Using Model Organism Databases (MODs)

Stacia R. Engel[1] and Kevin A. MacPherson[1]

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

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.

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

How to cite this article:
Engel, S.R., and MacPherson, K.A. 2016. Using model organism databases (MODs). *Curr. Protoc. Essential Lab. Tech.* 13:11.4.1-11.4.22.
doi: 10.1002/cpet.4

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

Current Protocols Essential Laboratory Techniques 11.4.1-11.4.22, November 2016
Published online November 2016 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpet.4

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 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., 2010). Other well-known MODs include The Arabidopsis Information Resource (TAIR; Rhee, 2000; Lamesch et al., 2010), Mouse Genome Informatics (MGI; Blake et al., 2009; Shaw, 2009), the Rat Genome Database (RGD; Twigger et al., 2007; Karolchik et al., 2012), 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, if a single Web page indicates that a gene is expressed when a specific sugar is introduced into the growth medium, while also indicating that the protein encoded by that gene contains a DNA-binding domain, a researcher viewing this page could infer that the gene may encode a transcription factor that activates other genes whose products are 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 (Gene Ontology Consortium, 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 (Gene Ontology Consortium, 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/locus/S000003387/go*). 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 three tools found at many MODs: JBrowse, Textpresso, and InterMine. JBrowse and Textpresso were both 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). InterMine was developed by the Micklem lab at the University of Cambridge (Smith et al., 2012). The goal of both the GMOD and InterMine projects is to generate 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. JBrowse is a genome browser (Skinner et al., 2009), Textpresso is a full-text literature search application (Müller et al., 2004), and InterMine is a data warehousing system built for the integration and analysis of complex biological datasets (Smith et al., 2012).

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 JBrowse. Basic Protocol 3 describes how to perform a full-text literature search using Textpresso. Basic Protocol 4 explains how to analyze gene lists using YeastMine, an InterMine-based data warehouse.

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.

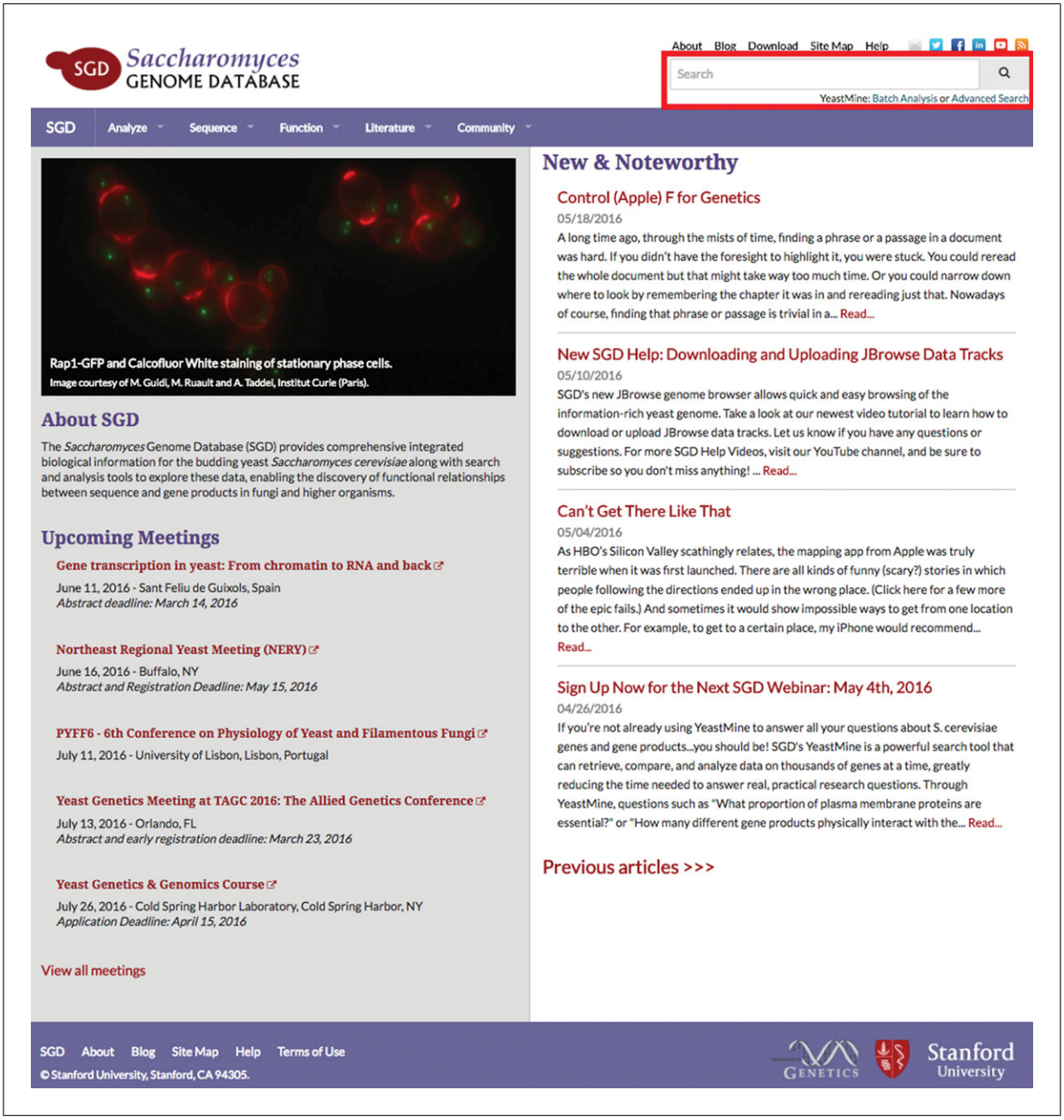

Figure 11.4.1 The SGD home page (*http://www.yeastgenome.org*), like most MOD home pages, is the main point of entry to the Web site. The home page lists news items and announcements, and provides links to different areas and tools provided by SGD. The database Search box and link to YeastMine are indicated by the red rectangle.

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

   Common MOD resources available in SGD are accessible via the toolbar that runs across the top of most SGD Web pages, and are divided into broad categories: Analyze, Sequence, Function, Literature, and Community (Fig. 11.4.1). Each category subtopic

listing is linked to the corresponding page within the SGD Web site. Popular tools are included, such as BLAST, Pattern Matching, Design Primers, Restriction Mapper, Genome Browser, Biochemical Pathways, and Full-text Search.

3. 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 the home page and most other Web pages that accesses a simple search.

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

6. If you do not find what you are looking for, you may wish to try alternative Search Options, such as an Advanced Search.

   Most MODs provide Advanced Search tools that can be used for more refined or complex searches (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.

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

9. 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. *Gene model(s) and nucleotide sequences:* A gene summary will typically display a listing of exons, introns, regulatory features, etc., with chromosomal and relative

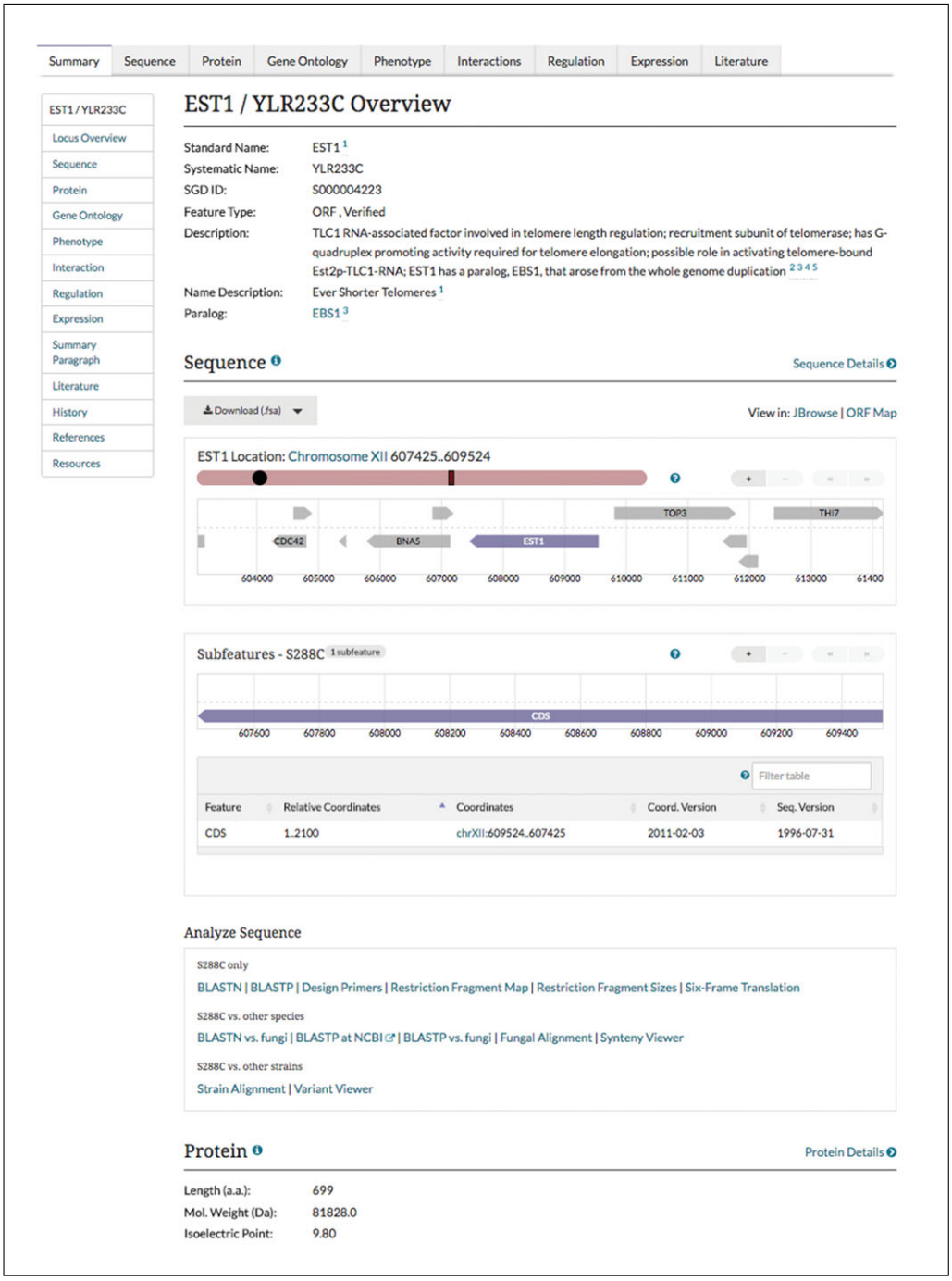

Figure 11.4.2 A locus summary page in SGD showing the different types of information included on a typical gene page. A portion of the page at the top that contains the SGD toolbar and Search box is not shown. A portion of the page at the bottom that contains the following sections is not shown: Gene Ontology, Phenotype, Interaction, Regulation, Expression, Summary Paragraph, Literature, History, References, and Resources.

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, coding and genomic sequences, and sequence analysis tools such as BLAST, FASTA, etc. Many MODs will also provide version dates for gene models and sequences, and/or a "History" section describing any past annotation changes to the gene model or sequence.

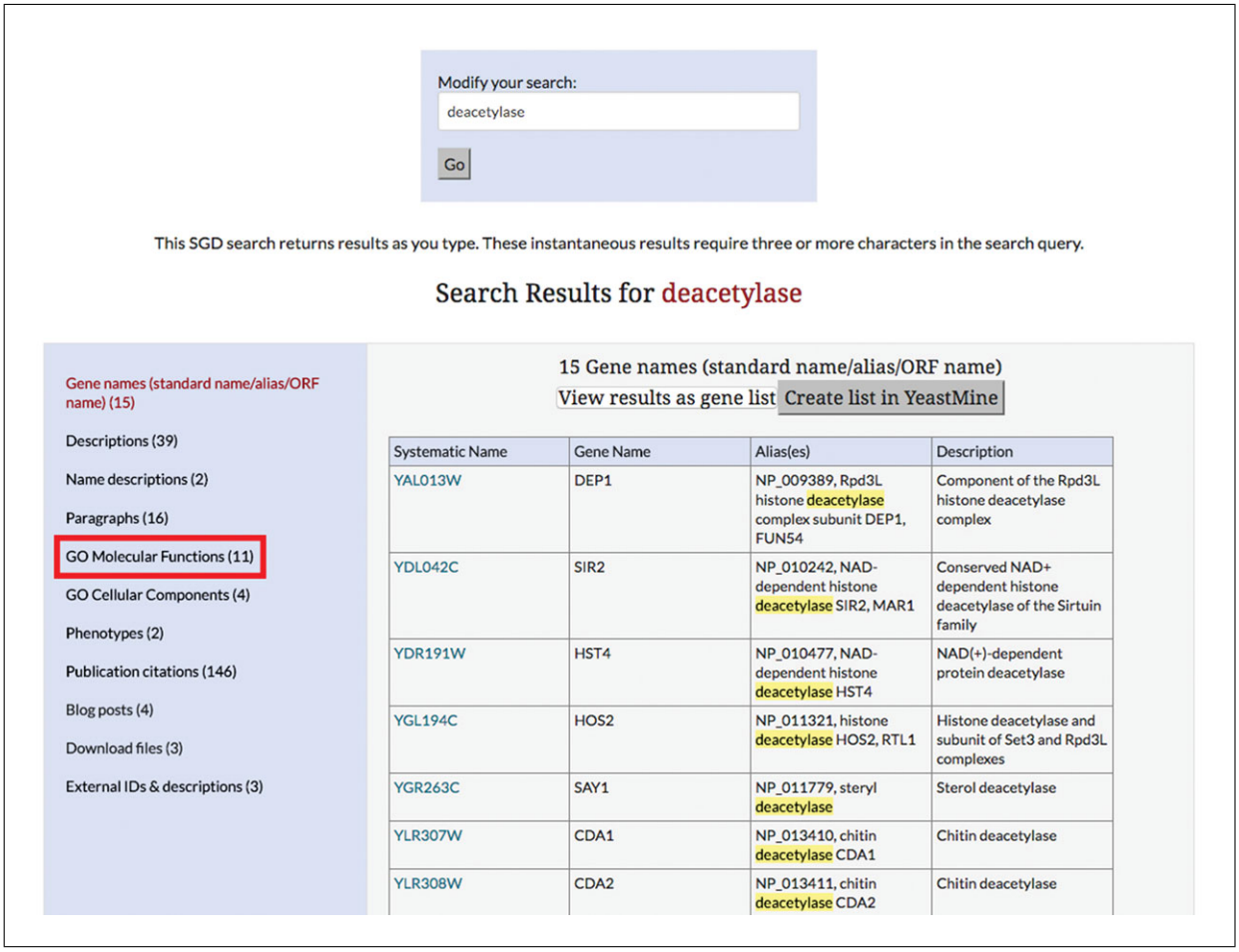

Figure 11.4.3 Search Results for gene/protein information and functional annotations in SGD from a simple search using the word "deacetylase." The GO Molecular Function category is indicated by the red rectangle.

c. *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.

d. *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.

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.

10. 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

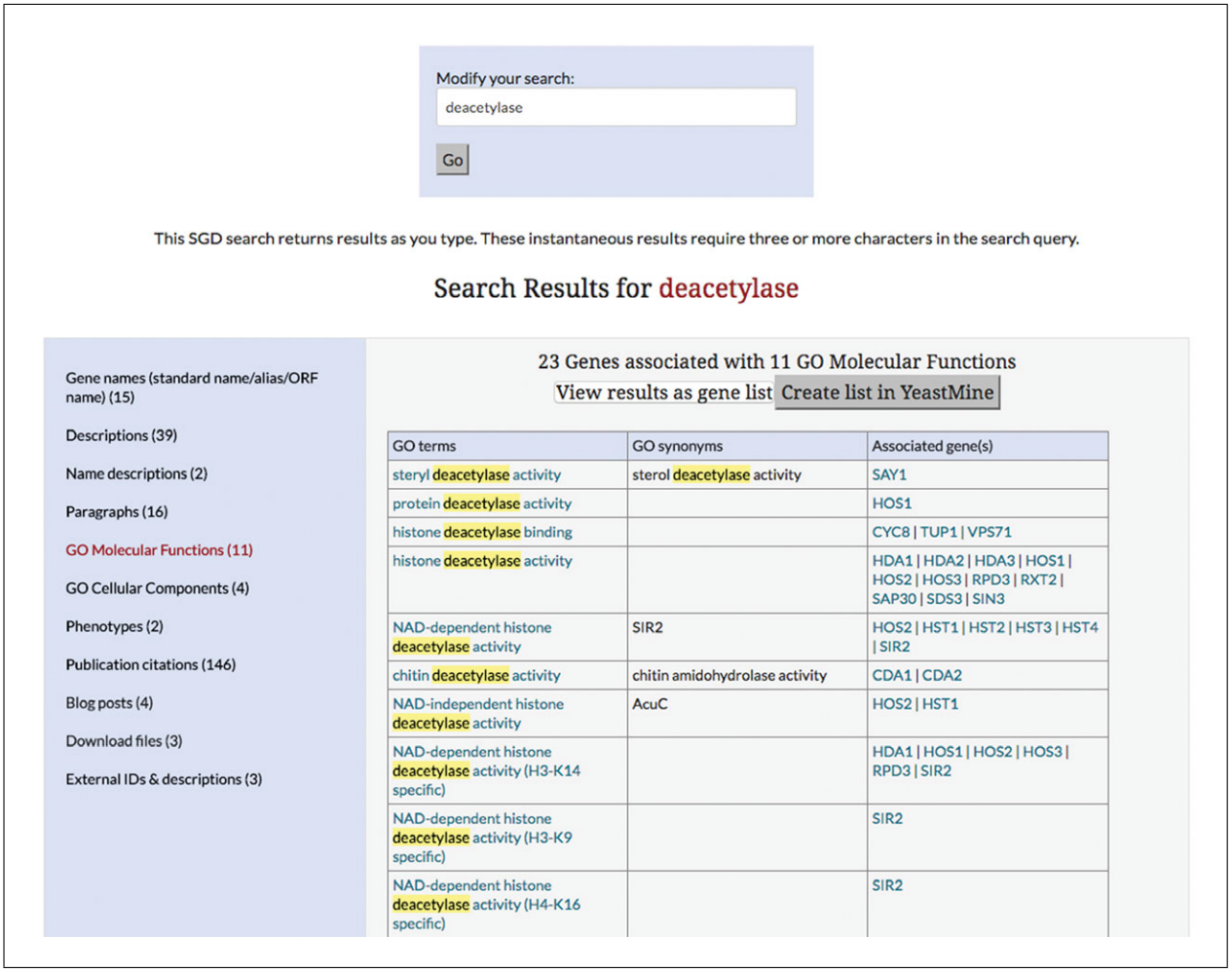

Figure 11.4.4 Search Results for Gene Ontology molecular function terms that contain the word “deacetylase.”

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).

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

12. 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.

13. 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 “deacetylase” returns a Search Results page listing hits in different categories that match the term “deacetylase” (Fig. 11.4.3).

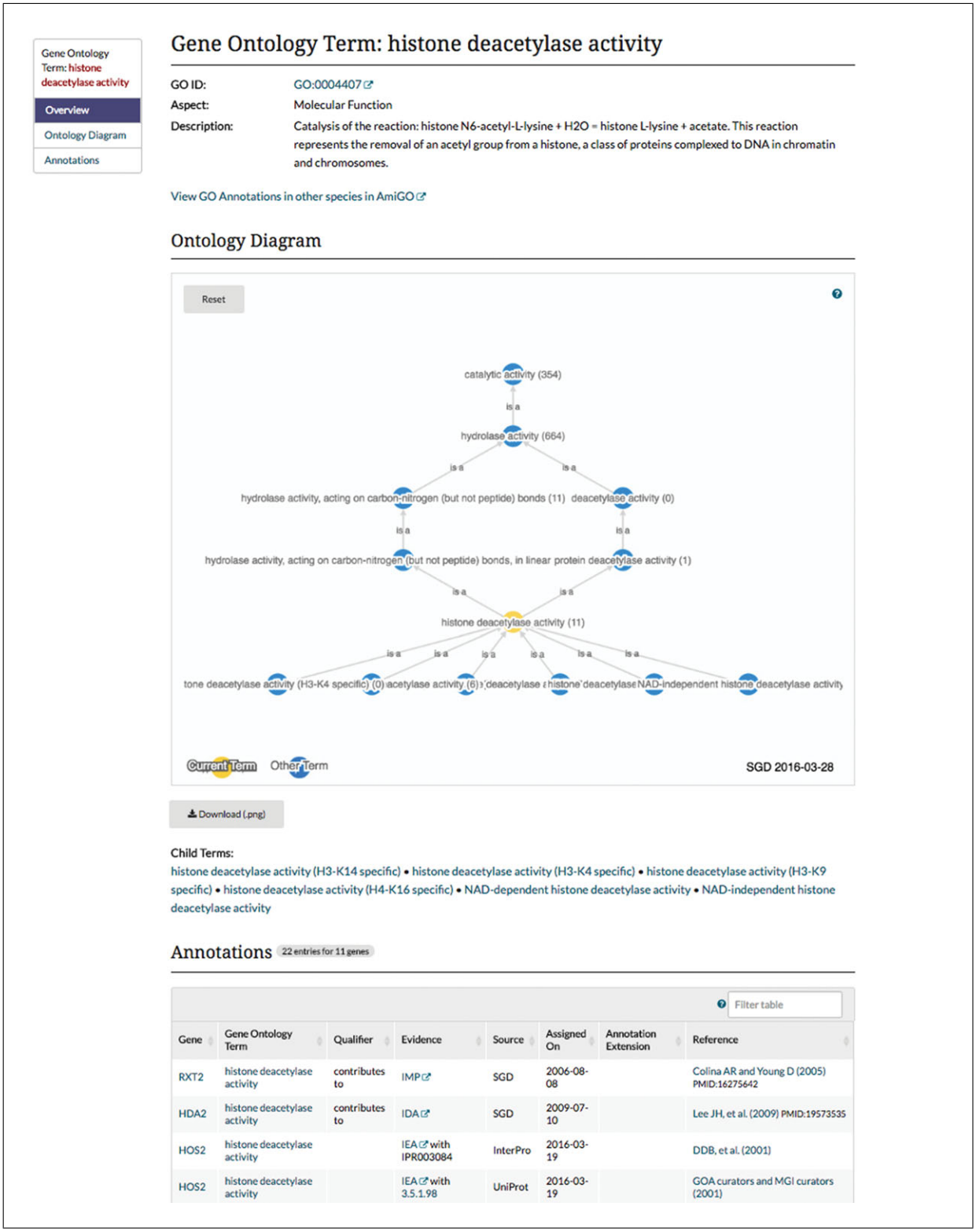

Figure 11.4.5 The top portion of the Gene Ontology detail page in SGD for the molecular function term "histone deacetylase activity."

Clicking on the "Molecular Functions" category displays the various Gene Ontology molecular function terms that contain the word "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).

OBTAINING A SEQUENCE FROM JBrowse

BASIC PROTOCOL 2

Many MODs display genomic features and other annotations using the JBrowse genome browser. JBrowse enables users to examine and scroll through genomic regions of choice, as well as view and download nucleotide sequences (Skinner et al., 2009).

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 Sequence menu in the toolbar that runs across the top of the home page. This menu contains links to various DNA and protein sequence analysis tools, such as BLAST and Gene/Sequence Resources, as well as JBrowse.
3. Click on the Genome Browser link to launch JBrowse (*http://www.yeastgenome.org/browse/;* Fig. 11.4.6).

   The JBrowse genome browser is also accessible via SGD gene summary pages by clicking on the JBrowse link in the Sequence section (Fig. 11.4.2).

4. To view a genomic region, enter a search term of choice in the search box near the top of the page and click 'Go' (Fig. 11.4.6). Users can search using a protein name, gene name, locus identifier, or other landmark.

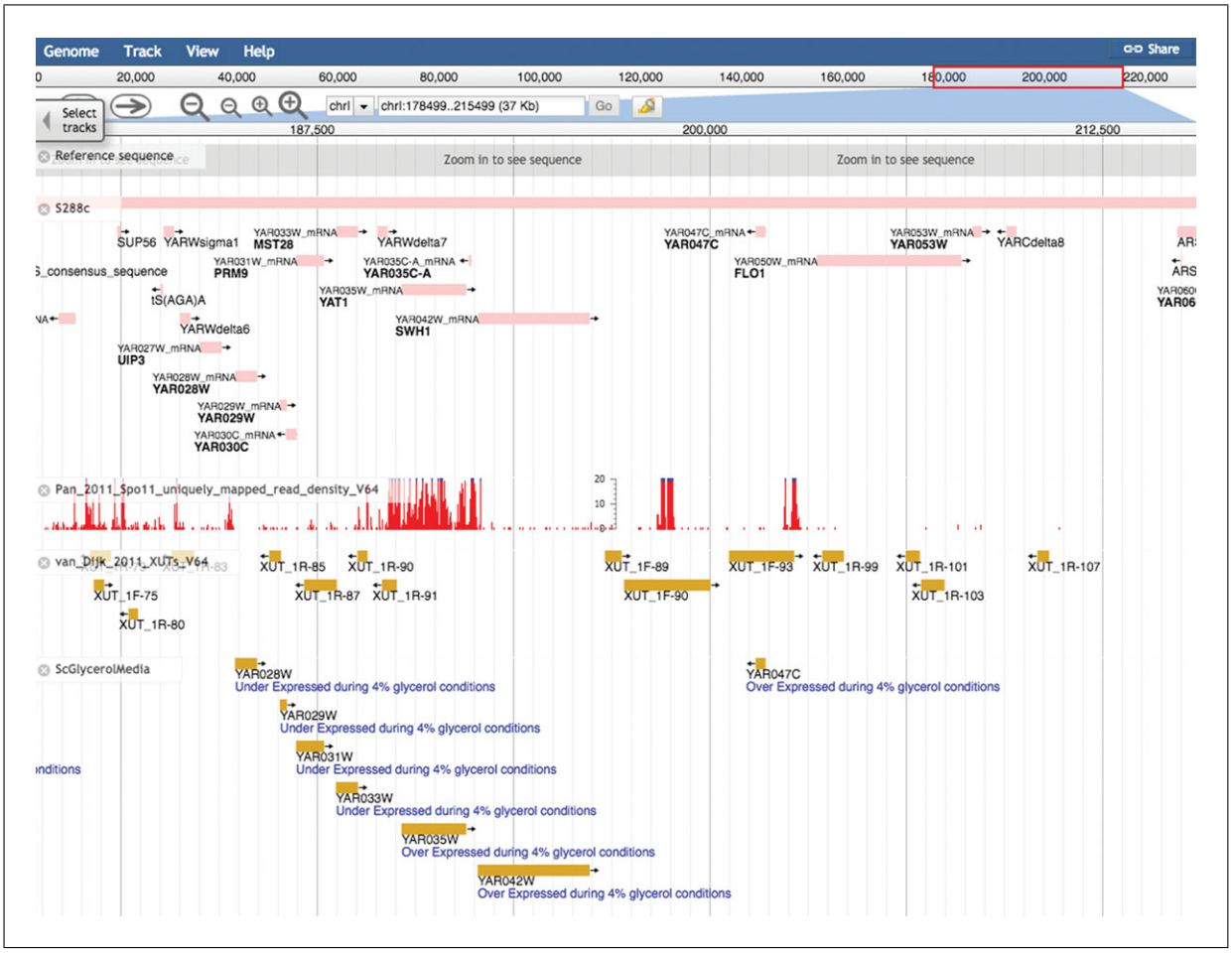

Figure 11.4.6 A view of a portion of *S. cerevisiae* chromosome I in SGD's version of the JBrowse genome browser. JBrowse allows queries using the region search box, download of DNA sequences, adjustment of the viewing window, and customization of the data tracks shown.

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, protein names, accession numbers, systematic locus identifiers, enzymatic activities, etc. Examples of searchable terms in SGD's JBrowse include chrIII, chrV:80,000..120,000, SGS1, YCR065W, ribosome, and sulfite reductase.

Examine a specific gene or region of the genome

5. Adjust the size of the genomic region shown in the main display by using the magnifying glass buttons in the top left portion of the page (Fig. 11.4.6).

   The "−" and "+" buttons allow fine control of the zoom level. For further control over the size and location of the genomic region displayed, click and drag the mouse cursor across a specific region in the overview bar (Fig. 11.4.6). Alternatively, click and drag your cursor over a region in the main display while holding down the shift button.

6. Scroll the genomic region shown in the main display panel left or right by using the arrow buttons located in the navigation bar at the top left of the page (Fig. 11.4.6).

   For finer scroll control, click and hold your mouse cursor anywhere on the main display and drag it to the left or right. Or, click and drag the red box in the overview bar (Fig. 11.4.6).

7. Customize the main display panel further by turning desired data tracks on or off using the appropriate checkboxes within the "Select Tracks" menu, accessible from the "Select tracks" button at top left (Fig. 11.4.6).

   The Select Tracks panel contains tracks that can be added to the main display. These tracks show genomic features, translation frames, transcription start sites, non-coding RNAs, nucleosome positions, etc. To customize the order in which tracks are shown, drag and drop the title bars on the left side of the page (Fig. 11.4.6).

View and download a specific nucleotide sequence

8. To download nucleotide sequence, open the configuration menu for the "Reference sequence" data track and select "Save track data" (Fig. 11.4.6).

   To save the sequence of the region currently displayed in the main panel, choose 'Visible region'. To download the sequence for the entire chromosome that contains the region currently displayed, choose 'Whole reference sequence'. The 'View' button opens a display window with the sequence in FASTA format—copy/paste into a text editor to save the sequence. Some Web browsers offer a 'Save' option, which will download the sequence directly without the need for a copy/paste step. For more information, see the online help at http://gmod.org/wiki/JBrowse.

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.

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

2. Open the Literature menu in the toolbar that runs across the top of the home page and click on the Full-text Search link to open the Textpresso search engine.

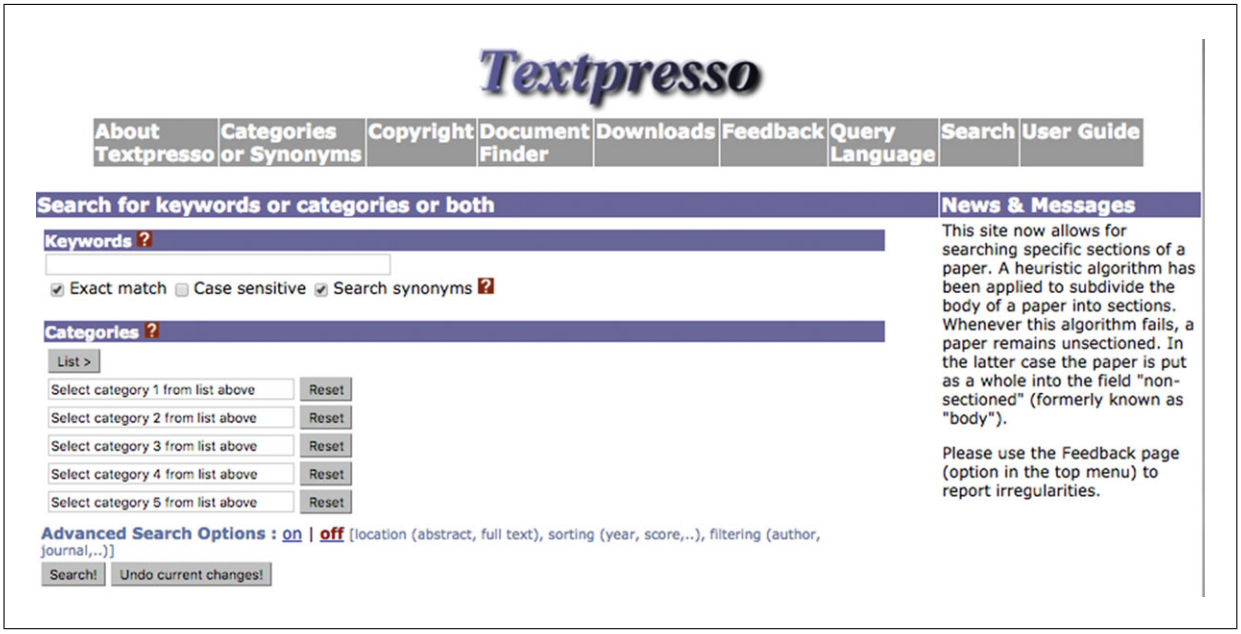

Figure 11.4.7 Entry page in SGD for the Textpresso text-mining system.

3. Enter a word, phrase, gene or protein name(s), author name(s) etc., into the Keywords box and click the Search button (Fig. 11.4.7).

   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." "Search synonyms" allows those sentences containing the synonyms of your search keywords to be returned, in addition to those containing the search keywords.

4. Examine the results of the search (Fig. 11.4.8).

   Results include the number of text hits and matched documents (e.g., "464 matches found in 102 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 or XML 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.

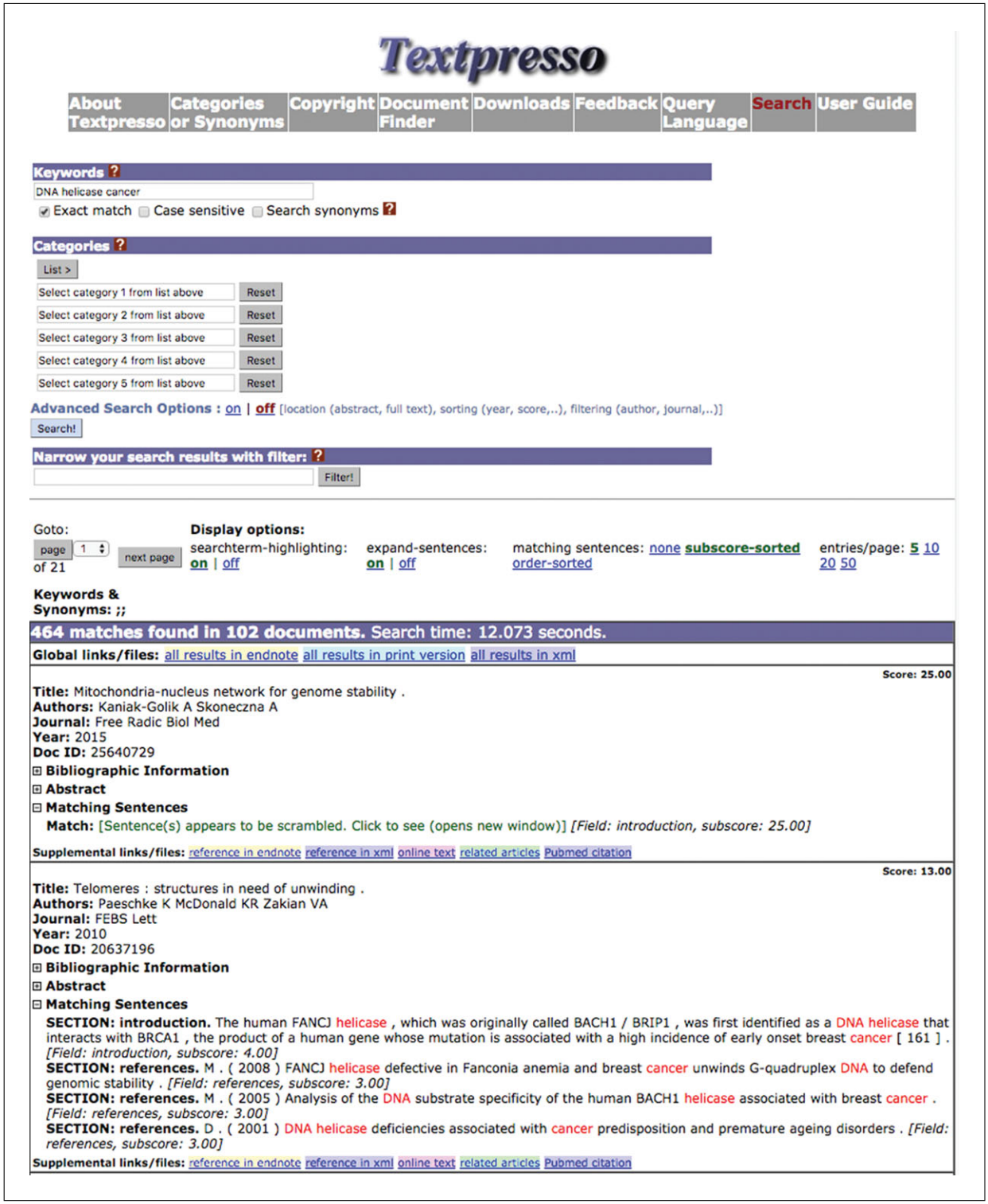

Figure 11.4.8 The top portion of a results page in SGD's version of Textpresso for the query "DNA helicase cancer."

BASIC PROTOCOL 4

Using Model Organism Databases (MODs)

USING InterMine TO PERFORM COMPLEX DATA QUERIES

Several MODs use the InterMine open-source data warehouse system (*http://www.intermine.org*) to create vast, searchable databases that integrate multiple types of data from different sources. These data warehouses provide powerful search and retrieval tools that enable users to access and analyze biological data sets through predefined templates, customizable Web queries, and user-created lists (Smith et al., 2012). There are "Mines" already in place for most model organisms, including yeast

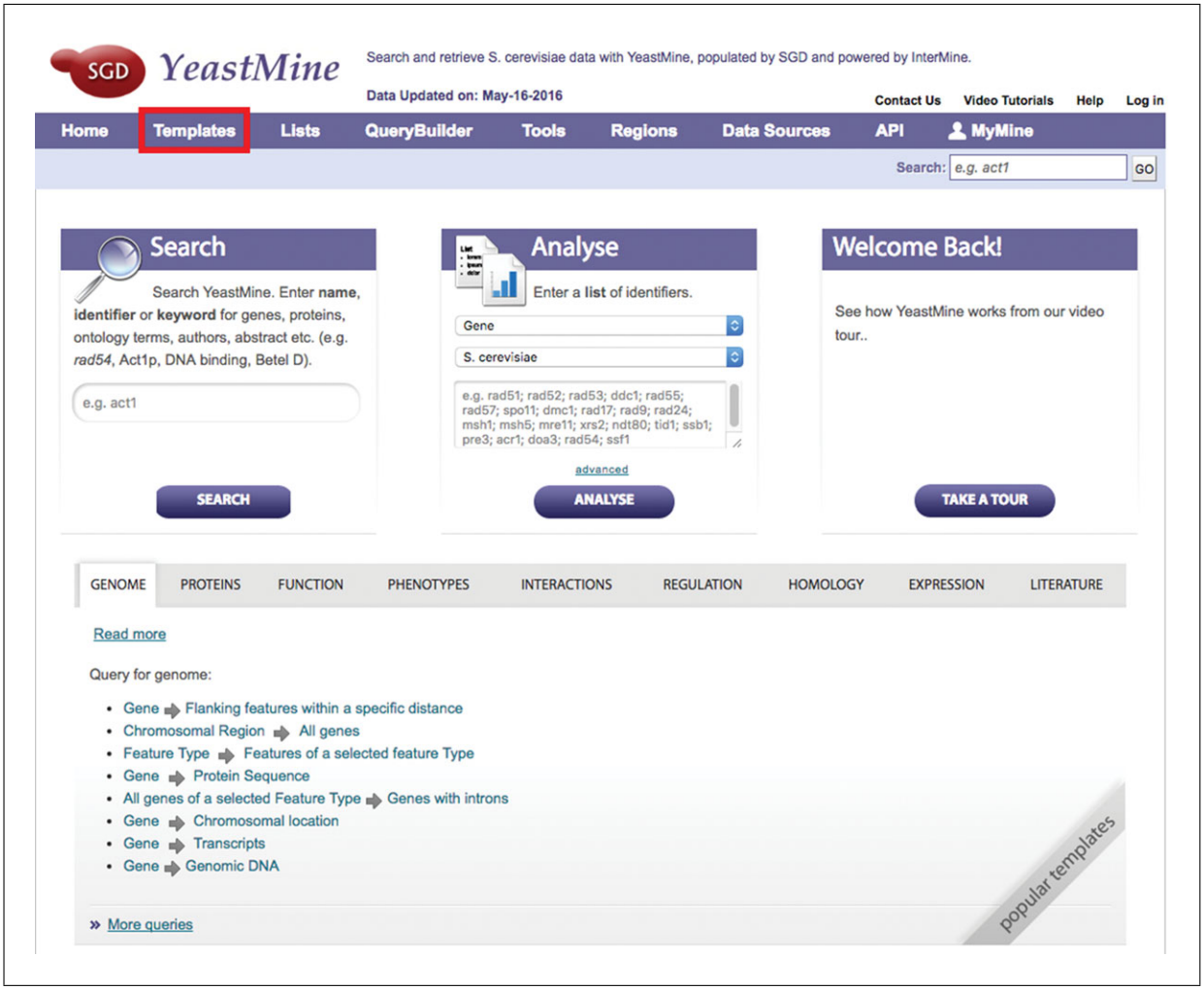

Figure 11.4.9 The top portion of the YeastMine homepage (*http://yeastmine.yeastgenome.org*). The Templates menu is indicated by the red rectangle.

(*http://yeastmine.yeastgenome.org*), worm (*http://intermine.wormbase.org*), fly (*http://www.flymine.org*), zebrafish (*http://www.zebrafishmine.org*), rat (*http://ratmine.mcw.edu*), and mouse (*http://www.mousemine.org*). These environments are separate from, but complementary to, the MODs, and allow navigation of different data types, exploration of features shared between gene products, and connections to information provided by other MODs. The ability to customize InterMine software allows MODs to incorporate data types that are common to other implementations of InterMine and then extend their own data models to provide additional curated details that are uniquely captured at each MOD. To demonstrate these functions, this protocol uses SGD's YeastMine as an example. YeastMine can be used to efficiently retrieve information such as Gene Ontology annotations, mutant phenotypes, genetic and physical interactions, expression, homology, and more. More information about the InterMine project, as well as links to a number of different InterMine-powered data warehouses for different organisms, can be found at the InterMine homepage (*http://www.intermine.org*).

Running a query in YeastMine

1. Open the SGD home page (*http://www.yeastgenome.org*) in a Web browser (Fig. 11.4.1).
2. Click on the YeastMine link at the top right on the home page to open the YeastMine data warehouse (Fig. 11.4.1).

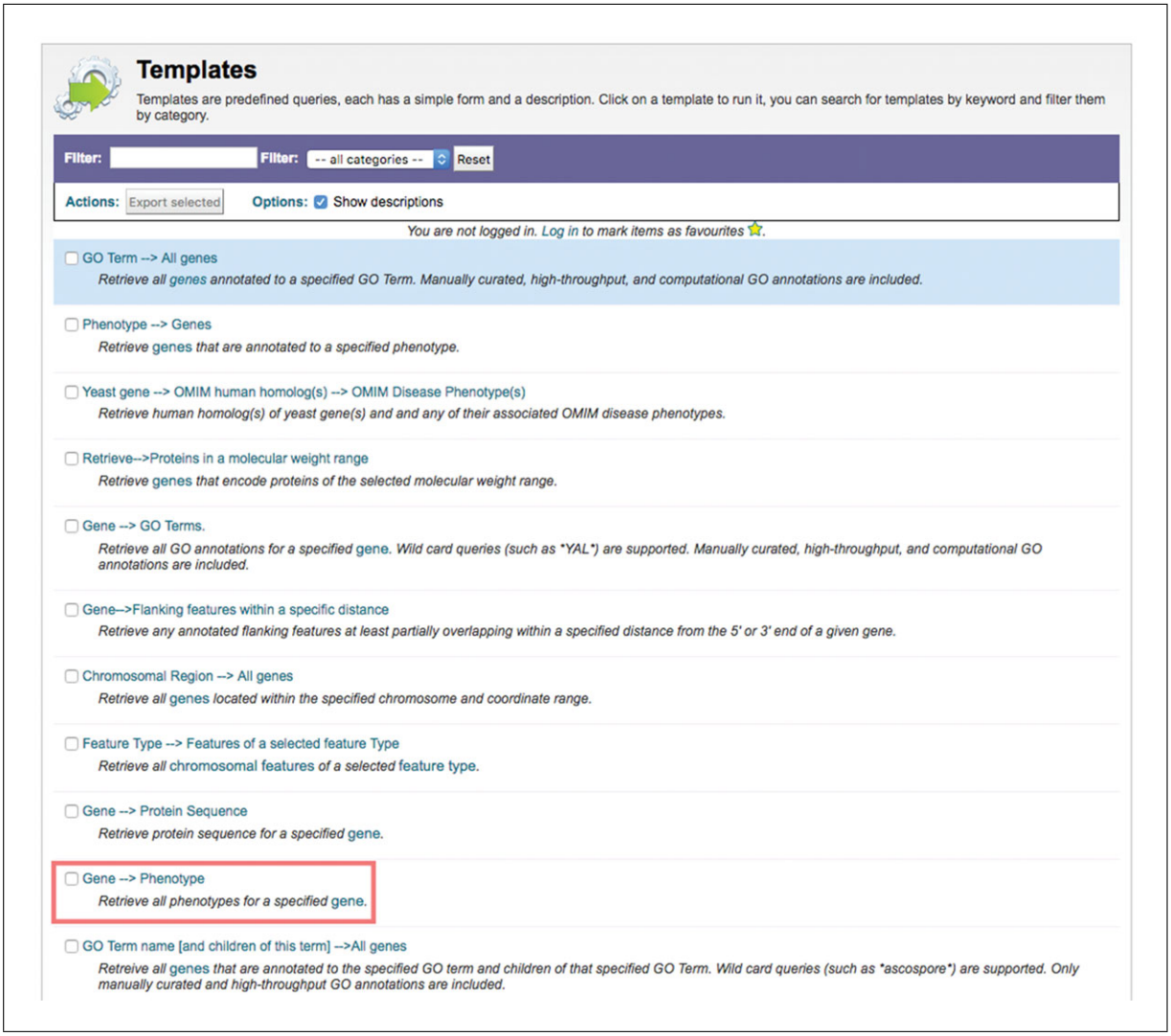

Figure 11.4.10 A portion of the Templates page in YeastMine. The query "Gene → Phenotype" is indicated by the red rectangle.

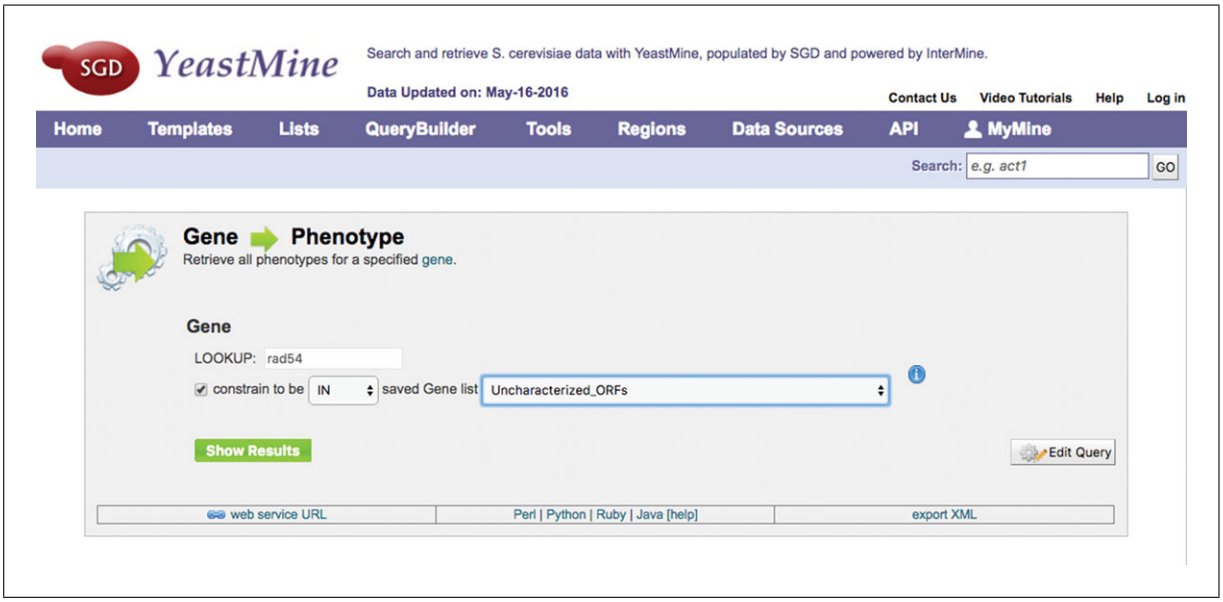

Figure 11.4.11 Gene → Phenotype query page in YeastMine.

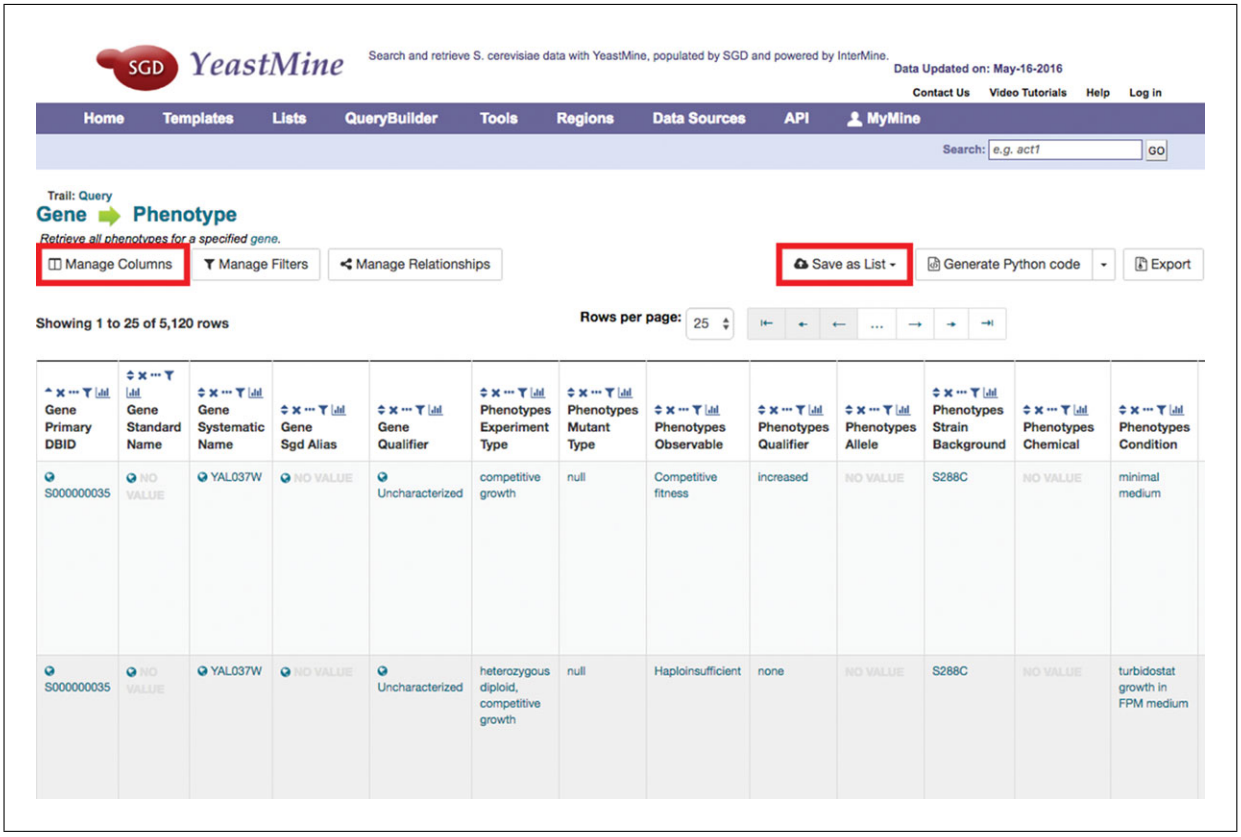

Figure 11.4.12 The results page for a Gene → Phenotype query. The Save as List and Manage Columns buttons are indicated by the red rectangles. Only a portion of the query results table is shown.

The YeastMine home page is divided into several different sections including a section of query templates in the main body of the page, predefined gene lists from different types of resources, and a search box (Fig. 11.4.9).

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

3. Click on the 'Templates' link in the toolbar that runs across the top of the YeastMine homepage (Fig. 11.4.9).

   Templates are predefined queries; each has a simple form and a description. You can also search for templates by keyword, or filter them by category (e.g., Proteins, Literature, Phenotypes). For each template query, the left side of the arrow represents input while the right side represents output.

4. Click on a template link to start the query (Fig. 11.4.10).

   For example (Fig. 11.4.10), the "Gene → Phenotype" query will retrieve every phenotype annotated to the gene(s) in the input list.

5. Examine the input options available. Define the input, then click on the Show Results button (Fig 11.4.11).

   In this example (Fig. 11.4.11), the pre-defined list "Uncharacterized_ORFs" was selected from the drop-down menu. Note that queries can also be run using a single gene name.

6. Explore the results of your query.

Managing query results

7. To add, remove, or rearrange columns on a query results page, use on the Manage Columns button located above the left side of the table (Fig. 11.4.12).

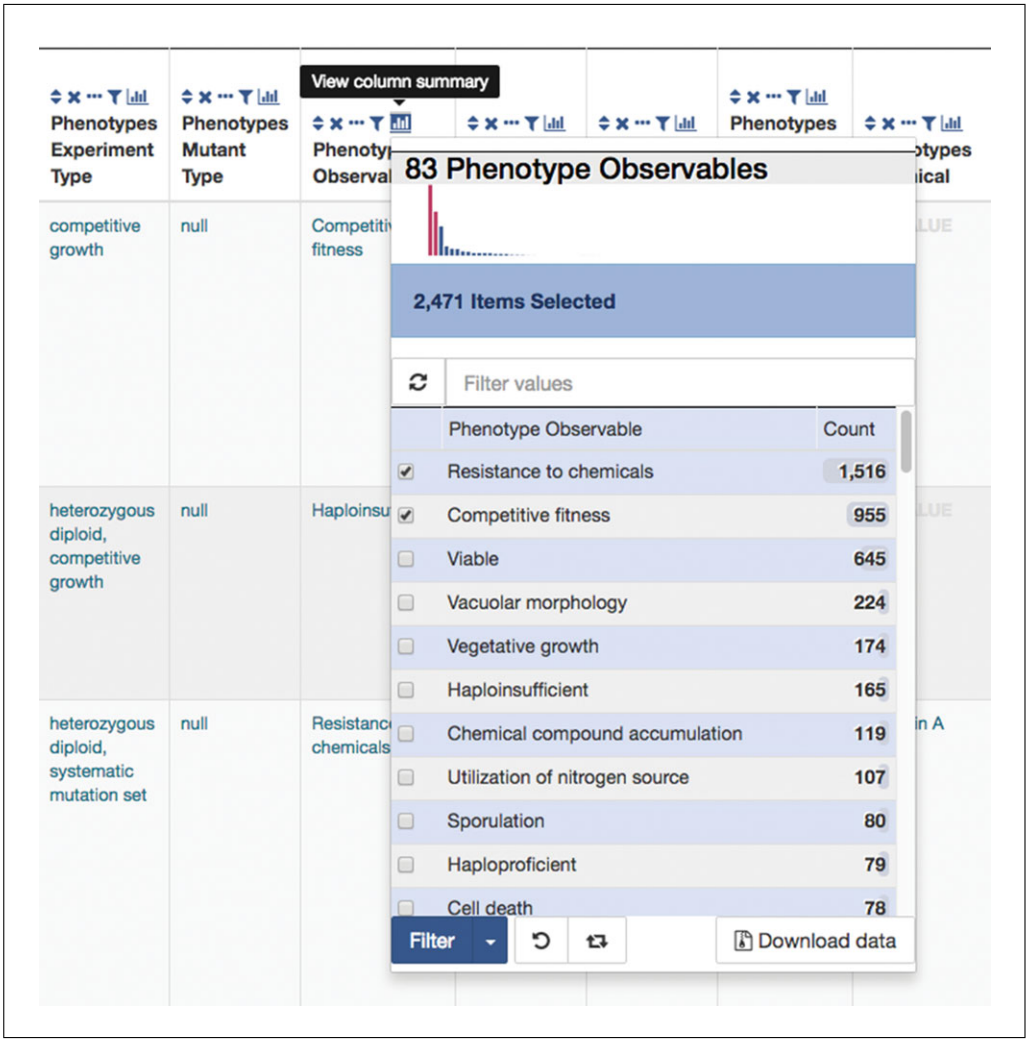

Figure 11.4.13 The View Column Summary function on the Phenotype Observables column. The View Column Summary button is located directly above the column title and is represented by a bar graph icon.

Each category of data is arranged into a column that can be moved around or deleted using the Manage Columns button. Additional columns can also be added using the Manage Columns feature.

8. To sort data, use the small blue icons located at the top of each column. These icons enable sorting/filtering of column values and toggling of column visibility, and also present column summaries.

   To filter the values in a column, use the View Column Summary feature (Fig. 11.4.13). Check desired values on which to filter, and then hit the Filter button. For example, using View Column Summary on the Phenotypes Observable column filters the table for specific observables.

Saving query results into a list

9. To save the results of your query into a list, use the Save as List button located above the right side of the table (Fig. 11.4.14).

   Each "Save as List" option saves a different portion of the table. Selecting the option "Gene," for example (Fig. 11.4.14), stores the genes on the left side of the table into a list of genes. To confirm which portion of the table will be saved, simply hover the cursor over each Save as List option. This will highlight the portion that will be saved into a YeastMine list.

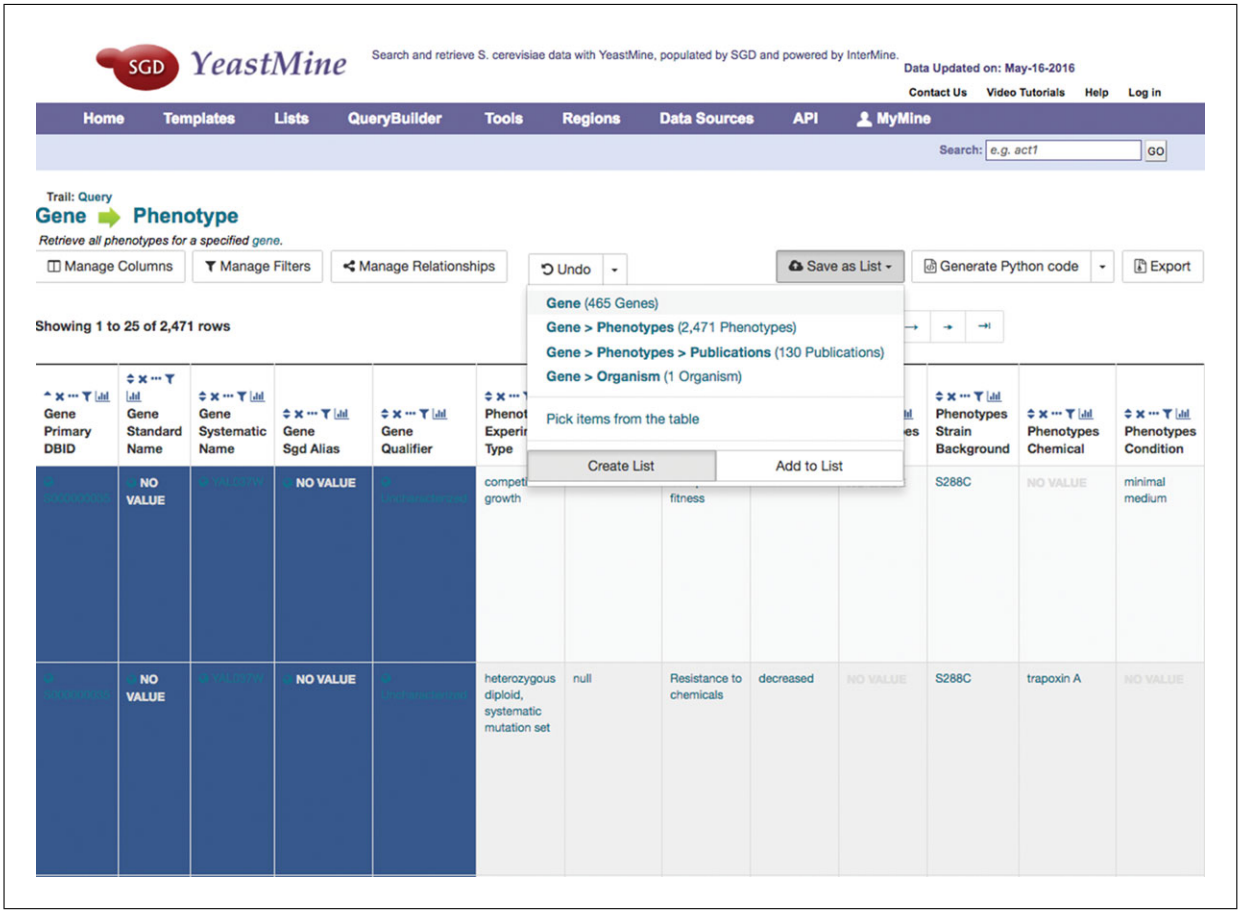

Figure 11.4.14 Use of the "Save as List" function to save the genes on the left side of the table (see blue highlight) into a YeastMine list. Only a portion of the query results table is shown.

10. To access your saved lists, click on the "Lists" link in the toolbar that runs across the top of the YeastMine home page (Fig. 11.4.9). Then, click on the "View" link located on the upper left portion of the page to be taken to the Lists page.

    The Lists page presents your saved lists along with the default lists provided by SGD. Note that lists are saved in a session-dependent manner. To save lists for future use, log in to create/use a MyMine account.

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, with a 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;* Stover and Cavalcanti, 2014), 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 10 to 13). To extract the entire coding region of the DNA helicase, plus flanking sequences, she uses JBrowse (Basic Protocol 2). To see which DNA helicases alter chronological lifespan in yeast when mutated, Elizabeth creates a list of DNA helicases in YeastMine and runs it through a Genes → Phenotype query (Basic Protocol 4, steps 3 to 8). 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, Command-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 are 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*: JBrowse (Skinner and Holmes, 2010), MGI (Shaw, 2009), RGD (Laulederkind et al., 2012), TAIR (Lamesch et al., 2010), WormBase (Schwarz and Sternberg, 2006), and the UCSC Genome Browser (Karolchik et al., 2012). 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. doi: 10.1093/nar/gki003.

Arnaud, M.B., Chibucos, M.C., Costanzo, M.C., Crabtree, J., Inglis, D.O., Lotia, A., Orvis, J., Shah, P., Skrzypek, M.S., Binkley, G., Miyasato, S.R., Wortman, J.R., and Sherlock, G. 2010. The *Aspergillus* Genome Database, a curated comparative genomics resource for gene, protein, and sequence information for the *Aspergillus* research community. *Nucleic Acids Res.* 38:D420-D427. doi: 10.1093/nar/gkp751.

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. doi: 10.1002/0471250953.bi0702s23.

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. doi: 10.1093/nar/gkn886.

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. doi: 10.1093/nar/26.1.73.

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. doi: 10.1093/bib/5.1.9.

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. doi: 10.1093/nar/25.1.63.

Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. *Nucleic Acids Res*. 36:440-444. doi: 10.1093/nar/gkm883.

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. doi: 10.1126/science.274.5287.546.

Karolchik, D., Hinrichs, A.S., and Kent, W.J. 2012. The UCSC genome browser. *Curr. Protoc. Bioinform.* 40:1.4.1-1.4.33. doi: 10.1002/0471250953.bi0104s40.

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*. 32:D332-D333. doi: 10.1093/nar/gkh138.

Lamesch, P., Dreher, K., Swarbreck, D., Sasidharan, R., Reiser, L., and Huala, E. 2010. Using The *Arabidopsis* Information Resource (TAIR) to find information about *Arabidopsis* genes. *Curr. Protoc. Bioinform.* 30:1.11.1-1.11.51. doi: 10.1002/0471250953.bi0111s30.

Laulederkind, S.J.F., Hayman, G.T., Wang, S.-J., Lowry, T.F., Nigam, R., Petri, V., Smith, J.R., Dwinell, M.R., Jacob, H.J., and Shimoyama, M. 2012. Exploring genetic, genomic, and phenotypic data at the rat genome database. *Curr. Protoc. Bioinform.* 40:1.14.1-1.14.27. doi: 10.1002/0471250953.bi0114s40.

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. doi: 10.1371/journal.pbio.0020309.

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. doi: 10.1002/(SICI)1097-0061(19990930)15:13<1365::AID-YEA468>3e3.0.CO;2-U.

Rhee, S.Y. 2000. Bioinformatic resources, challenges, and opportunities using *Arabidopsis* as a model organism in post-genomic era. *Plant Physiol*. 124:1460-1464. doi: 10.1104/pp.124.4.1460.

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. doi: 10.1002/0471250953.bi0108s14.

Shaw, D.R. 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. doi: 10.1002/0471250953.bi0107s25.

Skinner, M. E. and Holmes, I. H. 2010. Setting up the JBrowse genome browser. *Curr. Protoc. Bioinform.* 32:9.13.1-9.13.13. doi: 10.1002/0471250953.bi0913s32.

Skinner M.E., Uzilov A.V., Stein L.D., Mungall, C.J., and Holmes, I.H. 2009. JBrowse: A next-generation genome browser. *Genome Res*. 19:1630-1638. doi: 10.1101/gr.094607.109.

Smith, R.N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., Lyne, M., Lyne, R., Kalderimis, A., Rutherford, K., Stepan, R., Sullivan, J., Wakeling, M., Watkins, X., and Micklem, G. 2012. InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics* 28:3163-3165. doi: 10.1093/bioinformatics/bts577.

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. doi: 10.1093/nar/29.1.82.

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. doi: 10.1101/gr.403602.

Stover, N. A. and Cavalcanti, A. R. 2014. Using NCBI BLAST. *Curr. Protoc. Essen. Lab. Tech.* 11:11.1.1-11.1.35. doi: 10.1002/9780470089941.et1101s08.

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. doi: 10.1093/nar/gkl988.

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://www.aniseed.cnrs.fr

Ascidian Network for In Situ 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://porteco.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://hymenopteragenome.org

Hymenoptera Genome Database (BeeBase, NasoniaBase, Ant Genomes Portal): Database of genes and genomes of Apis mellifera, Nasonia vitripennis, and other Hymenopterans.

http://intermine.org

InterMine: Open-source data warehouse system that enables the creation of biological databases that integrate multiple types of data from different sources.

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.pombase.org

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.