

# Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*

Jan Ihmels<sup>1,3</sup>, Ronen Levy<sup>1,3</sup> & Naama Barkai<sup>1,2</sup>

Cellular networks are subject to extensive regulation, which modifies the availability and efficiency of connections between components in response to external conditions. Thus far, studies of large-scale networks have focused on their connectivity, but have not considered how the modulation of this connectivity might also determine network properties. To address this issue, we analyzed how the coordinated expression of enzymes shapes the metabolic network of *Saccharomyces cerevisiae*. By integrating large-scale expression data with the structural description of the metabolic network, we systematically characterized the transcriptional regulation of metabolic pathways. The analysis revealed recurrent patterns, which may represent design principles of metabolic gene regulation. First, we find that transcription regulation biases metabolic flow toward linearity by coexpressing only distinct branches at metabolic branchpoints. Second, individual isozymes were often separately coregulated with distinct processes, providing a means of reducing crosstalk between pathways using a common reaction. Finally, transcriptional regulation defined a hierarchical organization of metabolic pathways into groups of varying expression coherence. These results emphasize the utility of incorporating regulatory information when analyzing properties of large-scale cellular networks.

Regulatory and metabolic functions of cells are mediated by networks of interacting biochemical components. Although most specific functions involve a limited set of components, several studies have shown that connectivity among proteins or metabolites extends far beyond the limits of individual function<sup>1–5</sup>. How separate functional modules are defined and isolated from each other to ensure their proper function under diverse physiological conditions is a central issue in understanding cellular organization<sup>6</sup>. A particularly relevant system is the cell's metabolic network, where hundreds of substrates are interconnected through biochemical reactions. Although such an interconnected design could in principle lead to the simultaneous flow of substrates in numerous directions, in practice metabolic fluxes pass through specific pathways. Moreover, the flux is readily optimized to maximize metabolic efficiency under different conditions<sup>7–9</sup>. Control of metabolic flow involves a variety of well-studied mechanisms, including allosteric interactions and covalent modifications affecting enzymatic activity. Recent genome-wide expression studies have revealed the prominent role of transcription in regulating metabolic flow in response to specific perturbations<sup>10,11</sup>. However, it remains unclear how and to what extent the modulation of enzyme expression determines functional metabolic units and how this affects the global properties of the metabolic network.

Here we have addressed this issue by analyzing the metabolic network of *S. cerevisiae*. Once we had established that genes associated with similar metabolic function are likely to exhibit a similar expression pattern, we characterized the regulation of genes associated with adjacent metabolic reactions. We then focused on reactions defining metabolic branch points as well as on those catalyzed by several

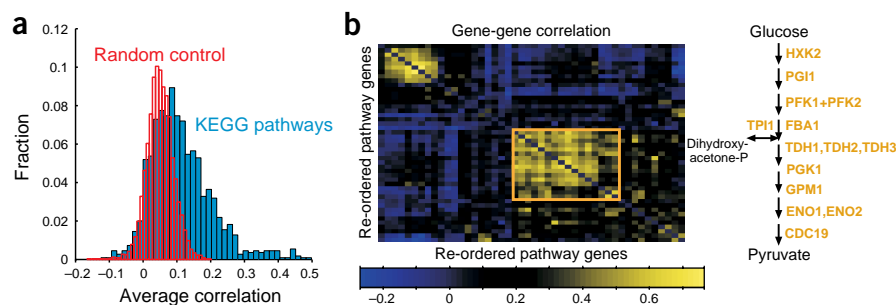
isozymes, and we provide a detailed account of the coregulation of enzymes associated with such reactions in central metabolic pathways. Several recurring regulatory patterns were observed that may represent general design principles of metabolic gene regulation. Next we extended the analysis to groups of genes associated with particular metabolic pathways. We provide a comprehensive database describing the coregulated genes and the regulatory conditions associated with most metabolic pathways (see authors' website; URL at end of Methods). In addition, this analysis revealed a higher-order organization of metabolic pathways into a hierarchy of groups with varying expression coherence. We discuss properties of this hierarchical organization and its relationship with the metabolite-based hierarchical modularity reported recently<sup>12</sup>.

## RESULTS

### Linear arrangement of coexpressed enzymes

To systematically examine the transcriptional coregulation of metabolic genes, we assembled a large dataset of over 1,000 genome-wide expression profiles describing the genome-wide transcription response of *S. cerevisiae* cells to a variety of environmental perturbations<sup>13,14</sup>, genetic mutations<sup>15</sup> and natural processes (Supplementary Table 1). Previous work has shown that functionally linked genes are often coexpressed<sup>16–19</sup>. To investigate the extent to which genes associated with the same metabolic function are coregulated, we considered first the classification of genes into metabolic pathways as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database<sup>20</sup>. As expected, we found that, on average, pairs of genes associated with the same metabolic pathway show a similar expression pattern,

<sup>1</sup>Department of Molecular Genetics and <sup>2</sup>Department of Physics of Complex Systems, Weizmann Institute of Science, 76100 Rehovot, Israel. <sup>3</sup>These authors contributed equally to this work. Correspondence should be addressed to N.B. (naama.barkai@weizmann.ac.il).



**Figure 1** Correlation between genes of the same metabolic pathway. **(a)** Distribution of the average correlation between genes assigned to the same metabolic pathway in the KEGG database. The distribution corresponding to random assignment of genes to metabolic pathways of the same size is shown for comparison. Importantly, only genes coding for enzymes were used in the random control. **(b)** Genes of the glycolysis pathway (according to the KEGG database) were clustered and ordered based on the correlation in their expression profiles. Shown here is the matrix of their pair-wise correlations. The cluster of highly correlated genes (orange frame) corresponds to genes that encode the central glycolysis enzymes. The linear arrangement of these genes along the pathway is shown at right.

reflected by a high correlation coefficient (Fig. 1a). However, we also noticed that typically only a subset of the genes assigned to a given pathway are in fact coregulated. For example, of the 46 genes assigned to the glycolysis pathway in the KEGG database, only 24 show a correlated expression pattern (Fig. 1b).

We asked if genes composing the coregulated subpart of the pathway are found in random positions within the metabolic route or, alternatively, if the coregulated genes define a particular metabolic path. In the case of glycolysis, the coregulated genes are arranged along the central part of the pathway (Fig. 1b). This was also true for most other pathways we examined (see authors' website). Moreover, detailed analysis of coregulation in central metabolic pathways showed that the coexpressed enzymes are often arranged in a linear order, corresponding to a metabolic flow that is directed in a particular direction (Fig. 2a; see authors' website). This is in contrast to the full structural descriptions of most metabolic pathways, which often contain numerous junctions that allow for alternative metabolic routes.

To examine more systematically whether coregulation enhances the linearity of metabolic flow, we analyzed the coregulation of enzymes at metabolic branch-points. Using the KEGG database, we identified all metabolites that are associated with three distinct reactions. Based on the directionality of the associated reactions, such junctions could integrate metabolic flow (convergent junction) or could allow the flow to diverge in two directions (divergent junction; see junction classification in Fig. 2b). Even in divergent junctions, however, metabolites will flow in both directions only when the enzymes catalyzing all three reactions are expressed. Alternatively, metabolic flux can be biased in one particular direction if the enzyme catalyzing one of the emanating reactions is not expressed.

We found that in the majority of divergent junctions, only one of the emanating branches is significantly coregulated with the incoming reaction that synthesizes the metabolite (Fig. 2b; see authors' website). Moreover, in cases where the incoming reaction was associated with several gene products (isozymes), distinct outgoing branches were often preferentially coexpressed with alternative isozymes (linear-switch; Fig. 2b). Such an arrangement corresponds to a linear metabolic flow, whose directionality can be switched in a condition-specific manner. Analysis of coregulation in junctions that allow metabolic flow in a larger number of directions demonstrated that there also, only a few of the emanating branches are coregulated with the incom-

ing branch (not shown). We conclude that transcription regulation is used to enhance the linearity of metabolic flux, by biasing the flow toward only a few of the possible routes. This is also reflected in the connectivity distribution of the network (Fig. 2c).

### Regulation of isozymes

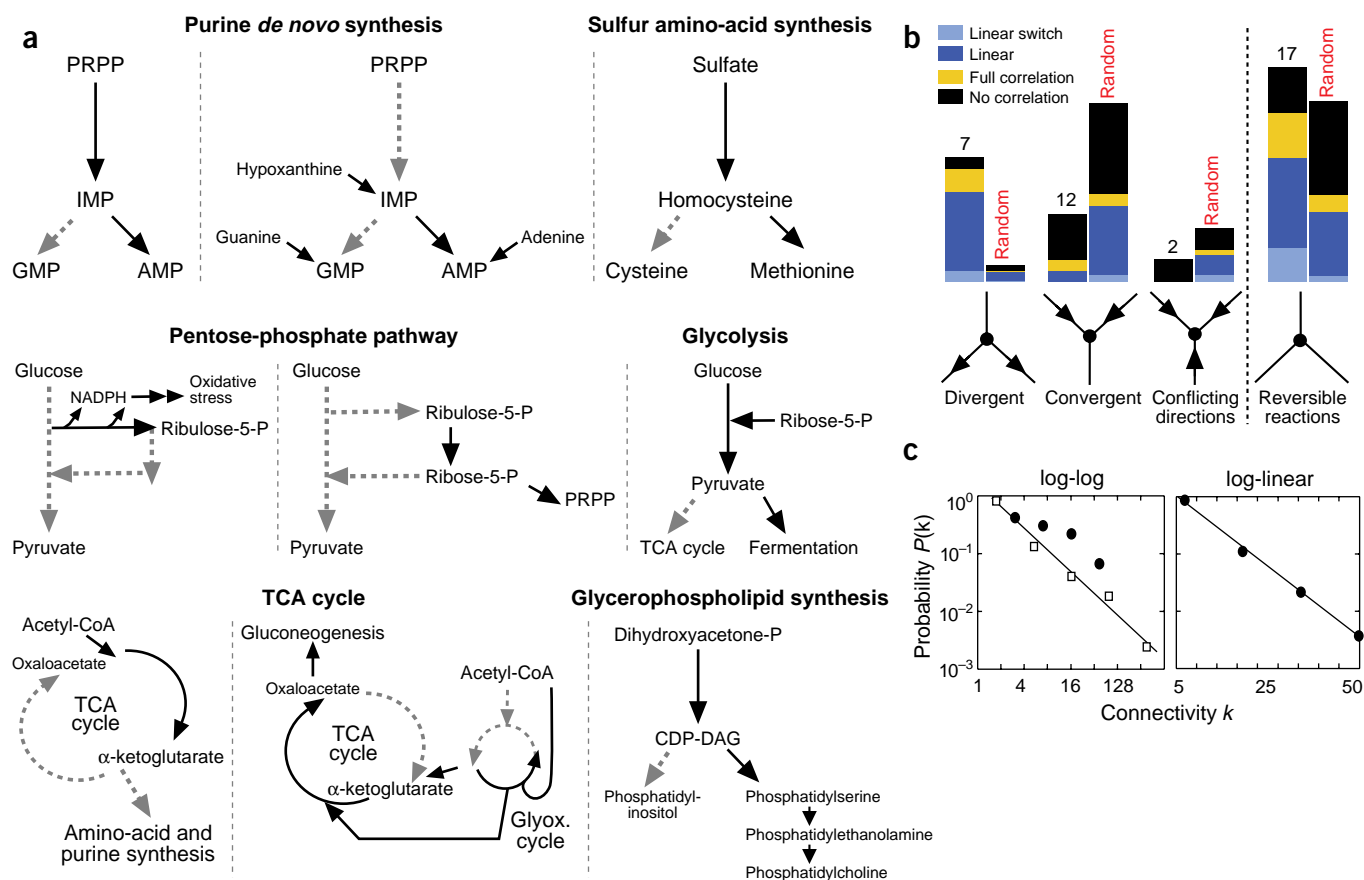
The observation that isozymes at junction points are often preferentially coexpressed with alternative reactions prompted us to investigate their role in the metabolic network more systematically. Isozymes could provide redundancy, which may be important for buffering the metabolic balance against genetic mutations and for amplifying metabolic production (Fig. 3a). Alternatively, isozymes could be dedicated to distinct processes using a common reaction and thus reduce cross-talk and unwanted interactions between separate metabolic pathways. It is not clear which

of these possible roles is more prevalent on a genomic scale. We reasoned that coexpression could be used to distinguish between the two alternatives, because redundant enzymes are expected to be coexpressed under the same conditions, whereas pathway-dedicated enzymes are likely to be individually expressed with the alternative processes using the shared reactions.

We analyzed the expression patterns of isozymes associated with central metabolic pathways. Most members of isozyme pairs were separately coregulated with alternative processes (Fig. 3b; see authors' website). To examine this result further, we also studied the regulatory pattern of all gene pairs associated with the same reaction according to the KEGG database. The majority of these pairs were differentially coregulated with alternative reactions, in agreement with the behavior observed at central metabolic pathways (Fig. 3c; see authors' website). This finding is supported by previous reports that specific members of isozyme families are induced by environmental stress, whereas other members of the same families are not<sup>14</sup>. Our results suggest that the primary role of isozyme multiplicity is to allow for differential regulation of reactions that are shared by separated processes. Dedicating a specific enzyme to each pathway may offer a way of independently controlling the associated reaction in response to pathway-specific requirements, at both the transcriptional and the post-transcriptional levels.

### Genes coexpressed with metabolic pathways

Our results above describe coexpression properties of genes catalyzing adjacent metabolic reactions. We next extended the analysis to groups of enzymes involved in specific metabolic pathways. To this end, we first refined the pathway information provided by the KEGG database in an extensive literature survey and created a dataset of metabolic pathways that includes the majority of carbohydrate metabolism and energy generation, lipid metabolism, amino acid biosynthesis, ribonucleotide synthesis, cofactor synthesis, metal metabolism and phosphate metabolism. Using this refined database, we characterized the coregulated subpart of each metabolic pathway and identified relevant experimental conditions that induce or repress the expression of the pathway genes. In addition, we associated additional genes showing similar expression profiles with each pathway. The algorithm used is the signature algorithm (ref. 19; see Methods for details). It requires as input a set of genes, some of which are expected to be coregulated. In



**Figure 2** Coexpressed enzymes often catalyze a linear chain of reactions. **(a)** Coregulation between enzymes associated with central metabolic pathways. Each branch corresponds to several enzymes whose identity is given on our web page. In the cases shown, only one of the branches downstream of the junction point is coregulated with upstream genes. **(b)** Coregulation pattern in three-point junctions. All junctions corresponding to metabolites that participate in exactly three reactions (according to the KEGG database) were identified and the correlations between the genes associated with each such junction were calculated. The junctions were grouped according to the directionality of the reactions, as shown. Divergent junctions, which allow the flow of metabolites in two alternative directions, predominantly show a linear coregulation pattern, where one of the emanating reaction is correlated with the incoming reaction (linear regulatory pattern) or the two alternative outgoing reactions are correlated in a context-dependent manner with a distinct isozyme catalyzing the incoming reaction (linear switch). By contrast, the linear regulatory pattern is significantly less abundant in convergent junctions, where the outgoing flow follows a unique direction, and in conflicting junctions that do not support metabolic flow. Most of the reversible junctions comply with linear regulatory patterns. Indeed, similar to divergent junctions, reversible junctions allow metabolites to flow in two alternative directions. Reactions were counted as coexpressed if at least two of the associated genes were significantly correlated (correlation coefficient  $>0.25$ ). As a random control, we randomized the identity of all metabolic genes and repeated the analysis. **(c)** The connectivity of a given metabolite was defined as the number of reactions connecting it to other metabolites. Shown are the distributions of connectivity between metabolites in an unrestricted network ( $\square$ ) and in a network where only correlated reactions are considered ( $\bullet$ ). In accordance with previous results<sup>25</sup>, the connectivity distribution between metabolites follows a power law (log-log plot). In contrast, when coexpression is used as a criterion to distinguish functional links, the connectivity distribution becomes exponential (log-linear plot). We verified that neither the functional exponential form nor the range of the distribution depend strongly on the correlation threshold used (see authors' website).

the present work, these input sets correspond to genes associated with particular pathways. The output contains the coregulated part of the input and additional coregulated genes, together with the set of conditions where the coregulation is realized. Noncoregulated genes in the input set do not appear in the output.

Numerous genes that are not directly involved in enzymatic steps clustered with defined metabolic pathways. We examined in detail the function of these genes, and found that most are engaged in processes that were linked to the associated pathways. In particular, we observed that genes encoding transporters were often associated with metabolic pathways using the transported metabolite (Supplementary Fig. 1). We further found that transcription factors were often assigned to the metabolic pathways they regulate (Supplementary Fig. 2). These observed strong functional connections suggest that uncharacterized

open reading frames can be assigned a functional link based on coexpression with specific metabolic pathways. Indeed, several such links generated by our analysis are noteworthy. For example, a gene encoding a putative transcription factor (YGR067C, containing a two tandem zinc-finger motif) was coexpressed with genes of the glyoxylate cycle, suggesting its putative involvement in their regulation. Another example is YGL186C, which is coexpressed with the purine *de novo* biosynthesis pathway and also has sequence similarity to the purine/cytosine transporter Fcy2p<sup>21</sup>, suggestive of its putative role as a purine transporter. Similarly, YJL200C is coexpressed with the lysine biosynthesis pathway and may catalyze its second step (conversion of homocitrate to homo-*cis*-aconitate) because of its similarity to ACO1, which catalyzes a similar reaction (conversion of citrate to *cis*-aconitate). This prediction has been independently suggested elsewhere<sup>22</sup>,

based on the finding that YJL200C deletion strains cannot grow on media lacking lysine.

### Hierarchical modularity in the metabolic network

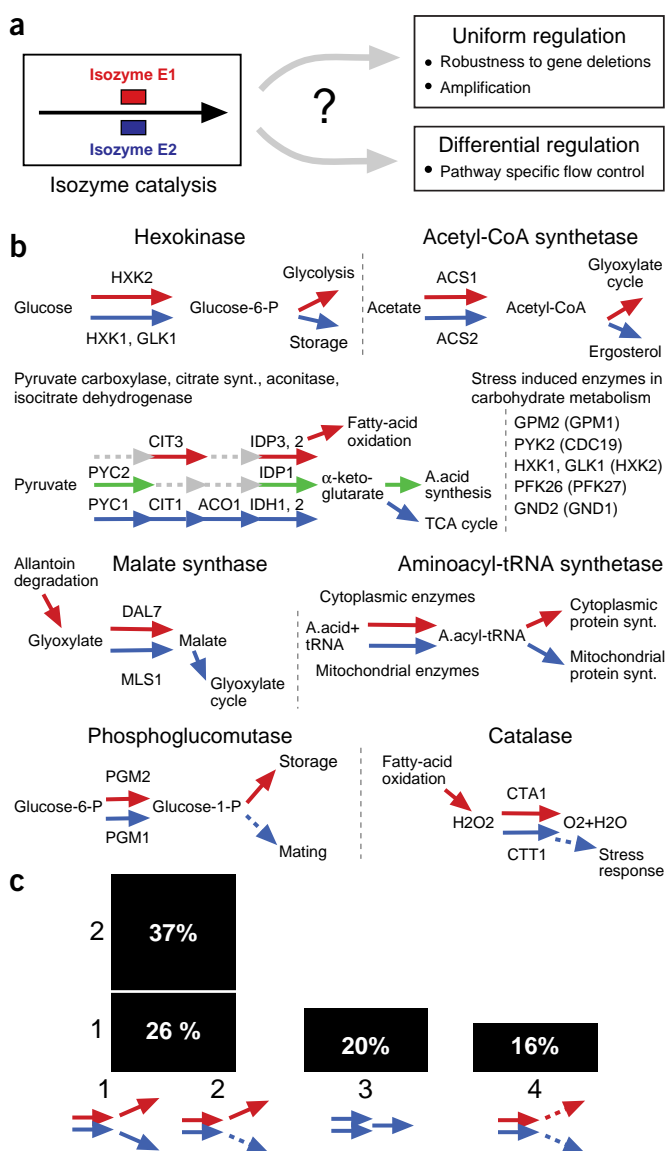
Coexpression analysis revealed a strong tendency toward coordinated regulation of genes involved in individual metabolic pathways. We next asked if transcription regulation also defines a higher-order metabolic organization, through coordinated expression of distinct metabolic pathways. An indication that such organization may exist was provided by our observation that feeder pathways (which synthesize metabolites) are frequently coexpressed with pathways using the synthesized metabolites (Supplementary Fig. 3). To explore a possible higher-level organization, we used a recently developed method (the iterative signature algorithm; ref. 23 and J.I., unpublished data) that is designed to reveal hierarchies of coregulatory units of varying expression coherence (Methods).

Our analysis identified a clear hierarchical structure, which is summarized by the module tree shown in Figure 4. Each box in this figure represents a group of coregulated genes (transcription module), where the stringency of coregulation is given by the indicated resolution parameter. Strongly correlated modules (left) include a small number of genes and can usually be associated with a very specific function, whereas moderately correlated modules (right) are larger and their function is less coherent. Lines connect modules that are of similar content obtained at different resolution. Importantly, distinct modules may include common genes, and when two branches of the module tree merge, the new module is not necessarily their union. Rather, the merging of two branches indicates that the associated modules are induced under similar conditions. It should be noted, in particular, that the low-resolution modules are generally different from the metabolic pathways and may not include the original pathway genes (J.I., unpublished data). Modules were annotated manually according to the function of the associated genes (see authors' website).

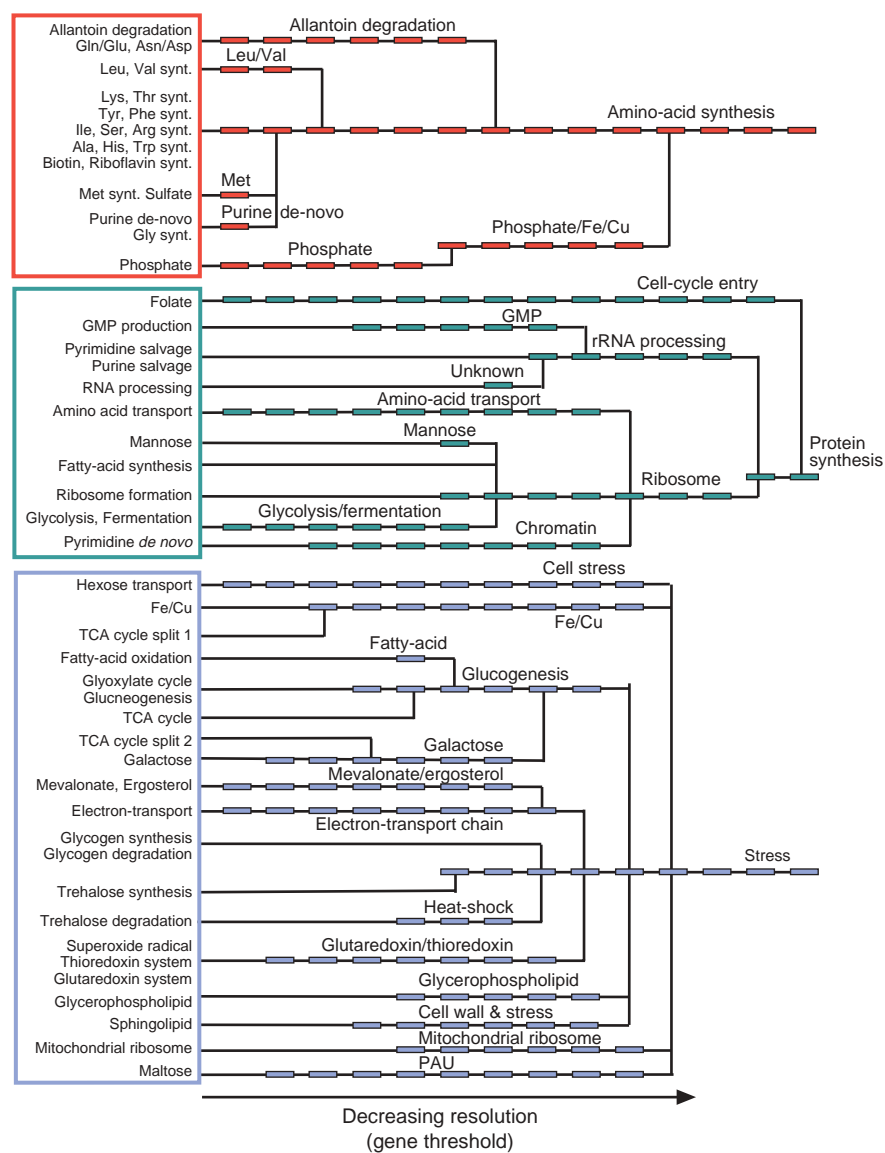
The most notable aspect of the hierarchical organization is the convergence of all pathways to one of just three low-resolution modules, associated with amino acid biosynthesis, protein synthesis and stress. Although amino acids serve as building blocks for proteins, the expression of genes mediating these two processes is clearly uncoupled. This lack of correlation may reflect the association of rapid cell growth (which triggers enhanced protein synthesis) with rich growth conditions, where amino acids are readily available and do not need to be synthesized. Amino acid biosynthesis genes, on the other hand, are required when external amino acids are scarce. In support of that, a group of amino acid transporters converged to the protein synthesis module, together with other pathways required for rapid cell growth (glucose fermentation, nucleotide synthesis and fatty acid synthesis).

Amino acid biosynthesis is also required for diverse metabolic processes, and its adjustment may thus be tuned to accommodate internal cellular requirements. Sensitivity to internal growth conditions may also account for the convergence of the purine *de novo* synthesis and the phosphate metabolism pathways to this module. Although the products of both pathways are required primarily for DNA synthesis, their regulation may be related to their role in producing ATP and thus dictated primarily by internal growth requirements. Indeed, in the purine *de novo* pathway, only the subpart leading to AMP, but not that leading to GMP, is coexpressed with the pathway (Fig. 2a). We note that a common transcription factor (Pho2) is known to regulate the expression of genes from both groups, as well as some genes involved in amino acid biosynthesis.

Pathways that directly mediate the defense against stress (superoxide radical removal, glutaredoxin and thioredoxin systems and trehalose



**Figure 3** Differential regulation of isozymes. (a) Two possible functions of isozymes associated with the same metabolic reaction. An isozyme pair could provide redundancy which may be needed for buffering genetic mutations or for amplifying metabolite production. Redundant isozymes are expected to be coregulated. Alternatively, distinct isozymes could be dedicated to separate biochemical pathways using the associated reaction. Such isozymes are expected to be differentially expressed with the two alternative processes. (b) Differential regulation of isozymes in central metabolic pathways. Arrows represent metabolic pathways composed of a sequence of enzymes (see authors' website for gene identity). Coregulation is indicated with the same color (e.g., the isozyme represented by the green arrow is coregulated with the metabolic pathway represented by the green arrow). (c) Regulatory pattern of all gene pairs associated with a common metabolic reaction (according to the KEGG database). All such pairs were classified into one of four classes: parallel (1), where each gene is correlated with a distinct connected reaction (a reaction that shares a metabolite with the reaction catalyzed by the respective gene pair); selective (2), where only one of the enzymes shows a significant correlation with a connected reaction; and converging (3), where both enzymes were correlated with the same reaction. Correlations coefficients  $>0.25$  were considered significant. To be counted as parallel, rather than converging, we demanded that the correlation with the alternative reaction be  $<80\%$  of the correlation with the preferred reaction (see authors' website).



**Figure 4** Hierarchical modularity in the metabolic network. This hierarchy was derived by applying the iterative signature algorithm to the metabolic pathways shown, and decreasing the resolution parameter (coregulation stringency) in small steps (see Methods). The gene content of all modules is given on the authors' website.

metabolism) converged to the stress module. Also associated with the same branch were pathways required for nonfermenting metabolism or other mitochondrial functions (TCA cycle, electron transport chain, glyoxylate cycle, gluconeogenesis, mitochondrial ribosome formation, and the mevalonate pathway leading to the production of quinone and heme required in the mitochondria). This association may indicate that cells trigger gene expression associated with the general environmental stress response to protect against the generation of free radicals produced during mitochondria respiration.

#### Global properties shaped by transcription regulation

It was recently shown that the structural connectivity between metabolites also imposes a hierarchical organization of the metabolic network<sup>12</sup>. This previous analysis was based on connectivity between substrates, considering all potential connections, whereas our analysis

is based on coexpression of enzymes. Intriguingly, related metabolic pathways were clustered together in both cases. Several important differences between the two hierarchical trees should be noted, however. First, although in the substrate-based hierarchy all carbohydrate metabolisms were grouped together, coexpression appears to separate the preferred fermentable sugar metabolism (e.g., glucose and mannose) from the metabolism of nonfermentable sugars. Similarly, although fatty acid synthesis and degradation were grouped together in the substrate-based hierarchy, coexpression associated these processes with different parts of the tree. Finally, we also observed that the large-scale structures of the two hierarchical trees differ. In particular, nucleotide metabolism emerged as a distinct branch in the substrate-based analysis, whereas in our analysis it was merged with protein synthesis and other processes involved in rapid cell growth. In contrast, the clear separation between protein synthesis and amino acid biosynthesis, which we identified as a central feature of the regulatory pattern, is not evident in the structure-based hierarchy.

Lastly, we examined whether large-scale topological features are altered when regulatory relationships are introduced. It was previously shown that the fraction of metabolites that are connected to other metabolites in the metabolic network by exactly  $k$  reactions decays as a power law  $P(k) \sim k^{-\gamma}$ , where  $\gamma$  is a network-specific constant (ref. 24), corresponding to a scale-free network topology. We found that when coexpression is used as a criterion to distinguish the connections that are functional under the same cellular contexts, this distribution becomes exponential (Fig. 2c), corresponding to a network structure with a defined scale of connectivity. This result reflects the reduction in the complexity of the metabolic flow, achieved by coexpressing only a few of the branches at each metabolic junction (compare Fig. 2a and b).

#### DISCUSSION

We report here a systematic analysis of how transcription regulation shapes the metabolic network of *S. cerevisiae*. Our analysis provides a detailed account of transcriptional regulation of most metabolic pathways, by specifying the coregulated genes within their regulatory context. Detailed information about a variety of pathways can be accessed using an interactive application provided on the authors' website.

A main finding that emerged from our analysis is that transcription leads the metabolic flow toward linearity. The structural description of the metabolic network is far from linear, but contains an abundance of branch points, reflecting the need for flexibility and diversity of metabolic flow. The coexpression pattern of enzymes participating at such branch points, however, suggests that many possible branches are in fact suppressed in the actual context-dependent map. This suppres-

sion reduces metabolite dissipation and ensures a more efficient metabolic flow.

It would appear that the activity of a linear metabolic pathway could be modified by altering the expression of just one, or a few, of its enzyme constituents (e.g., an enzyme catalyzing its rate-limiting step). However, we found that coregulated enzymes are often arranged in a linear order, suggesting that coexpression itself is in fact used to enhance the linearity of metabolic routes. The importance of coordinately modifying the full chain of enzymes is underlined by several experiments demonstrating that fluxes in central metabolic pathways, such as the glycolytic pathway in yeast, are largely robust to changes in the concentration of individual enzymes<sup>25</sup>. More importantly, metabolic flux was altered only when the concentration of all enzymes was changed in a coordinated manner<sup>25</sup>. Although the mechanism underlying this robustness is not known, it probably involves post-translational compensatory mechanisms resulting from the interaction between pathway components. These results emphasize the need for coordinately regulating the expression of all pathway components to obtain a change in metabolic flux.

Our second main finding is that transcriptional regulation entails a higher-order structure of the metabolic network. In contrast to the naive way of thinking about isolated and equivalent pathways, our analysis revealed a hierarchical organization of metabolic pathways into groups of decreasing expression coherence. Interestingly, hierarchical modularity was also described recently based on the analysis of hard-wired connectivity between metabolites<sup>12</sup>. This emergence of related hierarchical structures from two conceptually distinct analyses suggests the universality of this type of organization. At the same time, the distinct properties of the two ensuing hierarchies point to important differences between structural and functional constraints imposed on the metabolic network.

In conclusion, we have shown that transcription regulation is prominently involved in shaping the metabolic network of *S. cerevisiae* in response to changing conditions. As more genomic data accumulate in other organisms, it will be interesting to see if similar principles apply to metabolic networks of bacteria or higher eukaryotes.

## METHODS

**Expression dataset.** Our dataset was composed of over 1,000 conditions, including environmental stresses, profiles of deletion mutants and natural processes such as cell cycle. A full list of references is given in **Supplementary Table 1**. The expression data was organized in two separate normalizations:  $E_G(g,c)$  and  $E_C(g,c)$  are the log-expression ratios of gene  $g$  in condition  $c$  normalized over genes and conditions, respectively, such that  $\langle E_G \rangle_{\text{all } g} = 0$ ,  $\langle [E_G]^2 \rangle_{\text{all } g} = 1$  for each  $c$  and  $\langle E_C \rangle_{\text{all } c} = 0$ ,  $\langle [E_C]^2 \rangle_{\text{all } c} = 1$  for each  $g$ . Here,  $\langle \rangle$  denotes the average value.

**Directionality of reactions in the KEGG database.** The KEGG data files do not include the directionality of the reactions, but this information is available in the pathway images provided by KEGG. We thus added this information by visual inspection.

**Calculation of correlation coefficients.** Throughout the analysis, Pearson correlation coefficients were used as a measure of similarity between expression profiles. The coefficient  $r$  for two expression vectors  $A$  and  $B$  is defined as  $r = \sum(A_n - \langle A \rangle) \cdot (B_n - \langle B \rangle) / \sqrt{[\sum(A_n - \langle A \rangle)^2 \cdot \sum(B_n - \langle B \rangle)^2]}$ , where  $\langle \rangle$  denotes the average value.

**Statistics of three-point junctions.** The KEGG database was searched for metabolic compounds that are involved in exactly three reactions. Only reactions associated with enzymes that exist in *S. cerevisiae* were considered. In the cases where several reactions were catalyzed by the same enzymes, one representative reaction was chosen, such that all junctions considered were composed of precisely three reactions, catalyzed by distinct enzymes. Each three-junction was

categorized according to the correlation pattern found between enzymes catalyzing its branches (Fig. 2b). Correlation coefficients larger than 0.25 were considered significant. As a random control, the analysis was repeated, but the directionality of the reactions and the catalyzing genes were replaced randomly. See the supplementary table on the authors' website for the details of the individual compounds and reactions. All junctions were inspected visually to control for errors in the database; as a result, we removed four junctions.

**Statistics of isozyme regulation.** We searched the KEGG database for reactions  $R_i$  catalyzed by more than one gene product  $\{E_j\}$  in *S. cerevisiae*. For each such reaction, we identified all reactions that share with it at least one common metabolite (connected reactions). We then measured the correlation between the genes associated with these connected reactions and all enzymes  $\{E_j\}$  catalyzing the original reaction. The correlation pattern was classified according to the four types shown in **Figure 3**. Correlation values exceeding 0.25 were considered significant. To ensure that only proper isozymes and not complex members or homologous genes are considered, we excluded all genes whose one-line description contained one of the keywords 'Complex', 'Subunit', 'Strong similarity', 'Component', 'Homolog' or 'Maltose'. In addition, common compounds that participate in a very large number of reactions (ATP, ADP, protein, acceptor,  $H_2O$ ,  $CO_2$ , NADH,  $NAD^+$ , NADPH,  $NADP^+$ , orthophosphate, pyrophosphate and AMP) were excluded from the analysis. Details of each reaction, including its classification and the correlation coefficient found between the associated genes, are given on the accompanying authors' website.

**Recurrent signature algorithm.** The signature algorithm is described in detail in a previous publication<sup>19</sup>. It consists of two steps. First, all conditions in the dataset are scored by their average expression over the input set  $G_m$  of genes:  $s_c = \langle E_G \rangle_{g \text{ in } G_m}$ . The conditions whose absolute score  $|s_c|$  exceeds the condition threshold  $t_C$  are selected. This set of conditions is denoted  $C_{\text{out}}$ . Second, all genes are scored by the weighted average over  $C_{\text{out}}$ :  $s_g = \langle s_c \cdot E_C \rangle_{c \text{ in } C_{\text{out}}}$ . The genes with a score  $s_g$  greater than the gene threshold  $t_G$  are selected. This set of genes, together with their associated score, defines the output of the signature algorithm.

**Hierarchical structure: iterative signature algorithm.** The method used to reveal hierarchical structure between transcription modules is an iterative extension of the signature algorithm described above. It is described in detail in a separate work (ref. 23 and J.L., unpublished data). A transcription module consists of a set of coregulated genes (a subset  $G_m$  of all genes  $G$ ) and an associated set of regulating conditions (a subset  $C_m$  of all conditions  $C$ ). Optionally, each gene  $g$  and each condition  $c$  may also be characterized by scores  $s_g$  and  $s_c$ , respectively, that indicate their relative importance. If no preference is given to any of the genes or conditions, all scores are set to unity. The defining property of a transcription module is self-consistency, which is introduced as follows. First, we assign new scores to both genes and conditions that reflect their actual degree of association with the module. The gene score is the average expression of each gene over the module conditions, weighted by the condition score:  $s_g = \langle s_c \cdot E_C \rangle_{c \text{ in } C_m}$ . Analogously, the condition score is the weighted average over the module genes,  $s_c = \langle s_g \cdot E_G \rangle_{g \text{ in } G_m}$ . Self-consistency implies that the genes of the module are exactly those genes of the dataset that receive the highest scores  $s_g$ , while the module conditions are those conditions of the dataset with the highest scores  $s_c$ . To identify transcription modules, we iteratively apply the signature algorithm until convergence to a fixed point of the algorithm is reached. By definition, these fixed points satisfy the self-consistency criterion. The gene threshold  $t_G$  functions as a resolution parameter that controls the minimum level of coregulation stringency among the module genes. The threshold values are given in units of the expected standard deviation (corresponding to uncorrelated genes or conditions). Starting from the highest resolution, we applied the iterative signature algorithm to 69 distinct gene sets, each corresponding to the genes of a specific pathway. The resulting fixed points represent the transcription modules that are closest to the original pathway genes. The procedure is then repeated at a lower threshold, by iterating from the fixed points obtained in the previous resolution. In those cases where the iterations yielded a fixed point with fewer than five genes, the original pathway genes were used for the iterations at the lower threshold. As the threshold is lowered, new genes are included in each module. Modules in general remain fixed points over a range of thresholds, with gradual changes in their gene and condition content. At a

specific resolution, however, the modules lose their stability as genes from related but originally separate modules are included. This leads to either fusion of two modules or convergence of the weaker towards the stronger, revealing the modular hierarchy of coregulation shown in Figure 4. The range of thresholds considered was  $t_G = 5.4-1.5$ .

**Connectivity distributions.** For every metabolic compound in the KEGG database, we identified all reactions  $\{R_i\}$  involving the compound, together with the associated enzymes  $\{E_i\}$  in *S. cerevisiae*. For each such reaction, the correlation coefficient between its catalyzing enzyme  $E$  and those of the remaining reactions of  $\{R_i\}$  was measured. Reactions that are catalyzed by an enzyme whose correlation coefficient with  $E$  exceeded the threshold of 0.25 were considered connected. All distinct connection patterns were recorded for all reactions in  $\{R_i\}$ . The restricted connectivity is the number of connected reactions in each distinct pattern. The structural connectivity of the unrestricted network corresponds to the number of reactions  $R_i$ . Note that in the restricted network, there are in general several connectivity numbers associated with each metabolite, reflecting context-specific regulation. By contrast, in the case of structural connectivity, a single number is associated with each metabolite. The distributions of the restricted and the structural connectivities are shown (Fig. 2c).

**URL.** In addition to the supplementary information attached to this paper on the *Nature Biotechnology* website, an accompanying website containing supplementary material is available at <http://barkai-serv.weizmann.ac.il/MetabolicNetworks>.

*Note: Supplementary information is available on the Nature Biotechnology website.*

#### ACKNOWLEDGMENTS

We thank Sven Bergmann, Avigdor Eldar and Benny Shilo for comments on the manuscript. This work was supported by US National Institutes of Health grant no. A150562, by the Israeli Science Ministry and by the Y. Leon Benozzi Institute for Molecular Medicine. N.B. is the incumbent of the Soretta and Henry Shapiro career development chair at the Weizmann Institute of Science.

#### COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 18 August; accepted 29 October 2003  
Published online at <http://www.naturebiotechnology.com/>

1. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
2. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).

3. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
4. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
5. Lee, T.I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
6. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
7. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
8. Ibarra, R.U., Edwards, J.S. & Palsson, B.O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186–189 (2002).
9. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E.D. Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**, 190–193 (2002).
10. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
11. Miki, R. *et al.* Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* **98**, 2199–2204 (2001).
12. 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).
13. Causton, H.C. *et al.* Remodeling of yeast genome expression in response to environmental changes. *Mol. Biol. Cell* **12**, 323–337 (2001).
14. Gasch, A.P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
15. Hughes, T.R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
16. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
17. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
18. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
19. Ihmels, J. *et al.* Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* **31**, 370–377 (2002).
20. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
21. Nelissen, B., De Wachter, R. & Goffeau, A. Classification of all putative permeases and other membrane plurispansers of the major facilitator superfamily encoded by the complete genome of *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **21**, 113–134 (1997).
22. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
23. Bergmann, S., Ihmels, J. & Barkai, N. The Iterative Signature Algorithm for the analysis of large scale gene expression data. *Phys. Rev. E* **67**, 031902 (2003).
24. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
25. Schaaff, I., Heinisch, J. & Zimmermann, F.K. Overproduction of glycolytic enzymes in yeast. *Yeast* **5**, 285–290 (1989).