D442–D445 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj117

Genome Snapshot: a new resource at the *Saccharomyces* Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome

Jodi E. Hirschman, Rama Balakrishnan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Eurie L. Hong, Michael S. Livstone[1], Robert Nash, Julie Park, Rose Oughtred[1], Marek Skrzypek, Barry Starr, Chandra L. Theesfeld, Jennifer Williams, Rey Andrada, Gail Binkley, Qing Dong, Christopher Lane, Stuart Miyasato, Anand Sethuraman, Mark Schroeder[1], Mayank K. Thanawala, Shuai Weng, Kara Dolinski[1], David Botstein[1] and J. Michael Cherry*

Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA and
[1]Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University,
Washington Road, Princeton, NJ 08544, USA

Received September 13, 2005; Revised and Accepted October 18, 2005

ABSTRACT

Sequencing and annotation of the entire *Saccharomyces cerevisiae* genome has made it possible to gain a genome-wide perspective on yeast genes and gene products. To make this information available on an ongoing basis, the *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org/) has created the Genome Snapshot (http://db.yeastgenome.org/cgi-bin/genomeSnapShot.pl). The Genome Snapshot summarizes the current state of knowledge about the genes and chromosomal features of *S.cerevisiae*. The information is organized into two categories: (i) number of each type of chromosomal feature annotated in the genome and (ii) number and distribution of genes annotated to Gene Ontology terms. Detailed lists are accessible through SGD's Advanced Search tool (http://db.yeastgenome.org/cgi-bin/search/featureSearch), and all the data presented on this page are available from the SGD ftp site (ftp://ftp.yeastgenome.org/yeast/).

INTRODUCTION

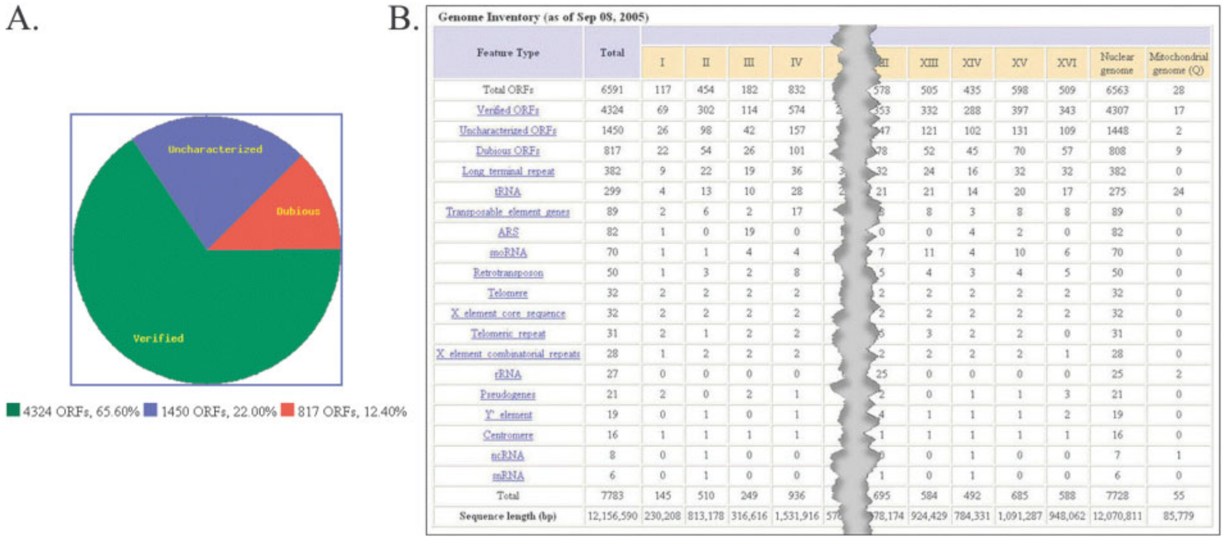
The knowledge gained from the sequencing and extensive annotation of the *Saccharomyces cerevisiae* genome over the past decade has enabled researchers to take a genome-wide view of yeast genes and proteins. Because it is now possible to present information about the *S.cerevisiae* genome in its entirety, basic questions about protein function and cellular processes can be studied on a larger scale by searching for and studying every gene involved, rather than just one or a few genes. Moreover, researchers new to studies in *S.cerevisiae* can begin their inquiry at the genome level and be led to individual genes, rather than the more classical approach of starting with a phenotype and determining the genes involved.

To provide a composite overview of the genes and proteins of *S.cerevisiae*, the *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) has created the Genome Snapshot (http://db.yeastgenome.org/cgi-bin/genomeSnapShot.pl). The Genome Snapshot presents a graphical summary of this genomic information, which is also available through SGD's ftp site (ftp://ftp.yeastgenome.org/yeast/) and Advanced Search form (http://db.yeastgenome.org/cgi-bin/search/featureSearch). The current status of the entire *S.cerevisiae* genome is displayed on a single page and organized into two categories: (i) an inventory of the different types of chromosomal features found in *S.cerevisiae* and (ii) a summary of the distribution of gene product annotations among Gene Ontology (GO) terms, which reflect the molecular function, biological role and cellular localization of the gene products (1).

Organization of the Genome Snapshot

The Genome Snapshot includes information on genomic features as well as summaries of GO annotations. The data are organized into four sections on the page.

*To whom correspondence should be addressed. Tel: +1 650 723 7541; Fax: +1 650 725 1534; Email: cherry@genome.stanford.edu



Genome Inventory (as of Sep 08, 2005)

Feature Type	Total	I	II	III	IV		XII	XIII	XIV	XV	XVI	Nuclear genome	Mitochondrial genome (Q)
Total ORFs	6591	117	454	182	832		578	505	435	598	509	6563	28
Verified ORFs	4324	69	302	114	574	2	353	332	288	397	343	4307	17
Uncharacterized ORFs	1450	26	98	42	157		47	121	102	131	109	1448	2
Dubious ORFs	817	22	54	26	101		78	52	45	70	57	808	9
Long terminal repeat	382	9	22	19	36	3	32	24	16	32	32	382	0
tRNA	299	4	13	10	28	2	21	21	14	20	17	275	24
Transposable element genes	89	2	6	2	17		8	8	3	8	8	89	0
ARS	82	1	0	19	0		0	0	4	2	0	82	0
snoRNA	70	1	1	4	4		7	11	4	10	6	70	0
Retrotransposon	50	1	3	2	8		5	4	3	4	5	50	0
Telomere	32	2	2	2	2		2	2	2	2	2	32	0
X element core sequence	32	2	2	2	2		2	2	2	2	2	32	0
Telomeric repeat	31	2	1	2	2		5	3	2	2	0	31	0
X element combinatorial repeats	28	1	2	2	2		2	2	2	2	1	28	0
rRNA	27	0	0	0	0		25	0	0	0	0	25	2
Pseudogenes	21	2	0	2	1		2	0	1	1	3	21	0
Y' element	19	0	1	0	1		4	1	1	1	2	19	0
Centromere	16	1	1	1	1		1	1	1	1	1	16	0
ncRNA	8	0	1	0	0		0	0	1	0	0	7	1
snRNA	6	0	1	0	0		1	0	1	0	0	6	0
Total	7783	145	510	249	936		695	584	492	685	588	7728	55
Sequence length (bp)	12,156,590	230,208	813,178	316,616	1,531,916	57	78,174	924,429	784,331	1,091,287	948,062	12,070,811	85,779

Figure 1. Chromosomal features annotated in the *S.cerevisiae* genome at SGD. (**A**) Graphical View of Protein Coding Genes (http://db.yeastgenome.org/cgi-bin/genomeSnapshot.pl#pieChart). ORFs are classified by SGD as 'Verified', 'Uncharacterized' and 'Dubious'. 'Verified' ORFs are those for which experimental evidence demonstrates that a gene product is produced in *S.cerevisiae*. 'Uncharacterized' ORFs have orthologs in at least one other species and are likely to encode proteins although experimental proof has not yet been published. 'Dubious' ORFs are those unlikely to encode a protein because they are not conserved in closely related *Saccharomyces* species, and because no data exist demonstrating that a protein is produced. (**B**) The Genome Inventory (http://db.yeastgenome.org/cgi-bin/genomeSnapshot.pl#genomeInventory). A total count of each feature type in the genome as well as a count of each feature type on each chromosome is displayed in this table. (Note that data for chromosomes IV through XIV are not shown in the figure.) Definitions for each feature type can be found in SGD's Glossary. Clicking on any feature type initiates a search that uses SGD's Advanced Search tool (http://db.yeastgenome.org/cgi-bin/search/featureSearch) to find all the features in SGD of that type. This table also lists the current length of each chromosome in base pair.

Graphical view of protein-coding genes

The total number of biologically significant open reading frames (ORFs) in *S.cerevisiae* has been the subject of debate since the genome sequence was finished, and estimates continue to change as new experimental evidence becomes available (2–5). In August 2003, SGD curators began classifying ORFs into three categories, 'Verified', 'Uncharacterized' and 'Dubious', according to the degree of certainty that they actually encode proteins. The assignment to one of these categories is based on manual review of experimental evidence and sequence conservation.

Dubious ORFs are not conserved in closely related *Saccharomyces* species, and no experimental evidence currently exists that a gene product is produced. Many, but not all, Dubious ORFs overlap other ORFs. Uncharacterized ORFs have orthologs in at least one other species, suggesting that they are bona fide protein-coding genes, but to date there is no experimental evidence to support this. Verified ORFs are those for which experimental evidence demonstrates that a gene product is produced in *S.cerevisiae*.

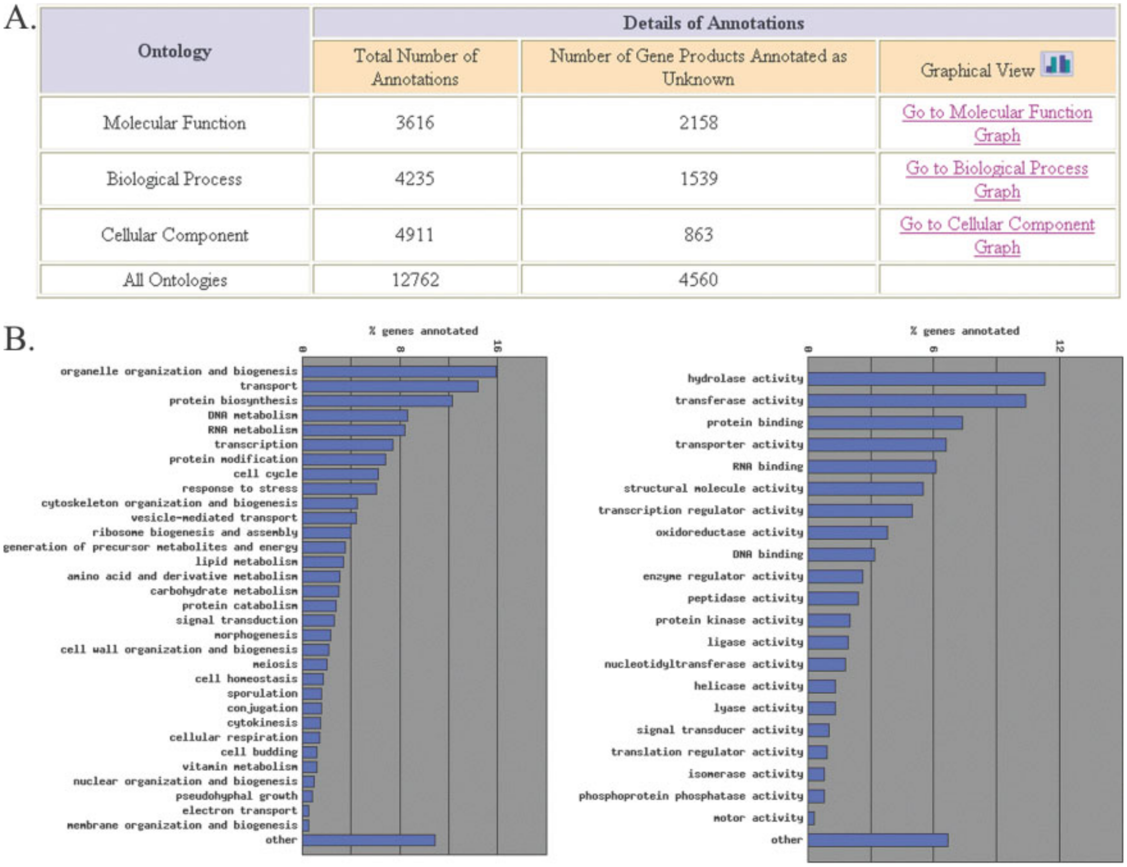
It is important to note that these classifications are meant to provide a current best estimate of the coding capacity of the genome. The classifications are fluid and are updated continually as new experimental results become available, during the normal process of literature curation; it is typical for a designation to change from Uncharacterized to Verified as more information is published. The color-coded pie chart at the top of the Genome Snapshot summarizes the current classification status of all the ORFs in the yeast genome (Figure 1A). Over the period from August 2003 through June 2005, the number of Verified ORFs increased by 238, from 4065 to 4303. The number of Dubious ORFs stayed relatively constant over this time period, decreasing from 820 to 818.

Genome Inventory

The Genome Inventory is in tabular form and displays several types of information for each chromosome and for the genome as a whole (Figure 1B). Although the complete sequence for *S.cerevisiae* was first published in 1996 (6), annotation of the genome remains an ongoing process (2–4,7). In the past 5 years, the yeast genome catalog has expanded to include 426 new protein-coding genes and 255 other non-protein-coding features (centromeres, ARS sequences, tRNA genes, rRNA genes, etc.). The Genome Inventory lists all types of chromosomal features currently annotated in SGD and provides a count for each type. Clicking on any feature type initiates a search that uses SGD's Advanced Search tool (http://db.yeastgenome.org/cgi-bin/search/featureSearch) to find all the features in SGD of that type. This table also lists the current length of each chromosome in base pairs. Information for all chromosomal features is available in the SGD_features.tab file at SGD's ftp site (ftp://ftp.yeastgenome.org/yeast/chromosomal_feature/).

Summary of Gene Ontology (GO) annotations

A tabular overview summarizes the status of GO annotation at SGD (Figure 2A). The first column 'Total Number of Annotations' shows the total number of *S.cerevisiae* gene products (protein and RNA gene products) currently annotated to each of the three GO ontologies: Biological Process, Molecular Function and Cellular Component (1,8) (see http://www.geneontology.org/ for detailed information about the GO project). The 'Total Number of Annotations'



A.

Ontology	Details of Annotations		
	Total Number of Annotations	Number of Gene Products Annotated as Unknown	Graphical View
Molecular Function	3616	2158	Go to Molecular Function Graph
Biological Process	4235	1539	Go to Biological Process Graph
Cellular Component	4911	863	Go to Cellular Component Graph
All Ontologies	12762	4560	

Figure 2. Gene Ontology annotation at SGD. (**A**) Summary of GO annotations (http://db.yeastgenome.org/cgi-bin/genomeSnapshot.pl#goAnnotations). The column, 'Total Number of Annotations', refers to the total number of *S.cerevisiae* gene products (protein and RNA gene products) currently annotated to one or more terms (other than 'unknown') in each of the three GO ontologies: Biological Process, Molecular Function and Cellular Component. The number of gene products annotated to 'unknown' for any ontology is provided in the second column. The third column offers links to the graphs shown in Figure 2B. (**B**) Distribution of gene products by process, function and component (http://db.yeastgenome.org/cgi-bin/genomeSnapshot.pl#barCharts). Shown are percentages of gene products annotated to a specific term that maps up the ontology to a yeast GO Slim term. The yeast GO Slim is a high-level subset of GO terms that allows grouping of genes into broad categories (see text for details). Annotations to 'unknown' are excluded. (Note that the Cellular Component graph is not shown.)

column does not include annotations to the three terms that represent lack of knowledge at this time, 'molecular function unknown', 'biological process unknown' or 'cellular component unknown'; these numbers are reported in the 'Number of Gene Products Annotated as Unknown' column. Note that an 'unknown' annotation indicates that there are no published data available for that gene product, in that category. Not included in these counts, for either column, are GO annotations made for ORFs classified as 'Dubious', or for features of the type 'Pseudogene', 'Not in systematic sequence of S288C', or 'Not physically mapped'. GO annotations for *S.cerevisiae* gene products are included in the gene_association.sgd file available at SGD's ftp site (ftp://ftp.yeastgenome.org/yeast/literature_curation/).

Distribution of gene products by process, function and component

This last section of the Genome Snapshot displays distributions, presented as bar graphs, of gene products annotated to GO terms other than 'unknown' (Figure 2B; Cellular Component graph not shown). Instead of providing the distribution based on the specific term to which the gene product was annotated, which would make the graphs quite large and difficult to interpret, gene products are more broadly grouped using yeast GO Slim terms. The yeast GO Slim is a subset of GO terms that apply to broad categories of cell processes, functions or components (and as such are placed at a high level in their respective ontologies). Yeast GO Slim terms allow grouping of genes into categories, some of which are very general (e.g. 'DNA metabolism' and 'nucleus') and others that are tailored to yeast biology (e.g. 'sporulation' and 'bud'). For example, DNA replication is a high-level term in the Biological Process ontology that has a number of more granular child terms. Genes annotated to these child terms are all involved in the process of DNA replication and thus are grouped into the GO Slim category of that name. The GO Slim terms for each of the GOs (Molecular Function, Biological Process or Cellular Component) are listed on separate graphs, along with the percentage of *S.cerevisiae* gene products (proteins and RNAs) annotated to a specific term that maps up the ontology to that GO Slim term. Since some gene products are annotated to GO terms that map to more than one GO Slim term, there is some redundancy in these distributions. Lists of ORFs annotated to these GO terms may be accessed using the GO Slim Mapper (http://db.yeastgenome.org/cgi-bin/GO/goTermMapper) or the Advanced Search tool (http://db.yeastgenome.org/cgi-bin/search/featureSearch). An alternative processing of the same data, with some redundancies removed, is available from the go_slim_mapping.tab file on the SGD ftp site (ftp://ftp.yeastgenome.org/yeast/literature_curation/).

SUMMARY

The wealth of information describing the genes and proteins of *S.cerevisiae* has both necessitated and made possible the creation of SGD's new Genome Snapshot, a constantly updated overview of the genome. This page answers a range of questions, from the most basic 'How many genes in *S.cerevisiae* encode proteins?' to more complex questions regarding which and how many genes are involved in particular cellular processes. By making accessible lists of Uncharacterized ORFs, it points researchers to some of the many intriguing questions that remain to be answered about the yeast genome and biological processes. Finally, Genome Snapshot documents the characterization of the genome, both by tracking annotation of ORFs to GO terms and by tracking increases in the number of Verified ORFs. SGD is committed to continually expanding its resources to increase the ease of access to information about *S.cerevisiae* and welcomes all comments from the research community toward this end. Please send any suggestions about the Genome Snapshot or any other tool at SGD to: yeast-curator@genome.stanford.edu.

ACKNOWLEDGEMENT

The SGD project is supported by the NHGRI grant 5P41HG001315-11 (J.M.C.). The SGD annotations to the Gene Ontology are partially supported by NHGRI grant 5P41HG002273-05 (J. Blake, The Jackson Laboratory, Bar Harbor, Maine). Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
2. Brachat,S., Dietrich,F.S., Voegeli,S., Zhang,Z., Stuart,L., Lerch,A., Gates,K., Gaffney,T. and Philippsen,P. (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. *Genome Biol*., **4**, R45.
3. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
4. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
5. Blandin,G., Durrens,P., Tekaia,F., Aigle,M., Bolotin-Fukuhara,M., Bon,E., Casaregola,S., de Montigny,J., Gaillardin,C., Lepingle,A. *et al*. (2000) Genomic exploration of the hemiascomycetous yeasts: 4. The genome of *Saccharomyces cerevisiae* revisited. *FEBS Lett*., **487**, 31–36.
6. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al*. (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.
7. Oshiro,G., Wodicka,L.M., Washburn,M.P., Yates,J.R.,III, Lockhart,D.J. and Winzeler,E.A. (2002) Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res*., **12**, 1210–1220.
8. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al*. (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*., **30**, 69–72.