D700–D705 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 21 November 2011*
doi:10.1093/nar/gkr1029

Saccharomyces Genome Database: the genomics resource of budding yeast

J. Michael Cherry*, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng and Edith D. Wong

Department of Genetics, Stanford University, Stanford, CA, 94305-5120, USA

Received September 15, 2011; Revised October 19, 2011; Accepted October 21, 2011

ABSTRACT

The *Saccharomyces* Genome Database (SGD, http://www.yeastgenome.org) is the community resource for the budding yeast *Saccharomyces cerevisiae*. The SGD project provides the highest-quality manually curated information from peer-reviewed literature. The experimental results reported in the literature are extracted and integrated within a well-developed database. These data are combined with quality high-throughput results and provided through Locus Summary pages, a powerful query engine and rich genome browser. The acquisition, integration and retrieval of these data allow SGD to facilitate experimental design and analysis by providing an encyclopedia of the yeast genome, its chromosomal features, their functions and interactions. Public access to these data is provided to researchers and educators via web pages designed for optimal ease of use.

INTRODUCTION

Saccharomyces cerevisiae has been widely utilized in the exploration of biochemistry, molecular biology, cell biology and systems biology because of the ease with which it can be grown and manipulated, the extensive conservation of its genes and pathways with those of higher organisms, and the powerful genetic techniques that it offers. Many new genome-wide technologies—the creation of bar-coded systematic deletion sets; large-scale detection of protein–protein and genetic interactions and subcellular localizations; transcriptome, proteome and metabolome analysis—were first developed using yeast before being more widely applied to other organisms. For these reasons *S. cerevisiae* research serves as a model for a variety of processes that occur in humans, many of which are disease-associated. Direct parallels can be drawn between the two organisms in areas such as lipid metabolism (1), the unfolded protein response (2), mitochondrial metabolism (3), prion development (4) or ageing (5), to name just a few possible examples. The core users of *Saccharomyces* Genome Database (SGD) include those conducting research on *Saccharomyces* species as well as those exploring the genetics and cellular biology of other fungal genera that are important to basic research, human health and industry. Researchers studying larger organisms, including models such as *Drosophila* and *Caenorhabditis*, as well as plants and humans, represent growing communities that look to SGD for information when their research leads to genes with similarity to one of the many that are already well characterized in yeast. Educators and students in genetics and cellular biology comprise another large community that SGD serves, as do bioinformatics scientists who perform genome-wide computational analyses, for either yeast or comparative studies.

In this update the organization and enhancement to the SGD resource is presented in three parts: the representation and integration of experimental results into a rich biological model, new query and display tools that unlock discovery of this information, and new social media outlets to stay informed of new features and data provided by SGD.

REPRESENTATION OF EXPERIMENTAL RESULTS: CAPTURING, MAINTAINING AND INTEGRATING HIGH-QUALITY ANNOTATIONS

We define curation as the comprehensive and accurate manual extraction of experimental results from

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Email: cherry@stanford.edu

peer-reviewed literature. At SGD, curation has been, and will continue to be, a core responsibility that we conduct supplying a valued service to biomedical research communities. Providing comprehensive information for gene products, derived via literature curation, is of special value because it aids experimental design, facilitates gene function discovery and enhances gene product annotations provided by other genomic resources.

Structured controlled vocabularies (i.e. ontologies) include precise definitions representing conserved common biological concepts, thereby ensuring a consistent interpretation of results being represented. Controlled vocabularies facilitate the integration of diverse data types and promote computational analysis. SGD uses controlled vocabularies to describe experimental results selected from the complete literature on budding yeast. This approach has allowed us to maintain the gold standard of annotations describing the complete representation of yeast chromosomal features, their functions and interactions. The quality of our literature annotations as a whole is addressed regularly with group curation exercises to ensure that information is entered into SGD accurately and consistently. In addition, we are actively researching and developing new methods for quality control of Gene Ontology (GO) functional annotations (6). These efforts in ensuring consistency and quality, coupled with a quick review of new papers after publication, ensure the gold standard annotations are in pace with scientific progress, as defined by experimental results.

Gene Ontology (GO) annotations

Manual literature curation provides the highest quality representation of published experimental results. At SGD, annotation using GO terms captures biological information regarding specific gene products in a searchable, computable form that can be readily compared across organisms. Manual GO annotations for specific gene products are chosen carefully within the context of all available biological knowledge, and represent the best possible summary of the most relevant information. The annotations for each gene product are reviewed as a whole, and updated as new information becomes available, a time- and resource-intensive process.

GO annotations may also be assigned via computational methods based on criteria such as protein domains, HMMs, or patterns of experimental data, allowing the rapid annotation of multiple genes using consistent parameters. Limitations of computational predictions do exist: such as incomplete genomic coverage; the use of too-general GO terms; and the inherent biases of different methods (6). In 2007, SGD began including computational predictions from the Gene Ontology Annotation (GOA) project at UniProtKB (7,8), whose InterPro tools rely in part on the presence of experimentally characterized protein domains with manually curated connections to GO terms.

Researchers often transfer GO functional annotations between homologous proteins, thus the accuracy of SGD's literature annotations is of utmost importance (9,10). In order to identify efficient methods to keep GO annotations comprehensive and current, we explored whether the large set of computational predictions could be leveraged to find inaccuracies or omissions in our manual literature-based GO annotations. In an initial feasibility study (6), we examined a subset of genes that had a manual GO annotation that was less specific than the computational prediction. Specifically we selected genes that lacked published functional information (manual 'unknown' annotations) but had InterPro computational predictions. Review of the publications for these genes found that a significant proportion of the manual annotations could be updated to a defined, experimentally supported function. In contrast, review of the literature for a comparable control set of genes (those with manual 'unknown' annotations, but without corresponding InterPro computational predictions) did not identify as many possible improvements. Thus, we have demonstrated that the presence of InterPro computational predictions can effectively identify annotations that need to be updated.

We expanded this strategy so that each literature-based annotation for a gene was compared to computational predictions generated by different methods in each of the three GO aspects (Molecular Function, Biological Process, and Cellular Component) (Park,J., Costanzo,M.C., Balakrishnan,R., Cherry,J.M. and Hong, E.L. CvManGO, a method for leveraging computational predictions to improve literature-based GO annotations. *submitted to Database for Biocuration 2012 virtual issue*.). The comparison between literature-based annotations and computational predictions identified two types of literature-based annotations that needed review: those that were less specific than the computational predictions and those that do not have the same lineage to the root term as the computational predictions. Our goal is to develop an automated method to identify and prioritize genes whose annotations may need to be re-examined based on the number of literature-based annotations that need review as well as the distance between the GO terms used by the literature-based and computational annotations, the number of computational predictions that suggest a gene should be reviewed, and the date the gene was last reviewed. This strategy will be applicable to GO annotations for any organism for which both literature-based and computational annotation is performed.

Phenotypes

Complete descriptions of mutant phenotypes are also represented using a controlled vocabulary known as the Ascomycete Phenotype Ontology (APO; available from the OBO Foundry, http://www.obofoundry.org). Each phenotype annotation in SGD is represented as an observable, which describes the entity or process, paired with a qualifier that describes the change in that entity or process. For example, if a mutant strain has increased sensitivity to a chemical compound, relative to wild-type, in which the observable is captured as 'resistance to chemicals' and the qualifier is 'decreased'. Chemical compounds, which are often used to elicit phenotypes, are specified using a

controlled vocabulary called the ChEBI ontology (11). Phenotypes representing over 17 000 classical and high-throughput phenotype assays have been described and are available from SGD (12).

Biochemical pathways

Biochemical pathways are manually curated by SGD and provided using the Pathway Tools browser version 15.0 (13). The SGD biochemical pathways data set for *S. cerevisiae*, one of the most highly curated data sets among all Pathway Tools data sets available, is the gold standard for budding yeast; SGD supports an ongoing effort to update and enhance these data. The Pathway Tools interface provides a complete description of each pathway, with molecular structures, E.C. numbers and full reference listing. The updated pathways browser provides several enhanced features, including download of a list of genes found in a pathway for further analysis with other tools available at SGD. The pathway browser is hyperlinked via the 'Pathways' section of the Locus Summary page. The Pathway display is available from http://pathway.yeastgenome.org.

Nomenclature

SGD continues to maintain the *S. cerevisiae* genomic nomenclature. Our job is to promote the community-defined nomenclature standards and to ensure that the agreed-upon guidelines are followed in naming new genes or assigning new names to previously identified genes. Community guidelines state that the first published name for a gene becomes the standard name. However, prior to publication, a gene name may be registered and displayed in SGD in order to notify the community of its intended use. If there are disagreements or naming conflicts, we communicate with the relevant researchers within the community and negotiate an agreement whenever possible. The majority of those working on the gene in question must agree to any nomenclature change before it is implemented in SGD. In addition to maintaining genetic names, SGD ensures that the names of ORFs, ARS elements, tRNAs and other chromosomal features also conform to agreed-upon formats. Over the past 2 years 154 new gene names have been assigned and 21 community-initiated name changes have been processed.

Maintenance of the S288C reference genome

SGD maintains, updates and distributes the *S. cerevisiae* reference genome sequence of the commonly used lab strain S288C. The original reference sequence, released in April 1996 (14), was the first available eukaryotic genome sequence and has been maintained by SGD ever since (*S. cerevisiae* S288C reference genome sequence NCBI accession numbers: NC_001133, NC_001134, NC_001135, NC_001136, NC_001137, NC_001138, NC_001139, NC_001140, NC_001141, NC_001142, NC_001143, NC_001144, NC_001145, NC_001146, NC_001147 and NC_001148). Between its original release and 2010, there have been a total of 239 changes to the sixteen nuclear chromosomes. In 2010, a newly determined reference sequence was released, from strain AB972, a direct descendant of S288C and one of the strains used in the original 1996 published sequence (14). The new reference sequence (version 64) was determined by mapping next-generation sequence reads to the previous reference (version 63) then manually inspecting any disagreements between the new sequence and the old reference. The AB972 sequence was released in February 2011 as the new reference genome for *S. cerevisiae* strain S288C, providing even higher quality sequence from a single consistent strain background for all 16 nuclear chromosomes. Concurrent with the update of the reference sequence we have adopted a reference versioning system; the current version of the reference sequence with chromosomal feature annotation was released as version 64.1. The first part of the version number increments when the sequence changes and second part increments when there are changes to the gene models. The current and all previous versions can be downloaded from SGD (http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases). With this comprehensive update in place, we plan that future differences between the reference and sequenced genes will primarily be represented as variations from the reference. However, a reference genome must take into consideration the best case for that strain and not simply the sequence of one individual; therefore, we will accommodate corrections to non-functional alleles unique to S288C. We have confidence that in the near future, the genomes of additional isolates of AB972 and other direct derivatives of S288C will be determined. Differences between these genomes will also be provided as alleles or observed variation of the current reference sequence.

Continued annotation and analysis of the *S. cerevisiae* genome

Our understanding of the *S. cerevisiae* genome is no longer limited to a few commonly studied strains such as S288C (genomics) (15), W303 (cell cycle and budding) (16) and SK1 (sporulation and meiosis) (17). Nextgen sequencing methods have already provided the genomic sequences of tens of *S. cerevisiae* laboratory and industrial strains (http://downloads.yeastgenome.org/sequence/strains), and will soon provide the genomic sequences for thousands of wild isolates. As with other model organisms, comparative genomics of many isolates will provide a new understanding of the full genetic constituent parts of a species. We are compiling *S. cerevisiae* genomes available from NCBI, selecting genomes based on scientific or industrial relevance and sequence quality. We currently provide 26 strain genomes available for download and include them in our BLAST and pattern matching (PatMatch) search tools. In addition, homologs to protein-coding genes present in the reference strain have been automatically predicted for 18 genomes. The predicted ORF sequence as well as a ClustalW alignment of all predicted ORFs in all strains is available from the Locus Summary page for each protein-coding ORF.

We are developing a pan-genomic representation of *S. cerevisiae* that includes all protein encoding genes, and their variation, found within any strain of

S. cerevisiae. SGD has long provided so called 'not in S288C' genes beginning with those that were genetically defined. We are now expanding this to include genes that have been observed within the genome of any *S. cerevisiae* stain. Genes can be lost as isolated populations adapt to their environment, and the pan-genomic representation of *S. cerevisiae* provides ready access to these genes, in context. For example, the strain S288C, source of the reference sequence, is known to be missing several well studied genes involved in sugar utilization such as *SUC1* (encoding one of several invertases, sucrose hydrolyzing enzymes), *MEL1* (encoding an alpha-galactosidase required for catabolic conversion of melibiose to glucose) and *BIO1* (encoding a pimeloyl-CoA synthetase) [SUC family (18); MEL family (19); BIO family (20)]. In addition to the reference genome there are currently twelve other genomes with annotations at GenBank and UniProt. The growth environment of commercial yeast strains, grape juice, is harsher than that faced by laboratory strains. The difference in growth environment is likely represented by the different repertoire of gene functions. The commercial wine yeast EC1118 (21) lacks 111 genes present in S288C and contains 34 that are not found in S288C. These 34 genes of EC1118 not present in S288C include predicted hexose transporters, drug transporters, a putative zinc-finger transcription factor, a methyltransferase, an oxidoreductase and nucleotide transporters. The genomic sequence from another wine strain, AMRI1631 (22), is reported to contain several other genes not found in S288C. The addition of these non-S288C genes to SGD will expand the catalog of functions that can be explored by yeast genetics.

As we add additional 'not in S288C' genes, we will adopt a standard gene categorization for all genes in the pan-genome. Genes will be classified into one of three types: 'core' genes that are found in all strains, 'frequent' genes indicating the gene is only found in some strains, and 'unique' genes for those that are only observed in one stain. Because many of the genomes have not been finished, their GenBank records may contain partial coding regions, representing gene fragments or other ambiguous annotations. Thus many of these genomes will require extensive reannotation. SGD will not add a gene to the pan-genome unless it meets very stringent annotation standards. Because of these issues we cannot appropriately estimate the number of core and frequently occurring genes within the species. However, based on the lack of sequence similarity between genes in an annotated strain and S288C, we roughly estimate that 1300 genes could be considered in the 'unique' class of genes. Therefore, the creation of the pan-genome will not be a trivial task but it is reasonable that we will be successful.

Incorporation of high-throughput data sets

Expression analysis at SGD has a new powerful interface and many new data sets. The new interface uses a tool called SPELL [Serial Pattern of Expression Levels Locator, http://spell.yeastgenome.org (23)], which facilitates the rapid identification of the most informative data sets and co-expressed genes based on patterns of expression shared with the query gene(s) from a comprehensive collection of almost 400 data sets. The expression analysis tool can be accessed from the Locus Summary via the Expression tab and the Expression Summary histogram. Implementation of a new data pipeline and the SPELL interface was created through collaboration with the SGD Colony at Princeton University.

Budding yeast has long been used as the initial system for developing new high-throughput methodologies; as a result, we have a wealth of data types to integrate with specific gene products. SGD uses the GBrowse genome browser (24) to display diverse types of genomic information, with our most recent efforts focusing on chromatin structure, including histone modifications, histone variants and nucleosome organization; chromosomal maintenance sites for meiotic recombination and origins of replication; transcription regulation, including RNA Polymerase II occupancy and transcription factor binding from ChIP-chip and ChIP-seq assays; and the transcriptome, including RNA-seq information and 3′-UTR analyses. SGD's genome browser is provided from http://browse.yeastgenome.org. New high-throughput results are frequently added to the browser, and all data sets are available for download in standard formats (currently 60 published data sets) from the SGD file download site, http://downloads.yeastgenome.org (Chan,E.T. and Cherry,J.M. Considerations for creating and annotating the budding yeast Genome Map at SGD: A progress report. *submitted to Database for Biocuration 2012 virtual issue*).

NEW QUERY AND DISPLAY TOOLS: ACCESS TO INTEGRATED BUDDING YEAST INFORMATION

All literature-based manually curated annotations, sequence and high-throughput data sets are accessible to users via multiple access points and methods. Data are integrated for display on SGD's locus-specific pages and can be analyzed using web-based tools, or downloaded for custom use.

Downloading gene-specific and data set-specific annotations

Enhancements to SGD's locus specific pages now allow download of detailed information. For example, the phenotype data is available by selecting the 'Download Data' button available from the Phenotype tab of the Locus pages. A tab-delimited file of all interaction data for a gene can be obtained with the 'Download Unfiltered Data' or via the 'Download options' link, which passes the list of genes to YeastMine, where the researchers can explore SGD's rich collection of associated data, including genetic and physical interactions.

Advanced query

YeastMine (http://yeastmine.yeastgenome.org; Balakrishnan, R., Park,J., Karra,K., Hitz,B.C., Binkley,G., Hong,E.L., Sullivan,J., Micklem,G. and Cherry,J.M. YeastMine – An integrated data warehouse of *S. cerevisiae* data as a

multi-purpose tool-kit at SGD. *submitted to Database for Biocuration 2012 virtual issue*) is a multifaceted search and retrieval tool that provides access to diverse types of curated and systematic data. Searches can be initiated with a list of genes, a list of GO terms, or results from multiple searches of a variety of data types. At each step, the results can be combined for further analysis. Queries can be customized by modifying predefined templates, or by creating completely new searches via the QueryBuilder. Customized queries facilitate access and retrieval of specific types of data, without requiring additional programming support at SGD. The search and its results can be saved or downloaded in customizable file formats. YeastMine can currently query the following data types: GO annotations, chromosomal features, co-ordinates, phenotype annotations, interaction data, orthologs, gene expression and curated literature. We are automating the update of YeastMine from the central SGD databases on a weekly basis. YeastMine will continue to expand with more data types, such as high-throughput chromosome data sets, biochemical pathways and enhanced data presentation. YeastMine was developed using the Intermine environment (25) and is being employed by other model organism databases and the modENCODE project (26) for *Drosophila* and *Caenorhabditis* genomic data. YeastMine has many other features such as Web Services and customized widgets that allow enhanced data display and retrieval. Through the integration of new high-throughput data and nextgen sequencing results, implementation of data models and advanced search via YeastMine, and enhancement of data presentation with Gbrowse, SGD has become the modENCODE-like hub for yeast related data (Chan,E.T. and Cherry,J.M. Considerations for creating and annotating the budding yeast Genome Map at SGD: A progress report. *submitted to Database for Biocuration 2012 virtual issue*).

Literature search tools

The Textpresso [http://textpresso.yeastgenome.org (27)] search tool is available to search approximately 50 000 papers collected by the SGD project. SGD Textpresso provides access to the full-text and abstracts of papers that have been retrieved for potential curation. Although not all such papers were found to include sufficient yeast-focused information to be included within SGD, these papers may still be generally useful and are thus included in the Textpresso catalog. While full-text search is available from many other sources, Textpresso provides the unique ability to quickly and effectively explore a literature collection using very specific, complex queries, such as finding all sentences containing a description of two interacting genes.

KEEPING CURRENT WITH CHANGES AT SGD: ANNOUNCEMENTS AND SOCIAL MEDIA

News announcements

SGD distributes a quarterly newsletter via our web site and direct email. The newsletter focuses on new data and tool development at SGD. We also provide significant community news such as future scientific meetings and new laboratory resources. Blog-style news announcements will be available on the SGD home page providing news noteworthy for the fungal genetics researcher. We are updating our software to allow reporting of news about yeast genomics, notable awards to community members, publication of highly significant results and new methods. All posts are also distributed via RSS feed.

Social media

SGD is now on Facebook and Twitter. Users can 'Like' us on Facebook or follow @yeastgenome on Twitter to receive updates about new features, tips on using SGD, real-time meeting updates and other interesting tidbits about yeast.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Kara Dolinski, Peter Koppstein and Lance Parsons, Princeton University, for help with data preparation and implementation of SPELL at SGD and Kara Dolinski and David Botstein, Princeton University, for encouragement and thoughtful advice. To the Troyanskaya laboratory, Princeton University, for creating the expression data visualization tool, SPELL. Gos Micklem, Julie Sullivan and Richard Smith, University of Cambridge UK, for development and maintenance of complex data search and display environment, InterMine. Most importantly, the SGD project would not be possible without the continued communication and support of the worldwide community of yeast researchers who provide suggestions for improvement and notice of issues in the data and tools the authors provide. The authors also wish to acknowledge the interactions and discussions with all the members of the Gene Ontology Consortium. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health.

FUNDING

National Human Genome Research Institute of the US National Institutes of Health: Saccharomyces Genome Database project (grant number P41 HG001315); YeastMine development [PI: Gos Micklem, grant number R01 HG004834); Gene Ontology Consortium (PI: Judith Blake, grant number P41 HG002273). Funding for open access charge: NHGRI (grant number P41 HG001315).

Conflict of interest statement. None declared.

REFERENCES

1. Petranovic,D., Tyo,K., Vemuri,G.N. and Nielsen,J. (2010) Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism. *FEMS Yeast Res.*, **10**, 1046–1059.

2. Goeckeler,J.L. and Brodsky,J.L. (2010) Molecular chaperones and substrate ubiquitination control the efficiency of endoplasmic reticulum-associated degradation. *Diabetes Obes. Metab.*, **12(Suppl.)**, 32–38.
3. Rinaldi,T., Dallabona,C., Ferrero,I., Frontali,L. and Bolotin-Fukuhara,M. (2010) Mitochondrial diseases and the role of the yeast models. *FEMS Yeast Res.*, **10**, 1006–1022.
4. Bharadwaj,P., Martins,R. and Macreadie,I. (2010) Yeast as a model for studying Alzheimer's disease. *FEMS Yeast Res.*, **10**, 961–969.
5. Barros,M.H., da Cunha,F.M., Oliveira,G.A., Tahara,E.B. and Kowaltowski,A.J. (2010) Yeast as a model to study mitochondrial mechanisms in ageing. *Mech. Ageing Dev.*, **131**, 494–502.
6. Costanzo,M.C., Park,J., Balakrishnan,R., Cherry,J.M. and Hong,E.L. (2011) Using computational predictions to improve literature-based Gene Ontology annotations: a feasibility study. *Database*, **2011**, bar004.
7. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009 – an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
8. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.
9. Gaudet,P., Livstone,M.S., Lewis,S.E. and Thomas,P.D. (2011) Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief. Bioinform.*, doi: 10.1093/bio/bbr042.
10. Reference Genome Group of the Gene Ontology Consortium. (2001) The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput. Biol.*, **5**, e1000431.
11. de Matos,P., Alcantara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2009) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
12. Costanzo,M.C., Skrzypek,M.S., Nash,R., Wong,E., Binkley,G., Engel,S.R., Hitz,B., Hong,E.L. and Cherry,J.M. (2009) New mutant phenotype data curation system in the *Saccharomyces* Genome Database. *Database*, doi: 10.1093/database/bap001.
13. Karp,P.D., Paley,S.M., Krummenacker,M., Latendresse,M., Dale,J.M., Lee,T.J., Kaipa,P., Gilham,F., Spaulding,A., Popescu,L. *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **11**, 40–79.
14. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.
15. Mortimer,R.K. and Johnston,J.R. (1986) Genealogy of principal strains of the yeast genetic stock center. *Genetics*, **113**, 35–43.
16. Fan,H.Y., Cheng,K.K. and Klein,H.L. (1996) Mutations in the RNA polymerase II transcription machinery suppress the hyper-recombination mutant hpr1 delta of *Saccharomyces cerevisiae*. *Genetics*, **142**, 749–759.
17. Kane,S.M. and Roth,R. (1974) Carbohydrate metabolism during ascospore development in yeast. *J. Bacteriol.*, **118**, 8–14.
18. Carlson,M. and Botstein,D. (1983) Organization of the SUC gene family in *Saccharomyces*. *Mol. Cell Biol.*, **3**, 351–359.
19. Naumov,G., Turakainen,H., Naumova,E., Aho,S. and Korhola,M. (1990) A new family of polymorphic genes in *Saccharomyces cerevisiae*: alpha-galactosidase genes MEL1-MEL7. *Mol. Gen. Genet.*, **224**, 119–128.
20. Hall,C. and Dietrich,F.S. (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, **177**, 2293–2307.
21. Novo,M., Bigey,F., Beyne,E., Galeote,V., Gavory,F., Mallet,S., Cambon,B., Legras,J.L., Wincker,P., Casaregola,S. *et al.* (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. *Proc. Natl Acad. Sci. USA*, **106**, 16333–16338.
22. Borneman,A.R., Forgan,A.H., Pretorius,I.S. and Chambers,P.J. (2008) Comparative genome analysis of a Saccharomyces cerevisiae wine strain. *FEMS Yeast Res.*, **8**, 1185–1195.
23. Hibbs,M.A., Hess,D.C., Myers,C.L., Huttenhower,C., Li,K. and Troyanskaya,O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
24. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
25. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., Mclaren,P., North,P. *et al.* (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol.*, **8**, R129.
26. Celniker,S.E., Dillon,L.A.L., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
27. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.