# PROPHECY—a database for high-resolution phenomics

Luciano Fernandez-Ricaud, Jonas Warringer, Elke Ericson, Ilona Pylvänäinen[1], Graham J. L. Kemp[2], Olle Nerman[1] and Anders Blomberg*

Department of Cell and Molecular Biology, Lundberg Laboratory, Göteborg University Medicinaregatan 9c, 41390 Göteborg, Sweden, [1]Department of Mathematical Statistics and [2]Department of Computing Science, Chalmers University of Technology, SE-41296 Göteborg, Sweden

## ABSTRACT

**The rapid recent evolution of the field phenomics—the genome-wide study of gene dispensability by quantitative analysis of phenotypes—has resulted in an increasing demand for new data analysis and visualization tools. Following the introduction of a novel approach for precise, genome-wide quantification of gene dispensability in *Saccharomyces cerevisiae* we here announce a public resource for mining, filtering and visualizing phenotypic data—the PROPHECY database. PROPHECY is designed to allow easy and flexible access to physiologically relevant quantitative data for the growth behaviour of mutant strains in the yeast deletion collection during conditions of environmental challenges. PROPHECY is publicly accessible at http://prophecy.lundberg.gu.se.**

## INTRODUCTION

A fundamental approach in determining the cellular role of functionally unclassified genes is to determine the consequences of gene loss. Recently, systematic approaches towards the targeted inactivation of each gene in a genome have been completed in model organisms such as the yeast *Saccharomyces cerevisiae* (1) and the nematode *Caenorhabditis elegans* (2). These initiatives have opened the door to phenomics; a novel field with the aim of providing systematic descriptions of phenotypic characteristics of gene deletion mutants on a genome-wide scale. We have recently reported on a high-resolution system aimed at the exact quantification of the consequences of the loss of every individual gene in the yeast genome (3). In this system each strain is microcultivated in isolation providing individual growth curves that are reduced to precise, quantitative data in a fully automated

manner. Thus, the accumulated data reflect the importance of every gene in the deletion collection with regard to the time required to adapt to an environmental challenge, the rate of growth and the efficiency of growth. Here, we introduce and
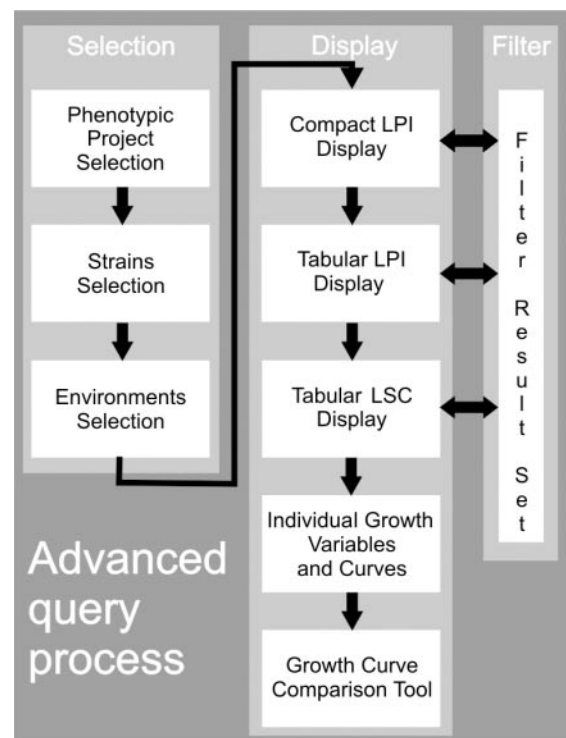


**Figure 1.** The Advanced Query data selection process: overview of the Advanced Query process in PROPHECY that is designed to guide the user through the selection criteria in a wizard/assistant based fashion via a linear querying process by (i) selecting a dataset, (ii) selecting strains/genes and (iii) selecting environmental stress conditions. Phenotypic data can then be displayed at different levels of abstraction (LPI, LSC and growth variables and curves). The resulting sets can be filtered by degree of growth defect or functional classification.

*To whom correspondence should be addressed. Tel: +46 31 7732588; Fax: +46 31 7732599; Email: anders.blomberg@gmm.gu.se

describe the yeast phenomics database PROPHECY—PROfiling of PHEnotypic Characteristics in Yeast (http://prophecy.lundberg.gu.se)—designed to mine, filter and visualize genome-wide gene dispensability data in an easy to use manner. PROPHECY, constituting the first online resource for yeast high-resolution growth data, allows for the visualization and quantitative evaluation of phenotypes with the simultaneous integration of external functional genomics data.

## DATA STANDARDIZATION

PROPHECY stores and evaluates phenotypes of deletion strains on the basis of growth behaviour during environmental challenges in a microcultivation format (4). PROPHECY provides a precise measure of gene-by-environment interactions by quantifying three standard variables of growth: the rate of growth reflected in the generation time, the efficiency of growth reflected in the cell density at stationary phase and the time to adapt to an environmental challenge reflected in the length of the lag phase. These values provide quantitatively precise data with high physiological resolution on the importance of every investigated gene during growth in defined environments. To reduce experimental variance and provide long-term standardization, eight reference strains (wild-type) are included in each experiment [one experiment corresponds to the analysis in one instrument of 200 strains in

a particular environment (4)]. The growth behaviour of each deletion strain is related to the average behaviour of these eight reference strains, forming reference strain normalized growth measures, termed logarithmic strain coefficients ($LSC_{rate}$, $LSC_{adaptation}$ and $LSC_{efficiency}$) as defined by Warringer *et al.* (3). Mean variance (coefficient of variation) is reduced from 7.2 to 0.5% for $LSC_{rate}$, from 12.3 to 1.7% for $LSC_{efficiency}$ and from 15.2 to 1.6% for $LSC_{adaptation}$ by this normalization/standardization procedure [mean variance over 18 experiments over a period of 2 years and using five different instruments (3)].

Furthermore, to provide a quantitative measure of the specific gene-by-environment interactions and to compensate for general growth defects observed even under favourable growth conditions, LSC from conditions of environmental stress are related to LSC from favourable (no stress) conditions; LSCs with and without stress are thus combined into logarithmic phenotypic indexes ($LPI_{rate}$, $LPI_{adaptation}$ and $LPI_{efficiency}$) (3).

## DATABASE DESIGN AND IMPLEMENTATION

PROPHECY follows a centralized data integration philosophy—all data reside in a single data repository. Thus, not only the locally generated phenotypic data but also data from external data sources, notably MIPS (5), SGD (6) and GRID (7), are stored centrally. All external data sources are formatted and
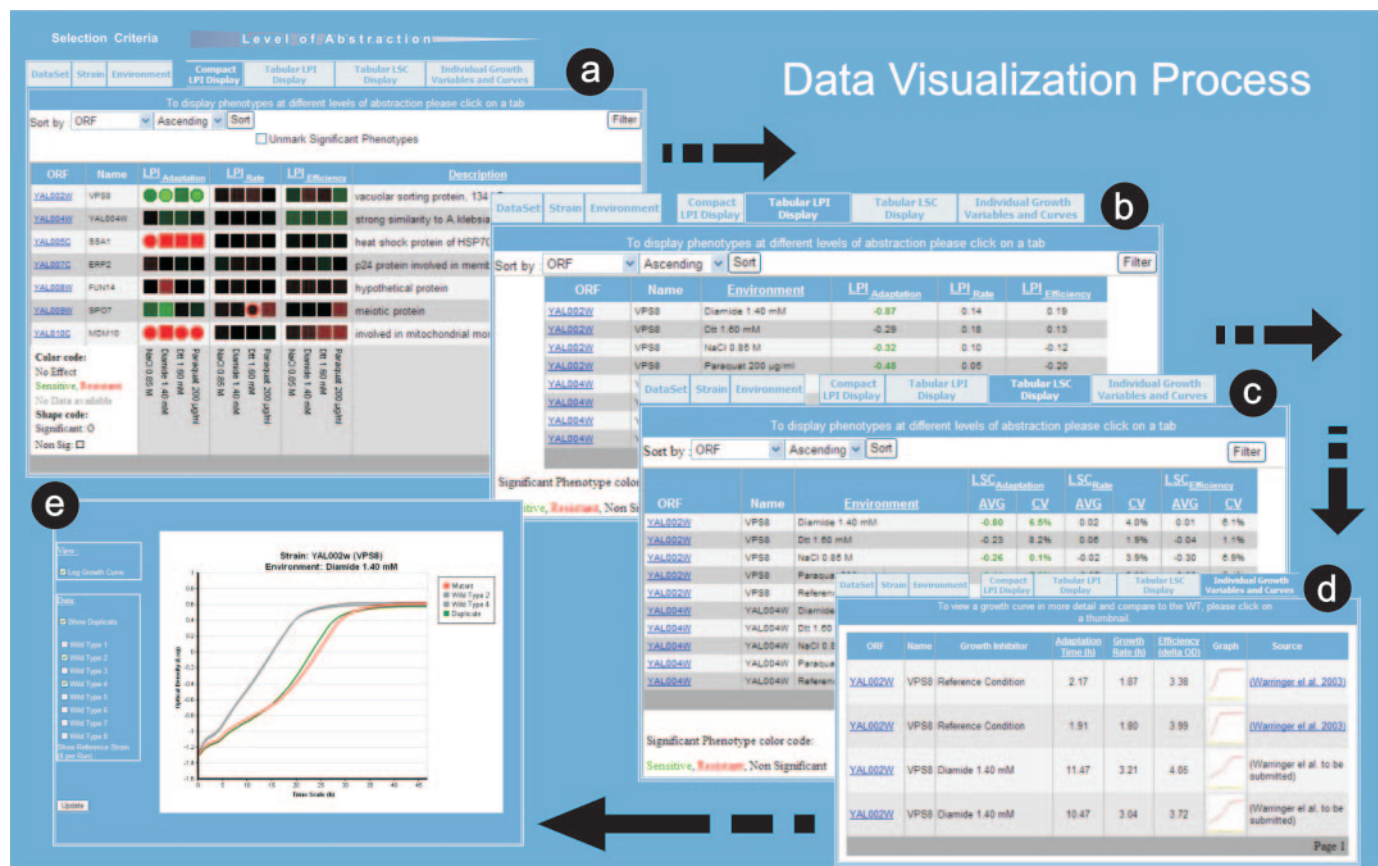


**Figure 2.** The different levels of data display: data output formats following the Advanced Query. Data are presented at progressively lower levels of abstraction: (**a**) the graphical Compact LPI Display, (**b**) the Tabular LPI Display, (**c**) the Tabular LSC Display, (**d**) the Individual Growth Variables and Curves display and (**e**) the Growth Curve Comparison Tool display (activated by clicking on the growth curve thumbnails).

mapped onto the specific data structure of the PROPHECY database. The PROPHECY data structure follows a relational data storage model and is implemented using a Microsoft SQL Server 2000. All PROPHECY features are developed using Microsoft.NET technology (http://www.microsoft.com/vstudio) in a Windows environment. PROPHECY consists of a database server, a web server and a data mining application, and the system is implemented in Visual Basic.NET (http://www.microsoft.com/vbasic) and CS.NET (http://www.microsoft.com/vcsharp) while the web interface is implemented using the ASP.NET (http://www.asp.net) framework.

## DATABASE USER INTERFACE

### Data query

The phenotypic data in the database can be queried through the web interface in two ways: (i) Quick Search visualizes data for one gene at a time, where the user enters either the open reading frame (ORF) name or the gene name to be investigated or (ii) Advanced Query allows analysis of many gene deletions at the same time and provides options for filtering, which requires input of a number of detailed selection criteria (Figure 1). The PROPHECY selection interface is designed so that the user is guided from step to step until the query is complete. When using the Advanced Query the user selects (i) a phenotypic project (ii) any number of genes, pasted into an entry box or selected from a full gene list and (iii) any number of environmental challenges selected from a list.

### Data display

After the selection criteria are established the quantitative phenotypic data are displayed at three levels of abstraction, the LPI, the LSC and the actual growth variables for each of the three physiological windows—adaptation time, rate of growth and efficiency of growth. At the highest level of abstraction, the LPI level, the data provide a quantitative measure of the importance of a gene in a particular environment in relation to the importance of the gene in favourable conditions, and in relation to how the environment influences the reference strain. Thus, an LPI = 0 indicates that the gene-by-environment response of the mutant is identical to the response of the wild-type. An LPI < 0 indicates an environmentally related growth defect of the deletion strain, the lower the value the graver the environment-specific growth defect. An LPI > 0 indicates that the deletion strain is tolerant to that particular environment in comparison to the reference strain. Data at the level of LPI are represented in two displays, first condensed in a Compact LPI Display [Figure 2, (a)], where gene data are visualized in a graphical format, i.e. colour (green=sensitive, red=tolerant) and shape coded (square=non-significant, round= significant at $P < 0.001$), and then in a Tabular LPI Display where the actual LPI values are found [Figure 2, (b)]. The next lower level of abstraction, the Tabular LSC Display [Figure 2, (c)], provides the LSC values from which the LPI values were derived. LSC in itself is also a measure of the fitness of the deletion strain as compared to the reference strain with no regard to the corresponding relation in favourable conditions. The lowest level of abstraction is reached at the growth variables, where raw, non-normalized, non-standardized growth variables, i.e. adaptation time, rate of growth and efficiency of growth, are provided for all individual strain replicates. Data at the growth variable level are also displayed in two formats; the actual growth variables represented in the tabular form and the individual growth curves (which form the core of the database) represented in graphical format [Figure 2, (d)]. Clicking on the growth curve thumbnails brings up a curve comparison tool where the duplicates for each mutant together with all the eight wild types for that particular run can be visually compared [Figure 2, (e)], displaying all the growth curves that are the base for the LSC value for that particular gene deletion.

### Data filtering and integration features

Advanced Query mode in PROPHECY features filtering capabilities that facilitate multidimensional data mining.
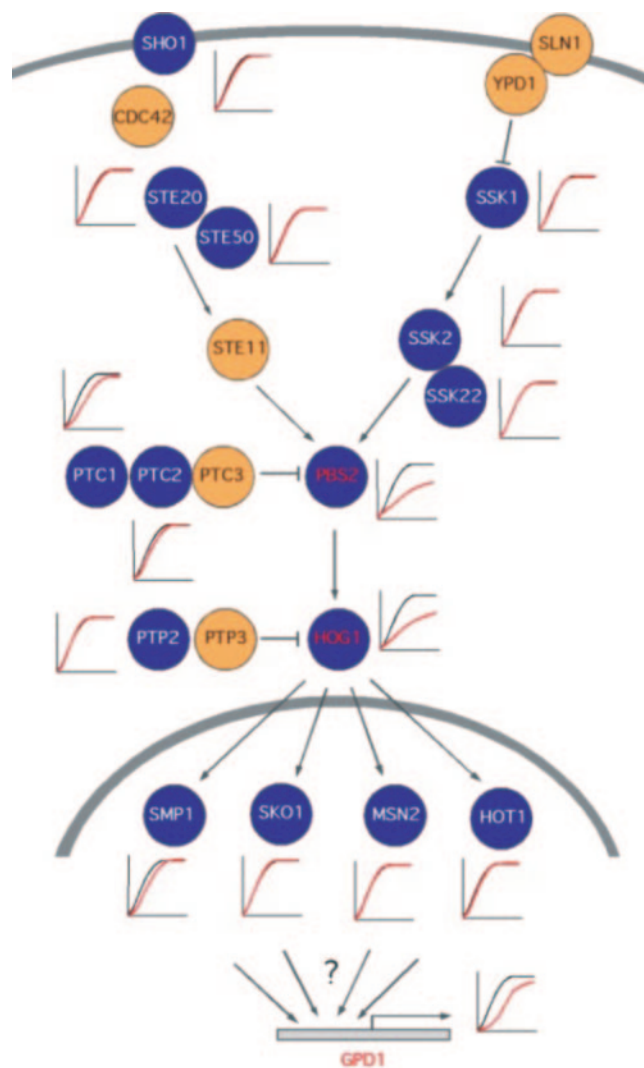


**Figure 3.** Functional classification filtering allows dissection of biochemical pathways: visualizing the growth behaviour of the components of a signalling pathway—the HOG pathway—during saline stress. Deletion strains are represented by red growth curves, representative reference strain by black curves. Gene names in red indicate significant (LSC < 0; $P < 0.001$) phenotypes. Yellow circles indicate strains not present in the yeast deletion strain collection.
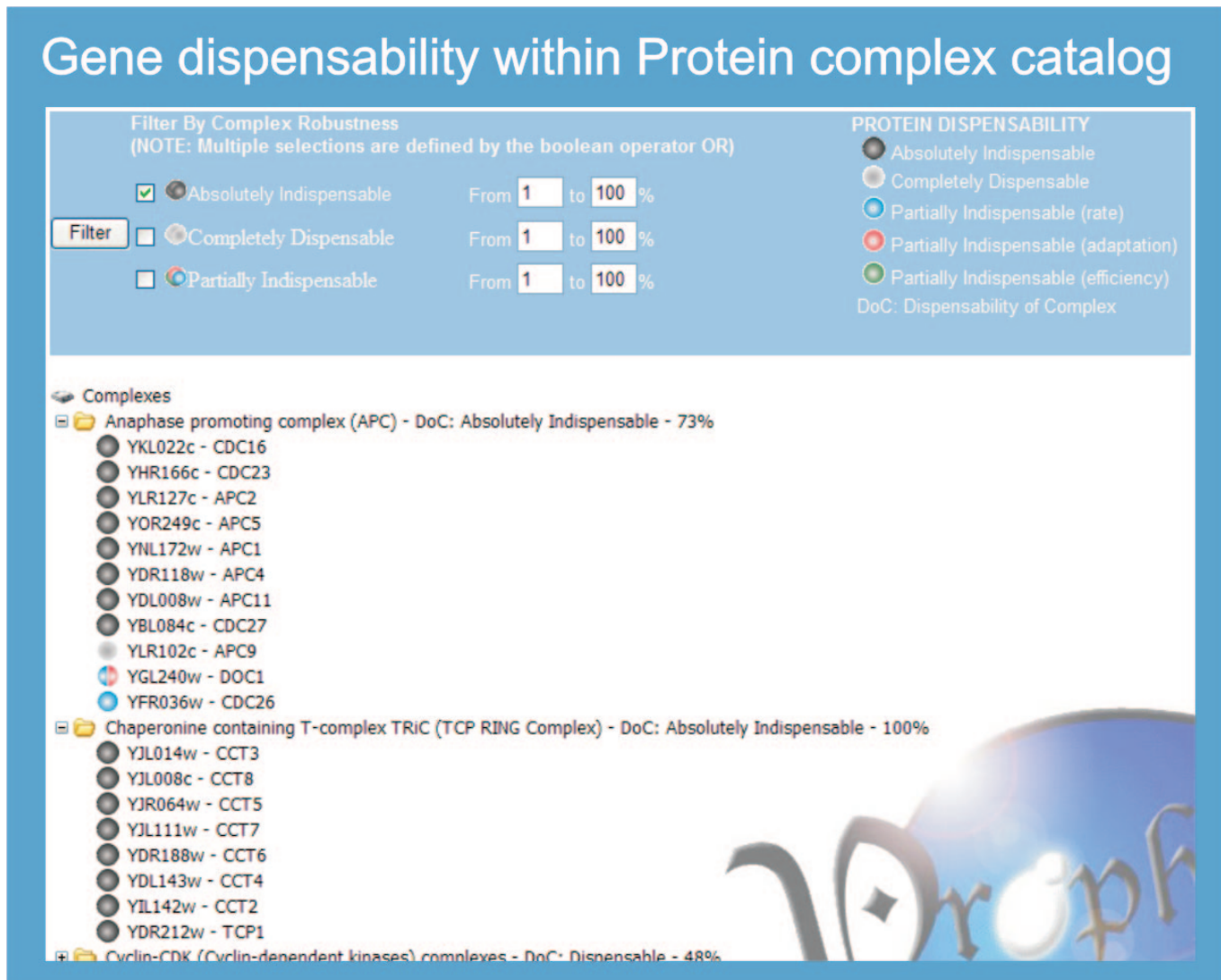
**Figure 4.** Integrating gene dispensability and protein complex data: visualizing robustness within protein complexes by integrating MIPS protein complex data with PROPHECY gene dispensability data. The protein complex catalogue can be filtered by the degree of phenotypic homogeneity in the complex. The robustness of the protein complexes are classified according to the majority of phenotypes scored for the removal of individual complex components.

Currently, PROPHECY provides two modes of filtering: (i) filtering by degree of growth defect, either on the level of LPI or on the level of LSC and (ii) filtering by functional classification using the MIPS functional classification catalogue (5). These filters may be applied separately or in combination, the latter allowing holistic dissection of phenotypic behaviour within discrete biochemical pathways and cellular processes at different levels of resolution. One example of such a dissection is portrayed in Figure 3 where the growth behaviour during saline stress from the deletion of components of the High Osmolarity Glycerol (HOG) signalling pathway and suggested downstream transcriptional activators is visualized. The HOG pathway, which resides in the stress response subcategory, is believed to consist of two parallel, functionally partially overlapping branches that are joined into one at the level of Pbs2p (Figure 3). This view is substantiated by the observed phenotypes as the upper parallel branches are individually dispensable during saline growth while deletion of *PBS2* and *HOG1*, the supposedly unbranched part of the pathway, confers very grave and similar growth defects (Figure 3).

Another integration feature in PROPHECY allows the user to explore the robustness of protein complexes, i.e. how well-known protein complexes tolerate loss of individual complex components. This option, selected from the main menu, couples the latest release of the MIPS protein complex catalogue to the PROPHECY data on gene dispensability, thus providing a tool for dissection of the phenotypic consequences of gene loss within discrete protein complexes. The resulting graphical display denotes a complex as either (i) absolutely indispensable—meaning that the majority of its components are absolutely required for growth (essential for viability), (ii) partially dispensable—meaning that the majority of its components are of significant importance (LSC < 0; $P < 0.001$) for uncompromised growth, (iii) completely dispensable—meaning that the majority of its components are not of significant importance for uncompromised growth, or (iv) phenotypically heterogeneous, i.e. no phenotypic class is in majority within

the complex. The search results are further visually detailed by phenotypic annotation of each complex component (Figure 4). Filtering for different ranges of protein complex robustness is possible (Figure 4). The protein complex robustness feature is currently only available for phenotypic data (LSC) derived from growth during favourable conditions.

### Future perspectives

PROPHECY is continuously updated with growth data derived from additional environmental challenges when articles are published and data made public. Most of these data are, and will be, genome-wide and correspond to haploid deletion strains. Parts of the data, however, will relate to smaller subsets of genes (100–600) analysed during a wide array of environmental stress conditions as well as to diploid, heterozygote or homozygote deletion strains. The web interface will also be continuously updated to provide increased flexibility and power in data integration and visualization, thus allowing, e.g. the display of gene dispensability within the biochemical pathway in a graphical format, and to incorporate novel filtering mechanisms, among them filtering for expression changes, subcellular localization and data on genetic interactions.

## REFERENCES

1. 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.
2. Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.
3. Warringer,J., Ericson,E., Fernandez,L., Nerman,O. and Blomberg,A. (2003) High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc. Natl Acad. Sci. USA*, **100**, 15724–15729.
4. Warringer,J. and Blomberg,A. (2003) Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae*. *Yeast*, **20**, 53–67.
5. Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
6. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M. *et al.* (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
7. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol.*, **4**, R23.