BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2290–2297
doi:10.1093/bioinformatics/btm327

Systems biology

Detection of eQTL modules mediated by activity levels of transcription factors

Wei Sun[1], Tianwei Yu[2] and Ker-Chau Li[1,3,*]

[1]Department of Statistics, University of California at Los Angeles, Los Angeles, California, [2]Department of Biostatistics, Emory University, Atlanta, Georgia, USA and [3]Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, ROC

Received on December 23, 2006; revised and accepted on June 17, 2007

Advance Access publication June 28, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Studies of gene expression quantitative trait loci (eQTL) in different organisms have shown the existence of eQTL hot spots: each being a small segment of DNA sequence that harbors the eQTL of a large number of genes. Two questions of great interest about eQTL hot spots arise: (1) which gene within the hot spot is responsible for the linkages, i.e. which gene is the quantitative trait gene (QTG)? (2) How does a QTG affect the expression levels of many genes linked to it? Answers to the first question can be offered by available biological evidence or by statistical methods. The second question is harder to address. One simple situation is that the QTG encodes a transcription factor (TF), which regulates the expression of genes linked to it. However, previous results have shown that TFs are not overrepresented in the eQTL hot spots. In this article, we consider the scenario that the propagation of genetic perturbation from a QTG to other linked genes is mediated by the TF activity. We develop a procedure to detect the eQTL modules (eQTL hot spots together with linked genes) that are compatible with this scenario.

Results: We first detect 27 eQTL modules from a yeast eQTL data, and estimate TF activity profiles using the method of Yu and Li (2005). Then likelihood ratio tests (LRTs) are conducted to find 760 relationships supporting the scenario of TF activity mediation: (DNA polymorphism → *cis*-linked gene → TF activity → downstream linked gene). They are organized into 4 eQTL modules: an amino acid synthesis module featuring a *cis*-linked gene LEU2 and the mediating TF Leu3; a pheromone response module featuring a *cis*-linked gene GPA1 and the mediating TF Ste12; an energy-source control module featuring two *cis*-linked genes, GSY2 and HAP1, and the mediating TF Hap1; a mitotic exit module featuring four *cis*-linked genes, AMN1, CSH1, DEM1 and TOS1, and the mediating TF complex Ace2/Swi5. Gene Ontology is utilized to reveal interesting functional groups of the downstream genes in each module.

Availability: Our methods are implemented in an R package: eqtl.TF, which includes source codes and relevant data. It can be freely downloaded at http://www.stat.ucla.edu/~sunwei/software.htm

Abbreviations: eQTL (expression Quantitative Trait Loci); TF (Transcription Factor); QTG (Quantitative Trait Gene); SNP (Single Nucleotide Polymorphism); FDR (False Discovery Rate); SGD (Saccharomyces Genome Database); LRT (Likelihood Ratio Test).

*To whom correspondence should be addressed.

Contact: kcli@stat.ucla.edu

Supplementary information: http://www.stat.ucla.edu/~sunwei/yeast_eQTL_TF/supplementary.pdf

1 INTRODUCTION

The eQTL studies have been applied in several model organisms and human recently (Brem *et al*., 2002; Chesler *et al*., 2005; Morley *et al*., 2004; Petretto *et al*., 2006; Schadt *et al*., 2003; Stranger *et al*, 2005; Wang *et al*., 2006). These works have shown that gene expression level is inheritable. A gene expression profile can be *cis*-linked to a local eQTL around the gene itself or *trans*-linked to a distant eQTL. Several genes' expression profiles can be linked to a small region, which is commonly referred to as an eQTL hot spot. A hot spot and the genes linked to it will be called an eQTL module hereafter. eQTL modules are the building blocks in constructing the gene expression linkage network.

While many eQTL modules can be detected statistically, (Bing and Hoeschele, 2005; Kulp and Jagalur, 2006) the molecular mechanism of associating the DNA polymorphism in the eQTL hot spot to the *cis*- or trans-linked genes is still poorly understood (Rockman and Kruglyak, 2006). One fundamental question concerns what roles the transcription factors (TFs) are playing. Yvert *et al*. (2003) showed that there is no TF enrichment in eQTL hot spots, thus eQTL module regulation must involve genes other than TFs. The following is a likely scenario, which we shall focus on. It highlights the importance of the TFs' activities.

(1) One or more *cis*-linked genes in module A are affected by the DNA polymorphism in the corresponding eQTL hot spot.

(2) Either these genes encode TFs; or their gene products interact with the activities of some TFs at the protein level.

(3) The affected TFs control the expression of their target genes.

(4) Therefore, some of the target genes are linked to module A.

Which modules may follow this scenario? We set out to identify such modules in a yeast eQTL data (Brem *et al*., 2005a, b) using

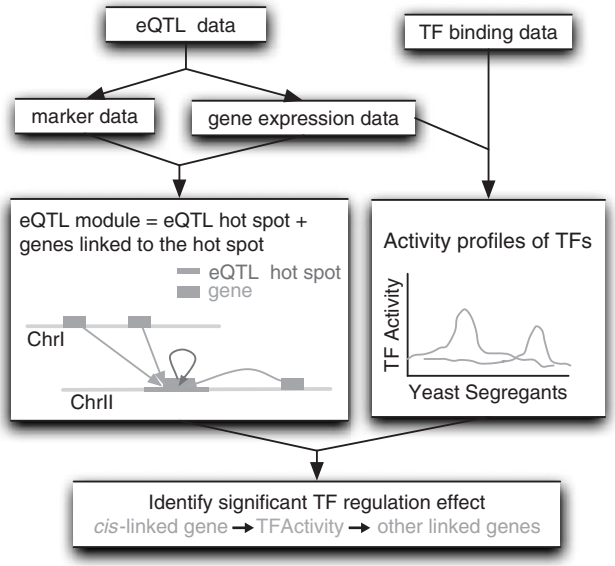

Fig. 1. A strategy of detecting the eQTL modules that are mediated by TF activities.

a computational approach sketched in Figure 1. In addition to the eQTL data, we also employ TF-binding data in order to estimate TF activity. As many authors have reported, transcription level of a TF may not be a good indicator of its protein activity (Rustici *et al*., 2005; Vleugel *et al*., 2004) for reasons, such as post-translation modification, protein translocation and so on. Although the TF activity profile is difficult to measure directly, it can be estimated from gene expression profiles in combination with genome-wide TF-binding data (Liao *et al*., 2003), or in combination with genome-wide TF-binding data and the known TF-TG (target gene) relationships from literature (Yu and Li, 2005). We estimate TF activity profile with the method of Yu and Li (2005). After that, we formulate the proposed TF activity mediation scenario into a statistical model and conduct likelihood ratio tests (LRTs) (Vuong, 1989) to compare this model with two other models representing the contrasting situations.

2 METHODS

2.1 Detect eQTL module

Before detecting eQTL modules, we need to identify significant and non-redundant linkages. Significant linkages are identified according to false discovery rate (FDR), which is calculated based on permutations (Brem *et al*., 2002). Because of linkage disequilibrium, profiles of neighboring markers tend to have high correlations. Thus if one gene is linked to one marker, it may also be linked to the adjacent markers. Such redundant linkages need to be eliminated before proceeding with further analysis. One simple solution for eQTL data from a cross of inbred experimental organism is to keep the most significant linkage for each gene expression trait in each chromosome and discard others (Wang *et al*., 2006). In the next step, we select the enriched markers, to which more genes are linked than expected by chance. Then, the eQTL hot spots can be detected using a marker merging procedure shown in Supplementary Figure 1a. The basic idea is to initiate a hot spot from an enriched marker M and to extend it iteratively. During the extension, we ask whether an adjacent marker M_k belongs to this hot spot or not. To answer this question, for each gene G that is linked to marker M_k, we carry out an LRT (Vuong, 1989) to see if the hypothesis H_0 holds or not where the statement of H_0 conveys the situation that, as a candidate marker for mapping the expression of gene G, M_k is as good as M statistically. If H_0 cannot be rejected for a majority of the genes linked to M_k, we merge M_k to this linkage hot spot. Details of LRT and the simulation (including power analysis) used to choose the decision boundary are described in Supplementary Materials. We also compare the eQTL hot spots identified from the yeast eQTL data by our approach and those identified by the method of Brem *et al*. (2002) and Schadt *et al*. (2003) using a fixed bin size for all the hot spots (Supplementary Table 3).

2.2 Estimate TF activity

Since the unobserved TF activity level in general differs from its expression level, we employ a two-stage constrained space factor analysis (Yu and Li, 2005) to estimate TF activity. Three sets of input data are required: a set of high confidence TF-TG (target gene) relationship data, a set of low confidence TF-TG relationship data and the set of gene expression data of main interest. In this article, we use the same TF-TG relationship data as Yu and Li (2005): the high confidence TF-TG relationship data comes from the literature and the low confidence data comes from the genome-wide TF-binding study (Harbison *et al*., 2004). The gene expression dataset comes from 112 yeast segregants (Brem *et al*., 2005a, b). If two or more TFs work together as a complex, e.g. Hap2/Hap3/Hap4/Hap5, we generate the activity profile for the complex instead of each individual TF.

2.3 Scenario modeling and identification

Given an eQTL module, use GC to denote the expression level of one *cis*-linked gene, M to denote the genotype of the marker where GC is *cis*-linked, GT to denote the expression level of any gene in the module other than GC and TA to denote a TF's activity. We consider three models that describe their relationships. The first one, causal model, represents the proposed scenario for eQTL-module regulation, while the other two are the competitive models with equal model complexity, namely having the same number of model parameters as the causal model:

- causal model: $(M \rightarrow GC) \rightarrow TA \rightarrow GT$
- reactive model: $(M \rightarrow GC) \rightarrow GT \rightarrow TA$
- conditional independence model: $GT \leftarrow (M \rightarrow GC) \rightarrow TA$

where $(M \rightarrow GC)$ represents the genetic perturbation. No arrow points to $(M \rightarrow GC)$ because protein activity or gene expression cannot change DNA sequence, and the expression of *cis*-linked genes are tightly controlled by the DNA variation (Schadt *et al*., 2005; Zhu *et al*., 2004). Here, each model corresponds to a conditionally independent restriction: GT is independent of $(M \rightarrow GC)$ given TA for causal model; TA is independent of $(M \rightarrow GC)$ given GT for reactive model; TA is independent of GT given $(M \rightarrow GC)$ for conditional independence model. Due to the different conditionally independent restrictions, likelihoods of different models have different decompositions. Thus, we can compare different models by their likelihoods. For the purpose of likelihood comparison, we can replace $M \rightarrow GC$ with GC since it appears in all the three models. Then the three models to be compared are:

- causal model: $GC \rightarrow TA \rightarrow GT$

Table 1. Likelihoods of the three models

Model	Likelihood
Causal	$L_c = \Pi_t\, l(TA_t \mid GC_t)\, l(GT_t \mid TA_t)$
Reactive	$L_r = \Pi_t\, l(GT_t \mid GC_t)\, l(TA_t \mid GT_t)$
Conditional independence	$L_i = \Pi_t\, l(TA_t \mid GC_t)\, l(GT_t \mid GC_t)$

One common term, $l(GC_t)$ is skipped in the three likelihoods.

- reactive model: $GC \rightarrow GT \rightarrow TA$
- conditional independence model: $GT \leftarrow GC \rightarrow TA$

An arrow in these models indicates only the direction of the relation. There could be potential intermediate players in these models. For example, there can be a hidden signal transduction gene G_1 lying between GC and TA : $GC \rightarrow G_1 \rightarrow TA \rightarrow GT$. We model the pairwise relation between GC, TA and GT by simple linear regression:

$$TA = \alpha_0 + \alpha_1 GC + \varepsilon_1,\ \theta_{TA|GC} = \{\alpha_0, \alpha_1, \sigma_1\}$$
$$GT = \beta_0 + \beta_1 GC + \varepsilon_2,\ \theta_{GT|GC} = \{\beta_0, \beta_1, \sigma_2\}$$
$$TA = \mu_0 + \mu_1 GT + \varepsilon_3,\ \theta_{TA|GT} = \{\mu_0, \mu_1, \sigma_3\}$$
$$GT = \nu_0 + \nu_1 TA + \varepsilon_4,\ \theta_{GT|TA} = \{\nu_0, \nu_1, \sigma_4\}$$

where $\varepsilon_k \sim N(0, \sigma_k^2)$, $(k = 1,2,3,4)$, and $\theta_{.|.}$ are unknown model parameters. The third and the fourth linear models describe the same bivariate distribution by exchanging the input and output variables. It is easy to write down the conditional densities for each linear model. For example,

$$l(TA_t \mid GC_t) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(TA_t - \alpha_0 - \alpha_1 GC_t)^2}{2\sigma_1^2}\right),$$

where t is sample index. From these conditional densities, the likelihood functions for the proposed TF regulation scenario and the two competitive models can be obtained easily, see Table 1.

The LRTs for non-nested models (Vuong, 1989) are conducted to identify the model that is significantly better than the other two models. If the result of LRT is insignificant, the corresponding triplet (GC, TF, GT) is labeled as an in-differential case.

3 RESULT

We apply our method to a yeast eQTL dataset of 112 yeast segregants generated from a cross of two inbred strains: BY and RM (Brem *et al.*, 2005a, b). The dataset includs expression profiles of 6229 gene expression traits and genotype profiles of 2956 SNP markers. We use student's t-test to quantify the significance of linkages. The cutoff P-value $4e^{-5}$ is chosen according to FDR 5%. Redundant linkages are removed as described in Method section. We end up with a total of 3162 linkages, which associate 2674 gene expression traits with 838 markers. Supplementary Figure 3 shows the number of genes linked to each marker.

The specific locations and sizes of eQTL hot spots are detected using the procedures described in Method section. Each linkage hot spot is initiated from an enriched maker, to which more than seven genes are linked. This cutoff value 7 is the 95% of binomial distribution with $n = 3162$ and $p = 1/838$. Altogether 27 eQTL hot spots/modules are obtained (Table 2). Table 2 also lists the *cis*-linked genes for each eQTL module. We define a linkage as *cis*-linkage if the corresponding marker is located within 10 kb distance of the gene, the same definition used by Brem *et al.* (2002). The lengths of the 27 eQTL hot spots vary from 1 base pair (only one marker in the eQTL module) to 73 kb with a median of ~20 kb (Supplementary Fig. 4). Figure 2 shows the eQTL modules detected in chromosome 2. The genes within each eQTL module can be found in the R package: eqtl.TF.

Table 2. eQTL modules

ID	Module symbol	N	*Cis*-linked genes
1	chr2_mod1	267	AMN1, ARA1, CNS1, CSH1, DEM1, TBS1, TOS1, YBR141C
2	chr2_mod2	67	ECM2, GIP1, NRG2, TAT1, TIP1, UBP14, YBR064W
3	chr2_mod3	19	N/A
4	chr3_mod1	92	FRM2
5	chr3_mod2	46	MATALPHA1, MATALPHA2, TAF2, YCR041W
6	chr3_mod3	75	LEU2
7	chr3_mod4	22	N/A
8	chr3_mod5	12	GLK1, RNQ1
9	chr4_mod1	29	N/A
10	chr4_mod2	25	YDR539W, YDR541C, YRF1-1
11	chr5_mod1	36	YER119C-A
12	chr5_mod2	36	GEA2, NPP2, URA3
13	chr5_mod3	12	UBP5, YER139C, YER140W
14	chr5_mod4	13	N/A
15	chr7_mod1	21	N/A
16	chr8_mod1	82	HSE1, YHL010C
17	chr8_mod2	40	GPA1, LAG1
18	chr12_mod1	134	GSY2, HAP1
19	chr12_mod2	47	YLR455W, YLR462W
20	chr12_mod3	10	NEJ1
21	chr13_mod1	36	N/A
22	chr13_mod2	30	MDM1
23	chr14_mod1	353	RHO2, TOP2, YNL089C, YPT53
24	chr14_mod2	132	AQR1, LAT1
25	chr14_mod3	8	POR1, YNL058C
26	chr15_mod1	277	ATG19, HAL9, PHM7, YOL087C
27	chr15_mod2	43	CAT5, YOR131C

N is the number of genes within the eQTL module.

After detecting the eQTL modules, we want to identify the potential TFs responsible for the module regulation. We first use the TF-binding data (Harbison *et al.*, 2004) to decide whether a TF binds to a gene or not (setting TF-binding P-value cutoff to 10^{-3}, the one used by Harbison *et al.*). Then for each eQTL module, we identify the TFs that bind more genes in the module than expected by chance. The degree of enrichment is quantified by hypergeometric distribution with P-value cutoff 0.01 (Table 3). We further quantify the relevance of a selected TF to an eQTL module by checking the correlation between the binding strength BS (−log(binding P-value)) and the linkage strength LS (−log(linkage P-value[1])) for all the genes of the eQTL

[1]Each gene only has only one linkage P-value describing the linkage strength into one hot spot since we have removed the redundant linkages.

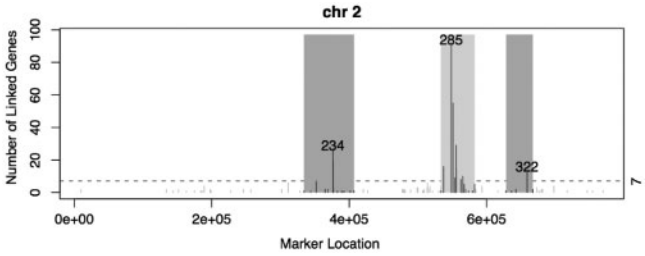

Fig. 2. eQTL modules in yeast chromosome 2. Each vertical line corresponds to a marker. Each colored rectangle corresponds to an eQTL hot spot. The number within each rectangle is the ID of the marker, from which the eQTL hot spot is initiated. eQTL hot spots are initiated from markers with more than seven linkages.

Table 3. TFs related with each eQTL module

Module symbol	TF	q	m	n	k	p	cor	p_cor
chr2_mod1	Swi5	15	120	6023	266	1.9E−04	0.31	2.7E−07
	Ace2	12	92	6051	266	5.6E−04	0.70	2.1E−40
chr3_mod2	Cha4	6	11	6132	43	3.7E−11	0.46	2.5E−03
	Zap1	6	22	6121	43	5.6E−09	0.43	4.2E−03
	Arr1	4	18	6125	43	5.9E−06	0.40	8.8E−03
	Hap4	6	75	6068	43	1.2E−05	0.44	3.3E−03
	Hir2	3	16	6127	43	1.7E−04	0.46	2.4E−03
	Aro80	3	27	6116	43	8.3E−04	0.45	2.9E−03
chr3_mod3	Leu3	7	24	6119	75	8.9E−09	0.69	5.2E−12
chr5_mod1	Phd1	3	65	6078	36	6.3E−03	0.52	1.1E−03
	Azf1	2	21	6122	36	6.5E−03	0.76	9.2E−08
	Cup9	2	21	6122	36	6.5E−03	0.46	5.3E−03
	Rox1	3	67	6076	36	6.8E−03	0.52	1.3E−03
	Nrg1	3	73	6070	36	8.7E−03	0.57	3.3E−04
chr8_mod2	Dig1	15	45	6098	40	0.0E+00	0.46	2.7E−03
	Ste12	15	57	6086	40	0.0E+00	0.50	1.1E−03
	Tec1	6	39	6104	40	1.4E−07	0.45	4.0E−03
chr12_mod1	Hap1	41	149	5994	134	0.0E+00	0.63	3.7E−16
chr12_mod2	Gat3	25	57	6086	46	0.0E+00	0.42	4.4E−03
	Rgm1	3	8	6135	46	2.1E−05	0.41	5.4E−03
chr13_mod1	Gcn4	5	75	6068	36	6.7E−05	0.69	3.7E−06

q is the number of genes within the eQTL module and bound by the given TF. m is the total number of genes bound by the TF. n is the number of genes not bound by the TF. k is the total number of genes with TF-binding data within the eQTL module. This number may be smaller than the total number of genes within the eQTL module in Table 2 due to missing of TF-binding data in some genes. P is the hypergeometric P-value based on q, n, m and k. cor is the correlation between TF's binding strengths (−log(P-value)) and linkage strengths (−log(P-value)) for the genes within the eQTL module and p_cor is the P-value of the correlation, which is assessed by linear regression.

module (Table 3). Specifically, we fit a linear model $\mathrm{LS} = \alpha + \beta \mathrm{BS} + \varepsilon(\varepsilon \sim N(0, \sigma_\varepsilon^2))$, and use the P-value of the regression coefficient β (at the 0.01 level) to trim out the weakly-related TFs. Table 3 shows the selected TF candidates for further analysis.

Taking the intersection of Tables 2 and 3, we find seven modules with both *cis*-linked genes and TF candidates: chr2_mod1, chr3_mod2, chr3_mod3, chr5_mod1, chr8_mod2, chr12_mod1 and chr12_mod2. The activity profiles of 97 TFs are estimated (including 10 complexes; see Method section).

Table 4. Detected causal relation modules

Module symbol	*Cis*-linked gene	TF	causal	react	con.ind	in-diff
chr2_mod1	AMN1	Ace2/Swi5	136	0	1	130
chr2_mod1	ARA1	Ace2/Swi5	27	0	0	240
chr2_mod1	CNS1	Ace2/Swi5	19	0	4	244
chr2_mod1	CSH1	Ace2/Swi5	131	0	0	136
chr2_mod1	DEM1	Ace2/Swi5	122	0	0	145
chr2_mod1	TBS1	Ace2/Swi5	58	0	0	209
chr2_mod1	TOS1	Ace2/Swi5	158	0	0	109
chr3_mod3	LEU2	Leu3	49	0	1	25
chr8_mod2	GPA1	Ste12	14	0	1	25
chr12_mod1	GSY2	Hap1	18	0	0	116
chr12_mod1	HAP1	Hap1	28	0	0	106

causal is the number of significant causal relations, react is the number of significant reactive relations, cond.ind is the number of significant conditional independence relations and in-diff is the number of in-differential relations.

The information of these TFs together with their (estimated) activity profiles can be found in the R package: eqtl.TF. For these seven modules, we test the significance of causal relation for each triplet $\{GC, TA, GT\}$, where GC is the expression profile of a *cis*-linked gene, TA is the activity profile of a TF and GT is the expression profile of any gene in the module other than GC. In total, we find 760 causal relationships that support the TF activity mediation scenario (Table 4). Here one model is called significantly better than the other two models if the LRT P-values are smaller than 0.01. The four modules harboring the detected high confidence causal relationships are discussed next. The complete result is listed in Supplementary Table 4. If we change the LRT P-value cutoff to 0.05, more causal relationships can be detected, but they are still within the same four modules (Supplementary Table 5).

Leu3 mediates causal relations in amino acids synthesis module, chr3_mod3. Among the 75 genes of this module, only LEU2 is *cis*-linked. LEU2 encodes beta-isopropylmalate dehydrogenase, the enzyme that catalyzes the third step in leucine biosynthesis (Andreadis *et al.*, 1984). According to Table 3, the enriched TF related to this module is Leu3, which regulates the transcription of genes involved in branched-chain amino acid synthesis (Friden and Schimmel, 1988). There are 49 significant causal relationships, one significant conditional independence relationship, and no significant reactive relationship in this module (Table 4).

We conduct a Gene Ontology (GO) (Ashburner *et al.*, 2000) analysis for the detected 49 downstream genes, using the GO term finder in SGD (Hong *et al.*, 2007). The most significantly enriched terms is 'amino acid biosynthesis' (10 out of 49 genes, P-value 3.3E−7)[2]. In contrast,

[2]In order to make the P-value comparable to the P-value from the 25 remaining genes, we choose 25 genes from the 49 genes to test for enrichment. Specifically, we order the 49 genes by LRT P-values comparing causal model against the other two models and choose the 25 genes with smallest P-values. The enriched GO term is still 'amino acid biosynthesis' (9 out of 25 genes, P-value 4.5e−9)

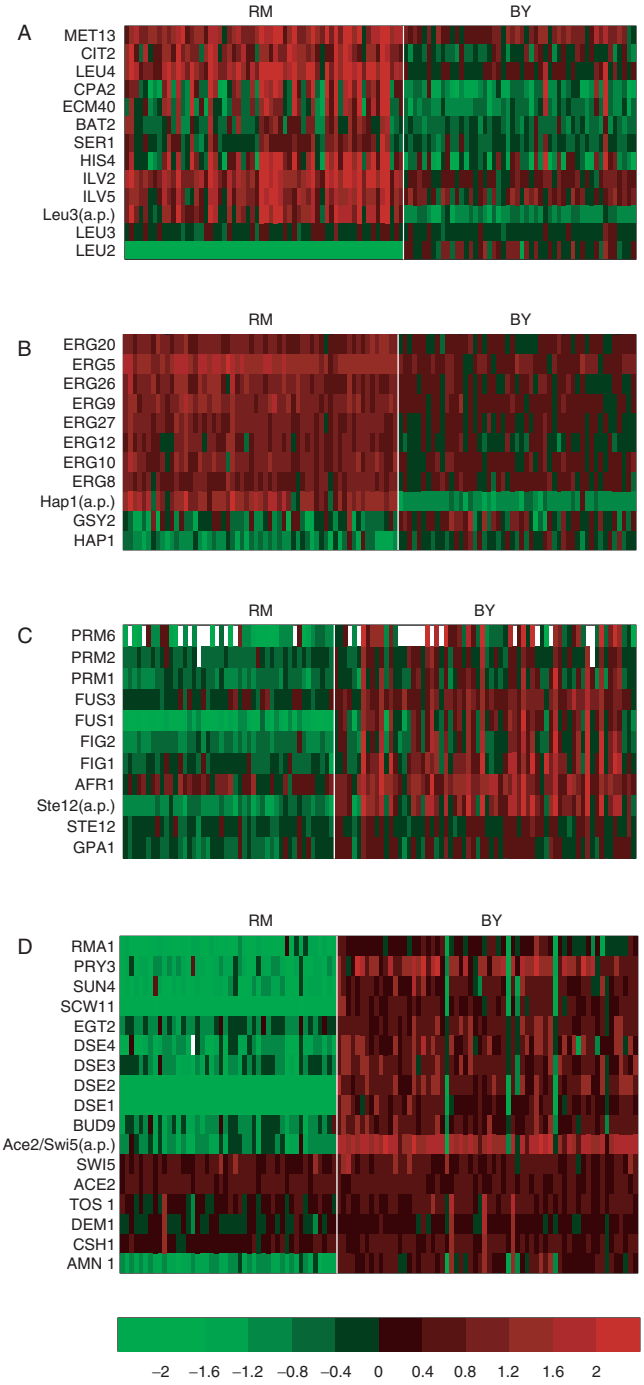

Fig. 3. **(a)** Expression profile of LEU2, LEU3 and 10 downstream genes that participate in amino acid biosynthesis, and the activity profile of Leu3 (Leu3(a.p.)), grouped by the genotype of LEU2. **(b)** Expression profile of HAP1, GSY2 and 8 downstream genes that participate in ergosterol biosynthesis, and activity profile of Hap1 (Hap1(a.p.)), grouped by genotype of HAP1. **(c)** Expression profiles of GPA1, STE12 and 8 downstream genes response to pheromone, and activity profile of Ste12 (Ste12(a.p.)), grouped by genotype of GPA1. **(d)** Expression profiles of AMN1, CSH1, DEM1, TOS1, ACE2, SWI5 and 10 down stream genes linked to AMN1 locus identified by Yvert *et al.* (2003), and activity profile of Ace2/Swi5, grouped by genotype of AMN1.

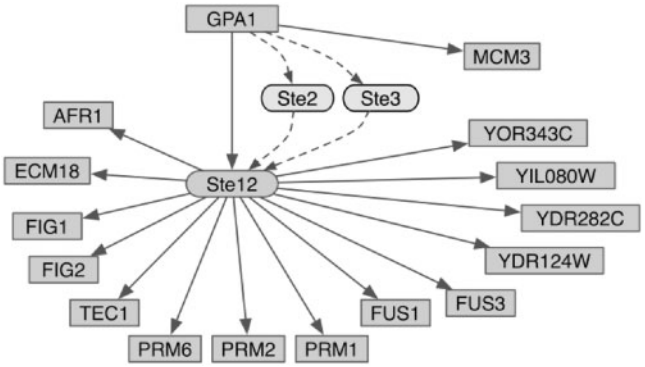

Fig. 4. eQTL module of genes linked to GPA1 locus, those genes with in-differential relationships are not shown in this figure. All the significant relationships are causal except for MCM3. The expression of MCM3 is independent with the activity of Ste12 given the expression of GPA1.

for the remaining 25 (=75−49−1, excluding LEU2 itself) genes, no enriched GO term is found. Thus the genes supporting causal model show a better functional enrichment. Changing LRT *P*-value for model comparison to 0.05 yields similar GO term enrichments (see Supplementary Materials).

Figure 3a compares the expression patterns of the 10 downstream genes participating in amino acid biosynthesis with the expression pattern of LEU2. LEU2 is not expressed in RM strain because this strain is LEU2-deleted. The elevated expressions of amino acid biosynthesis genes in RM strain suggest an interesting compensation effect due to the loss of LEU2. This is consistent with the phenotype of LEU2, 'Null mutant is viable, leucine auxotroph' (SGD) (Hong *et al.*, 2007). According to the established literature, Leu3 does bind to LEU2 (Friden and Schimmel, 1988). The expression profile of LEU3 is not linked to the LEU2 locus and Figure 3a confirms that the gene expression profile of LEU3 bears little similarity with the expression profiles of its binding targets. However, the estimated activity profile of Leu3 [denoted by Leu3(a.p.) in Fig. 3a] shows a good coherent pattern with the expression of its target genes.

In summary, all of the above discussion supports a causal scenario for this eQTL module: the genetic disruption of LEU2 perturbed the TF activity of Leu3, possibly via feedback control and the perturbed TF activity affects the expression of the downstream genes.

Hap1 mediates causal relations in energy-source control module, chr12_mod1. There are 134 genes in this module, of which two are *cis*-linked, HAP1 and GSY2. Only one TF, Hap1, is related to this module according our pre-selection (Table 3). Hap1 is a zinc finger TF involved in the complex regulation of gene expression in response to the levels of heme and oxygen (Pfeifer *et al.*, 1989). Our result indicates that both GSY2 and HAP1 can be the causative *cis*-linked genes for this module (Table 4). Biologically, HAP1 is of course more likely to be the causative gene. Hap1 is known to repress transcription from its coding gene HAP1 (Hon *et al.*, 2005), suggesting a negative correlation between its expression profile and its activity profile. Brem *et al.* (2002) showed that there is a Ty1 insertion in BY allele

of HAP1 that reduces its activity. The expression profile of HAP1 and estimated Hap1 activity profile shown in Figure 3b are consistent with these biological facts.

The gene product of GSY2 is a metabolic enzyme, glycogen synthase. Gsy2 may affect the activity of Hap1 via Bmh1 since previous studies have identified protein–protein interactions between Hap1 and Bmh1 (Krogan *et al*., 2002), and between Gsy2 and Bmh1(Ho *et al*., 2006). Bmh1 is involved in the regulation of many processes including exocytosis and vesicle transport, Ras/MAPK signaling, rapamycin-sensitive signaling (van Hemert *et al*., 2001). The expression of GSY2 is known to be induced by glucose limitation, nitrogen starvation and environmental stress. Thus this module may signify the cellular control of energy resources including sugars and lipid, which is further supported by the following GO analysis.

Out of the 46 cases supporting the causal model with either HAP1 or GSY2 as the *cis*-linked gene, there are 5 overlapping cases. We conduct a GO analysis for the 41 downstream genes, and identify enrichments in ergosterol biosynthesis (8 out of 41 genes, $P = 1.79E{-}10$) and cellular lipid metabolism (14 out of 41 genes, $P{=}1.32E{-}9$)[3]. Consistent enrichments are observed with 0.05 as LRT *P*-value cutoff (Supplementary Materials). There are 26 genes involved in ergosterol biosynthesis, of which 20 genes are included in this module. Eight of them favor causal model with *P*-value smaller than 0.01, of which the expression profiles are shown in Figure 3b, and 14 of them favor causal model with *P*-value smaller 0.05.

Ste12 mediates causal relations in pheromone response module, chr8_mod2. Among the 40 genes in this module, 2 are *cis*-linked: GPA1 and LAG1. Three TFs: Dig1, Ste12 and Tec1 may mediate the signal transduction in this module (Table 3). No causal relationship is detected for Dig1 and Tec1 (Supplementary Table 4). In contrast, Ste12 supports a total of 14 causal relationships, 1 conditional independence relationship and no reactive relationship. All of the causal relationships start with GPA1 as the *cis*-linked causative gene (Figure 4).

GPA1 encodes GTP-binding alpha subunit of the heterotrimeric G protein that couples to pheromone receptors (Guo *et al*., 2003). GO analysis for the 14 downstream genes shows the enrichment in the GO term 'response to pheromone' (8 of 14 genes, *P*-value $9.43E{-}11$) and in one of its ancestor term 'conjugation' (8 of 14 genes, *P*-value $7.57E{-}10$). The over-representation of these two GO terms is consistent with GPA1's biological function. The expression profiles of the eight downstream genes responding to pheromone are shown in Figure 3c. For the remaining 24 (=40−14−2, excluding GPA1 and STE12) genes, the enriched term is 'conjugation' (6 out of 24 genes, *P*-value $2.8E{-}04$). Consistent over-representation of GO terms can be found if we use 0.05 as LRT *P*-value cutoff. Yvert *et al*. (2003) have confirmed the functional role of a missense mutation of GPA1 from RM strain in pheromone signaling, and suggested this mutation may affect the binding of Gpa1 to two pheromone receptors Ste2 and Ste3, leading to expression level change of genes response to pheromone. However, none of Gpa1, Ste2 and Ste3 has regulation role. In addition, for the eQTL dataset that we are investigating here, expression of STE2 and STE3 are affected by an interaction between MAT locus and GPA1 locus (Brem *et al*., 2005b), and the effect of MAT locus is much stronger so that STE2 and STE3 do not co-express with GPA1 or other linked genes. Thus, it remains a question how Gpa1 or Ste2/Ste3 may affect the expression of other genes linked to GPA1 locus.

Our analysis points to Ste12 as the TF that regulates the expression of the downstream genes for this module (Table 4). This is consistent with the biological knowledge that in the pheromone response pathway, signals initiated from Gpa1, Ste2 and Ste3 propagate through the MAPK signaling cascade that reach the TF Ste12 (Veiga *et al*., 2006; Wang and Dohlman, 2004). In this case, the gene expression of Ste12 is also linked to GPA1 locus, i.e. Ste12 is one of the *trans*-linked genes in this module. The activity profile of Ste12 correlates well with its expression profile (correlation = 0.71, Fig. 3c), which is consistent with the fact that Ste12 binds the DNA sequence of itself (Harbison *et al*., 2004).

Ace2/Swi5 mediate causal relations in the mitotic-exit network module, chr2_mod1. The pre-selected TFs for this module are Ace2 and Swi5 (Table 3). Ace2 activates the expression of the early G1-specific genes and Swi5 activates the transcription of the genes expressed at the G1 or M/G1 boundary of the cell cycle (Dohrmann *et al*., 1996). Because Ace2 and Swi5 bind the same DNA sequences *in vitro* with similar affinities, and they share most target genes *in vivo* (Dohrmann *et al*., 1996), we treat them as one TF complex in this study.

Among the 267 genes of this module, 8 are *cis*-linked. Our analysis shows a total of 651 cases supporting the TF activity mediation scenario. In contrast, there is no case supporting reactive model and only five cases supporting conditional independence model (Table 4). Among the seven *cis*-linked genes supporting at least one causal relation, AMN1 has the strongest *cis*-linkage. AMN1 encodes a protein required for daughter cell separation, multiple mitotic checkpoints and stability (Wang *et al*., 2003). Previous works (Ronald *et al*., 2005; Yvert *et al*., 2003) have confirmed the causal role of AMN1. Yvert *et al*. (2003) identified one coherent expressed cluster of 12 genes linked to AMN1 locus, among which 8 are previously reported to have daughter cell specific expression. Yvert *et al*. (2003) also proposed an explanation how Amn1 affects the expression of the daughter cell-specific genes: Amn1 regulates the mitotic exit network (MEN), MEN activates Ace2, and Ace2 in turn regulates genes specifically expressed in daughter cells during budding. This explanation is consistent with our result of choosing Ace2/Swi5 as the mediator. Among the 12 genes identified by Yvert *et al*. (2003), excluding AMN1 itself and one gene ISR1, which is linked to a locus near to, but different from AMN1 locus in this bigger yeast cross, causal relations are identified in 8 of the remaining 10 genes (Supplementary Table 7, Fig. 3d).

[3]There are 91 genes that do not fit causal model with either HAP1 or GSY2 as *cis*-linked gene at *P*-value 0.01. From these 91 genes, we choose a subset of 41 genes with most insignificant LRT result (for comparing causal model with the other two models) and test for functional enrichment. No significant GO term enrichment is found (Supplementary Materials).

GO terms overrepresented by the 136 down-stream genes of the AMN1 to Ace2/Swi5-activity causal path include 'ribosome biogenesis' (58 of 136 genes, *P*-value 1.95*E*−40) and one of its ancestor terms 'organelle organization and biogenesis' (71 of 136 genes, *P*-value 2*E*−17). No significantly enriched GO term is found for the remaining 130 (=267−136−1, excluding AMN1 itself) genes.

4 DISCUSSION

In this study, we propose a procedure to detect eQTL modules and identify related TFs. The novelties of this approach are: (1) it makes a clear distinction between the unobserved TF activity profile and the observed TF gene expression profile; (2) external ChIP-Chip data and TF target gene information from the literature are utilized; (3) causal modeling is employed to formulate the scenario of TF activity mediation; (4) a likelihood ratio test is introduced to rule out cases where the scenarios of TF activity mediation are compatible with two contrasting scenarios, thus reducing the chance of detecting false positives and (5) the eQTL hot spots are allowed to have variable lengths.

The approach of causal relation identification by model comparison has been used by Schadt *et al.* (2005) in order to dissect the relation between DNA variation, gene expression and clinical trait. There are several differences between our approach and Schadt *et al.*'s approach. First we conduct formal statistical tests to aid the model identification, while Schadt *et al.* compared different models by AIC. Secondly, Schadt *et al.* did not consider the conditional independence model. Instead, they used what they called the 'independent model' as a competitive model. Their 'independent model' is indeed a complete model, in which any two of the three variables are dependent. There is at least one obvious drawback of omitting the conditional independence model. Some in-differential cases, in which the causal model and the conditional independence model are comparable, will be classified as causal model when conditional independence model is omitted (Supplementary Table 8).

The LRT that we used in comparing the three models (of equal complexity) admits the complete model as the ground truth because it contains all three sub-models. The goal of the test is not to conclude how well a sub-model fits the data (Vuong, 1989). Rather it is used to see which one of them fits the data better. The finding that the causal model outperforms the other two sub-models provides good evidence that TF regulation is a feasible scenario. It is also an encouraging starting point for more complex model building to improve the goodness of fit. We further compare the goodness of fit of our three models with the complete model (Supplementary Table 9). We see that the majority of the significant causal relationships detected by our approach (Table 4) do fit the data well. For instance, out of the 136 causal relationships with AMN1 as the causative gene for the Ace2/Swi5 mediated module, 84 are accepted when testing against the complete model. For Ste12, Hap1 and Leu3 modules discussed earlier, the numbers are 10 out of 14, 23 out of 28 and 29 out of 49, respectively. In addition, the genes favoring causal model by our approach exhibits better GO enrichment than the genes favoring causal model obtained by Schadt's approach or by the approach of comparing four models including complete model (see Supplementary Materials).

TF regulation is only one explanation why many genes are linked to an eQTL hot spot. We should not rule out alternative biological mechanisms for better explanations. Nevertheless, the genes favoring TF mediation scenario (causal model) do exhibit GO enrichments in specific biological processes, which are consistent with the functions of the *cis*-linked genes and the mediating TFs. However, without additional information, we cannot be really sure about excluding any compatible models.

The identification of causal TFs of eQTL modules reveal more subtle aspects about gene expression linkages. For medical applications, this could lead to the discovery of more options to correct genetic defects. One of the major objectives of eQTL studies is to identify the genetic determinants/modifiers of complex traits, which are likely related with many genes. The identification of the upstream TFs responsible for mediating the expression of a good portion of these genes allows us to manipulate the combined effect of the downstream genes, which may in turn affect the disease trait.

There is room to improve our procedure. For example, incorporating two or more TFs (TF complexes) simultaneously for each eQTL module may help identify more causal relationships. But this would require more sophisticated statistical methods to estimate the unobserved TF activities, a direction worth further investigation.

ACKNOWLEDGEMENTS

We thank Dr Shinsheng Yuan and Dr Rachel Brem for discussions. We are also grateful to the two reviewers for their helpful comments and suggestions. This work is supported by NSF grants DMS-0201005, DMS-0104038 and DMS-0406091. K.-C.L.'s work is also supported in part by an internal grant from Academia Sinica, Taipei, Taiwan, ROC.

Conflict of Interest: none declared.

REFERENCES

Andreadis,A. *et al.* (1984) Yeast LEU2. Repression of mRNA levels by leucine and primary structure of the gene product. *J. Bio. Chem.*, **259**, 8059–8062.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.

Bing,N. and Hoeschele,I. (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics*, **170**, 533–542.

Brem,R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.

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

Brem,R.B. *et al.* (2005b) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, **436**, 701–703.

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.

Dohrmann,P.R. *et al.* (1996) Role of negative regulation in promoter specificity of the homologous transcriptional activators Ace2p and Swi5p. *Mol. Cell. Biol.*, **16**, 1746–1758.

Friden,P. and Schimmel,P. (1988) LEU3 of Saccharomyces cerevisiae activates multiple genes for branched-chain amino acid biosynthesis by binding to a common decanucleotide core sequence. *Mol. Cell. Biol.*, **8**, 2690–2697.
Guo,M. *et al.* (2003) The yeast G protein alpha subunit Gpa1 transmits a signal through an RNA binding effector protein Scp160. *Mol. Cell*, **12**, 517–524.
Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
Ho,Y. *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
Hong,E.L. *et al.* (2007) Saccharomyces Genome Database. http://db.yeastgenome.org/ (Accessed on Feb 18th, 2007).
Hon,T. *et al.* (2005) The heme activator protein Hap1 represses transcription by a heme-independent mechanism in Saccharomyces cerevisiae. *Genetics*, **169**, 1343–1352.
Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
Kulp,D.C. and Jagalur,M. (2006) Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics*, **7**, 125.
Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.
Morley,M. *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
Petretto,E. *et al.* (2006) Heritability and tissue specificity of expression quantitative trait loci. *Genet.*, **2**, e172.
Pfeifer,K. *et al.* (1989) Functional dissection and sequence of yeast HAP1 activator. *Cell*, **56**, 291–301.
Rockman,M.V. and Kruglyak,L. (2006) Genetics of global gene expression. *Nat. Rev. Genet.*, **7**, 862–872.
Ronald,J. *et al.* (2005) Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet.*, **1**, e25.
Rustici,G. *et al.* (2005) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.*, **36**, 809–817.
Schadt,E.E. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297–302.
Schadt,E.E. *et al.* (2005) An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.*, **37**, 710–717.
Stranger,B.E. *et al.* (2005) Genome-wide asssociations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
van Hemert,M.J. *et al.* (2001) Yeast 14-3-3 proteins. *Yeast*, **18**, 889–895.
Veiga,D.F. *et al.* (2006) Gene networks as a tool to understand transcriptional regulation. *Genet. Mol. Res.*, **5**, 254–268.
Vleugel,M. *et al.* (2004) No amplifications of hypoxia-inducible factor-1alpha gene in invasive breast cancer: a tissue microarray study. *Cell. Oncol.*, **26**, 347–351.
Vuong,Q.H. (1989) Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.
Wang,Y. and Dohlman,H.G. (2004) Pheromone signaling mechanisms in yeast: a prototypical sex machine. *Science*, **306**, 1508–1509.
Wang,S. *et al.* (2006) Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.*, **2**, e15.
Wang,Y. *et al.* (2003) Exit from exit: resetting the cell cycle through Amn1 inhibition of G protein signaling. *Cell*, **112**, 697–709.
Yu,T. and Li,K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, **21**, 4033–4038.
Yvert,G. *et al.* (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.*, **35**, 57–64.
Zhu,J. *et al.* (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.*, **105**, 363–374.