Letter

Systematic condition-dependent annotation of metabolic genes

Tomer Shlomi,[1,4] Markus Herrgard,[2] Vasiliy Portnoy,[2] Efrat Naim,[1] Bernhard Ø. Palsson,[2] Roded Sharan,[1] and Eytan Ruppin[1,3,4]

[1]*School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel;* [2]*Department of Bioengineering, University of California, San Diego, California 92093-0412, USA;* [3]*School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel*

The task of deriving a functional annotation for genes is complex as their involvement in various processes depends on multiple factors such as environmental conditions and genetic backup mechanisms. This study employs a large-scale model of the metabolism of *Saccharomyces cerevisiae* to investigate the function of yeast genes and derive a condition-dependent annotation (CDA) for their involvement in major metabolic processes under various genetic and environmental conditions. The resulting CDA is validated on a large scale and is shown to be superior to the corresponding Gene Ontology (GO) annotation, by showing that genes annotated with the same CDA term tend to be more coherently conserved in evolution and display greater expression coherency than those annotated with the same GO term. The CDA gives rise to new kinds of functional condition-dependent metabolic pathways, some of which are described and further examined via substrate auxotrophy measurements of knocked-out strains. The CDA presented is likely to serve as a new reference source for metabolic gene annotation.

[Supplemental material is available online at www.genome.org.]

In recent years, high-throughput techniques have provided a wealth of data on the expression and activity of genes and proteins. The task of inferring the involvement of gene products in various cellular processes, commonly referred to as functional annotation, is a major goal of current biological research. It involves the definition of a set of biological functions, termed "ontology," and the association of the gene products with ontology terms. The most comprehensive and commonly used ontology is the Gene Ontology (GO), consisting of 20,000 terms and numerous associated gene products (Ashburner et al. 2000). Both the ontology and the corresponding annotation are constantly updated based on various experimental manifestations of gene function.

The involvement of a gene product in a specific process depends on multiple factors, such as the environmental conditions (Giaever et al. 2002; Brown et al. 2006) and the genetic background (Hartman et al. 2001; Tong et al. 2004; van Swinderen and Greenspan 2005). Recent work has provided several lines of evidence for the condition-dependent nature of gene function in *Saccharomyces cerevisiae* based on large-scale phenotypic screens and gene deletion phenotypes across multiple growth media (Harrison et al. 2007). However, the GO annotation does not adequately reflect this basic condition-dependency of gene function as it strives to maintain uniformity in the experimental conditions underlying the annotation. To achieve such uniformity, the GO consortium has instructed annotators that gene products should be annotated with terms reflecting their activity in standard experimental conditions (http://www.geneontology.org). The determination of what these normal experimental conditions are for any particular organism is then left to the annotator's judgment. Consequently, the current ontology specification lacks an overall comprehensive treatment of the various factors that affect gene function.

The annotation of metabolic genes is a particularly difficult task, arising from the high level of dependency between the function of individual metabolic enzymes that form the overall complex network of biochemical reactions. Yet, inspecting the sources of GO annotation for metabolic genes reveals that most annotations (56%) are based solely on the involvement of genes in classical biochemical pathways (Traceable Author Statement evidence code; Supplemental Fig. 1). Twenty-six percent of the annotations arise from gene knockout experiments measuring various metabolic phenotypes (Inferred from Mutant Phenotype evidence code). As no single "normal" condition is enforced, these experiments span a large range of environmental conditions and genetic backgrounds, whose identities are not reflected in the annotation. More complex phenotypic experiments involving the knockout of multiple genes to identify genetic backup mechanisms account for only 3% of the annotations (Inferred from Genetic Interaction evidence code). Notably, such experiments that determine epistatic interactions between genes are still difficult to conduct on a genome-wide scale especially in multiple growth conditions.

In this study, we employ a genome-scale model of cellular metabolism to investigate the function of genes under multiple environmental and genetic conditions, deriving a condition-dependent annotation (CDA) of metabolic genes. The CDA associates genes with terms representing metabolic processes under multiple conditions. Specifically, we employ constraint-based modeling that uses stoichiometric, thermodynamic, and flux capacity constraints to predict a space of possible flux distributions attainable by the metabolic network under various environmental and genetic conditions. Flux Balance Analysis (FBA) is a specific constraint-based optimization method that is commonly used to find flux distributions that minimize or maximize a defined cellular objective such as its biomass production rate. Here we employ numerous optimization criteria to explore the organ-

[4]Corresponding authors.
E-mail shlomito@post.tau.ac.il; fax +972-3-640-9357.
E-mail ruppin@post.tau.ac.il; fax +972-3-640-9357.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6678707.

17:1626–1633 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

ism's ability to synthesize metabolites that contribute to biomass formation under multiple conditions. Constraints-based models have been successfully used previously for predicting various metabolic phenotypes such as growth, uptake rates, by-product secretion, knockout lethality, and pathway activity across different conditions (Edwards and Palsson 2000; Schilling et al. 2000; Schuster et al. 2000; Edwards et al. 2001; Ibarra et al. 2002; Famili et al. 2003; Forster et al. 2003; Mahadevan and Schilling 2003; Kuepfer et al. 2005; Shlomi et al. 2005).

We derive a CDA for metabolic genes of the yeast *S. cerevisiae*. The resulting CDA spans two dimensions representing growth media (minimal/rich) and the availability of oxygen, and accounts for genetic backups in the form of isozymes and alternative pathways. The CDA obtained is compared with the standard GO annotation, generating novel annotation predictions, which are validated in a large-scale manner using gene conservation and expression data. To gain insight on the dynamic metabolic mechanisms underlying the annotation, we derive functional pathways, which are condition-dependent, network-based representations of the biosynthetic processes. We then conduct growth phenotyping of single and double gene deletion strains in auxotrophic growth conditions to examine these novel functional pathways.

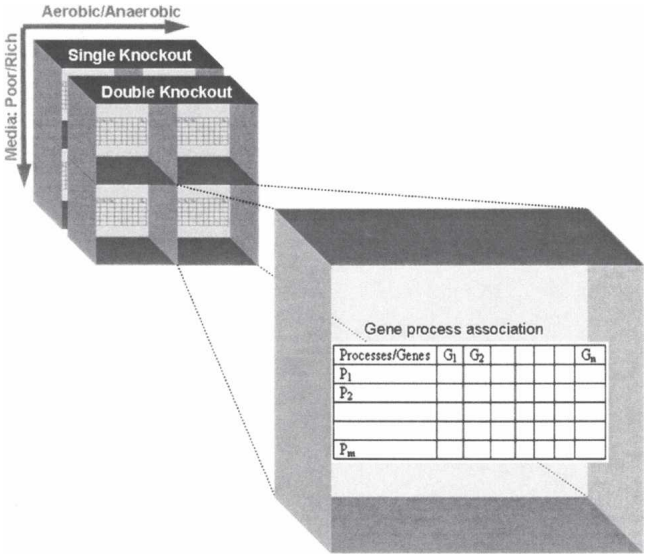

Figure 1. Condition-dependent annotation (CDA). A schematic representation of the CDA as a set of associations between genes and processes under various conditions. The basic element is a table specifying an association between genes and processes. This condition-dependent association table is the content of each entry of the CDA. These entries are spanned in turn by two dimensions representing growth media and the availability of oxygen (but additional dimensions can be added in principle in accordance with the data available). We consider two annotation systems, for single- and for double-knockouts simulations.

Results

Deriving a condition-dependent annotation of yeast metabolism

We have used a genome-scale, metabolic network model of the yeast *S. cerevisiae* (Duarte et al. 2004) to derive a condition-dependent annotation for the involvement of genes in various metabolic processes under multiple conditions. We focused on processes that synthesize the 38 different compounds that form the bulk of yeast biomass, such as amino acids, nucleotides, lipids, etc. The annotation protocol involved simulating the effect that gene knockouts have on different biosynthesis processes, in accordance with common experimental procedures for gene annotation, such as auxotrophy tests (i.e., tests of the inability of an organism to synthesize a particular organic compound required for its growth) (Cherest et al. 1993) and measurements of metabolite concentrations following genetic mutations (Farkas et al. 1991). Specifically, to determine whether a given gene contributes to the production of a certain compound under a certain condition, we compute the maximal production rate of the compound when a gene is present in the model and after it is knocked out and measure the drop in the compound's production rate (Methods). To identify a contribution of a gene to the production of a certain compound that is backed up by isozymes or alternative pathways, the annotation protocol involved a systematic knockout of all genes pairs in the model while measuring the resulting drop in production rate (Methods). Notably, a previous study has shown that FBA can reliably predict condition-specific backup mechanisms (Harrison et al. 2007). The various simulated conditions are spanned by two dimensions representing poor and rich media, and aerobic and anaerobic conditions (Fig. 1). Overall, we analyzed the contribution of 750 metabolic genes to 38 biosynthetic processes, under two media (poor and rich) and two environments (aerobic and anaerobic) using single- and double-knockout analysis.

We define a gene's multifunctionality level as the number of processes it contributes to in a given condition. The distribution of the multifunctionality levels of all genes under single-knockout conditions exhibits a bimodal shape that peaks at levels 1 and 38. That is, most genes are annotated as being involved either in a small number of processes or in almost all processes (Supplemental Fig. 2). A similar bimodal distribution of environmental specificity of predicted synthetic lethal interactions was previously observed (Harrison et al. 2007). To obtain annotations for genes that denote their contribution to specific processes (in contrast to "housekeeping" genes, which contribute across the board) we excluded genes with a multifunctionality level greater than a threshold of five in a given environmental condition. We note that the GO annotation of metabolic genes shows a similar multifunctionality pattern with <7% of the genes annotated to more than five processes. The resulting CDA consists of a total of 651 associations between 233 genes and the 38 ontology terms describing the different metabolic processes under the conditions examined here. It can be accessed via the supplemental Web site, http://www.cs.tau.ac.il/~shlomito/CDA.

Examining the annotation within the different CDA conditions reveals the importance of each annotation dimension (Table 1). The annotation obtained with single knockout simulations in rich media significantly varies between aerobic and anaerobic conditions, with the aerobic condition providing 85% of the total annotations (obtained in either aerobic or anaerobic condition), and the anaerobic condition providing only 62%. Similarly, the dimension representing growth media is important for single-knockout experiments in anaerobic conditions, as we get only 87% and 60% of the annotations when considering either poor or rich media, respectively. Interestingly, single knockouts in aerobic conditions, which form the basis for much of the existing GO annotations for yeast, are somewhat insensitive to the specific choice of growth medium, as both poor and

Table 1. CDA statistics

	Aerobic	Anaerobic	Both
Single knockouts			
Poor media	214 (111)	206 (98)	247 (121)
Rich media	195 (105)	141 (87)	229 (124)
Both	221 (114)	236 (115)	279 (135)
Double knockouts			
Poor media	203 (82)	245 (95)	333 (117)
Rich media	156 (78)	178 (86)	261 (112)
Both	233 (95)	343 (130)	428 (147)
Total			
Poor media	417 (176)	451 (175)	564 (212)
Rich media	351 (169)	319 (156)	463 (198)
Both	452 (189)	532 (202)	651 (233)

The number of functional annotations (and the corresponding number of genes in parenthesis) considering the different dimensions of the CDA.

rich media provide >88% of the total annotations in these conditions. The double-knockout analysis more than doubled the number of functional annotations and exhibits even higher variation between different conditions. These results clearly show the importance of a CDA that considers the involvement of genes in various processes, under multiple environmental conditions and genetic backup levels.

The similarity and differences between CDA and GO

Gene Ontology contains specific terms representing the process of synthesizing each of the 38 essential biomass compounds that are required for growth according to the metabolic network model. Yeast GO annotation consists of 361 associations between 199 metabolic genes and these ontology terms (Methods). The overlap between GO and the CDA, when considering all conditions, is highly significant with 179 common annotations (hyper-geometric P-value $<1 \times 10^{-300}$). Focusing on specific CDA conditions, we find that 60% of the CDA annotations in single-knockout aerobic conditions appear in GO, covering 33% of the latter annotations. In the single-knockout anaerobic conditions (which are less common within the experiments underlying GO) only 49% of the annotations appear also in GO, covering 19% of the latter. In the double-knockout conditions, <29% of annotations appear also in GO, suggesting that double-knockout experiments would significantly enrich functional annotations by revealing functional contributions masked by genetic backup mechanisms. An inspection of GO annotations that are not included in the CDA shows that in some cases they were identified in experiments conducted under growth media other than the poor or rich media considered here, or with experiments involving high-order knockouts (e.g., triple or quadruple knockouts) (Johnson et al. 1994). In some cases the GO annotations that are not included in the CDA are predicted to be highly nonspecific as the corresponding genes are predicted to contribute to a high number of processes and thus may potentially reflect overly specific GO annotations (Supplemental Table 1). Overall, the CDA consists of 472 novel annotations for 158 unique genes whose biological plausibility is further examined below.

The number of novel annotations varies across the different metabolic processes and the different dimensions of the annotation (Supplemental Table 2; Supplemental Fig. 3). For example, CDA extends the current GO annotation for amino acid biosynthetic processes by 83% under anaerobic conditions, while extending it by only 60% in aerobic conditions. Examining the known GO annotation of the novel CDA predictions, we find that 43% of the novel CDA predictions have a corresponding GO annotation within a category of closely related terms (Fig. 2). Almost all of these novel predictions (96%) are within the amino acids category (Supplemental Fig. 4). For example, several genes annotated in GO as involved in methionine biosynthesis are also annotated as involved in cysteine and isoleucine biosynthesis in the CDA. Other genes are annotated very differently than in GO. For example, we find several genes annotated in GO as involved in nucleotide biosynthesis and are predicted to be involved in the production of the amino acids histidine, cysteine, and methionine.

A network representation of the CDA under single knockout, poor media, aerobic and anaerobic conditions, and its comparison with GO is shown in Figure 3. The network is clustered with distinct sets of genes annotated in each category. In the lipid biosynthesis cluster, we see that many known GO annotations for ergosterol biosynthesis are found in the CDA only in aerobic conditions. These condition-specific annotations reflect the known dependency of the ergosterol biosynthesis pathway on the availability of oxygen (Deytieux et al. 2005). This cluster is connected to the amino-acids cluster via *MET6* (annotated in GO as involved in methionine biosynthesis), reflecting the dependency of the ergosterol biosynthesis pathway on the cofactor S-adenosylmethionine. Genes annotated as involved in nucleotide biosynthesis in the CDA are organized in two clusters for purine and pyrimidine biosynthesis. The amino acids biosynthesis cluster is the largest cluster. It is connected to the purine and pyrimidine biosynthesis clusters via shared precursor steps in the biosynthetic pathways (*ADE3*, *ADE12*, and *ADE13* for purines and histidine; *CPA1* and *CPA2* for pyrimidines and arginine). Another link between these clusters is formed by *FUM1* (fuma-

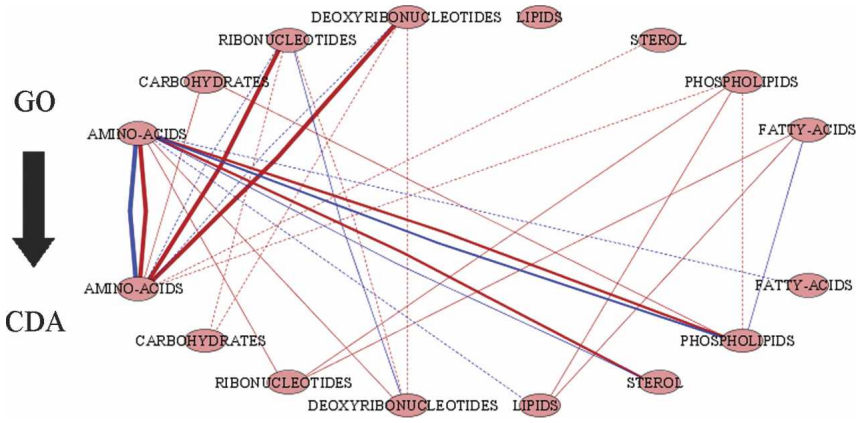

Figure 2. Novel CDA predictions of known GO annotations. Nodes represent GO terms (*top* curve) and CDA terms (*bottom* curve). An edge between GO term x and CDA term y represents a set of novel CDA associations of genes annotated in GO as involved in term x, and annotated with term y in CDA. The width of the edge represents the set size. Blue and red edges represent annotations obtained with single and double knockouts, respectively.

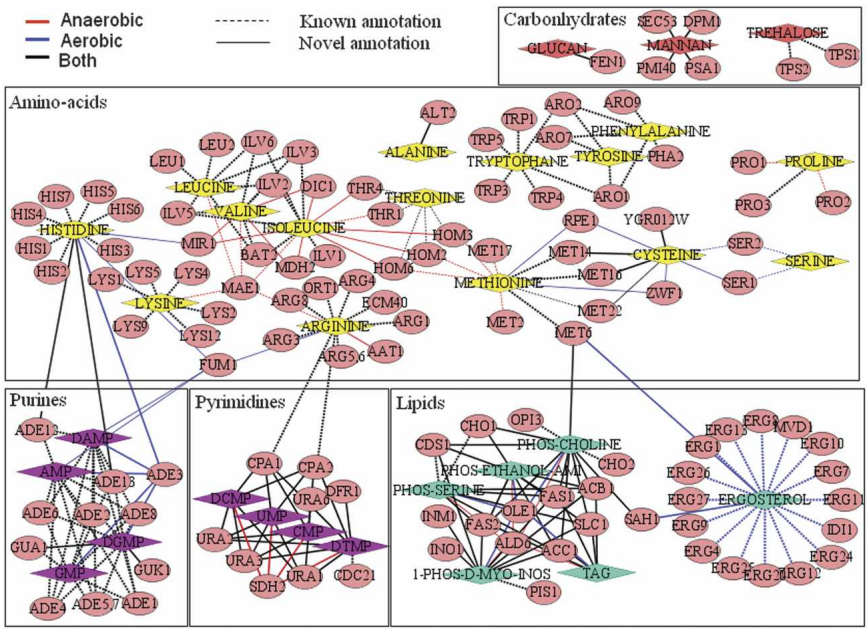

Figure 3. A network view of CDA in single knockout, poor media conditions, in aerobic and anaerobic conditions. Genes are marked with red circular nodes, and process terms with colored diamond nodes. Edges connecting between genes and processes denote annotation associations. Dotted lines represent annotations that are also present in GO. Wide edges represent an essential contribution of a gene to a process. Blue, red, and black edges represent contribution under aerobic, anaerobic, and both conditions, respectively. Note that several CDA annotations (e.g., in the pyrimidine biosynthesis cluster) have corresponding GO annotations that are highly nonspecific (Methods) and are hence considered here as novel.

rase) which contributes to the production of purines as well as to histidine and arginine biosynthesis under aerobic conditions by recycling the fumarate produced by these pathways.

Conservation coherency, expression coherency, multifunctionality, and evolutionary rate of CDA vs. GO annotated genes

Genes involved in the same biological processes have been previously shown to be coherently conserved in evolution (Pellegrini et al. 1999). Based on this observation, we tested the biological plausibility of CDA in comparison to GO by computing the conservation coherency of the sets of genes that are associated with similar terms in each annotation. To compute conservation coherency we used phylogenetic profiles of yeast metabolic genes in a set of 10 fungal genomes (Methods). The conservation coherency score is defined as the mean similarity in phylogenetic profiles between genes annotated with the same term; a significance *P*-value is computed by comparing this score to that obtained for a random annotation (i.e., random association of genes with terms; Methods). We find that the CDA is significantly coherently conserved to a higher extent than the corresponding GO annotation under all single-knockout conditions (Fig. 4A; Supplemental Table 3). The conservation coherency of the CDA in the double-knockout conditions is lower than that of the single-knockout conditions but is still highly statistically significant (Supplemental Table 4). The lower conservation coherency in the double-knockout conditions results from anti-correlated phylogenetic profiles of genes that back up each other's function (e.g., isozymes; Supplemental Section 1). Removing isozymes from the analysis increases the conservation coherency scores for the double-knockout CDA above the scores obtained with the single knockouts (Supplemental Table 4). The conservation coherency analysis conducted here is similar to the approach described in Dolan et al. (2005), where cross-species comparison was used to assess the consistency of GO annotations provided by different annotation groups.

We next compared the expression coherency of CDA and GO annotated genes. Expression coherency is defined as the mean similarity in expression patterns across different conditions between genes annotated with the same term, with a significant *P*-value computed by comparison to the expression coherency of a random annotation (Methods). CDA has a significantly higher expression coherency score than that of GO under all single-knockout conditions (Fig. 4B; Supplemental Table 5). CDA obtained with double knockouts is significantly coherently expressed, but to a lower extent (Supplemental Table 6). This lower expression coherency score is partially attributed to the availability of isozymes that tend to have anti-correlated expression patterns (Ihmels et al. 2004; Kafri et al. 2005). Removing isozymes from the analysis in turn increases the expression coherency score of the CDA under all conditions (Supplemental Table 6). These higher conservation and expression coherency scores of the CDA annotation in comparison with GO are robust to different choices of the multifunctionality threshold level (Supplemental Fig. 5).

Previously it was shown that gene multifunctionality level (as reflected by GO) is correlated with its degree of pleiotropy, the latter measured by the extent to which its deletion affects survival under multiple different environmental conditions (He and Zhang 2006). We repeated the same analysis here focusing on CDA- and GO-derived multifunctionality measures. There is a significant correlation between multifunctionality, as it is reflected by the CDA and pleiotropy (Spearman = 0.17, *P*-value = 0.03), and no correlation between GO multifunctionality (computed only for the corresponding metabolic genes) and pleiotropy. Furthermore, we find a significant negative correlation (Spearman = -0.22, *P*-value = 0.01) between the CDA multifunctionality measure and the gene evolutionary rate, in accordance to the negative correlation between pleiotropy and evolutionary rate shown in He and Zhang (2006). We find no corresponding correlation between the GO multifunctionality measure and conservation rate.

CDA-derived functional pathways

To gain insight on the metabolic mechanisms underlying biosynthetic processes, we use the CDA to derive condition-dependent, network-based representations of various processes, referred to as "functional pathways." These functional pathways can be automatically generated by integrating static metabolic

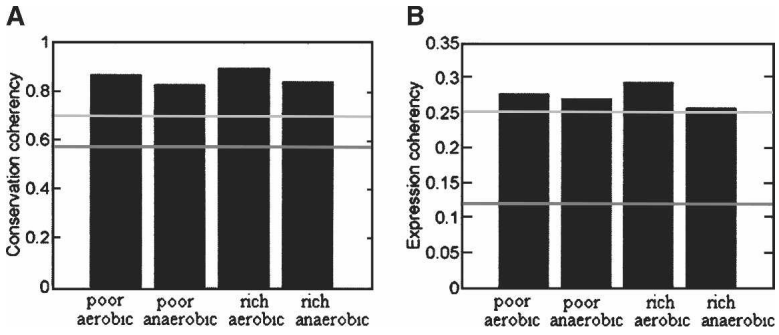

Figure 4. The conservation (*A*) and expression (*B*) coherency of the CDA under different conditions. The black and gray lines represent the coherency score obtained for a random annotation and for GO, respectively.

network structure with the CDA condition-dependent annotation (Methods). We examined in detail the functional pathways of alanine, proline, and glutamine biosynthesis, focusing on their condition-dependent nature. To further support the biological plausibility of specific novel functional pathways predicted by CDA, we performed phenotyping of relevant single and double gene deletion strains in auxotrophic growth conditions.

The functional pathway for alanine biosynthesis under poor media for single and double, aerobic and anaerobic conditions is shown in Figure 5A. Using single knockouts, only the gene *ALT2* (a putative cytoplasmic alanine transaminase) is predicted to contribute to alanine biosynthesis in both aerobic and anaerobic conditions. Under aerobic conditions, six additional genes belonging to the tryptophan biosynthetic pathway and the kynurenine pathway for tryptophan degradation are predicted to contribute to alanine biosynthesis based on the double-knockout analysis. Under anaerobic conditions their contribution vanishes due to oxygen dependence of the tryptophan degradation pathway. To identify potential additional alanine biosynthesis pathways, we tested whether the *ALT2* deletion strain is an alanine auxotroph in anaerobic conditions (data not shown). The deletion strain showed no growth defect in these conditions, suggesting the existence of an additional alternative pathway. The most likely backup for *ALT2* would be provided by *ALT1*, a mitochondrial isozyme of *ALT2*. *ALT1* was previously considered to be noncontributing to alanine production in the cytoplasm, due to the lack of a known mitochondrial alanine transporter. However, our experimental results suggest that there is an uncharacterized alanine transporter that allows mitochondrially synthesized alanine to be utilized outside the mitochondria in an *ALT2* deletion strain.

The functional pathway of proline biosynthesis consists of all the genes annotated in GO as being involved in proline biosynthesis as well as several novel CDA predictions (Fig. 5B). Proline is synthesized from L-glutamate gamma-semialdehyde, which in turn can be synthesized either through the standard proline biosynthesis pathway involving *PRO1* and *PRO2* gene products, or through the arginine catabolic pathway involving the *CAR1* and *CAR2* gene products. The arginine catabolic pathway is inactive when a preferred nitrogen source (e.g., ammonium sulphate) is available, but in the absence of preferred nitrogen sources it allows yeast to utilize arginine as a nitrogen source in aerobic conditions (Dubois et al. 1978). CDA predicts that *CAR2*, which codes for an ornithine transaminase, is also involved in proline biosynthesis, providing a backup function for *PRO1/PRO2* (Fig. 5B). In agreement with this prediction, the auxotrophy test for the *CAR2/PRO2* double-knockout strain shows a significant drop in maximum optical density (OD) in minimal media compared with the maximum OD measured in a minimal media supplemented with proline (Fig. 5C). Furthermore, we find a synthetic sick interaction between *CAR2* and *PRO2* in minimal media, with the double knockout showing a larger drop in maximum OD compared to the single knockouts. The fact that the *CAR2/PRO2* double knockout remains viable (though with a lower maximum OD) suggests the existence of an additional, uncharacterized parallel pathway that bypasses the *CAR2* deletion (Supplemental Fig. 6). Auxotrophy tests have also validated the CDA prediction of *CIT1* and *CIT3* involvement in glutamine biosynthesis (Supplemental Fig. 7).

Discussion

Utilizing a metabolic network model of *S. cerevisiae*, we derive a systematic condition-dependent annotation of yeast metabolic genes, associating genes with major metabolic processes under various genetic and environmental conditions. The resulting CDA, which promises to serve as a new reference source for metabolic gene annotation, is highly dependent on the growth media (poor or rich), the availability of oxygen, and whether single or double knockouts are employed. Under these conditions, the CDA maps annotations for 233 genes that are specifically contributing to some biosynthetic processes, and additional annotations for 62 genes that contribute to many processes and whose contribution is hence considered nonspecific (out of a total of 750 genes in the model). Considering that the annotation of many genes in the CDA was determined based on specific conditions (Table 1), we expect that extending the CDA to a variety of growth media and employing high-order knockouts (Deutscher et al. 2006) would significantly increase the number of annotated genes. The overlap between the CDA and GO is highly significant, with CDA achieving the highest GO enrichment under aerobic rich media, the common environmental condition in many of the experiments underlying GO. Indeed, FBA (like most model) predictions are prone to a certain level of false-positive and false-negative errors, but overall they have shown to provide a clear biological signal in recent, condition-dependent double-knockout studies (Deutscher et al. 2006; Harrison et al. 2007). Most importantly, a large-scale validation of the predicted CDA shows that genes annotated with the same term tend to be more coherently conserved in evolution and display greater expression coherency than the corresponding GO annotation, thus establishing the superiority of the new annotation over GO for metabolic genes, and testifying to its veracity.

The common view of metabolic pathways as static, distinct entities with a defined function may be misleading, considering the interconnectivity of different pathways through shared cofactors and metabolites. Here, we extend the classical notion of metabolic pathways into functional pathways, which are condition-dependent, network-based representations of biosynthetic processes. We examine in detail the functional pathways for alanine, glutamine, and proline biosynthesis, and demonstrate their

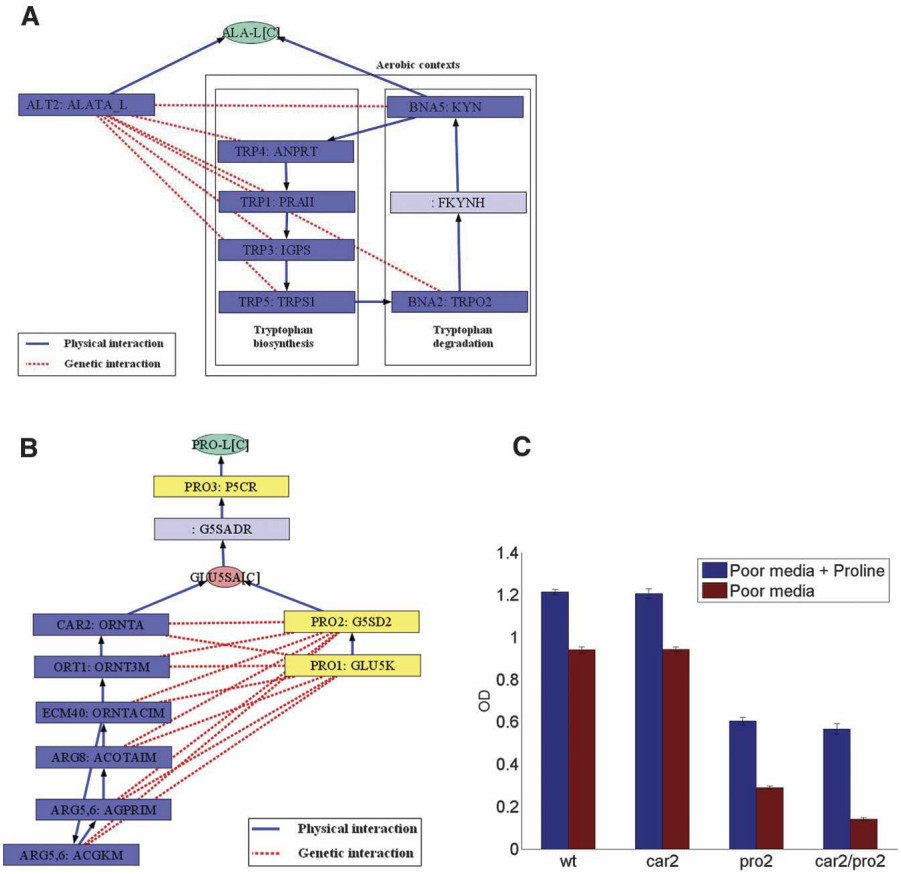

Figure 5. Functional pathways of alanine and proline biosynthesis as reflected by CDA. (*A*) A functional pathway of alanine biosynthesis under aerobic and anaerobic poor media. Rectangular nodes represent metabolic reactions, specifying names of the coding genes and names of the reactions. The circular node represents the metabolite alanine. Blue edges represent physical interactions between enzymes, in the form of a metabolite that is the product of one enzyme and the substrate of the other. Red edges represent genetic interactions between genes that are specific to alanine biosynthesis. (*B*) A functional pathway of proline biosynthesis in aerobic minimal media condition. (*C*) Proline auxotrophy experiment for the *CAR2* and *PRO2* single deletion, and *CAR2/PRO2* double deletion strains. The experiments show a significant drop in maximum optical density (OD) for the *CAR2/PRO2* double mutant strain when proline is removed from the growth medium. The results show the existence of a synthetic sick interaction between *CAR2* and *PRO2* in minimal media that lacks proline.

condition specificity. Experimental growth phenotyping of single and double gene deletion strains has verified the presence of the predicted, previously uncharacterized full and partial backup mechanisms in these pathways.

The CDA is inspired but different from the multidimensional genome annotation framework of Reed et al. (2006), where the use of the term dimension refers to different kinds of gene descriptors, such as the functional connectivity between genes and their three-dimensional organization. In difference, the dimensions in the CDA represent independent (mostly environmental) factors that affect gene involvement in different cellular processes. Specifically, we consider dimensions representing the growth medium, the availability of oxygen, and knockout level, giving rise to a set of common conditions for yeast growth. Future work may consider additional conditions by adding new dimensions to the annotation or by adding new coordinates to dimensions used here. Such an interesting new annotation dimension may represent genotypes of different yeast strains, allowing for a systematic comparison of gene function across strains. Additional coordinates for the growth media dimension may represent intermediate environments with different combinations of nutrients, providing further insight on the dependency of gene function on the presence of specific media nutrients.

Major efforts have been recently made to identify genetic interactions on a large scale, both experimentally (Tong et al. 2004; Schuldiner et al. 2005) and computationally (Segre et al. 2005; Deutscher et al. 2006). The annotation approach described here utilizes double-knockout simulations to identify genetic interactions between genes that are both process-specific and condition-specific. Overall, we find that genetic interactions display higher condition specificity than process specificity, as 37% of the interactions are predicted to be specific to a single process, whereas 64% of the interactions are specific to a single condition (Supplemental Fig. 8), further supporting the notion of strong condition dependency of genetic interactions recently reported in (Harrison et al. 2007). In summary, gene annotation remains a fundamental and open challenge, considering that complex processes are the result of the interaction of many genes and numerous factors. The condition-dependent annotation described here presents a new step in addressing this challenge for metabolic genes.

Methods

Metabolic network analysis

The metabolic network model of Duarte et al. (2004) consists of 1062 metabolites, 1149 reactions, and 750 genes. The model specifies a growth reaction that contains the following 38 essential biomass precursors: (1) Amino acids: methionine, aspartate, glutamate, glutamine, asparagine, alanine, proline, arginine, serine, cysteine, glycine, histidine, threonine, lysine, valine, tyrosine, tryptophan, phenylalanine, leucine, and isoleucine; (2) Carbohydrates: glycogen, trehalose, mannan, and glucan; (3) Nucleotides: UMP, GMP, CMP, AMP, DGMP, DAMP, DTMP, and DCMP; and (4) Fatty acids, sterols, lipids, and phospholipids: triacylglycerol, ergosterol, phosphatidylcholine, phosphatidylethanolamine, 1-phosphatidyl-D-myo-inositol, phosphatidylserine.

Flux Balance Analysis (FBA) was used to compute the production rate of each biomass precursor under various growth media and genetic environments. To simulate the production of a given metabolite, a new exchange reaction representing the secretion of this metabolite is added to the model, and the flux through this reaction is maximized. For the single-knockout annotation we systematically knocked out each gene and considered it as contributing to the production of a certain metabolite if its knockout reduced the metabolite's production rate in >20%. For the double-knockout annotation, we knocked out all gene pairs whose genes were noncontributing in the single-knockout

experiments and considered a pair as contributing if the joint knockout reduced the metabolite's production rate in >20% (a similar threshold was used in Deutscher et al. 2006). All of the results presented here are robust to the choice of this production rate threshold (we repeated the analysis for thresholds of 10% and 30%; data not shown). Furthermore, adding a lower bound on the biomass production (and hence accounting for the resources required for the production of all growth enabling biomass compounds) when computing the maximal production rate of some compound (without knockouts) has negligible effect on the identification of gene contributions (data not shown). Poor and rich media as well as aerobic and anaerobic conditions were modeled by varying the bounds on the uptake rates of various nutrients.

GO annotation extraction

Yeast GO annotation was downloaded from the SGD database. For each of the 38 biomass compounds included in the model, we identified the corresponding GO term representing the process of synthesizing this compound (Supplemental Data). For each GO term, we associated all genes that are either annotated to this term or annotated to its ancestors (situated on the pathway leading to the root) in the GO hierarchy. This extraction method was applied to obtain a comprehensive process annotation, as in numerous cases genes that are known to be involved in a specific process are annotated with one of its ancestor terms in the GO hierarchy (e.g., in the case of valine, leucine, and isoleucine biosynthesis, only the parent term called "branched chain amino acid biosynthesis" is associated with genes that are known to be involved in the synthesis of all three amino acids). As in the CDA construction, to eliminate annotations that are highly nonspecific we excluded genes with a multifunctionality level greater than a threshold of five.

Phylogenetic profiling analysis

Ten sequenced fungal genomes (*S. cerevisiae*, *Candida albicans*, *Candida glabrata*, *Cryptococcus neoformans*, *Debaryomyces hansenii*, *Encephalitozoon cuniculi*, *Eremothecium gossypii*, *Kluyveromyces lactis*, *Schizosaccharomyces pombe*, *Yarrowia lipolytica*) were used to construct phylogenetic profiles. The conservation coherency of a metabolic annotation (i.e., either GO or the CDA) is defined as the mean similarity of phylogenetic profiles between genes annotated to the same term. Similarity between phylogenetic profiles is computed using a Jaccard measure (intersection/union). The statistical significance of the conservation coherency score is computed by comparison to similar scores obtained for randomly drawn annotations, preserving the same number of genes annotated with each term.

Expression coherency analysis

Gene expression measurements were obtained from Stanford microarray database (Ball et al. 2005) and included 973 conditions. Expression coherency of metabolic annotation is computed as the mean pairwise Pearson correlations between genes annotated to the same term. The statistical significance of the expression coherency score is computed by comparison to similar scores obtained for randomly drawn annotations, preserving the same number of genes annotated with each term.

Functional pathways construction

A functional pathway, representing the process of synthesizing a certain compound under a certain environment, is a subgraph of the metabolic network (in a "reaction graph" representation) augmented with functional information on genetic backups. In this representation, nodes represent metabolic reactions (displayed along with their associated coding genes) and directed edges represent the existence of a metabolite that is produced by one reaction and consumed by the other. Metabolites participating in more than two reactions are represented as additional vertices in the graph. Process-specific genetic interactions between gene pairs are represented as additional edges between the genes.

A functional pathway describing the production of a target compound under a given condition is derived via the following algorithm:

1. For each gene, *g*, annotated in the CDA as involved in the compound's production under this condition:
   a. Use FBA to predict steady-state flux distribution that maximizes the production rate of the compound.
   b. Let *G* denote the subset of the metabolic network consisting of active metabolic reactions based on the predicted flux distribution (having non-zero flux) with all currency metabolites removed (Supplemental Table 7).
   c. Find the shortest path between the target compound and the gene *g* (via breadth first search in *G*) and add it to the functional pathway.
2. For each gene, *g*, annotated in the CDA as involved in the compound's production using double knockout under this condition, that is backed up by gene *b*:
   a. Repeat steps a–c while knocking out the backup gene *b*.
   b. Add a synthetic sick interaction between *g* and *b* to the functional pathway.

Experimental procedures

The single gene deletion strains were in the BY4741 strain background (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) and were obtained from Open Biosystems. The *CIT3/CIT1* and *CAR2/PRO2* double gene deletion strains were constructed in a single gene deletion strain background (*cit3Δ* or *car2Δ*) using the standard one-step method using *LEU2* as a marker (Christianson et al. 1992; Baudin et al. 1993). All the single and double deletion strains were verified using PCR with the appropriate gene and marker-specific primers (Winzeler et al. 1999). Growth rates of wild type and gene deletion strains were evaluated by using the Bioscreen C system (Thermo Labsystems). Minimal medium (Van Hoek et al. 1998) with the appropriate auxotrophic supplements for the BY4741 background (methionine, leucine, histidine, and uracil) and additional supplements specific to each experiment (alanine, glutamine, proline, ornithine) was used to test growth in aerobic conditions. In the experiments involving the car2Δ, pro2Δ, and car2Δpro2Δ strains (Fig. 5; Supplemental Material) minimal media without ammonium sulphate was used in order to avoid nitrogen catabolite repression of proline uptake (methionine, leucine, and histidine were provided in excess as nitrogen sources). Maximum specific growth rates and ODs were determined using the Bioscreen C system based on at least three independent 48-h growth curves obtained as described in Herrgard et al. (2006).

Acknowledgments

We thank Kai Tan for providing the phylogentic profiles, Pep Charusanti for help with verifying knockout strains, and David Deutscher and Elhanan Borenstein for very helpful discussions. T.S. is supported by an Eshkol Fellowship from the Israeli Ministry of Science. R.S. was supported by an Alon Fellowship. E.R.'s research is supported by the Yishayahu Horowitz Center for Complexity Science, the Israeli Science Foundation (ISF), the Ger-

man-Israeli Foundation for scientific research and development (GIF), and the Tauber fund. M.H. and V.P. are supported by the National Institutes of Health (RO1 GM071808) and the National Science Foundation (BES-0331342).

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Ball, C.A., Awad, I.A., Demeter, J., Gollub, J., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Matese, J.C., Nitzberg, M., Wymore, F., et al. 2005. The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* **33:** D580–D582. doi: 10.1093/nar/gki006.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., and Cullin, C. 1993. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21:** 3329–3330. doi: 10.1093/nar/21.14.3329.

Brown, J.A., Sherlock, G., Myers, C.L., Burrows, N.M., Deng, C., Wu, H.I., McCann, K.E., Troyanskaya, O.G., and Brown, J.M. 2006. Global analysis of gene function in yeast by quantitative phenotypic profiling. *Mol. Syst. Biol.* **2**. doi: 10.1038/msb4100043.

Cherest, H., Thomas, D., and Surdin-Kerjan, Y. 1993. Cysteine biosynthesis in *Saccharomyces cerevisiae* occurs through the transsulfuration pathway which has been built up by enzyme recruitment. *J. Bacteriol.* **175:** 5366–5374.

Christianson, T.W., Sikorski, R.S., Dante, M., Shero, J.H., and Hieter, P. 1992. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110:** 119–122.

Deutscher, D., Meilijson, I., Kupiec, M., and Ruppin, E. 2006. Multiple knockouts analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* **38:** 993–998.

Deytieux, C., Mussard, L., Biron, M.J., and Salmon, J.M. 2005. Fine measurement of ergosterol requirements for growth of *Saccharomyces cerevisiae* during alcoholic fermentation. *Appl. Microbiol. Biotechnol.* **68:** 266–271.

Dolan, M.E., Ni, L., Camon, E., and Blake, J.A. 2005. A procedure for assessing GO annotation consistency. *Bioinformatics* **21 (Suppl. 1):** i136–i143.

Duarte, N.C., Herrgard, M.J., and Palsson, B.O. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14:** 1298–1309.

Dubois, E., Hiernaux, D., Grennon, M., and Wiame, J.M. 1978. Specific induction of catabolism and its relation to repression of biosynthesis in arginine metabolism of *Saccharomyces cerevisiae*. *J. Mol. Biol.* **122:** 383–406.

Edwards, J.S. and Palsson, B.O. 2000. The *Escherichia coli* MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci.* **97:** 5528–5533.

Edwards, J.S., Ibarra, R.U., and Palsson, B.O. 2001. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19:** 125–130.

Famili, I., Forster, J., Nielsen, J., and Palsson, B.O. 2003. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci.* **100:** 13134–13139.

Farkas, I., Hardy, T.A., Goebl, M.G., and Roach, P.J. 1991. Two glycogen synthase isoforms in *Saccharomyces cerevisiae* are coded by distinct genes that are differentially controlled. *J. Biol. Chem.* **266:** 15602–15607.

Forster, J., Famili, I., Palsson, B.O., and Nielsen, J. 2003. Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*. *OMICS* **7:** 193–202.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.

Harrison, R., Papp, B., Pal, C., Oliver, S.G., and Delneri, D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl. Acad. Sci.* **104:** 2307–2312.

Hartman, J.L.T., Garvik, B., and Hartwell, L. 2001. Principles for the buffering of genetic variation. *Science* **291:** 1001–1004.

He, X. and Zhang, J. 2006. Toward a molecular understanding of pleiotropy. *Genetics* **173:** 1885–1891.

Herrgard, M.J., Lee, B.S., Portnoy, V., and Palsson, B.O. 2006. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res.* **16:** 627–635.

Ibarra, R.U., Edwards, J.S., and Palsson, B.O. 2002. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420:** 186–189.

Ihmels, J., Levy, R., and Barkai, N. 2004. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **22:** 86–92.

Johnson, D.R., Knoll, L.J., Levin, D.E., and Gordon, J.I. 1994. *Saccharomyces cerevisiae* contains four fatty acid activation (FAA) genes: An assessment of their role in regulating protein N-myristoylation and cellular lipid metabolism. *J. Cell Biol.* **127:** 751–762.

Kafri, R., Bar-Even, A., and Pilpel, Y. 2005. Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* **37:** 295–299.

Kuepfer, L., Sauer, U., and Blank, L.M. 2005. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15:** 1421–1430.

Mahadevan, R. and Schilling, C.H. 2003. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5:** 264–276.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96:** 4285–4288.

Reed, J.L., Famili, I., Thiele, I., and Palsson, B.O. 2006. Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7:** 130–141.

Schilling, C.H., Letscher, D., and Palsson, B.O. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203:** 229–248.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. 2005. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507–519.

Schuster, S., Fell, D.A., and Dandekar, T. 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18:** 326–332.

Segre, D., Deluna, A., Church, G.M., and Kishony, R. 2005. Modular epistasis in yeast metabolism. *Nat. Genet.* **37:** 77–83.

Shlomi, T., Berkman, O., and Ruppin, E. 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci.* **102:** 7695–7700.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813.

Van Hoek, P., Van Dijken, J.P., and Pronk, J.T. 1998. Effect of specific growth rate on fermentative capacity of baker's yeast. *Appl. Environ. Microbiol.* **64:** 4226–4233.

van Swinderen, B. and Greenspan, R.J. 2005. Flexibility in a gene network affecting a simple behavior in *Drosophila melanogaster*. *Genetics* **169:** 2151–2163.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.

Received May 5, 2007; accepted in revised form August 6, 2007.