[19] *Saccharomyces* Genome Database

By LAURIE ISSEL-TARVER, KAREN R. CHRISTIE, KARA DOLINSKI, REY ANDRADA, RAMA BALAKRISHNAN, CATHERINE A. BALL, GAIL BINKLEY, STAN DONG, SELINA S. DWIGHT, DIANNA G. FISK, MIDORI HARRIS, MARK SCHROEDER, ANAND SETHURAMAN, KANE TSE, SHUAI WENG, DAVID BOTSTEIN, and J. MICHAEL CHERRY

Introduction

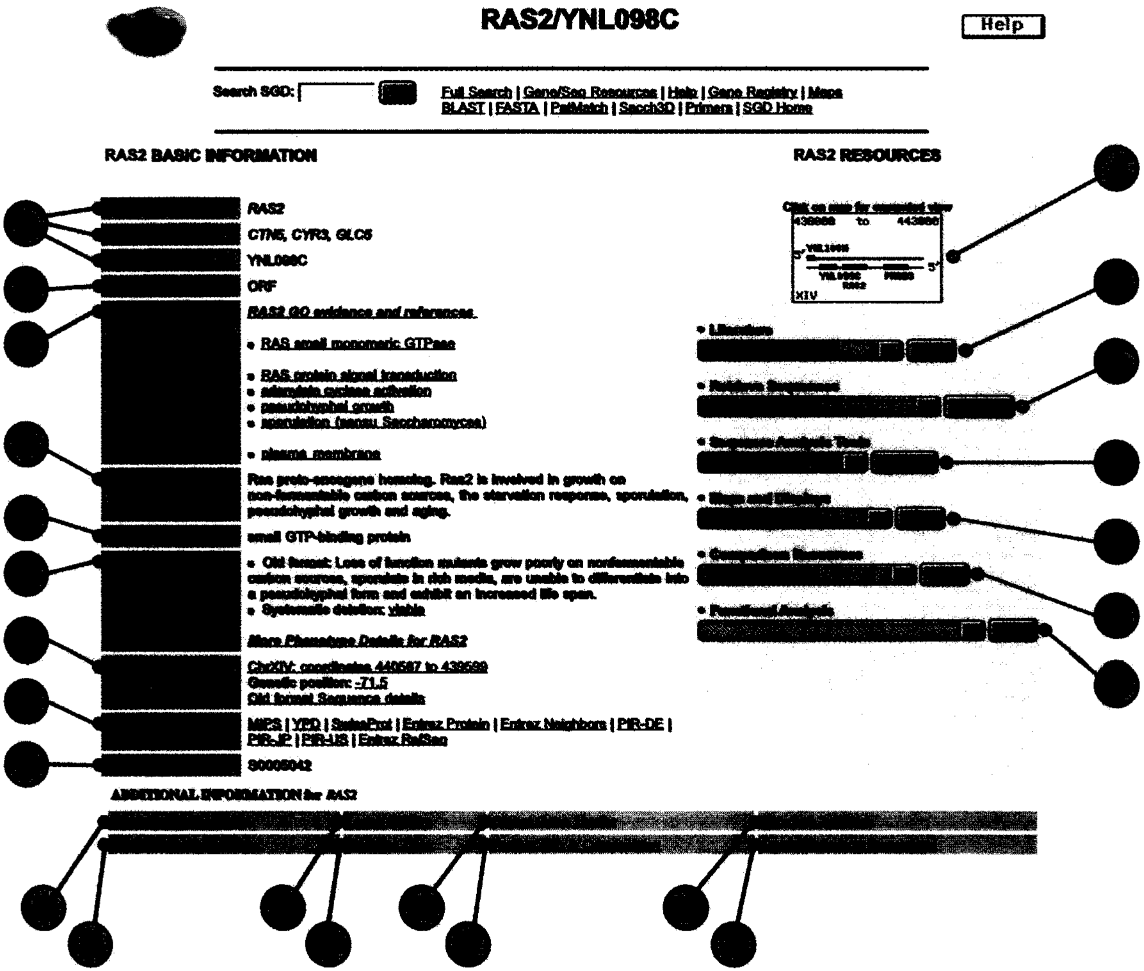
The goal of the *Saccharomyces* Genome Database (SGD) is to provide information about the genome of this yeast, the genes it encodes, and their biological functions. The genome sequence of *S. cerevisiae* provides the structure around which information in SGD is organized; value is added to the sequence by careful biological annotation drawn from a number of sources. SGD curates and stores information about budding yeast DNA and protein sequences, genetics, cell biology, and the associated community of researchers. SGD also provides search and analysis tools designed to help researchers mine the data for pieces or patterns of biological information relevant to their interests. A continuing challenge for the staff of SGD is to present up-to-date information about yeast genes in a format that is intuitive and useful to biomedical researchers, while responding to the needs of this community by providing resources and tools for exploring the data in new ways.

This chapter describes the organization of SGD, the sources of the data stored in SGD, some methods for retrieving information from the database, connections SGD has with outside databases and non-yeast research communities, and SGD's repository of yeast community information. This is not a complete overview of the database, as SGD contains hundreds of tools, including specialized sequence analysis programs, and new tools are always being added. As of this writing, several new tools and Web interfaces for DNA and protein sequences are being developed, along with enhanced database navigation methods. New tools for comparison of the *S. cerevisiae* genome with other fungal genomes will also soon be available. To explore the resources currently available at SGD, visit the Web site at http://genome-www.stanford.edu/Saccharomyces/.

Locus-Centered Organization of SGD

The systematic sequencing project[1] defined the set of genes and non-open reading frame (ORF) features (centromeres, tRNAs, etc.) around which information in SGD is organized. SGD's Locus pages display basic information about a locus and provide links to further information and resources (Fig. 1). The basic information

0076-6879/02 $35.00



RAS2/YNL098C

Search SGD:

Full Search | Gene/Seq Resources | Help | Gene Registry | Maps
BLAST | FASTA | PatMatch | Sacch3D | Primers | SGD Home

RAS2 BASIC INFORMATION

RAS2

CTN5, CYR3, GLC5

YNL098C

ORF

RAS2 GO evidence and references.

- RAS small monomeric GTPase

- RAS protein signal transduction
- adenylate cyclase activation
- pseudohyphal growth
- sporulation (sensu Saccharomyces)

- plasma membrane

Ras proto-oncogene homolog. Ras2 is involved in growth on non-fermentable carbon sources, the starvation response, sporulation, pseudohyphal growth and aging.

small GTP-binding protein

- Old format: Loss of function mutants grow poorly on nonfermentable carbon sources, sporulate in rich media, are unable to differentiate into a pseudohyphal form and exhibit an increased life span.
- Systematic deletion: viable

More Phenotype Details for RAS2

ChrXIV: coordinates 440587 to 439599
Genetic position: -71.5
Old format Sequence details

MIPS | YPD | SwissProt | Entrez Protein | Entrez Neighbors | PIR-DE |
PIR-JP | PIR-US | Entrez RefSeq

S0005042

ADDITIONAL INFORMATION for *RAS2*

RAS2 RESOURCES

Click on map for expanded view

- Literature
- Retrieve Sequences
- Sequence Analysis Tools
- Maps and Displays
- Comparison Resources
- Functional Analysis

FIG. 1. The *RAS2* Locus Page. (A) The names and aliases for a locus are listed, with an indication of whether a name is standard or reserved. (B) Feature Type indicates whether a locus is an ORF, TyORF, LTR, tRNA, RNA gene, or centromere. (C) Gene Ontology terms assigned to a gene are listed, and each term links to a page showing all yeast genes annotated to that term (see Fig. 5). There is also a link from the locus page to a table that lists the references that were used for assigning each GO term, and the type of evidence from that reference (direct assay, genetic interaction, etc.) that supported the assignment. (D) The Description field lists general information about the gene. (E) The Gene Product field lists the protein or gene product that the gene encodes. (F) Mutant phenotypes of a gene are listed. Clicking on a phenotype leads to a table of other yeast genes that share the phenotype. (G) The Position field lists the chromosomal coordinates of the gene as well as its genetic position. The chromosomal coordinates link to the chromosomal features map, and the genetic position links to the combined physical and genetic map. (H) External links provide additional sources of information about a gene. These links take users outside of SGD. (I) SGDIDs are unique database identifiers for *S. cerevisiae* loci. (J) The mini ORF map shows the chromosomal features located near the locus. Loci encoded on the Watson and Crick strands are shown separately and color coded according to strand and feature type. (K) The pull-down Literature menu gives users access to SGD-curated items (the Literature Guide and Gene Summary) as well as an external link to a PubMed search for that gene name. (L) Users may specify a sequence to retrieve, choosing from options that include DNA (with introns), coding sequence, ORF translation, and 6-frame translation. Other options include custom retrieval specified by the user, retrieving all associated sequences, and retrieving lists of restriction fragment sizes. (M) Sequence Analysis Tools include BLASTN, BLASTP, FASTA nt, and FASTA aa analyses. This menu also allows retrieval of a restriction map of the sequence or the Design Primers tool. (N) Maps and Displays retrieves several locus-centered maps or tables, including a chromosomal features map or table, the Physical & Genetic Map, Physical Map, and the Physical/Genetic Map Ratios. (O) Comparison Resources allow users to retrieve information about genes in other organisms that show significant similarity to the yeast gene described on this locus page. (P) Users can retrieve gene-specific results from a fast-growing list of published functional analysis studies. (Q) The Gene Summary Paragraph is a summary of published biological information for a gene and its product that is designed to familiarize both yeast and non-yeast researchers with the general facts and important subtleties regarding a locus. (R) Expression Connection allows users to search the results of several microarray studies for gene expression data. (S) Locus History provides notes about the locus, which may alert the user to contradictory information in the literature or to potentially confusing gene names. For reserved gene names, the Locus History includes the reservation date and expiration date. (T) Two-point genetic mapping data tables are available for many loci. (U) Global Gene Hunter searches several online databases for locus-specific information. (V) The Protein Info and Composition page provides a great deal of information about protein sequence, chemistry, and more. (W) Function Junction searches functional analysis project sites for locus-specific results. (X) Gene/Sequence Resources allows users to access biological information, table/map displays, and sequence analysis and retrieval options for a locus.

about a locus includes the standard gene name, the systematic ORF name, and any aliases; Gene Ontology[2,3] annotations describing the gene product's molecular functions, biological processes, and cellular components; additional brief information about the locus and gene product; phenotype information; the position of the locus on the genetic and physical maps; and links to information about the locus in other databases. The assembled resources available for a locus include links to the scientific literature (SGD's Literature Guide, a PubMed search for that locus name, and a curator-composed Gene Summary if available); several options for locus-specific DNA or protein sequence retrieval; sequence analysis tools; genetic and physical map displays; genome comparison resources for finding homologs of the locus in other organisms; and a large set of functional analysis links which connect the user to data for that locus from several published functional analysis studies. Also present on the locus page is a clickable mini-ORF map, showing the locus and any adjacent chromosomal features. Arrayed along the bottom of the locus page is a third series of links: the locus' Gene Summary provides a brief, in-depth description of the gene and its gene product; Locus History contains gene nomenclature information, a list of researchers associated with the locus, a history of updates to the locus sequence or coordinates, and other pertinent information; Global Gene Hunter allows users to locate information about the locus from several different online databases; Function Junction searches for results specific to that locus from several genome-wide functional analysis projects; Expression Connection simultaneously searches the results of several published microarray studies for gene expression results for the locus; Mapping Data provides genetic and physical map information for the locus; Protein Information and Composition contains data generously provided by the YPD[4]; Researchers provides a list of Colleagues associated with the locus; and Gene/Sequence Resources gives users many options for accessing information about a locus' sequence, map position, biology, and more. As research into yeast genes provides new and different types of data, the locus information will evolve to reflect those changes.

Sources of Information in SGD

SGD obtains biological data from many different sources. Curators, who are Ph.D. biologists, are trained to maintain and validate information within the

[1] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver, *Science* **274,** 546, 563 (1996).

[2] Gene Ontology Consortium, *Genome Res.* **11,** 1425 (2001).

[3] Gene Ontology Consortium, *Nat. Genet.* **25,** 25 (2000).

[4] M. C. Costanzo, J. D. Hogan, M. E. Cusick, B. P. Davis, A. M. Fancher, P. E. Hodges, P. Kondu, C. Lengieza, J. E. Lew-Smith, C. Lingner, K. J. Roberg-Perez, M. Tillberg, J. E. Brooks, and J. I. Garrels, *Nucleic Acids Res.* **28,** 73 (2000).

database, process the information to check for accuracy and consistency, and then post the data for public view. The majority of the data in SGD are freely available for download via ftp (ftp://genome-ftp.stanford.edu/pub/yeast/). SGD information sources include the yeast genome systematic sequencing project, published literature, individual users, genome-wide functional studies, and scientific databases.

Systematic Sequencing Project

The genome sequence of *S. cerevisiae,* completed in 1996,[1] is at the heart of SGD. Curators keep the genome sequence up-to-date by incorporating individual corrections made by researchers as reported in the literature or communicated directly to SGD curators, and by incorporating the more extensive changes that result from larger-scale resequencing efforts. These sequence updates are performed in collaboration with MIPS,[5] and in consultation with the original systematic sequencing groups. A compendium of the changes made to the original genome sequence is available in sequence update tables at the SGD Web site (http://genome-www.stanford.edu/Saccharomyces/sequenceupdates.html). The systematic sequencing effort also defined a set of known and predicted open reading frames (ORFs); curators at SGD revise that set of ORFs by incorporating previously unpredicted (often small) ORFs that researchers have experimentally identified and designating as "questionable" those ORFs that research indicates are unlikely to be transcribed genes. SGD strives to provide laboratory researchers and computational biologists with an accurate, up-to-date representation of the yeast genome. Generally, this represents a synthesis of the systematic sequencing entries in GenBank. However, at times the SGD-provided sequence is ahead of the international DNA databanks.

Published Literature

The biological information about yeast genes found in SGD is largely derived from the published literature. Weekly searches of PubMed incorporate newly published papers describing named *S. cerevisiae* genes into SGD and are critical to keeping the database current. SGD adds value to this resource by assigning papers associated with a gene to categories according to the topics that each paper covers, thus creating a comprehensive Literature Guide (previously called Gene Info) for each gene. This Literature Guide helps direct researchers to particularly relevant publications. Curators also use published literature to choose Gene Ontology[2,3] terms to describe gene products and to compose Gene Summary paragraphs (see below). Besides acting as a valuable resource for yeast biologists, SGD's literature services help researchers from outside the community take advantage of the wealth

[5] H. W. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil, *Nucleic Acids Res.* **28,** 37 (2000).

of experimental evidence available for *S. cerevisiae,* by allowing them to quickly focus on their particular area of interest.

Individual Users

Much of the data contained in SGD has come directly from individual users. The SGD curators daily incorporate information sent by researchers to update and improve all aspects of the database. There are two main routes by which users communicate information to SGD: Web-based forms and email to the curators.

Web-Based Forms. There are forms on SGD's Web site that allow users to submit information directly to curators. One of the most commonly used is the Gene Registry form (http://genome-www4.stanford.edu/cgi-bin/SGD/registry/geneRegistry). In 1994 Robert Mortimer transferred the task of maintaining the nomenclature of *S. cerevisiae* genes, the Gene Name Registry, to SGD. Yeast researchers can reserve a gene name or register a published gene name by submitting a completed Gene Registry form. This form accepts an explanation of the acronym, the identity of the corresponding open reading frame (ORF), any aliases for that gene, a description of the encoded gene product, any phenotypes associated with mutations in the gene, references associated with the gene, and any other comments the researcher would like listed with the gene entry in SGD. Curators process the form first by searching several databases to ensure that the proposed gene name has not been previously used for another *S. cerevisiae* gene, and then by reviewing the submitted data to ensure that the resulting database entries will be clear and useful to all users.

The Colleague Submission/Update form (http://genome-www4.stanford.edu/cgibin/SGD/colleague/colleagueSearch) is another frequently used form on SGD's Web site. This form allows users to create or update their contact information in SGD, thus providing other members of the research community with telephone and mail information, Web page addresses, a description of research interests, lists of co-workers, and other relevant data.

E-mail to Curators. SGD encourages the submission of information via e-mail to the yeast curator address (*yeast-curator@genome.stanford.edu*). Users enrich SGD by providing updated information about genes and sequences, and by making suggestions about the content and features provided. User scrutiny of database contents benefits the community by ensuring that SGD remains accurate and up-to-date. Questions submitted by researchers about retrieving particular kinds of data from SGD and other resources have been an important source of inspiration for the development of new tools. Pattern matching and sequence retrieval tools have been created, and existing tools have been improved, in response to user requests.

Genome-Wide Studies

Genome-wide analyses of genes and gene products have created enormous datasets rich with information about individual genes. As the number of large-scale experiments has grown, so has the need to create intuitive and useful methods for accessing and analyzing the results of those experiments. Consequently, one of SGD's high priorities is to find ways to assimilate increasing numbers of datasets from gene expression, systematic deletion, and functional analysis projects, to make them readily available and easy to navigate. To present the data from genome-wide analyses SGD either incorporates the data into its own database or provides users with gene-specific links to external databases.

Other Databases

Other scientific databases are an important source of information for SGD. These databases include those that provide specifics about yeast genes (such as descriptions of gene products, sequences, intron/exon boundaries, promoters, and tRNAs), those that provide a framework for classifying gene products, and those that provide information about homologs of yeast genes in other organisms. Some of the external databases that SGD relies on most heavily include PubMed, GenBank, YPD, MIPS, and SwissProt.[4–8]

A database collaboration that has contributed greatly to the biological information available in SGD is the Gene Ontology (GO).[2,3] SGD is one of the founding members of GO, a collaboration among several model organism databases whose objective is to produce shared, structured vocabularies for the biological description of gene products in any organism. Consortium members are developing three independent networks of terms, collectively called Gene Ontology, in which biological concepts and the relationships between them are specified. One ontology describes the molecular functions a gene product carries out, another describes the biological processes in which a gene product is involved, and the third describes the cellular components where a gene product is found. The short phrases that describe a gene product's function, process, and cellular component are called GO terms. In addition to developing the ontologies, the member databases of the GO Consortium are using GO terms to annotate gene products in their respective model organisms and contributing the annotations to a shared central resource.

Accessing Biological Information at SGD

There are several possible starting points and paths a researcher might use to get biological information about yeast genes. Options for identifying genes of interest and discovering connections between genes include retrieving information such as sequences, expression patterns, phenotypes, associated key words, or Gene

[6] A. Bairoch and R. Apweiler, *Nucleic Acids Res.* **28,** 45 (2000).

[7] D. L. Wheeler, C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp, *Nucleic Acids Res.* **28,** 10 (2000).

[8] K. Dolinski, C. A. Ball, S. A. Chervitz, S. S. Dwight, M. A. Harris, S. Roberts, T. Roe, J. M. Cherry, and D. Botstein, *Yeast* **14,** 1453 (1998).

Ontology annotations relevant to the user's interests. Once the appropriate genes have been identified, the user's options for exploring the genes' biology begin at the locus pages. In particular the user should read Gene Summary paragraphs, explore the literature through the Literature Guide, and browse the hyperlinks to other databases provided as external links. For navigating easily through the database, the SGD Search box at the top of almost every page allows users to do a quick search using query terms such as gene names, Colleague names, GO terms, gene product names, and more.

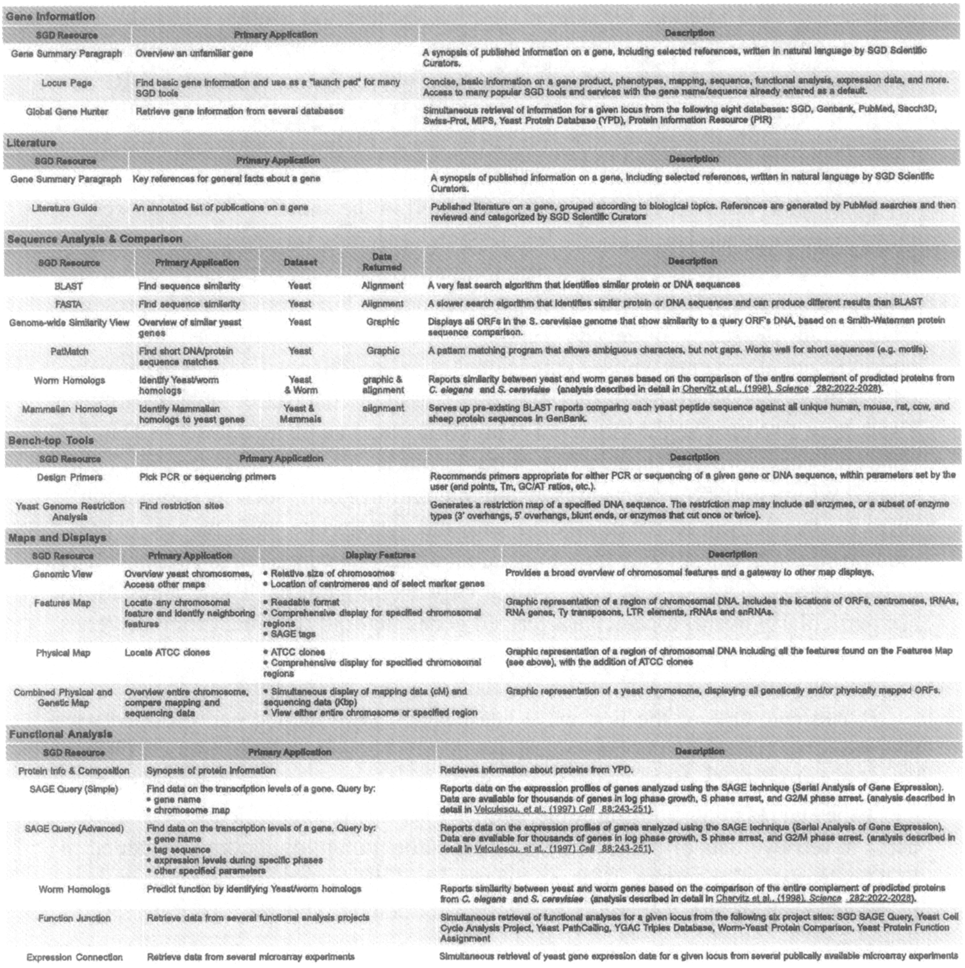
Starting Points

To start learning about the tools provided by SGD we recommend the Resource Guide, available at the URL http://genome-www.stanford.edu/Saccharomyces/resource_guide.html (Fig. 2). This provides a listing of resources for investigating gene information, the scientific literature, sequence analysis options, bench-top tools, genetic and physical maps, and genome-wide functional analysis studies. There are innumerable strategies researchers might use to explore yeast genes, and SGD hopes to aid its users in as many of these as possible. A few likely starting points for identifying yeast genes relevant to a researcher's interests are described below.

Sequence. Because genes with similar sequences often perform similar molecular functions (although perhaps in different pathways), sequence comparisons provide a powerful method of identifying genes with common functions. Similarity to a characterized yeast gene product can give significant clues about the function of an uncharacterized gene, whether from yeast or from another organism.

Sequence comparisons against *S. cerevisiae* sequences can be performed at the SGD Web site (http://genome-www2.stanford.edu/cgi-bin/SGD/nph-blast2sgd). Users input any DNA or protein query sequence and choose from among several BLAST options. The *S. cerevisiae* DNA datasets that can be queried include the complete genomic sequence including mitochondrial DNA, ORF coding DNA, intergenic DNA, ORF upstream flanking sequences, and the set of all *S. cerevisiae* DNA sequences found in GenBank.[7] This set includes the individual results of the systematic sequencing efforts as well as those from the many laboratories which represent the yeast community. Protein datasets that can be queried include translations of all *S. cerevisiae* ORFs, and all *S. cerevisiae* protein sequences from GenPept, PIR, and Swissprot.[6,7,9] Several parameters can be modified to customize the BLAST search, and results can be returned to the user in a variety of formats. The results can be viewed immediately or optionally sent via e-mail to an address the user defines.

[9] W. C. Barker, J. S. Garavelli, H. Huang, P. B. McGarvey, B. C. Orcutt, G. Y. Srinivasarao, C. Xiao, L. S. Yeh, R. S. Ledley, J. F. Janda, F. Pfeiffer, H. W. Mewes, A. Tsugita, and C. Wu, *Nucleic Acids Res.* **28**, 41 (2000).



Gene Information

SGD Resource	Primary Application	Description
Gene Summary Paragraph	Overview an unfamiliar gene	A synopsis of published information on a gene, including selected references, written in natural language by SGD Scientific Curators.
Locus Page	Find basic gene information and use as a "launch pad" for many SGD tools	Concise, basic information on a gene product, phenotypes, mapping, sequence, functional analysis, expression data, and more. Access to many popular SGD tools and services with the gene name/sequence already entered as a default.
Global Gene Hunter	Retrieve gene information from several databases	Simultaneous retrieval of information for a given locus from the following eight databases: SGD, Genbank, PubMed, Sacch3D, Swiss-Prot, MIPS, Yeast Protein Database (YPD), Protein Information Resource (PIR)

Literature

SGD Resource	Primary Application	Description
Gene Summary Paragraph	Key references for general facts about a gene	A synopsis of published information on a gene, including selected references, written in natural language by SGD Scientific Curators.
Literature Guide	An annotated list of publications on a gene	Published literature on a gene, grouped according to biological topics. References are generated by PubMed searches and then reviewed and categorized by SGD Scientific Curators

Sequence Analysis & Comparison

SGD Resource	Primary Application	Dataset	Data Returned	Description
BLAST	Find sequence similarity	Yeast	Alignment	A very fast search algorithm that identifies similar protein or DNA sequences
FASTA	Find sequence similarity	Yeast	Alignment	A slower search algorithm that identifies similar protein or DNA sequences and can produce different results than BLAST
Genome-wide Similarity View	Overview of similar yeast genes	Yeast	Graphic	Displays all ORFs in the S. cerevisiae genome that show similarity to a query ORF's DNA, based on a Smith-Waterman protein sequence comparison.
PatMatch	Find short DNA/protein sequence matches	Yeast	Graphic	A pattern matching program that allows ambiguous characters, but not gaps. Works well for short sequences (e.g. motifs).
Worm Homologs	Identify Yeast/worm homologs	Yeast & Worm	graphic & alignment	Reports similarity between yeast and worm genes based on the comparison of the entire complement of predicted proteins from *C. elegans* and *S. cerevisiae* (analysis described in detail in Chervitz et al., (1998). *Science* 282:2022-2028).
Mammalian Homologs	Identify Mammalian homologs to yeast genes	Yeast & Mammals	alignment	Serves up pre-existing BLAST reports comparing each yeast peptide sequence against all unique human, mouse, rat, cow, and sheep protein sequences in GenBank.

Bench-top Tools

SGD Resource	Primary Application	Description
Design Primers	Pick PCR or sequencing primers	Recommends primers appropriate for either PCR or sequencing of a given gene or DNA sequence, within parameters set by the user (end points, Tm, GC/AT ratios, etc.).
Yeast Genome Restriction Analysis	Find restriction sites	Generates a restriction map of a specified DNA sequence. The restriction map may include all enzymes, or a subset of enzyme types (3' overhangs, 5' overhangs, blunt ends, or enzymes that cut once or twice).

Maps and Displays

SGD Resource	Primary Application	Display Features	Description
Genomic View	Overview yeast chromosomes, Access other maps	• Relative size of chromosomes • Location of centromeres and of select marker genes	Provides a broad overview of chromosomal features and a gateway to other map displays.
Features Map	Locate any chromosomal feature and identify neighboring features	• Readable format • Comprehensive display for specified chromosomal regions • SAGE tags	Graphic representation of a region of chromosomal DNA. Includes the locations of ORFs, centromeres, tRNAs, RNA genes, Ty transposons, LTR elements, rRNAs and snRNAs.
Physical Map	Locate ATCC clones	• ATCC clones • Comprehensive display for specified chromosomal regions	Graphic representation of a region of chromosomal DNA including all the features found on the Features Map (see above), with the addition of ATCC clones
Combined Physical and Genetic Map	Overview entire chromosome, compare mapping and sequencing data	• Simultaneous display of mapping data (cM) and sequencing data (Kbp) • View either entire chromosome or specified region	Graphic representation of a yeast chromosome, displaying all genetically and/or physically mapped ORFs.

Functional Analysis

SGD Resource	Primary Application	Description
Protein Info & Composition	Synopsis of protein information	Retrieves information about proteins from YPD.
SAGE Query (Simple)	Find data on the transcription levels of a gene. Query by: • gene name • chromosome map	Reports data on the expression profiles of genes analyzed using the SAGE technique (Serial Analysis of Gene Expression). Data are available for thousands of genes in log phase growth, S phase arrest, and G2/M phase arrest. (analysis described in detail in Velculescu, et al., (1997) *Cell* 88:243-251).
SAGE Query (Advanced)	Find data on the transcription levels of a gene. Query by: • gene name • tag sequence • expression levels during specific phases • other specified parameters	Reports data on the expression profiles of genes analyzed using the SAGE technique (Serial Analysis of Gene Expression). Data are available for thousands of genes in log phase growth, S phase arrest, and G2/M phase arrest. (analysis described in detail in Velculescu, et al., (1997) *Cell* 88:243-251).
Worm Homologs	Predict function by identifying Yeast/worm homologs	Reports similarity between yeast and worm genes based on the comparison of the entire complement of predicted proteins from *C. elegans* and *S. cerevisiae* (analysis described in detail in Chervitz et al., (1998). *Science* 282:2022-2028).
Function Junction	Retrieve data from several functional analysis projects	Simultaneous retrieval of functional analyses for a given locus from the following six project sites: SGD SAGE Query, Yeast Cell Cycle Analysis Project, Yeast PathCalling, YGAC Triples Database, Worm-Yeast Protein Comparison, Yeast Protein Function Assignment
Expression Connection	Retrieve data from several microarray experiments	Simultaneous retrieval of yeast gene expression date for a given locus from several publically available microarray experiments

FIG. 2. The SGD Resource Guide provides a listing of resources for investigating gene information, the scientific literature, sequence analysis options, benchtop tools, genetic and physical maps, and genome-wide functional analysis studies.

Comparisons using a *S. cerevisiae* sequence as the initial query sequence are easy to implement using SGD's flexible retrieval tools. Users can choose to retrieve the sequence of any locus via a pop-up menu from the appropriate locus page, or they may begin with a tool (Gene/Sequence Resources) that allows customization of sequence retrieval for a desired locus or for a chosen region of yeast DNA by specifying the chromosomal coordinates. Gene or ORF sequences can be customized by choosing whether to include introns or flanking sequences of user-specified lengths. Of course options include the ability to retrieve the reverse complement of a specified DNA sequence. Researchers can also retrieve all sequences associated with a particular locus, including the systematic ORF sequence, sequences from mapped cosmids, and individual GenBank entries. Protein sequences encoded by the systematic ORFs are available; in addition, restriction maps are available with 6-frame translations of a specified sequence.

Expression Pattern. Because genes involved in the same or related processes may have coordinated regulation of expression, searching for genes that share similar (or diametrically opposed) expression patterns may provide clues about the roles of those genes in the cell. SGD provides access via a tool called Expression Connection[10] to many published genome-wide expression studies that can be queried to identify genes whose expression is coordinated (positively or negatively) with a query gene or ORF. Users may query a single dataset or several at once. Users may also browse the clustered expression data in a given dataset to scan for genes with an expression pattern, for example, one resulting from a stimulus or correlated with a cell-cycle phase or developmental program, such as sporulation[11] (Fig. 3). In addition to being labeled with the appropriate gene names, expression profiles also show Gene Ontology annotations that can give researchers clues to the biology of genes that have similar expression patterns. These annotations are particularly useful for uncharacterized genes that fall within a group of genes with correlated expression, as the annotations of characterized genes with similar expression patterns may hint at the cellular roles for the uncharacterized genes.

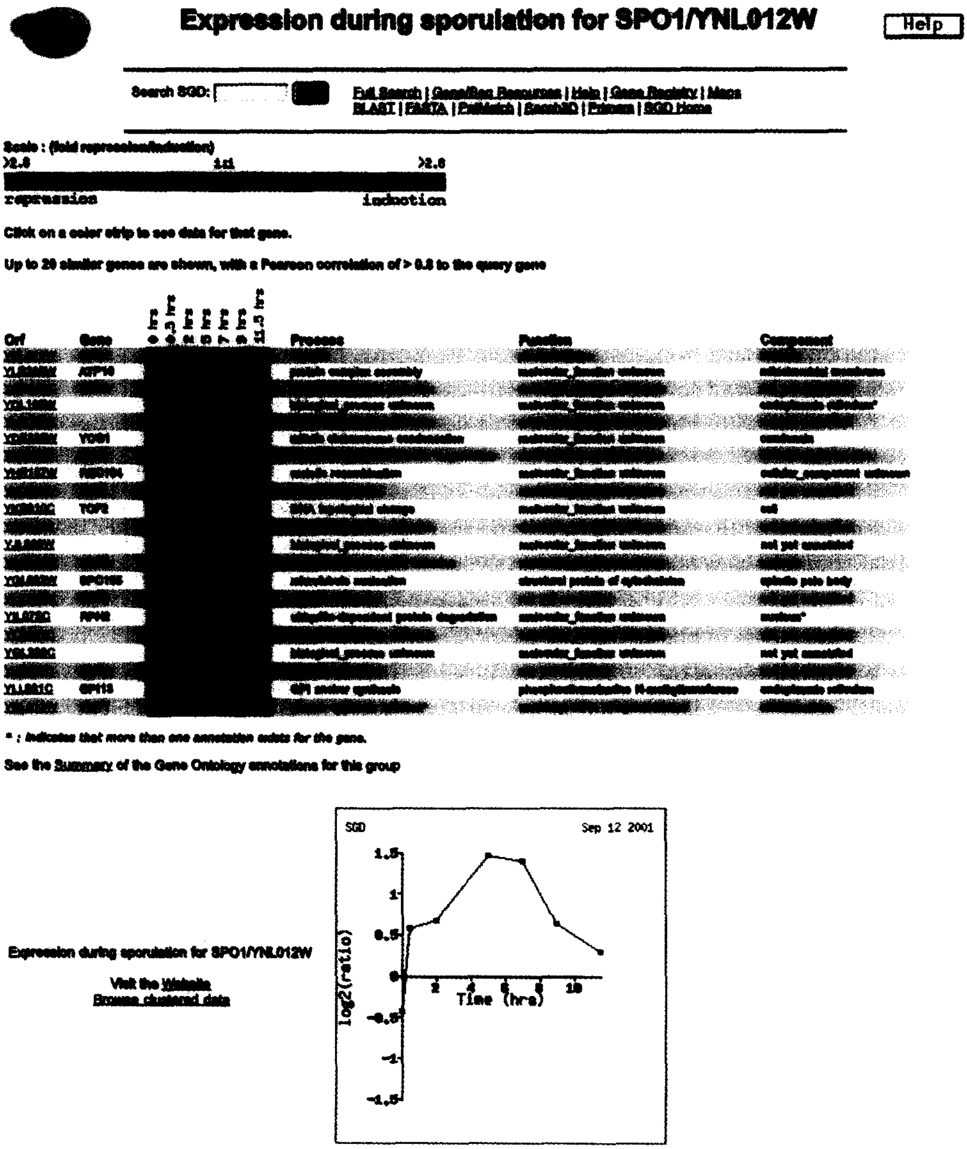
Phenotype. One of the first ways yeast genes were named and grouped was according to common mutant phenotype, partly because related mutant phenotypes could indicate that genes participate in a common process, and partly because they were isolated during comprehensive screening experiments. SGD allows users to retrieve lists of genes that share the same mutant phenotype. Included in SGD's display of phenotype data are the results of the systematic deletion project, in which each of the *S. cerevisiae* ORFs was deleted and the resulting strains analyzed.[12]

Text Search. There are occasions when a user may be interested in finding out biological information about a specific topic rather than beginning with a specific sequence or experimental result. For instance, the user may want to obtain information on the general subject of "chitin." In this case, a text search of the database for "chitin" will retrieve all information associated with this word. Examples of

[10] C. A. Ball, H. Jin, G. Sherlock, S. Weng, J. C. Matese, R. Andrada, G. Binkley, K. Dolinski, S. S. Dwight, M. A. Harris, L. Issel-Tarver, M. Schroeder, D. Botstein, and J. M. Cherry, *Nucleic Acids Res.* **29,** 80 (2001).

[11] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, *Science* **282,** 699 (1998).

[12] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, S. Whelen Dow, S. H. Friend, C. J. Roberts, T. Ward, R. W. Davis *et al., Science* **285,** 901 (1999).



Expression during sporulation for SPO1/YNL012W

Help

Search SGD:

Full Search | Gene/Seq Resources | Help | Gene Registry | Maps
BLAST | FASTA | PatMatch | SacchDB | Primers | SGD Home

Scale : (fold repression/induction)
>2.8 1:1 >2.8
repression induction

Click on a color strip to see data for that gene.

Up to 20 similar genes are shown, with a Pearson correlation of > 0.8 to the query gene

Orf	Gene	0 hrs	0.5 hrs	2 hrs	5 hrs	7 hrs	9 hrs	11.5 hrs	Process	Function	Component
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
YLR393W	ATP10								protein complex assembly	molecular_function unknown	mitochondrial membrane
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
[illegible]									biological_process unknown	molecular_function unknown	endoplasmic reticulum*
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
YDR325W	YCG1								mitotic chromosome condensation	molecular_function unknown	condensin
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
YHR157W	REC104								meiotic recombination	molecular_function unknown	cellular_component unknown
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
[illegible]	TOP2								DNA topological change	molecular_function unknown	cell
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
[illegible]									biological_process unknown	molecular_function unknown	not yet annotated
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
[illegible]	SPC110								microtubule nucleation	structural protein of cytoskeleton	spindle pole body
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
YIL075C	RPN2								ubiquitin-dependent protein degradation	molecular_function unknown	nucleus*
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
[illegible]									biological_process unknown	molecular_function unknown	not yet annotated
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]
YLL031C	GPI13								GPI anchor synthesis	phosphoethanolamine N-methyltransferase	endoplasmic reticulum
[illegible]	[illegible]								[illegible]	[illegible]	[illegible]

* : Indicates that more than one annotation exists for the gene.

See the Summary of the Gene Ontology annotations for this group

Expression during sporulation for SPO1/YNL012W

Visit the Website
Browse clustered data

FIG. 3. Data showing the 20 genes with expression most similar to *SPO1* during sporulation [S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herskowitz, *Science* **282,** 699 (1998)]. Expression information like this is most easily accessed using SGD's Expression Connection tool, or from the appropriate locus page.

the types of information that can be retrieved using a text search include locus, sequence, descriptions, phenotype, gene product, GO terms, paper abstracts, and colleague information. The Text Search is available from the "Full Search" page. Because text searches provide the ability to scan many different types of biological data at once and can therefore be very powerful, SGD has made an effort to store its data in a way that allows it to be efficiently queried. At a simple level, the association of such items as gene product and GO terms with a locus entry means that a text query may associate a locus with a given biological topic. Similarly, the association of keywords with colleagues allows one to find which colleagues might be doing research on a specific subject (for instance, querying for "chitin" brings back several colleague entries in which this word is listed as a keyword on the colleague page). On a more complex level, Gene Summary paragraphs are written in a markup language, hidden from the users, which serves to break the paragraph down into different biological topics. One purpose for designing the paragraphs this way is so that a text query can bring back a specific section of text that contains the search word and is already marked as being relevant to a specific biological topic. In early 2002 this feature will become more powerful, after the installation of a new text processing system.

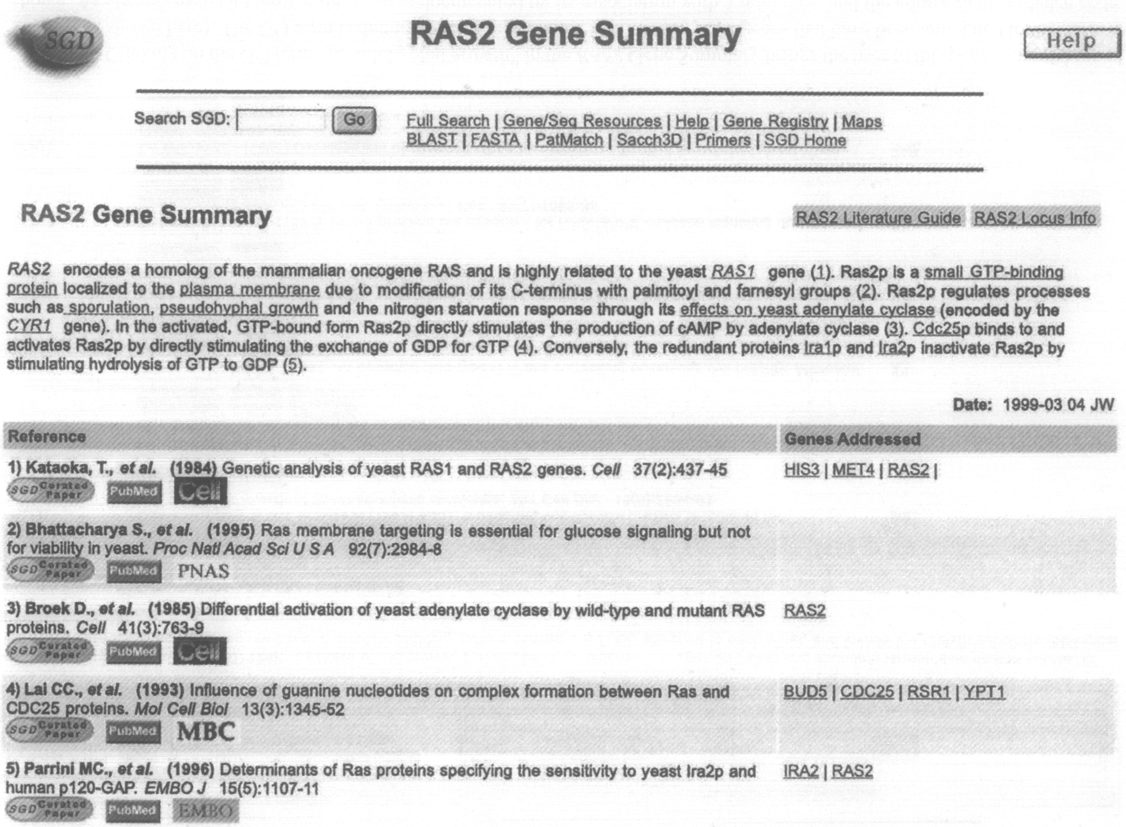
After Search: Information about Genes

As mentioned above, SGD's locus pages provide concise information about genes and gene products, including Gene Ontology annotations, phenotype descriptions, and links to a variety of resources. A few of the most useful resources SGD provides for exploring the biology of yeast genes are discussed below.

Gene Summaries. A Gene Summary is a short synopsis of the published biological information about a gene and its product and is designed to familiarize yeast and non-yeast researchers with the general facts and important subtleties regarding a locus (Fig. 4). The SGD curators compose Gene Summary paragraphs using natural language and a controlled vocabulary based on the Gene Ontology described above. A few recent publications are selected so the resulting paragraph is a snapshot of the current understanding of the gene, rather than an exhaustive review.

The first instance of each Gene Ontology term used in a Gene Summary is marked by curators so that it can serve as a link to a list of other genes that have been annotated to the same term (Fig. 5). Each sentence of a Gene Summary is also marked by curators according to the topics covered in the sentence, so that the summaries can be easily searched and parsed according to content.

Literature. SGD contains a set of those research papers (from PubMed and other sources) that are relevant to yeast biology. As an ongoing process, curators create and update a Literature Guide for each yeast gene that has been described



RAS2 Gene Summary

Full Search | Gene/Seq Resources | Help | Gene Registry | Maps
BLAST | FASTA | PatMatch | Sacch3D | Primers | SGD Home

RAS2 Gene Summary

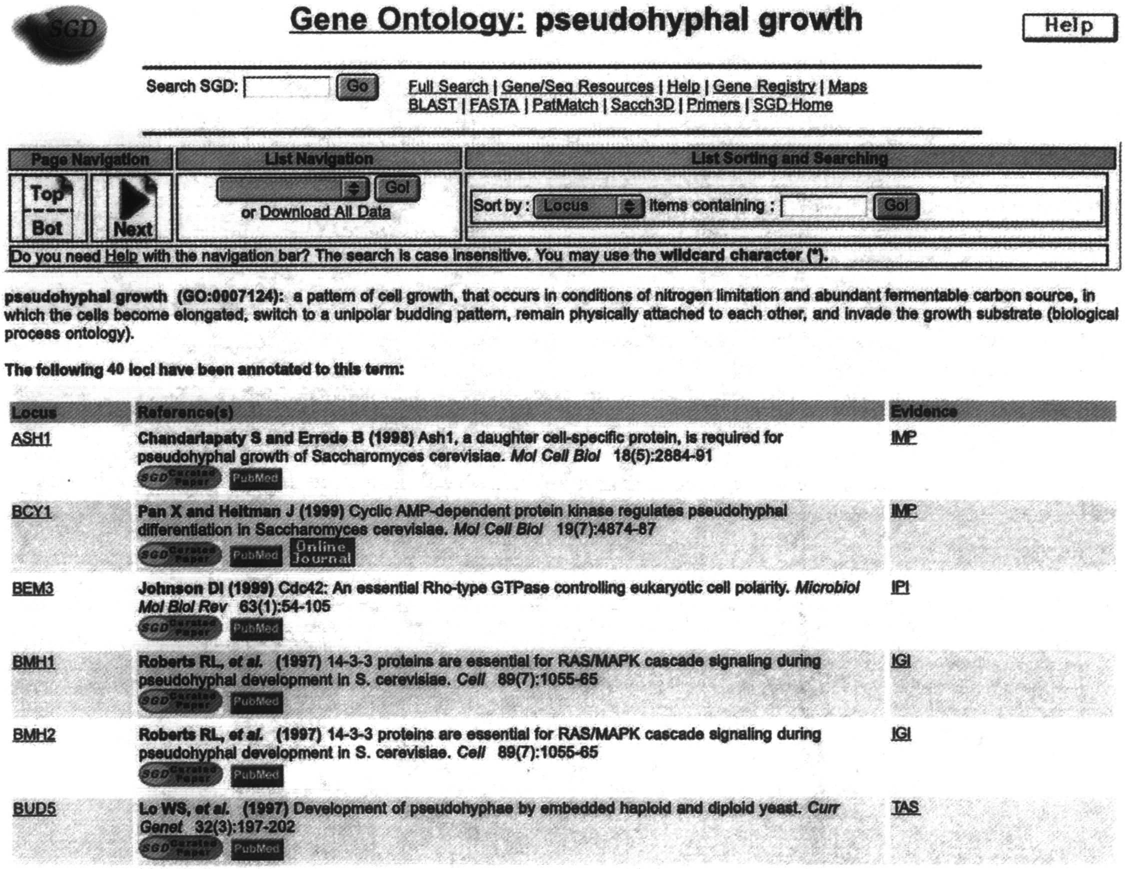
RAS2 Literature Guide | RAS2 Locus Info

RAS2 encodes a homolog of the mammalian oncogene RAS and is highly related to the yeast *RAS1* gene (1). Ras2p is a small GTP-binding protein localized to the plasma membrane due to modification of its C-terminus with palmitoyl and farnesyl groups (2). Ras2p regulates processes such as sporulation, pseudohyphal growth and the nitrogen starvation response through its effects on yeast adenylate cyclase (encoded by the *CYR1* gene). In the activated, GTP-bound form Ras2p directly stimulates the production of cAMP by adenylate cyclase (3). Cdc25p binds to and activates Ras2p by directly stimulating the exchange of GDP for GTP (4). Conversely, the redundant proteins Ira1p and Ira2p inactivate Ras2p by stimulating hydrolysis of GTP to GDP (5).

Date: 1999-03 04 JW

Reference	Genes Addressed
1) Kataoka, T., *et al.* (1984) Genetic analysis of yeast RAS1 and RAS2 genes. *Cell* 37(2):437-45	HIS3 | MET4 | RAS2 |
2) Bhattacharya S., *et al.* (1995) Ras membrane targeting is essential for glucose signaling but not for viability in yeast. *Proc Natl Acad Sci U S A* 92(7):2984-8	
3) Broek D., *et al.* (1985) Differential activation of yeast adenylate cyclase by wild-type and mutant RAS proteins. *Cell* 41(3):763-9	RAS2
4) Lai CC., *et al.* (1993) Influence of guanine nucleotides on complex formation between Ras and CDC25 proteins. *Mol Cell Biol* 13(3):1345-52	BUD5 | CDC25 | RSR1 | YPT1
5) Parrini MC., *et al.* (1996) Determinants of Ras proteins specifying the sensitivity to yeast Ira2p and human p120-GAP. *EMBO J* 15(5):1107-11	IRA2 | RAS2

FIG. 4. The Gene Summary for *RAS2* includes links to lists of genes that share GO terms, and to the locus pages of genes mentioned in the paragraph. There is also a list of references used in composing the summary. The reference display includes the list of yeast genes addressed in each publication.



Gene Ontology: pseudohyphal growth

Do you need Help with the navigation bar? The search is case insensitive. You may use the **wildcard character (*)**.

pseudohyphal growth (GO:0007124): a pattern of cell growth, that occurs in conditions of nitrogen limitation and abundant fermentable carbon source, in which the cells become elongated, switch to a unipolar budding pattern, remain physically attached to each other, and invade the growth substrate (biological process ontology).

The following 40 loci have been annotated to this term:

Locus	Reference(s)	Evidence
ASH1	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	IMP
BCY1	Pan X and Heitman J (1999) Cyclic AMP-dependent protein kinase regulates pseudohyphal differentiation in Saccharomyces cerevisiae. *Mol Cell Biol* 19(7):4874-87	IMP
BEM3	Johnson DI (1999) Cdc42: An essential Rho-type GTPase controlling eukaryotic cell polarity. *Microbiol Mol Biol Rev* 63(1):54-105	IPI
BMH1	Roberts RL, *et al.* (1997) 14-3-3 proteins are essential for RAS/MAPK cascade signaling during pseudohyphal development in S. cerevisiae. *Cell* 89(7):1055-65	IGI
BMH2	Roberts RL, *et al.* (1997) 14-3-3 proteins are essential for RAS/MAPK cascade signaling during pseudohyphal development in S. cerevisiae. *Cell* 89(7):1055-65	IGI
BUD5	Lo WS, *et al.* (1997) Development of pseudohyphae by embedded haploid and diploid yeast. *Curr Genet* 32(3):197-202	TAS

FIG. 5. Clicking on the GO term "pseudohyphal growth" in the *RAS2* Gene Summary brings the user to this page (only the top of the page is shown here). The GO term is defined on this page, and a list of other yeast genes that have been annotated to this term is shown. As always, each GO term assignment is documented by its association with a reference, and the appropriate evidence code.

in a publication (Fig. 6). By categorizing papers according to topics addressed (e.g., cellular location, protein sequence features), the guides are intended to help researchers search through the literature about a given gene quickly and efficiently. Each paper in SGD is searched for the mention of all gene names in its title or abstract. Accessing a paper allows the user to identify a group of related genes.

Other Databases. SGD provides gene-specific connections to many databases that contain important information about yeast biology. These databases provide further information about the yeast genome (MIPS), gene products (YPD and SwissProt), gene sequences (GenBank), and more.[4-8]

Connecting to Larger Biological Community

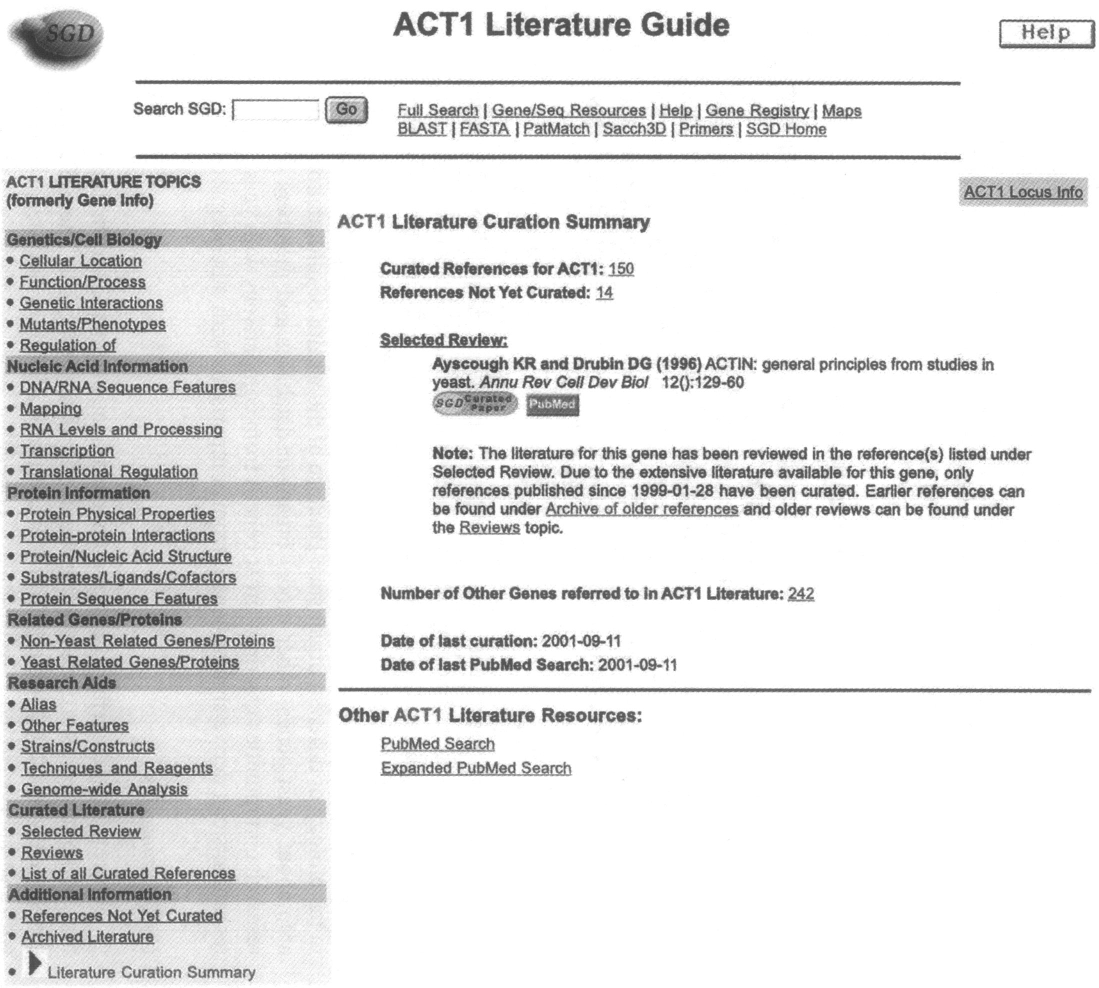
As researchers studying yeast and other organisms deepen their understanding of biological processes and the roles specific genes play in those processes, the ability to make sophisticated comparisons among different research organisms becomes increasingly important. SGD is taking several steps to facilitate the flow of biological information to and from the yeast community: annotating yeast genes to a universal framework, composing Gene Summaries to describe yeast genes, and making the database easily accessible for data mining by other scientific databases.

Annotating Yeast Genes

By annotating yeast genes to the universal framework provided by the Gene Ontology Consortium described above, comparisons of the molecular functions, biological processes, and cellular components of gene products can be made within and across species bounds[2,3] (Fig. 7). In combination with sequence comparisons, these annotations provide a powerful tool for studying similarities and differences in the biology of different organisms. Because of the wealth of information available about yeast genes, other model organism databases and members of their research communities can draw great benefit from comparisons with yeast. Yeast researchers can derive similar benefit from comparisons with other organisms.

Gene Summaries

As previously described, curator-composed Gene Summaries are brief descriptions of the current state of knowledge about individual yeast genes. The summaries are written with a target readership of those educated in biology but not necessarily with a yeast background. It is particularly hoped that these summaries will provide researchers studying other organisms a convenient entree to the body of knowledge compiled by yeast scientists. In composing the summaries, SGD curators emphasize any known relationships between the yeast genes and genes from other organisms. In some cases, such as yeast genes with a human disease gene homolog,



ACT1 Literature Guide

Search SGD: Go

Full Search | Gene/Seq Resources | Help | Gene Registry | Maps
BLAST | FASTA | PatMatch | Sacch3D | Primers | SGD Home

ACT1 LITERATURE TOPICS
(formerly Gene Info)

Genetics/Cell Biology
- Cellular Location
- Function/Process
- Genetic Interactions
- Mutants/Phenotypes
- Regulation of

Nucleic Acid Information
- DNA/RNA Sequence Features
- Mapping
- RNA Levels and Processing
- Transcription
- Translational Regulation

Protein Information
- Protein Physical Properties
- Protein-protein Interactions
- Protein/Nucleic Acid Structure
- Substrates/Ligands/Cofactors
- Protein Sequence Features

Related Genes/Proteins
- Non-Yeast Related Genes/Proteins
- Yeast Related Genes/Proteins

Research Aids
- Alias
- Other Features
- Strains/Constructs
- Techniques and Reagents
- Genome-wide Analysis

Curated Literature
- Selected Review
- Reviews
- List of all Curated References

Additional Information
- References Not Yet Curated
- Archived Literature
- ▶ Literature Curation Summary

ACT1 Locus Info

ACT1 Literature Curation Summary

Curated References for ACT1: 150

References Not Yet Curated: 14

Selected Review:

Ayscough KR and Drubin DG (1996) ACTIN: general principles from studies in yeast. *Annu Rev Cell Dev Biol* 12():129-60

SGD Curated Paper PubMed

Note: The literature for this gene has been reviewed in the reference(s) listed under Selected Review. Due to the extensive literature available for this gene, only references published since 1999-01-28 have been curated. Earlier references can be found under Archive of older references and older reviews can be found under the Reviews topic.

Number of Other Genes referred to in ACT1 Literature: 242

Date of last curation: 2001-09-11

Date of last PubMed Search: 2001-09-11

Other ACT1 Literature Resources:

PubMed Search

Expanded PubMed Search

the homologous gene named in the summary is hotlinked to a database outside SGD where readers can learn more about that related gene.

Use of SGDIDs

Another way in which SGD facilitates the free exchange of biological information among databases is by making the database easy to retrieve information from, and reliable to connect with. One component of this strategy is the use of SGDIDs, unique identifiers for elements of the genome. Using SGDIDs as accession numbers for genes and other features of the *S. cerevisiae* genome prevents problems when loci change names; the SGDIDs are a stable means of connecting entries in our database with entries in other databases, remaining unaffected by nomenclature changes.

Yeast Community Information at SGD

SGD was designed to be a centralized resource for the yeast community, and an important part of that role is to provide a forum for the collection and display of community-related information.

Colleague Information

One feature designed to facilitate communication among yeast researchers is a searchable database of colleague information. SGD users may choose to enter their contact information, a description of research interests, a list of collaborators, and relevant Web site addresses. SGD also provides a separate list of links to several yeast laboratories located around the world (http://genome-www4.stanford.edu/cgi-bin/SGD/colleague/yeastLabs/yeastLabs.pl). This list provides links to each laboratory's Web site, and to colleague information for the laboratory's Principal Investigator. In addition to these links, the above list also displays key words and gene names associated with the Principal Investigator's research. The list's search option allows users to search the list using the Principal Investigator's name, institution, gene name, or key words.

FIG. 6. SGD's Guide to the Literature for *ACT1*. The left-hand column of all of Literature Topics pages lists the various categories of biological information that were found for that locus in the PubMed abstracts. There are 32 different topics that are currently in use. A topic will be missing from the list of Literature Topics if no abstract associated with the locus has made reference to that kind of information. This column functions as a navigation bar between the individual topics and additional information including the Literature Curation Summary. The Literature Curation Summary is the starting page to access the Literature Topics. It gives the curation status, with the numbers of curated and uncurated references, the date of last curation, and the date of the last systematic search of PubMed. Any notes or information specific to the curation of the locus are found on this page, as are a link to SGD's Gene Summary Paragraph, if available, and links to PubMed to search for references that mention the locus.

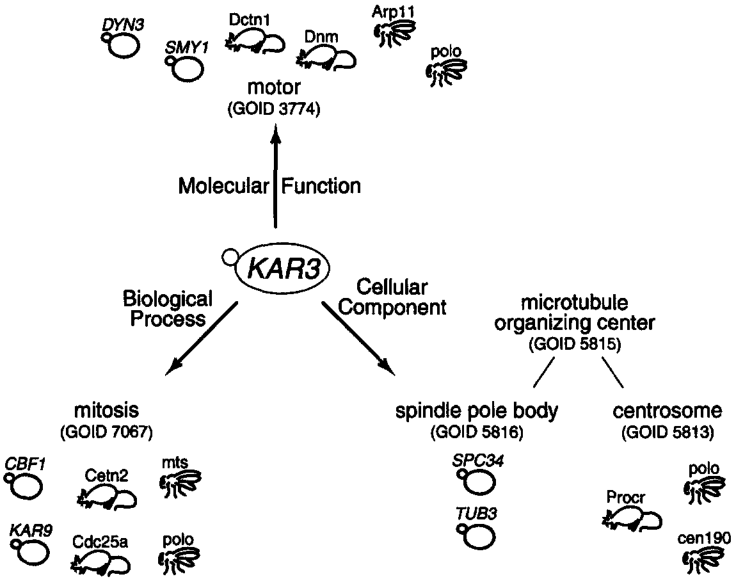


FIG. 7. GO connects across species boundaries. The controlled vocabularies of the Gene Ontology project provide a way to identify genes in multiple species with similar annotations. Here, the yeast gene, *KAR3,* is used as an example. *KAR3* is annotated to GO terms, each with a unique identifier (GOID), in each of the three ontologies: Molecular Function, Biological Process, and Cellular Component. As genes in other organisms (here we show only mouse and fly) are annotated using the same controlled vocabularies, use of these GO terms allows identification of other genes from yeast and other organisms which are involved in the same functions, processes, or structures. One benefit of the organization of the GO vocabularies and annotations is highlighted by the Cellular Component ontology annotations in this example. *KAR3* is annotated to the term "spindle pole body," an instance of the parent term "microtubule organizing center." In flies and mice, the microtubule organizing center is a different structure, the centrosome, represented by the GO term "centrosome." The ontologies show that the yeast genes *KAR3, SPC34,* and *TUB3,* the mouse gene *Procr,* and the fly genes *polo* and *cen190* are all involved in organizing microtubules.

Meetings and Community Resources

Upcoming yeast conferences and courses are listed at SGD, with links for further information and registration. For some past meetings, abstracts can be searched and lists of participants are available. SGD maintains a list of other Web sites that could be of use to yeast researchers, including the yeast "Virtual Library" of Web sites, and several databases and functional analysis Web sites. SGD also provides a searchable archive of the BioSci Yeast Newsgroup.