eQED: an efficient method for interpreting eQTL associations using protein networks

Silpa Suthram1,2, Andreas Beyer2,3, Richard M Karp4, Yonina Eldar5 and Trey Ideker1,2,*

1 Bioinformatics Program, University of California San Diego, La Jolla, CA, USA, 2 Department of Bioengineering, University of California San Diego, La Jolla, CA, USA, 3 Biotechnology Center, TU Dresden, Dresden, Germany, 4 Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA, USA and 5 Department of Electrical Engineering, Technion, Haifa, Israel

* Corresponding author. Department of Bioengineering, University of California San Diego, 9500 Gilman Dr 0412, La Jolla, CA 92093-0412, USA.
Tel.: +1 858 822 4558; Fax: +1 858 822 4246; E-mail: trey@bioeng.ucsd.edu

Received 14.12.07; accepted 21.12.07

Introduction

The technique of expression quantitative trait loci (eQTLs) is becoming increasingly widespread for revealing the genetic loci in control of specific changes in gene expression (Brem and Kruglyak, 2005; Schadt et al, 2005). eQTLs are a variant of the more basic concept of quantitative trait loci, which measures the association between a quantitative phenotype (such as height and weight) and a panel of polymorphic genetic markers distributed across the genome (Griffiths, 2002). For the special case of eQTL analysis, the phenotype of interest is a gene expression level measured with DNA microarrays (Brem and Kruglyak, 2005). Since a microarray monitors expression levels of all genes, separate statistical tests are performed to compute scores of association of each genetic marker with each gene expression level.

Two of the core challenges (Rockman and Kruglyak, 2006; Schadt and Lum, 2006) in understanding and explaining eQTL associations are

1. Fine mapping: Due to the spacing of genetic markers and/or linkage disequilibrium, several genes can reside near each marker. Typically, no more than one of these genes is responsible for the observed expression phenotype. Identifying the true causative gene requires additional data, since all genes at a locus are indistinguishable based on the eQTL measurements alone.

2. Lack of mechanistic explanation: A gene–phenotype association typically lends little insight into the underlying molecular mechanism for the association.

Several bioinformatic approaches have been proposed recently to address these two issues (Schadt et al, 2005; Kulp and Jagalur, 2006; Lee et al, 2006; Tu et al, 2006; Perez-Enciso et al, 2007). For the problem of 'fine mapping', the main bioinformatic focus has been on predicting which genes within a given locus are the true regulators of expression of the target phenotype. For instance, Kulp and Jagalur (2006) sought to infer the true causal genes using a Bayesian network model.
constructed from expression correlations detected within the eQTL profiles. Another powerful approach has been to complement eQTLs with data on physical molecular interactions. Tu et al (2006) modeled each eQTL association as a sequence of transcriptional and protein–protein interactions (PPIs) that transmits signals from the locus to the affected target. This method is promising since it prioritizes candidate genes by their network proximity to the affected target gene and also provides a model of the underlying regulatory pathways. In addition, assembly of protein interaction networks is a burgeoning area in genomics and the amount and quality of protein interaction data are rapidly improving. Integrating eQTL data with additional independent information may significantly reduce the noise and improve the statistical power of the analysis (Beyer et al, 2007).

Here, we describe a new integrative approach (named ‘eQTL electrical diagrams’ or eQED), which also combines eQTL data with protein interaction networks but predicts the true causal gene at each locus with substantially higher accuracy than the previous method. eQED models the flow of information from a locus to target genes as electric currents through the protein network. Currents can be simulated simultaneously for all loci influencing a target, allowing multiple loci to reinforce each other when they fall along a common regulatory pathway.

Results and discussion

Definition of terms

In what follows, the genes near a polymorphic genetic marker are called candidate genes, and the genes with an associated change in expression are called targets. The particular candidate gene that is truly responsible for the downstream change in expression of a target is called the true causal gene. Collectively, the set of candidate genes near a marker defines a locus. Finally, the proteins and their interactions in the protein network are referred to as nodes and edges, respectively.

Open problems motivated by the previous method

For a given locus and associated target, the Tu et al method works by executing a random walk through the protein network starting at the target. At every step of the walk, the next edge to be followed depends on its predefined weight (see Materials and methods). The walk ends when it reaches one of the candidate genes in the locus. The random walk is repeated 10 000 times, and the candidate gene that is visited most often is predicted to be the true causal gene. Figure 1A shows a sample network, while Figure 1B shows a sample random walk on this network according to the Tu et al approach. Gene L3 is visited most often and, hence, is reported as the causal gene.

Given that the protein network is large, many random walks must be executed for the predictions to be accurate. Moreover, a single random walk from the target to any candidate gene may require many steps. These two issues can lead to random walk simulations that last a prohibitively long time. In Tu et al, the authors make a key approximation that allows them to achieve feasible simulation times: they constrain the path taken by each random walk to be acyclic (i.e. no genes can be revisited). As a consequence, many walks result in ‘dead ends’ unable to reach any candidate gene, but all walks are at least relatively short. This ‘greedy’ approximation may lead to different predictions from typical random walk models (Doyle and Snell, 1984), which may affect their accuracy. In addition, since biological networks are scale free (Albert-László Barabási, 1999), they contain a large number of dead ends (i.e. nodes with a single edge). The many dead ends greatly reduce the absolute number of visits to the candidate genes, thereby reducing the overall confidence in the final causal gene prediction for a given number (e.g. 10 000) of walks.

The eQED model

The eQED approach seeks to address the above open problems by replacing the random walk model with a framework based on electric circuits. There is considerable prior work establishing the equivalence between electric networks and random walks (see Materials and methods). The eQTL associations and the corresponding protein network are abstracted as an analog electric circuit model grounded at a given target gene. The weights on the edges of the molecular network are modeled as conductances (1/resistance) in the electric circuit. The $P$-values of association between each genetic locus and expression of the target are modeled as independent sources of current. An electric circuit abstraction is constructed for every locus–target association (which we call the single-locus model, Figure 1C). Further details of the model are provided in the Materials and methods section.

After solving the circuit for currents, the causal gene is predicted as the one with the highest current running through it. Analyzing the network as an electric circuit provides a deterministic ‘steady-state’ solution, in contrast to a stochastic random walk. Moreover, the number of dead-end nodes in the network does not affect the final result as the total current through them is always zero (Figure 1C).

Application to eQTL associations in yeast

As a proof of principle, we applied the eQED approach to analyze the results of a genome-wide eQTL study in yeast by Brem and Kruglyak (2005). This study reported associations between 2956 genetic markers and 5727 gene expression levels measured across 112 yeast strains (Materials and methods). All locus–target pairs with a gene association $P$-value $\leq 0.05$ were considered; within this set, we selected only those loci containing more than one candidate gene (i.e. for which the true causal gene was ambiguous). At the same time, we assembled a pooled interaction network consisting of 17171 transcriptional and PPIs reported in previous large-scale studies (Materials and methods). Given this network, the set of locus–target pairs was further filtered to include only those loci for which at least two of their candidate genes had at least one transcriptional or PPI, yielding a total of 131 863 locus–target pairs. The single-locus model of eQED was applied to each locus–target pair, and a causal gene prediction was made in each case. This step-by-step procedure is diagrammed in Figure 2.

To estimate the accuracy of the predictions, we compiled a set of ‘gold standard’ cause–effect pairs from two large gene
knockout expression profiling studies in yeast, Hughes et al (2000) and Hu et al (2007), as well as from a gene overexpression study by Chua et al (2006). In these studies, strains harboring a single gene knockout or overexpression construct (the ‘true causal gene’) had been analyzed using whole-genome microarrays to identify a resulting set of differentially expressed genes (the ‘targets’). We filtered these three data sets to include only those causal gene–target pairs that were present in the molecular network used by eQED and for which the causal gene was associated with the target gene at $P < 0.05$ in Brem and Kruglyak (see Materials and methods). The resulting gold-standard set contained 548 causal gene–target pairs.

Table 1 reports the number of correct predictions of the causal gene for each method. The single-locus model of eQED correctly predicted 392 of the 548 gold standards (72% accuracy). In comparison, the approach by Tu et al (2006) achieved 50% accuracy. Both methods performed substantially better than random selection of a gene at a locus, which achieved 22% accuracy.

**Combining multiple loci**

In our model, given a target gene and a corresponding significant marker, there exists only one causal gene. However, in eQTL studies, the expression level of a target gene typically has significant associations with more than one marker (and thus more than one causal gene). If these causal genes fall along common regulatory pathways, considering multiple loci together in the same eQED model might increase our confidence in the causal gene predictions. Motivated by these considerations, we explored a second circuit model, called multiple-loci eQED, in which currents were included for all significant loci associated with a target (see Materials and methods). For example (Figure 1E), assume the target T associates significantly with two loci. In the single-locus model, eQED would predict a causal gene based on one of the loci, while in the multiple-loci model, the causal gene would be identified based on the combined expression levels from all significant loci.
Electric circuit analysis of eQTL data
S Suthram et al

model, we would investigate the two associations separately, but in the multiple-loci model their information is processed as a single circuit. Figure 1F shows a schematic of the multiple-loci model from Figure 1E. For each locus considered, the causal gene is predicted as the one having the highest current flowing through it.

The accuracy of the multiple-loci eQED model was estimated using the same gold-standard data set used for the single-locus model. As shown in Table 1 and Supplementary Table 1, the multiple-loci model boosted prediction accuracy substantially over the single-locus case (80 versus 72%). Combining information from all significant loci for a given target also reduces computation time, as all loci are processed in a single eQED simulation instead of multiple runs.

Predicting the direction of signaling along protein interactions

A direct consequence of the electric circuit model is that the currents on the wires of the network suggest a direction of information flow in the biological system. In the case of transcriptional interactions, the current is restricted to flow from the transcription factor (TF) to the regulated gene, and not vice versa (Materials and methods). In contrast, the direction of information flow along PPIs is not predetermined, since the underlying biochemical measurements typically report only whether an interaction exists, not its functional consequences. Therefore, for PPIs in particular, eQED provides a means of predicting the direction of signal transmission.

eQED induces a current on each PPI in the network. Repeated application over all targets yields a distribution of current values for each interaction. This distribution can be analyzed to determine whether the current is predominantly positive or negative (prior to the analysis, positive and negative directions of flow are defined arbitrarily for each interaction). We evaluated three simple methods for summarizing this distribution of currents, by using either (1) the most extreme current; (2) the sum of currents or (3) the skewness of the current distribution. Each of these three methods yielded a single value per interaction whose sign was interpreted as the predicted direction and whose magnitude could be used to rank the predictions in order of confidence.

Table 1 Causal gene prediction accuracya

| Methods        | Number of correct predictions |
|----------------|------------------------------|
| Random         | 118                          |
| Tu et al       | 262                          |
| Shortest pathb | 351                          |
| eQED (single loci) | 392                      |
| eQED (multiple loci) | 438                      |

aAll predictions were tested against a gold-standard data set of 548 causal gene–target pairs compiled from yeast gene expression knockout studies by Hughes et al (2000) and Hu et al (2007) and a gene overexpression study by Chua et al (2006).

bA naive method in which the causal gene is selected to be the gene at the locus that is connected by the shortest path to the target.
To assess the performance of directionality prediction, we once again compiled a set of gold standards, consisting of PPIs for which the signaling directions are known. A total of 408 gold-standard interactions were obtained, including 103 signaling interactions reported in the Kyoto Encyclopedia of Genes and Genomes (Kanehisa, 2002) or the Munich Information center for Protein Sequences (Mewes et al., 2006), as well as the 596 (top 10%) highest confidence kinase–substrate interactions reported in a systematic analysis of phosphorylation by Ptacek et al. (2005). Figure 3A shows the accuracy of the three methods at recapitulating the known directions of signaling. Although the ‘sum of currents’ method yielded very high accuracy (>80%) for the 40 highest ranking predictions, the ‘most extreme current’ method retained moderate accuracy (generally >75%) out through the best 80 predictions (corresponding to the largest area under the curve). In contrast to these first two methods, the third method based on ‘skewness’ was not an accurate predictor of directionality.

On the basis of this analysis, we used the ‘most extreme current’ method to predict directionality of information flow for all PPIs in the eQED network. A total of 368 predictions (with absolute most extreme current ≥623, corresponding to the 75% accuracy mark above) are provided in Supplementary Table 2.

**Prediction of regulatory pathways**

The currents computed by eQED provide an estimate of the influence of each protein interaction on the regulation of the target gene. To reveal how individual high-current interactions might assemble into regulatory pathways, we sought to connect each causal gene to its target by finding an optimal path through the network, defined as the shortest route with the highest total sum of currents across its interactions. The union of all optimal paths leading from each predicted causal gene into a given target reveals its regulatory network. We also filtered the regulatory network to include only those PPIs that have a predicted direction of influence (see previous section). Figure 3B–D shows the regulatory network obtained for three example target genes: HMG2, AAD15 and ARG5/6. Although the causal genes are often at the head of each path comprising the regulatory network, in some cases a path contains a chain of causal genes in series. For instance, both ARO80 and RLR1 associate significantly with the target ARG5/6 and share the same regulatory pathway. This is a direct consequence of

---

**Figure 3** Inferred pathways and directionality prediction. (A) The accuracy of the direction prediction methods. The ‘gold’ standard protein interactions were ranked according to the different metrics (x-axis), and the cumulative percent accuracy represented as y-axis. (B–D) The regulatory networks for three example target genes. The nodes colored in shades of red correspond to predicted causal genes. The intensity of color corresponds to their P-value of association with the target.
integrating the information about all significant loci when running eQED (multiple-loci model). As a result, the casual genes not only reinforce each other but also increase the overall confidence of the underlying regulatory network.

**Application to gene association studies in humans**

During the past few years, a substantial body of eQTL data has been generated in higher eukaryotes, including a number of studies in mouse and Arabidopsis thaliana (see www.gene-network.org). Large eQTL studies are now also available for humans (Dixon et al., 2007; Goring et al., 2007; Stranger et al., 2007). All of these datasets associate genetic loci with gene expression levels without explicitly identifying the causal genes at each locus, raising the important question of whether they could be identified using an integrative network-based approach such as eQED.

Clearly, a network-based analysis of human eQTLs will require a substantial database of protein–protein and transcriptional interactions. In terms of PPIs, several large networks have recently been mapped for humans (Rual et al., 2005; Stelzl et al., 2005; Mathivanan et al., 2006). The remaining hurdle is the availability of large-scale measurements of transcriptional interactions. Although no systematic study has yet been published, several such efforts are underway using systematic chromatin immunoprecipitation experiments in human cell lines and *in vitro* technologies such as the protein binding microarray (Berger et al., 2006). As these networks become available, the success of eQED in yeast suggests that it may also provide a powerful means for identifying human disease genes and their associated transcriptional regulatory pathways in higher eukaryotes.

**Materials and methods**

**Electric circuit and random walks**

There is considerable literature establishing the analogy between random walks and electric networks (Doyle and Snell, 1984; Faloutsos and Tomkins, 2004; Newman, 2005). In particular, Doyle and Snell (1984) showed that there always exists a random walk equivalent of linear electrical circuits. Random walks on a network can be abstracted as a Markov chain and consequently, be represented using a transition state matrix. Consider an electric network $E$ where the conductance on an edge $(x, y)$ is represented by $C_{xy}$. A random walk can then be defined on $E$, which has the transition state probabilities: $P_{xy} = C_{xy} / C$, where $C = \sum_{x,y} C_{xy}$ and $N(x)$ is the set of neighbors of $x$ in the network.

Since an electric network is a connected graph, it is possible to travel between any two states. A Markov chain with such a property is known as an *ergodic chain*. For an ergodic chain represented by the transition matrix $P$, there exists a fixed vector $\mathbf{w} = \{w_1, w_2, \ldots, w_n\}$, such that $w_i = w_{ij}P_{ij}$, where the component $w_i$ represents the steady-state proportion of times the walker remains in state $i$. In the case of random walks derived from electric networks, it can be shown that $w_i = C_i / C$, where $C = \sum_x C_x$.

An ergodic chain is called *time reversible* if $w_i P_{xy} = w_j P_{yx}$. Thus, in the case of the random walk derived from an electric circuit,

$$w_i P_{xy} = \frac{C_{xy}}{\sum_x C_x} \frac{C_{yx}}{\sum_y C_y} = \frac{C_{xy} C_{yx}}{\sum_x C_x \sum_y C_y} = \frac{C_{xy} C_{yx}}{C_x C_y} = w_j P_{yx}$$

As a result, the random walk $P$ is also time reversible. Finally, using the above properties we can show that when a unit current flows into an electric network at node ‘$a$’ and leaves at node ‘$b$’, then the amount of current through any intermediary node or edge is proportional to the expected number of times a random walker will pass through that node or edge (see Doyle and Snell (1984) for details).

We demonstrate this equivalence using the sample network of Figure 1A. Figure 1C is the electric network model of the sample network. Here, we add a new node $L$, which is connected to all the candidate genes at the locus. The edges connecting $L$ to $L_1$, $L_2$ and $L_3$ have infinite conductance and for all purposes, $L$ is no different from any of $L_1$, $L_2$ or $L_3$. The conductance on the remaining edges is equal to their weight in the sample network. The target gene $T$ is treated as ‘ground’ for the electric network. There is an independent source of current sending 10 000 A of current into the network at $L$. We solve the network using Kirchhoff’s and Ohm’s Laws (Irwin and Wu, 1999) to get the currents through each edge and node. Figure 1D shows the sample network represented as a random walk derived from the electric network of Figure 1C. The random walk is repeated 10 000 times. The number of times each edge and node was visited in the random walk converges to the amount of current through those edges and nodes in the electric network (Figure 1C and D, and Supplementary Information).

**eQTL associations**

Yeast eQTLs were obtained from Brem and Kruglyak (2005), consisting of whole genome expression data for 112 yeast strains, which were genotyped across 2956 genetic markers. Genetic similarity between strains, referred to as population substructure, can lead to false-positive relationships where the observed phenotype correlates well with the phylogenetic relationships between the strains and the markers do not predict phenotype beyond the phylogeny. We corrected for population substructure problem using the method of Zhao et al. (2007). The resulting marker–gene associations were converted to gene–gene associations by assigning genes to their nearest marker (within 10 kb) on the genome. Finally, all genes assigned to the same marker were defined to belong to the same locus.

**High-confidence physical interaction network**

Protein–protein interactions were obtained from a modified form of the STRING database (Search Tool for the Retrieval of Interacting Proteins, version 6.3) (von Mering et al., 2005), extended to incorporate additional information on potential interactions. STRING reports a confidence score for each protein interaction based on numerous experimental and computational evidences. We implemented a naive Bayes classifier that takes the STRING score as one line of evidence. As a second line of evidence, we incorporated quantitative genetic interactions from Collins et al. (2006) who analyzed double mutants to detect both aggravating and alleviating genetic interactions. Genetic interactions may also be used as indirect predictors of physical protein interactions (Kelley and Ideker, 2005; Ye et al., 2005). As a third and final line of evidence, we used recently published protein interaction data (Gavin et al., 2002; Krogan et al., 2006) that were not included in the 6.3 version of STRING. For fitting the parameters of the model, a positive training set of 11 814 distinct interactions was created from pairs of proteins falling within known pathways recorded in the Kyoto Encyclopedia of Genes and Genomes (Kanehisa, 2002) as well as from small-scale binary physical interactions and protein complexes from the Munich Information center for Protein Sequences (Mewes et al., 2006). The negative training set of 35 676 interactions was obtained by randomly pairing proteins.

**Pooling with transcriptional interactions**

A pooled molecular interaction network was constructed by merging the above PPIs with TF–DNA interactions obtained from Beyer et al. (2006). This study combined several lines of evidence in a Bayesian framework to assign log-likelihood scores to each TF–DNA link. The 11 513 TF–DNA interactions with log-likelihood scores > 3.0 were included in the final set.
To ensure that all interactions in the network (PPI and TF–DNA) represented physical binding events (as opposed to functional linkages), we required that each included interaction has been reported in at least one experiment, indicating direct physical interaction between the proteins. In addition, to enrich for interactions within regulatory pathways, the network was restricted to regulatory proteins. We included proteins that were assigned to the following MIPS categories: (1) regulation of glycolysis and gluconeogenesis, (2) regulation of electron transport and membrane-associated energy conservation, (3) regulation of respiration, (4) regulation of energy conversion/regeneration, (5) regulation of DNA processing, (6) mitotic cell cycle and cell cycle control, (7) transcriptional control, (8) regulation of splicing, (9) translational control, (10) protein fate (folding, modification, destination), (11) regulation of metabolism and protein function, (12) cellular communication and signal transduction mechanism, (13) cell rescue, defense and virulence, (14) cellular sensing and cell cycle and cell cycle control, (7) transcriptional control, (8) regulation (2) regulation of electron transport and membrane-associated energy conversion.

MIPS categories: (1) regulation of glycolysis and gluconeogenesis, (2) regulation of electron transport and membrane-associated energy conservation, (3) regulation of respiration, (4) regulation of energy conversion/regeneration, (5) regulation of DNA processing, (6) mitotic cell cycle and cell cycle control, (7) transcriptional control, (8) regulation of splicing, (9) translational control, (10) protein fate (folding, modification, destination), (11) regulation of metabolism and protein function, (12) cellular communication and signal transduction mechanism, (13) cell rescue, defense and virulence, (14) cellular sensing and cell cycle and cell cycle control, (7) transcriptional control, (8) regulation

where u and v are any nodes in the network, and t is the target gene, and V is the voltage on the nodes of the electric circuit. Here, equations (1) and (2) are derived from Ohm’s Law which states that the current flowing through any two points is directly proportional to the voltage difference and the conductance between them. Further, equation (3) corresponds to Kirchoff’s current law in electric circuit theory which states that the total sum of current through any point in the circuit is zero (Irwin and Wu, 1999). The wires of a simple resistive circuit (as shown in Figure 1C) do not have explicit directionality, such that current can flow in either direction. However, the molecular network used in this study includes TF–DNA interactions that, by definition, transmit signal from the TF to the DNA and not vice versa. Electrical circuits account for directed links by using diodes, which constrain current to flow in one direction only. Equations (4) and (5) are constraints to ensure that the current only flows in the correct direction on known directed edges. For instance, let (u, v) be a directed edge with the signal going from u to v: If V(u) > V(v), then to minimize the objective function, d(u, v) will take the value (V(u)−V(v)). As a result, the equation becomes the same as (1). However, if V(u) < V(v), then due to (5), d(u, v) will be equal to 0, implying that there is no current on that edge. We implemented the above linear programming approach in Matlab (http://www.mathworks.com/) using the MOSEK package Version 5 (http://www.mosek.com/).

Supplementary information
Supplementary information is available at the Molecular Systems Biology website (www.nature.com/msb).

Acknowledgements
We are grateful to Hyun Kang and Dr Eleazar Eskin for the population substructure correction of the data set used in this study. Finally, we thank Kai Tan and Sourav Bandyopadhyay for critical reading of the manuscript. This research was supported by the National Center for Research Resources (RR018627, SS, TI), the National Institute of General Medical Sciences (GM70743-01, TI), a David and Lucille Packard Fellowship award (TI) and the Klaus Tschira Foundation (AB).
Electric circuit analysis of eQTL data
S Suthram et al

Perez-Enciso M, Quevedo JR, Bahamonde A (2007) Genetical

Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27: 39–54

Perez-Enciso M, Quevedo JR, Bahamonde A (2007) Genetical genomics: use all data. BMC Genomics 8: 69

Molecular Systems Biology is an open-access journal published by European Molecular Biology Organization and Nature Publishing Group. This article is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 Licence.