2-4-2019

Inferring Gene RegulatoryNetworks from a Populationof Yeast Segregants

Chen Chen
Purdue University

Tony R. Hazbun
Purdue University, thazbun@purdue.edu

Min Zhang
Purdue University

Recommended Citation

Chen, C., Zhang, D., Hazbun, T.R. et al. Inferring Gene Regulatory Networks from a Population of Yeast Segregants. Sci Rep 9, 1197 (2019). https://doi.org/10.1038/s41598-018-37667-4

SCIENTIFIC REPORTS

OPEN

Inferring Gene Regulatory Networks from a Population of Yeast Segregants

Received: 31 July 2018
Accepted: 30 November 2018
Published online: 04 February 2019

Chen Chen[1], Dabao Zhang[1,3], Tony R. Hazbun[2,3] & Min Zhang[1,3]

Constructing gene regulatory networks is crucial to unraveling the genetic architecture of complex traits and to understanding the mechanisms of diseases. On the basis of gene expression and single nucleotide polymorphism data in the yeast, *Saccharomyces cerevisiae*, we constructed gene regulatory networks using a two-stage penalized least squares method. A large system of structural equations via optimal prediction of a set of surrogate variables was established at the first stage, followed by consistent selection of regulatory effects at the second stage. Using this approach, we identified subnetworks that were enriched in gene ontology categories, revealing directional regulatory mechanisms controlling these biological pathways. Our mapping and analysis of expression-based quantitative trait loci uncovered a known alteration of gene expression within a biological pathway that results in regulatory effects on companion pathway genes in the phosphocholine network. In addition, we identify nodes in these gene ontology-enriched subnetworks that are coordinately controlled by transcription factors driven by trans-acting expression quantitative trait loci. Altogether, the integration of documented transcription factor regulatory associations with subnetworks defined by a system of structural equations using quantitative trait loci data is an effective means to delineate the transcriptional control of biological pathways.

Gene expression is a fundamental step in the flow of information from an organism's genotype to phenotype. The genetic information encoded in an organism's DNA is transferred into a functional gene product (e.g., protein) via the process of gene expression, and gene expression leads to the formation of the organism's phenotype. Gene expression have been found to be associated with a broad range of complex traits and diseases[1], and thus play an important role in determining an organism's development. Numerous efforts have been made to map phenotypes to gene expression in order to dissect their genetic basis.

Genes rarely act in isolation; instead, they interact with each other and make up gene regulatory networks to function as a whole[2]. The study of this mechanism is crucial for understanding the properties and functions of genes, which help reveal the genetic architecture of complex traits and diseases. Although genetic experiments can be conducted to discover interactions among genes, this approach can be costly and time consuming. Alternatively, measurements of gene expression levels reveal gene expression patterns in a specific condition and can be exploited to infer gene regulatory networks. Various approaches have been proposed to infer gene regulatory networks using gene expression data, such as relevance networks[3–7], Bayesian networks[8–11], Gaussian graphical models[12–15], and many others.

Recent advances in sequencing technologies make it feasible to obtain both whole-genome genotype and gene expression for each individual, i.e., genetical genomics data[16]. Combining genetics with gene expression reveals additional information on genetic structure and holds great promise for improving the accuracy of gene regulatory network inference. Numerous genetical genomics experiments, such as the Genotype-Tissue Expression (GTEx) project[17], have been conducted to collect genetical genomics data.

Much effort has been devoted to using genetical genomics data for genome-wide association (GWA) analysis of gene expression, i.e., expression quantitative trait loci (eQTL) mapping[18]. Mapping of eQTL intends to elucidate variation of expression traits attributed to genomic variation, and to identify chromosomal loci (i.e., eQTL)

[1]Department of Statistics, Purdue University, West Lafayette, IN, 47907, USA. [2]Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, IN, 47907, USA. [3]Purdue University Center for Cancer Research, Purdue University, West Lafayette, IN, 47907, USA. Correspondence and requests for materials should be addressed to D.Z. (email: zhangdb@purdue.edu) or T.R.H. (email: thazbun@purdue.edu) or M.Z. (email: minzhang@purdue.edu)

of genetic polymorphisms associated to the expression of a gene under investigation. An eQTL located within the region of the gene under investigation is called a cis-eQTL, otherwise it is called a trans-eQTL. While the cis effects of a gene represent direct regulations, indirect regulations of trans-eQTL are likely caused by interactions among genes. These eQTL provide insight on the functional sequences of the gene expression, and thus an indirect interrogation of the functional landscape of gene regulations[19].

Gene regulatory networks can be characterized using a system of structural equations[20], with each equation describing the causal effects of cis-eQTL and the regulatory effects of other genes on a given gene. Such a framework makes it feasible to take a genome-wide survey and to directly reveal interactions among genes. Application of structural equations in genetical genomics studies have been previously demonstrated[21–24]. Two studies are applicable to constructing gene regulatory networks for a small number of genes[21,22]. However, genetical genomics experiments usually collect whole-genome gene expressions for a very limited number of samples, therefore the number of genes is much larger than the sample size. For such consideration, another study[23] proposed to apply the adaptive lasso[25] to construct a sparse gene regulatory network. An additional approach instead proposed to maximize a penalized likelihood for constructing a sparse gene regulatory network[24].

Here we construct gene regulatory networks in yeast via building up a large system of structural equations with the two-stage penalized least squares (2SPLS) method[26]. We applied the 2SPSLS method to an eQTL dataset derived from a cross between a wild yeast vineyard strain and a laboratory strain[27]. Fitting one linear model for each gene at each stage, the 2SPLS method develops optimal prediction of a set of conditional expectations at the first stage, and consistent selection of regulatory effects from massive candidates at the second stage. It is computationally fast and allows for parallel implementation, outperforming the adaptive lasso based algorithm[23], and the sparsity-aware maximum likelihood algorithm[24], in terms of both accuracy and speed, for identifying regulatory effects in different network structures. This parallel implementation makes it feasible to evaluate the significance of regulatory effects via the bootstrap method. Using this approach we identified subnetworks that were enriched in gene ontology categories suggesting an extrinsic regulatory mechanism controlling these biological networks. Our eQTL predictions uncovered a known alteration of gene expression within a biological pathway that results in regulatory effects on companion pathway genes in the phosphocholine network. In addition, we delineate how nodes in these subnetworks are coordinately controlled by a transcription factor driven by trans-acting eQTL. For example, we detail how a proteasomal subnetwork is controlled by the *RPN4* transcription factor, via a trans-acting eQTL, resulting in the coordinated expression of genes in this subnetwork.

Results and Discussion

Identified cis-eQTL. To investigate and demonstrate the utility of cis-eQTL to infer regulatory interactions among genes, we performed a genome-wide survey of the budding yeast, *Saccharomyces cerevisiae*. We used a well-established dataset that involved a cross between a laboratory strain (BY4716) and a wild yeast strain (RM11-1A) isolated from a California vineyard. At a significance level of 0.05, we identified 409 genes (out of a total of 5,727 genes), with significant cis-eQTL (Table S1 has each p-value listed). The set of cis-eQTL for each gene was filtered to control the pairwise correlation under 0.90, and then was further filtered to keep a maximum of three cis-eQTL that have the strongest association with the corresponding gene expression. Detailed results are provided in Supplementary Information (Table S1).

Constructed gene regulatory networks. The constructed network includes a total of 409 nodes and 5,068 edges respectively (Table S2). Among 260 edges repeatedly identified in more than 80% of the 10,000 bootstrap data sets, 258 edges, including 226 positive and 32 negative regulations, were in the 5,068 edges constructed from the original data set. The edges formed a number of subnetworks, among which 12 identified subnetworks have more than 5 genes (Table S3). We examined the 12 subnetworks for gene set enrichment using DAVID and found enrichments in gene ontology categories within each subnetwork (Table S4).

Figure 1 shows the largest subnetwork formed by these 260 edges, other constructed subnetworks are listed in Supplementary Information (Table S3). This large subnetwork (subnetwork 1) was subjected to YeastMine analysis to identify gene ontology enrichments and pathways[28]. This analysis revealed that 17 genes in subnetwork 1 are involved in a variety of biosynthetic pathways (p-value = 4.17E-07) and synthesis of secondary metabolites (Table S5). Many genes within this subnetwork are involved in amino acid synthesis and we also observed a subset of connected genes that were closely associated with phosphocholine metabolism. The enrichment in gene ontology terms for the subnetworks demonstrated that using the 2SPLS method of constructing regulatory cis-eQTL results in identification of clusters of genes with common biological function. The closely connected nodes with genes of common function suggest that genetic polymorphisms commonly result in compensating regulatory events of companion genes.

Comparison to existing databases (STRING and BioGRID). To investigate the constructed gene regulations with involvement of downstream protein-protein interactions, we compared the subnetworks to the known and predicted protein-protein interactions in the STRING database (http://string-db.org/)[29]. Developed by a consortium of institutions, the current version of STRING collects information of 9,643,763 proteins from 2,031 organisms. The comparison demonstrated common and enriched processes that parallel the gene ontology enrichments detected via DAVID analysis. For example, subnetwork 6 yielded a highly connected set of nodes that involved proteasome subunits and associated proteins reflecting the molecular architecture of the proteasome complex and this subnetwork is further analyzed in this report. Analysis of Subnetwork 1 with STRING database also revealed that *CHO1*, *ITR1* and *OPI3* are interconnected identically to the phosphocholine network discussed in the following section (highlighted in yellow of Fig. 1). Similar results were obtained when comparing to BioGRID using the YeastMine tool (Table S5)[28,30]. These striking examples of similar network organization observed in STRING with our predictions validated our approach and prompted the examination and integration

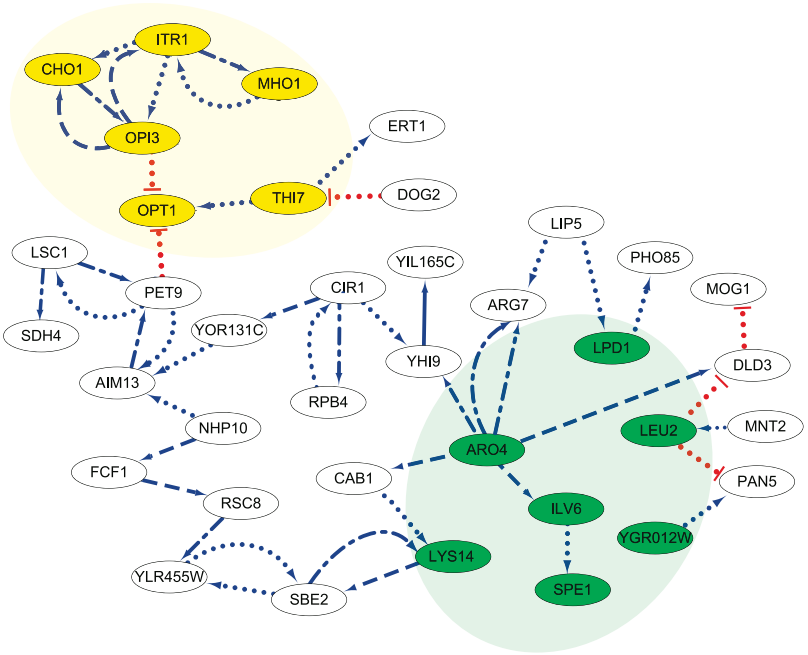

Figure 1. The largest gene regulatory subnetworks in yeast. While the dotted, dash-dotted, dashed, and solid lines implied the corresponding connections were constructed respectively in [80%, 90%), [90%, 95%), [95%, 100%), and 100% of the bootstrapping data sets, the blue arrow- and red bar-headed lines indicate up and down regulations, respectively. Highlighted in yellow is the Inositol subnetwork in which several genes involved in the CDP-DAG/phosphocholine pathway are coordinately repressed by exogenous inositol. Within the amine biosynthetic process subnetwork highlighted in green, *LEU2, LPD1, YGR012W, LYS14, ILV6*, and *ARO4* are involved in multiple biosynthetic processes (as shown in Table S5).

of these subnetworks with the literature and other functional genomics database information such as mRNA profiling.

The Phosphocholine subnetwork. All of the genes in the phosphocholine subnetwork (highlighted in yellow of Fig. 1), except for *OPT1*, have similar patterns of regulation and are repressed by the presence of inositol or choline in yeast growth medium. The majority of the genes (*MHO1*, *ITR1*, *CHO1* and *OPI3*) are involved in lipid metabolism and are subject to transcriptional regulation by the Opi1 repressor[31]. Strikingly, two of these genes are in a linear metabolic pathway converting cytidine diphosphate diacylglycerol (CDP-DAG) to phosphocholine (*CHO1* and *OPI3*) (Fig. 2)[32]. *ITR1* encodes a transporter that imports exogenous inositol from the growth media. The function of *MHO1* is unclear, but the gene has been shown to be synthetic lethal with *PLC1*, an enzyme involved in the production DAG and inositol trisphosphate (IP3)[33]. The eQTL-based prediction of reciprocal positive regulation between genes within the DAG-phosphocholine pathway indicates a regulatory interdependence of these genes (*MHO1*, *ITR1*, *CHO1* and *OPI3*). Interestingly, these genes are coordinately controlled by the Ino2-Ino4 transcription factor complex via the inositol sensitive upstream activating element (UAS-INO) but additional regulation may be exerted based on mRNA abundance level of pathway components. For example, *CHO1* mRNA stability increased in response to respiratory deficiencies leading to increased phosphatidylserine levels and activities of other CDP-DAG pathway enzymes[34]. The regulatory mechanisms involved for phospholipid synthesis are complex and include biochemical regulation by several phospholipid precursors and products including phosphatidic acid (PA) and CDP-DAG[35]. PA helps to sequester the Opi1 repressor away from the nucleus[36] and elevated levels of CDP-DAG favors the Opi1-mediated repression of genes under control of the UAS-INO element[35], shown in Fig. 2.

In addition, inositol-based regulation has been observed to control various metabolic pathways involved in membrane biogenesis including the activation of *OPT1*, an oligopeptide and glutathione transporter encoding gene[31]. The prediction that *OPI3* negatively regulates *OPT1* expression is consistent with the opposite effects of inositol on these two genes. An examination of the expression pattern of *OPT1* and *OPI3* shows the strong anti-correlated expression pattern between these genes (Fig. 3A). The inferred gene-gene relationships for this phosphocholine subnetwork demonstrate the utility of our eQTL analysis to delineate biologically relevant pathways. In addition, our analysis implicated that a poorly characterized gene, *MHO1*, may have a functional role in the phosphocholine pathway.

Examination of the sequence of the RM and BY parental strains for the genes in the phosphocholine subnetwork revealed a lack of nonsynonymous polymorphisms within the *OPI3* gene and the presence of four single nucleotide polymorphisms (SNPs) in the upstream promoter region (500 bp from the ATG). The identical amino acid sequence of Opi3 present in the RM and BY strains suggests that the differences between strains is due to expression level of the protein but not due to any differences in protein stability or activity. One of the SNPs was

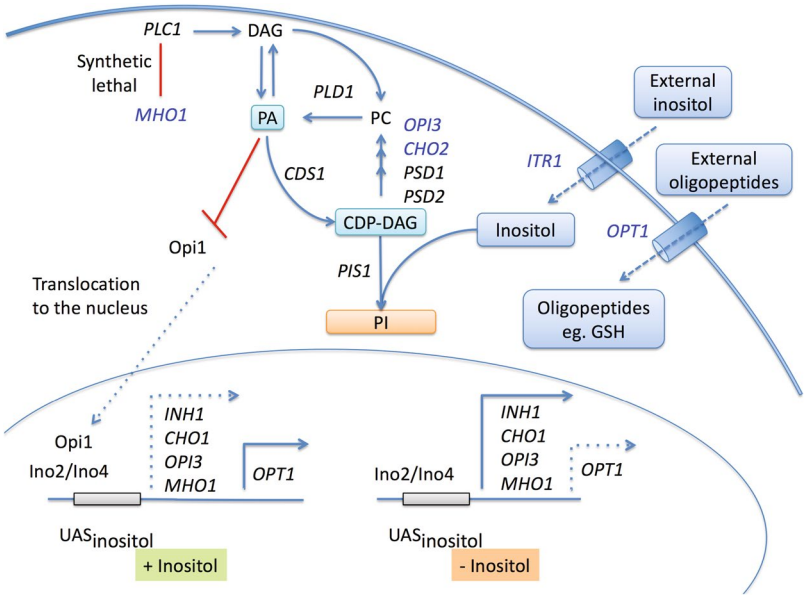

Figure 2. The pertinent features of the phosphocholine pathways. The CDP-DAG phosphocholine pathway shows the involvement of genes implicated in the eQTL–based phosphocholine subnetwork (Blue font) - *CHO1*, *OPI3* and *ITR1* (transport of external inositol). PA inhibits the Opi1 repressor translocation to the nucleus. Low levels of PA result in translocation of Opi1 to the nucleus and the association and repression of the Ino2/Ino4 heterodimeric transcription factor. Low levels of inositol result in activation of transcription of several phosphocholine pathway genes and *MHO1* and repression of *OPT1*.

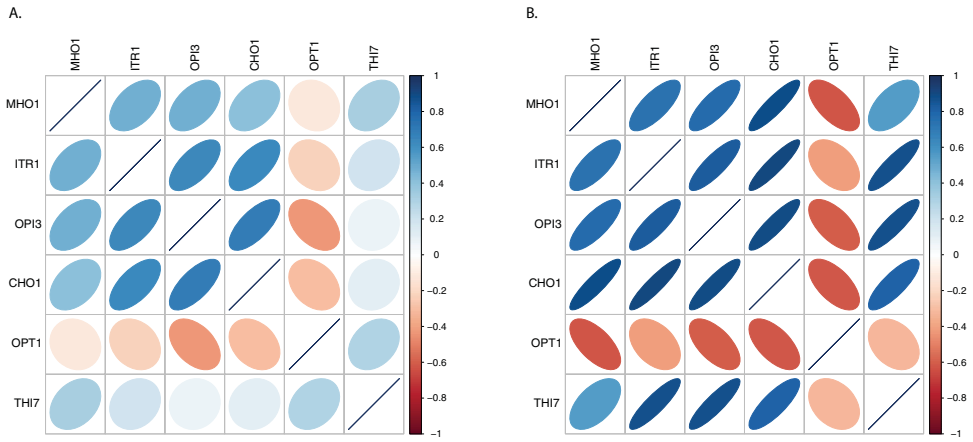

Figure 3. Correlation of expression for genes in the phosphocholine network. (**A**) Pairwise correlation plot between the 6 genes in the phosphocholine subnetwork for the eQTL expression data from parental strain replicates[27]. (**B**) Pairwise correlation plot between the 6 genes involved in phosphocholine subnetwork for independent expression datasets from SPELL[39]. The color indicates the direction of the correlations (blue indicates positive and red indicates negative) and the shape represents the strength of correlation.

located at the -1 position upstream to the start codon, which is a position demonstrated to affect gene expression level. The adenine nucleotide in the BY strain favors a higher expression level compared to guanine for the RM parent based on large scale analysis of variant nucleotides at the -3 to -1 position relative to the start codon[37]. This is reflected in the overall expression levels observed for mRNA levels in the eQTL expression data set from Serial Pattern of Expression Levels Locator (SPELL) database[38]: ~1.5 fold lower expression for 12 RM parent values compared to a BY reference pool (see Tables 1 and S4). The *CHO1* gene exhibited an expression difference of 1.2 fold or lower between the RM and BY parents. Genes with similar mRNA levels between the parent strains do not harbor SNPs that are driving the expression differences evident in the segregant progeny strains suggesting the presence of trans-acting SNPs as discussed in the proteasome subnetwork section below. In addition to SNPs in the promoter region, the other genes in the network exhibited nonsynonymous polymorphisms using the Variant Viewer analysis tool[39], as shown in Table 1.

Gene	Nonsynonymous SNPS	SNPS in Promoter REGION[a]	RM/BY Fold Change	P-Value[b]
CHO1	A9T; L234F	6 (−78; −79; −213; −228; −375; −451)	1.24*	0.02
ITR1	C521F	2 (−211; −286)	0.98	ns
MHO1	A331T; F164I	4 (−141; −169; −224; −285)	1.11**	0.002
OPI3	None	4 (−1; −389; −395; −450)	1.51**	0.008
OPT1	A200V; V439I	4 (−108; −142; −143; −333)	0.98	ns

Table 1. Summary of SNPs and gene expression difference between RM and BY strains for genes in the phosphocholine network. [a]The total number of SNPs in the promotor region within 500 bp upstream of the gene start. [b]P-value calculated by comparing 12 RM parent strains to 6 BY parent strains (ns = not significant).

Validation of expression patterns using independent datasets. From the SPELL database, we input all 6 genes from the phosphocholine subnetwork to identify expression profiling experiments that had correlated data for the query genes. This approach resulted in 7 datasets with relevance weighting larger than 1.0% compared to all other experimental datasets. Among these, several datasets had missing data or very low levels of expression for the 6 genes of interest with the exception of 3 datasets, which were subjected to further analysis. We calculated the pairwise correlation between these 6 genes and visualized the correlation matrix using the R package "corrplot" (https://cran.r-project.org/web/packages/corrplot/index.html) for one of these data sets that focused on hypo-osmotic shock[40]. The pairwise correlation plot[41] is presented in Fig. 3B. This independent expression data set demonstrated the strong anti-correlation between *OPT1* and the other genes within the phosphocholine subnetwork, which is consistent with the prediction of negative regulation of *OPT1* by *OPI3*. Other genes in the network demonstrated similar correlation plots to the eQTL data from parental replicates with the exception of the *THI7-OPT1* pair, which appears to be regulated differently in hypo-osmotic conditions. The *THI7* gene encodes a transporter that facilitates the uptake of thiamine and is upregulated in the hypo-osmotic experiment whereas it is down-regulated in the RM strain compared to the BY parent strain. The regulatory relationship between *THI7-OPT1* pair appears complex and is altered depending on environmental conditions and stress.

The Proteasome subnetwork. Analysis of the genes in subnetwork 6 indicated enrichment in ubiquitin-dependent protein catabolic processes (p-value = 1.25E-04 which is adjusted to 0.014 by applying the Bonferroni method), shown in Table S4. This subnetwork included 4 genes that encode proteasomal subunits. The network structure indicated extensive reciprocal regulation between proteasomal genes (Fig. 4A). The proteasome has key roles in cellular homeostasis and is subject to multiple regulatory mechanisms[42]. This reciprocal regulation predicted by our eQTL analysis is consistent with a proposed feedback circuit in which the *RPN4* transcription factor upregulates proteasomal genes but is also degraded by the proteasome. A similar feedback mechanism exists in higher eukaryotes because deletion of the regulatory S5a/Rpn10/p54 subunit results in extreme and coordinate upregulation of other proteasomal genes[43]. Additional studies with RNA interference in *Drosophila* indicate that knockdown of gene expression of a proteasomal subunit results in upregulation of the companion subunit mRNAs[44,45]. A mechanism underlying mRNA upregulation in higher eukaryotes appears to be dependent upon the 5′ untranslated mRNA region[46]. These and other studies have culminated in a model where factors such as proteotoxic stress, proteasome inhibitors and proteasomal gene mutations have been documented to upregulate proteasome levels via *RPN4* -mediated transcription. *RPN4* is a transcription factor that specifically binds to the *P*roteasome *A*ssociated *C*ontrol *E*lement (PACE) found in most proteasome genes[47,48] resulting in coordinate regulation of many proteasome genes (Fig. 4B). The positive regulation predictions between proteasome genes outlined in subnetwork 6 (Fig. 4A) may reflect this coordinate regulation. The RM and BY parent strain gene expression data, 6 BY parent strains and 12 RM parent strains, indicated similar expression levels[27] between the proteasomal genes (Fig. 4C) suggesting that trans-acting polymorphisms are driving the expression differences evident in the segregant progeny strains. The other three genes in this network (*CCT2*, *SEN1* and *SMF1*) have differing expression levels between RM and BY parent strains. The prevalence of trans-acting eQTL has been documented and previously reported for this dataset between 22–48%[49]. The regulatory events observed in subnetwork 6 maybe controlled by *RPN4* because six nodes (*RPN6*, *CDC53*, *RPN5*, *SPT16*, *RPN1* and *RPT5*) have documented regulations by *RPN4* based on the YEASTRACT database[50], shown in Table S7. The edges in this network may reflect the timing of expression driven by *RPN4* and not the direct regulation of one proteasomal gene by another proteasomal gene. Further examination of all the subnetworks using the YEASTRACT database shows several networks that are controlled by one or more transcription factors (Table S7). In total, this proteasome subnetwork example demonstrates that interpretation of eQTL regulatory information must be integrated with heterologous information such as transcription factor activity. This integrated approach recapitulates the biological networks controlled by transcription factors.

Conclusions

In this work, we constructed gene regulatory networks in yeast via establishing a large system of structural equations. By integrating genomic information into gene regulatory network construction, we identified subnetworks that were enriched in gene ontology categories revealing regulatory mechanisms controlling these biological pathways. Our eQTL predictions uncovered a known alteration of gene expression within a biological pathway that results in regulatory effects on companion pathway genes in the phosphocholine network. In addition, we delineate how nodes in these subnetworks are coordinately controlled by a transcription factor driven by

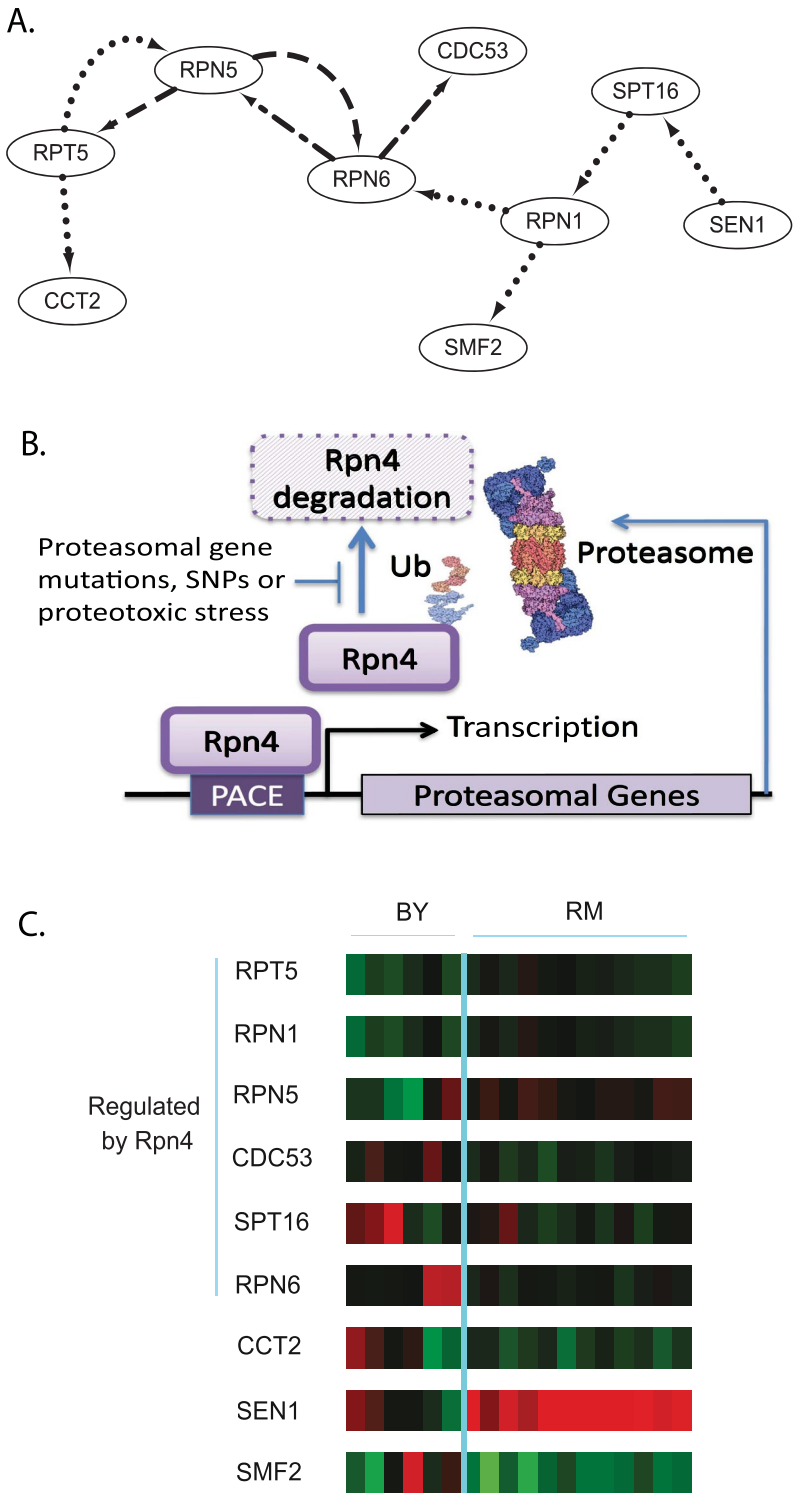

Figure 4. Proteasomal subnetwork is subject to feedback regulation. (**A**) Subnetwork 6 contains four proteasomal genes and other genes enriched for ubiquitin-dependent protein catabolic processes. (**B**) Feedback regulation model depicting the control of proteasomal gene transcription. The *RPN4* transcription factor binds to the promoter of proteasomal genes via the PACE DNA site and initiates proteasomal gene transcription. The *RPN4* transcription factor is modified by ubiquitin (Ub) and degraded by the proteasome. Mutations to proteasomal genes, SNPs or proteotoxic activity result in the inhibition of *RPN4* degradation. (**C**) Heat map depicting the expression level of each strain (6 BY parent strains and 12 RM parent strains[27]) for genes in the proteasomal subnetwork. Six genes within the network have evidence of regulation by *RPN4*. The *RPN4*-regulated genes do not exhibit any difference between BY and RM parent strains suggesting that trans-acting eQTL are impacting expression in segregant strains. Note other genes in the network do demonstrate different expression levels between the parent strains.

trans-acting eQTL. Hence, directionality of the edges in the subnetworks may reflect the timing of transcription control of these related genes. We expect that it is possible to build regulatory networks with increased size and accuracy with more extensive datasets of eQTL. For example, several studies have used additional quantitative traits, multi-parent crosses and also integrated other phenotypic markers such as metabolite levels in probing yeast biological networks[51–54]. This study demonstrates that 2SPLS analysis provides insight on understanding regulatory relationships among genes, which reveal the genetic architecture of complex traits and diseases.

Materials and Methods

eQTL analysis. We analyzed a yeast data set with 112 segregants from a cross between two strains BY4716 and RM11-la[27]. The study measured mRNA expression combined with genotyping data (2,956 SNPs) from the 112 haploid segregant progeny from the BY4716 and RM11-la cross. The data were obtained from the Gene Expression Omnibus[55] (GEO; http://www.ncbi.nlm.nih.gov/projects/geo/) with a GEO accession number of GSE1990. A total of 5,727 genes were measured for their expression values, and detailed procedure of normalization was previously described[27]. Briefly, base 2 logarithm transformation of the gene expression ratio (sample/BY4716 reference) was calculated and averaged over duplicated samples. The data were then normalized using MAANOVA package[56]. As previously described[27], the missing genotype information of the available 2,956 markers was imputed using sample mean prior to analysis. To identify eQTL for each gene, the expression of each gene was regressed against all markers in the gene and within 500 bp upstream of the genetic region, using a simple linear regression model.

Network construction. Denoting the expression values of p genes as $Y = (Y_1, \ldots, Y_p)$ and the genotypic values of q polymorphisms as $X = (X_1, \ldots, X_q)$, we characterized the gene regulatory network using a system of structural equations,

$$Y = Y\Gamma + X\Psi + \mathrm{E}, \tag{1}$$

where the $p \times p$ matrix Γ has zero diagonal elements and contains gene regulatory effects, the $q \times q$ matrix Ψ contains causal genomic effects from cis-eQTL, and E is an $n \times p$ matrix of error terms. We assume that X and E are independent of each other, and each component of E is independently distributed as normal with zero mean while its rows are identically distributed.

With the expression levels of the 409 genes and the genotypes of the selected cis-eQTL for each of 112 segregants, we applied the 2SPLS method[26] to establish the system (1) for constructing a gene regulatory network in yeast. Fitting a single regression model for each endogenous variable at each stage, 2SPLS employs the ridge regression at the first stage to obtain consistent estimation of a set of conditional expectations, and the adaptive lasso[25] at the second stage to consistently identify regulatory effects among a huge number of candidates.

To evaluate the reliability of constructed gene regulations, we generated a total of 10,000 bootstrap data sets (each with 112 segregants) by randomly sampling the original data with replacement, and applied 2SPLS to each data set to infer the gene regulatory network.

SPELL - *S. cerevisiae*. To validate the results using independent datasets, we searched the SPELL database (http://spell.yeastgenome.org/)[38]. The phosphocholine subnetwork genes were entered into SPELL and experimental datasets were identified that had expression data for all genes and were highly ranked with relevance weighting larger than 1.0%. Using this approach, we identified three datasets for analysis and demonstrated independent validation of the predicted phosphocholine subnetwork structure.

Identification of controlling transcription factors. A curated database of yeast transcription factors was used to identify transcription factors that are associated with regulating genes within subnetworks. The Yeast Search for Transcriptional Regulators And Consensus Tracking (YEASTRACT) database includes over 163,000 regulatory associations curated from the literature[50]. Genes within each subnetwork were used as the input gene list to search for transcription factors that are documented or potentially regulate gene within the list. Genes were considered to have a regulatory association with the transcription factor if there was documented DNA binding evidence plus expression evidence. The transcription factors were ranked by percentage of genes regulated by the respective transcription factor and the output for each subnetwork was included in the Supporting Information.

Data Availability

While the gene expression information can be found at Gene Expression Omnibus database with accession no. GSE1990, the genotype data are provided in the Supplemental Material with permission from Leonid Kruglyak. The gene expression of 12 RM and 6 BY parent strains are collected from Serial Pattern of Expression Levels Locator (SPELL) database (http://spell.yeastgenome.org/)[38]. The gene expression from the hypo-osmotic shock experiment[39] can be downloaded from https://spell.yeastgenome.org/search/dataset_details/1002.

References

1. Dermitzakis, E. T. From gene expression to disease risk. *Nat Genet* **40**, 492–493, https://doi.org/10.1038/ng0508-492 (2008).
2. Mani, R. *et al*. Defining genetic interaction. *Proc Natl Acad Sci USA* **105**, 3461–3466, https://doi.org/10.1073/pnas.0712255105 (2008).
3. Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci USA* **97**, 12182–12186, https://doi.org/10.1073/pnas.220392197 (2000).
4. Luo, F. *et al*. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* **8**, 299, https://doi.org/10.1186/1471-2105-8-299 (2007).

5. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13**, 328, https://doi.org/10.1186/1471-2105-13-328 (2012).
6. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255, https://doi.org/10.1126/science.1087447 (2003).
7. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* **4**, Article17, https://doi.org/10.2202/1544-6115.1128 (2005).
8. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601–620, https://doi.org/10.1089/106652700750050961 (2000).
9. Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. & Young, R. A. Combining location and expression data for principled discovery of genetic regulatory network models. *Pac Symp Biocomput*, 437–449 (2002).
10. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17**(Suppl 1), S215–224 (2001).
11. Werhli, A. V. & Husmeier, D. Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol* **6**, Article15, https://doi.org/10.2202/1544-6115.1282 (2007).
12. Dobra, A. *et al*. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212 (2004).
13. Schafer, J. & Strimmer, K. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764, https://doi.org/10.1093/bioinformatics/bti062 (2005).
14. Toh, H. & Horimoto, K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* **18**, 287–297 (2002).
15. Yin, J. & Li, H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Ann Appl Stat* **5**, 2630–2650, https://doi.org/10.1214/11-AOAS494 (2011).
16. Jansen, R. C. & Nap, J. P. Genetical genomics: the added value from segregation. *Trends Genet* **17**, 388–391 (2001).
17. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585, https://doi.org/10.1038/ng.2653 (2013).
18. Schadt, E. E. *et al*. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302, https://doi.org/10.1038/nature01434 (2003).
19. Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. *Nat Rev Genet* **12**, 277–282, https://doi.org/10.1038/nrg2969 (2011).
20. Bollen, K. A. *Structural Equations with Latent Variables*. (John Wiley & Sons, Incorporated, 1989).
21. Xiong, M., Li, J. & Fang, X. Identification of genetic networks. *Genetics* **166**, 1037–1052 (2004).
22. Liu, B., de la Fuente, A. & Hoeschele, I. Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* **178**, 1763–1776, https://doi.org/10.1534/genetics.107.080069 (2008).
23. Logsdon, B. A. & Mezey, J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Comput Biol* **6**, e1001014, https://doi.org/10.1371/journal.pcbi.1001014 (2010).
24. Cai, X., Bazerque, J. A. & Giannakis, G. B. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Comput Biol* **9**, e1003068, https://doi.org/10.1371/journal.pcbi.1003068 (2013).
25. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429, https://doi.org/10.1198/016214506000000735 (2006).
26. Chen, C., Ren, M., Zhang, M. & Zhang, D. A two-stage penalized least squares method for constructing large systems of structural equations. *Journal of Machine Learning Research* **19,** 40–73 (2018).
27. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* **102**, 1572–1577, https://doi.org/10.1073/pnas.0408709102 (2005).
28. Balakrishnan, R. *et al*. YeastMine–an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford)* **2012**, bar062, https://doi.org/10.1093/database/bar062 (2012).
29. Szklarczyk, D. *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**, D362–D368, https://doi.org/10.1093/nar/gkw937 (2017).
30. Chatr-Aryamontri, A. *et al*. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**, D470–478, https://doi.org/10.1093/nar/gku1204 (2015).
31. Santiago, T. C. & Mamoun, C. B. Genome expression analysis in yeast reveals novel transcriptional regulation by inositol and choline and new regulatory functions for Opi1p, Ino2p, and Ino4p. *J Biol Chem* **278**, 38723–38730, https://doi.org/10.1074/jbc.M303008200 (2003).
32. Henry, S. A., Gaspar, M. L. & Jesch, S. A. The response to inositol: regulation of glycerolipid metabolism and stress response signaling in yeast. *Chem Phys Lipids* **180**, 23–43, https://doi.org/10.1016/j.chemphyslip.2013.12.013 (2014).
33. Schlatter, I. D. *et al*. MHO1, an evolutionarily conserved gene, is synthetic lethal with PLC1; Mho1p has a role in invasive growth. *PLoS One* **7**, e32501, https://doi.org/10.1371/journal.pone.0032501 (2012).
34. Choi, H. S. & Carman, G. M. Respiratory deficiency mediates the regulation of CHO1-encoded phosphatidylserine synthase by mRNA stability in Saccharomyces cerevisiae. *J Biol Chem* **282**, 31217–31227, https://doi.org/10.1074/jbc.M705098200 (2007).
35. Carman, G. M. & Han, G. S. Regulation of phospholipid synthesis in the yeast Saccharomyces cerevisiae. *Annu Rev Biochem* **80**, 859–883, https://doi.org/10.1146/annurev-biochem-060409-092229 (2011).
36. Loewen, C. J. *et al*. Phospholipid metabolism regulated by a transcription factor sensing phosphatidic acid. *Science* **304**, 1644–1647, https://doi.org/10.1126/science.1096083 (2004).
37. Dvir, S. *et al*. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci USA* **110**, E2792–2801, https://doi.org/10.1073/pnas.1222534110 (2013).
38. Hibbs, M. A. *et al*. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23**, 2692–2699, https://doi.org/10.1093/bioinformatics/btm403 (2007).
39. Sheppard, T. K. *et al*. The Saccharomyces genome database variant viewer. *Nucleic Acids Res* **44**, D698–702, https://doi.org/10.1093/nar/gkv1250 (2016).
40. Segal, E. *et al*. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**, 166–176, https://doi.org/10.1038/ng1165 (2003).
41. R package 'corrplot': Visualization of a correlation matrix (Version 0.80) (2016).
42. Livneh, I., Cohen-Kaplan, V., Cohen-Rosenzweig, C., Avni, N. & Ciechanover, A. The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. *Cell Res* **26**, 869–885, https://doi.org/10.1038/cr.2016.86 (2016).
43. Szlanka, T. *et al*. Deletion of proteasomal subunit S5a/Rpn10/p54 causes lethality, multiple mitotic defects and overexpression of proteasomal genes in Drosophila melanogaster. *J Cell Sci* **116**, 1023–1033 (2003).
44. Lundgren, J., Masson, P., Realini, C. A. & Young, P. Use of RNA interference and complementation to study the function of the Drosophila and human 26S proteasome subunit S13. *Mol Cell Biol* **23**, 5320–5330 (2003).
45. Wojcik, C. & DeMartino, G. N. Analysis of Drosophila 26 S proteasome using RNA interference. *J Biol Chem* **277**, 6188–6197, https://doi.org/10.1074/jbc.M109996200 (2002).
46. Lundgren, J., Masson, P., Mirzaei, Z. & Young, P. Identification and characterization of a Drosophila proteasome regulatory network. *Mol Cell Biol* **25**, 4662–4675, https://doi.org/10.1128/MCB.25.11.4662-4675.2005 (2005).
47. Schmidt, M. & Finley, D. Regulation of proteasome activity in health and disease. *Biochim Biophys Acta* **1843**, 13–25, https://doi.org/10.1016/j.bbamcr.2013.08.012 (2014).

48. Mannhaupt, G., Schnall, R., Karpov, V., Vetter, I. & Feldmann, H. Rpn4p acts as a transcription factor by binding to PACE, a nonamer box found upstream of 26S proteasomal and other genes in yeast. *FEBS Lett* **450**, 27–34 (1999).
49. Ronald, J., Brem, R. B., Whittle, J. & Kruglyak, L. Local regulatory variation in Saccharomyces cerevisiae. *PLoS Genet* **1**, e25, https://doi.org/10.1371/journal.pgen.0010025 (2005).
50. Teixeira, M. C. *et al.* The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Res* **42**, D161–166, https://doi.org/10.1093/nar/gkt1015 (2014).
51. Breunig, J. S., Hackett, S. R., Rabinowitz, J. D. & Kruglyak, L. Genetic basis of metabolome variation in yeast. *PLoS Genet* **10**, e1004142, https://doi.org/10.1371/journal.pgen.1004142 (2014).
52. Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T. L. & Kruglyak, L. Finding the sources of missing heritability in a yeast cross. *Nature* **494**, 234–237, https://doi.org/10.1038/nature11867 (2013).
53. Cubillos, F. A. *et al.* High-resolution mapping of complex traits with a four-parent advanced intercross yeast population. *Genetics* **195**, 1141–1155, https://doi.org/10.1534/genetics.113.155515 (2013).
54. Cubillos, F. A. *et al.* Identification of Nitrogen Consumption Genetic Variants in Yeast Through QTL Mapping and Bulk Segregant RNA-Seq Analyses. *G3 (Bethesda)* **7**, 1693–1705, https://doi.org/10.1534/g3.117.042127 (2017).
55. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207–210 (2002).
56. Wu, H., Kerr, M. K., Cui, X. & Churchill, G. A. In *The Analysis of Gene Expression Data: Methods and Software* (eds Giovanni Parmigiani, Elizabeth S. Garrett, Rafael A. Irizarry, & Scott L. Zeger) 313–341 (Springer New York, 2003).

Acknowledgements

We would like to thank Drs. Rachel Brem and Leonid Kruglyak for providing the data and answering questions related to the data. TH was funded by a phase I and II grant from the Purdue University Center for Cancer Research (NIH grant P30CA023168). This project was partially supported by NSF grant IIS-0844945 to DZ, a Purdue University internal equipment program grant to DZ and MZ, and a grant from the Mildred Elizabeth Edmundson Research Grant of Women's Global Health Institute at Purdue University and Indiana CTSI to MZ. The authors gratefully acknowledge the support of the Cancer Care Engineering (CCE) project, a joint effort between the Oncological Sciences Center (Purdue Center for Cancer Research, NCI P30CA23168) in the Purdue University Discovery Park and the Indiana University Melvin and Bren Simon Cancer Center (NCI P30CA082709). Support for the CCE project is gratefully acknowledged from the Walther Cancer Foundation, NIH (UL1RR025761), DOD (USAMRMC (CDMRP) W81XWH-008-1-0065, 9107003) and the Regenstrief Foundation. Publication of this article was funded in part by Purdue University Libraries Open Access Publishing Fund.

Author Contributions

D.Z. and M.Z. conceived and designed the study. C.C. carried out the data analysis under the supervision of D.Z. and M.Z. T.H. contributed to the interpretation of the results from yeast data analysis. All authors wrote and revised the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-37667-4.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.