Gene Ontology Development and Implementation at the *Saccharomyces* Genome Database

E. L. Hong, S. Weng, K. Dolinski, R. Balakrishnan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, L. Issel-Tarver, A. Sethuraman, C. L. Theesfeld, G. Binkley, C. Lane, M. Schroeder, S. Dong, R. Andrada, D. Botstein, and J. M. Cherry

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305.

The *Saccharomyces* Genome Database (SGD; http://genome-www.stanford.edu/Saccharomyces/) is a model organism database that provides information about the genes and proteins of the budding yeast *S. cerevisiae* and develops resources that allow the scientific community to analyze and retrieve this information. SGD uses the controlled vocabulary terms of the Gene Ontology (GO) to describe the molecular function, biological process, and cellular component of a gene product. The GO terms have a specific relationship to each other that is determined by a parent/child structure. SGD has developed resources that use this structure to identify shared parent terms for a group of gene products. The GO Term Finder (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder) and the GO Term Mapper (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermMapper) find common function, process, or cellular localization terms for a set of genes while the Feature Search (http://genome-www4.stanford.edu/cgi-bin/SGD/search/featureSearch) retrieves gene products that have a similar function, process, or localization.

Introduction

The *Saccharomyces* Genome Database (SGD; http://genome-www.stanford.edu/Saccharomyces/) is a public resource for the scientific community about the budding yeast *S. cerevisiae*. It provides access to the sequence as well as information about the genetics, molecular biology, and biochemistry of this model organism (Ball et al., 2000; Ball et al., 2001; Cherry et al., 1998; Chervitz et al., 1999; Dwight et al., 2002; Weng et al., 2003). The database is organized around individual *S. cerevisiae* genes (such as open reading frames (ORFs) and tRNAs) and genetic elements (such as centromeres and ARS elements). Each gene or genetic element is presented as a "locus page". The locus page is a central location to find a gene's name, retrieve and analyze its sequence, find literature associated with the gene, and learn about its biological role in the cell. As information of a gene shifts from identifying its sequence to understanding the biology of its gene product, it is curated on the locus page using the Gene Ontology (GO) (Ball et al., 2000).

GO provides three controlled vocabularies that describe the molecular function, biological process, and cellular component of a gene product (Ashburner et al., 2000; The GO Consortium, 2001). The molecular function ontology contains terms that represent the specific task performed by that individual gene product, such as a helicase activity or a kinase activity. The biological process ontology contains terms that describe the biological goals in which the molecular function of the gene product participates, such as DNA repair or purine metabolism. The cellular component ontology contains terms that define the sub-cellular location of

the gene product, such as the nucleus or an origin of replication complex. Since GO is a controlled vocabulary, the phrasing of terms may be different from that found in scientific literature. Therefore, to clarify their scope, all GO terms have definitions and some terms have commonly used phrases as synonyms.

The terms have a specific relationship to each other. The relationship is determined by the directed acyclic graph (DAG) structure in which each term can have multiple parents and zero or more children (Figure 1). This structure allows more complexity than a standard hierarchy. In this structure, a child term is considered a more specific term than its parent term(s). Because each child term inherits characteristics of each parent term, all relationships between the terms must be true.

The ontologies are developed and refined so that the terms and the relationships reflect that current state of biological knowledge. As of May 2003, the ontologies contained more than 10,000 terms. Current versions of the ontology are available through the GO FTP site (ftp://ftp.geneontology.org/pub/go/) or an anonymous CVS repository (http://www.geneontology.org/#cvs).

Members of the GO Consortium are responsible for the development of the ontologies as well as using the ontologies to annotate the gene products of sequenced model organisms. Current members of the GO Consortium include organizations that maintain model organism databases for *S. cerevisiae, D. melanogaster, M. musculus, A. thaliana, C. elegans, S. pombe, R. norvegicus*, and *D. discoideum* (Dwight et al., 2002; Glockner et al., 2002; Hill et al., 2001; Rhee et al., 2003; Wood and Bahler, 2002) and institutes that manage genomic and proteomic information such as the European Bioinformatics Institute (EBI), Swiss-Prot, InterPro, TrEMBL, the Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute, The Institute for Genomic Research (TIGR), and Gramene (Camon et al., 2003; Haft et al., 2003; Ware et al., 2002). The annotations between GO terms and gene products made by these groups are publicly available on the GO website (http://www.geneontology.org/#annotations).

The GO annotations by member groups follow the guidelines determined by the GO Consortium (The GO Consortium, 2001). The most specific GO term is used to annotate a gene product based on the available information about that gene product. Since the characteristics of the more general terms are inherited by the more specific term, all terms used to annotate a gene product are checked to make sure the relationships between terms accurately describe the biology of the gene product. If the relationships are not an accurate reflection of the function, process, or component of the gene product, the ontology is refined. Whenever possible, a GO term is used to annotate a gene product based on published experimental evidence. The annotation is then associated with a literature citation and an evidence code that describes the type of experiment used to make the GO annotation. For example, evidence codes exist to indicate that the GO annotation was assigned based on a mutant phenotype or on the basis of sequence similarity to a gene product in another organism.

Although GO annotations provide a literature-based synopsis of the biology of a single gene product, they also facilitate the identification of relationships between gene products in the same organism or in different organisms. In order to facilitate genome-wide studies of *S. cerevisiae*, SGD has developed resources that take advantage of the GO annotations and structure of the ontology.

Current Status of GO Annotations at SGD

The GO terms are annotated to gene products in SGD following the guidelines determined by the GO consortium. Since the ontologies and the available biological information are dynamic, GO annotations at SGD are regularly reviewed and updated. As of April 2003, all ORFs have a complete set of GO annotations. These annotations are freely available in tab-delimited files from two sources: the SGD FTP site (ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/) and the GO Consortium (http://www.geneontology.org/#annotations).

Display of GO Annotations

One of the advantages of GO is that it provides a quick summary of the biology of a gene product. At SGD, this information is displayed on the locus page (Figure 2). Consistent with SGD's emphasis on literature-based annotations, the link labeled "GO evidence and references" leads to a page that lists every GO annotation, the references used to assign the annotations, the evidence codes, and the dates the annotations were last updated (Figure 3).

The GO term name is a link to the GO term page at SGD (Figure 4). Each term page lists the GO term, its unique ID, any synonyms, and its definition. The GO term page lists all genes at SGD that have been annotated to that term, along with their associated references and evidence codes. Since the relationship of a term to its parent terms and child terms is important to understanding the scope of the term, the structure of the ontology is displayed at the top of the term page. The GO term being viewed is written in red. A brown box indicates there are no gene products at SGD that have been annotated to that GO term. A blue box indicates there are gene products at SGD that have been annotated to that GO term. The number in the blue box indicates the number of gene products directly annotated to that term; the number in parentheses indicates the number of genes that are annotated to children of that term. The GO term page of a different term may be viewed by clicking on the term name. The blue arrow is a link that redraws the ontology to display more parent terms of the GO term being viewed.

New Resources at SGD

In addition to providing a summary of the biology of a gene product, significant relationships between gene products can be identified by taking advantage of the DAG structure of the ontology. Since gene products are annotated to the most specific GO term possible and each GO term can have multiple parents, it may not be immediately evident whether there is a common parent term that is shared by a set of GO terms and what that common parent term might be. In addition, since GO terms can have multiple child terms, it is difficult to identify all gene products that might be annotated to terms that are children of a GO term. To simplify this task, SGD has developed three resources that identify significant relationships between gene products by tracing the lineage of the GO terms that have been annotated to them.

Two of the new SGD tools identify a common parent term that is shared by a group of gene products. The GO Term Finder (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermFinder) finds the most significant GO term that is a parent of all the GO terms to which the genes in an input list have been annotated. The list of genes can be entered manually or uploaded from a text file. To improve the performance of the search, the tool considers one ontology (function, process, or component) at a time (Figure 5A). The result of the search appears as a modified version of the GO tree view that is used on the GO term pages. The genes are listed below the GO terms that are directly annotated to the gene product. In addition, the boxes containing the GO term name are color-coded based on the significance of that term as a common GO term (Figure 5B). The warmer colors indicate that the GO term is a more significant common parent term. The significance of a GO term is determined by the frequency that term occurs as a parent to the GO terms used in annotating a groups of genes as compared to the number of times that term is associated with other genes in the genome. The table, located at the bottom of the results page, summarizes the statistical results of the search (Figure 5C). This information can be downloaded in a tab-delimited file for future reference or further analysis. The Stanford Microarray Database has created a generic version of the GO Term Finder uses GO annotations for any genome. It is available to download at CPAN (http://search.cpan.org/author/ SHERLOCK/).

The GO Term Mapper (http://genome-www4.stanford.edu/cgi-bin/SGD/GO/goTermMapper) is similar to the GO Term Finder in that it identifies a common GO term that is a parent of the GO terms to which the

genes in an input list have been annotated. The difference is that GO Term Mapper traces the lineage of a GO term to a GO-Slim term. GO-Slim is a selection of higher-level GO terms that represent major branches of an ontology. Given a list of genes, one GO-Slim ontology or a specific subset of GO-Slim terms is selected (Figure 6A). The results display the genes that are annotated directly or indirectly to the selected GO-Slim terms (Figure 6B). Because a GO term can have multiple parents and a gene product can have multiple GO annotations, a gene can appear to be a descendant of multiple GO-Slim terms.

The third new SGD tool simplifies the problem of identifying all gene products that have related functions, are involved in the same processes, or have similar subcellular localizations. The Feature Search (http://genome-www4.stanford.edu/cgi-bin/SGD/search/featureSearch) is an advanced search that searches for genes that have been annotated to children of a selected GO-Slim term. The Feature Search takes as input any chromosomal feature type and one or more GO-Slim terms (Figure 7A). It will return the chromosomal features that have been annotated directly or indirectly to all the selected GO-Slim terms (Figure 7B).

There is a need to identify shared characteristics within clusters of genes that are the generated by genome-wide analyses, such as microarray experiments. By taking advantage of the structure of GO, the new resources at SGD are powerful tools for addressing this need. Because the strength of the tools is a reflection of the quality of the GO annotations at SGD, SGD is committed to reviewing and updating the GO annotations as more biological information becomes available.

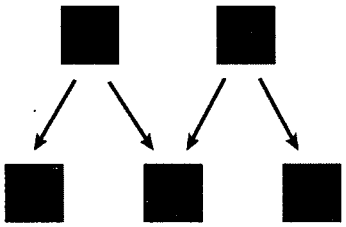

Fig. 1. An illustration of the directed acyclic graph (DAG) structure of the Gene Ontology (GO). The blue boxes represent terms in the ontology and the arrows point to child terms. Each term in the ontology can have multiple child terms. Each term can have more than one parent term.

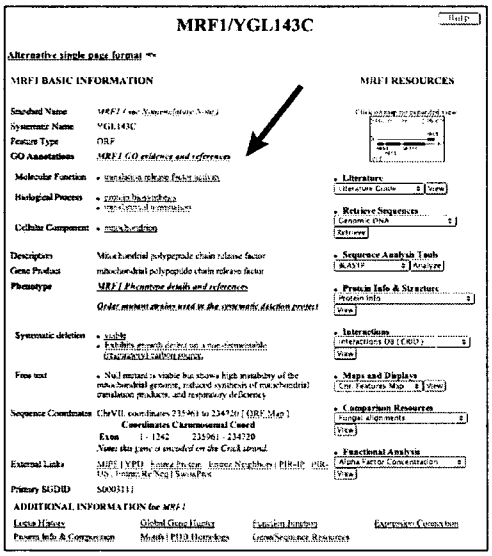

Fig. 2. The locus page of MRF1/YGL143C from SGD. The locus page offers basic information about a gene or chromosomal feature, such as its names, its GO annotations (pointed out by the arrow), its mutant phenotype, and its chromosomal coordinates. In addition, the locus page provides links to retrieve and analyze the sequence of MRF1 as well as to links to other biological information.

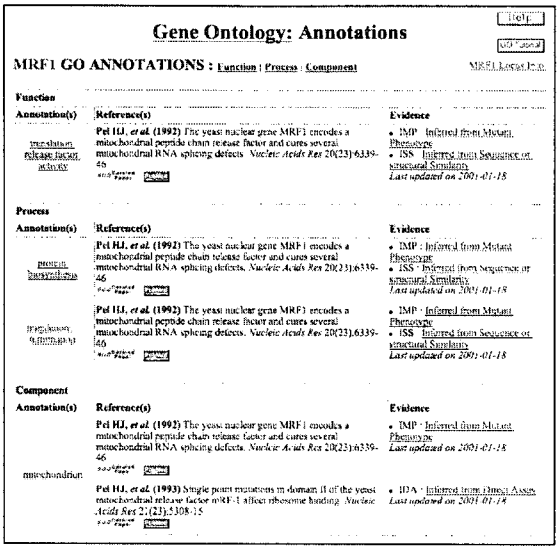

Fig. 3. The GO annotations page for MRF1/YGL143C. The molecular function, biological process, and cellular component GO annotations for MRF1 are listed along with the literature citation used to make the annotation and the evidence code. Help resources are located in the upper right corner of the page. The "HELP" button leads to a page of general information about GO and GO Tools at SGD. The "GO Tutorial" button leads to a step-by-step guide to viewing GO annotations at SGD.

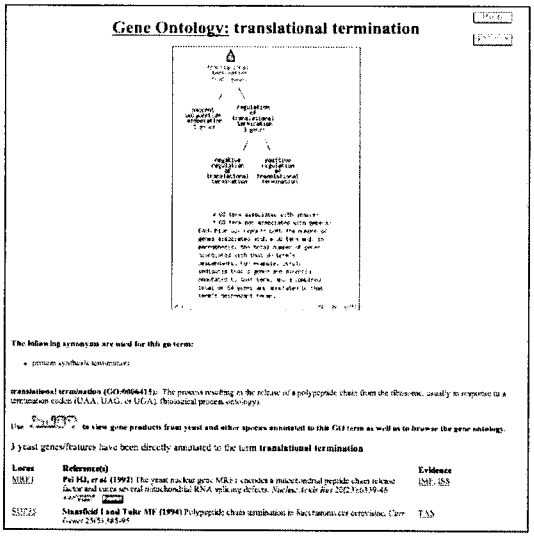

Fig. 4. The GO term page for GO term "translational termination". The top half of the GO term page displays the term name and a graphical representation of its context in the ontology. The blue arrow will redraw the ontology by displaying the parent terms of the ontology. The bottom half of the GO term page displays a term's synonyms, its unique numeric identifier, its definition, and the genes that have been annotated to that term.

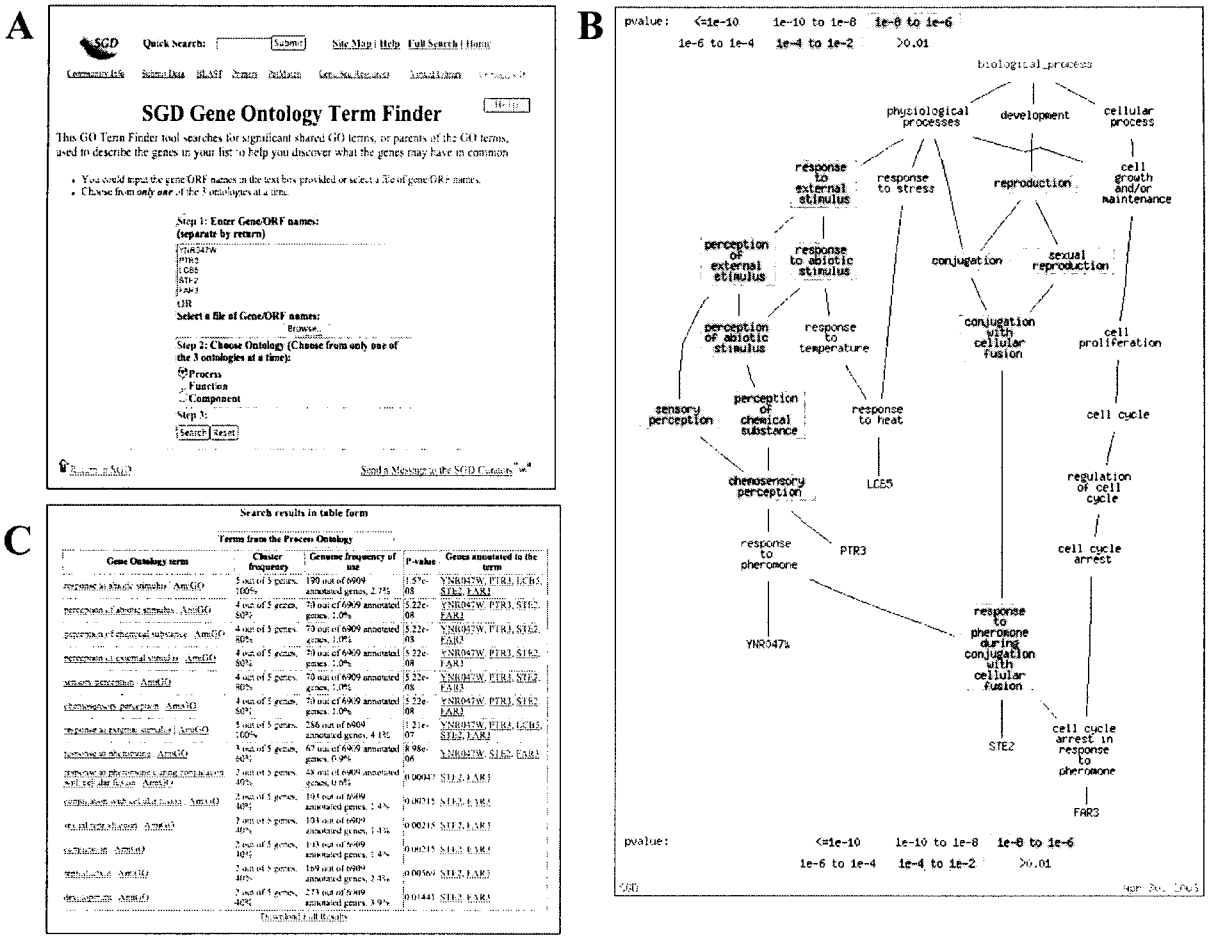

Fig. 5. The GO Term Finder finds the GO term that is shared by a group of genes. (A) A list of gene or ORF names can either be entered manually or uploaded from a text file. One ontology is searched at a time. (B) At the top of the results page, the ontology is drawn. The genes are listed below the GO term used to annotated it. The boxes are color-coded based on the significance of the term as a shared GO term. (C) At the bottom of the results page, a table summarizes the statistical results of the search.

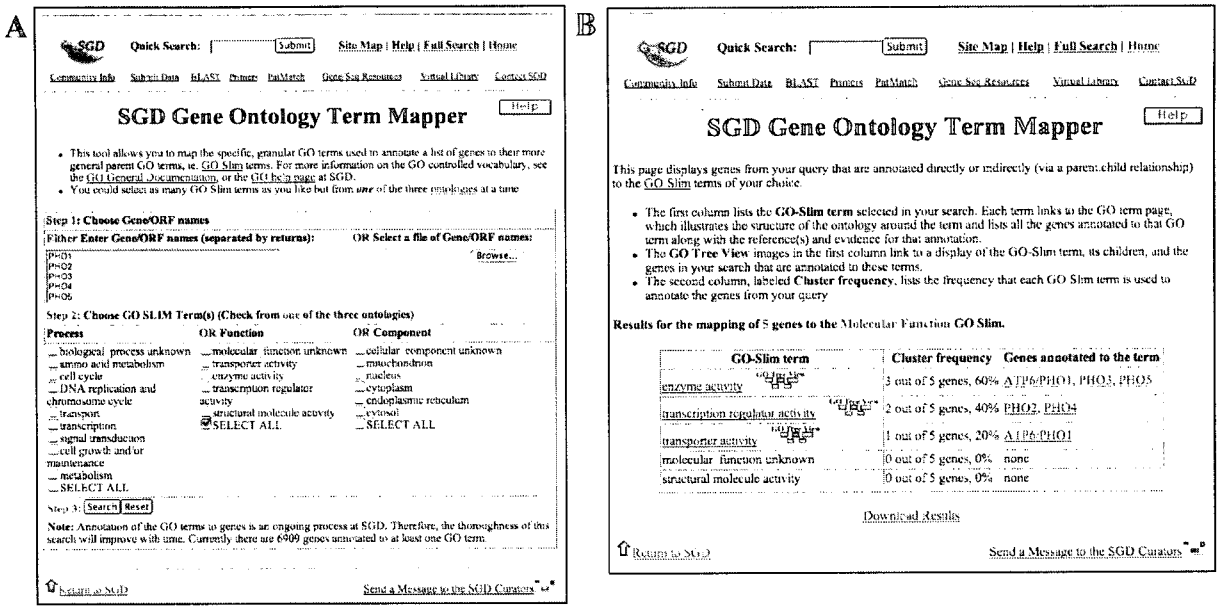

Fig. 6. The GO Term Mapper identifies a common GO-Slim term that is a parent of the GO terms to which the genes in an input list have been annotated. (A) A list of gene or ORF names can either be entered manually or uploaded from a text file. An ontology or a subset of GO-Slim is selected for the search. (B) The results page is a table which shows which genes have annotated to the selected GO-Slim terms.

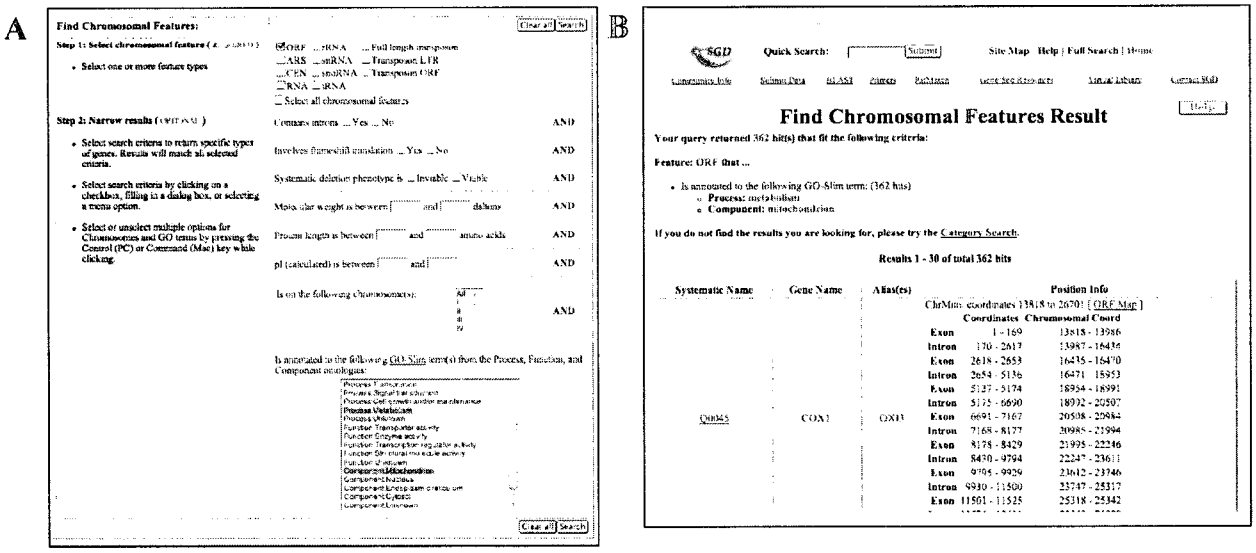

Fig. 7. The Feature Search is an advanced search that returns a list of genes based on their GO annotations. (A) A chromosomal feature type is selected as well as the GO-Slim terms of interest. (B) The search retrieves all genes that have been annotated to the children of the selected GO-Slim terms.

References

1. 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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the

unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-9.

2. Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Kasarskis, A., Scafe, C. R., Sherlock, G., Binkley, G., Jin, H., Kaloper, M., Orr, S. D., Schroeder, M., Weng, S., Zhu, Y., Botstein, D., and Cherry, J. M. (2000). Integrating functional genomic information into the Saccharomyces genome database. Nucleic Acids Res *28*, 77-80.
3. Ball, C. A., Jin, H., Sherlock, G., Weng, S., Matese, J. C., Andrada, R., Binkley, G., Dolinski, K., Dwight, S. S., Harris, M. A., Issel-Tarver, L., Schroeder, M., Botstein, D., and Cherry, J. M. (2001). Saccharomyces Genome Database provides tools to survey gene expression and functional analysis data. Nucleic Acids Res *29* 80-1.
4. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J. Cox, A., and Apweiler, R. (2003). The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res *13*, 662-72.
5. Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD: Saccharomyces Genome Database. Nucleic Acids Res *26*, 73-9.
6. Chervitz, S. A., Hester, E. T., Ball, C. A., Dolinski, K., Dwight, S. S., Harris, M. A., Juvik, G., Malekian, A., Roberts, S., Roe, T., Scafe, C., Schroeder, M., Sherlock, G., Weng, S., Zhu, Y., Cherry, J. M., and Botstein, D. (1999). Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. Nucleic Acids Res *27*, 74-8.
7. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L. Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). Nucleic Acids Res *30*, 69-72.
8. Glockner, G., Eichinger, L., Szafranski, K., Pachebat, J. A., Bankier, A. T., Dear, P. H., Lehmann, R., Baumgart, C., Parra, G., Abril, J. F., Guigo, R., Kumpf, K., Tunggal, B., Cox, E., Quail, M. A., Platzer, M., Rosenthal, A., and Noegel, A. A. (2002). Sequence and analysis of chromosome 2 of Dictyostelium discoideum. Nature *418*, 79-85.
9. The GO Consortium. (2001). Creating the gene ontology resource: design and implementation. Genome Res *11*, 1425-33.
10. Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. Nucleic Acids Res *31*, 371-3.
11. Hill, D. P., Davis, A. P., Richardson, J. E., Corradi, J. P., Ringwald, M., Eppig, J. T., and Blake, J. A. (2001). Program description: Strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. Genomics *74*, 121-8.
12. Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res *31*, 224-8.
13. Ware, D. H., Jaiswal, P., Ni, J., Yap, I. V., Pan, X., Clark, K. Y., Teytelman, L., Schmidt, S. C., Zhao, W., Chang, K., Cartinhour, S., Stein, L. D., and McCouch, S. R. (2002). Gramene, a tool for grass genomics. Plant Physiol *130*, 1606-13.
14. Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S. S., Engel, S., Fisk, D. G., Hong, E., Issel-Tarver, L., Sethuraman, A., Theesfeld, C., Andrada, R., Binkley, G., Lane, C., Schroeder, M., Botstein, D., and Michael Cherry, J. (2003). Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. Nucleic Acids Res *31*, 216-8.
15. Wood, V., and Bahler, J. (2002). How to get the best from fission yeast genome data. Comparative and Functional Genomics *3*, 282-8.