nature genetics

Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks

Jun Zhu[1], Bin Zhang[1], Erin N Smith[2,3], Becky Drees[4], Rachel B Brem[5], Leonid Kruglyak[2], Roger E Bumgarner[4] & Eric E Schadt[1]

A key goal of biology is to construct networks that predict complex system behavior. We combine multiple types of molecular data, including genotypic, expression, transcription factor binding site (TFBS), and protein–protein interaction (PPI) data previously generated from a number of yeast experiments, in order to reconstruct causal gene networks. Networks based on different types of data are compared using metrics devised to assess the predictive power of a network. We show that a network reconstructed by integrating genotypic, TFBS and PPI data is the most predictive. This network is used to predict causal regulators responsible for hot spots of gene expression activity in a segregating yeast population. We also show that the network can elucidate the mechanisms by which causal regulators give rise to larger-scale changes in gene expression activity. We then prospectively validate predictions, providing direct experimental evidence that predictive networks can be constructed by integrating multiple, appropriate data types.

Large-scale genetic, transcriptomic, proteomic and metabolomic datasets have enabled researchers to decipher the biological function of individual genes, pathways, and, more generally, biological networks that drive complex phenotypes. However, the progress toward uncovering the mechanisms by which these genes lead to complex phenotypes has progressed at a slower rate. More recently, significant progress has been made by integrating multiple sources of data sampled from human and experimental populations to reconstruct networks that are predictive of complex phenotypes. A number of studies in a variety of species have demonstrated that predictive networks can be built by leveraging naturally occurring DNA variation to determine how such variation influences gene expression and other molecular phenotypes. By examining the effects of naturally occurring DNA variation on gene expression in segregating populations, other phenotypes can be examined with respect to these same DNA variations and ordered relative to the genes to infer causality[1–4]. Network reconstructions based on protein–protein interaction data[5], metabolomic data[6] and literature data[7] are also now becoming more routine. The common theme among these reconstruction efforts is the organization of vast amounts of molecular data into networks that capture fundamental properties of complex systems in states that give rise to complex phenotypes.

Although advances in the application of network reconstruction algorithms to high-dimensional biological data are being applied to a number of distinct data types, such as protein–protein interaction data[5], metabolomic data[6] and published gene-gene relationship data[7], no systematic studies have exploited the advantages that can be gained by combining genotypic, gene expression, protein–protein interaction and DNA–protein binding data to reconstruct whole gene networks. Although comprehensive forms of these types of data are very rare at present, yeast is one such system in which all of these data types currently exist. Here, we apply recent advances in coexpression and Bayesian network reconstruction methods to large-scale yeast datasets in order to create yeast gene networks capable of predicting complex system behavior. Specifically, we combine multiple types of large-scale molecular data, including genotypic, gene expression, TFBS and PPI data that were previously generated from a number of yeast experiments, to reconstruct causal, probabilistic gene networks. We demonstrate that the integration of these diverse data types in the context of a genetic cross[8] enhances the predictive power of the resulting gene networks beyond what could be achieved by reconstructing gene networks on the basis of expression data alone. We show that the functional subnetworks represented in the more integrated networks are significantly enriched for genes under the control of common genetic loci (that is, expression quantitative trait loci (eQTL) hot spots[9]). We also demonstrate that such networks can lead directly to the identification of the causal regulators for these different functional subnetworks, with many of the predictions providing a putative mechanistic understanding of how the causal regulators give rise to larger-scale changes in gene expression activity.

[1]Rosetta Inpharmatics, LLC, Seattle, Washington 98109, USA. [2]Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Princeton University, Carl Icahn Laboratory, Princeton, New Jersey 08544, USA. [3]Department of Molecular and Cellular Biology, Box 357275, University of Washington, Seattle, Washington 98195, USA. [4]Department of Microbiology, Box 358070, University of Washington, Seattle, Washington 98195,USA. [5]Department of Molecular and Cell Biology, 304A Stanley Hall #3220, University of California, Berkeley, Berkeley, California 94720, USA. Correspondence should be addressed to E.E.S. (eric_schadt@merck.com).

Received 11 September 2007; accepted 14 April 2008; published online 15 June 2008; doi:10.1038/ng.167

Table 1 Network modules identified from the yeast cross are enriched for GO categories (columns 3–6) and eQTL hot spots (columns 7–10)

Module color[a]	Module size	GO category type[b]	GO category	GO category size (overlap)	GO enrichment nominal *P* value[c]	Chr.	Within-chr. genome coordinate	eQTL hot spot size (overlap)	eQTL enrichment nominal *P* value[c]
Turquoise	1,208	BP	Cytoplasm organization and biogenesis	169 (153)	7.47×10^{-58}	2	550,000	186 (144)	2.87×10^{-39}
Blue	369	BP	Organic acid metabolism	235 (94)	4.44×10^{-34}	3	70,000	50 (46)	1.61×10^{-42}
Brown	290	BP	Protein biosynthesis	292 (98)	2.63×10^{-41}	14	450,000	206 (107)	2.92×10^{-69}
Yellow	282	BP	Generation of precursor metabolites and energy	258 (46)	2.09×10^{-8}	15	170,000	182 (134)	6.06×10^{-122}
Green	84	MF	Transferase activity	428 (20)	0.0012	2	570,000	25 (5)	0.00021
Red	83	BP	Generation of precursor metabolites and energy	168 (44)	6.54×10^{-39}	15	570,000	25 (21)	2.28×10^{-32}
Black	44	BP	Lipid metabolism	149 (20)	3.31×10^{-17}	12	650,000	52 (34)	2.37×10^{-60}
Pink	43	BP	Intracellular transport	275 (14)	1.41×10^{-6}	14	450,000	206 (19)	1.92×10^{-13}
Magenta	39	MF	RNA binding	140 (18)	3.21×10^{-16}	8	90,000	31 (4)	0.00028
Purple	37	BP	Chromosome organization and biogenesis (sensu Eukaryota)	200 (7)	0.0033	12	1,050,000	38 (31)	9.07×10^{-64}
Green-yellow	31	CC	Endoplasmic reticulum	213 (15)	2.39×10^{-11}	12	670,000	68 (5)	0.00022
Tan	29	BP	Response to chemical stimulus	153 (3)	0.12	5	110,000	24 (13)	4.84×10^{-23}
Cyan	27	BP	Response to chemical stimulus	153 (9)	7.55×10^{-7}	3	210,000	33 (23)	5.26×10^{-56}
Salmon	27	BP	Biopolymer catabolism	140 (13)	2.71×10^{-12}	12	670,000	68 (2)	0.089
Midnight blue	23	BP	Reproduction	160 (15)	7.64×10^{-16}	8	110,000	38 (23)	4.66×10^{-50}

[a]Module colors for the yeast coexpression network correspond to the modules identified in **Supplementary Figure 1**. [b]BP, biological process; MF, molecular function; CC, cellular component. [c]Nominal *P* values represent the significance of the Fisher's exact test statistic under the null hypothesis that the frequency of the indicated gene set is the same between a reference set of all genes comprising the coexpression network (3,662 genes) and the set of genes comprising the network module. The column 6 *P* value corresponds to Gene Ontology gene sets for the functional categories indicated. Given that 75 GO categories were tested, we set a *P*-value threshold of 0.05/75 = 0.00067 (Bonferroni-adjusted threshold) for significance. The column 9 *P* value corresponds to the eQTL hot spot gene sets. Given that 23 eQTL hot spots were tested, we set a *P*-value threshold of 0.05/23 = 0.0022 (Bonferroni-adjusted threshold) for significance.

Finally, we prospectively test and experimentally validate a number of these predictions, providing direct experimental evidence that predictive networks can be constructed via the integration of multiple, appropriate data types.

RESULTS

We assembled genotypic and expression data from 112 segregants obtained from a yeast cross between the BY and RM strains of *Saccharomyces cerevisiae* (referred to here as the BXR cross)[8]. Of the 5,740 genes represented on the microarrays used in this study, 5,180 were supported as having been sampled from a normal distribution, thus satisfying important assumptions about the use of the Pearson correlation statistic to reconstruct coexpression networks. From this set, we selected 3,662 informative genes for the construction of Bayesian networks for further network analysis (**Supplementary Methods** online). We also gathered transcription factor binding site (TFBS) data derived from multiple sources[10,11] and protein–protein interaction (PPI) data by combining the Saccharomyces Genome Database and Database of Interacting Proteins yeast databases to improve the structured priors for the Bayesian network reconstructions, as detailed below. From these data sources, we constructed four networks—one coexpression network and three different Bayesian networks—each of which took advantage of increasing amounts of data. Coexpression networks represent correlations among expression traits, whereas probabilistic, causal Bayesian networks based on integrative methods we have developed previously[4,12] represent causal relationships among genes. We test the predictive power of different Bayesian networks constructed from increasing amounts of data by comparing predictions made from the networks to experiments that were independent of the experiments used to construct the networks as well as to experiments we carried out prospectively to specifically test the predictions.

The yeast coexpression network

We reconstructed the coexpression network for the set of 3,662 informative genes identified in the BXR cross using a previously described weighted coexpression network algorithm[13]. A number of studies have demonstrated previously that coexpression networks are both scale free and modular[2,14], thus highlighting functional components of the network that are often associated with specific biological processes. Therefore, to identify modules comprised of highly interconnected expression traits within the coexpression network, we examined the topological overlap matrix[15] associated with this network. **Supplementary Figure 1** online depicts a hierarchically clustered topological overlap map in which the most highly interconnected modules in the network are readily identified. To identify gene modules (subnetworks) formally from the topological overlap map, we used a previously described algorithm that ensures that genes in any given module are maximally interconnected relative to genes in other modules (**Supplementary Methods**)[2].

From the BXR topological overlap map (**Supplementary Fig. 1**), 15 modules were identified. We tested each of these modules for gene enrichment using the yeast gene ontology (GO) categories for biological processes, molecular functions and cellular components. **Table 1** lists the most significantly enriched GO category for each module. Thirteen modules were significantly enriched for at least one GO category, indicating that the BXR coexpression network is organized into functional units. For example, 20 of the 149 genes annotated as belonging to the lipid metabolism GO biological process category fell into the black module, comprised of only 44 genes (Fisher's exact test *P* value = 3.31×10^{-17}).

To assess whether modules in the yeast coexpression network are enriched for genes controlled by common genetic loci, we carried out a genome-wide linkage scan on each of the expression traits to map eQTL for each of the expression traits. After partitioning the yeast

Table 2 Network modules identified from the yeast cross are enriched for genes that bind common transcription factors (columns 3–5) and for gene expression changes induced by specific gene knockout strains or chemical compounds (columns 6–8)

Module color[a]	Module size	Transcription factor	TFBS target gene set size (overlap)	TFBS enrichment nominal *P* value[b]	KO gene or compound	KO signature size (overlap)	KO enrichment nominal *P* value[b]
Turquoise	1,208	*FHL1*	94 (68)	3.01×10^{-15}	*CDC42*	553 (230)	0.0000026
Blue	369	*GCN4*	204 (108)	1.92×10^{-58}	*ANP1*	318 (119)	2.71×10^{-44}
Brown	290	*ABF1*	270 (34)	0.0037	*RML2*	395 (69)	2.46×10^{-11}
Yellow	282	*SKN7*	202 (37)	2.99×10^{-7}	*CKA2*	141 (55)	7.36×10^{-27}
Green	84	*REB1*	237 (14)	0.00082	*BUB1*	231 (39)	7.89×10^{-26}
Red	83	*HAP4*	66 (27)	5.01×10^{-29}	*DSK2*	671 (44)	5.83×10^{-13}
Black	44	*HAP1*	110 (26)	7.92×10^{-30}	TERBINAFINE	145 (16)	2.58×10^{-12}
Pink	43	*REB1*	237 (6)	0.056	*MTC7*	248 (8)	0.0072
Magenta	39	*OPI1*	42 (5)	0.000067	*SWI5*	48 (13)	3.71×10^{-16}
Purple	37	*YAP5*	79 (11)	1.26×10^{-10}	*SWI4*	560 (29)	1.36×10^{-17}
Green-yellow	31	*HSF1*	52 (3)	0.0092	*KAR2*	597 (20)	1.88×10^{-9}
Tan	29	*SPT2*	40 (3)	0.0036	*FAR1*	7 (4)	1.09×10^{-7}
Cyan	27	*MCM1*	61 (7)	1.71×10^{-7}	*STE12*	51 (8)	1.45×10^{-9}
Salmon	27	*RPN4*	64 (8)	9.51×10^{-9}	*UBR2*	29 (4)	0.000049
Midnight blue	23	*DIG1*	196 (18)	1.61×10^{-19}	*DIG1, DIG2*	331 (20)	1.07×10^{-18}

[a]Module colors for the yeast coexpression network correspond to the modules identified in **Supplementary Figure 1**. [b]Nominal *P* values represent the significance of the Fisher's exact test statistic under the null hypothesis that the frequency of the indicated gene set is the same between a reference set of all genes comprising the coexpression network (3,662 genes) and the set of genes comprising the network module. The column 5 *P* value corresponds to transcription factor target gene sets, and the column 8 *P* value corresponds to the knockout or compound signature gene sets.

genome into 603 bins, each 20 kb in length, we identified 23 bins in chromosomes 1, 2, 3, 5, 8, 12, 13, 14 and 15 that contained at least 15 eQTLs each (defined as eQTL hot spots), 22 of which were identical to the previously reported results from this dataset[16]. As shown in **Table 1**, all but one of the 15 modules in the coexpression network are significantly enriched for linking to at least one bin, thus suggesting extensive pleiotropic effects of QTLs on expression traits in the different bins, given each bin represents at least one QTL affecting multiple gene expression traits. These results also suggest that much of the correlation structure in the different modules is driven by common genetic loci representing perturbations on a particular part of the coexpression network, as we have shown in mouse[17]. For example, there are 52 genes linked to the bin located on chromosome 12 between base-pair positions 640,000 and 660,000, and 34 of these genes are located in the black module (Fisher's exact test *P* value = 2.37×10^{-60}). Therefore, the joint analysis of the genotypic and coexpression data highlights that common genetic perturbations in this cross affect the expression activity of subnetworks of genes, which in turn affect the activity of important biological processes. For all subsequent analyses, we merged the 23 eQTL-enriched bins into the 13 previously reported eQTL hot-spot regions for the BXR cross[16].

The colocalization of entire network modules to common genetic loci suggests that the coherence achieved within these modules is at least partially driven by genes in the modules that are more directly under the control of common factors. Although the eQTL hot spots were originally reported as not enriched for linking to loci harboring transcription factors (TF)[16], the eQTL hot spots could be driven by genes that are not TFs, but that affect TF activity. We found that 14 of the 15 modules in the coexpression network were significantly enriched (all Fisher's exact test *P* value <0.01) for at least one TF target gene set (**Table 2**). For example, 26 of 44 genes in the black module are annotated as having a Hap1 binding site (Fisher's exact test *P* value = 7.92×10^{-30}). Notably, the gene encoding the Hap1 TF physically resides in the chromosome 12 eQTL hot-spot region and has a strong *cis* eQTL linked to this same region.

Naturally occurring DNA variations that influence expression traits provide one way to test for coherence in the coexpression network modules. However, directed single-gene perturbation experiments can also be leveraged for this purpose. We generated knockout signatures from a previously published yeast compendium dataset[18] and found that all of the network modules were significantly enriched for at least one knockout signature gene set (**Table 2**). For example, the midnight blue module depicted in **Supplementary Figure 1** is comprised of 23 genes, most of which function in yeast mating; 20 of these genes overlap with the signature from the double knockout of *DIG1* and *DIG2* (Fisher's exact test *P* value = 1.07×10^{-18}), which are repressors of pheromone responsive transcription. These data not only demonstrate how targeted gene perturbations can affect entire subnetworks, but they also suggest that when genes in a highly connected module are perturbed, these genes tend to move a significant proportion of the module, reflecting a higher degree of causal connectivity among highly interconnected genes than has been previously appreciated[17].

Inferring clique communities from PPI networks

Protein–protein interaction data provide a complementary view of molecular interactions at play in living systems. Unlike in coexpression networks, edges in the PPI networks represent putative physical interactions between two proteins. Combination of the SGD and DIP yeast PPI data yields a network comprised of 4,833 nodes (distinct proteins) and 15,345 edges between these nodes. To compare the PPI and coexpression networks, we compared the correlation distribution of gene pairs in the coexpression network that were also linked in the PPI network with the correlation distribution of an equal number of randomly selected gene pairs from the coexpression network. Although the correlation distribution derived from the PPI-linked genes is shifted slightly higher compared to randomly selected gene expression trait pairs, most pairs linked by PPI are not highly correlated (**Supplementary Fig. 2** online).

The lack of correlation between gene expression trait pairs that correspond to genes connected in the PPI network may be because the different data types reflect different domains of information. The lack of correlation may also be due to a high false-positive rate in the PPI data, given that recent studies have explicitly demonstrated that PPI data generated from high-throughput experiments are not very

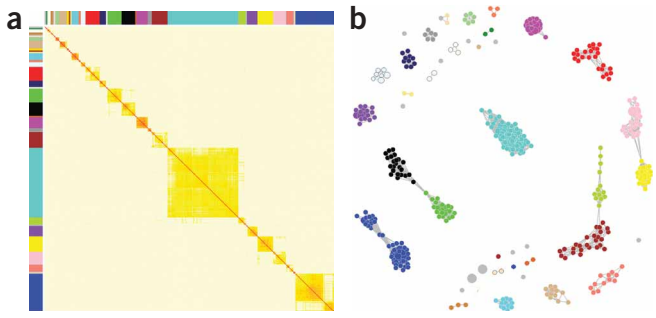

Figure 1 A generic approach to identifying clique communities in the PPI network. (**a**) Hierarchical clustering over the clique-clique similarity matrix heat map derived from a network of the 492 k-cliques with $k \geq 5$. Cliques in the rows and columns are sorted by an agglomerative hierarchical clustering algorithm. The clique network clearly displays strong modularity under hierarchical organization. Each of the colored bars along the top horizontal and left vertical axes represents a network module. (**b**) The clique community network. Each node represents a clique and each link indicates that the two connected cliques have a similarity greater than 0.5 on the basis of the Tanimoto measure. Nodes that do not belong to any module are colored gray. Of the clique communities represented in this plot, 74% overlap the set of stable protein complexes.

specific[19]. To identify higher-confidence interactions in the PPI data that may better inform the coexpression network, we used a method to explore the global structural organization of the PPI network in terms of overlapping local network neighborhoods, referred to here as 'clique communities' (**Supplementary Methods**)[20].

Applying this method to the yeast PPI data, we identified 492 cliques comprised of 5 or more proteins. A total of 112 clique communities representing 2,477 unique links in the PPI network were identified from these 492 cliques (**Fig. 1**). Of the 4,833 proteins in the PPI network, only 840 (~17%) are represented in the 112 highly interconnected clique communities. Compared with the full PPI network, a larger portion of the gene pairs connected in the PPI clique communities are also connected in the coexpression network (**Supplementary Fig. 2**). We compared the clique communities with manually curated stable protein complexes in MIPS[21]. Of the 74 protein complexes consisting of 5 or more proteins in this set, 72% (53/74) of them significantly overlapped with the clique communities, and 74% of the clique communities significantly overlapped the set of stable protein complexes. The clique communities were generally significantly enriched for genes involved in protein complexes, with 330 genes involved in protein complexes (out of 674 represented in the PPI network) overlapping the set of genes comprising the clique communities (a 3.5-fold enrichment; $P = 1.54 \times 10^{-96}$). The increased overlap between the PPI and coexpression networks likely reflects the higher confidence links represented in the clique communities of the PPI network, as well as the biological need for proteins that are in complexes to be regulated to similar levels.

Our findings are consistent with previous reports[22,23]. However, whereas these previous reports focused on known protein complexes[23] and establishing associations between gene expression clusters and protein interaction clusters[22], here we have systematically uncovered known and unknown complexes from the PPI data. Because PPI data can be noisy, the clique-community analysis serves as a filter to uncover not only the underlying building blocks (cliques) of the PPI network, but also their high-level organization (communities)[20].

Reconstructing probabilistic causal networks in yeast

Coexpression networks highlight how biological networks are organized into functional modules that are under the control of common genetic loci and transcription factors, as well as how DNA variations at specific loci perturb these network modules and in turn induce changes in higher-order phenotypes. Despite these and other advantages, coexpression networks do not provide explicit details on the connectivity structure among genes in the network and do not represent causal associations. A number of efforts have sought to integrate different data types, such as gene expression, genotype, PPI, TFBS and literature data, using a Bayesian approach[24–26]. However, the networks resulting from such efforts are still comprised of modules from which causal information or detailed mechanisms at the gene level are not easily derived. However, both simulation and experimental results have demonstrated that Bayesian networks reconstructed by incorporating genetic data lead to predictive networks[12,27]. Here we extend this approach by integrating diverse data types, including gene expression, genotype, TFBS and PPI data.

We used a Bayesian network approach to construct three networks: (i) a Bayesian network based on expression data alone (BN_{raw}), (ii) a Bayesian network based on expression and eQTL data (BN_{qtl}) and (iii) a Bayesian network based on expression, eQTL, TFBS and PPI data (BN_{full}). BN_{qtl} was constructed by incorporating eQTL information from the BXR cross as prior evidence that two genes are causally related (**Supplementary Methods**). To complement the use of the eQTL data, we constructed BN_{full} by incorporating TFBS data derived from high-quality ChIP-chip experiments, phylogenetic conservation[10], protein complex data and eQTL data. Manually curated protein complexes[21] and complexes identified by the clique-community analysis described above for the PPI network were leveraged to enhance the TFBS target gene sets[28]. If at least half of the genes in a protein complex carried a given TFBS, then all genes in the complex were added to that TFBS gene set (**Supplementary Table 1** online). In the Bayesian network reconstruction process, this extended TFBS dataset was considered as prior evidence that two genes are causally related (**Supplementary Methods**).

Testing whether the Bayesian networks are predictive

There were 3,779, 3,712 and 3,645 links in BN_{raw}, BN_{qtl} and BN_{full}, respectively. All three Bayesian networks were highly similar with respect to edge overlap when the edge direction was not considered, demonstrating that the different networks captured the covariance structure for the expression traits to a similar degree. In fact, BN_{raw} and BN_{qtl} had 3,335 edges in common (~90% overlap) and BN_{qtl} and BN_{full} had 3,335 edges in common (~86% overlap). To test the relative power of these networks to predict system behavior, we examined whether the networks could recapitulate known biological processes. We tested whether the networks could predict the GO categories, whether the networks could predict genes regulated by different transcription factors and whether the networks could predict the expression responses to gene knockout signatures represented in the yeast compendium dataset[18].

First, we found that similar coherence with respect to GO categories was achieved in all networks. A total of 75 GO terms (46 GO biological processes, 16 GO cellular components and 13 GO molecular functions categories) were searched for enrichment in each of the Bayesian networks. We found that 26, 27 and 22 signature sets were significantly enriched (permutation P value <0.01) in BN_{raw}, BN_{qtl} and BN_{full}, respectively (**Supplementary Table 2** online). These results demonstrate that each of the Bayesian networks captured a significant proportion of the associations among genes known to operate in

Table 3 Causal regulators identified in the original publication on the BXR cross and predicted using the different Bayesian networks described in the main text

				Bayesian network		
eQTL hot spot	Hot spot chr.	Hot spot base-pair position	Yvert *et al.* predictions[16]	BN_{raw}	BN_{qtl}	BN_{full}
1	2	390,000	None predicted	*AGP2*	None predicted	None predicted
2	2	560,000	*AMN1, MAK5*	*AMN1*	*AMN1*	*TBS1, TOS1, ARA1, CSH1, SUP45, CNS1, AMN1*
3	2	710,000	None predicted	None predicted	None predicted	None predicted
4	3	100,000	*LEU2*	*CIT2, MATALPHA2*	*LEU2, MATALPHA1, CIT2*	*LEU2, ILV6, NFS1, CIT2, MATALPHA1*
5	3	230,000	*MATALPHA1*	None predicted	*MATALPHA1*	*MATALPHA1*
6	5	130,000	*URA3*	None predicted	*URA3*	*URA3*
7	8	130,000	*GPA1*	*ARN2, SPO11*	*GPA1*	*GPA1*
8	12	680,000	*HAP1*	None predicted	*HAP1*	*HAP1*
9	12	1,070,000	*SIR3*	*YRF1-4*	*YRF1-4, YRF1-5*	*YRF1-4, YRF1-5, YLR464W*
10	13	70,000	None predicted	*SMA2*	*SMA2*	None predicted
11	14	503,000	None predicted	*TOP2*	*SAL1, TOP2*	*SAL1, TOP2*
12	15	180,000	None predicted	*NDJ1*	*PHM7, HAL9, SKM1*	*PHM7*
13	15	590,000	*CAT5*	None predicted	None predicted	None predicted

The 'none predicted' designation indicates that no causal regulator could be identified for the indicated hot spot.

common pathways. The GO categories represent the association-based relationships between genes rather than the cause-effect relationships reflected in gene-specific perturbation data such as the yeast compendium data[18]. Therefore, this result suggests that the eQTL and TFBS data contribute mainly to making causal inferences.

Second, we found that BN_{qtl} and BN_{full} predict TF targets significantly better than BN_{raw}. We identified 15, 19 and 30 TF target sets that were significantly enriched in BN_{raw}, BN_{qtl} and BN_{full}, respectively (**Supplementary Table 2**). Because BN_{full} was constructed using the TFBS data as priors, it is unfair to compare this network to the other two networks. However, the extent of enrichment between BN_{qtl} and BN_{raw} was significantly different (Wilcoxon signed-rank test $P = 0.0002$), indicating that eQTLs enhance the power of the Bayesian networks to infer causal associations.

Third, we found that BN_{qtl} and BN_{full} predict knockout signatures better than BN_{raw}. One of the true tests of a causal network is the ability to predict what genes will change in response to a specific perturbation. Single and double gene perturbation experiments enable testing the predictive power of a network in this way. We used a previously published yeast knockout compendium[18] consisting of 300 expression profiles from 287 knockout strains and 13 chemical perturbation experiments to carry out this test for each of the Bayesian networks. We found that 92, 111 and 116 signature sets were significantly enriched (permutation P value <0.01) in BN_{raw}, BN_{qtl} and BN_{full}, respectively (**Supplementary Table 2**). In addition, the significance values of the enrichments for BN_{qtl} and BN_{full} were much greater than that for BN_{raw} (Wilcoxon signed-rank test $P = 1.09 \times 10^{-5}$; **Supplementary Fig. 3** online).

Dissecting eQTL hot spots using Bayesian networks

Like transgenics, gene knockouts and other artificial perturbations, eQTLs represent perturbations that affect gene expression traits. In some cases, a given QTL may have pleiotropic effects on a number of expression traits, leading to eQTL clusters that colocalize to a common genetic locus (known as an eQTL hot spot). From the BXR data used to reconstruct the Bayesian networks, we have previously identified 13 eQTL hot spots, with 9 putative regulators proposed for 8 of the hot spots that were based on genes with known biological functions and *cis* eQTLs that were coincident with these hot spots (**Table 3**)[16]. We identified all of the gene expression traits linked to each of the 13 hot spot regions and then searched each of these gene sets for enrichment in subnetwork structures in BN_{raw}, BN_{qtl} and BN_{full}, similar to the tests for enrichments carried out above. All but two small eQTL hot spots were enriched in subnetworks of BN_{raw}, BN_{qtl} and BN_{full} (**Supplementary Table 3** online).

One objective method for assessing the predictive power of each Bayesian network is to use the different networks to infer the causal regulators for each of the eQTL hot spots. To identify causal regulators for a given hot spot, we selected genes that gave rise to a putative *cis* eQTL in the corresponding eQTL hot spot region. For this set of putative regulators, we defined the signature for each regulator as the set of genes in the subnetwork that could be reached by the putative regulator following directed links throughout the entire network. The signature for each putative regulator was then intersected with the set of genes linked to the corresponding hot spot region. If the significance of the overlap was significant, we declared the putative regulator as a regulator of the hot spot and associated subnetwork.

We predicted the causal regulators for the eQTL hot spots represented in BN_{raw}, BN_{qtl} and BN_{full}. Causal regulators were inferred from BN_{raw} for seven of the eQTL hot spots, with two of these regulators matching those previously identified in the BXR cross[16]. However, ten causal regulators were inferred from BN_{qtl} and six of these matched regulators that were previously identified in this cross[16]. Causal regulators were also inferred from BN_{full} for nine eQTL hot spots. Again, six of the regulators matched those previously identified for this cross (**Table 3**). For eQTL hot spots 5 and 6, BN_{raw} failed to predict the correct regulators when the regulators were known, whereas BN_{qtl} and BN_{full} correctly identified the regulators in these instances (**Table 3**). These results support that although there is strong similarity between the different Bayesian networks, when we compare the extent of local network enrichments for genes operating in known pathways, the networks constructed by incorporating the eQTL, TFBS and PPI data as prior information on the relationships among pairs of genes have greater power to infer causal regulators for validated signature gene sets. We examined 9 of the 13 previously identified

eQTL hot spots[16] in the BXR cross for which a causal regulator could be identified in BN_{full}. The percentage of variance for each eQTL hot spot that can be explained by a causal regulator in BN_{full} is summarized in **Supplementary Table 4** online. The network was used to elucidate the mechanism driving the eQTL hot spot activity, and we used experimental methods to validate prospectively all of the new predictions.

In a number of cases, causal regulators for eQTL hot spots previously proposed and/or tested[16] were predicted by BN_{qtl} and BN_{full}, thereby serving as positive controls. For example, our analyses indicated that the mitotic exit regulator AMN1 is causal for the variation of a subset of transcripts linking to eQTL hot spot 2, which has been confirmed in previous work[16]. We also made detailed predictions regarding the mechanism of variation of additional transcripts linking to this hot spot (**Supplementary Data** online). Our analyses identified GPA1 as the causal regulator for hot spot 7 and the transcription factor HAP1 as the causal regulator for hypoxia-related transcripts in hot spot 8, both consistent with previous work[16]. We used previously reported microarray results to experimentally confirm the causal regulator roles predicted for these genes (**Supplementary Data**). It is important to note that for each of these regulators, hypotheses in previous studies were generated on the basis of manual curation, whereas our methods are data driven and objective, predicted by BN_{full}, which was constructed by integrating eQTL, expression, TFBS and PPI data.

We also examined hot spot 4, which encompasses *LEU2*, encoding the leucine biosynthetic enzyme. The RM parent of the BXR cross contains an engineered *LEU2* deletion, whereas BY is wild-type; previous analyses identified expression effects in the cross that are likely the result of this perturbation[16]. In keeping with this, in BN_{full} we observed a subnetwork enriched for amino acid biosynthesis pathways (enrichment $P = 1.70 \times 10^{-47}$) containing genes whose expression linked to the hot spot 4 region. *LEU2* was predicted as a causal regulator for this subnetwork, confirming as above that our inference methods correctly identify known biological effects in the cross (**Fig. 2a**).

To investigate the mechanism by which *LEU2* deletion causes expression changes, we first noted enrichment in the subnetwork of genes with binding sites for LEU3 (enrichment $P = 1.42 \times 10^{-8}$) and GCN4 ($P = 3.81 \times 10^{-12}$), consistent with the known roles of these transcription factors in regulation of amino acid biosynthesis genes. We also observed strong overlap between the subnetwork and the *GCN4* knockout gene expression signature (**Fig. 2b**; $P = 1.04 \times 10^{-75}$). We profiled the expression of a *LEU2* knockout strain and found strong overlap between this signature and the hot spot, as expected ($P = 4.91 \times 10^{-18}$); again, known GCN4 and LEU3 targets were enriched in the signature, suggesting that LEU3 and GCN4 are likely to be key mediators of the *LEU2* deletion effects. Notably, the amino acid biosynthesis gene *ILV6* also lies in the hot spot 4 region and its expression shows strong linkage in *cis*; in fact, our methods inferred *ILV6* as an additional key causal regulator for the hot spot (**Fig. 2a**). In particular, *ILV6* was predicted to be causal for *GCN4*. We profiled the expression of an *ILV6* knockout strain and found significant overlap with the hot spot 4 subnetwork ($P = 4.03 \times 10^{-52}$), including upregulation of *GCN4* and many of its targets (**Fig. 2b**). Our inference of *ILV6* as a major regulator can be explained by either of two models. Naturally occurring polymorphisms in *ILV6* and the engineered deletion of *LEU2* may together act as the genetic causes of expression variation linked to hot spot 4. Alternatively, the *LEU2* deletion may be the sole genetic difference responsible for the hot spot, causing expression changes in *ILV6* that in turn trigger downstream effects. In either case, our data confirm that variation in *ILV6* expression can influence expression of genes in the hot spot 4 subnetwork as predicted.

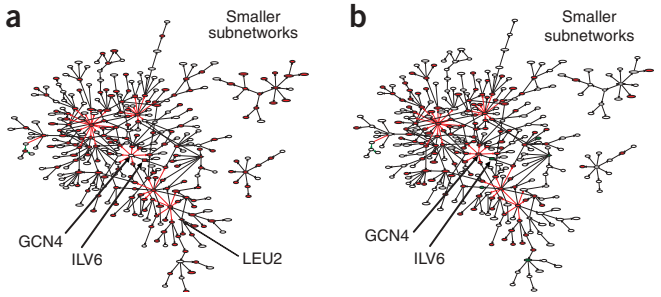

Figure 2 eQTL hot spot 4 subnetworks. (**a**) Subnetworks in BN_{full} enriched for expression traits linked to eQTL hot spot 4. Of the 3,662 genes comprising BN_{full}, 203 link to eQTL hot spot 4. There are 309 genes comprising the three subnetworks shown here, and 170 of these genes link to eQTL hot spot 4 (red nodes), a nearly tenfold enrichment over what was expected by chance (empirical $P < 10^{-8}$). LEU2 and ILV6 were identified as the primary causal regulators for the large subnetwork, as described in the text. ILV6 is supported as causal for GCN4 in the large subnetwork. (**b**) The *ILV6* knockout signature is enriched in the large eQTL hot spot 4 subnetwork. Of the 635 genes in the *ILV6* knockout signature, 432 were represented in BN_{full}, and 129 of these overlapped the large eQTL hot spot 4 subnetwork (colored nodes), a nearly fourfold enrichment over what was expected by chance (Fisher's exact test $P = 4.04 \times 10^{-52}$). The red and green nodes represent genes that are up- and downregulated, respectively, in the *ILV6* knockout signature (note that *GCN4* is upregulated).

We also analyzed eQTL hot spots with no previously identified causal regulators. Two BN_{full} subnetworks were found to be enriched for genes in hot spot 12 (see **Supplementary Fig. 4** online for the large subnetwork). The larger of the two subnetworks was enriched for carbohydrate metabolism ($P = 4.87 \times 10^{-14}$), whereas the smaller network was enriched for amino acid biosynthesis ($P = 1.56 \times 10^{-10}$). The large subnetwork is comprised of 452 genes, which were enriched for targets of the MSN2 (represented by PIL1 in the network) and SKN7 transcription factors, and also showed more modest enrichment of the targets of the transcription factors MSN4, which responds to stress; CIN5, which responds to salt stress; XBP1, which responds to starvation or stress; and REB1 (represented by ENP2). The main transcription factor for the small subnetwork is GCN4. The small subnetwork overlaps the hot spot 4 subnetwork just described. In this case, the amino acid biosynthesis processes are likely responding to stress rather than causing stress.

Our method predicted a single causal regulator for the large subnetwork: *PHM7*, a gene of unknown function that is regulated by phosphate levels[29]. A *cis*-acting polymorphism is known to affect *PHM7* expression between RM and BY[30], suggesting a model in which differential expression of *PHM7* in the BXR progeny drives expression variation of genes in hot spot 12. The MSN2 and MSN4 transcription factors are known to mediate the expression response to phosphate and to more general stressors[31], suggesting that variation in *PHM7* might influence expression of stress-related genes in the BXR cross through these factors. To test the role of *PHM7* in regulation of hot spot 12 genes, we deleted *PHM7* in the BY background and measured genome-wide expression via microarray, of this strain and wild-type BY, under phosphate-limited conditions. We found that 1,329 transcripts showed a significant difference between the two strains; this set

was enriched for targets of MSN2 ($P = 1.66 \times 10^{-4}$) and of the upstream regulator SOK2 ($P = 7.11 \times 10^{-4}$). Of the genes changing in response to the *PHM7* deletion, 817 are in BN_{full}, and 155 of those are in the hot spot 12 subnetwork (**Supplementary Fig. 4b**; enrichment $P = 2.70 \times 10^{-10}$), confirming that variation in *PHM7* can influence expression of genes in the subnetwork as predicted. We note that 65% of the genes in the hot spot 12 subnetwork are not in the *PHM7* signature, suggesting that *PHM7* may be one of multiple causal, polymorphic regulators. As we showed for eQTL hot spot 4 where multiple causal regulators may be at play, validation of *PHM7* as a causal regulator does not exclude the existence of other causal regulators.

Although our analysis well demonstrates the ability of expression-based networks to identify causal regulators with polymorphic transcription, such networks may fail to identify causal regulators with polymorphic protein function. A large BN_{full} subnetwork was found to be enriched for genes in hot spot 11 (**Supplementary Fig. 5** online). Our methods inferred the putative mitochondrial transporter SAL1 as the main causal regulator of this subnetwork. The BY strain bears a frameshift allele in *SAL1*, which truncates the protein product by 49 residues relative to that in RM and results in a loss of its function as a suppressor of a deletion of adenine nucleotide translocase[32]. To test the role of the *SAL1* polymorphism in expression variation in the BXR cross, we introduced the RM version of the *SAL1* coding region into the BY genome as a replacement at its own locus, and in parallel, we introduced the BY region into the RM background by the same method. We measured expression in these strains via microarray and computed expression changes for each gene between the engineered strains and their wild-type counterparts. The expression changes for genes not in the subnetwork were close to zero (mean = –0.0114, s.d. = 0.2326), whereas the expression changes for genes in the subnetwork were slightly larger (mean = 0.1209, s.d. = 0.2358), although not significantly so at the 0.05 level. However, the two-sample Kolmogorov-Smirnov test shows that the distributions of the two groups are significantly different ($P = 1.57 \times 10^{-22}$), suggesting that SAL1 may have a minor, causal regulatory role.

To expand our search for causal regulators in this case, we used a complementary approach to identify regulators that may not be detected by gene expression methods[3]. We examined genes located in the hot spot region that gave rise to *cis* eQTL and that harbored coding polymorphisms at highly conserved amino acids that induce known phenotypic changes in yeast (**Supplementary Data**). *MKT1* was the only gene in this hot spot giving rise to a *cis* eQTL and harboring a coding polymorphism for a highly conserved amino acid (D30G) previously shown to induce phenotypic changes in yeast[33,34]. Therefore, we used a previously generated strain carrying the RM version of the functional polymorphism (Gly30) in the BY genome (YAD350 (ref. 33)) and profiled expression changes in synthetic media to test whether *MKT1* was a causal regulator for this hot spot. The eQTL hot spot 11 subnetwork was significantly enriched for genes in the *MKT1* allele swap signature, with 698 of the genes in this signature overlapping the set of 3,662 genes used to construct the network, and 124 of these overlapping the subnetwork comprised of 317 genes (Fisher's exact test P value = 1.86×10^{-18}). This result validates MKT1 as a major causal regulator for hot spot 11.

DISCUSSION

Previous yeast network reconstructions have focused on a more limited number of genes in order to make the reconstruction tractable[24,28,35]. Our integrative analysis combined large-scale genotype, gene expression, PPI and TFBS data to construct networks comprised of more than 50% of the genes in the yeast genome using a novel Bayesian network reconstruction method. The relative utility of the resulting networks was highlighted by predicting responses to independent experimental perturbations and the known biology of the system. Specifically, we demonstrated that networks constructed by incorporating genetic, TFBS and PPI data were more predictive than a network constructed from expression data alone. Further, our method of integrating diverse data was also demonstrated to predict previously unknown interactions, which in turn led to the identification of genes that are not well annotated, but that nevertheless serve as causal regulators for eQTL hot spots.

The modules emerging from the coexpression network were shown to elucidate the functional relevance of the different components of the network. Of particular interest is our finding that the coexpression network overlapped poorly with the PPI network, suggesting that the PPI and coexpression data reflect complementary views of the system, that the PPI data generated via high-throughput experiments is not very specific[19] or a combination of the two. We were able to identify structures in the PPI network that overlapped well with the coexpression network only after we performed a clique-community analysis on the PPI network to define the core, highly interconnected substructures of this network. Through this analysis, we found that the overlapping structures were enriched for stable protein complexes, likely explaining the good correlation between the PPI clique communities and corresponding coexpression network modules.

The gold standard for assessing the predictive power of any network model is prospectively validating predictions made from such a model. We queried the different Bayesian networks constructed using progressively more data (BN_{raw}, BN_{qtl} and BN_{full}) to predict the causal regulators of the subnetworks enriched for genes linked to the different eQTL hot spots in the BXR cross[16]. BN_{full} was demonstrated to be the most predictive network, and five of the predictions of previously unknown interactions made using BN_{full} were prospectively tested experimentally, and all of these predictions were validated, thus confirming the predictive power of the integrated network to elucidate the regulatory control of some of the subnetworks. These results are also consistent with a large-scale simulation study we conducted to assess the extent to which genetic information could improve the accuracy of Bayesian networks based on gene expression data in a segregating populations[27].

The integrative reconstructions carried out in our study represent only the beginning steps needed to construct large-scale, accurate whole-genome networks. A number of important limitations will need to be addressed to further enhance the accuracy of this type of network. First, the Bayesian network algorithm used in this study does not permit loops, making it difficult to represent some types of feedback, which are obviously an important control mechanism in any biological system. Second, Bayesian networks do not effectively represent time-series data[36]. These issues might be addressed by using dynamic Bayesian networks, which explicitly include a temporal representation of the interaction between nodes. Third, we reconstructed networks from a limited amount of data generated from a single population and under only a single biological condition. Given the impact genetic background and environment can have on network structure, with the connectivity structure of a network varying as a function of genetic background and environment, populations representing different genetic backgrounds in different environmental contexts will have to be studied to assess the impact on network structure. However, these and other issues notwithstanding, our results support that the construction of large-scale whole-gene networks based on genetic, gene expression, TFBS, PPI and related types

of large-scale data can lead to networks that are capable of predicting complex system behavior.

METHODS

Accession codes. NCBI GEO: all gene expression data generated for this study have been deposited under accession number GSE11111. The gene expression data for the BXR cross have been previously deposited into the GEO database under accession number GSE1990.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

Work at Princeton was supported by National Institute of Mental Health grant R37 MH059520 and a James S. McDonnell Foundation Centennial Fellowship to L.K., and Center grant P50GM071508 from the National Institute of General Medical Science to the Lewis-Sigler Institute. We thank J. Whittle for generating *GPA1* allele swap data, S. Iyer for performing some of the deletion strain validation experiments and A. Deutschbauer (Lawrence Berkeley National Laboratory) for providing us with the MKT1 allele swap strain YAD350. We would also like to thank R. Ireton for her careful reading and editing of this manuscript.

AUTHOR CONTRIBUTIONS

E.N.S., B.D., R.B.B. and L.K. constructed and characterized the genetically modified yeast strains. J.Z., B.Z. and E.E.S. carried out the coexpression and Bayesian network analyses and performed bioinformatic analyses. E.N.S., B.D., R.B.B., L.K. and R.E.B. aided in the data analysis. All authors were involved in the study design and interpretation of the experimental results, and discussed the results and commented on the manuscript. J.Z., B.Z. and E.E.S. designed the study, developed methods, analyzed the data and wrote the paper.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Kulp, D.C. & Jagalur, M. Causal inference of regulator-target pairs by gene mapping of expression phenotypes. *BMC Genomics* **7**, 125 (2006).
2. Lum, P.Y. *et al.* Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. *J. Neurochem.* **97** (Suppl. 1), 50–62 (2006).
3. Mehrabian, M. *et al.* Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat. Genet.* **37**, 1224–1233 (2005).
4. Schadt, E.E. *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37**, 710–717 (2005).
5. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Nielsen, J. & Oliver, S. The next wave in metabolome analysis. *Trends Biotechnol.* **23**, 544–546 (2005).
7. Rajagopalan, D. & Agarwal, P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics* **21**, 788–793 (2005).
8. 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 (2005).
9. Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
10. MacIsaac, K.D. *et al.* An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 113 (2006).
11. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
12. Zhu, J. *et al.* An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet. Genome Res.* **105**, 363–374 (2004).
13. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
14. Ghazalpour, A. *et al.* Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* **2**, e130 (2006).
15. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
16. Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* **35**, 57–64 (2003).
17. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
18. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
19. Deeds, E.J., Ashenberg, O. & Shakhnovich, E.I. A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci. USA* **103**, 311–316 (2006).
20. Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
21. Guldener, U. *et al.* MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.* **34**, D436–D441 (2006).
22. Ge, H., Liu, Z., Church, G.M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29**, 482–486 (2001).
23. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12**, 37–46 (2002).
24. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555–1558 (2004).
25. Lee, S.I., Pe'er, D., Dudley, A.M., Church, G.M. & Koller, D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl. Acad. Sci. USA* **103**, 14062–14067 (2006).
26. Workman, C.T. *et al.* A systems approach to mapping DNA damage response pathways. *Science* **312**, 1054–1059 (2006).
27. Zhu, J. *et al.* Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS Comput. Biol.* **3**, e69 (2007).
28. Tan, K., Shlomi, T., Feizi, H., Ideker, T. & Sharan, R. Transcriptional regulation of protein complexes within and across species. *Proc. Natl. Acad. Sci. USA* **104**, 1283–1288 (2007).
29. Ogawa, N., DeRisi, J. & Brown, P.O. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell* **11**, 4309–4321 (2000).
30. Ronald, J., Brem, R.B., Whittle, J. & Kruglyak, L. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet.* **1**, e25 (2005).
31. Giots, F., Donaton, M.C. & Thevelein, J.M. Inorganic phosphate is sensed by specific phosphate carriers and acts in concert with glucose as a nutrient signal for activation of the protein kinase A pathway in the yeast *Saccharomyces cerevisiae*. *Mol. Microbiol.* **47**, 1163–1181 (2003).
32. Chen, X.J. Sal1p, a calcium-dependent carrier protein that suppresses an essential cellular function associated with the Aac2 isoform of ADP/ATP translocase in *Saccharomyces cerevisiae*. *Genetics* **167**, 607–617 (2004).
33. Deutschbauer, A.M. & Davis, R.W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.* **37**, 1333–1340 (2005).
34. Sinha, H., Nicholson, B.P., Steinmetz, L.M. & McCusker, J.H. Complex genetic interactions in a quantitative trait locus. *PLoS Genet.* **2**, e13 (2006).
35. Pe'er, D., Regev, A., Elidan, G. & Friedman, N. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* **17** (Suppl 1), S215–S224 (2001).
36. Ong, I.M., Glasner, J.D. & Page, D. Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics* **18** (Suppl 1), S241–S248 (2002).