From genetics to gene networks:

Evaluating approaches for integrative analysis of genetic marker and gene expression data for the purpose of gene network inference

Bing Liu\textsuperscript{1,2,*}, Alberto de la Fuente\textsuperscript{2,3,*} and Ina Hoeschele\textsuperscript{1,2,§}

\textsuperscript{1}Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061

\textsuperscript{2}Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0477

\textsuperscript{3}CRS4 Bioinformatica, Parco Scientifico e Tecnologico, POLARIS, Edificio 1, Loc. Piscina Manna, 09010, Pula (CA), Italy

*These authors contributed equally to this work

§Corresponding author

Email addresses:

BL: bliu@vt.edu
ALF: alf@crs4.it
IH: inah@vbi.vt.edu
Abstract

Background
In a Genetical Genomics experiment a segregating population of hundreds of individuals is expression profiled for thousands of genes and genotyped for hundreds or thousands of genetic markers. Expression Quantitative Trait Locus (eQTL) mapping treats gene expression profiles as quantitative traits and identifies genomic regions causally affecting gene expression. With the identified eQTL regions, using DNA sequence information, genes physically located in an eQTL confidence region are identified as candidate causal regulators of the gene expression profiles. After using local structural models to identify regulator-target pairs for each eQTL, an Encompassing Directed Network (EDN) of causal relationships among genes can be constructed. The objective of the present work is to evaluate several eQTL mapping approaches and local structural models in their ability to construct an EDN.

Results
Several eQTL mapping approaches were evaluated: Single Profile Analysis (SPA), Principal Components (PC) mapping, and cis- and trans-eQTL mapping. Results on a Genetical Genomics dataset from yeast showed that both PC-mapping and cis- and trans-eQTL mapping greatly increased power for eQTL detection as compared to SPA. For regulator-target pair identification, our local structural models performed well on simulated data, except for one case where some genes had extremely high and other genes low heritability. For the yeast data set, an EDN was constructed based on the combined results from PCA and cis- and trans-mapping. It included 28,609 regulator-target pairs.
Conclusions

Our goal is gene network inference, and we present a method for the construction of an encompassing directed network based on a Genetical Genomics experiment. Given the EDN, the task of gene network reconstruction then reduces to the search for a set of optimal network structures nested within the EDN. EDN construction is based on eQTL mapping and identification of regulator-target pairs for each eQTL region. Compared to SPA, PC-mapping and cis- and trans-mapping strongly increase the power of identifying eQTL and regulator-target pairs. Hence using a combination of these methods is recommended. Epistatic interactions among eQTL remain to be incorporated. Identification of regulator-target pairs is based on local structural models, and this step would benefit from the incorporation of external information.

Background

Gene networks can be represented as graphical models in which nodes are genes, and undirected and directed edges correspond to correlations and causal influences between the nodes, respectively. Gene networks are coarse-grained descriptions of cellular physiology in the sense that only relationships between gene expression levels are modeled, and many other components such as proteins and metabolites are not explicitly taken into account. Nevertheless, gene networks are system-level descriptions of cellular physiology and will improve our understanding of the genetic architecture of complex traits and diseases. Gene networks have many practical applications [1], including the discovery of direct drug targets [2, 3]. It has been shown that some classical concepts in genetics, such as dominance and epistasis, can be understood in terms of networks and their properties [4, 5]. Many strategies have been proposed to obtain gene networks from gene expression data, including
probabilistic models, (e.g., [6, 7]), time series analysis using linear models, (e.g., [8, 9]), partial correlation analysis, (e.g., [10-13]) and several perturbation approaches (e.g., [3, 14-16]). Different approaches to gene network inference are reviewed by [1, 17-19].

Causal inference for genes can be achieved with a strategy of creating targeted perturbations (interventions) and measuring the responses of gene expression levels to those perturbations. It has been shown that such an approach can provide a reliable identification of gene networks [3, 15, 16]. There are two major types of targeted perturbation experiments: one uses one-at-a-time, specific perturbations (e.g., [20-22]), and the other uses naturally occurring multi-factorial perturbations in segregating populations (Genetical Genomics, [23-25]). We focus our network inference on Genetical Genomics data. In a Genetical Genomics experiment, a segregating population of hundreds of individuals is expression profiled for thousands of genes and genotyped for hundreds or thousands of genetic markers. The variation in the expression levels of genes is influenced by the variation in many polymorphisms (genotypes) across the genome. The genotypes can thus be regarded as natural, multifactorial perturbations [23-25] resulting in different gene expression “phenotypes”, and a causal relationship can therefore be established between the measured genotypes and gene expression phenotypes. In contrast to the approaches using specific experimental perturbations, in Genetical Genomics we do not know where the perturbations arise and we must identify their origin. This can be achieved by expression Quantitative Trait Locus (eQTL) mapping, which treats the gene expression profiles in a segregating population as quantitative traits and performs Quantitative Trait Locus (QTL) mapping on those traits. QTL mapping identifies the
polymorphic genomic regions having significant effects on a quantitative trait. Compared to the traditional QTL mapping, eQTL mapping is performed on a much larger scale: there are thousands to ten thousands of correlated expression traits (etraits). The result of eQTL analysis is the knowledge that certain genomic regions likely have causal effects on the expression levels of particular genes. Then, by using DNA sequence information, genes located in an eQTL region can be identified as candidate causal regulators of the genes whose expression levels are affected by that eQTL. After the identification of the candidate regulatory genes in each eQTL, a network of causal regulatory relationships among genes can be constructed, which we call an Encompassing Directed Network (EDN). The constructed EDN can be represented as a graph consisting of gene nodes and eQTL nodes. The directed edges in the EDN correspond to causal influences of the eQTL and candidate regulator genes on the target genes. A set of sparser networks well supported by the data can be found by searching within the network topology space defined by the EDN. The main purpose of the present work is to evaluate several eQTL mapping approaches and local structural models for the selection of regulator-target pairs in their ability to construct an EDN from the eQTL results. Sparsification of the EDN using a Structural Equation Modeling (SEM) approach will be addressed in our other contribution in this issue of BMC Genomics [26]. Very briefly, SEM represents a network both algebraically with a system of multiple equations and graphically. SEM is a generalization of linear regression, where each variable can be a response variable in one equation and a predictor variable in another equation. SEM is closely related to the frequently used framework of Bayesian networks [26], but it does not have the limitation of reconstructing only acyclic networks. In our case, the SEM consists of a system of structural equations, one equation for the etrait of each target gene
expressing it as a linear combination of the expression levels of all its regulator genes, its eQTL effects and noise (and sometimes certain interaction effects). For trans-mapping and regulator-target pair identification, we use single equations expressing the etrait of a target gene as a linear combination of the expression levels of some of its regulator genes, some of its eQTL effects and noise (and some interactions). We therefore refer to these single equations as local structural models or equations (evaluating regulator-target relationships of a single target in the presence of no, some or all of the regulators of the same target), while the SEM represents a global structural model (evaluating regulator-target relationships in the context of the entire network).

Expression QTL (eQTL) analysis is often performed on each etrait separately with adjustment for multiple testing based on the false discovery rate (FDR) [27] control, e.g. [28-30]. This approach will be referred to as Single Profile Analysis (SPA). SPA overlooks the fact that etraits are correlated and therefore does not have optimal power for the detection of pleiotropic eQTLs (i.e., eQTLs affecting multiple etraits). Multi-trait mapping is known to be more powerful than single trait mapping for detecting pleiotropic QTLs (e.g., [31]), but it is not computationally feasible for thousands of etraits. For small sets of correlated traits, QTL mapping of Principal Components (PC) has been shown to be equivalent to multi-trait mapping, without the drawback of increased computational complexity [32]. Several groups have used PCs for QTL or eQTL mapping in different experimental settings [33-39], but they analyzed only a small number of traits. Here, we perform Principal Component Analysis (PCA) on the entire set of 4,589 filtered yeast genes [29], as well as on subsets of genes obtained by gene clustering.
Expression QTL mapping and regulator gene identification can be performed much more effectively by taking into account two distinct types of genetic regulation: cis- and trans-regulation. In the case of cis-regulation, the cis-eQTL affects a particular etrait X (expression level of gene X) and is located at the physical location of gene X on a chromosome. The polymorphism of a cis-eQTL likely corresponds to a promoter region polymorphism of the gene (e.g. [40, 41], [24]). If gene X regulates the expression of some other genes, the eQTL that cis-affected gene X will have an indirect effect on the expression of those genes through gene X [41]. Such indirect effects have been referred to as cistrans effects [42]. Trans-eQTLs influence the expression levels of genes, but do not need to be co-located with any of these genes. The polymorphism of a trans-eQTL likely comes from a coding region polymorphism in a regulator gene located at the eQTL (e.g. [43], [24]). While a trans-eQTL does not affect the expression level of the regulatory gene, the coding region polymorphism affects the kinetic properties of the regulatory protein encoded by that gene, which in turn affects the expression levels of the targets.

Since by definition the location of a cis-eQTL must physically coincide with the location of the gene whose etrait is affected, only the marker(s) closest to the location of an etrait’s gene are tested to detect cis-eQTLs (e.g. [40, 41]). For network inference, such cis-linked etraits are not very informative. As shown on mouse data [41], the secondary targets of the cis-eQTLs, the so-called cistrans regulated etraits, can be obtained by testing the effects of the identified cis-eQTL regions on other etraits.
Trans-affected target etraits are affected by the eQTL genotype and the etrait of the candidate regulator gene simultaneously. Therefore, it was proposed [42] that in order to specifically detect trans-eQTLs, mapping is best performed by including candidate “regulatory” etraits in the QTL model. These authors performed interval mapping on any etrait $i$ with a model including the effects on etrait $i$ of another etrait $j$, the genotype at the physical location of gene $j$, and the etrait-by-genotype interaction. We performed mapping of trans-eQTLs also by including the candidate etrait in the model, but with a regression model and the Intersection-Union-Test (IUT) [44, 45] to test whether the eQTL genotype and the etrait of the candidate regulator gene both significantly affect the target etrait.

The problem of identifying candidate regulatory genes from eQTL confidence regions has been approached by using partial correlation tests [30], analysis of the between-strain Single Nucleotide Polymorphisms (SNPs) [46], assessing the extent of eQTL overlap between any two etraits [47], and more recently using a stochastic model incorporating protein-protein interaction data [48]. We used local structural models separately for each eQTL to identify the regulator-target pairs, taking into account that an eQTL may affect a target through cis, cistrans or trans regulation.

In contrast with previous work (e.g. [41, 42]), in this contribution we consider cis, cistrans and trans regulations jointly with the goal of reconstructing an EDN that defines the network search space for a network reconstruction method which we report on in a separate contribution elsewhere in this issue of BMC Genomics [26].
Results

Single Profile Analysis

With a 5% FDR p-value threshold of 0.000264, a total of 666 significant combined eQTL regions and 6,264 individual eQTL-target pairs were detected (combined eQTL regions resulted from combining eQTL regions affecting different etraits but having at least 80% overlap in their locations). The sizes of the eQTL regions were relatively wide (median 84 kb), which in some cases can be due to multiple linked QTL. The median size of the eQTL regions from three-marker regression decreased to 43 kb, the number of eQTL regions increased to 797, and the number of significant eQTL-target pairs increased to 6,729.

Principal Components Mapping

First, PCA was performed on all 4,589 etraits. Based on the scree plot which shows the fraction of total variance in the data explained by each PC, 20 PCs were selected for eQTL mapping. The plot of the sorted gene loadings of the first 10 PCs (Figure 1) shows that many genes contribute to each PC. Therefore, PC mapping based on PCA of all genes was only able to detect major eQTL affecting a relatively large number of genes. With PCA on all 4,589 filtered genes, 38 combined eQTL regions were detected (median CI 84 kb), including all major eQTL regions affecting relatively large numbers of genes identified earlier with analyses on single profiles or clustered profiles [43].

(Insert Figure 1 here)
When analyzing PCs computed from separate PCA of the gene subsets corresponding to 100 clusters, after three-marker regression, a total of 250 combined eQTL regions (median size 37 kb) were detected. The eQTL regions detected with PCA on all etraits were also detected here. Next, SPA was performed on these 250 eQTLs, with a 5% FDR adjusted p-value cutoff of 0.00012. A total of 10,316 eQTL-target pairs were detected.

**Cis-eQTL Mapping**

For cis-mapping, as expected, controlling FDR at the 5% level resulted in a considerably less stringent p-value threshold (0.0139), compared with the SPA threshold. After three-marker regression, a total of 578 combined cis-eQTL regions (median size 36 kb) were detected. We then searched for cistrans-affected etraits of these eQTLs. The 5% FDR-adjusted p-value cutoff at this stage was 0.000412. A total of 7,481 eQTL-target pairs were found.

**Trans-eQTL Mapping**

Trans-mapping appeared to greatly increase the power to detect eQTL. Using the IUT with 5% level FDR threshold, 41,309 significant candidate regulator-target pairs were identified. Figure 2 presents representative profiles of an etrait on two chromosomes. Red lines represent the SPA profile and its threshold, and blue lines represent the trans-mapping profile and its threshold. The trans-mapping profile is raised considerably above the SPA profile and this increase more than compensates for an increase in the threshold value.

*(Insert Figure 2 here)*
The interaction between eQTL and candidate regulator \((b_3\text{ in equation (1)})\) gene did not appear to be important. Out of all tests performed, only 0.08% had a significant eQTL by regulator gene interaction with FDR control at the 5% level for this term. Out of the tests with significant IUT, 4.94% had p-values smaller than 0.01, and 0.43% had p-values smaller than the FDR cutoff from all tests.

**Comparison in terms of eQTL detection power**

We compared the eQTL mapping methods by counting the overlap between any two eQTLs detected with two different methods and affecting the same trait. We considered any overlap, 50% or more overlap, and 99% or more. For comparing trans-mapping results with the other methods, an overlap or agreement was counted when the trans-eQTL marker was inside of an eQTL region detected by the other methods, or located immediately next to it with no other marker in between.

The percentages of eQTLs that had any overlap between two different methods are shown in Table 1. The percentages were not very different when a 50% or 99% overlap between eQTL regions was required (results hence not shown). Most eQTL-target pairs detected in SPA overlapped with eQTLs identified in cis-mapping and PC mapping. These effects were likely either cis or cistrans. Only 24% of the SPA eQTL regions contained or were next to markers identified by trans-mapping, an expected finding given that the SPA results included many cis- and cistrans linkages. Only 3% of the SPA eQTL-target pairs and 2% of the cis-mapping eQTL-target pairs were not detected by any of the other methods. For PC-mapping, 24% of its eQTLs did not overlap with eQTLs from the other methods, suggesting some advantage of this pleiotropic approach. Of all trans-eQTLs, 87% did not overlap with eQTLs from the
other methods, which indicated the high power of trans-eQTL mapping and supported the fact that the other methods mostly find cis and cistrans effects.

(*Insert Table 1 here*)

**Regulator-target pair identification**

A total of 6,723 eQTL-target pairs were detected using SPA. After searching for regulators in the eQTL confidence intervals, 6,276 regulator-target pairs involving 3,050 genes were found, with 1,192 regulators and 2,544 targets. From PC-mapping, for the 10,316 eQTL-target pairs, 9,843 regulator-target pairs were retained, involving 3,581 genes, with 1,143 regulators and 3,262 targets. A total of 7,481 eQTL-target pairs were found by cis-mapping, and 6,090 regulator-target pairs involving 3,034 genes were found, with 1,099 regulators and 2,562 targets. After local sparsification of the trans-mapping results, the 41,309 candidate regulator-target pairs reduced to 15,835 pairs involving 3,858 genes with 1,433 regulators and 3,682 targets.

The percentages of common regulator-target pairs between different eQTL mapping methods are shown in Table 2. Overlap between methods in terms of regulator-target pairs was approximately only half of the overlap in terms of eQTL-target pairs (Table 1). For SPA and cis-mapping, 41% of all regulator-target pairs were not in common with other methods, while for PC-mapping this number increased to 62%. For trans-mapping, most (95%) of the regulator-target pairs detected by trans-mapping were not identified by any of the other methods. These other methods, which did not include the regulator etrait in the model, had rather limited power to detect trans-regulations.
EDN construction

Since the combined eQTL mapping results from PC-, cis- and trans-mapping detected most eQTLs found in SPA and many more, we constructed an EDN based on these results, which included 28,609 regulator-target pairs. The network consisted of 4,274 genes nodes. The remaining 315 genes did not receive any inputs nor were they affecting any other genes. A total of 2,118 genes were regulators of at least one target, among which 158 did not receive any inputs. A total of 4,116 genes were targets having at least one regulator, among which 2,156 did not affect any other genes in the network. A total of 1,960 genes occurred both as regulators and targets. There were 135 instances of reciprocal regulation present (gene A directly affects gene B and vice versa). Gene PHM7 had the most targets: 468; gene YLR152C had the most regulators: 32.

The confirmed regulators or strong candidate regulator genes for the 13 eQTLs with widespread transcriptional effects identified in YVERT et al. (2003) were investigated in this EDN. Amn1, a confirmed regulator gene with widespread influence [43], was found to be a top cistrans regulator with 408 cistrans targets. The strong candidate regulator MAK5 with five coding region polymorphisms between the two parental strains [43] had 110 trans targets. Another confirmed regulator gene GPA1 [43] had 60 targets, about half of which are trans-targets. The genes LEU2 and URA3 (auxotrophic markers deleted in one of the parental strains) [43] had 98 (most were cistrans) and 32 (most were cistrans) targets, respectively. The heme-dependent
transcription factor HAP1, which has a Ty insertion in one of the parental strains [28, 43], had 141 (100 cistrans, others were trans) targets.

**Simulation results on regulator-target pair identification**

The results of our regulator-target pair identification from simulated data for the single eQTL network in Figure 3 are summarized in Table 3 in terms of power and false discovery rate for four types of simulated regulatory effects (see Figure 3 and Methods), which demonstrate that the procedure works well, with the exception of a case where some genes have extremely high and other genes low heritability (column 5 in Table 3). This problem was actually due to one of the cis-linked genes (G3) having very small residual variance and being assigned as a regulator for other genes incorrectly.

*(Insert Table 3 here)*

**Discussion**

Several different methods for eQTL mapping and regulator gene selection were evaluated in terms of their ability to construct an encompassing directed network (EDN) for gene network inference. The combined eQTL mapping results from PC-, cis- and trans-mapping detected most eQTLs found in SPA and many more. Methods which utilize multiple correlated traits in some form (PC-mapping, trans-mapping) exhibited a strong increase in the power of eQTL detection over SPA. For PC-mapping, 24% of all identified eQTLs were not detected by other methods, while for trans-mapping 87% were not found by other methods. Trans-mapping detected more than six times the number of eQTL-target pairs found by SPA, and more than four
times the number found with PC-mapping. However, the number of regulator-target pairs from trans-mapping was only increased by a factor of less than four over SPA and two over PC-mapping. For trans-mapping, the number of regulator-target pairs was less than half the number of identified eQTL-target pairs, indicating that many of these effects were indirect or more distant and were identified due to the greater sensitivity of this method.

Further development of eQTL mapping methods which utilize multiple correlated etraits appears to be warranted (e.g., performing cistrans and trans mapping on PCs). There were larger differences among eQTL mapping methods in the selection of regulator-target pairs than in the identified eQTL-target pairs. Therefore, incorporation of “external” biological information such as SNP presence in candidate regulators [46], information on protein-protein interactions [48], or information on pathway relationships [49] should be very helpful at the regulator-target selection step.

We have performed trans-eQTL mapping with a regression model incorporating the effects of a single candidate regulator etrait, its closest DNA marker and the etrait-marker interaction, or by incorporating the effects of two candidate regulator genes, their markers and etrait-marker interactions. While we have found that the marker-etrait interactions do not seem to be important, one could also consider the inclusion of additional candidate regulators, as well as interactions between the two markers (or regulators).
Sample size calculations (via simulation) should be performed for Genetical Genomics experiments, as in Kim et al. [50], with the most efficient methods for eQTL mapping and regulator-target pair identification to ensure sufficient power while containing the large expense of these experiments.

An EDN constructed as described in this study still contains many direct regulations that are actually indirect, as well as multiple candidate regulators for some eQTLs and targets. It is therefore important to perform further sparsification of the network by a search within the (constrained) network space defined by the EDN. Bayesian network analysis has been used for this purpose [46, 47], although it does not permit the reconstruction of networks with cycles. In our subsequent contribution [26], we therefore report on the use of Structural Equation Modeling to reconstruct cyclic networks in the constrained topology space of a Genetical Genomics experiment.

**Conclusions**

In this contribution we consider cis, cistrans and trans regulations jointly with the goal of constructing an EDN that defines the search space for further gene network inference. We present a method for EDN construction using several approaches for eQTL mapping and local structural models for regulator-target pair identification. Because PC-, cis- and trans-mapping combined detected most of the eQTLs found by SPA and many more, we constructed the EDN based on the eQTL results from these methods. The resulting EDN is very dense due to the high power of the eQTL mapping methods used. It still contains many direct regulations that are actually indirect, as well as multiple candidate regulators for some eQTLs and targets. It is
therefore important to perform further sparsification of the network, which is discussed in our subsequent contribution [26].

**Methods**

The methodology we discuss here can be applied to any organism where a segregating population is extensively marker genotyped and expression profiled, and where DNA sequence information is available. Currently several such datasets have been produced, most notably for yeast [29] and mouse [51]. For evaluation purposes we analyzed the yeast Genetical Genomics dataset [29]. After removing the 20% of genes with the lowest etrait variability from the original data, our dataset contained etraits for 4,589 genes and genotypes for 2,956 genetic markers on 112 haploid offspring from a cross between a laboratory and a wild strain. Observations with missing marker genotypes were excluded.

**Single Profile Analysis**

Marker linkages were tested using the Kruskal-Wallis test [52]. QTL confidence intervals (CIs) were obtained by searching for the closest markers on either side of the significant eQTL marker that show a decrease in the logarithm-of-odds (LOD) score of at least 1 [53]. LOD 1 intervals are approximately equivalent to a 96.8% CI [54]. The Kruskal-Wallis test statistic follows an approximate chi-square distribution under the null hypothesis, and approximate LOD scores were computed by dividing the Kruskal-Wallis test statistic by $2 \times \ln(10)$. If multiple eQTLs on the same chromosome had significant effects on the same etrait, they had to be separated by at least two insignificant markers to be regarded as different eQTLs. To identify chromosomal regions affecting multiple etraits, the eQTL regions of two different etraits were
combined into a single region if the two eQTLs were located at the same marker or their CIs overlapped by over 80%.

The nominal p-values were calculated based on normality assumptions. Rebaï [55] showed that even if the data are not normally distributed, a normal approximation will not give misleading results if the distribution is not too extreme. We verified this assumption using permutation tests on 200 genes and observed that the nominal p-values were very close to the permutation-based p-values and were slightly more conservative for over 90 percent of all genes with small p-value (< 0.02) and only very slightly smaller (< 0.00001) for the remaining genes with small p-values. Therefore, nominal p-values were used in this study. The p-values were adjusted for multiple testing by controlling the FDR using the BH-procedure [27].

Sliding three marker regression [30, 56] was performed to refine the original LOD CIs. A marker to be tested was fitted together with a pair of flanking markers in a regression model. The marker of interest has an expected nonzero partial regression coefficient if and only if at least one QTL is located between the flanking markers [57]. To select the appropriate flanking markers, we chose the closest markers having at least 20 recombinants with the tested marker. The number of recombinants required was determined based on a compromise between achieving sufficient power for the tested marker and sufficient proximity of the flanking markers to block the effects of the linked eQTL. A test statistic profile was obtained for all markers in each original LOD CI, and new confidence regions within each original CI were identified by significant regions separated by at least two non-significant markers.
Principal Components Mapping

PCs were first computed on the total set of 4,589 etraits. Subsequently, to detect eQTLs affecting smaller subsets of genes, we clustered genes and applied PCA separately to the clusters. We used k-means with absolute correlation as the distance measure to cluster genes into 100 subsets with the software Cluster 3.0 (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm). We chose 100 clusters because we found considerably more eQTLs with 100 than with fewer (e.g. 20) clusters, and because with more than 100 clusters, many small clusters were broken into clusters containing only one or two genes. The number of genes in each cluster varied from 3 to 232, with an average of 46 genes per cluster. An eigenvalue cutoff of 1.5 was used to determine how many PCs to retain for each cluster, so that the PCs from different clusters contained a similar amount of information. Then, eQTL mapping was performed on these “composite etraits” (PCs) rather than on the individual etraits in the same way as in SPA. An eQTL affecting a PC is assumed to be a common regulator of all etraits with high loadings on the affected PC. However, there was no clear cutoff for “high” loadings (Figure 1). Therefore, instead of choosing an arbitrary cutoff, all etraits were individually tested as in the SPA but only for the identified PC-eQTL regions.

Cis-eQTL Mapping

Cis-acting QTL effects were tested using the same non-parametric test as in SPA. For those etraits with a significantly linked cis-marker, LOD confidence intervals were obtained using the LOD results from the SPA. We searched both sides of the tested marker for a LOD 1 drop and recorded the maximum LOD marker. If the gene of the tested etrait fell outside the CI and was more than 10 kb away from the maximum, we discarded the eQTL. Local maxima were taken to be cis-eQTLs and were combined in
the same way as in the SPA. In the cistrans analysis, we tested all etraits for the effects of the identified cis-eQTL.

**Trans-eQTL Mapping**

For trans-eQTL mapping, we used a regression approach, as the Kruskal-Wallis test [52] cannot incorporate the etrait of a candidate regulator gene. Our regression model for the etrait of gene \( i \) includes the effects on etrait \( i \) of a candidate regulator \( j \neq i \), the genotype of the marker closest to the physical location of gene \( j \) and the regulator etrait by genotype interaction term:

\[
y_{in} = b_1y_{jn} + b_2x_{jn} + b_3y_{jn}x_{jn} + \varepsilon_{in}
\]

where \( y_{in} \) is the deviation of etrait value \( i \) in observation \( n \) from its mean; \( y_{jn} \) is the deviation of the potential regulator etrait \( j \) in observation \( n \) from its mean; \( x_{jn} \) is the deviation of the genotype value (0 or 1) of the marker closest to candidate regulator gene \( j \) from its mean, and \( \varepsilon_{in} \) represents the residual. Regression coefficients \( b_1 \) and \( b_2 \) must both be significantly different from zero for gene \( j \) to be declared as a trans-regulator of gene \( i \), as determined by the Intersection Union Test (IUT) [44, 45]. The null hypothesis of the IUT is that either \( b_1 \) or \( b_2 \) is zero, or both are zero, and the IUT rejects the null hypothesis if and only if all individual \( H_{0k} \) (\( H_{01}: b_1 = 0; H_{02}: b_2 = 0 \)) have been rejected. In our case, the IUT is performed by calculating a p-value for both individual null hypotheses, retaining the larger of the two p-values for each target, and performing BH FDR adjustment on the retained p-values.

There are two reasons why the trans-analysis based on model (1) and the IUT might give misleading results: The presence of a cis-eQTL affecting the target and
multicollinearity between $y_j$ and $x_j$. We therefore did not consider any candidate regulator whose closest marker had a recombination rate of 0.25 or less with the marker closest to the target etrait. For each gene we calculated the correlation between its expression profile and the genotype profile of its closest marker. The distribution of correlation values was very close to normal with only 1.11% (0.15%) of all genes having absolute values exceeding 0.65 (0.9). We also performed multicollinearity tests including the condition index (e.g., [58]; a condition index > 30 indicates a multicollinearity problem). In our data the highest condition index was 12 and most values were much lower (between 1 and 3). The values indicate that our trans-mapping results should essentially be unaffected by multicollinearity.

**Identification of regulator-target pairs for SPA, PCA and cis-mapping**

We used local structural models to select regulator genes in each of the identified eQTL CIs. The candidate regulator selection was performed in three steps: 1) Identification of the detected cis-linked etraits that were most likely truly cis-linked and those that were probably secondary (cistrans) effects, 2) Identification of the detected trans-affected etraits that were probably cistrans-affected and those that were more likely trans-affected, and 3) Search for the candidate regulator among the genes physically located in the eQTL confidence interval for each of the likely trans-affected etraits.

**Distinguishing cis from cistrans**

Some of the etraits found to be cis-regulated based on the fact that they were affected by an eQTL whose CI overlapped with the physical location of the gene may not be truly cis-affected. Such a gene may be cistrans regulated through a cis-affected gene,
or trans regulated by some coding region polymorphism in a gene located near the target gene. We tested whether a potentially cis-affected gene was likely truly cis-affected using model (1) but omitting the interaction term, letting $y_{in}$ be the value of the potentially cis-affected target etrait $i$, $y_{jn}$ the value of another potentially cis-affected regulator $j$ of $i$, and $x_{jn}$ the genotype of the marker at which the peak test statistic of the eQTL CI occurs. If $y_i$ is actually cistrans-affected through $y_j$, then $b_2$ should not be significantly different from zero with $y_j$ included in the regression equation. These tests were carried out for all identified cis-affected etraits in an eQTL. If for an etrait $i$, $b_2$ remained significant for all etraits $j$, then it was identified as a “true” cis-affected etrait. Since the effects were already identified as significant in the eQTL analysis, no multiple testing adjustment was applied and a p-value cutoff of 0.05 was used.

**Distinguishing trans from cistrans**

After having identified the likely cis-affected etraits, we focused on the etraits trans-affected by the eQTL. The trans-affected etraits can be either truly trans-affected or cistrans-affected through a cis-affected regulator. Using model (1) again, $y_{in}$ is now a trans-affected etrait and $y_{jn}$ is a cis-affected etrait identified in step 1. Cistrans regulation is indicated by $b_2$ not being significantly different from zero. If $b_2$ remains significant for all cis-affected etraits $j$, then gene $i$ is identified as a likely trans-affected etrait.
Selecting candidate trans-regulators in the same eQTL region

To find the candidate regulator(s) for a likely trans-affected etrait \( i \) among all genes physically located in the eQTL region, for target etrait \( i \) we fitted model (1) with any candidate regulator etrait \( j \) located in the eQTL region and the eQTL marker (without the interaction term), and any additional candidate etrait \( p \). The additional candidate etrait was included to examine whether the regulator-target correlation was due to some indirect mechanism. For each candidate-target pair \( i \& j \), the null hypothesis is that at least one of the \( b \) coefficients of \( j \) is not significantly different from zero after having a different candidate etrait \( p \) in the model. Therefore, we retained the maximum p value of all the \( b \) coefficients of the candidate regulator of each target as in (another application of the IUT principle; see also [11, 59]). We used a p-value cut off of \((0.05 / \text{number of candidate regulators})\) to control the family-wise error rate at 0.05 for all tests performed for each eQTL-target pair. A candidate regulator with significant maximum p-value was retained as a regulator of its target.

Identification of regulator-target pairs for trans-mapping

In the trans-eQTL mapping analysis based on model (1), some of the identified regulator genes may be false positives, because the regulator etrait might be significant due to some correlating mechanisms other than the hypothesized direct causal relationship between the regulator and the target, and/or the marker might be significant due to the linkage with another polymorphism. For example, regulator \( j \) might directly affect another regulator \( k \), which in turn affects target \( i \), without a direct causal relationship between \( j \) and \( i \). To eliminate such cases, we performed some “local sparsification” for each target etrait. For each target etrait \( i \) with at least two identified regulators, for each identified regulator \( j \) of etrait \( i \), we included another identified regulator etrait \( k \) of \( i \) and its nearest marker in the regression model:
\[ y_{in} = (b_{1j}y_{jn} + b_{2j}x_{jn} + b_{3j}y_{jn}x_{jn}) + (b_{1k}y_{kn} + b_{2k}x_{kn} + b_{3k}y_{kn}x_{kn}) + e_{in} \]  

(2)

The terms for \( i \) and \( j \) are the same as in model (1), with additional etrait, marker and interaction terms for gene \( k \). The interaction terms \( (b_3) \) were usually dropped from the model due to a lack of significance, and the marker for regulator \( k \) \( (b_{2k}x_{kn}) \) was not included in the model when its recombination rate with the marker for regulator \( j \) was less then 0.25. Since the candidate regulators included in the model were already identified as significant in the trans-mapping, no multiple testing adjustment was performed and the IUT was applied at a \( p \)-value cutoff of 0.05. If the IUT for regulator \( j \) was not significant, then we discarded gene \( j \) as a regulator.

**EDN construction**

The eQTL mapping and target-regulator pair selection steps resulted in three lists of causal regulatory relationships: (1) a list containing all identified cis-regulations (eQTL A affects gene A located in its confidence region), (2) a list containing all cis-trans regulations (cis-regulated gene A regulating gene B), and (3) a list containing all trans-regulations (gene A regulating gene B and eQTL A affecting gene B (but not gene A)). To construct an encompassing directed network (EDN), we assembled all the identified and retained regulator-target relationships, which consisted of directed edges (representing causal influences) from eQTLs to cis-regulated target genes, from cis-regulated genes to cistrans regulated target genes, from trans-regulator genes to target genes and from trans-eQTLs to target genes. The EDN consisted of two types of nodes: continuous nodes for the genes (etraits), and discrete nodes for the eQTLs (genotypes). Despite our “local sparsification” (see for example the section *Identification of regulator-target pairs for trans-mapping*), the network obtained from
eQTL analysis and regulator-target pair identification contains (many) direct regulations that are actually only indirect, which need to be eliminated for final network inference. We therefore call this network an ‘encompassing’ directed network, consistent with the terminology of an encompassing (regression) model in statistics, which contains a maximal set of regressors (edges in our case) some of which need to be deleted as their effects are zero.

**Simulation study on regulator-target pair identification**

We evaluated our regulator-target pair selection by simulation. For a population of 112 individuals (as in the yeast data), we simulated an eQTL region containing three eQTL causal polymorphisms and several candidate regulator and target genes. This local network is depicted in Figure 3, with G (Q) representing a gene (eQTL). The target list for the eQTL region is T = [G2, G3, G4, G5, G6, G7, G8]. Gene G1 is the only candidate trans-regulator, while genes G3, G4, G6 and G7 are candidate cis-regulators. There are four types of regulations: one true trans-regulation (from G1 and Q1 to G2); two true cis-regulations (Q2 to G3 and Q3 to G6); two true cistrans-regulations of targets located in the eQTL region (Q2 to G3 to G4 and Q3 to G6 to G7); and two true cistrans-regulations of targets not located in the eQTL region (Q2 to G3 to G5 and Q3 to G6 to G8).

(Insert Figure 3 here)

Data were simulated with linear regression models with regression coefficients fixed at the value of 1 and residual standard deviations (SD) set to 0.125, 0.25 or 0.5 (one value for all genes, or for genes with odd numbers SD = 0.5 or 0.25, and for genes
with even numbers SD = 0.25 or 0.125). For a gene directly regulated by an eQTL, the model was $y = bx + e = x + e$, where $x$ is QTL genotype (0/1), variance due to the eQTL was equal to 0.25, and heritability = $0.25/(0.25+SD^2) = 0.941$, 0.80 or 0.50 for the three SD values, respectively. For a gene indirectly regulated by an eQTL (Q2→G3→G4), the model was $y_2 = b(bx + e_1) + e_2 = x + e_1 + e_2$, and heritability = $0.25 / (0.25 + 2SD^2) = 0.889$, 0.667, and 0.333. The three causal polymorphisms in the eQTL region had order Q1 - Q2 - Q3 (see Figure 3) with recombination rate $r = 0.0$ or $r = 0.09$ between adjacent polymorphisms. A total of 1000 data replicates were simulated and analyzed for each of several combinations of SD and $r$ values chosen (see Table 3).

**Authors' contributions**

Dr. Bing Liu performed the analyses for SPA, cis-mapping, PC-mapping, and comparison of the results, and she drafted the manuscript. Dr. Alberto de la Fuente performed the analyses for trans-mapping, regulator-target pair identification, EDN construction, and the simulation study for regulator-target pair identification. Dr. Ina Hoeschele conceived and directed the work.

**Acknowledgements**

We thank Rachel Brem and Leonid Kruglyak for sharing the genotype data with us and for providing the raw data of the spotted microarray experiments at the National Center for Biotechnology Information / Gene Expression Omnibus website, [http://www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo). This work was supported by National Science Foundation cooperative agreement DBI-0211863.
References

26. Liu B, de la Fuente A, Hoeschele I: Gene network inference via structural equation modeling in Genetical Genomics experiments. BMC Genomics 2007(This issue!).


Figures

Figure 1 - Sorted gene loadings of the first 10 Principal Components (PCs) of all 4,589 filtered genes

For each of the first ten PCs of the filtered genes, the genes were sorted based on their loading on the PC. The loadings were scaled by dividing by the maximum loading for that PC.

Figure 2 - Test statistic profiles of an etrait on two chromosomes

Red lines: SPA profile and its threshold. Red arrows: Locations of significant eQTLs detected by SPA. Blue lines: Trans-mapping profile and its threshold. Blue arrows: Locations of significant QTLs detected by trans-mapping. Approximate LOD test statistics were computed by dividing the test statistics by 2*ln(10).

Figure 3 - The network model used in the simulation study

Black squares with starting letter Q: causal polymorphisms in the same eQTL region
Black squares with starting letter G: genes located in the eQTL region. White squares with starting letter G: genes not located in the eQTL region but affected by it. Solid arrows: true trans-regulations; dashed arrows: true cis-regulations; dotted arrows: true cistrans regulations of target genes located in the eQTL region; dashed-dotted arrows: true cistrans regulations of target genes not located in the eQTL region. Solid lines among the Q polymorphisms represent direct genetic linkage (recombination rate equal to 0 or to 0.09).
Tables

Table 1. The percentages of eQTL-target pairs identified by a particular eQTL mapping method (row) whose confidence intervals had any overlap with other eQTL mapping methods (columns) or no overlap (last column)

<table>
<thead>
<tr>
<th></th>
<th>SPA</th>
<th>Cis-mapping</th>
<th>PC-mapping</th>
<th>Trans-mapping</th>
<th>No overlap with any method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>83%</td>
<td>86%</td>
<td>24%</td>
<td></td>
<td>3%</td>
</tr>
<tr>
<td>Cis-mapping</td>
<td>91%</td>
<td>77%</td>
<td>21%</td>
<td></td>
<td>2%</td>
</tr>
<tr>
<td>PC-mapping</td>
<td>65%</td>
<td>52%</td>
<td>18%</td>
<td>24%</td>
<td></td>
</tr>
<tr>
<td>Trans-mapping</td>
<td>12%</td>
<td>8%</td>
<td>11%</td>
<td>87%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The percentage of regulator-target pairs identified based on a particular eQTL mapping method (row) that are in common with another eQTL mapping method (columns) or with no other method (last column)

<table>
<thead>
<tr>
<th></th>
<th>SPA</th>
<th>Cis-mapping</th>
<th>PC-mapping</th>
<th>Trans-mapping</th>
<th>No overlap with any method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>41%</td>
<td>41%</td>
<td>7%</td>
<td></td>
<td>41%</td>
</tr>
<tr>
<td>Cis-mapping</td>
<td>42%</td>
<td>41%</td>
<td>6%</td>
<td></td>
<td>41%</td>
</tr>
<tr>
<td>PC-mapping</td>
<td>26%</td>
<td>26%</td>
<td>5%</td>
<td></td>
<td>62%</td>
</tr>
<tr>
<td>Trans-mapping</td>
<td>3%</td>
<td>2%</td>
<td>3%</td>
<td></td>
<td>95%</td>
</tr>
</tbody>
</table>

Table 3 - Results from a simulation study on regulator-target pair identification in a single eQTL region with three causal polymorphisms, and with multiple candidate regulator and target genes (true network structure is in Figure 3)

<table>
<thead>
<tr>
<th></th>
<th>SD=0.5</th>
<th>SD=0.25</th>
<th>SD=0.125</th>
<th>SD=0.5/0.125</th>
<th>SD=0.5/0.25</th>
<th>SD=0.25/0.125</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cis-link Power (%)</td>
<td>100, 100</td>
<td>100, 100</td>
<td>100, 100</td>
<td>55.3, 59.85</td>
<td>89.4, 98.65</td>
<td>97.8, 98.5</td>
</tr>
<tr>
<td>Cis-link FDR (%)</td>
<td>0.6, 0.9</td>
<td>0.7, 0.67</td>
<td>0.67, 0.57</td>
<td>0.48, 0.53</td>
<td>0.6, 0.72</td>
<td>0.62, 0.57</td>
</tr>
<tr>
<td>Cis-reg cis Power (%)</td>
<td>99, 99</td>
<td>99, 99</td>
<td>99, 99</td>
<td>54.8, 59.4</td>
<td>88.6, 97.8</td>
<td>97.2, 97.8</td>
</tr>
</tbody>
</table>
Power: percentage of replicate data sets in which the regulation type was found; FDR: percentage of replicate data sets in which a regulation of a certain type was found that did not exist in the underlying network; Cis-link: cis-regulation of target in eQTL region; Cis-reg: cistrans-regulation of target not in eQTL region; Cis-reg cis: cistrans-regulation of target in eQTL region; Trans-reg: trans-regulation. For the last three columns, even numbered gene nodes (Figure 3) received the left amount of error variance and odd numbered nodes the right amount. The two numbers in each cell correspond to 0% recombination and 9% recombination among the three causal polymorphisms in the single eQTL region, respectively. A p-value cutoff of 0.01 was used.

### Additional files

**Additional file 1 – APPENDIX: DATA PREPROCESSING**

pdf file online at http://www.vbi.vt.edu/inah.

**Additional file 2 – The Encompassing Directed Network**

The network can be found online in several file formats at http://www.bioinformatica.crs4.org/Members/alf/bmc/.