Published online 3 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D577–D581***
doi:10.1093/nar/gkm909

Gene Ontology annotations at SGD: new data sources and annotation methods

Eurie L. Hong[1], Rama Balakrishnan[1], Qing Dong[1], Karen R. Christie[1], Julie Park[1], Gail Binkley[1], Maria C. Costanzo[1], Selina S. Dwight[1], Stacia R. Engel[1], Dianna G. Fisk[1], Jodi E. Hirschman[1], Benjamin C. Hitz[1], Cynthia J. Krieger[1], Michael S. Livstone[2], Stuart R. Miyasato[1], Robert S. Nash[1], Rose Oughtred[2], Marek S. Skrzypek[1], Shuai Weng[1], Edith D. Wong[1], Kathy K. Zhu[1], Kara Dolinski[2], David Botstein[2] and J. Michael Cherry[1,*]

[1]Department of Genetics, Stanford University, Stanford, CA, USA and [2]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA

Received September 17, 2007; Accepted October 6, 2007

ABSTRACT

The *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) collects and organizes biological information about the chromosomal features and gene products of the budding yeast *Saccharomyces cerevisiae*. Although published data from traditional experimental methods are the primary sources of evidence supporting Gene Ontology (GO) annotations for a gene product, high-throughput experiments and computational predictions can also provide valuable insights in the absence of an extensive body of literature. Therefore, GO annotations available at SGD now include high-throughput data as well as computational predictions provided by the GO Annotation Project (GOA UniProt; http://www.ebi.ac.uk/GOA/). Because the annotation method used to assign GO annotations varies by data source, GO resources at SGD have been modified to distinguish data sources and annotation methods. In addition to providing information for genes that have not been experimentally characterized, GO annotations from independent sources can be compared to those made by SGD to help keep the literature-based GO annotations current.

INTRODUCTION

Since 2001, the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) has used the Gene Ontology (GO) to annotate gene products in the budding yeast *Saccharomyces cerevisiae* (1,2). GO consists of three sets of structured, controlled vocabularies, also known as ontologies: the Molecular Function ontology describes the activities of gene products; the Biological Process ontology places molecular functions in a biological context; and the Cellular Component ontology describes the subcellular localizations of gene products (3). The selection of a GO term from one of these ontologies to annotate a gene product must be supported by a reference, such as a peer-reviewed research article or an abstract, as well as by an evidence code that describes the type of evidence present in that reference (4).

At SGD, results from traditional experimental methods published in the scientific literature are the primary sources of evidence used to support the GO annotation of gene products. If no experimental data are available for a gene, it is annotated to the terms 'biological_process', 'molecular_function' or 'cellular_component' (the root terms of the three ontologies) with the evidence code 'ND' to indicate there are 'No Biological Data Available'. While this does not describe the biology of the gene product, it indicates that no experimental results are available in the published literature at the time of annotation (Table 1). Using this curatorial process, every *S. cerevisiae* gene product has been assigned at least one GO term in each of the three ontologies since 2003.

In recent years, results from comparative sequence and genomic studies, as well as analyses of functional genomic and proteomic data, have provided valuable insights into the biological roles of gene products, especially when data from traditional experimental approaches are unavailable (5,6). In order to provide greater access to these results, SGD now incorporates these data as GO annotations. Because the process of assigning GO annotations from high-throughput experimental data and computational predictions differs

*To whom correspondence should be addressed. Tel: 650 723 7541; Fax: 650 725 1534; Email: cherry@stanford.edu

Table 1. Summary of annotation methods, sources and evidence codes used for GO annotations at SGD

Annotation method	Data source (No. of annotations)	Evidence code
Manually curated*	SGD (35 684) UniProt (93) MGI (8)	IDA: Inferred from Direct Assay IGI: Inferred from Genetic Interaction IMP: Inferred from Mutant Phenotype IPI: Inferred from Physical Interaction IEP: Inferred from Expression Pattern ISS: Inferred from Sequence/Structural Similarity IC: Inferred by Curator RCA: Reviewed Computational Analysis NAS: Non-traceable Author Statement TAS: Traceable Author Statement ND: No Biological Data Available
High-throughput*	SGD (4203)	IDA, IMP, IGI, IPI, IEP
Computational	UniProt (30959)	IEA: Inferred from Electronic Annotation

*Annotations generated by the manually curated and high-throughput methods are available from the GO Consortium (http://www.geneontology.org/GO.current.annotations.shtml). The total numbers of annotations are current as of September 2007. Numbers of manually curated annotations from GOA UniProt are cumulative since the January 2007 GOA UniProt data release. Because GOA UniProt compiles GO annotations from many sources, GO annotations are assigned by GOA UniProt and the Mouse Genome Informatics group (MGI; http://www.informatics.jax.org/). Numbers of Computational annotations from UniProt are from the June 2007 GOA UniProt data release. Documentation about evidence codes is available at http://www.geneontology.org/GO.evidence.shtml.

from the process of assigning annotations from traditional experimental studies, GO annotations in SGD are now distinguished by their annotation method.

INCORPORATING HIGH-THROUGHPUT DATA AT SGD

Traditional experimental methods, focusing on in-depth characterization of small numbers of genes, have been and will continue to be the primary source of evidence for GO annotations. However, modern techniques allow experiments to be designed on a genome-wide scale, generating data for large numbers of genes. SGD now assigns GO annotations based on data from such high-throughput experiments. These data sources have been particularly valuable in providing a nearly comprehensive set of Cellular Component GO annotations: from the GO annotation summary on SGD's Genome Snapshot, 5474 of 6301 gene products have been assigned at least one Cellular Component GO term as of September 2007, and 2238 of these are supported by data from high-throughput methods (7–9).

INCORPORATING GO ANNOTATIONS FROM GOA UNIPROT

In addition to data from high-throughput experimental methods, GO annotations can also be generated by computational analyses. For example, the Gene Ontology Annotation Project generates computationally predicted GO annotations for UniProt proteins based on sequence similarity algorithms (GOA UniProt; http://www.ebi.ac.uk/GOA/) (10,11). In order to provide greater access to these predictions, GOA UniProt annotations are now incorporated into SGD. Because these computationally predicted GO annotations are added without being reviewed in the context of literature-based GO annotations, they retain the 'Inferred from Electronic Annotation' ('IEA') evidence code assigned by GOA UniProt (Table 1).

Note that GOA UniProt also compiles literature-based GO annotations from many data sources (10). These annotations are also available at SGD, along with their original evidence codes and data sources, but are reviewed for redundancy with current SGD GO annotations before being incorporated (Table 1).

DIFFERENTIATING ANNOTATION METHODS

In addition to GO annotations derived from the manual curation of traditional experimental approaches published in the literature, SGD now contains GO annotations derived from data from high-throughput experiments as well as computational predictions provided by GOA UniProt, creating a central repository for all *S. cerevisiae* GO annotations. Although all of these annotations are supported by references and evidence codes, the basis for any differences among the GO annotations for any given gene may not be immediately clear. The curation process used for assigning GO annotations from these data varies according to the experimental approach. Therefore, in order to indicate how the data were curated, and to facilitate identification and comparison of these annotations, each GO annotation is now categorized in one of three annotation methods: manually curated, high-throughput or computational (Table 1).

The manually curated method indicates that the evidence in a publication has been individually reviewed to generate an annotation. Types of evidence can include experimental results in published literature that focuses on single genes or small sets of genes, author statements in a publication and sequence similarities that have been analyzed by the authors [for examples, see (12,13) shown in Figure 1B)].

The high-throughput method indicates that, although the evidence for a subset of results from a high-throughput or genome-wide experimental approach may have been reviewed, results for each gene product in the dataset have

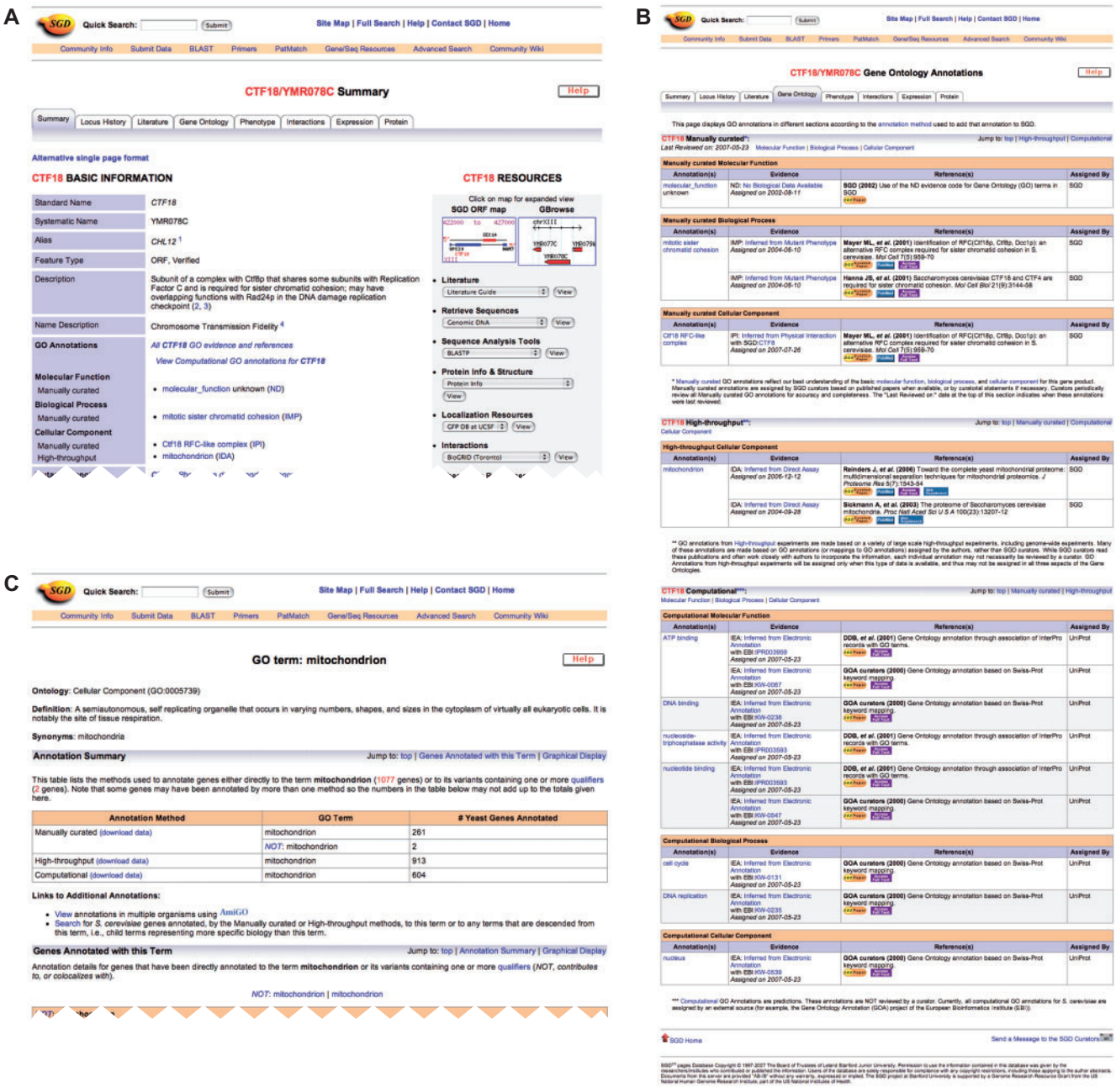

Figure 1. Modifications to SGD interfaces to display the different GO annotation methods and data sources. (**A**) Manually curated and high-throughput GO annotations are individually listed on the Locus Summary, and the computational GO annotations are available by the 'View Computational GO annotations' hyperlink. (**B**) The phrase 'GO Evidence and References' hyperlinks to the GO Annotations page, which is subdivided into three sections listing the reference and evidence code for each annotation, as well as additional supporting data used to make the prediction, such as the InterPro domain and the source of the data. (**C**) From the Locus Summary and the GO Annotation pages, each GO term is hyperlinked to its GO Term page, which lists every gene annotated to that term in SGD and provides the definition, any synonyms and a graphical representation of the GO structure for that GO term. A table summarizes the number of genes annotated to that term using each annotation method, and includes links to download data. Below this table, the genes annotated to this term are listed along with their relevant reference, evidence code and annotation method.

not been individually reviewed. Generally, this annotation method includes data from experimental approaches in which all significant results were produced using the same condition or analysis [for examples, see (7,8)].

In contrast, annotations generated by the computational method are not supported by direct experimental evidence and are not individually reviewed. These annotations include predictions generated by sequence similarity algorithms or by the integrated computational analyses of different sets of high-throughput experimental data that have not been individually reviewed [(for examples, see (11,14–17)].

All literature-based GO annotations from SGD and GOA UniProt are classified either as manually curated

or high-throughput. Computational predictions provided by GOA UniProt are classified as computational (Table 1).

MODIFICATIONS TO INTERFACES

SGD has changed several web interfaces in order to display data sources and annotation methods. The Locus Summary lists each manually curated and high-throughput GO annotation and indicates when computational GO annotations are available (Figure 1A). The phrases 'All GO Evidence and References' and 'View Computational GO annotations' are both hyperlinked to a detailed Gene Ontology Annotations page, which is subdivided into sections according to each annotation method. Because annotations no longer come solely from SGD, an 'Assigned by' column now indicates the data source (Figure 1B).

From the Locus Summary and GO Annotations pages, each GO term is hyperlinked to its GO Term page, which now lists all annotation methods used to generate that annotation for a particular gene. Annotations may be downloaded, according to annotation method, from the summary table at the top of the page (Figure 1C).

To ensure that data analyzed at SGD or by others in the scientific community are based on GO annotations supported by evidence in the published literature, only manually curated and high-throughput GO annotations are publicly available from the GO Consortium (http://www.geneontology.org/GO.current.annotations.shtml). They are also the default annotation sets used for SGD's GO Term Finder (http://www.yeastgenome.org/TermFinder) and GO Slim Mapper (http://www.yeastgenome.org/SlimMapper).

FUTURE DIRECTIONS

SGD will continue to update manually curated GO annotations as new experimental data are published and will add more sources of high-throughput and computational GO annotations. Discrepancies between annotations may become evident as GO annotations are made from different data sources and annotation methods. These differences can help refine GO and individual annotations by indicating areas in the ontology that require modification and gene products whose annotations need to be reviewed and updated to reflect the current literature. SGD will use this method of comparison to identify under-annotated gene products and areas in the GO structure that need to be reviewed.

SUMMARY

The incorporation of annotations from additional data sources makes SGD a central source for *S. cerevisiae* GO annotations. Differentiating these annotations by annotation method distinguishes what has been experimentally determined for each gene from what has only been computationally predicted. This knowledge will spur experimental research by contributing valuable information for genes that have not been experimentally characterized, and by suggesting additional roles for others (6).

SGD is committed to maintaining high-quality GO annotations and welcomes all comments or questions. Please contact us at: yeast-curator@genome.stanford.edu.

ACKNOWLEDGEMENTS

The SGD project is supported by a P41 grant from the NHGRI HG001315 (J.M.C.) and through the GO Consortium P41 grant from NHGRI HG002273 (co-PI J.M.C). Funding to pay the Open Access publication charges for this article was provided by the National Human Genome Research Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Gene Ontology Consortium, (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, 326.
2. Dwight,S.S., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
3. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
4. Gene Ontology Consortium, (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
5. Dolinski,K. (2005) Changing perspectives in yeast research nearly a decade after the genome sequence. *Genome Res.*, **15**, 1611–1619.
6. Pena-Castillo,L. and Hughes,T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**, 7–14.
7. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
8. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R. *et al.* (2002) Subcellular localization of the yeast proteome. *Gen. Dev.*, **16**, 707–719.
9. Hirschman,J.E., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G., Hong,E.L., Livstone,M.S. *et al.* (2006) Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, **34**, D442–D445.
10. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
11. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
12. Hanna,J.S., Kroll,E.S., Lundblad,V. and Spencer,F.A. (2001) *Saccharomyces cerevisiae* CTF18 and CTF4 are required for sister chromatid cohesion. *Mol. Cell. Biol.*, **21**, 3144–3158.
13. Mayer,M.L., Gygi,S.P., Aebersold,R. and Hieter,P. (2001) Identification of RFC(Ctf18p, Ctf8p, Dcc1p): an alternative RFC complex required for sister chromatid cohesion in *S. cerevisiae*. *Mol. Cell*, **7**, 959–970.
14. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. *et al.* (2003)

A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
15. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
16. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
17. Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.