

Special issue: Gene Ontology for microbiologists

# Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns

Karen R. Christie, Eurie L. Hong and J. Michael Cherry

Department of Genetics, 300 Pasteur Drive, Stanford University Medical School, Stanford, CA 94305-5120, USA

**The quest to characterize each of the genes of the yeast *Saccharomyces cerevisiae* has propelled the development and application of novel high-throughput (HTP) experimental techniques. To handle the enormous amount of information generated by these techniques, new bioinformatics tools and resources are needed. Gene Ontology (GO) annotations curated by the *Saccharomyces* Genome Database (SGD) have facilitated the development of algorithms that analyze HTP data and help predict functions for poorly characterized genes in *S. cerevisiae* and other organisms. Here, we describe how published results are incorporated into GO annotations at SGD and why researchers can benefit from using these resources wisely to analyze their HTP data and predict gene functions.**

## Gene Ontology annotations aid functional genomics

*Saccharomyces cerevisiae* was the first eukaryotic organism whose nuclear genome was completely sequenced [1]. This paved the way for the development of strain collections in which every protein-coding gene in the genome was modified – for example, by deletion, tagging with green fluorescent protein (GFP) or engineering for overexpression [2–4]. Coupled with advances in technology that allow transcribed regions of the genome to be detected by microarrays or protein abundance to be detected by mass spectrometry, these resources have enabled researchers to experimentally survey the *S. cerevisiae* genome and proteome [5–7].

The pioneering position of *S. cerevisiae* as a model organism in the genomics era is based not only on its experimental tractability and a complete genome sequence but also on the fact that the extensive literature is curated using Gene Ontology (GO), which enables researchers to make sense of large quantities of data [8]. The GO Consortium has developed and continues to update three structured, controlled vocabularies to describe a gene product: molecular function, biological process and cellular component [9] (Box 1). With these three vocabularies, GO provides a common language – used by a growing number of research projects and information resources working in different model organisms – to describe the functions of gene products from many species [10]. This widespread use

has facilitated the comparison of shared functions among hundreds of organisms, the functional annotation of newly sequenced genomes and the analysis of many types of data. GO annotations have become the primary resource used to facilitate the annotation of microarray expression profiles, protein interaction networks and regulatory modules [8,10]. The interested reader can find more articles on GO and its applications in this issue of *Trends in Microbiology*.

Since 2001, the *Saccharomyces* Genome Database (SGD) has used GO to provide descriptions, or annotations, of the functional roles of gene products in *S. cerevisiae* based on the published literature (<http://www.yeastgenome.org/>) [11]. In 2003, *S. cerevisiae* became the first organism with at least one GO annotation in each of the three GO vocabularies for every protein-coding and RNA gene. In this article, we describe how these GO annotations are curated at SGD to represent the current state of knowledge about the biology of *S. cerevisiae*, as well as how the scientific community has used these annotations. Because we cannot provide a comprehensive review of the entire body of *S. cerevisiae* literature that uses GO annotations (over 700 publications, as of April 2009) here, we highlight applications that facilitate the functional characterization of genes in *S. cerevisiae* and other organisms. We also describe why understanding the process of making GO annotations can improve the results produced by these applications.

## How GO annotations are made at SGD

The core of a GO annotation comprises a gene product, a GO term from one of the three vocabularies, a literature reference and an evidence code (Box 1,2) [12]. At SGD, GO annotations for all genes are curated from the primary research literature. This means that a curator – a PhD-level biologist who is an expert at abstracting information from the literature – has read the published work and determined the appropriate GO annotation(s) to describe the experimental results in that paper. For this purpose, all available literature for a gene is reviewed to identify experimental data and sequence-based predictions that characterize its molecular activity, cellular localization or biological role (Figure 1a,b). Thus, the set of manually curated GO annotations for all protein-coding and RNA genes represents the current collective view of the yeast research community.

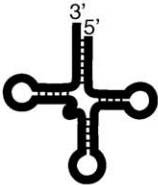
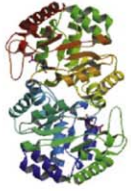
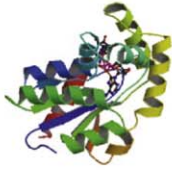
Corresponding author: Cherry, J.M. ([cherry@stanford.edu](mailto:cherry@stanford.edu)).

### Box 1. Key elements of a Gene Ontology annotation

The Gene Ontology (GO) develops three structured vocabularies, also referred to as GO aspects [9]. The Molecular Function vocabulary represents basic activities, such as catalysis or binding. The Biological Process vocabulary represents the larger cellular goals that are accomplished by multiple molecular functions, such as signal transduction or pyrimidine metabolism. The Cellular Component vocabulary represents locations in the cell, from large structures such as the nucleus to smaller structures such as a protein complex. More information about these vocabularies is available from the GO website (<http://www.geneontology.org/GO.doc.shtml#ontologies>)

For many users, it is the association of GO terms with individual genes that makes GO useful. Scientific curators associate a specific GO term to a specific gene to create a GO annotation [12]. All genes that produce a gene product, whether protein or RNA, can be associated

with a GO term. In addition to a gene and a GO term, a GO annotation also includes the source of the information supporting the association, as well as an evidence code (Figure 1). The reference is usually a published paper with a PubMed ID but is sometimes an unpublished abstract describing a method of assigning GO annotations used within the GO Consortium. The evidence code indicates the type of evidence that supports the annotation (Box 2). To avoid possible confusion caused by multiple uses of the same gene name in the published literature or from GO terms with similar names, the annotations are made using unique alphanumeric IDs for genes, GO terms and references. More information about GO annotations is available from the GO website (<http://www.geneontology.org/GO.format.annotation.shtml>). The ribbon diagrams of URA3 [70] and URA6 [71] were contributed to the Protein Data Bank (PDB; [www.pdb.org](http://www.pdb.org)) [72].

Gene product	GO aspect	GO term	Evidence	Reference
 <p>SUP2 tRNA-Tyr (SGD:S000006778)</p>	Molecular function (MF)	Triplet codon-amino acid adaptor activity (GO:0030533)	TAS - traceable author statement	Hani J and Feldmann H (1998) tRNA genes and retroelements in the yeast genome. Nucleic Acids Res 26(3):689-96 (PMID:9443958)
 <p>URA3 orotidine-5'-phosphate decarboxylase (SGD:S000000747)</p>	Biological process (BP)	'de novo' pyrimidine base biosynthetic process (GO:0006207)	IMP - inferred from mutant phenotype	Lacroute F (1968) Regulation of pyrimidine biosynthesis in <i>Saccharomyces cerevisiae</i> . J Bacteriol 95(3):824-32 (PMID:5651325)
 <p>URA6 uridylylate kinase (SGD:S000001507)</p>	Cellular component (CC)	Cytoplasm (GO:0005737)	IDA - inferred from direct assay	Jong A, et al. (1993) Characteristics, substrate analysis, and intracellular location of <i>Saccharomyces cerevisiae</i> UMP kinase. Arch Biochem Biophys 304(1):197-204 (PMID:8391780)

TRENDS in Microbiology

**Figure 1.** Examples of *S. cerevisiae* GO annotations. Each row is an example of a GO annotation, which includes a protein or RNA gene product, a GO term, a reference and an evidence code (Box 2). The ribbon diagrams of URA3 [70] and URA6 [71] were contributed to PDB [72].

Some protein-coding genes also have GO annotations derived from high-throughput (HTP) experimental data (Figure 1c) or from computational prediction methods (Figure 1b,d) [13]. At present, RNA genes have only manually curated GO annotations because these genes are generally not included in HTP experiments.

As a pioneer model organism with a small, completely sequenced genome, there are a wide range of HTP studies

for *S. cerevisiae* (papers describing HTP or genomic studies are listed at <http://www.yeastgenome.org/cache/genome-wide-analysis.html>). Some of these HTP studies can be used to support GO annotations, whereas others are incorporated as alternative types of annotations: phenotypes are curated using SGD's new phenotype curation system [14], and the full dataset is available ([http://downloads.yeastgenome.org/literature\\_curation/phenotype\\_data.tab](http://downloads.yeastgenome.org/literature_curation/phenotype_data.tab)); curated

**Box 2. GO evidence codes**

The GO Consortium uses a small set of evidence codes to provide a general idea of the type of evidence supporting an annotation. Curator-assigned evidence codes require that a curator has read the paper or analyzed the data to use one of these codes. The curator-assigned evidence codes can be divided into four categories: experimental, computational analysis, author statement and curator statement (Table 1). Experimental evidence codes indicate experiments (mutants, genetic analyses, enzyme assays, physical interactions, etc.) reported in the paper cited. Computational analysis codes indicate annotations based on computational analyses of various types, often involving sequence data, high-throughput experimental data or a combination of multiple data types. Author statement codes indicate that the annotation is based on an author statement in a published paper, often a review. The statement is considered

traceable when another reference is cited and non-traceable when no reference is associated with that statement. Curator statement codes indicate judgments made by the curator based on understanding of the biology; for example, a gene shown to be a transcription factor for RNA polymerase II must be in the nucleus to function, so the curator can make an annotation to the term 'nucleus' using the IC (inferred by curator) code, or when an overall review of the literature indicates that there is no information, the ND (no biological data available) code can be used. There is also one code, IEA (inferred from electronic annotation), for use when annotations are made automatically by a computational method without curator review (e.g. running InterProScan and applying the *interpro2* go mapping file without any curatorial judgment to approve the resulting annotations).

**Table 1. Categories of GO evidence codes**

Evidence code category	Code	Evidence code full name	Type of evidence
<b>Curator-assigned evidence codes</b>			
Experimental	EXP	Inferred from experiment	Any experimental evidence
	IDA	Inferred from direct assay	Enzyme assays, <i>in vitro</i> reconstitution, immunofluorescence, etc.
	IPI	Inferred from physical interaction	2-Hybrid interactions, co-purification, co-immunoprecipitation, etc.
	IMP	Inferred from mutant phenotype	Mutations, allelic variation, phenotypes of altered expression, etc.
	IGI	Inferred from genetic interaction	Genetic suppression, synthetic lethality, complementation, etc.
	IEP	Inferred from expression pattern	Northerns, Westerns, microarray expression, etc.
Computational analysis	ISS	Inferred from sequence or structural similarity	Any sequence-based evidence
	ISO	Inferred from sequence orthology	Assertion of orthology to gene in another species
	ISA	Inferred from sequence alignment	Pairwise or multiple alignment
	ISM	Inferred from sequence model	Sequence models (e.g. Hidden Markov Models, tRNASCAN, InterPro domains, etc.)
	IGC	Inferred from genomic context	Operon structure, syntenic regions, pathway analysis, etc.
	RCA	Inferred from reviewed computational analysis	Predictions based on one or more data types
Author statement	TAS	Traceable author statement	Author statements citing a reference
	NAS	Non-traceable author statement	Author statements not citing a reference
Curator statement	IC	Inferred by curator	When a curator makes an inference based on another GO annotation
	ND	No biological data available	When there is no information available on that gene product
<b>Automatically assigned evidence codes</b>			
	IEA	Inferred from electronic annotation	From computational methods without curatorial involvement

protein-protein and genetic interactions are available from BioGRID (a curated database for interaction data; see <http://www.thebiogrid.org/>) [15]; and expression and functional genomic data are available via the Yeast Functional Genomics Database (<http://yfgdb.princeton.edu/>).

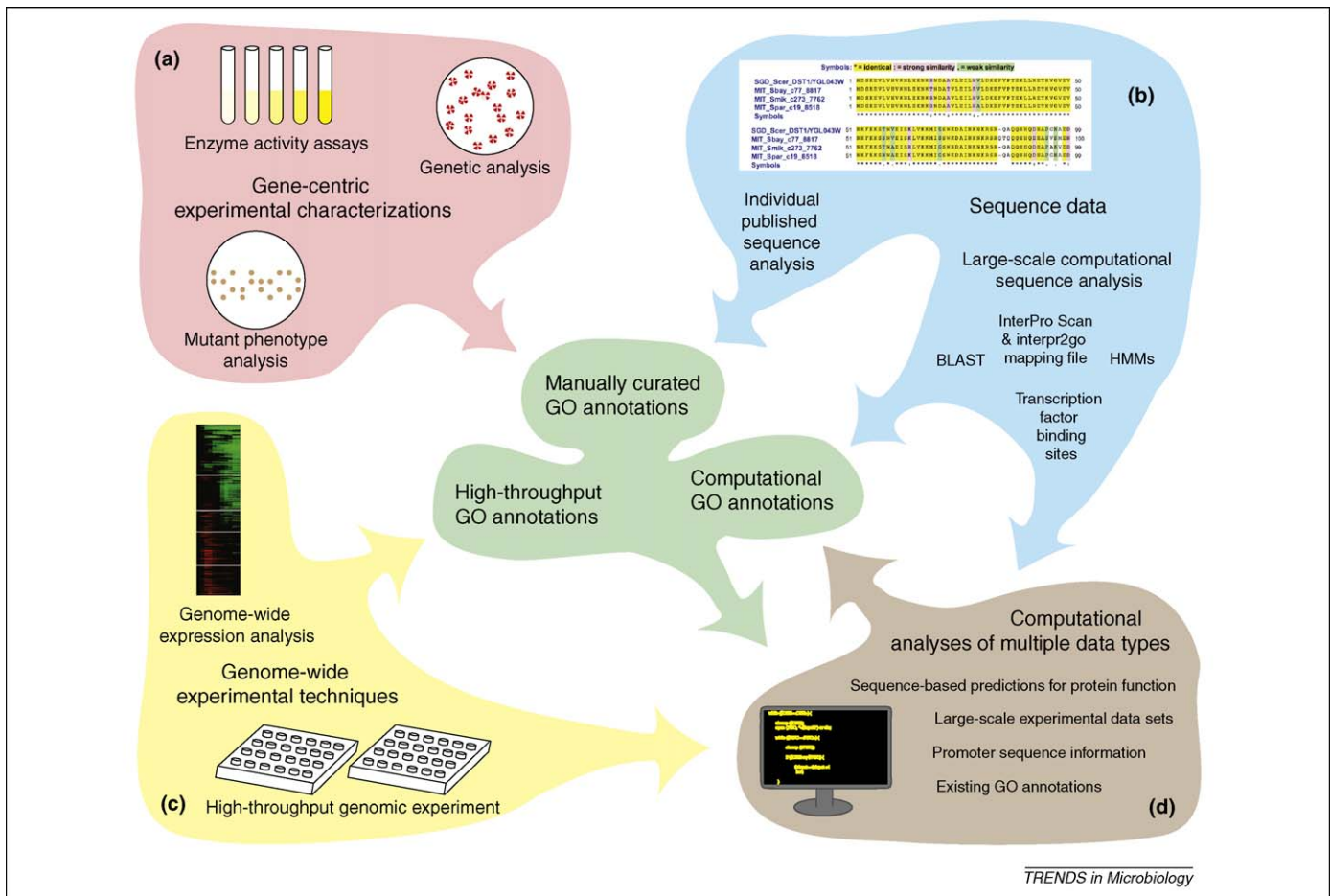
HTP data are incorporated as GO annotations when the data indicate a gene product is directly involved in the process being studied. We believe that HTP phenotype data often identifies many genes whose mutations affect broad processes owing to an indirect, downstream effect. For example, abnormal telomere length is a mutant phenotype observed for hundreds of genes [16,17]. Although the observation of shortened telomeres is biologically relevant, additional analysis must be done to judge whether the identified genes have a direct role in telomere maintenance [18,19]. Therefore, although these HTP datasets are used to make phenotype annotations in SGD [14], they are generally not represented as GO annotations. Similarly, large-scale expression studies are not used to assign GO annotations. For example, genes whose expression changes in response to sulfite are not annotated to the biological process terms 'sulfite metabolism' or 'sulfite detoxification' [20]. Although it is true that the

expression of these genes changes, it is not clear from these data which genes have a direct role in the response of the cell to the tested condition.

As a consequence of our selectivity in the use of HTP data, only ~40% of the genome's protein-coding genes have such GO annotations (Table 1). Almost all of these annotations are from studies that examined the localization of proteins using a single experimental method, such as visualization using a GFP marker or purification of an organelle [3,21,22], and the cellular component terms used are general ones, such as 'cytoplasm' or 'nucleus'. Fewer than 300 genes, ~5.5% of the protein-coding genes in the

**Table 1. Numbers of protein coding genes (5796 total) annotated by each annotation type for each Gene Ontology vocabulary, as of April 2009**

Annotation type	Gene Ontology vocabulary		
	Molecular function	Biological process	Cellular component
Manually curated	5771 (99.6%)	5770 (99.6%)	4696 (81.0%)
High-throughput	153 (2.6%)	172 (3.0%)	2344 (40.4%)
Computational	3388 (58.5%)	4554 (78.6%)	4395 (75.8%)



**Figure 1.** GO annotation types at SGD and sources of information. At SGD, GO annotations are made based on a wide range of published literature. Each GO annotation is further categorized with an annotation type: manually curated, high-throughput or computational [13]. (a) Manually curated GO annotations are made individually for each gene by curators reading the published literature describing experimental characterizations of that gene. We attempt to find experimental evidence whenever available. However, in our first pass through the genome to generate at least one annotation in each GO vocabulary for all genes, we sometimes made annotations from reviews using the TAS (traceable author statement) code (Box 2). We are working to replace these with annotations from the primary experimental papers with appropriate experimental evidence codes. (b) Sequence-based predictions can be classified as either manually curated or computational GO annotations. Sequence similarity comparisons from published papers are categorized as manually curated GO annotations because an expert in the field generated the comparison and a curator read the publication to determine the appropriate annotation. Predictions generated by the Gene Ontology Annotation group at the European Bioinformatics Institute are categorized as computational annotations because they are not reviewed by curators. (c) HTP annotations are made from published papers describing results of HTP experimental techniques. (d) Computational annotations are based on a variety of computational techniques, including sequence similarity and integrative analysis of experimental data. Computational methods that incorporate HTP data and sequence analysis should take caution to remove GO annotations derived from the source data to avoid including the information more than once.

genome, have molecular function or biological process annotations based on HTP data (Table 1).

In the absence of published literature describing focused or HTP experimental characterization of a gene, an annotation is made in each GO vocabulary using the ND (no biological data available) evidence code (Box 2). This indicates that the literature for the gene has been reviewed by curators and no information characterizing the role of the gene has been published.

GO annotations based on computational analyses (Figure 1b,d) were added to SGD in 2007 [13]. These annotations help researchers generate hypotheses of potential functions to test, particularly for experimentally uncharacterized genes. Two types of computational predictions available at SGD are protein domain predictions (from sequence analysis) and high-confidence predictions (based on integrated computational analyses of multiple HTP experimental datasets). The sequence-based predictions are provided by the Gene Ontology Annotation (GOA) group at the European Bioinformatics Institute (EBI) [23,24]. The

predictions based on integrated computational analyses of various types of HTP experiments and sometimes sequence or other information are produced by published algorithms [25,26]. Because these annotations are not individually reviewed by curators, we require these annotations be updated at least once a year. Computational annotations that have not been recalculated after one year are removed from SGD.

Thus, the manually curated set of GO annotations primarily represents the results of small-scale, gene-by-gene characterizations. For the majority of the protein-coding genes (over 90%), these are supplemented by computational predictions. For slightly less than half of the protein-coding genes (44%), there are also HTP annotations, mostly to cellular component terms (Table 1, Figure 1).

### Using GO annotations to advance *S. cerevisiae* experimental research

The availability of GO annotations in each of the three GO vocabularies for every *S. cerevisiae* protein-coding and

RNA gene has transformed the analysis methods available to bench biologists. As HTP resources and methods have become more widely available, the use of tools based on GO annotations has become more important for identifying a function, process or localization shared among a set of genes. The frequency of this type of usage is underestimated when searching the published literature (for instance, using PubMed): although authors might cite the reference for a specific GO analysis tool, it is rare to find a citation for the GO project or SGD as the source of the *S. cerevisiae* annotations. In fact, some researchers make no citation, demonstrating that the classification of genes using GO terms has become an accepted tool for molecular genetics.

One of the first suggested applications for GO is still widely used: the identification of a common biological role for genes that are part of an interesting cluster of microarray expression data [9,27]. However, there are many other experimental methods that produce lists of genes that can be analyzed with GO. Examples include gene sets having a genetic interaction with a target gene [28–30], genes whose mutants share a common phenotype [31–33], genes whose transcription levels might contribute to different morphological traits in different strain backgrounds [34], genes whose messenger RNAs (mRNAs) are poorly translated [35], protein interaction networks [36,37] and proteins that interact with a tagged protein or an mRNA [38,39].

Regardless of the type of experiment that generates the list of genes, many researchers use freely available tools to identify the function, process or localization that is enriched in the list. Such tools include GO Term Finder from SGD (<http://www.yeastgenome.org/TermFinder>) [40] and other analytical tools listed on the GO Consortium website (<http://www.geneontology.org/GO.tools.shtml>) [41]. Although each tool has its unique features, its input is typically a list of genes and its output is the identification of GO terms significantly shared by those genes. For instance, Georgiev and collaborators used the SGD GO Term Finder to discover that the Syh1p and Smy2p proteins (both containing a particular domain known as GYF) might be involved in mRNA catabolism, based on a list of proteins that interacted with them. This result enabled the researchers to test and confirm that these two GYF proteins localize to cytoplasmic mRNA processing bodies [42].

The use of GO annotations to identify the commonalities within a set of genes to make hypotheses for subsequent experiments has clearly become routine in the research community, but knowing which annotations are being used for the analysis and what types of evidence support these annotations can provide more accurate results [10,41] (Figure 1). For example, the GO Term Finder available at SGD does not use computationally predicted annotations when finding a function, process or localization shared among a list of genes. Excluding these computational predictions ensures that the analysis is based on annotations made from the primary literature, both small-scale and HTP experiments. Therefore, we advise that researchers should consider removing annotations made from computational or automated methods (including the RCA and IEA evidence codes; Box 2) when using other tools, to avoid propagating untested hypotheses.

The availability of GO annotations for *S. cerevisiae* and web-based tools that analyze gene lists based on these annotations have facilitated the analysis of HTP data. However, it is essential that the researcher understand how GO annotations are made to select the correct set of annotations to analyze their HTP results effectively.

### Extracting functional information from HTP data in *S. cerevisiae*

Any list of genes derived from an experimental assay might contain one or more *S. cerevisiae* genes that lack an informative GO annotation because of the absence of direct experimental evidence. Once a shared biological process has been identified for the characterized genes in the list, it has been common practice to transfer that process to the experimentally uncharacterized genes solely based on their presence in the same list [43]. Although this transfer of annotations can be misleading, the continued development of sophisticated algorithms has strengthened the predictive power of HTP data by using existing GO annotations in novel approaches.

Functional predictions for experimentally uncharacterized genes have benefited from the inclusion of *S. cerevisiae* GO annotations and the GO vocabularies as integral components of algorithms that analyze microarray and protein–protein interaction data. Some of the newer algorithms that group genes according to similar microarray expression patterns also include GO annotations to help generate biologically relevant clusters and improve functional predictions [44–46]. Not surprisingly, functional predictions using two or three GO vocabularies uncover details about the expression data more effectively than those using annotations from only one GO vocabulary [47]. In addition to using more than one vocabulary, integrating the relationships between GO terms defined in the GO biological process with protein–protein interactions improves the accuracy of the annotations [48].

More recent methods have taken an integrated approach, combining multiple types of experimental data to identify the functions of proteins [49]. For example, algorithms developed by the Troyanskaya, Marcotte and Roth groups analyze data from diverse experimental sources based on genes that have common GO annotations or use the annotations to describe the genes that have been grouped together based on data with similar patterns [25,26,50–52]. The utilization of multiple types of HTP experimental data (such as expression and protein–protein interaction datasets), in addition to more sophisticated uses of GO (such as including GO annotations from more than one vocabulary or taking advantage of the GO structure), might improve the functional predictions [26,53].

In addition to developing more sophisticated algorithms, an understanding of GO annotation practices and guidelines is essential to obtain the best quality results [10]. It is important not to include the same information twice, once as primary data and a second time as the GO annotation derived from it (Figure 1). For example, to avoid falsely emphasizing the significance of a single HTP dataset or the GO annotations derived from it, algorithms that combine annotations with HTP datasets must exclude any annotations derived from the publications describing those

HTP datasets. Similarly, algorithms that include protein domains should exclude GO annotations assigned based on the presence of those protein domains. However, researchers evaluating a new prediction algorithm might choose annotations from a single source as an appropriate comparison set (e.g. using all the annotations based on InterProScan, a tool that detects specific motifs and signatures in proteins [24], from the GOA group at EBI to benchmark a new algorithm based on protein domains).

New algorithms and bioinformatics tools have been developed to extract functional information from numerous HTP data. We advise those groups combining GO annotations with HTP datasets to review the references used to make the annotations to select appropriate annotations for their analysis.

### Use of *S. cerevisiae* annotations to predict gene functions in other organisms

An early prediction by Ashburner and collaborators was that GO would enable the transfer of functional annotations to newly sequenced genomes [9]. This vision has been realized with the use of *S. cerevisiae* annotations to make functional annotations for genes in newly sequenced genomes with small research communities. Based on sequence similarity, the *S. cerevisiae* annotations were transferred to genes of the filamentous fungus *Ashbya gossypii*; the fungal pathogens *Pneumocystis carinii*, *Sclerotinia sclerotiorum* and *Candida albicans*; and more distant organisms, such as the compost worm *Eisenia fetida* [54–58].

Because the majority of *S. cerevisiae* GO annotations are derived from experimental evidence, they have been used to determine the accuracy of predictions. For example, SGD's annotations have been used to validate functional predictions based on sequence similarity, by determining whether the predicted function matches the manually curated GO annotation [59]. Some researchers have even compared the results from their analysis based on the full set of annotations to those from the same algorithm run with only a subset of the annotations, as proof of concept that their method would be suitable for predicting gene functions for poorly annotated genomes [60]. In these methods, *S. cerevisiae* GO annotations provide a gold standard for measuring the accuracy of functional predictions. Once validated using *S. cerevisiae* GO annotations, new algorithms utilizing microarray expression data, protein–protein interactions, sequence similarity or a combination of these data can improve the functional predictions for genes from many other organisms.

Caution must be exercised in transferring annotations [43]. Genes in closely related species that seem to have a common evolutionary origin might have a conserved function, such as transcription factor activity, but be involved in regulating very different processes [61,62]. Snitkin and collaborators showed that phylogenetic profiling methods exploring the co-occurrence of multiple genes between genomes do not work well for eukaryotic genomes [63]. In summary, although GO annotations from *S. cerevisiae* have been successfully used to facilitate the annotation of other genomes, the question of which annotations should be transferred depends on the specific species and the role of the gene.

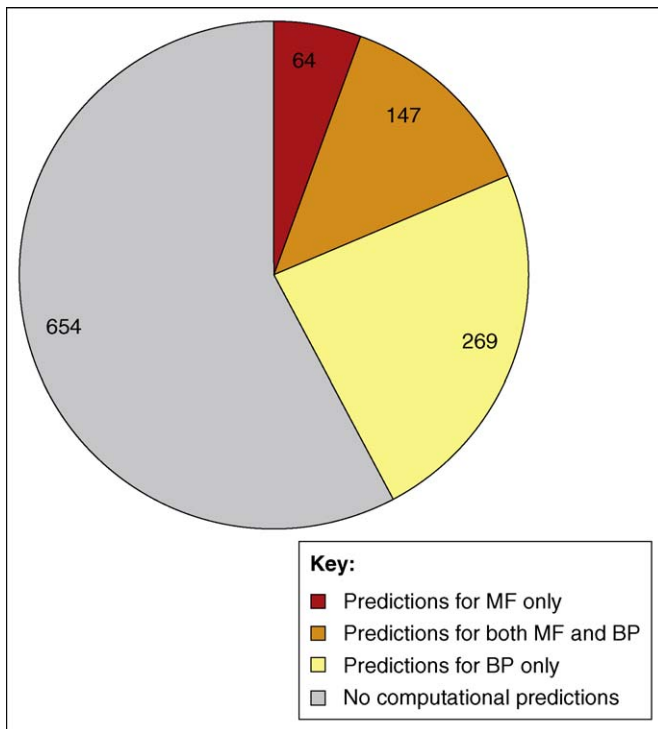
### Automated methods to extract information from the literature

Although they are informative, annotation transfers based on sequence similarity are still hypotheses for gene functions that need to be proven experimentally. Thus, the standard for functional annotations is a comprehensive set of GO annotations derived from the scientific literature. Unfortunately, the development of such a set might not be possible for model organism communities with a large body of literature but limited curation resources. Natural language processing and text mining can facilitate the identification of literature to be used for GO annotations and, thus, maximize the effectiveness of a small curatorial staff [64,65]. For instance, *S. cerevisiae* GO annotations have been used to validate a full-text analysis that identified papers supporting GO annotations in the molecular function vocabulary by searching for specific experimental methods [66]. Like gene function prediction algorithms, these tools – once developed and validated using *S. cerevisiae* GO annotations – can be used by other model organism communities.

### Continuing to improve the functional annotation of genes in *S. cerevisiae*

Intriguingly, the number of *S. cerevisiae* genes lacking any functional annotations has remained consistent through the years [67]. Despite the vast body of literature for this organism, 554 out of 5796 protein-coding genes (almost 10%) remain uncharacterized for all three GO vocabularies. Approximately 500 additional genes only have annotations to very general locations, such as 'cytoplasm', defined by HTP localization experiments. Thus, we believe that the set of uncharacterized genes is better represented by the number of genes with GO annotations indicating that no information is available in both the molecular function and the biological process vocabularies. As of April 2009, there are 1134 protein-coding genes in this group. Computational predictions based on either sequence similarity or integrated computational analysis of experimental and other data provide hypotheses about the biological process of only one-third of these 1134 genes and about the molecular functions of only one-fifth of them (Figure 2). Although broad terms like 'cytoplasm' and 'membrane' have been assigned for many of these genes by computational predictions, we again believe that these annotations are not informative about the gene's role in the cell. Thus, for the majority of these undercharacterized genes, there is not a prediction for either the function or the process, and we still have no inkling what role they have in the cell.

Although predicting gene functions based on the integrated analysis of multiple datasets can provide hypotheses for genes that lack GO annotations, this analysis is wholly dependent on the experimental conditions examined. Perhaps some uncharacterized genes cannot be classified because the experimental condition necessary to observe their function has not yet been examined. For example, to the best of our knowledge, no publications have reported HTP protein interaction networks or genetic interaction datasets during meiosis and sporulation in *S. cerevisiae*. Experiments during meiosis or sporulation



**Figure 2.** Computational predictions for the uncharacterized protein-coding genes in *S. cerevisiae*. Out of 5796 protein coding genes, 1134 of them have no published information with regard to their molecular function (MF) or their biological process (BP). Predictions for the biological process can be made for only 416 (36.7%) of them, and predictions for molecular function can be made for even fewer, only 211 (18.6%). For the majority (654, or 57.7%), no prediction can be made for either molecular function or biological process to provide hypotheses for biologists to test experimentally.

would provide additional data specific for these conditions, which could help identify any uncharacterized genes involved in these processes.

In March 2007, Peña-Castillo and Hughes [67] revisited a prediction made three years earlier [68] that all *S. cerevisiae* genes would have a function by mid-2007. However, they determined that 1253 genes, over 20% of the genome, were still classified as 'Uncharacterized' at SGD [67]. They identified ~200 genes that were found only in fungi [67]. Therefore, more research in other fungal species could help characterize some of the fungal-specific genes found in *S. cerevisiae*; of particular interest are those fungi that are studied specifically for their niche specialization, such as *C. albicans* (with respect to the medical implications of biofilm formation) or *Aspergillus fumigatus* (a common pathogen in immunocompromised patients). Another group of more than 150 uncharacterized genes contained sets of genes having at least 50% sequence similarity to each other [67]. Owing to their redundancy, the corresponding proteins will be difficult to analyze via single mutations, and those in groups with more than two members will probably remain resistant to characterization by techniques such as synthetic genetic analyses, which generally involve two mutations. However, for the majority of the uncharacterized genes, there is no clear single explanation for why they are refractory to characterization [67]. To begin to learn what these genes do might take more refined experimental genomic approaches, in addition to exploring other environmental conditions and

### Box 3. How to find functions for uncharacterized genes in *S. cerevisiae*

- Understand how GO annotations are made and identify which ones should be included in or excluded from your analyses.
- Use mutant phenotype data from SGD [14], as well as genetic and physical interactions from BioGRID [15], to complement the GO annotations.
- Generate datasets using different experimental conditions to expand the experimental conditions tested. For example, because no HTP protein interaction networks or genetic interaction datasets have been published for *S. cerevisiae* during meiosis, it is difficult to predict which uncharacterized genes might be involved in meiosis using current algorithms that group genes according to similar patterns across multiple datasets.
- Include RNA genes in HTP studies and computational predictions. New RNAs whose functions are not yet known have been reported and added to the set of genes in *S. cerevisiae* at SGD [73,74]. Including these RNA genes in HTP studies might facilitate identifying their functions.
- Algorithms should consider more than one GO vocabulary at a time and/or the structure of the GO vocabularies.

developing more sophisticated computational analyses that are enabled by GO, as discussed above.

### Concluding remarks and future directions

SGD strives to maintain a high-quality set of GO annotations that reflect the experimental literature to aid the efforts of the scientific community to generate new data and methods that facilitate the functional characterization of genes in *S. cerevisiae* and other organisms. To this end, SGD continues to review and update our oldest GO annotations as needed, based on current research. We also plan to replace all GO annotations derived solely from author statements with annotations supported by experimental results. In addition, future efforts will involve comparing computationally predicted GO annotations with manually curated ones to refine the manually curated set by identifying inaccurate or missing annotations. This comparison will also improve the accuracy of some computational prediction methods.

This review has focused on how *S. cerevisiae* GO annotations made by SGD have been used to analyze results from HTP experimental methods and predict functions of uncharacterized genes in *S. cerevisiae* and other organisms. However, GO annotations are also used to construct cellular pathways, construct protein interaction networks and build transcriptional regulatory networks to understand the budding yeast at the systems biology level [8,69]. Although GO annotations can provide a summary of the *S. cerevisiae* research literature, it is important for researchers to understand what types of data are represented by the annotations so they can use the information effectively and appropriately in their research (Box 3).

### Acknowledgements

We thank Maria Costanzo and Jodi Hirschman for their careful reading of the manuscript; Dianna Fisk, Julie Park and Rama Balakrishnan for their insightful comments on the illustrations; and the staff of SGD for their assistance with the literature search. We also thank three anonymous reviewers for their helpful comments to clarify the text. SGD is supported by the US National Human Genome Research Institute (NHGRI) (HG001315 to J.M.C., PI) and through the GO Consortium grant from NHGRI (HG002273 to J.M.C, co-PI).

## References

- 1 Goffeau, A. *et al.* (1996) Life with 6000 genes. *Science* 274, 563–567
- 2 Jones, G.M. *et al.* (2008) A systematic library for comprehensive overexpression screens in *Saccharomyces cerevisiae*. *Nat. Methods* 5, 239–241
- 3 Huh, W.K. *et al.* (2003) Global analysis of protein localization in budding yeast. *Nature* 425, 686–691
- 4 Winzler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906
- 5 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
- 6 Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21, 33–37
- 7 Bachi, A. and Bonaldi, T. (2008) Quantitative proteomics as a new piece of the systems biology puzzle. *J. Proteomics* 71, 357–367
- 8 Dolinski, K. and Botstein, D. (2005) Changing perspectives in yeast research nearly a decade after the genome sequence. *Genome Res.* 15, 1611–1619
- 9 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet* 25, 25–29
- 10 Rhee, S.Y. *et al.* (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515
- 11 Dwight, S.S. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* 30, 69–72
- 12 Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* 11, 1425–1433
- 13 Hong, E.L. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* 36, D577–D581
- 14 Costanzo, M.C. *et al.* (2009) New mutant phenotype data curation system in the *Saccharomyces* Genome Database. Database 2009, bap001
- 15 Breitkreutz, B.J. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* 36, D637–D640
- 16 Askree, S.H. *et al.* (2004) A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc. Natl. Acad. Sci. U. S. A.* 101, 8658–8663
- 17 Gathbonton, T. *et al.* (2006) Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.* 2, e35
- 18 Dubrana, K. *et al.* (2001) Turning telomeres off and on. *Curr. Opin. Cell Biol.* 13, 281–289
- 19 Rog, O. *et al.* (2005) The yeast VPS genes affect telomere length regulation. *Curr. Genet.* 47, 18–28
- 20 Park, H. and Hwang, Y.S. (2008) Genome-wide transcriptional responses to sulfite in *Saccharomyces cerevisiae*. *J. Microbiol.* 46, 542–548
- 21 Reinders, J. *et al.* (2006) Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *J. Proteome Res.* 5, 1543–1554
- 22 Sickmann, A. *et al.* (2003) The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13207–13212
- 23 Camon, E. *et al.* (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.* 32, D262–D266
- 24 Quevillon, E. *et al.* (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120
- 25 Huttenhower, C. and Troyanskaya, O.G. (2008) Assessing the functional structure of genomic data. *Bioinformatics* 24, i330–i338
- 26 Tian, W. *et al.* (2008) Combining guilt-by-association and guilt-by-profiling to predict *Saccharomyces cerevisiae* gene function. *Genome Biol.* 9 (Suppl 1), S7
- 27 Osborne, J.D. *et al.* (2007) Interpreting microarray results with gene ontology and MeSH. *Methods Mol. Biol.* 377, 223–242
- 28 Fillingham, J. *et al.* (2008) Chaperone control of the activity and specificity of the histone H3 acetyltransferase Rtt109. *Mol. Cell Biol.* 28, 4342–4353
- 29 Haarer, B. *et al.* (2007) Modeling complex genetic interactions in a simple eukaryotic genome: actin displays a rich spectrum of complex haploinsufficiencies. *Genes Dev.* 21, 148–159
- 30 Imbeault, D. *et al.* (2008) The Rtt106 histone chaperone is functionally linked to transcription elongation and is involved in the regulation of spurious transcription from cryptic promoters in yeast. *J. Biol. Chem.* 283, 27350–27354
- 31 Freimoser, F.M. *et al.* (2006) Systematic screening of polyphosphate (poly P) levels in yeast mutant cells reveals strong interdependence with primary metabolism. *Genome Biol.* 7, R109
- 32 Kramer, R.W. *et al.* (2007) Yeast functional genomic screens lead to identification of a role for a bacterial effector in innate immunity regulation. *PLoS Pathog.* 3, e21
- 33 Yu, L. *et al.* (2006) A survey of essential gene function in the yeast cell division cycle. *Mol. Biol. Cell* 17, 4736–4747
- 34 Nogami, S. *et al.* (2007) Genetic complexity and quantitative trait loci mapping of yeast morphological traits. *PLoS Genet.* 3, e31
- 35 Law, G.L. *et al.* (2005) The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. *Genome Biol.* 6, R111
- 36 Collins, S.R. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 6, 439–450
- 37 Yu, H. *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104–110
- 38 Colomina, N. *et al.* (2008) Whi3, a developmental regulator of budding yeast, binds a large set of mRNAs functionally related to the endoplasmic reticulum. *J. Biol. Chem.* 283, 28670–28679
- 39 Fleischer, T.C. *et al.* (2006) Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes Dev.* 20, 1294–1307
- 40 Boyle, E.I. *et al.* (2004) GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20, 3710–3715
- 41 Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587–3595
- 42 Georgiev, A. *et al.* (2007) Binding specificities of the GYF domains from two *Saccharomyces cerevisiae* paralogs. *Protein Eng. Des. Sel.* 20, 443–452
- 43 Friedberg, I. (2006) Automated protein function prediction – the genomic challenge. *Brief. Bioinform.* 7, 225–242
- 44 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
- 45 Tari, L. *et al.* (2009) Fuzzy c-means clustering with prior biological knowledge. *J. Biomed. Inform.* 42, 74–81
- 46 Brameier, M. and Wiuf, C. (2007) Co-clustering and visualization of gene expression data and gene ontology terms for *Saccharomyces cerevisiae* using self-organizing maps. *J. Biomed. Inform.* 40, 160–173
- 47 Nam, D. *et al.* (2006) ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics* 22, 2249–2253
- 48 Jiang, X. *et al.* (2008) Integration of relational and hierarchical network information for protein function prediction. *BMC Bioinformatics* 9, 350
- 49 Hughes, T.R. and Roth, F.P. (2008) A race through the maze of genomic evidence. *Genome Biol.* 9 (suppl. 1), S1
- 50 Troyanskaya, O.G. *et al.* (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci. U. S. A.* 100, 8348–8353
- 51 Chen, Y. and Xu, D. (2004) Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 32, 6414–6424
- 52 Lee, I. *et al.* (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. *PLoS One* 2, e988
- 53 Guan, Y. *et al.* (2008) Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol.* 9 (suppl. 1), S3
- 54 Cushion, M.T. *et al.* (2007) Transcriptome of *Pneumocystis carinii* during fulminate infection: carbohydrate metabolism and the concept of a compatible parasite. *PLoS One* 2, e423
- 55 Gattiker, A. *et al.* (2007) *Ashbya* Genome Database 3.0: a cross-species genome and transcriptome browser for yeast biologists. *BMC Genomics* 8, 9
- 56 Li, R. *et al.* (2004) Interaction of *Sclerotinia sclerotiorum* with a resistant *Brassica napus* cultivar: expressed sequence tag analysis identifies genes associated with fungal pathogenesis. *Fungal Genet. Biol.* 41, 735–753



- 57 Pirooznia, M. *et al.* (2007) Cloning, analysis and functional annotation of expressed sequence tags from the Earthworm *Eisenia fetida*. *BMC Bioinformatics* 8 (suppl. 7), S7
- 58 Arnaud, M.B. *et al.* Gene Ontology and the fungal pathogen *Candida albicans*. *Trends Microbiol*
- 59 Martin, D.M. *et al.* (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics* 5, 178
- 60 Biswas, S. *et al.* (2008) Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics* 9, 244
- 61 Borneman, A.R. *et al.* (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317, 815–819
- 62 Tuch, B.B. *et al.* (2008) Evolution of eukaryotic transcription circuits. *Science* 319, 1797–1799
- 63 Snitkin, E.S. *et al.* (2006) Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics* 7, 420
- 64 Camon, E.B. *et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 6 (suppl. 1), S17
- 65 Krallinger, M. *et al.* (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* 9 (suppl. 2), S8
- 66 Crangle, C.E. *et al.* (2007) Mining experimental evidence of molecular function claims from the literature. *Bioinformatics* 23, 3232–3240
- 67 Peña-Castillo, L. and Hughes, T.R. (2007) Why are there still over 1000 uncharacterized yeast genes? *Genetics* 176, 7–14
- 68 Hughes, T.R. *et al.* (2004) The promise of functional genomics: completing the encyclopedia of a cell. *Curr. Opin. Microbiol.* 7, 546–554
- 69 Hartwell, L.H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402, C47–C52
- 70 Miller, B.G. *et al.* (2000) Anatomy of a proficient enzyme: the structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc. Natl. Acad. Sci. U. S. A.* 97, 2011–2016
- 71 Muller-Dieckmann, H.J. and Schulz, G.E. (1995) Substrate specificity and assembly of the catalytic center derived from two structures of ligated uridylylate kinase. *J. Mol. Biol.* 246, 522–530
- 72 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
- 73 Kavanaugh, L.A. and Dietrich, F.S. (2009) Non-coding RNA prediction and verification in *Saccharomyces cerevisiae*. *PLoS Genet.* 5, e1000321
- 74 McCutcheon, J.P. and Eddy, S.R. (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucleic Acids Res.* 31, 4119–4128

## Celebrating Darwin: Evolution of Hosts, Microbes and Parasites

To commemorate the 200<sup>th</sup> anniversary of Charles Darwin's birthday (12<sup>th</sup> February, 1809), *Trends in Microbiology* is featuring several articles with evolutionary themes in the course of 2009, along with *Trends in Parasitology* and *Cell Host & Microbe*.

Although it is commonly assumed that Darwin had nothing to say about microbes, he did in fact say quite a lot. However, Darwin's impact on microbiological thinking of the late nineteenth century was negligible. These topics are the focus of an Opinion article by Maureen O'Malley, entitled 'What *did* Darwin say about microbes, and how did microbiology respond?', to be published in the next issue of *Trends in Microbiology* (August 2009).

See also 'Drug-resistance mechanisms in helminths: is it survival of the fittest?' by Mary W. Davey and colleagues in the July issue of *Trends in Parasitology*. Changes in drug targets, transport proteins and detoxification systems can make helminths more resistant to anthelmintics.

All the articles in the series are collected in the following webpage:  
<http://www.cell.com/trends/microbiology/Darwin>