Using Model Organism Databases (MODs)

Stacia R. Engel[1]

[1] Stanford University School of Medicine, Stanford, California

ABSTRACT

Model Organism Databases (MODs) represent the union of database technology and biology, and are essential to modern biological and medical research. Research communities are producing floods of new data, of increasingly different types and complexity. MODs assimilate this information from a wide variety of sources, organize it in a comprehensible manner, and make it freely available to the public via the Internet. MODs permit researchers to sort through massive amounts of data, providing access to key information that they might otherwise have overlooked. The protocols in this unit offer a general introduction to different types of data available in the growing number of MODs, and approaches for accessing, browsing, and querying these data. *Curr. Protoc. Essential Lab. Tech.* 1:11.4.1-11.4.17. © 2009 by John Wiley & Sons, Inc.

Keywords: Genome project • genetics • DNA sequence • gene model • protein function

OVERVIEW AND PRINCIPLES

Recent advances in DNA sequencing technologies over the past two decades have led to an increase in the number of fully sequenced genomes and other types of publicly available DNA sequences, which has in turn allowed a great expansion in the depth and breadth of experimental data available to today's researcher. In order to make the most of this information, it must be collected, vetted, collated, and made available to the relevant scientific community (i.e., it must be *curated*). This curation occurs within the context of Model Organism Databases (MODs), which are assuming increasing importance in all areas of biology.

"Model organisms" are nonhuman organisms that are typically used for biological research. The resulting data can be used as a framework for the interpretation and understanding of similar data from humans or other medically or economically important species. Popular model organisms include budding yeast, fruit flies, and laboratory mice, all of which contain genes that encode proteins and other gene products similar to those found in humans. Genetic manipulation of model organisms is generally the most efficient path to understanding the effects of mutations in their human homologs. Model organisms have become especially effective reference species because vast amounts of data have been generated, collected, and made freely available to the public research community.

History of Model Organism Databases

In order to help researchers sort through these mountains of data, crucial resources called Model Organism Databases (MODs) have been developed. Each MOD provides easy access to the diverse types of knowledge available for a particular model organism. Two of the earliest MODs were FlyBase and the *Saccharomyces* Genome Database (SGD), both of which were established in the early 1990s. FlyBase was started by Michael Ashburner and colleagues at Cambridge University, Harvard University, and Indiana University in 1992 as an effort to collate information regarding the genes and mutations of the fruit fly *Drosophila melanogaster*, one of the most intensely studied eukaryotic

Current Protocols Essential Laboratory Techniques 11.4.1-11.4.17, June 2009
Published online June 2009 in Wiley Interscience (www.interscience.wiley.com).
DOI: 10.1002/9780470089941.et1104s01

organisms (Gelbart et al., 1997). FlyBase has since expanded its taxonomic focus and now provides myriad types of genomic and molecular information for at least a dozen different *Drosophila* species.

SGD was founded by David Botstein and colleagues at Stanford University in 1993 (Dwight et al., 2004). *Saccharomyces cerevisiae* was the first eukaryote whose genome was completely sequenced, and at the time, it was the largest genome to be fully sequenced (Goffeau et al., 1996). The *S. cerevisiae* sequencing consortium was international in scope and involved hundreds of researchers. The completion of the genomic sequence revealed some 6000 protein-coding genes and hundreds of noncoding RNAs. The data in SGD are centered around this complete genome sequence and the genes contained therein (Cherry et al., 1998). While SGD began as a repository of literature references and sequence data, it has grown over the years to include droves of information regarding gene functions, mutant phenotypes, and yeast researchers, as well as data analysis tools, sequences of other yeast species, and a wiki for users. SGD has emerged as a model MOD, and, as such, has served as the basis for the creation of at least three other MODs: DictyBase (Kreppel et al., 2004), the *Candida* Genome Database (Arnaud et al., 2005), and the *Aspergillus* Genome Database (Arnaud et al., 2009). Other well-known MODs include The Arabidopsis Information Resource (TAIR; Rhee, 2000; Reiser and Rhee, 2005), Mouse Genome Informatics (MGI; Blake et al., 2009; Shaw, 2009), the Rat Genome Database (RGD; Karolchik et al., 2007; Twigger et al., 2007), and WormBase (Stein et al., 2001; Schwarz and Sternberg, 2006), to name just a few. A more comprehensive listing of online genetic databases is provided at the end of this unit.

MOD functions

All MODs serve a variety of functions, the most important among them being the organization and presentation of experimental data from disparate sources. Think of any particular MOD as the central hub of that organism's research community; they are designed to clearly and concisely present research regarding a key organism to all biologists, regardless of specialty. The strength of MODs lies in the fact that the data contained in them are meticulously curated from the primary literature by experts, thereby providing centralized, impartial summaries of various types of biological information for use by researchers. When organized well, the juxtaposition of different types of information within a MOD presents researchers with an expanded view of the roles of the genes and gene products within a cell, thereby facilitating the formulation and testing of new hypotheses. For example, showing on a single Web page that a gene is expressed when a new sugar is introduced into the growth medium and contains a DNA-binding domain may help a researcher infer that this gene encodes a transcription factor that activates genes needed to digest the sugar. Types of data typically presented at MODs are genomic sequence and mapping data, gene expression patterns and functional characterizations, homology data, mutant phenotypes, allele variants, quantitative trait loci (QTLs), biochemical pathways, protein structures, and historical nomenclatures, as well as the primary literature from which all of this information is derived. However, the exact kinds of data presented at a particular MOD depend entirely on the experiments researchers have performed using that organism.

MODs exist as service organizations rather than research organizations. The primary function of the scientific curators, the biological experts employed by the MOD, is not to perform experiments, but rather to facilitate the open exchange of scientific information. As such, they do not produce the data displayed by the MOD; instead, they obtain and present data from peer-reviewed journals, referencing the information to ensure validity and accountability. This occurs through the unbiased, standardized presentations of data and maintenance of close relationships with the communities they serve and with staff at other MODs.

The Gene Ontology

Most MODs foster relationships with other databases to share data, develop annotation tools, and ensure consistency of the biological annotation of homologs across species. The Gene Ontology (GO) is a well-known and useful product of these interactions. The GO Consortium is a collaborative effort composed of several MOD groups and other bioinformatics groups who have come together to develop controlled vocabularies for the annotation of gene products in a wide variety of organisms (Blake and Harris, 2008). These controlled vocabularies, known as ontologies, consist of standardized terms (i.e., kinase activity, transsulfuration, mitochondrion, etc.) with controlled definitions, and include all known relationships between the terms (a "histone kinase" is a type of "protein kinase;" a "protein kinase" is a type of "phosphotransferase" etc.). Since ontologies are collectively defined and maintained by the participating MODs, using terms in the ontology to describe biological entities in all species guarantees that the language used will be consistent across research groups and scientific communities. This uniformity in representing and communicating biological knowledge improves inferences that can be made from experimental data, simplifies computational searches, and allows users to find similar data and types of information in different MODs.

The GO ontologies are divided into three domains that are needed for gene annotation in all organisms: Molecular Function, Biological Process, and Cellular Component (Harris et al., 2008). Molecular Function refers to the tasks or activities performed by individual gene products, such as transcription factor, lyase activity, or electron carrier, etc. Biological Process describes broad biological series of reactions, such as mitosis, purine metabolism, or membrane docking, etc. Cellular Component encompasses subcellular locations, structures, or macromolecular complexes, such as nucleus, microtubule, or origin recognition complex, etc. The three ontologies together contain >20,000 terms (Harris et al., 2008). Terms in the structured vocabularies are used for the annotation of gene products (proteins or RNAs) based on published experimental evidence. The annotations made by MOD curators are incorporated into their own databases, and are provided to the GO Consortium for dissemination through its Web site (*http://www.geneontology.org*). For example, Ono et al. (1999) characterize the *CYS4* gene in *S. cerevisiae*, which codes for cystathionine beta-synthase, and provide evidence that this activity is involved in the biosynthesis of cysteine via a cystathionine intermediate. This paper was used by SGD curators to assign Molecular Function (cystathionine beta-synthase activity) and Biological Process (cysteine biosynthetic process via cystathionine) annotations to *CYS4*, and the paper is linked in the database to that information and prominently displayed for users (*http://www.yeastgenome.org/cgi-bin/GO/goAnnotation.pl?locus=cys4*). Though annotations are assigned by expert curators and are clearly referenced with the appropriate peer-reviewed paper, users are always encouraged to read the primary literature from which the data contained in MODs is curated.

MOD Tools

The content of active MODs is constantly being updated and expanded, both through the curation of newly published information and through the development of new data analysis tools and visualization interfaces. The main point of entry for most MODs is the home page, and the basic unit of organization typically focuses on individual genes. Users can perform basic searches using gene names or keywords, or more complex queries of various types of data using specially designed search interfaces. Data can also be analyzed using various Web-based applications, or downloaded in bulk via interfaces or FTP (File Transfer Protocol). A MOD will often provide a site map and online help documentation describing various aspects and available tools, as well as direct help for users via e-mail interaction with the MOD's scientific curators.

This unit provides basic protocols for accessing information about genes in MODs. The growing number of MODs and the various types of data and analysis tools available from them cannot all be covered in this unit. The aim of this set of protocols is to provide a general introduction to enable the novice user to gain entry into various characteristic MODs, then find and retrieve basic information about genes. The unit will explain simple uses for two tools found at many MODs, GBrowse and Textpresso, both of which were designed as part of the Generic Model Organism Database (GMOD) project (*http://gmod.org*). GMOD began as a collaboration between four established genome databases—SGD, FlyBase, MGD, and WormBase—to develop and provide generic database architecture and software to the scientific community under an open source policy (Stein et al., 2002). The goal of the GMOD project is to generate a set of independent software components that can be mixed and matched to set up MODs for newly sequenced genomes in an efficient and cost-effective manner, without unnecessary duplication of effort in the development of curation and visualization software. The result is that many MODs share common components, making it easier and more intuitive for users to navigate the different layouts at diverse MODs. GBrowse is the genome feature browser Web application produced by GMOD (Stein et al., 2002), whereas Textpresso is the full-text literature search Web application (Müller et al., 2004).

This unit is designed as a broad, general introduction to information contained in most MODs. Basic Protocol 1 covers how to view a MOD home page, do a simple database search, navigate a gene summary page, and find genes with similar functions. Basic Protocol 2 outlines how to obtain a sequence using GBrowse. Basic Protocol 3 describes how to perform a full-text literature search using Textpresso.

NOTE: Accessing model organism databases requires a standard computer connected to the Internet and an up-to-date Web browser.

BASIC PROTOCOL 1

GENERAL GUIDELINES FOR USING A MODEL ORGANISM DATABASE USING THE *SACCHAROMYCES* GENOME DATABASE AS AN EXAMPLE

MOD home pages provide entry points to the various features of the Web sites. This protocol introduces a MOD home page and its features, using the *Saccharomyces* Genome Database (SGD) as an example.

Navigating the SGD home page

1. Open the SGD home page, at *http://www.yeastgenome.org*, in a Web browser.

   The SGD home page, like most MOD home pages, is divided into several different sections, including a section for news and announcements in the main body of the page, a hyperlinked listing of different types of resources, and a search box (Fig. 11.4.1).

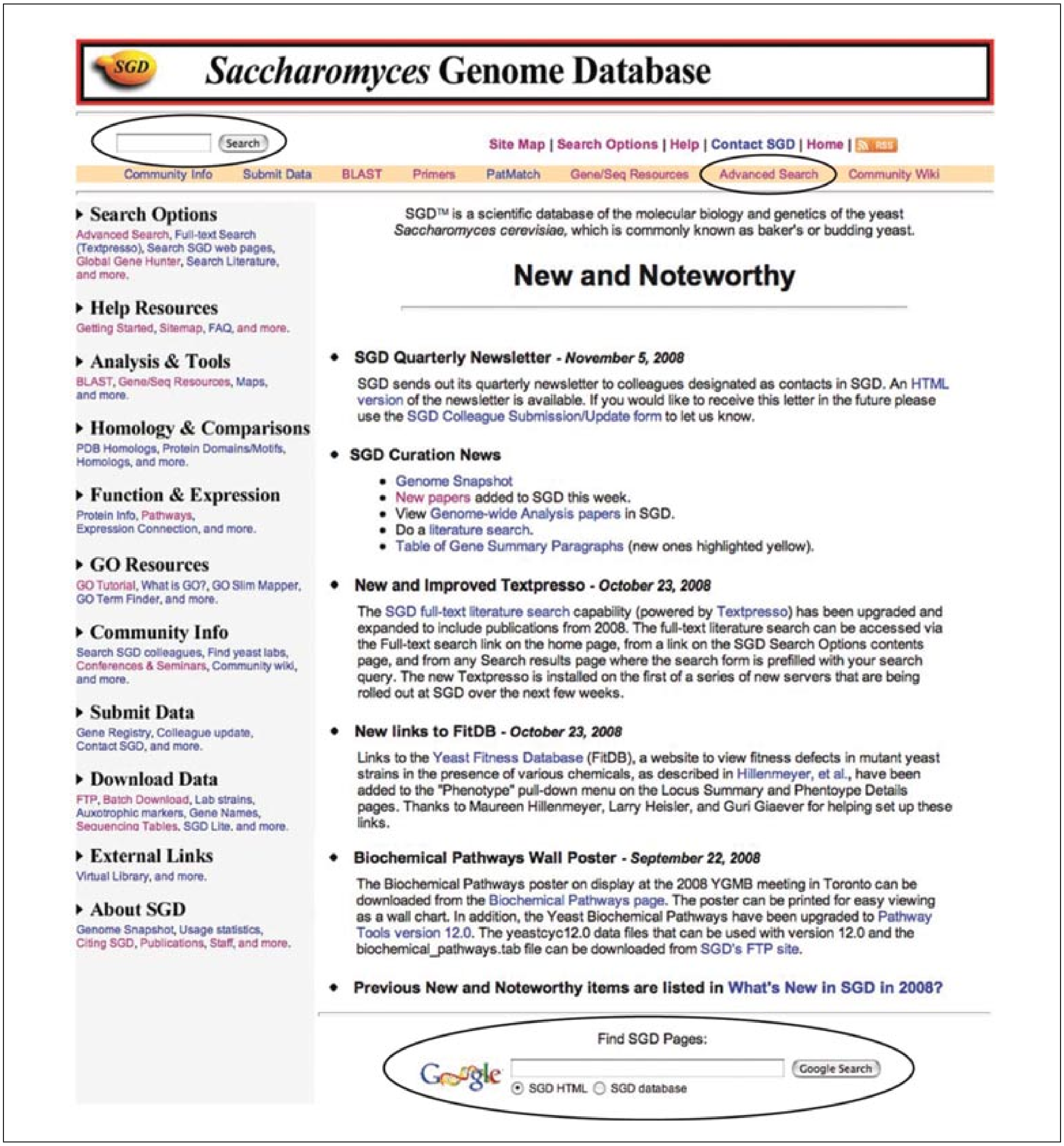
   Other MOD home page URLs are listed in Internet Resources at the end of this unit.

2. Read through the listing of the different types of resource categories available in SGD (Fig. 11.4.1).

   Common MOD resources available in SGD are accessible via the left-hand side of the home page, and are divided into broad categories: Search Options, Help Resources, Analysis & Tools, Homology & Comparisons, Function & Expression, GO Resources, Community Info, Submit Data, Download Data, External Links, and About SGD (Fig. 11.4.1). Each category and subtopic listing is linked to the corresponding section or page within the SGD Web site.

3. Explore the tools available via the toolbar located at the top of the page (Fig. 11.4.1).

   The toolbar running across the top of the SGD home page is present on most SGD Web pages, and includes links to popular tools and resources: Search, Site Map, Help, Contact SGD, RSS feed, Community Info, Submit Data, BLAST, Primers, Pattern Match, Gene and Sequence Resources, Advanced Search, and Community Wiki.



Saccharomyces Genome Database

SGD
Search
Site Map | Search Options | Help | Contact SGD | Home | RSS
Community Info Submit Data BLAST Primers PatMatch Gene/Seq Resources Advanced Search Community Wiki

Search Options
Advanced Search, Full-text Search (Textpresso), Search SGD web pages, Global Gene Hunter, Search Literature, and more.

Help Resources
Getting Started, Sitemap, FAQ, and more.

Analysis & Tools
BLAST, Gene/Seq Resources, Maps, and more.

Homology & Comparisons
PDB Homologs, Protein Domains/Motifs, Homologs, and more.

Function & Expression
Protein Info, Pathways, Expression Connection, and more.

GO Resources
GO Tutorial, What is GO?, GO Slim Mapper, GO Term Finder, and more.

Community Info
Search SGD colleagues, Find yeast labs, Conferences & Seminars, Community wiki, and more.

Submit Data
Gene Registry, Colleague update, Contact SGD, and more.

Download Data
FTP, Batch Download, Lab strains, Auxotrophic markers, Gene Names, Sequencing Tables, SGD Lite, and more.

External Links
Virtual Library, and more.

About SGD
Genome Snapshot, Usage statistics, Citing SGD, Publications, Staff, and more.

SGD™ is a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*, which is commonly known as baker's or budding yeast.

New and Noteworthy

- **SGD Quarterly Newsletter** - *November 5, 2008*

  SGD sends out its quarterly newsletter to colleagues designated as contacts in SGD. An HTML version of the newsletter is available. If you would like to receive this letter in the future please use the SGD Colleague Submission/Update form to let us know.

- **SGD Curation News**
  - Genome Snapshot
  - New papers added to SGD this week.
  - View Genome-wide Analysis papers in SGD.
  - Do a literature search.
  - Table of Gene Summary Paragraphs (new ones highlighted yellow).

- **New and Improved Textpresso** - *October 23, 2008*

  The SGD full-text literature search capability (powered by Textpresso) has been upgraded and expanded to include publications from 2008. The full-text literature search can be accessed via the Full-text search link on the home page, from a link on the SGD Search Options contents page, and from any Search results page where the search form is prefilled with your search query. The new Textpresso is installed on the first of a series of new servers that are being rolled out at SGD over the next few weeks.

- **New links to FitDB** - *October 23, 2008*

  Links to the Yeast Fitness Database (FitDB), a website to view fitness defects in mutant yeast strains in the presence of various chemicals, as described in Hillenmeyer, et al., have been added to the "Phenotype" pull-down menu on the Locus Summary and Phenotype Details pages. Thanks to Maureen Hillenmeyer, Larry Heisler, and Guri Giaever for helping set up these links.

- **Biochemical Pathways Wall Poster** - *September 22, 2008*

  The Biochemical Pathways poster on display at the 2008 YGMB meeting in Toronto can be downloaded from the Biochemical Pathways page. The poster can be printed for easy viewing as a wall chart. In addition, the Yeast Biochemical Pathways have been upgraded to Pathway Tools version 12.0. The yeastcyc12.0 data files that can be used with version 12.0 and the biochemical_pathways.tab file can be downloaded from SGD's FTP site.

- **Previous New and Noteworthy items are listed in What's New in SGD in 2008?**

Find SGD Pages:

Figure 11.4.1 The SGD home page (*http://www.yeastgenome.org*), like most MOD home pages, is the main point of entry to the Web site. The home page lists news items and announcements, and provides links to different areas and tools provided by SGD. Ovals indicate the database Search box, a link to the Advanced Search tool, and the Google-based html Web site search.

4. Click on a link of interest to go to that Web page.

Performing a simple database search

Most MODs will provide a search box on its home page and most other Web pages that accesses a simple search.

5. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).

6. Enter a word, phrase, gene or protein name, author name, etc., into the Search box (Fig. 11.4.1).

   Simple searches at most MODs search a wide variety of database fields, including gene and protein names, systematic nomenclature, functional annotations, cellular annotations, phenotypes, external IDs, literature, and researchers.

   If too many results are obtained, use more specific search terms. If too few results are returned, use more general search terms. Many MODs support simple searches using one or more wildcard characters (*).

7. If you do not find what you are looking for, you may wish to try alternative Search Options, such as an Advanced Search or a Web site html search.

   Most MODs provide Advanced Search tools that can be used for more refined or complex searches (Fig. 11.4.1). Many MODs also allow use of the Google search engine to search the words and phrases in static html pages directly from the MOD Web sites (Fig. 11.4.1).

Navigating an SGD gene summary page

The basic unit of organization of most MODs is the gene summary. A gene summary page will generally provide a synopsis of everything of biological significance that is known about a gene. For well-characterized genes, the information and summary can be quite complex. For genes about which little is known, the summary may be quite sparse. Gene summaries are continually updated as new data and information become available.

8. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).

9. Enter a gene name in the Search box (Fig. 11.4.1). A successful search will lead directly to the gene summary of the requested gene.

   Sometimes a gene name search will match more than one locus. In such cases, most MODs will display a list of gene search results. Click on the gene name to view the gene summary page.

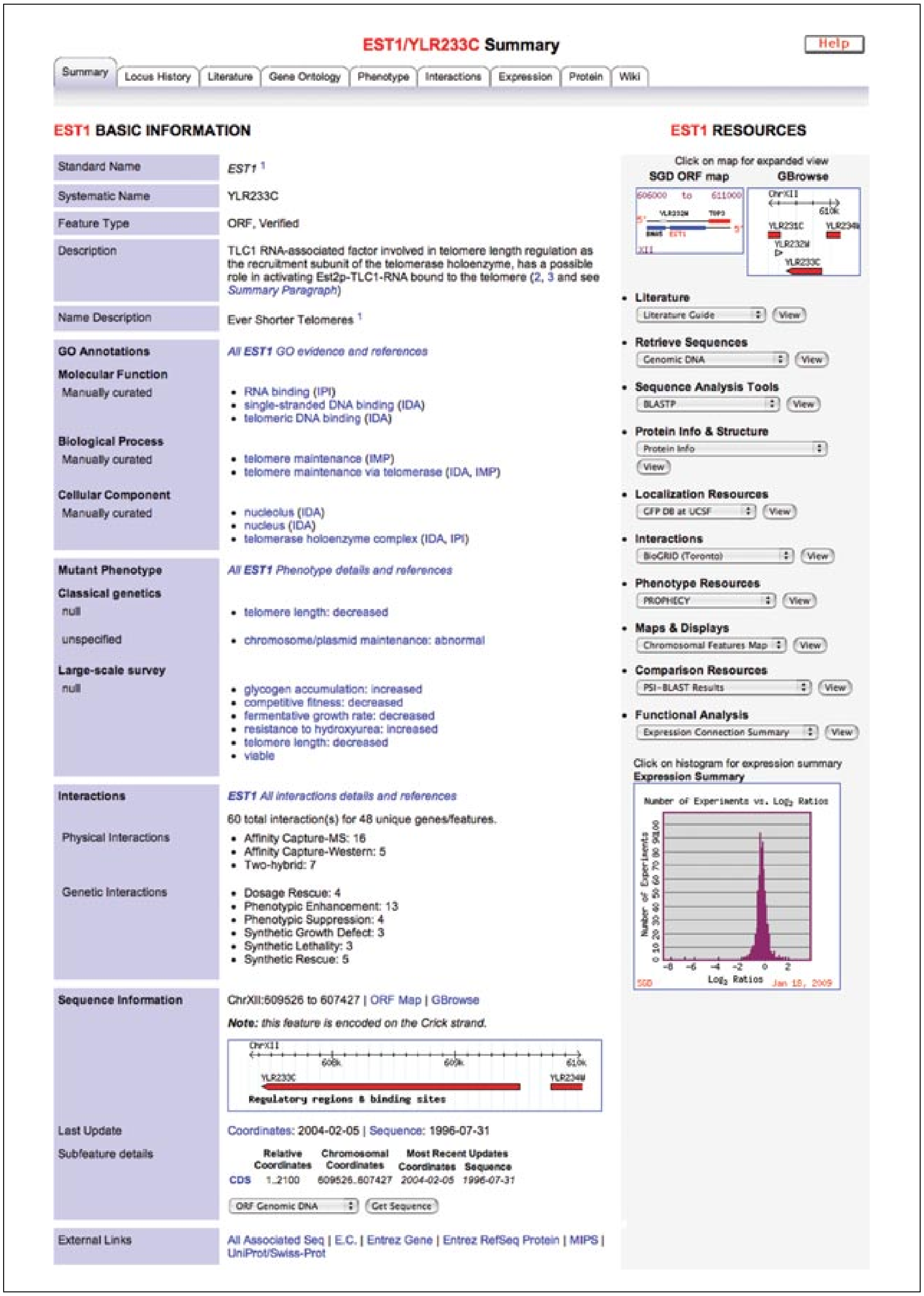
   Many MODs also support gene name searches using one or more wildcard characters (*), such as ABC.*

10. Explore the different types of information on the gene page (Fig. 11.4.2). In most cases, this will include the following in varying types and amounts.

    a. *Basic information:* General information on a gene page typically includes the standardized name given to the gene by the genome sequencing center or consortium, other published names given to the gene by researchers, chromosomal location, and a brief description of the function of the gene product.

    b. *Functional information:* Functional annotations on a gene page may include controlled vocabulary terms, such as from the Gene Ontology (Blake and Harris, 2008), which are used to describe a gene product's activity, any biological processes to which the gene product contributes, and the subcellular locations in which the gene product is found. Information related to a gene product's function may also include mutant phenotypes, genetic and physical interactions, or metabolic pathways in which the gene product participates.

    c. *Gene model(s) and nucleotide sequences:* A gene summary will typically display a listing of exons, introns, regulatory features, etc., with chromosomal and relative coordinates, as well as graphics depicting the gene model and its location on the appropriate chromosome. Some MODs provide details regarding alternative splice variants. Links are generally provided to genome browsers (Donlin, 2007), coding and genomic sequences, and sequence analysis tools such as BLAST,



EST1/YLR233C Summary

EST1 BASIC INFORMATION

Standard Name	*EST1* [1]
Systematic Name	YLR233C
Feature Type	ORF, Verified
Description	TLC1 RNA-associated factor involved in telomere length regulation as the recruitment subunit of the telomerase holoenzyme, has a possible role in activating Est2p-TLC1-RNA bound to the telomere (2, 3 and see *Summary Paragraph*)
Name Description	Ever Shorter Telomeres [1]
GO Annotations	*All EST1 GO evidence and references*
Molecular Function Manually curated	• RNA binding (IPI) • single-stranded DNA binding (IDA) • telomeric DNA binding (IDA)
Biological Process Manually curated	• telomere maintenance (IMP) • telomere maintenance via telomerase (IDA, IMP)
Cellular Component Manually curated	• nucleolus (IDA) • nucleus (IDA) • telomerase holoenzyme complex (IDA, IPI)
Mutant Phenotype	*All EST1 Phenotype details and references*
Classical genetics null	• telomere length: decreased
unspecified	• chromosome/plasmid maintenance: abnormal
Large-scale survey null	• glycogen accumulation: increased • competitive fitness: decreased • fermentative growth rate: decreased • resistance to hydroxyurea: increased • telomere length: decreased • viable
Interactions	*EST1 All interactions details and references* 60 total interaction(s) for 48 unique genes/features.
Physical Interactions	• Affinity Capture-MS: 16 • Affinity Capture-Western: 5 • Two-hybrid: 7
Genetic Interactions	• Dosage Rescue: 4 • Phenotypic Enhancement: 13 • Phenotypic Suppression: 4 • Synthetic Growth Defect: 3 • Synthetic Lethality: 3 • Synthetic Rescue: 5
Sequence Information	ChrXII:609526 to 607427 | ORF Map | GBrowse ***Note:*** *this feature is encoded on the Crick strand.*
Last Update	Coordinates: 2004-02-05 | Sequence: 1996-07-31
Subfeature details	see below
External Links	All Associated Seq | E.C. | Entrez Gene | Entrez RefSeq Protein | MIPS | UniProt/Swiss-Prot

	Relative Coordinates	Chromosomal Coordinates	Most Recent Updates Coordinates	Most Recent Updates Sequence
CDS	1..2100	609526..607427	*2004-02-05*	*1996-07-31*

ORF Genomic DNA ⇕ (Get Sequence)

EST1 RESOURCES

Click on map for expanded view

SGD ORF map GBrowse

- **Literature**
  Literature Guide ⇕ (View)
- **Retrieve Sequences**
  Genomic DNA ⇕ (View)
- **Sequence Analysis Tools**
  BLASTP ⇕ (View)
- **Protein Info & Structure**
  Protein Info ⇕ (View)
- **Localization Resources**
  GFP DB at UCSF ⇕ (View)
- **Interactions**
  BioGRID (Toronto) ⇕ (View)
- **Phenotype Resources**
  PROPHECY ⇕ (View)
- **Maps & Displays**
  Chromosomal Features Map ⇕ (View)
- **Comparison Resources**
  PSI-BLAST Results ⇕ (View)
- **Functional Analysis**
  Expression Connection Summary ⇕ (View)

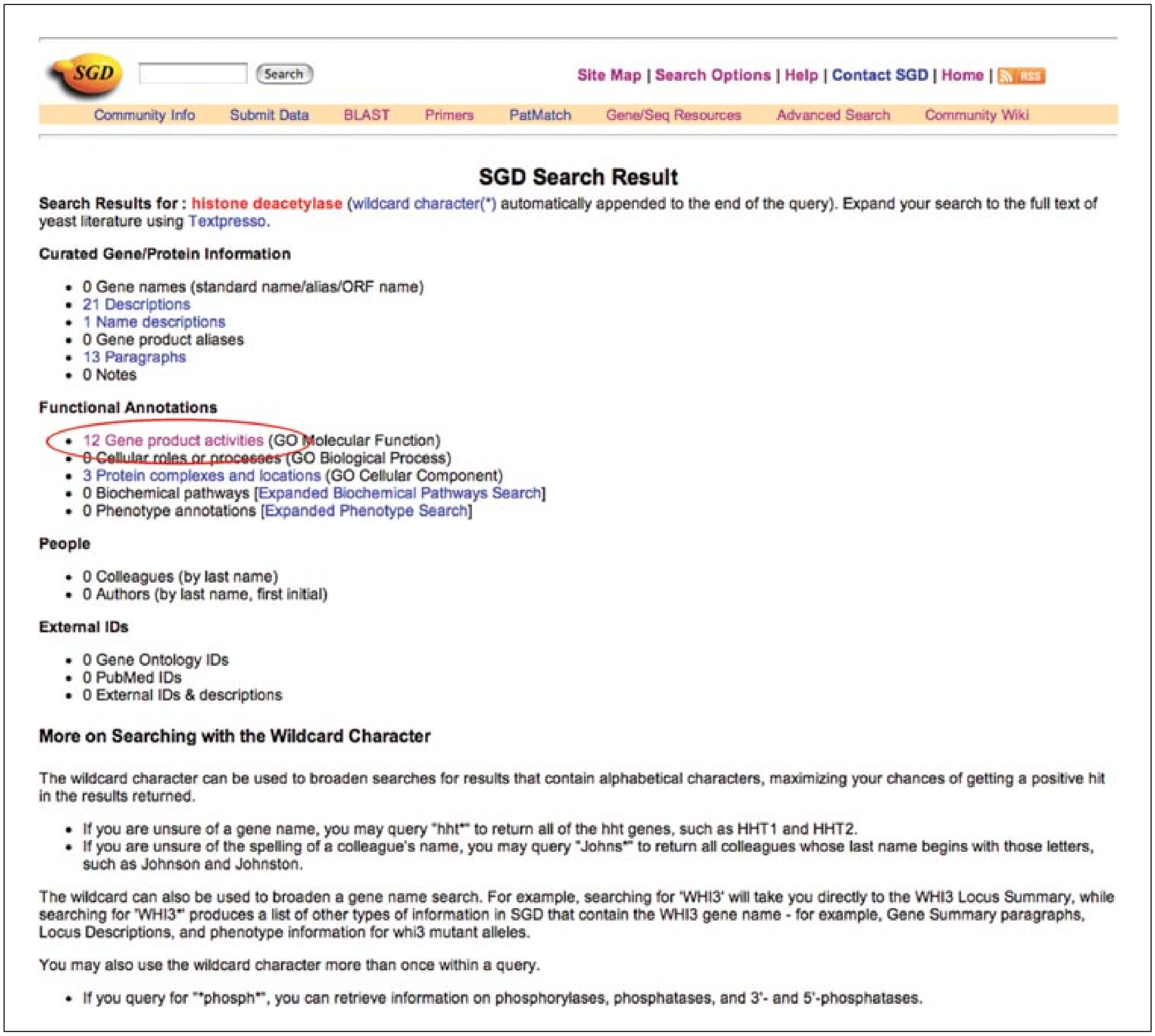
Click on histogram for expression summary

Expression Summary

Figure 11.4.2 A locus summary page in SGD showing the different types of information included on a typical gene page. A portion of the page at the top that contains the SGD toolbar and Search box is not shown. A portion of the page at the bottom that contains a gene summary paragraph and references is not shown.

FASTA, etc. Many MODs will also provide version dates for gene models and sequences, and/or a "Locus History" describing any past annotation changes to the gene model or sequence.

d. *Protein information:* For protein-coding genes, MODs will generally provide protein information such as amino acid sequence, conserved domains, protein structure and physical properties, post-translational modifications, etc.



SGD Search Result

Search Results for : histone deacetylase (wildcard character(*) automatically appended to the end of the query). Expand your search to the full text of yeast literature using Textpresso.

Curated Gene/Protein Information

- 0 Gene names (standard name/alias/ORF name)
- 21 Descriptions
- 1 Name descriptions
- 0 Gene product aliases
- 13 Paragraphs
- 0 Notes

Functional Annotations

- 12 Gene product activities (GO Molecular Function)
- 0 Cellular roles or processes (GO Biological Process)
- 3 Protein complexes and locations (GO Cellular Component)
- 0 Biochemical pathways [Expanded Biochemical Pathways Search]
- 0 Phenotype annotations [Expanded Phenotype Search]

People

- 0 Colleagues (by last name)
- 0 Authors (by last name, first initial)

External IDs

- 0 Gene Ontology IDs
- 0 PubMed IDs
- 0 External IDs & descriptions

More on Searching with the Wildcard Character

The wildcard character can be used to broaden searches for results that contain alphabetical characters, maximizing your chances of getting a positive hit in the results returned.

- If you are unsure of a gene name, you may query "hht*" to return all of the hht genes, such as HHT1 and HHT2.
- If you are unsure of the spelling of a colleague's name, you may query "Johns*" to return all colleagues whose last name begins with those letters, such as Johnson and Johnston.

The wildcard can also be used to broaden a gene name search. For example, searching for 'WHI3' will take you directly to the WHI3 Locus Summary, while searching for 'WHI3*' produces a list of other types of information in SGD that contain the WHI3 gene name - for example, Gene Summary paragraphs, Locus Descriptions, and phenotype information for whi3 mutant alleles.

You may also use the wildcard character more than once within a query.

- If you query for "*phosph*", you can retrieve information on phosphorylases, phosphatases, and 3'- and 5'-phosphatases.

Figure 11.4.3 Search Results for gene/protein information and functional annotations in SGD from a simple search using the phrase “histone deacetylase.”

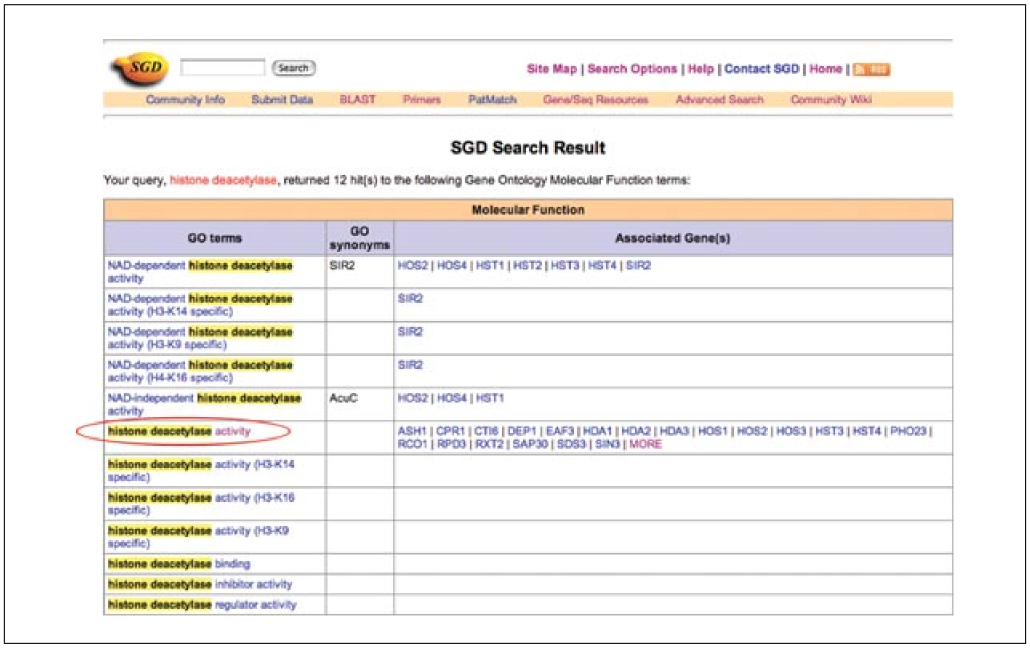
e. *External information:* Many MODs will augment the information they provide on gene pages by providing links to the primary literature or to external Web-based resources such as other databases or search engines, etc.

11. Click on a link of interest for more information.

Finding genes or proteins with similar functions

Many MODs use the Gene Ontology (GO) controlled vocabulary terms to describe a gene product’s molecular function, the biological processes for which it is required, and the cellular location and/or protein complexes in which it can be found. These controlled vocabulary terms can be exploited to find genes with similar functions, genes whose products participate in a specific biological process, or gene products that are found in specific cellular compartments. Because many MODs use these same Gene Ontology controlled vocabulary terms, they can also be used to make cross-species comparisons. For a more complete description of Gene Ontology terms, and ways in which they can be used, see Blake and Harris (2008).

12. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).

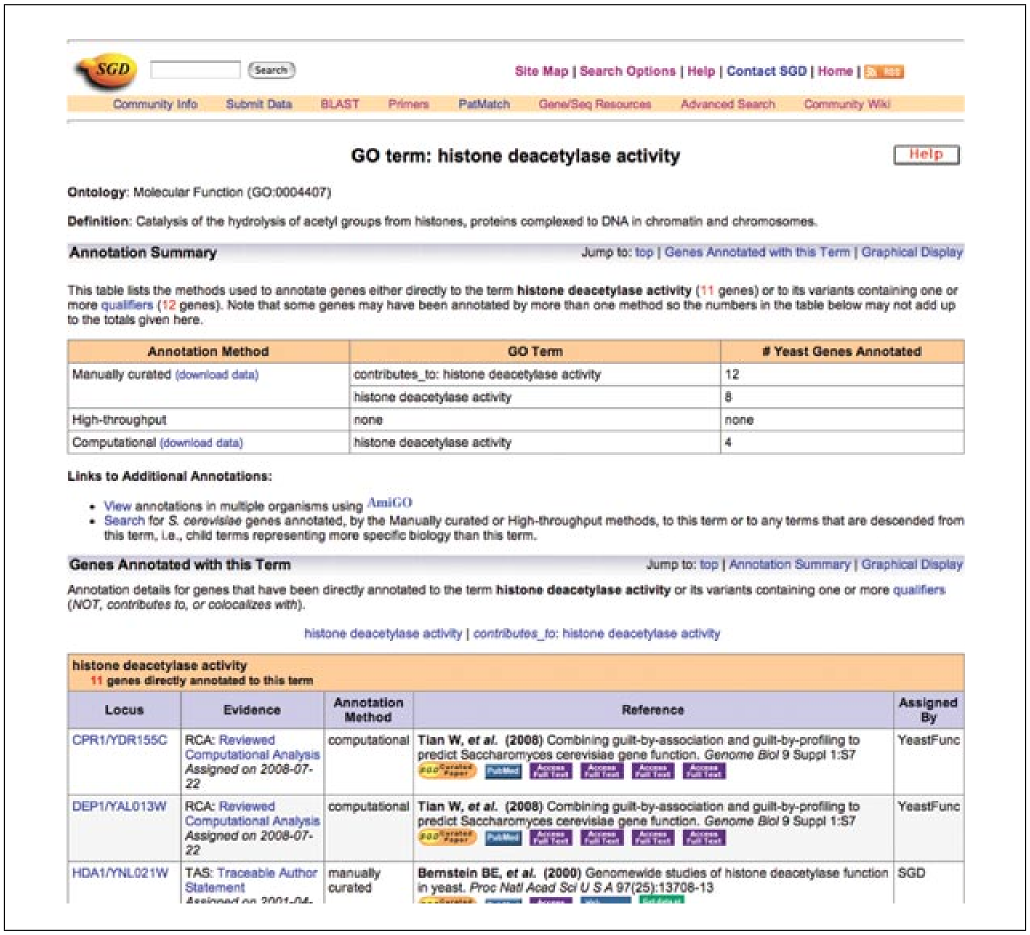


SGD Search Result

Your query, histone deacetylase, returned 12 hit(s) to the following Gene Ontology Molecular Function terms:

Molecular Function		
GO terms	GO synonyms	Associated Gene(s)
NAD-dependent histone deacetylase activity	SIR2	HOS2 | HOS4 | HST1 | HST2 | HST3 | HST4 | SIR2
NAD-dependent histone deacetylase activity (H3-K14 specific)		SIR2
NAD-dependent histone deacetylase activity (H3-K9 specific)		SIR2
NAD-dependent histone deacetylase activity (H4-K16 specific)		SIR2
NAD-independent histone deacetylase activity	AcuC	HOS2 | HOS4 | HST1
histone deacetylase activity		ASH1 | CPR1 | CTI6 | DEP1 | EAF3 | HDA1 | HDA2 | HDA3 | HOS1 | HOS2 | HOS3 | HST3 | HST4 | PHO23 | RCO1 | RPD3 | RXT2 | SAP30 | SDS3 | SIN3 | MORE
histone deacetylase activity (H3-K14 specific)		
histone deacetylase activity (H3-K16 specific)		
histone deacetylase activity (H3-K9 specific)		
histone deacetylase binding		
histone deacetylase inhibitor activity		
histone deacetylase regulator activity		

Figure 11.4.4 Search Results for Gene Ontology molecular function terms that contain the words "histone deacetylase."



GO term: histone deacetylase activity

Ontology: Molecular Function (GO:0004407)

Definition: Catalysis of the hydrolysis of acetyl groups from histones, proteins complexed to DNA in chromatin and chromosomes.

Annotation Summary Jump to: top | Genes Annotated with this Term | Graphical Display

This table lists the methods used to annotate genes either directly to the term **histone deacetylase activity** (11 genes) or to its variants containing one or more qualifiers (12 genes). Note that some genes may have been annotated by more than one method so the numbers in the table below may not add up to the totals given here.

Annotation Method	GO Term	# Yeast Genes Annotated
Manually curated (download data)	contributes_to: histone deacetylase activity	12
	histone deacetylase activity	8
High-throughput	none	none
Computational (download data)	histone deacetylase activity	4

Links to Additional Annotations:

- View annotations in multiple organisms using AmiGO
- Search for *S. cerevisiae* genes annotated, by the Manually curated or High-throughput methods, to this term or to any terms that are descended from this term, i.e., child terms representing more specific biology than this term.

Genes Annotated with this Term Jump to: top | Annotation Summary | Graphical Display

Annotation details for genes that have been directly annotated to the term **histone deacetylase activity** or its variants containing one or more qualifiers (*NOT, contributes to, or colocalizes with*).

histone deacetylase activity | contributes_to: histone deacetylase activity

histone deacetylase activity 11 genes directly annotated to this term				
Locus	Evidence	Annotation Method	Reference	Assigned By
CPR1/YDR155C	RCA: Reviewed Computational Analysis *Assigned on 2008-07-22*	computational	**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 SGD Curated Paper PubMed Access Full Text Access Full Text Access Full Text Access Full Text	YeastFunc
DEP1/YAL013W	RCA: Reviewed Computational Analysis *Assigned on 2008-07-22*	computational	**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 SGD Curated Paper PubMed Access Full Text Access Full Text Access Full Text Access Full Text	YeastFunc
HDA1/YNL021W	TAS: Traceable Author Statement *Assigned on 2001-04-*	manually curated	**Bernstein BE, *et al.* (2000)** Genomewide studies of histone deacetylase function in yeast. *Proc Natl Acad Sci U S A* 97(25):13708-13	SGD

Figure 11.4.5 The top portion of the Gene Ontology detail page in SGD for the molecular function term "histone deacetylase activity."

13. Perform a simple database search using keywords of choice, such as kinase, transcription, membrane, etc.

    Most MODs will return a search results page containing links to the various pages that match the search term. These pages are typically grouped into categories.

14. Click on a link of interest to see specific terms that match the keyword used for the search.

    For example, a search within SGD using the term "histone deacetylase" returns an intermediate Search Result page listing hits in different categories that match the term "histone deacetylase" (Fig. 11.4.3). Clicking on the "Gene product activities" link opens a page displaying the various Gene Ontology molecular function terms that contain the words "histone deacetylase," as well lists of genes whose products execute those activities (Fig. 11.4.4). Click on a term name, such as "histone deacetylase activity," to go to a page that describes that function and lists all genes annotated with that term (Fig. 11.4.5).

BASIC PROTOCOL 2

OBTAINING A SEQUENCE FROM GBROWSE

Many MODs display genomic features and other annotations using the GBrowse genome browser. GBrowse enables users to enter the genome by performing a text or sequence search, examine and scroll through genomic regions of choice, and view and download nucleotide sequences (Stein et al., 2002).

Search for a gene or region

1. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).

2. Click on the Analysis & Tools link at the top left of the home page to open the Analysis & Tools section of the Web site (*http://www.yeastgenome.org/ATContents.shtml*), which contains links to various DNA and protein sequence analysis tools, such as BLAST and Gene/Sequence Resources, as well as GBrowse.

3. Click on the Genome Browser link to launch GBrowse (*http://www.yeastgenome.org/cgi-bin/gbrowse/scgenome/*; Fig. 11.4.6).

    The GBrowse genome browser is also accessible via SGD gene summary pages by clicking on the GBrowse thumbnail at the top right of the page under Resources, the GBrowse thumbnail in the Sequence Information section near the center middle of the page, or the GBrowse link above the GBrowse thumbnail in the Sequence Information section (Fig. 11.4.2).

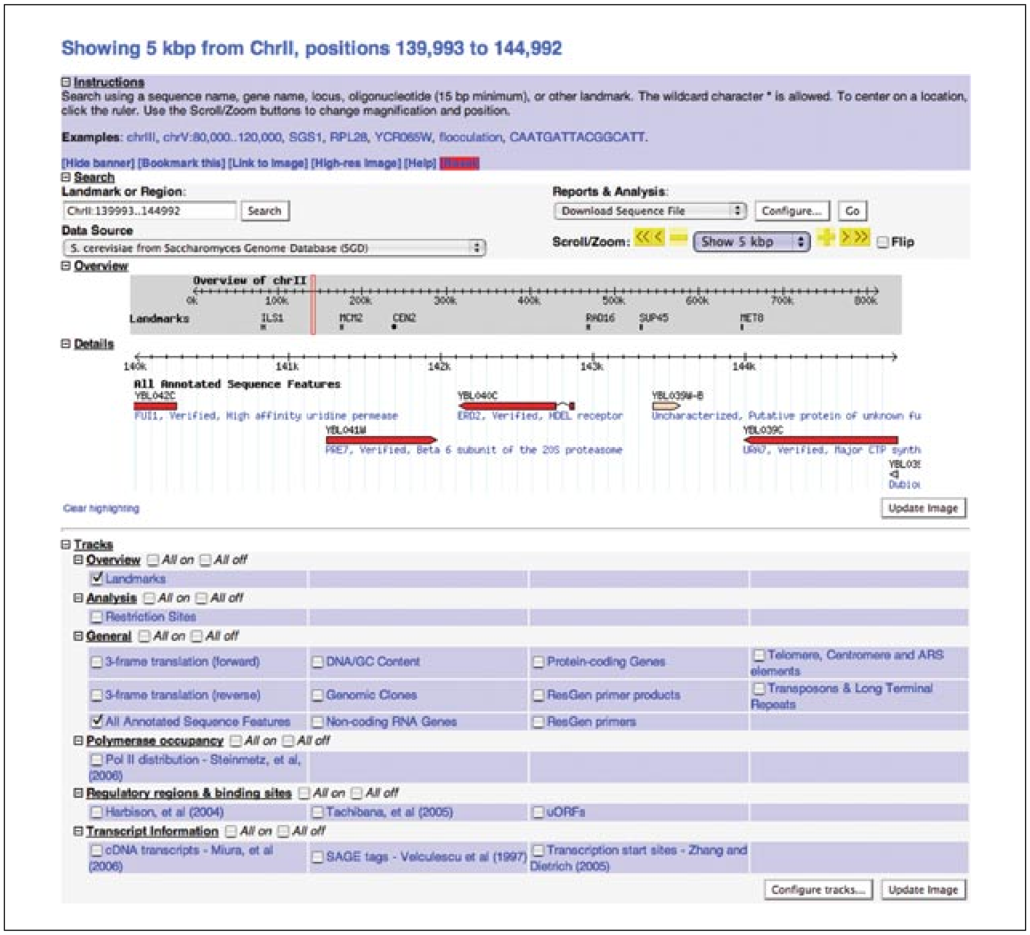
4. To view a genomic region, enter a search term of choice in the text field labeled Landmark or Region and click Search (Fig. 11.4.6). Users can search using a sequence or protein name, gene name, locus identifier, oligonucleotide sequence (15 bp minimum), or other landmark. The wildcard character (*) is allowed.

    The types of search terms allowed are configurable by each database administrator, but typically include chromosome names or numbers, clone or contig identifiers, gene names, accession numbers, systematic locus identifiers, enzymatic activities, etc. Examples include chrIII, chrV:80,000..120,000, SGS1, YCR065W, centromere, flocculation, ribosome, sulfite reductase, and CAATGATTACGGCATT.

Examine a specific gene or region of the genome

5. If more than one search result was generated, such as from a search using a text string or oligonucleotide sequence (Fig. 11.4.7), click on a link of choice to view the corresponding region of the genome.

    The Overview panel depicts the genomic context, typically an entire chromosome, in which the specific genomic feature or region resides. The section of the genome shown in the Details section is indicated with a red rectangle on the chromosome schematic. If the



Showing 5 kbp from ChrII, positions 139,993 to 144,992

⊟ Instructions
Search using a sequence name, gene name, locus, oligonucleotide (15 bp minimum), or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: chrIII, chrV:80,000..120,000, SGS1, RPL28, YCR065W, flocculation, CAATGATTACGGCATT.

[Hide banner] [Bookmark this] [Link to image] [High-res image] [Help] [Reset]

⊟ Search
Landmark or Region: ChrII:139993..144992 [Search]
Data Source: S. cerevisiae from Saccharomyces Genome Database (SGD)
Reports & Analysis: Download Sequence File [Configure...] [Go]
Scroll/Zoom: Show 5 kbp □ Flip

⊟ Tracks
⊟ Overview □ All on □ All off
☑ Landmarks
⊟ Analysis □ All on □ All off
□ Restriction Sites
⊟ General □ All on □ All off
□ 3-frame translation (forward) □ DNA/GC Content □ Protein-coding Genes □ Telomere, Centromere and ARS elements
□ 3-frame translation (reverse) □ Genomic Clones □ ResGen primer products □ Transposons & Long Terminal Repeats
☑ All Annotated Sequence Features □ Non-coding RNA Genes □ ResGen primers
⊟ Polymerase occupancy □ All on □ All off
□ Pol II distribution - Steinmetz, et al, (2006)
⊟ Regulatory regions & binding sites □ All on □ All off
□ Harbison, et al (2004) □ Tachibana, et al (2005) □ uORFs
⊟ Transcript Information □ All on □ All off
□ cDNA transcripts - Miura, et al (2006) □ SAGE tags - Velculescu et al (1997) □ Transcription start sites - Zhang and Dietrich (2005)

⊟ Overview
Overview of chrII
0k
100k
200k
300k
400k
500k
600k
700k
800k
Landmarks
ILS1
MCM2
CEN2
RAD16
SUP45
MET8
⊟ Details
140k
141k
142k
143k
144k
All Annotated Sequence Features
YBL042C
FUI1, Verified, High affinity uridine permease
YBL040C
ERD2, Verified, HDEL receptor
YBL039W-B
Uncharacterized, Putative protein of unknown fu
YBL041W
PRE7, Verified, Beta 6 subunit of the 20S proteasome
YBL039C
URA7, Verified, Major CTP synth
YBL03
Dubio
Clear highlighting
Update Image
Configure tracks...
Update Image

Figure 11.4.6 A view of a portion of *S. cerevisiae* chromosome II in SGD's version of the GBrowse genome browser. GBrowse allows queries using the Landmark or Region search box, download of DNA sequences using the Reports & Analysis pull-down, adjustment of the viewing window using the Scroll/Zoom menu, and customization of the tracks shown in the Details field.

section of genome shown in the Details section is narrow, the red rectangle may appear as a single red line.

6. Adjust the size of the genomic region shown in the Details panel by using the Zoom controls in the Scroll/Zoom menu at the top right of the Overview panel (Fig. 11.4.6).

   The "−" and "+" buttons allow fine control of the zoom level. The Zoom drop-down menu allows different preset choices for zoom level, typically from 100 bp to 1.5 Mbp.

7. Scroll the genomic region shown in the Details panel left or right by using the Scroll controls in the Scroll/Zoom menu at the top right of the Overview panel (Fig. 11.4.6).

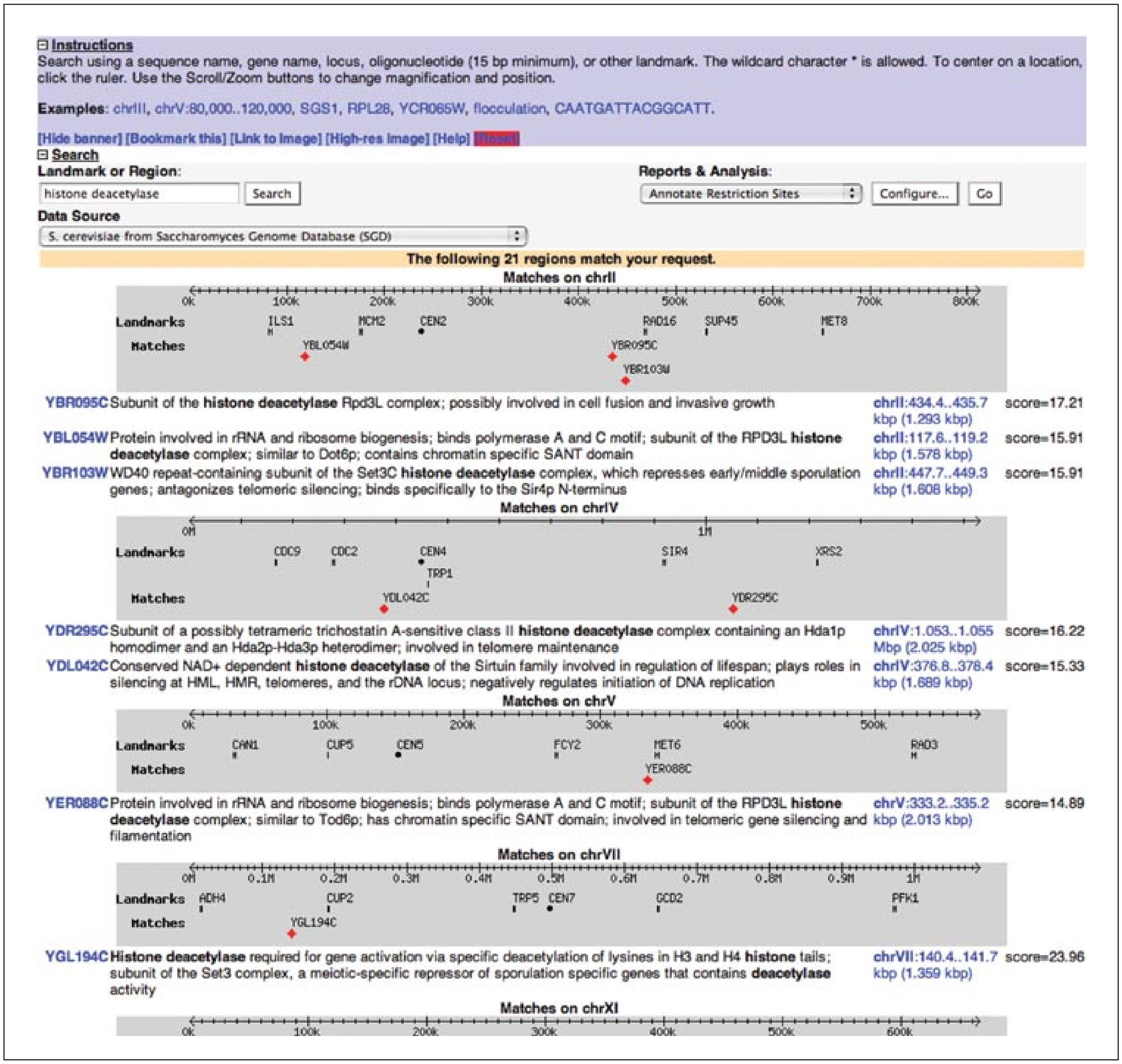
   The "≪" and "≫" buttons scroll the window one full length to the left or right. The "<" and ">" buttons scroll the window one-half length to the left or right.

8. Customize the Detail panel further by turning desired tracks on or off using the appropriate checkboxes (Fig. 11.4.6).

   The Details panel contains one or more tracks showing genomic features, clones, GC content, translation frames, transcription start sites, etc. The order in which the tracks are shown can be customized using the "Configure tracks" button. For more information, see Donlin (2007).

View and download a specific nucleotide sequence

9. To extract a particular nucleotide sequence, navigate the Details panel to the desired region of the genome.



⊟ **Instructions**

Search using a sequence name, gene name, locus, oligonucleotide (15 bp minimum), or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: chrIII, chrV:80,000..120,000, SGS1, RPL28, YCR065W, flocculation, CAATGATTACGGCATT.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ **Search**

Landmark or Region: histone deacetylase [Search]

Reports & Analysis: Annotate Restriction Sites [Configure...] [Go]

Data Source: S. cerevisiae from Saccharomyces Genome Database (SGD)

The following 21 regions match your request.

Matches on chrII

YBR095C Subunit of the **histone deacetylase** Rpd3L complex; possibly involved in cell fusion and invasive growth — chrII:434.4..435.7 kbp (1.293 kbp) — score=17.21

YBL054W Protein involved in rRNA and ribosome biogenesis; binds polymerase A and C motif; subunit of the RPD3L **histone deacetylase** complex; similar to Dot6p; contains chromatin specific SANT domain — chrII:117.6..119.2 kbp (1.578 kbp) — score=15.91

YBR103W WD40 repeat-containing subunit of the Set3C **histone deacetylase** complex, which represses early/middle sporulation genes; antagonizes telomeric silencing; binds specifically to the Sir4p N-terminus — chrII:447.7..449.3 kbp (1.608 kbp) — score=15.91

Matches on chrIV

YDR295C Subunit of a possibly tetrameric trichostatin A-sensitive class II **histone deacetylase** complex containing an Hda1p homodimer and an Hda2p-Hda3p heterodimer; involved in telomere maintenance — chrIV:1.053..1.055 Mbp (2.025 kbp) — score=16.22

YDL042C Conserved NAD+ dependent **histone deacetylase** of the Sirtuin family involved in regulation of lifespan; plays roles in silencing at HML, HMR, telomeres, and the rDNA locus; negatively regulates initiation of DNA replication — chrIV:376.8..378.4 kbp (1.689 kbp) — score=15.33

Matches on chrV

YER088C Protein involved in rRNA and ribosome biogenesis; binds polymerase A and C motif; subunit of the RPD3L **histone deacetylase** complex; similar to Tod6p; has chromatin specific SANT domain; involved in telomeric gene silencing and filamentation — chrV:333.2..335.2 kbp (2.013 kbp) — score=14.89

Matches on chrVII

YGL194C **Histone deacetylase** required for gene activation via specific deacetylation of lysines in H3 and H4 **histone** tails; subunit of the Set3 complex, a meiotic-specific repressor of sporulation specific genes that contains **deacetylase** activity — chrVII:140.4..141.7 kbp (1.359 kbp) — score=23.96

Matches on chrXI

Figure 11.4.7 An example of a text search (“histone deacetylase”) in SGD’s version of GBrowse that generates multiple results.

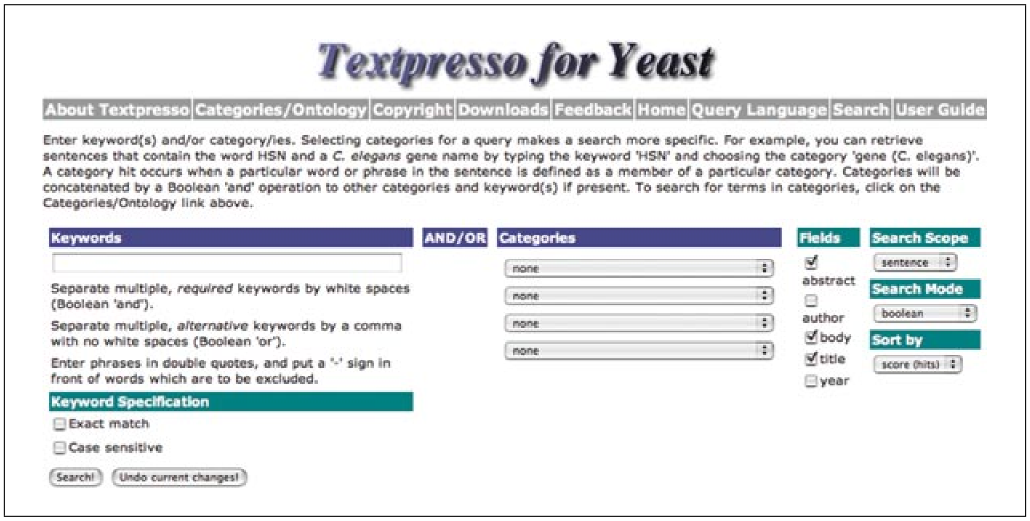
10. In the Reports & Analysis menu at the top right of the Overview panel, select Download Sequence File, and click the Go button (Fig. 11.4.6).

    The sequence of the region currently displayed in the Details panel will be downloaded to the Web browser. To save the sequence to file, select ”Save as” in the Web browser’s File menu, or from the keyboard use Ctrl-S (PC) or Apple-S (Mac). To download the sequence file in different formats, click the Configure button instead to select desired text output options, such as text or XML/HTML, or specific sequence file formats, such as FASTA, GCG, or GFF3. For more information, see Donlin (2007) or the online help at http://gmod.org/wiki/Gbrowse.

BASIC PROTOCOL 3

USING TEXTPRESSO TO SEARCH FULL TEXT PAPERS

Many MODs provide capacity for full-text literature searches using the Textpresso text-mining system designed under the GMOD project specifically for scientific literature (Müller et al., 2004). Textpresso combines full-text keyword or phrase searching with the use of categories to impart semantic context to the queries, allowing users to more efficiently uncover relevant information of interest.



Enter keyword(s) and/or category/ies. Selecting categories for a query makes a search more specific. For example, you can retrieve sentences that contain the word HSN and a *C. elegans* gene name by typing the keyword 'HSN' and choosing the category 'gene (C. elegans)'. A category hit occurs when a particular word or phrase in the sentence is defined as a member of a particular category. Categories will be concatenated by a Boolean 'and' operation to other categories and keyword(s) if present. To search for terms in categories, click on the Categories/Ontology link above.

Separate multiple, *required* keywords by white spaces (Boolean 'and').

Separate multiple, *alternative* keywords by a comma with no white spaces (Boolean 'or').

Enter phrases in double quotes, and put a '-' sign in front of words which are to be excluded.

□ Exact match

□ Case sensitive

none
none
none
none

☑ abstract

□ author

☑ body

☑ title

□ year

Search Scope: sentence

Search Mode: boolean

Sort by: score (hits)

Figure 11.4.8 Entry page in SGD for the Textpresso text-mining system.

1. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).

2. Click on the Textpresso link at the top left of the home page to open the Textpresso search engine.

3. Enter a word, phrase, gene or protein name(s), author name(s) etc., into the Keywords box and click the Search button (Fig. 11.4.8).

   Text strings used for searching in Textpresso can be single words, collections of words, or longer phrases and full sentences. Boolean operators are also supported. One or more optional predefined categories (such as allele, disease, gene, metabolic process, phenotype, etc.) may be selected to make the query more specific. Text fields (e.g., abstract, author, body, title, and year) to be searched can also be selected as the user considers appropriate. Queries can be further refined using checkboxes for "Exact match" and "Case sensitive."

4. Examine the results of the search (Fig. 11.4.9).

   Results include the number of text hits and matched documents (e.g., "102 matches in 96 documents"). Several pieces of information are provided for each matched document, including title, author(s), journal citation, year, publication type, PubMed ID, abstract, and matching sentences with search terms highlighted. Also provided are links to open the full-text of the document, search for related articles in PubMed, or download the reference information in EndNote format. If desired, display options can be customized to show only a subset of these different types of information.

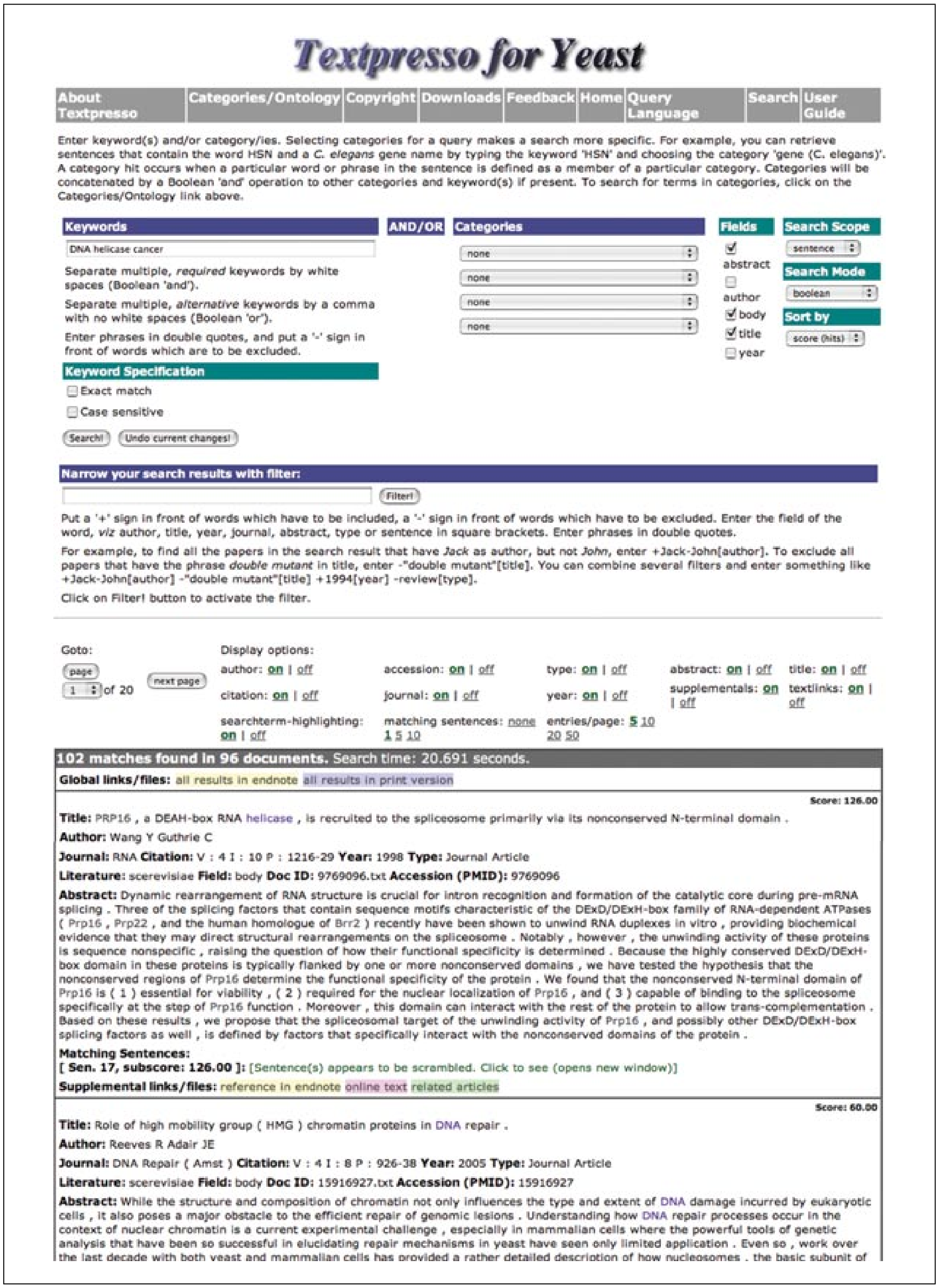
   a. *If there are too many results:* Narrow your search results using the filter.

      Add keywords to the original query using a "+" sign; use a "−" sign to indicate other words that should be excluded from the results. Enter the appropriate field after the word in square brackets (e.g., "+TOM7[abstract]"). Click the Filter button.

   b. *If there are too few results:* Use different or a smaller number of keywords in your search, and/or use fewer specified categories.

5. When reasonable search results are obtained, follow the appropriate links to open and read the full text of the extracted papers.

   Users are always encouraged to read the primary literature themselves rather than relying solely on phrases returned in search results.



Enter keyword(s) and/or category/ies. Selecting categories for a query makes a search more specific. For example, you can retrieve sentences that contain the word HSN and a *C. elegans* gene name by typing the keyword 'HSN' and choosing the category 'gene (C. elegans)'. A category hit occurs when a particular word or phrase in the sentence is defined as a member of a particular category. Categories will be concatenated by a Boolean 'and' operation to other categories and keyword(s) if present. To search for terms in categories, click on the Categories/Ontology link above.

Keywords: DNA helicase cancer

Separate multiple, *required* keywords by white spaces (Boolean 'and').

Separate multiple, *alternative* keywords by a comma with no white spaces (Boolean 'or').

Enter phrases in double quotes, and put a '-' sign in front of words which are to be excluded.

AND/OR

Categories: none; none; none; none

Fields: ☑ abstract ☐ author ☑ body ☑ title ☐ year

Search Scope: sentence

Search Mode: boolean

Sort by: score (hits)

Keyword Specification

☐ Exact match

☐ Case sensitive

Narrow your search results with filter:

Put a '+' sign in front of words which have to be included, a '-' sign in front of words which have to be excluded. Enter the field of the word, *viz* author, title, year, journal, abstract, type or sentence in square brackets. Enter phrases in double quotes.

For example, to find all the papers in the search result that have *Jack* as author, but not *John*, enter +Jack-John[author]. To exclude all papers that have the phrase *double mutant* in title, enter -"double mutant"[title]. You can combine several filters and enter something like +Jack-John[author] -"double mutant"[title] +1994[year] -review[type].

Click on Filter! button to activate the filter.

Goto:

1 of 20

Display options:

author: on | off   accession: on | off   type: on | off   abstract: on | off   title: on | off

citation: on | off   journal: on | off   year: on | off   supplementals: on | off   textlinks: on | off

searchterm-highlighting: on | off   matching sentences: none 1 5 10   entries/page: 5 10 20 50

102 matches found in 96 documents. Search time: 20.691 seconds.

Global links/files: all results in endnote all results in print version

Score: 126.00

Title: PRP16 , a DEAH-box RNA helicase , is recruited to the spliceosome primarily via its nonconserved N-terminal domain .

Author: Wang Y Guthrie C

Journal: RNA **Citation:** V : 4 I : 10 P : 1216-29 **Year:** 1998 **Type:** Journal Article

Literature: scerevisiae **Field:** body **Doc ID:** 9769096.txt **Accession (PMID):** 9769096

Abstract: Dynamic rearrangement of RNA structure is crucial for intron recognition and formation of the catalytic core during pre-mRNA splicing . Three of the splicing factors that contain sequence motifs characteristic of the DExD/DExH-box family of RNA-dependent ATPases (Prp16 , Prp22 , and the human homologue of Brr2) recently have been shown to unwind RNA duplexes in vitro , providing biochemical evidence that they may direct structural rearrangements on the spliceosome . Notably , however , the unwinding activity of these proteins is sequence nonspecific , raising the question of how their functional specificity is determined . Because the highly conserved DExD/DExH-box domain in these proteins is typically flanked by one or more nonconserved domains , we have tested the hypothesis that the nonconserved regions of Prp16 determine the functional specificity of the protein . We found that the nonconserved N-terminal domain of Prp16 is (1) essential for viability , (2) required for the nuclear localization of Prp16 , and (3) capable of binding to the spliceosome specifically at the step of Prp16 function . Moreover , this domain can interact with the rest of the protein to allow trans-complementation . Based on these results , we propose that the spliceosomal target of the unwinding activity of Prp16 , and possibly other DExD/DExH-box splicing factors as well , is defined by factors that specifically interact with the nonconserved domains of the protein .

Matching Sentences:
[Sen. 17, subscore: 126.00]: [Sentence(s) appears to be scrambled. Click to see (opens new window)]

Supplemental links/files: reference in endnote online text related articles

Score: 60.00

Title: Role of high mobility group (HMG) chromatin proteins in DNA repair .

Author: Reeves R Adair JE

Journal: DNA Repair (Amst) **Citation:** V : 4 I : 8 P : 926-38 **Year:** 2005 **Type:** Journal Article

Literature: scerevisiae **Field:** body **Doc ID:** 15916927.txt **Accession (PMID):** 15916927

Abstract: While the structure and composition of chromatin not only influences the type and extent of DNA damage incurred by eukaryotic cells , it also poses a major obstacle to the efficient repair of genomic lesions . Understanding how DNA repair processes occur in the context of nuclear chromatin is a current experimental challenge , especially in mammalian cells where the powerful tools of genetic analysis that have been so successful in elucidating repair mechanisms in yeast have seen only limited application . Even so , work over the last decade with both yeast and mammalian cells has provided a rather detailed description of how nucleosomes , the basic subunit of

Figure 11.4.9 The top portion of a results page in SGD's version of Textpresso for the query "DNA helicase cancer."

COMMENTARY

Understanding Results

The assimilation of vast amounts and different varieties of information within MODs provides rich context for researchers' results, allowing more in-depth interpretation and improved experimental design. The increased use of large-scale genomic and proteomic technologies means that more scientists are coming to SGD, or other MODs, sequence in hand, trying to make sense of information gleaned from expression profiles, suppressor screens, chromatin immunoprecipitation data, single nucleotide polymorphism data, etc. For example, Elizabeth is a researcher studying recombination in yeast. She performs a genetic screen and obtains a sequence that suppresses a topoisomerase deficiency. Using BLAST at SGD (see *UNIT 11.1*), Elizabeth pulls up a list of similar sequences, and then follows a link for the top hit to a gene summary page. From this gene summary page, Elizabeth learns that the sequence is that of a DNA helicase involved in DNA replication, meiotic chromosome segregation, replicative cell aging, and resistance to UV irradiation and DNA-damaging chemicals. Elizabeth also finds links to orthology sets including genes in other yeasts, plants, and mammals, as well as information indicating that mutations in the human orthologs have been implicated in cancer and premature-aging syndromes. To find other DNA helicases in yeast, or other genes involved in similar biological processes, Elizabeth follows Basic Protocol 1 (steps 12 to 14). To extract the entire coding region of the DNA helicase, plus flanking sequences, she uses GBrowse (Basic Protocol 2). Finally, to find more literature regarding the role of this helicase in DNA replication and cell aging in yeast, plus information regarding its use as a model gene for the study of human cancers and aging, Elizabeth uses Textpresso (Basic Protocol 3).

Troubleshooting

Upon visiting a new MOD for the first time, users may be thwarted by the lack of an obvious place to begin exploring, overwhelmed by busy pages, or stymied by unfamiliar interfaces. This need not be the case. Because all MODs provide at least a basic search box, to get started a new user can simply try a basic query for a gene name, author name, biological term, etc. (see Basic Protocol 1). This will provide entry into the database, and begin familiarizing the user with the MOD's look and feel. Searching for a gene or cellular component common to most organisms, like "actin" or "ribosome," will lead a user to a gene summary page or list of search results that can then be investigated further. If a page is too busy or visually overwhelming, users can use the Web browser's find-in-page function (Ctrl-F on a PC, Apple-F on a Mac) to locate familiar keywords such as "function," "phenotype," "literature," "sequence," etc. Complicated interfaces are usually accompanied by explanatory documentation (do a find-in-page for "help"), or will provide links to download data from an FTP site or other Web-based mirror site (e.g., *http://downloads.yeastgenome.org*). Lastly, if all else fails, users should not just leave the site; instead, do a find-in-page for "contact" and send an e-mail to the MOD's curators. The friendly and responsive curatorial staff is there not only to populate the database with skillfully organized scientific information, but also to help users use it.

Variations

Readers should remember that while many MODs use similar software and database structures, all MODs are unique and provide differing amounts and types of biological information, as well as different visualization and analysis tools. More detailed descriptions and protocols relating to these specific databases and related tools can be found in *Current Protocols in Bioinformatics*: GBrowse (Donlin, 2007), MGI (Shaw, 2009), RGD (Twigger et al., 2006), TAIR (Reiser and Rhee, 2005), WormBase (Schwarz and Sternberg, 2006), and the UCSC Genome Browser (Karolchik et al., 2007). Readers are encouraged to familiarize themselves with their favorite MODs by perusing their various Web pages, reading the online help documentation provided by the MOD, and contacting the MOD directly via e-mail with any specific questions. Some of the more widely used MODs are listed below in Internet Resources.

Literature Cited

Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Binkley, G., Lane, C., Miyasato, S.R., and Sherlock, G. 2005. The *Candida* Genome Database (CGD), a community resource for *Candida albicans* gene and protein information. *Nucleic Acids Res.* 33:D358-D363.

Arnaud, M.B., Costanzo, M.C., Skrzypek, M.S., Shah, P., Binkley, G., Miyasato, S.R., and Sherlock, G. 2009. *Aspergillus* Genome Database *http://www.aspergillusgenome.org/* (April 8, 2009).

Blake, J.A., and Harris, M.A. 2008. The Gene Ontology (GO) Project: Structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr. Protoc. Bioinform.* 23:7.2.1-7.2.9.

Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and the Mouse Genome Database Group. 2009. The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res.* 37:D712-D719.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* 26:73-79.

Donlin, M.J. 2007. Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinform.* 17:9.9.1-9.9.24.

Dwight, S.S., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dolinski, K., Engel, S.R., Feierbach, B., Fisk, D.G., Hirschman, J., Hong, E.L., Issel-Tarver, L., Nash, R.S., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Weng, S., Botstein, D., and Cherry, J.M. 2004. Saccharomyces genome database: Underlying principles and organisation. *Brief. Bioinform.* 5:9-22.

Gelbart, W.M., Crosby, M., Matthews, B., Rindone, W.P., Chillemi, J., Russo Twombly, S., Emmert, D., Ashburner, M., Drysdale, R.A., Whitfield, E., Millburn, G.H., de Grey, A., Kaufman, T., Matthews, K., Gilbert, D., Strelets, V., and Tolstoshev, C. 1997. FlyBase: A *Drosophila* Database. *Nucleic Acids Res.* 25:63-66.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. 1996. Life with 6000 genes. *Science* 274:546-567.

Harris, M., and the Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res.* 36:440-444.

Karolchik, D., Hinrichs, A.S., and Kent, W.J. 2007. The UCSC Genome Browser. *Curr. Protoc. Bioinform.* 17:1.4.1-1.4.24.

Kreppel, L., Fey, P., Gaudet, P., Just, E., Kibbe, W.A., Chisholm, R.L., and Kimmel, A.R. 2004. dictyBase: A new *Dictyostelium discoideum* genome database. *Nucleic Acids Res.* 2004 32:D332-D333.

Müller, H.M., Kenny, E.E., and Sternberg, P.W. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2:e309.

Ono, B.I., Hazu, T., Yoshida, S., Kawato, T., Shinoda, S., Brzvwczy, J., and Paszewski, A. 1999. Cysteine biosynthesis in *Saccharomyces cerevisiae*: A new outlook on pathway and regulation. *Yeast* 15:1365-1375.

Reiser, L. and Rhee, S.Y. 2005. Using the *Arabidopsis* Information Resource (TAIR) to find information about *Arabidopsis* genes. *Curr. Protoc. Bioinform.* 9:1.11.1-1.11.45.

Rhee, S.Y. 2000. Bioinformatic resources, challenges, and opportunities using *Arabidopsis thaliana* as a model organism in post-genomic era. *Plant Physiol.* 124:1460-1464.

Schwarz, E.M. and Sternberg, P.W. 2006. Searching WormBase for information about *Caenorhabditis elegans*. *Curr. Protoc. Bioinform.* 14:1.8.1-1.8.43.

Shaw, D. 2009. Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr. Protoc. Bioinform.* 25:1.7.1-1.7.14.

Stein, L.D., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. 2001. WormBase: Network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 29:82-86.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* 12:1599-1610.

Twigger, S.N., Smith, J.S., Zuniga-Meyer, A., and Bromberg, S.K. 2006. Exploring phenotypic data at the Rat Genome Database. *Curr. Protoc. Bioinform.* 14:1.14.1-1.14.27.

Twigger, S.N., Shimoyama, M., Bromberg, S., Kwitek, A.E., Jacob, H.J., and the RGD Team. 2007. The Rat Genome Database, update 2007 – Easing the path from disease to data and back again. *Nucleic Acids Res.* 35:D658-D662.

Internet Resources

http://www.agbase.msstate.edu

Resource for functional analysis of agricultural plant and animal gene products.

http://www.arabidopsis.org

The Arabidopsis Information Resource (TAIR): Database of genetic and molecular biology data for the plant Arabidopsis thaliana.

http://crfb.univ-mrs.fr/aniseed

Ascidian Network for InSitu Expression and Embryological Data (ANISEED): Database for Ciona intestinalis, C. savignyi, Halocynthia roretzi, and Phallusia mammillata.

http://agd.vital-it.ch

Ashbya Genome Database (AGD): Database of gene annotation and microarray data for Ashbya gossypii and Saccharomyces cerevisiae.

http://www.aspergillusgenome.org

Aspergillus Genome Database (AspGD): Resource for genomic sequence data and gene and protein information for Aspergilli.

http://bovinegenome.org

Database that integrates bovine genomics data with structural and functional annotations of genes and the genome.

http://www.candidagenome.org

Database that serves as a resource for genomic sequence data and gene and protein information for Candida albicans.

http://dictybase.org

Resource for the biology and genomics of the social amoeba Dictyostelium discoideum.

http://ecolihub.org

Centralized resource linking various E. coli online information services, databases, and Web sites.

http://flybase.org

Database of Drosophila genes and genomes.

http://gmod.org

Generic Model Organism Database (GMOD) project: Collection of open source software tools for creating genome-scale biological databases.

http://www.gramene.org

Data resource for comparative genome analysis in the grasses.

http://www.beebase.org

Hymenoptera Genome Database (BeeBase): Database of genes and genomes of Apis mellifera and Nasonia vitripennis.

http://www.informatics.jax.org

Mouse Genome Informatics: Resource for the laboratory mouse, providing genetic, genomic, and biological data for the study of human health and disease.

http://paramecium.cgm.cnrs-gif.fr

Database of genomic sequence and genetic data for Paramecium tetraurelia.

http://rgd.mcw.edu

Rat Genome Database (RGD): Database of laboratory rat genetic and genomic data, including information for quantitative trait loci, mutations, and phenotypes.

http://www.yeastgenome.org

Saccharomyces Genome Database (SGD): Scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae.

http://www.genedb.org/genedb/pombe

Schizosaccharomyces pombe GeneDB: Database of genetic features, functional annotations, and other information for fission yeast.

http://smedgd.neuro.utah.edu

Schmidtea mediterranea Genome Database (SmedGD): Database for information associated with the planarian genome.

http://www.textpresso.org

Text-mining system for scientific literature.

http://wfleabase.org

Web service that provides gene and genomic information for species of the genus Daphnia, commonly known as the water flea.

http://www.wormbase.org

Biology and genomic information for Caenorhabditis species.

http://zfin.org

Zebrafish Information Network: Database for the molecular biology and genetics of zebrafish.