Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO)

Selina S. Dwight, Midori A. Harris, Kara Dolinski, Catherine A. Ball, Gail Binkley, Karen R. Christie, Dianna G. Fisk, Laurie Issel-Tarver, Mark Schroeder, Gavin Sherlock, Anand Sethuraman, Shuai Weng, David Botstein and J. Michael Cherry*

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

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

The *Saccharomyces* Genome Database (SGD) resources, ranging from genetic and physical maps to genome-wide analysis tools, reflect the scientific progress in identifying genes and their functions over the last decade. As emphasis shifts from identification of the genes to identification of the role of their gene products in the cell, SGD seeks to provide its users with annotations that will allow relationships to be made between gene products, both within *Saccharomyces cerevisiae* and across species. To this end, SGD is annotating genes to the Gene Ontology (GO), a structured representation of biological knowledge that can be shared across species. The GO consists of three separate ontologies describing molecular function, biological process and cellular component. The goal is to use published information to associate each characterized *S.cerevisiae* gene product with one or more GO terms from each of the three ontologies. To be useful, this must be done in a manner that allows accurate associations based on experimental evidence, modifications to GO when necessary, and careful documentation of the annotations through evidence codes for given citations. Reaching this goal is an ongoing process at SGD. For information on the current progress of GO annotations at SGD and other participating databases, as well as a description of each of the three ontologies, please visit the GO Consortium page at http://www.geneontology.org. SGD gene associations to GO can be found by visiting our site at http://genome-www.stanford.edu/Saccharomyces/.

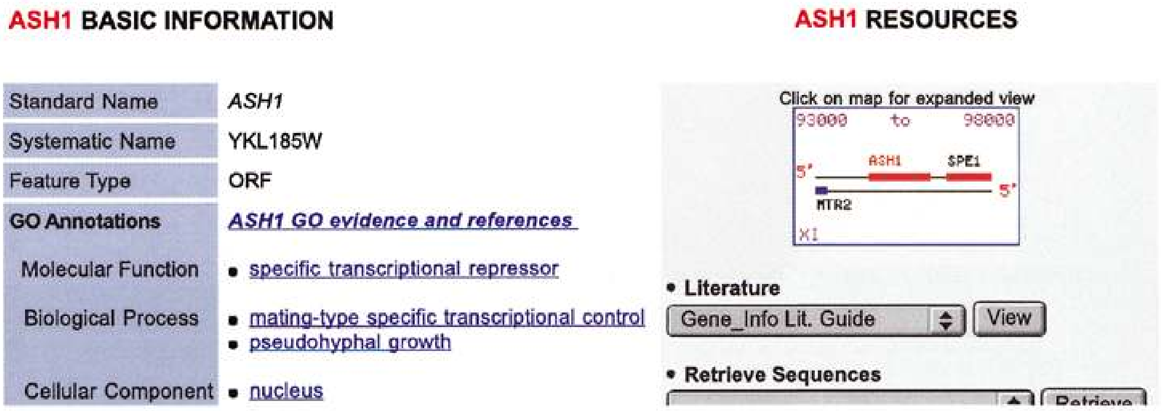
ANNOTATION GOALS AND GUIDELINES

The *Saccharomyces* Genome Database's (SGD's) (1–5) goal of annotating yeast genes to Gene Ontology (GO) (6,7) is to provide users with accurate information about the roles of gene products in the cell and their relationship to other gene products in yeast and other organisms. The availability of published experimental data for *Saccharomyces cerevisiae* as a model organism, and the participation of other organism databases (currently *Drosophila melanogaster*, *Mus musculus*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Schizosaccharomyces pombe*, *Dictyostelium discoideum* and *Plasmodium falciparum* and other parasites) and organizations [InterPro, SWISS-PROT, TrEMBL (8–10) and Compugen] in GO development and annotation make this possible. Complete annotation of *S.cerevisiae* genes to GO will allow users to find all genes, including those across species, which share the same (or related) annotation(s) for function, process and component.

GO consists of three ontologies, representing the fundamental aspects of gene products: molecular function, biological process and cellular component. Each ontology is structured such that specific terms are considered children of more broad terms. For instance, when describing localization, the cellular component term 'nucleus' may be considered more general than 'chromosome'. If a gene product is annotated to the cellular component term 'chromosome', then it is also implicitly annotated to 'nucleus', by virtue of the parent–child relationship between these GO terms. To appropriately model biological data, the structure allows for many-to-many relationships, such that nodes within the structure, representing individual biological concepts, may have many parents and many children, each connected by their relationship to one another (6). For instance, the process of 'DNA ligation' has parent terms of 'DNA recombination', 'DNA repair' and 'DNA-dependent DNA replication', as it is required for all of these processes. Gene products may be annotated to as many GO terms as needed, at the most specific levels possible, to reflect the current state of our understanding. Implicit to the integrity of this structure, all relationships between nodes must be true. The structure of GO allows relationships to be made between genes that share related as well as identical GO terms by exploring the surrounding GO structure. Another important clarification is that, while SGD GO associations are made between the gene/ORF and GO terms in the database, curators are actually annotating the gene product (rather than the gene itself) to the appropriate function, process or component term(s).

With these goals and considerations in mind, SGD curators annotate gene products to GO using the following guidelines:

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



ASH1 BASIC INFORMATION

Standard Name	*ASH1*
Systematic Name	YKL185W
Feature Type	ORF
GO Annotations	*ASH1 GO evidence and references*
Molecular Function	• specific transcriptional repressor
Biological Process	• mating-type specific transcriptional control • pseudohyphal growth
Cellular Component	• nucleus

ASH1 RESOURCES

Figure 1. GO display on the locus page at SGD. All existing GO terms to which a gene product has been annotated are listed on the gene product's locus page. Clicking on an individual GO term gives a list of all other *S.cerevisiae* loci which are associated with the GO term, along with the published references and corresponding evidence codes used to make the associations. Clicking on 'ASH1 GO evidence and references' results in a more detailed page that provides the references and evidence codes for each ASH1 annotation (Fig. 2).

(i) whenever possible, associations are made based on information obtained from published literature, (ii) associations are made to the most specific terms contained within the ontologies, (iii) each annotation requires a GO evidence code, and (iv) each annotation is associated with a literature citation. The third and fourth guidelines can be used to evaluate the confidence level of the association.

CREATING GO ASSOCIATIONS

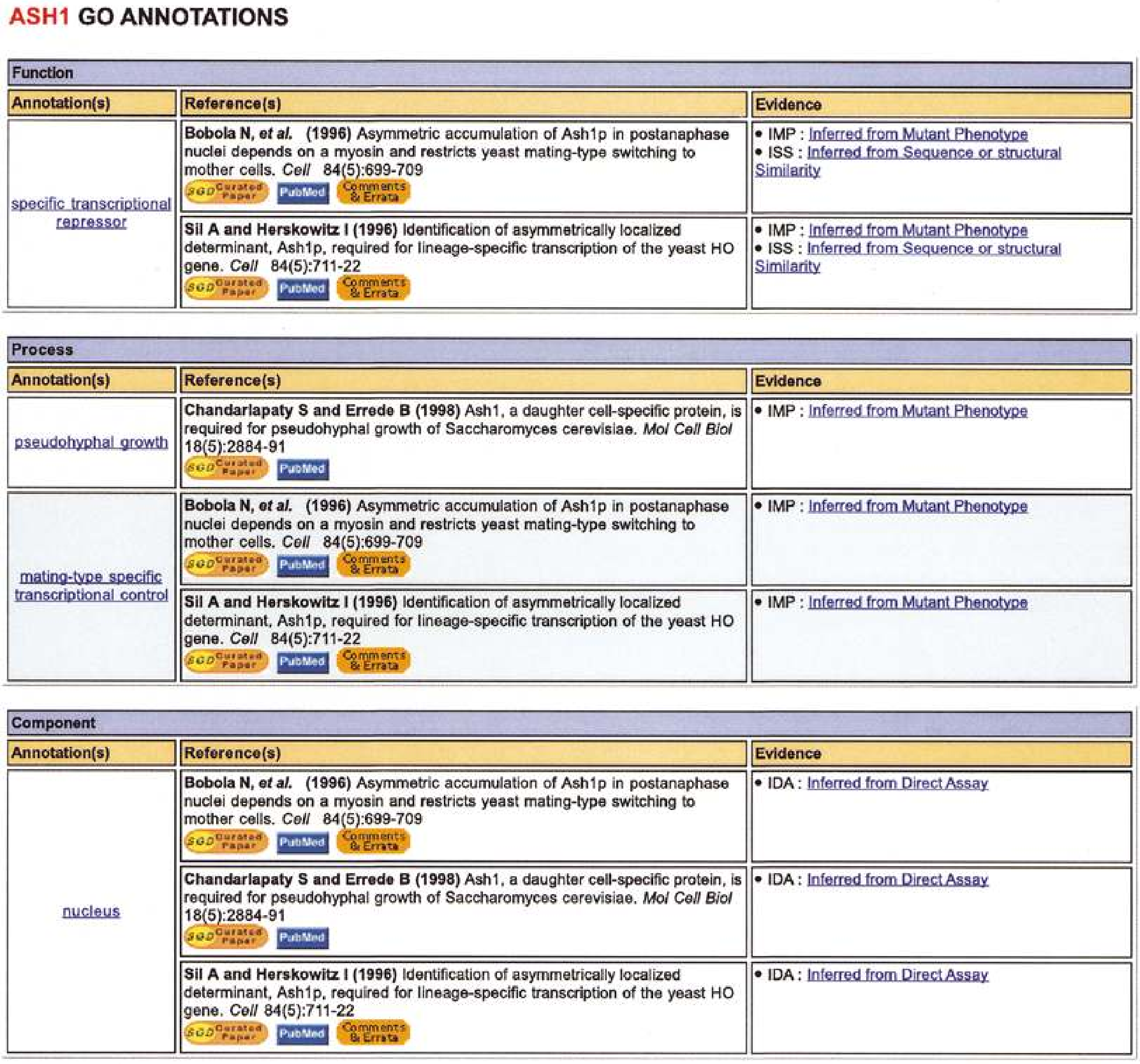
In the process of annotating genes to the Gene Ontology, SGD curators read published literature to capture information about a gene product's function(s), its role in biological process(es) and/or its cellular component(s). In cases where a gene product has a large amount of information associated with it, a review may serve as the primary source for information. The information found may be very general (e.g. overall process in which a gene is involved) or very specific (identification of the specific activity of a gene product). The next step is to browse ontologies to determine if there is an appropriate GO term (or terms) whose definition matches the information published for the gene. In many cases, an appropriate term exists. However, this process may also result in the suggestion for the modification of the GO structure and, in certain cases, for a localized restructuring of the existing ontology. In keeping with the philosophy of GO, it is possible to annotate a gene to more than one node within a specific ontology, thereby reflecting its multiple functions, roles or locations within the cell.

In addition to knowing the actual function, process or cellular component for a gene, it is also useful to know when a gene and/or its literature has been analyzed and no relevant information is found. For this purpose, the GO terms 'molecular_function unknown', 'biological_process unknown' and 'cellular_component unknown' exist. At SGD, these terms are used to distinguish between genes that have not yet been annotated, and those that have been checked for known functions and roles within the cell but no relevant information was found. Thus, before a gene has been investigated by curators with respect to GO, it will not have an associated GO annotation, whereas genes that have been investigated either by curators or electronic annotation (see below) will utilize the appropriate 'unknown' GO term(s). Either lack of evidence in the published literature, as determined by an SGD curator, or a statement from an author that specific information about a gene is unknown suggests the use of the appropriate 'unknown' GO term.

EVIDENCE CODES

As mentioned, evidence codes are integral to annotation of genes to GO. They themselves serve to annotate the association between the gene product and the GO; that is, they describe the certainty level of the association based on the evidence used to make the association. There are nine evidence codes that describe the type of information used to make the GO association. A list of these evidence codes and their guidelines for use are linked off the GO site (http://www.geneontology.org). Generally speaking, the evidence codes allow one to distinguish whether the association was made based on published information about mutant phenotype, sequence similarity, physical interaction, genetic interaction, expression assay or direct assay. In addition, the Traceable Author Statement (TAS) evidence code is used for statements made in review articles or books that are referenced by the author. The Non-traceable Author Statement (NAS) code is used by SGD for published statements that cannot be traced to a reference. One common use of NAS is for annotations to the 'unknown' terms when a curator has been unable to find information for a gene (in this case, SGD is used as a reference).

In SGD, annotations bearing the Inferred from Electronic Annotation (IEA) evidence code are the only GO associations in SGD that have not been reviewed by curators, and should therefore be regarded as approximate annotations. All currently existing IEA annotations have been assigned based on mapping previously existing terms in the database to GO terms. Specifically, all gene associations bearing the IEA evidence code represent exact matches between a previously existing term in SGD for a gene and an existing node within the GO structure. Thus far, there have been only two instances of IEA mapping at SGD, and both have utilized this conservative approach. The first involved mapping pre-existing SGD terms



ASH1 GO ANNOTATIONS

Function

Annotation(s)	Reference(s)	Evidence
specific transcriptional repressor	**Bobola N, *et al.* (1996)** Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. *Cell* 84(5):699-709 SGD Curated Paper PubMed Comments & Errata	• IMP : Inferred from Mutant Phenotype • ISS : Inferred from Sequence or structural Similarity
	Sil A and Herskowitz I (1996) Identification of asymmetrically localized determinant, Ash1p, required for lineage-specific transcription of the yeast HO gene. *Cell* 84(5):711-22 SGD Curated Paper PubMed Comments & Errata	• IMP : Inferred from Mutant Phenotype • ISS : Inferred from Sequence or structural Similarity

Process

Annotation(s)	Reference(s)	Evidence
pseudohyphal growth	**Chandarlapaty S and Errede B (1998)** Ash1, a daughter cell-specific protein, is required for pseudohyphal growth of Saccharomyces cerevisiae. *Mol Cell Biol* 18(5):2884-91 SGD Curated Paper PubMed	• IMP : Inferred from Mutant Phenotype
mating-type specific transcriptional control	**Bobola N, *et al.* (1996)** Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. *Cell* 84(5):699-709 SGD Curated Paper PubMed Comments & Errata	• IMP : Inferred from Mutant Phenotype
	Sil A and Herskowitz I (1996) Identification of asymmetrically localized determinant, Ash1p, required for lineage-specific transcription of the yeast HO gene. *Cell* 84(5):711-22 SGD Curated Paper PubMed Comments & Errata	• IMP : Inferred from Mutant Phenotype

Component

Annotation(s)	Reference(s)	Evidence
nucleus	**Bobola N, *et al.* (1996)** Asymmetric accumulation of Ash1p in postanaphase nuclei depends on a myosin and restricts yeast mating-type switching to mother cells. *Cell* 84(5):699-709 SGD Curated Paper PubMed Comments & Errata	• IDA : Inferred from Direct Assay
	Chandarlapaty S and Errede B (1998) Ash1, a daughter cell-specific protein, is required for pseudohyphal growth of Saccharomyces cerevisiae. *Mol Cell Biol* 18(5):2884-91 SGD Curated Paper PubMed	• IDA : Inferred from Direct Assay
	Sil A and Herskowitz I (1996) Identification of asymmetrically localized determinant, Ash1p, required for lineage-specific transcription of the yeast HO gene. *Cell* 84(5):711-22 SGD Curated Paper PubMed Comments & Errata	• IDA : Inferred from Direct Assay

Figure 2. Evidence code and reference display for locus specific GO annotations. Clicking on 'ASH1 GO evidence and references' from the ASH1 locus page (Fig. 1) returns to a page which lists each GO term and the reference(s) and evidence code(s) used for each ASH1 specific GO association. Links to the SGD, PubMed and existing full text journal entries for each publication, as well as any associated Comments & Errata, are provided for each supporting reference. Note that annotation to a single GO term may be supported by more than one reference. In addition, a single reference may offer more than one type of evidence for the annotation, in which case each evidence code is provided along with a link to its definition and selected examples.

for function and process to the GO molecular function and GO biological process ontologies. Because of the conservative approach of requiring exact matches, many of these were associations to the 'molecular_function unknown' and 'biological_process unknown' terms. However, 639 gene products received biological process and molecular function terms other than 'unknown'. The second set of IEA associations were produced by mapping Enzyme Commission (E.C.) terms that YPD (11–13) had associated with *S.cerevisiae* gene products to exact matches in the GO function ontology. This produced 353 new molecular function associations, all of which were something other than 'moleuclar_function unknown'.

The evidence codes thus allow users to weigh their confidence level in the information behind a gene's annotation to a given GO term. In keeping with this philosophy, it is possible for a gene to be annotated to two or more GO terms within different levels of a single ontology (e.g. to a parent and a child) because the annotations are made using different types of evidence. In each case, the appropriate reference is also linked to the annotation.

GO DISPLAY AT SGD

SGD has taken into account the importance of evidence codes and references in its display of GO information. Any existing

GO annotations for molecular function, biological process and/or cellular component for a gene are displayed on its locus page (Fig. 1). Each GO term is linked to a page that describes the GO term and lists all the *S.cerevisiae* loci annotated to that term, including evidence codes and references. On the locus page, there is also a prominent link (an example is entitled 'ASH1 GO evidence and references' in Fig. 1) which leads directly to a page listing all GO annotations for the gene, their associated evidence codes and their associated references. For an example of the information displayed, see Figure 2. Links to the SGD and PubMed entries for each reference, as well as any available full text journal links and/or web supplements are also provided, as is the case with all reference displays in SGD.

The current number of SGD annotations to each of the three ontologies (including and excluding those associated with the IEA evidence code) are listed on the GO site (http://www.geneontology.org). At SGD, a tab-delimited list of the existing GO annotations for yeast genes can be obtained from our ftp site: ftp://genome-ftp.stanford.edu/pub/yeast/data_dump/phenotype_go/.

Other existing data in the SGD Oracle database can now also be found at our ftp site at ftp://genome-ftp.stanford.edu/pub/yeast/data_dump/. The directories at this site are organized by logical groupings of data within the database, and are available at: http://genome-www4.stanford.edu/Saccharomyces/SGD/doc/db_specifications.html.

As stated earlier, SGD GO annotations are an ongoing process at SGD. It is through this and the continuing development of new resources and incorporation of new data that SGD hopes to allow its users to uncover relationships between gene products within *S.cerevisiae* and across species. For a list of features added to SGD within the past year, please visit our What's New site at http://genome-www.stanford.edu/Saccharomyces/whats_new00.html.

REFERENCES

1. 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.
2. 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 expresssion and functional analysis data. *Nucleic Acids Res.*, **29**, 80–81.
3. 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) Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature*, **387**, 67–73.
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) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
5. 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**, 74–78.
6. 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. *Nature Genet.*, **25**, 25–29.
7. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
8. Biswas,M., Kanapin,A. and Apweiler,R. (2001) Application of InterPro for the functional classification of the proteins of fish origin in SWISS-PROT and TrEMBL. *J. Biosci.*, **26**, 277–284.
9. Apweiler,R. (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Brief. Bioinform.*, **2**, 9–18.
10. Gasteiger,E., Jung,E. and Bairoch,A. (2001) SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.*, **3**, 47–55.
11. Costanzo,M.C., Hogan,J.D., Cusick,M.E., Davis,B.P., Fancher,A.M., Hodges,P.E., Kondu,P., Lengieza,C., Lew-Smith,J.E., Lingner,C., Roberg-Perez,K.J., Tillberg,M., Brooks,J.E. and Garrels,J.I. (2000) The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.*, **28**, 73–76.
12. Hodges,P.E., Payne,W.E. and Garrels,J.I. (1998) The Yeast Protein Database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **26**, 68–72.
13. Hodges,P.E., McKee,A.H., Davis,B.P., Payne,W.E. and Garrels,J.I. (1999) The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res.*, **27**, 69–73.