 Nucleic Acids Research, 2000, Vol. 28, No. 1 77–80

Integrating functional genomic information into the *Saccharomyces* Genome Database

Catherine A. Ball, Kara Dolinski, Selina S. Dwight, Midori A. Harris, Laurie Issel-Tarver, Andrew Kasarskis, Charles R. Scafe, Gavin Sherlock, Gail Binkley, Heng Jin, Mira Kaloper, Sidney D. Orr, Mark Schroeder, Shuai Weng, Yan Zhu, David Botstein and J. Michael Cherry*

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

Received October 1, 1999; Revised and Accepted October 7, 1999

ABSTRACT

The *Saccharomyces* Genome Database (SGD) stores and organizes information about the nearly 6200 genes in the yeast genome. The information is organized around the 'locus page' and directs users to the detailed information they seek. SGD is endeavoring to integrate the existing information about yeast genes with the large volume of data generated by functional analyses that are beginning to appear in the literature and on web sites. New features will include searches of systematic analyses and Gene Summary Paragraphs that succinctly review the literature for each gene. In addition to current information, such as gene product and phenotype descriptions, the new locus page will also describe a gene product's cellular process, function and localization using a controlled vocabulary developed in collaboration with two other model organism databases. We describe these developments in SGD through the newly reorganized locus page. The SGD is accessible via the WWW at http://genome-www.stanford. edu/Saccharomyces/

GENERAL FORMAT OF SGD's NEW LOCUS PAGE

The diverse information in the *Saccharomyces* Genome Database (SGD) is organized around individual genes. The 'locus page' that organizes the information about each gene is critically important to database users. Since the completion of the yeast genomic sequence (1), genome-scale experiments have become increasingly routine. To integrate the large datasets from these experiments with existing information about the 6200 yeast genes, SGD will introduce a redesigned locus page that presents all of the information in a format intuitive to biologists. The locus page will continue to serve as a central information source, from which users will be able to retrieve a large set of data about any gene with minimal navigation. The data on the locus page will also be accessible in tab-delimited and XML formats to facilitate automated data exchange. We present here an overview of the new information, concentrating on two new features, the Gene Summary Paragraph and the function, process and cellular component descriptions, beginning with the way in which these will be organized on the locus page.

To illustrate how SGD plans to incorporate some of the new data that has resulted from the completion of the yeast genomic sequence (1), a prototype of the new locus page for *MET16* is illustrated in Figure 1. The left side of Figure 1 lists the information most users want to know about a particular gene, while the blue box on the right side contains links to tools and resources similar to those currently available from SGD's Gene/Sequence Resources page (http://genome-www2.stanford. edu/cgi-bin/SGD/seqTools). Links to additional information and resources are displayed as orange buttons at the bottom of Figure 1.

BASIC INFORMATION

Information about a gene's standard name and aliases along with gene product and phenotype descriptions will continue to be displayed on SGD's new locus page. A controlled vocabulary to describe mutant phenotypes is being developed to facilitate quick and accurate searches for genes with similar phenotypes.

Three new descriptions will be added to the display on the locus page: function, process and cellular component. These descriptions will come from a controlled vocabulary created by a cross-species project to describe the biological roles of individual gene products. In an on-going collaboration, FlyBase (2), the Mouse Genome Database (MGD) (3) and SGD (4) are developing a Gene Ontology (Ashburner *et al*., manuscript in preparation and http://genome-www.stanford.edu/GO), a common vocabulary with defined relationships between controlled vocabulary terms that describes a gene product's biological objective, function and localization. The first category in the Gene Ontology is process, which describes the biological role or general cellular objective of the gene product. Examples of biological processes are 'karyogamy' and 'amino acid biosynthesis'. The second category is function, which describes the elemental activity or task performed by a gene product. 'DNA binding', 'ATPase' and 'microtubule motor' are examples of gene function. The third category, cellular component, describes subcellular structures, locations, and macromolecular complexes in which a gene product may be found.

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Fax: +1 650 723 7016; Email: cherry@genome.stanford.edu

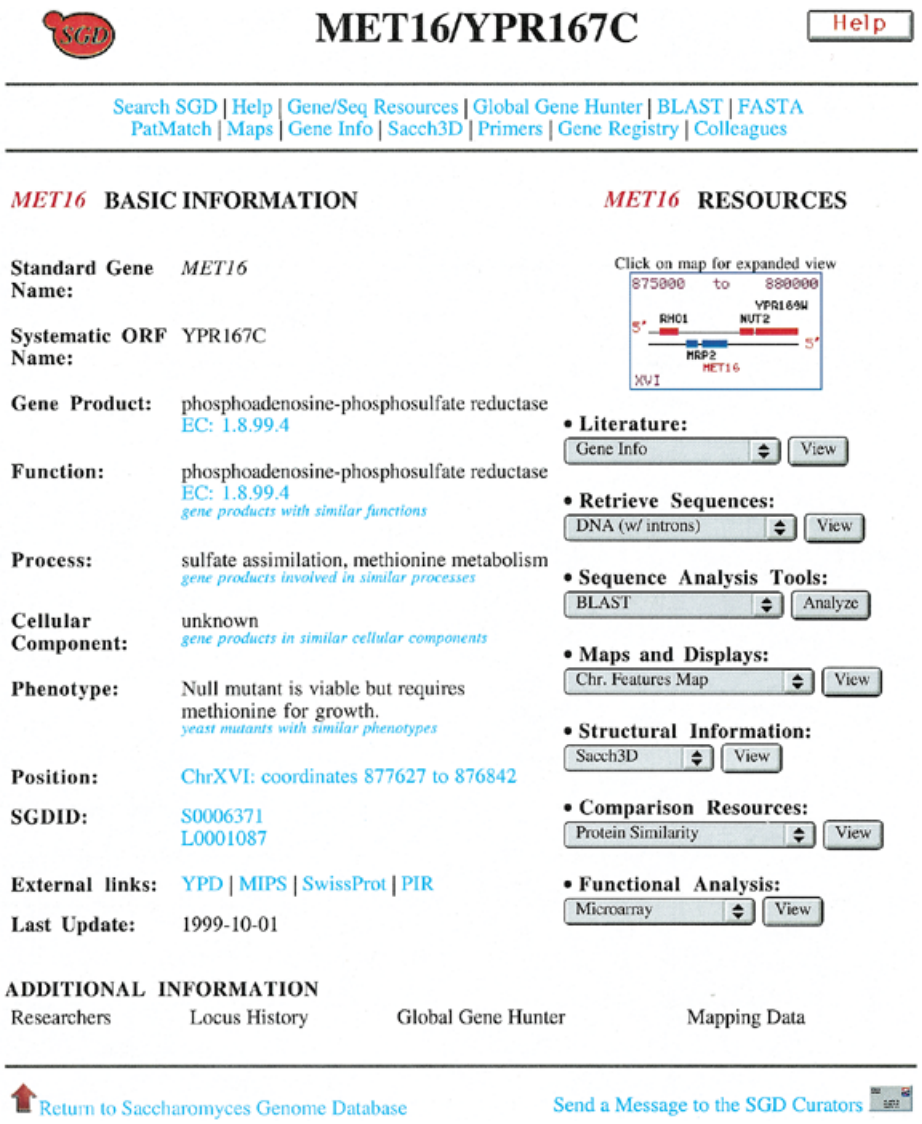

Figure 1. The new locus page at SGD. This new locus page for *MET16* shows how data will be organized to include several different types of information at a central location.

For instance, 'mitochondrial outer membrane' and 'spliceosome' are cellular components. Even at the level of a single gene, distinguishing between functions and processes enhances its annotation; for example, we can state that a gene encodes a protein kinase (a function) and is involved in cell cycle progression and cellular morphogenesis (two biological processes). This controlled vocabulary will provide a method to identify genes with similar processes, functions or localizations between species.

The usefulness of the distinction between process and function becomes still more apparent when attempting to interpret the results of large-scale experiments. As an example, Figure 2 shows a cluster of co-expressed yeast genes (originally published in ref. 5) with process and function annotated for each gene in the cluster. It is immediately obvious that genes whose products participate in a common process (in Fig. 2, methionine metabolism) tend to be co-expressed under these conditions, while function correlates much less with gene expression. There are other analyses, such as sequence comparisons, for which the function description will better illuminate relationships between genes or gene products.

GENE RESOURCES AND TOOLS

Shown in the blue box on the right side of Figure 1 are tools and resources to analyze a particular gene. Tools in the new locus page will be organized by topic and selected through pull-down menus. Two links to literature resources will be available: the Gene Info Literature Guide, a resource at SGD that organizes the literature by topic, and a hyperlink to PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) for papers that mention yeast *MET16*. All of our current DNA and protein sequence retrieval options will be available, including six frame translations with restriction maps and the non-systematic sequences available from GenBank (6). In addition, sequence analysis tools (4), maps (7) and comparison resources (8)

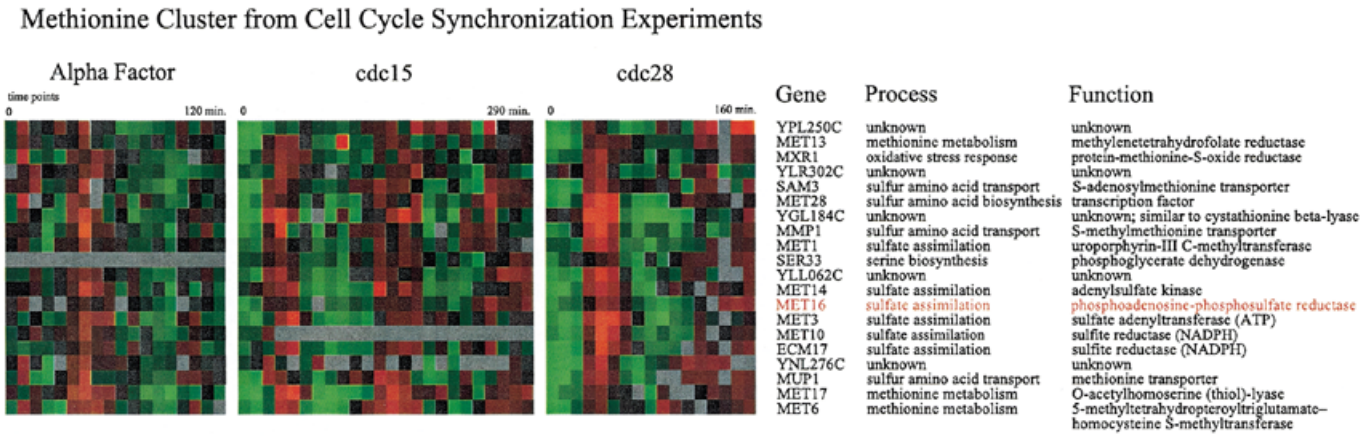

Figure 2. Process and function annotation of genes in a DNA microarray cluster. This cluster was originally published by Spellman *et al.* (5) and contains genes that are co-expressed in a cell cycle-dependent manner. The entry for *MET16* is highlighted in red. The gene name, process and function are shown for each gene in the cluster (cellular component annotation has been omitted for brevity). Genes correspond to the rows, and the time points of each experiment are the columns. The colors indicate changes in gene expression in synchronized cells relative to a reference sample. Red represents increased expression, green represents decreased expression and black denotes no change. Gray areas indicate missing data points. The clustering algorithm is described by Eisen *et al.* (9).

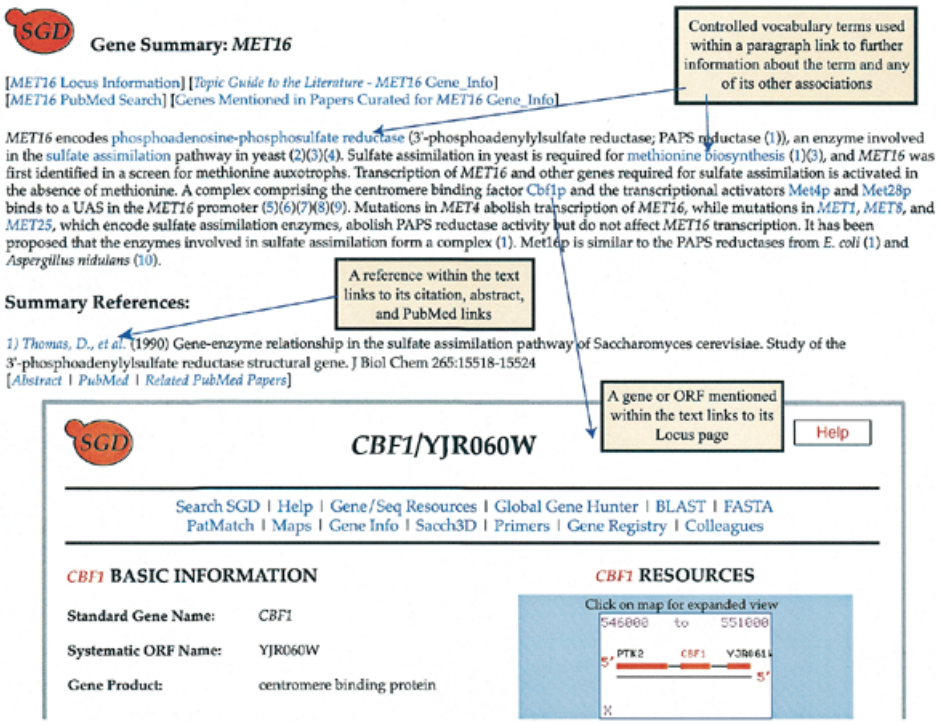

Figure 3. The Gene Summary Paragraph. The Gene Summary Paragraph, shown for *MET16*, summarizes important findings about a particular gene. From the paragraph page, there are links to the locus pages of other relevant genes (in this case, *CBF1*), as well as links to additional sources of information such as the Gene Ontology controlled vocabulary terms, references and PubMed links, and the SGD Literature Guide.

will still be provided. A valuable new resource will be the search of functional analysis data, such as DNA microarray analysis (5,9,10) and deletion studies (11), that will display the large-scale systematic results available for a particular gene.

SGD will continue to link to external databases such as YPD (12), MIPS (13), SWISS-PROT (14) and PIR (15,16) so that users can easily find complementary information from other sources. Additional information, such as mapping data and contact information for researchers who work on the selected gene, will be just a click away on the bottom of the page (Fig. 1).

GENE SUMMARY PARAGRAPH

Another new feature that will be incorporated at SGD is the Gene Summary Paragraph, which will be a concise summary of the major aspects of the gene's biology published in the scientific literature. An example illustrating the Gene Summary Paragraph for *MET16* is shown in Figure 3. In addition to being a resource for yeast biologists unfamiliar with a particular gene, these summaries will serve as an introduction to the gene and its product for researchers who may be entering SGD or the

area of yeast biology for the first time. As additional genomic sequencing projects are completed and new homologs in other species are identified, it will become increasingly important to convey such information to researchers who have limited expertise in yeast genetics and molecular biology. Written by PhD level biologists at SGD, Gene Summary Paragraphs will be extensively referenced. The list of references used to write the paragraph will be provided at the bottom of the page, along with links to the reference's abstract, PubMed record, and related PubMed papers. An exhaustive list of references for a given gene can be found in the Gene Info Literature Guide described above. Because SGD enjoys a high level of interaction with its users, we both invite and expect suggestions from the yeast research community to update and improve these gene summaries.

CONCLUSION

As biological research enters the genomic era, the types and amount of information available can be overwhelming. In anticipation of increasing data from large-scale functional analysis projects and the detection of new sequence homologs, SGD is consolidating and improving the presentation of gene-specific information. Specifically, SGD has entered a collaboration with FlyBase and the MGD to create the Gene Ontology which will describe the important components and relationships that reflect our current understanding of cell biology. It is expected that this cross-species project will allow database users to easily identify gene products with similar biological roles, protein functions or cellular localizations, within or between species, as well as improve the annotations within each of the collaborating databases. Additionally, SGD curators are creating short, concise Gene Summary Paragraphs that describe the published highlights of each gene's biology. The gene summaries will provide researchers with a convenient introduction to an unfamiliar topic or model organism.

ACKNOWLEDGEMENTS

S.G.D. is supported by a P41, National Resources, grant from the National Human Genome Research Institute at the US National Institutes of Health. C.R.S. was supported by Human Genome Training Grant post-doctoral fellowship #HG-00044.

REFERENCES

1. Goffeau,A. *et al.* (1997) *Nature*, **387**, 5.
2. FlyBase Consortium (1999) *Nucleic Acids Res.*, **27**, 85–88.
3. Blake,J.A., Richardson,J.E., Davisson,M.T. and Eppig,J.T. (1999) *Nucleic Acids Res.*, **27**, 95–98. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 108–111.
4. 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) *Nucleic Acids Res.*, **26**, 73–79.
5. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) *Mol. Biol. Cell*, **9**, 3273–3297.
6. Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J., Ouellette,B.F., Rapp,B.A. and Wheeler,D.L. (1999) *Nucleic Acids Res.*, **27**, 12–17. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 15–18.
7. Cherry,J.M., Ball,C., Weng,S., Juvik,G., Schmidt,R., Adler,C., Dunn,B., Dwight,S., Riles,L., Mortimer,R.K. and Botstein,D. (1997) *Nature*, **387**, 67–73.
8. 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) *Nucleic Acids Res.*, **27**, 74–78.
9. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
10. Ferea,T.L., Botstein,D., Brown,P.O. and Rosenzweig,R.F. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 9721–9726.
11. Winzeler,E.A., Shoemaker,D.D., Astromoff,A., Liang,H., Anderson,K., Andre,B., Bangham,R., Benito,R., Boeke,J.D., Bussey,H., Chu,A.M., Connelly,C., Davis,K., Dietrich,F., Dow,S.W., El Bakkoury,M., Foury,F., Friend,S.H., Gentalen,E., Giaever,G., Hegemann,J.H., Jones,T., Laub,M., Liao,H., Davis,R.W. *et al.* (1999) *Science*, **285**, 901–906.
12. Hodges,P.E., McKee,A.H., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) *Nucleic Acids Res.*, **27**, 69–73. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 73–76.
13. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S. and Frishman,D. (1999) *Nucleic Acids Res.*, **27**, 44–48. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 37–40.
14. Bairoch,A. and Apweiler,R. (1999) *Nucleic Acids Res.*, **27**, 49–54. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 45–48.
15. Srinivasarao,G.Y., Yeh,L.S., Marzec,C.R., Orcutt,B.C., Barker,W.C. and Pfeiffer,F. (1999) *Nucleic Acids Res.*, **27**, 284–285.
16. Srinivasarao,G.Y., Yeh,L.S., Marzec,C.R., Orcutt,B.C. and Barker,W.C. (1999) *Bioinformatics*, **15**, 382–390.