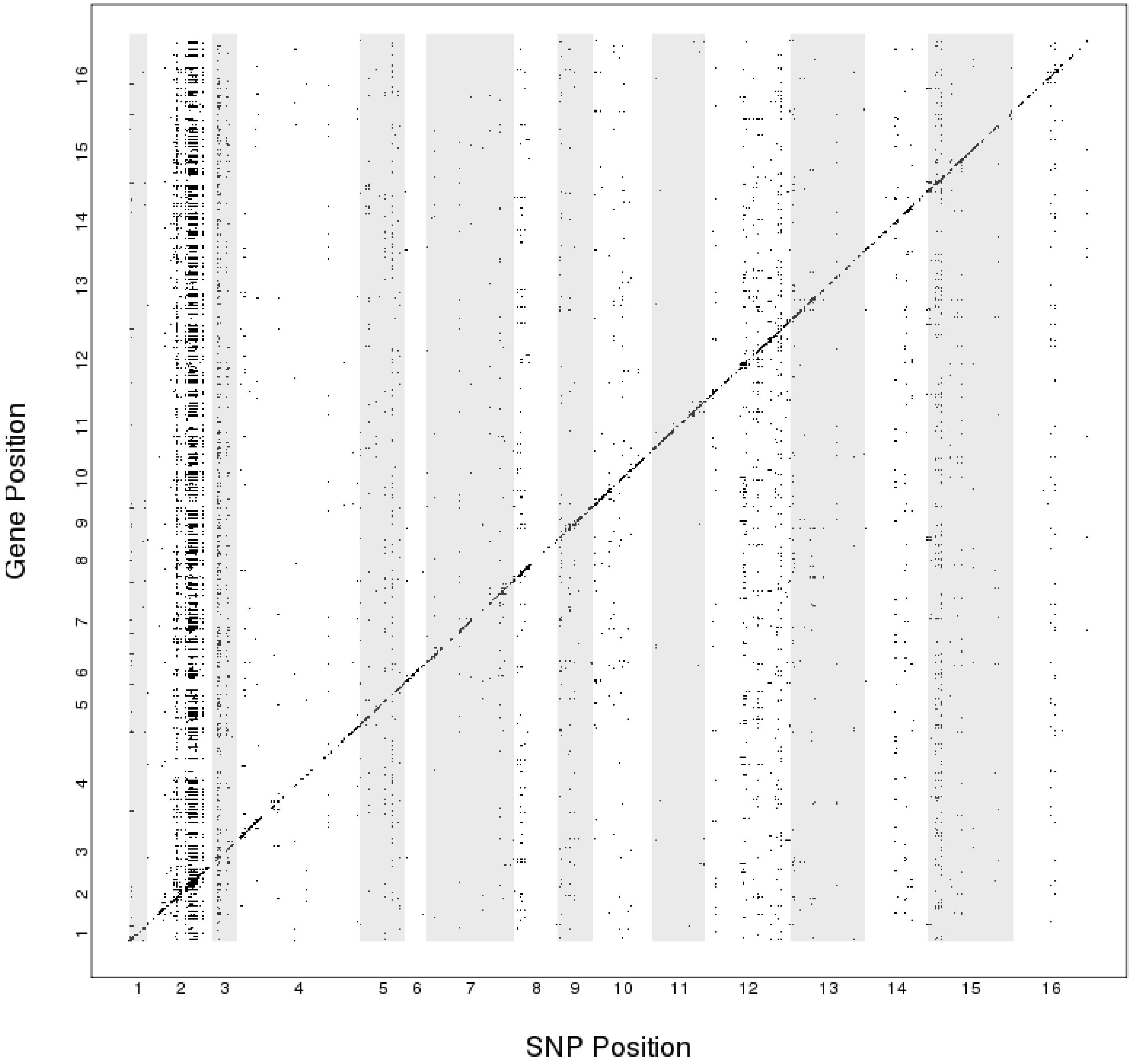


Figure S4 Local and distal eQTL linkages in yeast. X axis shows the genomic coordinates of marker variants; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.05.

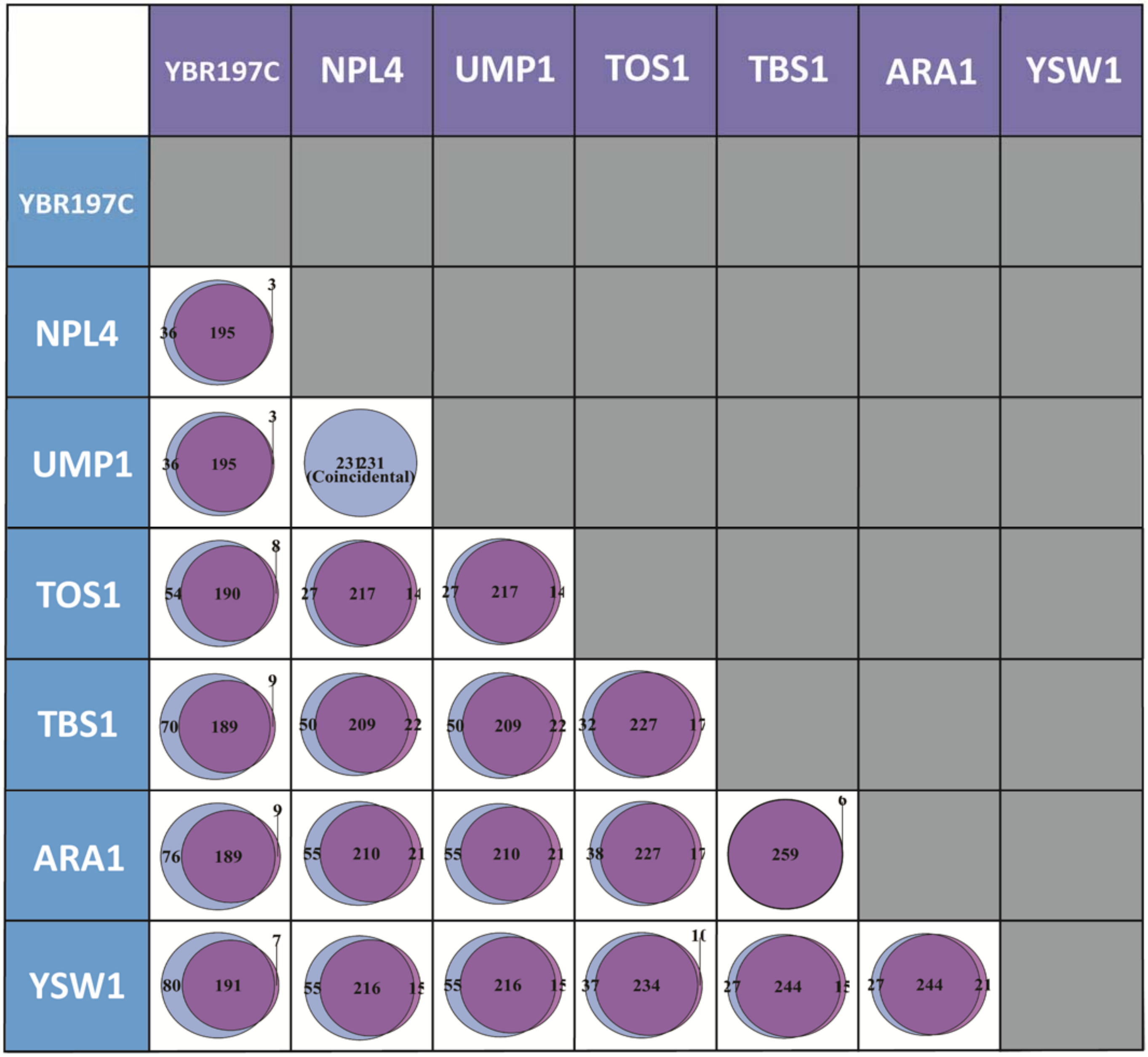


Figure S5 Pairwise overlap of target gene sets enriched for ribosomal annotation. Cell [i,j] shows the target gene overlap for between proposed regulators i, j.

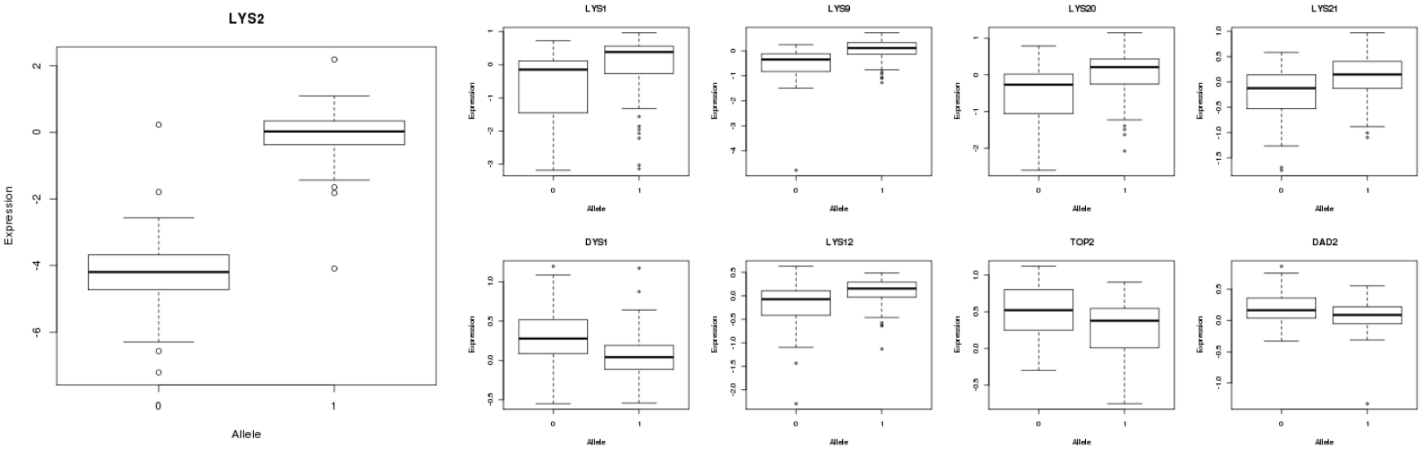


Figure S6 eQTL effects for LYS2 local regulatory variant and downstream genes. The allele associated with lower *LYS2* expression (“0”) is associated with lower expression of known *Lys14p* targets *LYS2*, *LYS1*, *LYS9*, *LYS20*, and *LYS21*. The same allele also associates with higher expression of three non-*LYS* genes containing *Lys14p* binding motifs (*DYS1*, *TOP2*, *DAD2*), and the *Lys14p* motif-containing *LYS12*.

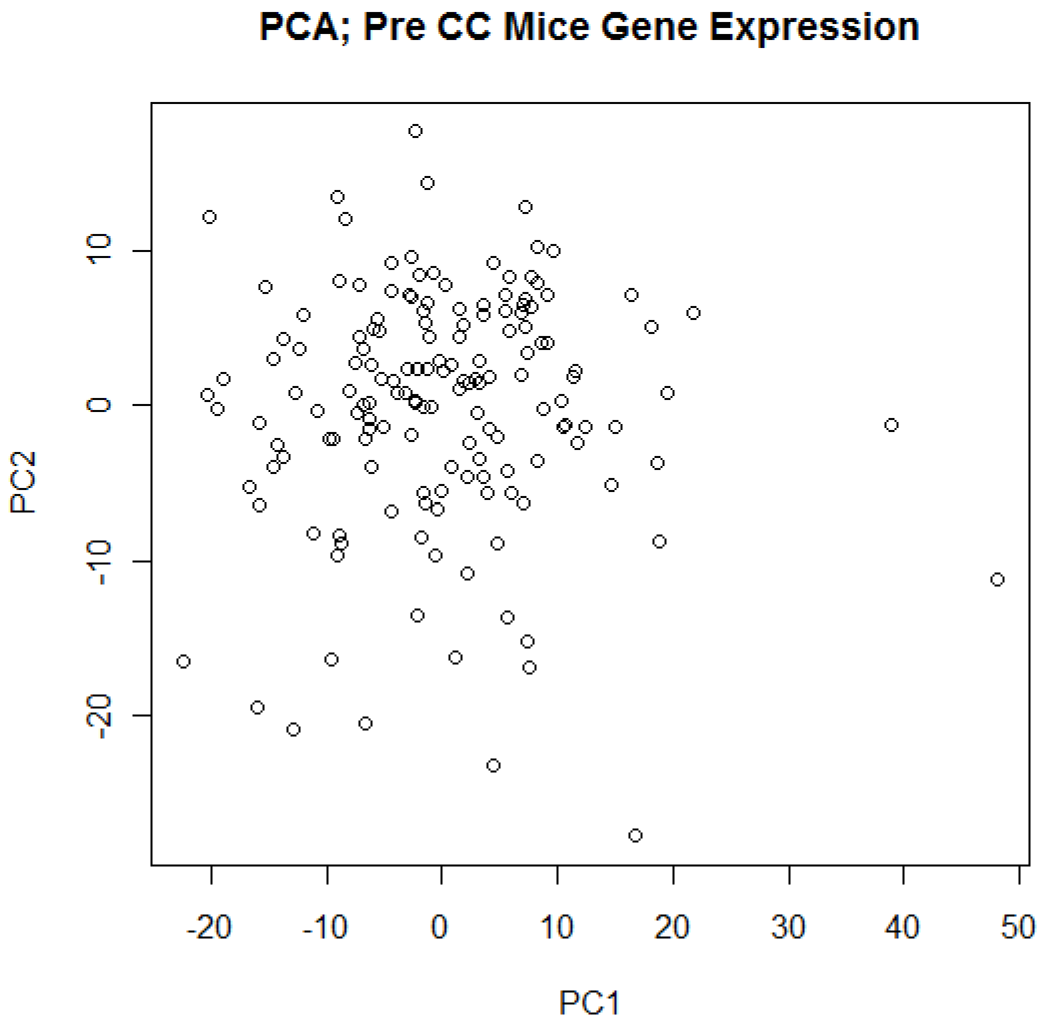


Figure S7 PCA analysis for pre-CC mice. Top two principal components for gene expression data in 156 pre-CC mice.

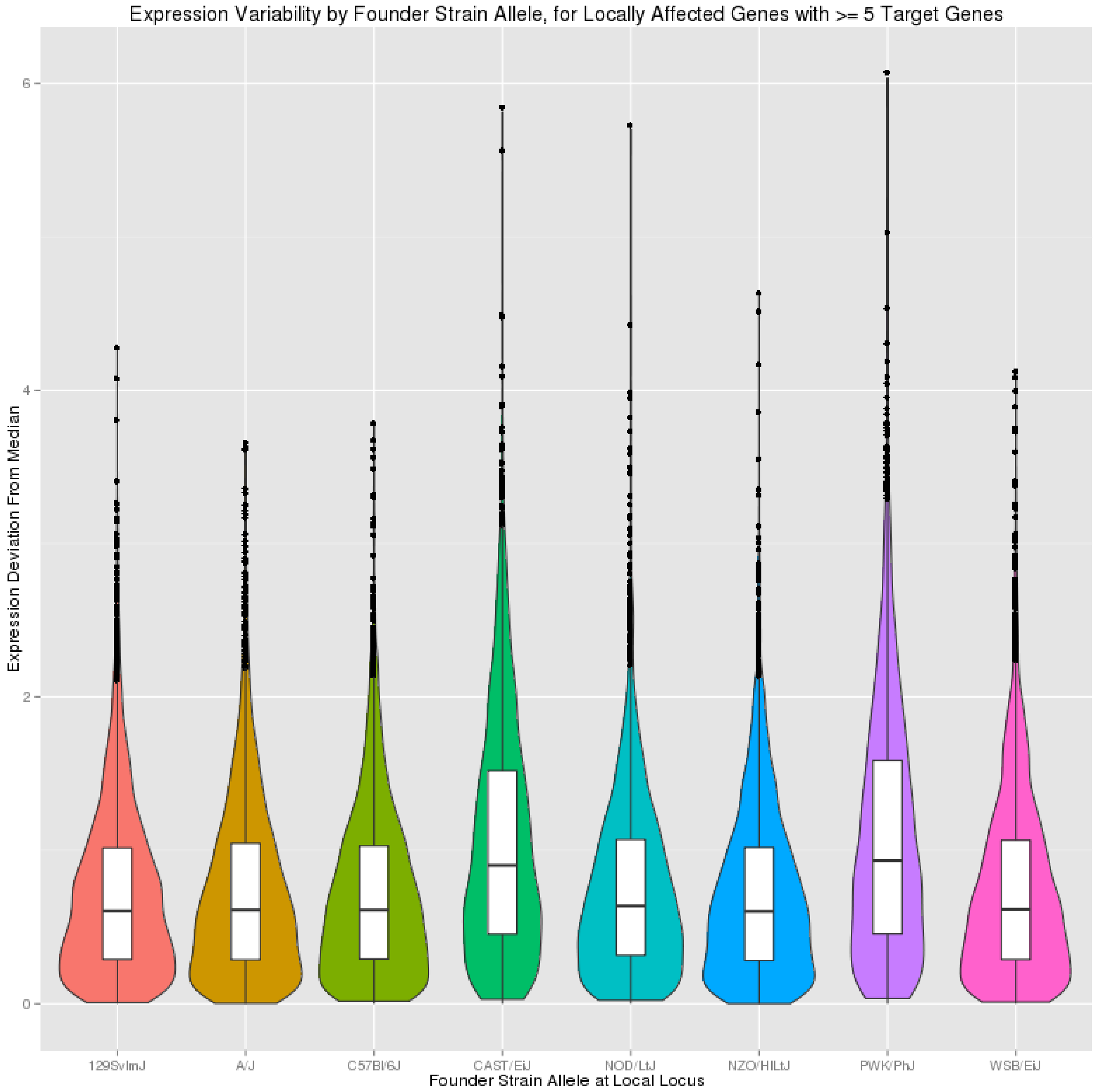


Figure S8 Expression variability by founder strain, for locally-regulated genes with at least 5 distal targets. Gene expression values were binned according to the genetic background of the locally-affected gene. Violin plot shows the level of variation compared to the overall sample expression medians, for each of the eight founder strains.

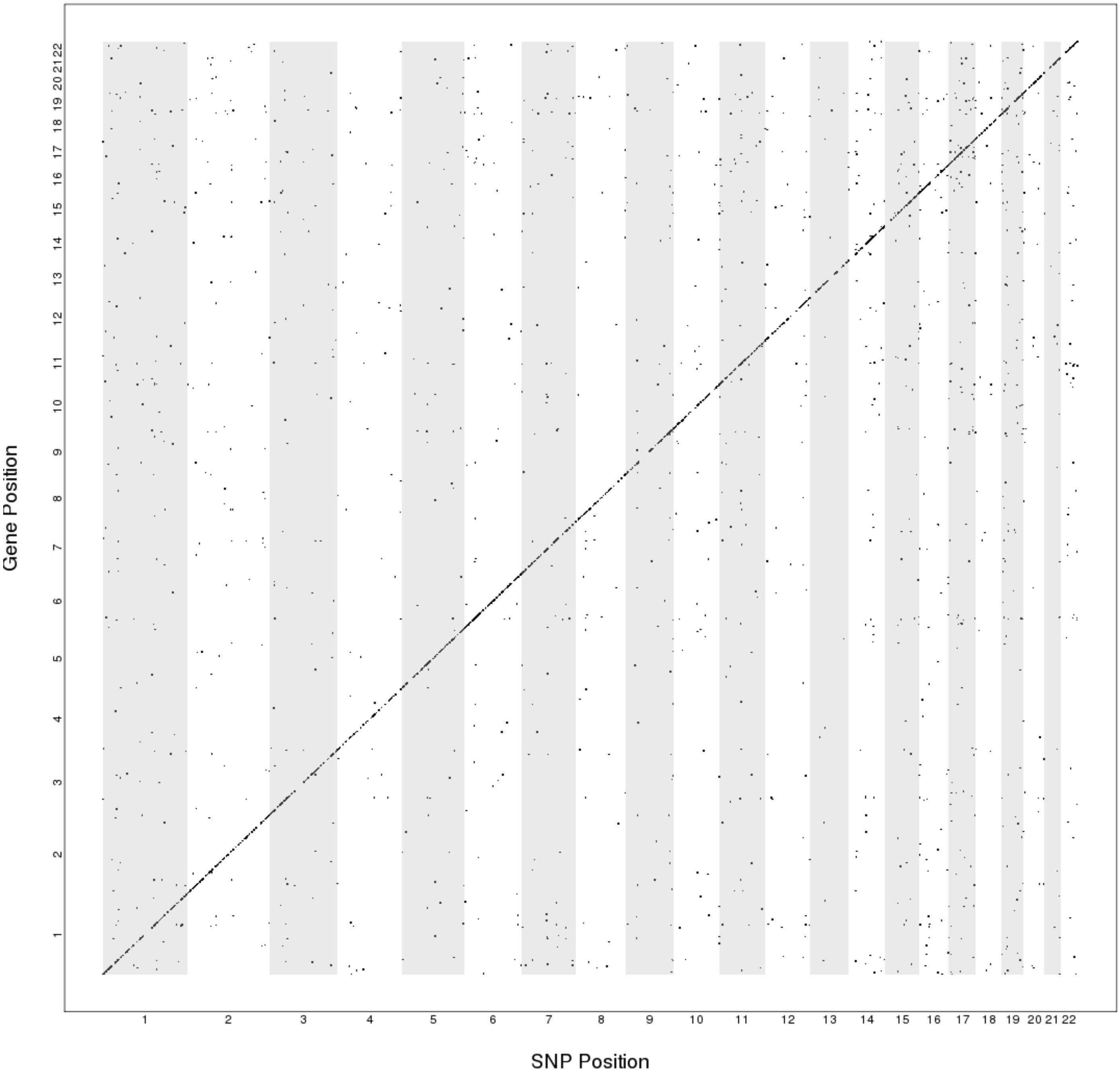


Figure S9 Local and distal eQTL linkages in human lymphoblastoid cell lines. X axis shows the genomic coordinates of SNPs; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.1.

Table S1 Sensitivity and specificity of partial correlation detection, by module. Column 1 shows the mean and standard deviation for the fraction of true edges detected, for ten simulated data sets. Column 2 estimates the intra-module false-edge detection rate. For each module, the ratio of false positive edges detected to the total number of possible false edges is reported.

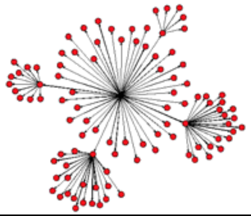
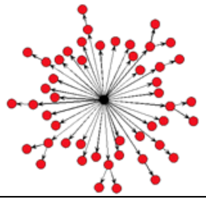
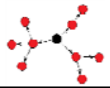
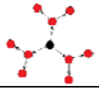
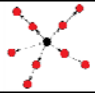
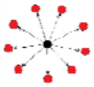
Module Topology	Mean(FracTP)± sd(FracTP)	Mean(FracFP)± sd(FracFP)
	0.94±0.018	0.092±0.0047
	0.79±0.015	0.16± 0.010
	1±0	0.13± 0.025
	1±0	0.14±0.022
	1±0	0.17±0.036
	1±0	0.22± 0.075

Table S2 Detected local eQTL effects, by method. FDR cutoff was set to 0.05. Counts are pooled for all 10 simulated data sets.

Method	True Positive	False Negative	False Positive
NetLIFT	442 (100%)	0	20
AllvsAll	442 (100%)	0	20
Trigger	442 (100%)	0	1653

Table S3 Hotspot detection rate for gene modules with eQTL at hub gene, in ten simulated data sets. A null distribution of maximum linkage counts were derived from the permuted data sets, with upper 95th quantile for each method listed in column 3. The mean number of identified associations for each module across all ten (non-permuted) data sets is listed in column 4.

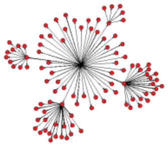
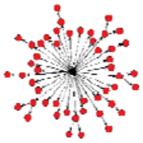
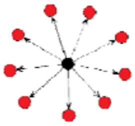
Method	Module	Num Associations Needed to Attain FWER 0.05	Mean Number of Associations Across Ten Simulations	Hotspot Detected
NetLIFT		3	66.4	10/10 (100%)
AllvsAll		1	1.6	4/10 (40%)
NetLIFT		3	39.3	10/10 (100%)
AllvsAll		1	0.9	6/10 (60%)
NetLIFT		3	9.0	10/10 (100%)
AllvsAll		1	0.2	2/10 (20%)

Table S4 Distribution of eQTL effects for local, distal eQTL, in 112 haploid yeast segregants using NetLIFT method (FDR < 0.05).

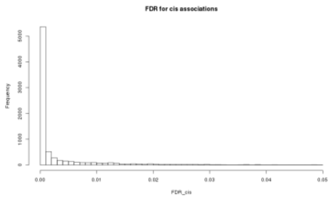
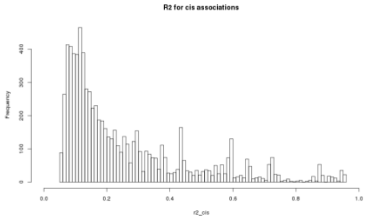
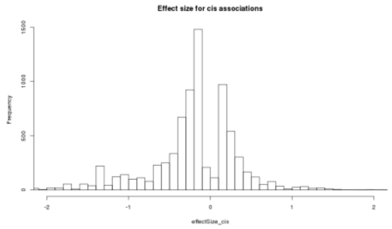
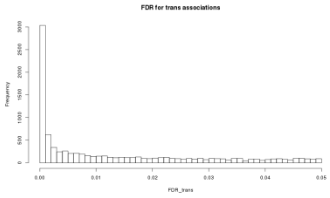
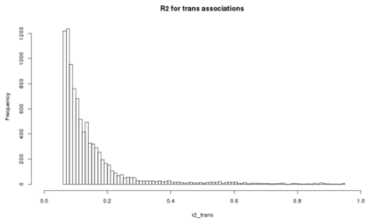
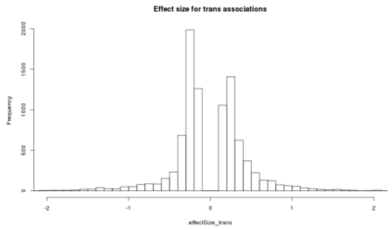
	Number (%)	FDR Distribution	R2 Distribution	Effect Size Distribution (β)
Local	1124 (19.9%)			
Distal	1642 (29.1%)			

Table S5 *trans* associations with growth associations

Available for download as an Excel file at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.114.167791/-/DC1