BIOINFORMATICS

Vol. 25 ISMB 2009, pages i15–i20
doi:10.1093/bioinformatics/btp189

Graph theoretical approach to study eQTL: a case study of *Plasmodium falciparum*

Yang Huang[1], Stefan Wuchty[2,†], Michael T. Ferdig[3] and Teresa M. Przytycka[1,*]

[1]National Center for Biotechnology Information, NLM, NIH, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894, [2]Northwestern Institute on Complexity, Northwestern University, 600 Foster Street, Evanston, IL 60201 and [3]Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, 107 Galvin Life Sciences, Notre Dame, IN 46556, USA

ABSTRACT

Motivation: Analysis of expression quantitative trait loci (eQTL) significantly contributes to the determination of gene regulation programs. However, the discovery and analysis of associations of gene expression levels and their underlying sequence polymorphisms continue to pose many challenges. Methods are limited in their ability to illuminate the full structure of the eQTL data. Most rely on an exhaustive, genome scale search that considers all possible locus–gene pairs and tests the linkage between each locus and gene.

Result: To analyze eQTLs in a more comprehensive and efficient way, we developed the Graph based eQTL Decomposition method (GeD) that allows us to model genotype and expression data using an eQTL association graph. Through graph-based heuristics, GeD identifies dense subgraphs in the eQTL association graph. By identifying eQTL association cliques that expose the hidden structure of genotype and expression data, GeD effectively filters out most locus–gene pairs that are unlikely to have significant linkage. We apply GeD on eQTL data from *Plasmodium falciparum*, the human malaria parasite, and show that GeD reveals the structure of the relationship between all loci and all genes on a whole genome level. Furthermore, GeD allows us to uncover additional eQTLs with lower FDR, providing an important complement to traditional eQTL analysis methods.

Contact: przytyck@ncbi.nlm.nih.gov

1 INTRODUCTION

The development of methods that allow us to uncover mechanisms of gene regulation and reconstruct gene regulatory networks is an important open problem in molecular biology. The advancement of high-throughput genotyping and gene expression platforms supports the analysis of expression quantitative trait loci (eQTL) as a tool to elucidate gene regulation. eQTL analysis considers expression of each gene as a quantitative trait and maps it to a genomic locus or marker. The genotype associated with a gene's expression level highlights the genome region carrying the DNA polymorphism impacting the expression. The polymorphism may reside in the gene's coding region or in a transcription factor binding site and could affect the expression level of its own or other genes in an inheritable way (Brem and Kruglyak, 2005; Monks *et al*., 2004; Petretto *et al*., 2006). Hence, a significant statistical linkage between a locus and a gene's expression suggests that the gene in question is regulated by the locus, which may hold a regulatory element or a regulator gene. Since the early work of Jansen and Nap (Jansen and Nap, 2001), eQTL has become a widespread technique to identify such regulatory associations and has been applied to several species including yeast (Brem and Kruglyak, 2005; Yvert *et al*., 2003), mouse (Bystrykh *et al*., 2005; Chesler *et al*., 2005) and human (Cheung *et al*., 2005; Stranger *et al*., 2005). Typically, these studies use genome-wide association studies (GWAS), considering loci spanning the genome and expression profiles of all genes in the organism. As a major advantage, simultaneous monitoring of thousands of gene expression traits provides unique and unbiased data and opens the possibility of constructing a global view of the underlying regulation machinery.

Despite the valuable insights that can be gained, current attempts to elucidate the structure of eQTLs still face many challenges. Only a few methods are available that model complete eQTL data to discover broader eQTL structure. The complex dependence of the variations of gene expression regulation on phenotypic differences nurtures the expectation that important information can be gained from considering more subtle relationships between genotype and expression. The large number of gene expression traits and genomic loci poses challenges for both computational efficiency and statistical power. Traditionally, an eQTL study tests the linkage between all genes' expression and all loci, adding up to millions of single statistical tests. For example, (Stranger *et al*., 2007) used 2 million single-nucleotide polymorphism (SNP) and more than 13 000 transcripts, leading to more than 10^{10} tests for all possible associations, a number that causes a serious multiple testing issue (i.e. the chance of false positives in a family of multiple hypothesis tests is higher than that of a single test). Consequently, that study was restricted to consider mainly *cis*-regulation—associations between SNPs and genes within 1 M bp of the SNP in question—which reduced the number of tests dramatically. While more complex regulation programs are of increasing interest (Storey *et al*., 2005), the combinatorial nature of such problems and the large number of loci call for improved methods that allow discovery of more complex regulation programs involving more than one locus and one gene.

To address such problems, we propose a novel method, GeD (Graph based eQTL Decomposition), to analyze eQTL data. Our method models the genotype, progeny and expression data as an eQTL association graph, a three–partite graph which is the union of two bipartite graphs. By simultaneously exploring two bipartite graphs, GeD discovers sets of dense subgraphs, called

*To whom correspondence should be addressed.
†Present address: NOB, NCI, NIH, 37 Convent Drive 1142E, Bethesda, MD 20892, USA.

eQTL association cliques, each containing a set of loci, a set of progenies and a set of genes. The progenies provide evidence that the set of loci may be associated with the set of genes. Such eQTL association cliques give a succinct representation of structures among loci, progenies and genes on a genome-wide scale. More importantly, each locus, progeny and gene can appear in more than one association clique, which depicts a complete picture of eQTL data. To find eQTL association cliques, GeD employs an efficient bipartite clique enumeration algorithm initially designed for building a concept lattice (Farach-Colton and Huang, 2008). The set of association cliques helps to select a small set of locus–gene pairs that are expected to have significant linkage; subsequently, statistical tests, including corrections for multiple testing, are performed for theses selected locus–gene pairs.

Testing GeD, we reanalyzed data from a recent eQTL study of the human malaria parasite *Plasmodium falciparum* (Gonzales *et al.*, 2008). While understanding regulatory programs of this parasite is of fundamental importance, successes in identifying specific transcription factors in *P.falciparum* have been limited. Gene expression of various *P.falciparum* strains does not vary significantly in response to perturbation (Rockman and Kruglyak, 2006); however, ubiquitous heritable expression patterns likely exist, although the association between loci and gene expression might be weak. Due to the difficulty of breeding and growing *P.falciparum* strains, only 34 progeny strains were used, a small number of strains that aggravates the detection of eQTL associations and increases the need for the more discerning methodology outlined here. Despite these challenges, Gonzales *et al.* successfully identified 1063 eQTLs with a FDR $\leq$0.24 in a genome-wide association study and showed several eQTL hotspots (Gonzales *et al.*, 2008).

Using GeD, we find that the size of eQTL association cliques is significantly different from random association cliques, and loci on different chromosomes tend to co-occur in some eQTL association cliques. In addition, by using eQTL association cliques, we detected 1327 eQTLs in *P.falciparum* with a FDR less than 0.05 without testing all possible locus–gene pairs, and new eQTL hotspots identified by GeD show several interesting biological characteristics.

2 MATERIALS AND METHODS

First we introduce the basic rationale of the GeD approach and present a detailed description of GeD. Finally, we describe the *P.falciparum* eQTL data we used to test our method.

2.1 Problem definition

In an eQTL experiment, we consider a set of progeny strains, often obtained from a cross between two parental strains with different genetic and phenotypic background. In our case, we only consider two possible genotypes (0, 1) for each locus and assign each locus l_j, to the parental strain the locus was inherited from. All strains can be partitioned into two groups by the genotype of a given locus, and we discretize the expression levels of each gene as being either 'up-regulated', 'unchanged' or 'down-regulated' (Fig. 1a).

To represent the above relationship between loci and genes, we define an eQTL association graph $\Gamma(G \cup S \cup L, E)$ as follows: The graph contains three sets of vertices (G, S, L), where L represents the genotypes of loci, S represents progeny strains, and G represents up- or down-regulated gene expression. Vertices g_{iu} and g_{id} indicate a gene g_i's up- or down-regulation and l_{j0} and l_{j1} represent the genotype at locus l_j as either 0 or 1. An edge

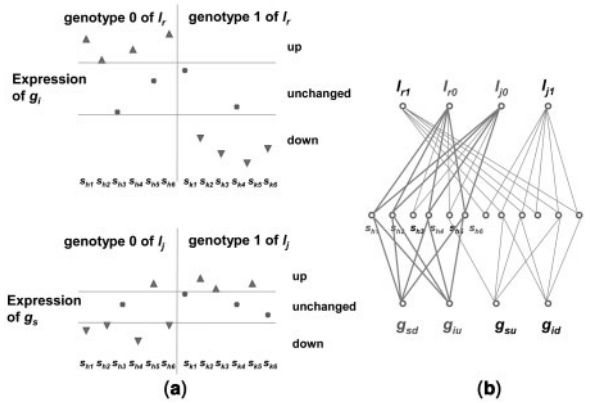

Fig. 1. (**a**) Each genomic locus l is assigned the genotype of the strain it was inherited from (0/1). In all strains s, we discretize expression levels of genes g as being either 'up', 'unchanged' or 'down'. Considering the genotypes of l, we observe different gene expression patterns, indicating different expression mechanisms in different strains. For example, the expression of g_i is up-regulated in strain s_{h1}, s_{h2}, s_{h4} and s_{h6}, and down-regulated in strain s_{k2}, s_{k3}, s_{k5} and s_{k6}. In (**b**) we show the corresponding eQTL association graph. Specifically, we find an association clique including genes g_{sd} and g_{iu}, strains s_{h1}, s_{h2}, s_{h4}, s_{h6} and loci l_{r0} and l_{j0}, shown in light grey. The edges of the association clique are drawn with wider lines.

between g_{iu}/g_{id} and a progeny strain s_k indicates g_i's expression is up- or down-regulated in strain s_k. An edge between l_{j0}/l_{j1} and a progeny strain s_k indicates the genotype of l_j is 0/1 in strain s_k. Note, that there are no edges between genes G and loci L. Our eQTL association graph can be viewed as a three-partite graph which is the union of two bipartite graphs $BG_1(L \cup S, E_1)$ and $BG_2(G \cup S, E_2)$.

The corresponding eQTL association graph in Figure 1a is shown in Figure 1b, where we consider the subgraph induced by g_{sd}, g_{iu}, s_{h1}, s_{h2}, s_{h4}, s_{h6}, l_{r0} and l_{j0}. We call such a subgraph an eQTL-association-clique, defined as $\Gamma_p = (G_p \cup S_p \cup L_p, E_p)$, $\forall g_{iu/d} \in G_p, \forall s_k \in S_p, (g_{iu/d}, s_k) \in E_p$ and $\forall l_{j0/1} \in L_p, \forall s_k \in S_p, (l_{j0/1}, s_k) \in E_p$. In other words, we require that G_p and S_p, and L_p and S_p are fully connected. Additionally, no such association clique $\Gamma_q = (G_q \cup S_q \cup L_q, E_q)$ exists, where G_q and S_q are fully connected, L_q and S_q are fully connected, and $G_q \cup S_q \cup L_q \supset G_p \cup S_p \cup L_p$. In other words, each eQTL association clique is a maximal subgraph that cannot be extended further, maintaining full connectivity. Similarly, an eQTL association clique can be viewed as the union of two dense bipartite subgraphs formed by $G_p \cup S_p$, and $L_p \cup S_p$, respectively. As defined, please note that in each eQTL association clique, $|G_p| \geq 1$ and $|L_p| \geq 1$. Furthermore, opposing loci l_{j0} and l_{j1}, or gene expression states g_{iu} and g_{id} can not appear in the same association clique.

It is easy to see that there can be four cases where a locus–gene pair (l_j, g_i) can appear in an association clique: The first case is that an up-regulated gene g_{iu} and a 0-genotyped locus l_{j0} are in an association clique while in the second case a down-regulated gene g_{id} and a 1-genotyped locus l_{j1} are in an association clique. We call these cases P_1. In a third case an up-regulated gene g_{iu} and a 1-genotyped locus l_{j1} are in an association clique while in the last case a down-regulated gene g_{id} and a 0-genotyped locus l_{j0} are in an association clique, cases we call P_2. We call the first two cases compatible since they both suggest that g_i's expression pattern is different in two groups defined by l_j's genotype-and vice versa.

Intuitively, a locus and a gene that co-appear in an association clique that has a large subset of strains are expected to be more closely associated. Therefore, we define the size of the progeny strain set in a subgraph of the association graph as support, *sp*. For a locus–gene pair (l_j, g_i) and an association clique with support *sp*, if g_{iu} and l_{j0} co-appear in the clique, we define the support provided by the clique sp_{ij}^{u0}. Similarly, we define

sp_{ij}^{d1}, sp_{ij}^{u1} and sp_{ij}^{d0}. Using these definitions, the support for pattern P_1 for (l_j, g_i) is $sp_{ij}^{P_1} = \max(sp_{ij}^{u0}) + \max(sp_{ij}^{d1})$, over all eQTL association cliques. Analogously, the support for P_2 is $sp_{ij}^{P_2} = \max(sp_{ij}^{u1}) + \max(sp_{ij}^{d0})$. Since P_1 and P_2 are opposites if we consider the linkage between l_j and g_i, we define $sp_{ij} = |sp_{ij}^{P_1} - sp_{ij}^{P_2}|$ as a rough measurement of the net support for the expectation that significant linkage between l_j and g_i exists.

2.2 Method

Based on these important heuristics, GeD performs the following steps to identify eQTL association cliques and to detect eQTL:

(i) Discretize (see below) gene expression levels and build an eQTL association graph $\Gamma(G \cup S \cup L, E)$, a union of bipartite graphs $BG_1(L \cup S, E_1)$ and $BG_2(G \cup S, E_2)$.
(ii) Find all maximal bipartite cliques in $BG_2(G \cup S, E_2)$.
(iii) For each maximal bipartite clique $BC(G_a \cup S_a, E_a)$, find all maximal bipartite cliques $BC(L_{ai} \cup S_{ai}, E_{ai})$ in the bipartite graph induced by S_a in $BG_1(L \cup S, E_1)$.
(iv) Identify sets G_{ai} where each vertex is connected to each vertex in S_{ai} appearing in $BG_2(G \cup S, E_2)$. If the subgraph $\Gamma(G_{ai} \cup S_{ai} \cup L_{ai}, E_{ai})$ has not been generated yet, output this graph as an eQTL association clique.
(v) For each locus–gene pair (l_j, g_i) appearing in one eQTL association clique, select the pair if its support value $\max(sp_{ij}^{P_1}, sp_{ij}^{P_2})$ and sp_{ij} meet criteria described below.
(vi) Among selected locus–gene pairs compute p-values of their association (adjusted for multiple testing).

In both steps (ii) and (iii), it is essential to enumerate bipartite cliques from a large bipartite graph efficiently. We apply an algorithm for building a concept lattice, which can be considered a hierarchical structure for organizing all bipartite cliques given a bipartite graph. Such structures have been used to compare gene expression matrices (Huang and Farach-Colton, 2007). The delay-time complexity—the time spent to compute each bipartite clique—of the algorithm is $O(n_1 n_2)$, where n_1 and n_2 are the size of two sets of vertices in the bipartite graph.

Here, we assume that the number of bipartite cliques in $BG_2(G \cup S, E_2)$ is lower than in $BG_1(G \cup S, E_1)$. If this is not the case, GeD starts from $BG_1(L \cup S, E_1)$ in step (ii); steps (iii) and (iv) are changed accordingly.

To obtain corrected p-value in the last step of GeD, we apply the method of Churchill and Doerge (Churchill and Doerge, 1994). For each gene g_i in a selected locus–gene pair from step (v), we maintain a locus list $(l_{j1}, l_{j2}, \ldots, l_{jd})$, where each locus in the list appears with g_i in one of selected locus–gene pairs. We randomly permute g_i's expression and compute the nominal p-value for the linkage between the random expression and a locus in the list and retain the smallest p-value. After repeating the process 1000 times, we use all retained p-values to approximate a null distribution. By comparing the nominal p-value from real data to the null distribution, we obtain the corrected p-value.

While numerous ways to discretize gene expression data (Becquet *et al*., 2002) transcription patterns of most genes in several major *P.falciparum* strains are very similar (Llinas *et al*., 2006). Therefore, we used a simple method (Quackenbush, 2002) that can be readily applied to our case. We computed the mean $\bar{m}$ and standard deviation *stdev* for each probe and define genes with expression levels $> \bar{m} + b^* stdev$ as 'up-regulated' and $< \bar{m} - b^* stdev$ as 'down-regulated'. Specifically, we set b to 1, allowing us to detect more variation in the gene expression. Another advantage of $b = 1$ is that each probe will be represented by at least one vertex in the association graph. In the worst case, the number of bipartite cliques in a bipartite graph is $\min(2^{n1}, 2^{n2}) - 2$, where n_1 and n_2 are the sizes of the two vertex sets of the bipartite graphs. Since thousands of vertices in G and L in the eQTL association graph exert extreme computational costs we only allow bipartite cliques with at least five progeny strains in step (ii) and (iii).

2.3 Materials

Utilizing *P.falciparum* eQTL data from the reference (Gonzales *et al*., 2008), we used 34 progeny strains obtained from a HB3xDd2 cross. Each progeny was genotyped at 329 microsatellite markers along 14 chromosomes. Expression levels were measured 18 h after the parasite invades human erythrocytes (RBCs), by 7665 probes representing 5150 ORFs.

3 RESULTS

As previously mentioned, eQTL association cliques allow us to determine the structure inherent in eQTL data. We first show the difference between association cliques obtained from the underlying eQTL data as well as from randomized data. Subsequently, we report eQTLs we determine in eQTL association cliques.

3.1 Size Distribution of eQTL Association Cliques

Applying GeD, we obtain 135 044 eQTL association cliques with support $sp \geq 5$. Overall, the support in eQTL association cliques ranges from 5 to 10. To generate random eQTL association cliques, we permuted the expression vector of each probe and applied GeD with the same parameters on the random data 100 times. We find 40 773, 20 393 and 5396 association cliques with support of 5, 6 and 7 in the real data. In random data, we find on average more association cliques ($sp = 5$: 77 200; $sp = 6$: 28 019; $sp = 7$: 5809). Applying a one-sample t-test between the number of association cliques in real and random data yielded $p < 10^{-11}$ in all three cases. Considering association cliques with sp 8, 9 and 10, we find 872, 84 and 5 in the real data. Compared to the random data, we analogously find significant differences. Specifically, we find on average 807, 74 and 4 random association cliques with the same supports in the randomized data with $p < 10^{-11}$ in the cases of support 8 and 9, and $p < 10^{-10}$ in the case of support 10.

Subsequently, we compared the number of association cliques in real and random data that have the same support sp and $|G|$, the number of probes. The number of random association cliques was significantly smaller except when $sp = 5$ or 6 or 7, $|G| = 1$, and $sp = 5$, $|G| = 2$. In Figure 2a we show the number of real association cliques and the average number of random association cliques with $sp = 6$ and several different number of probes $|G|$. Specifically, the largest eQTL association clique with $sp = 6$ has 87 probes.

A closer look revealed that, given support sp, $|G|$ and $|L|$, the number of association cliques in the random data was significantly larger than in the real data only when $|G|$ is small. For example, when $sp = 6$ and $|L| = 7$, the number of association cliques in the random data was larger only if $|G| = 1$. For a given a support value sp and the number of loci $|L|$, the number of association cliques in the random data was significantly larger than in the real data for most cases when $sp = 5$ or 6.We find similar results when $sp = 7$, $|L| < 12$, $sp = 8$, $|L| < 7$, and $sp = 9$, $|L| < 5$. Specifically, we show the number of random association cliques with $sp = 6$ and different numbers of loci $|L|$ in Figure 2b.

Since genotypes of adjacent loci are more similar than others, we expect that many loci in the same eQTL association clique are adjacent. Though this is frequently the case, we also find many co-appearing loci although they are on different chromosomes. For example, loci 2_0 on chromosome 2 and 12_45.8 on chromosome 12 co-appear with six loci on chromosome 3 in an eQTL association clique with support 10. We did not find such a result in the random data, suggesting that these loci tend to co-segregate and indicating

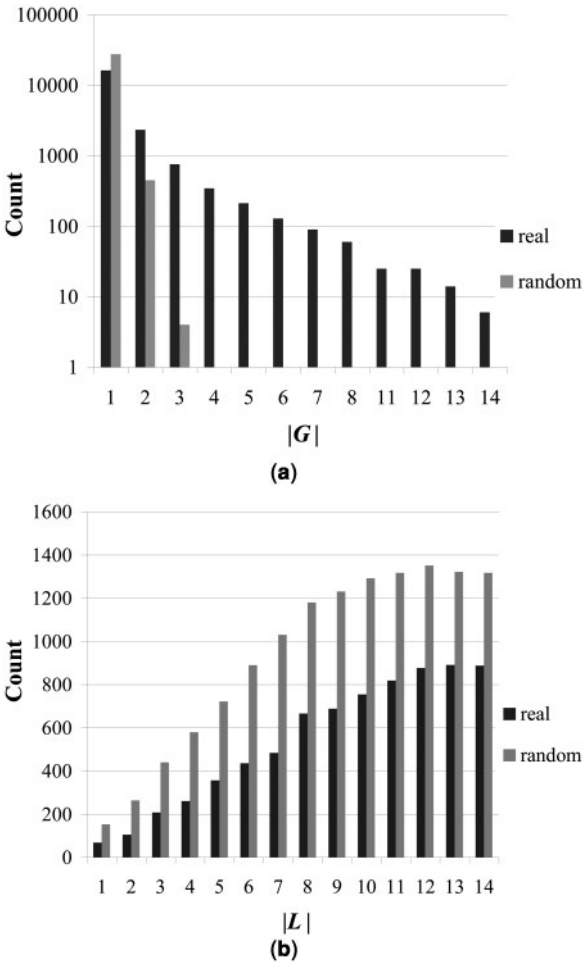

Fig. 2. Number of association cliques from real and random data with support 6. Numbers we obtained from randomized data were averaged over 100 runs. In (**a**) we perform the analysis varying $|G|$, the number of probes, while we show the analogous results with changing number of loci |L| in (**b**).

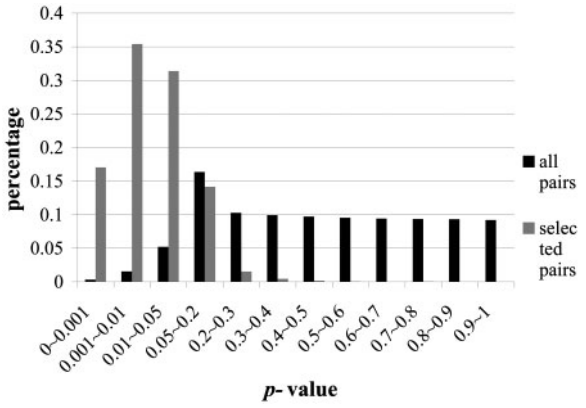

Fig. 3. Histogram of nominal P-values for all possible locus–gene pairs and pairs we selected from eQTL association cliques.

that a closer examination of their relation might be interesting with linkage disequilibrium based methods for *P.falciparum* (Su and Wootton, 2004).

3.2 eQTL detection

We have shown that a locus–gene pair appearing in two eQTL association cliques in a compatible way is more likely to have a significant linkage than those pairs that do not. Hence, we could use eQTL association cliques to select a small number of locus–gene pairs to be tested for linkage, many of which we expect to yield significant p-value. To this end, we used as criteria $\max(sp_{ij}^{P_1}, sp_{ij}^{P_2}) \geq 12$ and $sp_{ij} \geq 6$ in step (v) to select locus–gene pairs (l_j, g_i). Please note that each association clique has at least five progeny strains because we generate maximal bipartite cliques with at least five progeny strains. If we assume a locus–gene pair (l_j, g_i) with pattern P_1, then the minimum value for $sp_{ij}^{P_1}$ is 10 since $sp_{ij}^{P_1} = \max(sp_{ij}^{d1}) + \max(sp_{ij}^{u0})$, where $sp_{ij}^{u0} \geq 5$ and $sp_{ij}^{d1} \geq 5$. Note, that if we set this threshold too high, we potentially remove locus–gene pairs having significant linkage. In total, we selected 6232 locus–gene pairs. Figure 3 shows the histogram of nominal p-value computed by a two-sided T-test for the linkage of these selected pairs and all possible locus–gene pairs.

We observe that selected locus–gene pairs from association cliques yield significantly lower linkage p-value. Correcting p-values (see Methods section) and calculating FDRs (Storey and Tibshirani, 2003) we identified 2853 eQTLs ($p < 0.05$, FDR < 0.04). Identifying the most significant eQTLs, we used our set of association cliques we found in randomized data. With the same criteria, we obtained a list of locus–gene pairs from each set of randomized association cliques and applied a T-test to obtain p-values for these pairs. In this way, we obtained 100 groups of p-values from random data, allowing us to estimate an empirical null distribution, which is often more stringent than the null distribution obtained individually for each gene (Churchill and Doerge 1994). We required that each reported eQTL has a nominal p-value smaller than 90% of p-values in the empirical null distribution. Following this protocol, we found 1327 eQTLs for 513 probes (482 genes) and 231 loci. Previously, Gonzales *et al.* (Gonzales *et al.*, 2008) identified a set of 1063 eQTLs with FDR < 0.24 using standard GWAS. In Figure 4, we show the distributions of eQTLs identified by Gonzales *et al.* and 1327 eQTLs we obtained with GeD. We observe that the distribution of eQTLs detected by GeD is similar to the distribution of previously identified eQTLs, which were obtained by considering all possible locus–gene pairs. We also find 251 (∼25%) eQTLs that appear in both sets. Although the overlap is considerable the two sets are quite different, an observation that can be attributed to fundamental differences in the methods (see Discussion section). Both analyses show that there are several eQTL hotspots on chromosome 3, 5 and 7. Gonzales *et al.*, (Gonzales *et al.*, 2008) called a locus eQTL hotspot if there existed at least 14 linked probes at a particular locus. Analogously, we found 17 eQTL hotspots and discovered two/three new eQTL hotspots in the right/left subtelomeric region on chromosome 3. While two weak eQTL hotspots on chromosome 9 and 12 detected by Gonzales *et al.* did not appear in our result, we detected two new eQTL hotspots on chromosome 5 and 7. Note that the definition of a hotspots used in both studies does not differentiate between *cis*- and *trans*- links, and

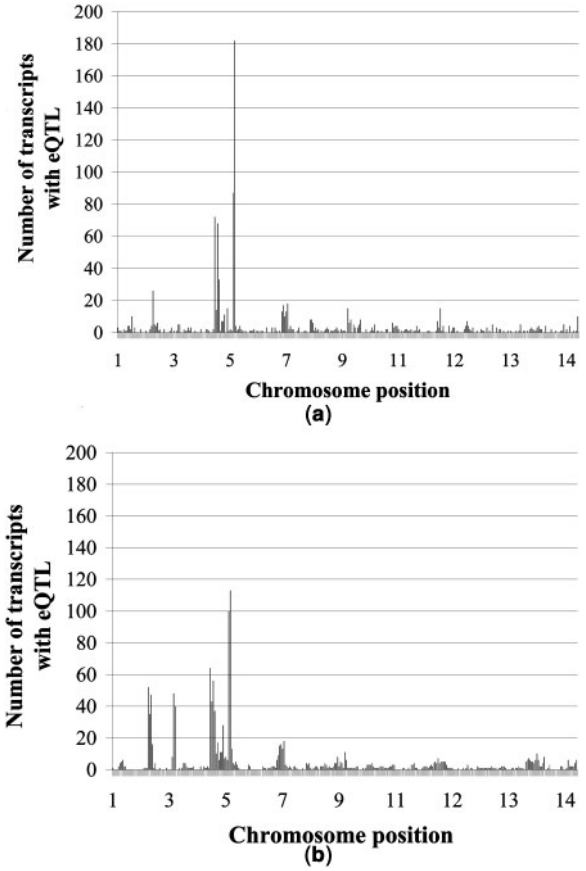

Fig. 4. In (**a**), we show genome-wide eQTL distributions in *P.falciparum* by testing all possible combination of loci and genes (Gonzales *et al*., 2008). In (**b**), we find similar, yet enriched patterns of eQTLs we detected with GeD.

the reported hotspots represent the combined effect of both types of regulation as well as that of the pattern of linkage disequilibrium.

Both subtelomeric regions on chromosome 3 are enriched with highly polymorphic surface antigen genes such as *cytoadherence linked asexual genes* (*CLAG*), *stevor* genes, and *var* genes (Gardner *et al*., 2002). While compelling, it remains to be experimentally determined if such polymorphic antigen genes are indeed regulated by eQTL hotspots we identified in the same region. Interestingly, it has been reported that the right telomere of chromosome 3 has an extended region of similarity with the right telomere of chromosome 2, and some pseudogene sequences in the regions were also preserved (Bowman *et al*., 1999). Such preservation in these rapidly evolving regions may imply that these subtelomeric regions are biologically significant (Bowman, *et al*., 1999), suggesting that the detection of additional eQTL hotspots in these regions provided more evidence for their importance in regulating the host-parasite interface. We also performed Gene Ontology term enrichment analysis for the target genes of newly detected eQTL hotspots using GOTermFinder (Boyle *et al*., 2004). We found that two hotspots show enriched GO terms referring to drug interaction and parasite-human invasion. The GO annotation of target genes of eQTL at locus 5_25.8 on chromosome 5 was enriched for drug binding ($p < 0.001$) and cis-trans isomerase activity ($p < 0.002$). The GO annotation of target genes of eQTLs at locus 3_14.3 on chromosome 3 was enriched for cytoadherence to microvasculature mediated by parasite protein and interaction with the host ($p < 0.03$).

4 DISCUSSION

We introduced a novel method—GeD—that integrates genotype, expression and progeny data, providing an analytical framework for the determination of gene regulation programs. In an eQTL association clique, vertices representing a locus' genotype are fully connected with vertices that represent progeny strains. Such a structure refers to the case that loci have the same genotype when restricted to these progeny strains. Analogously, vertices that represent genes are fully connected with vertices representing progeny strains, indicating that the corresponding progeny strains share the same gene expression patterns. As such, eQTL association cliques allow the determination of associations of loci, progeny strains and genes in a simple way. In addition, the number of progeny strains supports the linkage between loci and genes in the same association clique, which can help to detect eQTLs.

In this article we focused on the application of the eQTL association cliques to enhance eQTL discovery. However, eQTL association cliques have the potential to answer other questions as well. For example, loci that are not in linkage disequilibrium and co-occur in a highly supported clique might indicate functionally important co-segregation. Note that while loci that are in the same clique and are genomic neighbors are likely to be in linkage disequilibrium. However, the opposite case is not necessarily true. This observation should be useful in elucidating non-random properties of linkage disequilibrium. Additionally, eQTL association cliques may help the identification of loci and genes that are related in a certain phenotype. If the phenotype of progeny strains in an association clique is different from remaining progeny strains, the loci and genes in the corresponding association clique are the prime candidates that affect the phenotype in question.

Using eQTL association cliques might also help to uncover multiple locus linkage. For example, consider loci l_j and l_r and gene g_i, and four eQTL association cliques, where l_{j0} and l_{r0} appear with g_{iu} in one clique, l_{j0} and l_{r1} appear with g_{id} in another clique, l_{j1} and l_{r1} appear with g_{iu} in the third clique and l_{j1} and l_{r0} appear with g_{id} in the last clique. It is unlikely that l_j or l_r are associated with g_i individually because the genotype 0/1 of l_j is associated with both up- and down-regulated expression of g_i. The same rational holds for locus l_r. But since the joint genotype 00 and 11 of l_j and l_r is associated with up-regulation of g_i's expression, and joint genotype 01 and 10 of l_j and l_r is associated with down-regulation of g_i's expression, the two loci can have a significant epistatic interaction effect on g_i. By restricting our attention on loci in the same association clique, we can select a small set of triplets (l_j, l_r, g_i), which fit the above scenario, by simply counting association cliques. Testing the selected triples for epistatic effects reduces the number of statistical tests, $O(|L|^2|G|)$, required by an exhaustive search, where L is the locus set and G is the gene set.

In our method, we modeled underlying data using certain choices. First, discretizing expression data, a gene was considered differentially regulated if its expression level was at least one standard deviation away from its mean expression. This choice was dictated by its relative simplicity and applicability of that method to the data where differences in the expression levels are not expected

to be very large. Other methods of discretizing expression data will be considered in the future improvement of the method. Next, we chose to look at maximal cliques rather than other densely, yet not completely, connected subgraphs, allowing us to avoid the introduction of additional 'density' parameter. Furthermore, such an approach also allowed us to easily generate such clique-structures utilizing the efficient bipartite clique enumeration method (Farach-Colton and Huang, 2008). While bipartite cliques can potentially be replaced with bi-clusters, the best heuristic for the identification of such overlapping bi-clusters remains to be found. We conclude that our choices might potentially influence our ability to detect potential eQTLs. However, we made our choices as simple as possible and highlight the usability of our novel method.

We applied GeD to progeny data of *P.falciparum* and found that eQTL association cliques have very different structures and distributions compared to random association cliques. Using eQTL association cliques to select a small set of locus–gene pairs, we corroborated previously identified eQTLs, and significantly increased their number, including new eQTL hotspots. Preliminary analysis of the possible functional relevance of these new eQTL hotspots showed that some harbor important antigen genes while others include target genes involved in drug and parasite-host interactions. Compared to previous results, we conclude that GeD bolsters traditional eQTL analysis methods and provides new opportunities for the discovery of critical biological functions in *P.falciparum*. Approximately 25% of eQTLs in the two eQTL sets identified by GeD and Gonzales *et al.* (Gonzales *et al*., 2008) overlap, a difference that can be caused by several factors. First, Gonzales *et al.* applied an interval mapping method based on a complex Bayesian model for QTL detection (Sen and Churchill, 2001). Assuming each marker is the potential eQTL location, we in turn applied a two-sided T-test to determine linkage between markers and gene expression. To a certain extent, GeD may lose some information and consequently detection sensitivity due to the discretization of gene expression values and focus on relatively large eQTL association cliques. In contrast, the GWAS used by Gonzales *et al.* is likely to miss more subtle associations detected by our method because only the most significant eQTLs can pass multiple testing correction performed for all possible locus–gene pairs.

Our current implementation of GeD is designed for the analysis of the large data set of *P.falciparum*. However, the number of eQTL association cliques can increase exponentially with the number of loci and genes in the worst case. Therefore, the scalability of GeD to larger eQTL data sets containing thousands or even millions of loci remains to be tested. Specifically, in human studies where we have to deal with huge amount of expression and genomic data we expect strongly increasing computational costs, prompting the development of further heuristics and improved computational techniques that will allow us to tackle more challenging GWAS problems.

ACKNOWLEDGEMENTS

The authors thank John Wootton (NIH/NCBI) for stimulating discussions.

Funding: Intramural Research Program of the National Institutes of Health, National Library of Medicine; National Institutes of Health (AI071121 and AI055035 to M. T. F.).

Conflict of Interest: none declared.

REFERENCES

Becquet,C. *et al.* (2002) Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data. *Genome Biol.*, **3**, RESEARCH0067.

Bowman,S. *et al.* (1999) The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. *Nature*, **400**, 532–538.

Boyle,E.I. *et al.* (2004) GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

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

Bystrykh,L. *et al.* (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.*, **37**, 225–232.

Chesler,E.J. *et al.* (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.*, **37**, 233–242.

Cheung,V.G. *et al.* (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, **437**, 1365–1369.

Churchill,G.A. and Doerge,R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Farach-Colton,M. and Huang,Y. (2008) A linear delay algorithm for building concept lattices. *19th Symposium on Combinatorial Pattern Matching*. Springer-Verlag, pp. 204–216.

Gardner,M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature*, **419**, 498–511.

Gonzales,J.M. *et al.* (2008) Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol.*, **6**, e238.

Huang,Y. and Farach-Colton,M. (2007) Lattice based clustering of temporal gene-expression matrices. *7th SIAM International Conference on Data Mining*. SIAM, 398–409.

Jansen,R.C. and Nap,J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet.*, **17**, 388–391.

Llinas,M. *et al.* (2006) Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.*, **34**, 1166–1173.

Monks,S.A. *et al.* (2004) Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.*, **75**, 1094–1105.

Petretto,E. *et al.* (2006) Heritability and tissue specificity of expression quantitative trait loci. *PLoS Genet.*, **2**, e172.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.

Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.

Sen,S. and Churchill,G.A. (2001) A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.

Storey,J.D., *et al.* (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.*, **3**, e267.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Stranger,B.E. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.

Stranger,B.E. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.

Su,X.Z. and Wootton,J.C. (2004) Genetic mapping in the human malaria parasite *Plasmodium falciparum. Mol. Microbiol.*, **53**, 1573–1582.

Yvert,G. *et al.* (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, **35**, 57–64.