D736–D742 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 11 November 2017*
doi: 10.1093/nar/gkx1112

Saccharomyces genome database informs human biology

Marek S. Skrzypek, Robert S. Nash, Edith D. Wong, Kevin A. MacPherson, Sage T. Hellerstedt, Stacia R. Engel, Kalpana Karra, Shuai Weng, Travis K. Sheppard, Gail Binkley, Matt Simison, Stuart R. Miyasato and J. Michael Cherry*

Department of Genetics, Stanford University, Stanford, CA, 94305-5120 USA

Received September 15, 2017; Revised October 20, 2017; Editorial Decision October 21, 2017; Accepted October 24, 2017

ABSTRACT

The *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org) is an expertly curated database of literature-derived functional information for the model organism budding yeast, *Saccharomyces cerevisiae*. SGD constantly strives to synergize new types of experimental data and bioinformatics predictions with existing data, and to organize them into a comprehensive and up-to-date information resource. The primary mission of SGD is to facilitate research into the biology of yeast and to provide this wealth of information to advance, in many ways, research on other organisms, even those as evolutionarily distant as humans. To build such a bridge between biological kingdoms, SGD is curating data regarding yeast-human complementation, in which a human gene can successfully replace the function of a yeast gene, and/or vice versa. These data are manually curated from published literature, made available for download, and incorporated into a variety of analysis tools provided by SGD.

INTRODUCTION

The *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org) is a freely available online resource that serves the research community by curating knowledge about genes and proteins of the model organism budding yeast, *Saccharomyces cerevisiae*. Yeast had been an object of studies in the fields of genetics, biochemistry and cell biology for decades before it became the first eukaryote with a completely sequenced genome (1). Availability of the genome sequence, supported by all the previously collected information and aided by the development of a powerful arsenal of molecular biology techniques, has led to the accumulation of an enormous amount of knowledge, making yeast, arguably, one of the most thoroughly characterized model organisms. The constant development of new techniques that are validated and refined on yeast promises that this status will continue into the future.

The mission of SGD is to collect gene- and protein-related information from all publicly available sources. Research articles in scientific journals, expertly curated by SGD biocurators, remain the primary source of data. Large-scale datasets are also collected both manually and with automated scripts from publications, and also from repositories such as the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo) and Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra). The goal is to extract, validate, and organize the data coming from various sources into a uniform, comprehensive and up-to-date compendium that researchers can rely on in their work.

The wealth of information provided by SGD can also be leveraged outside of the research into yeast biology. Understanding the functions of a yeast gene may shed light on functions of an orthologous gene in another organism, even one as evolutionary distant as humans. It is often assumed that over the course of evolution, orthologous genes from different organisms retain at least some of their characteristics. Even though this 'ortholog-function conjecture' has many known exceptions, it remains a valuable and widely used method for predicting gene functions based on sequence similarity (2). Yeast and humans are separated by a billion years of evolution (3), yet there are thousands of recognizable orthologs between the two species (4). Moreover, there are hundreds of genes from one species that can functionally replace (complement) their orthologs in the other (5). Such functional complementation between genes from evolutionarily distant species indicates a significant conservation of function, opening up endless research possibilities to explore. A researcher can, for example, try to identify yeast metabolic pathways, regulatory circuits, or protein complexes the gene of interest is involved in, as well as mutant phenotypes and their potential disease links, and then start looking for similar connections in humans. Moreover, a human gene that functions in yeast cells gives access to all the power of yeast genetics that can be readily utilized.

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

Table 1. Number of unique complementation pairs and participating genes in SGD's functional complementation dataset. The functional complementation dataset includes curation by both SGD and Princeton Protein Orthology Database (P-POD); contributions by each group are indicated in the table

	Total	Curated by SGD	Curated by P-POD	Overlapping curation removed
Complementation pairs	1014	674	340	
Unique yeast genes	569	416	229	(645 total – 76 overlap)
Unique human genes	615	450	236	(686 total – 71 overlap)

For example, by establishing functional assays, researchers can experimentally dissect functional conservation to study the impact of human disease-associated variants on protein function (6).

In order to facilitate this research, SGD has been collecting yeast-human complementation data from the published literature. We have also collaborated with the Princeton Protein Orthology Database (P-POD; http://ppod.princeton.edu) to incorporate a set of data collected by their biocurators. The collective data are available for download from SGD and are also accessible by multiple search and analysis tools provided by SGD. This allows a researcher to not only find a human gene that functionally complements a yeast gene (or vice versa), but also to retrieve everything that is known about it.

YEAST-HUMAN FUNCTIONAL COMPLEMENTATION DATA IN SGD

A multifaceted strategy was used by SGD to identify experimentally determined functional complementation data in the yeast scientific literature. Publications were identified by searching PubMed with keywords ('human', 'complement', and 'baker's yeast'), searching SGD's instance of Textpresso (https://textpresso.yeastgenome.org; 7) using keywords ('human' and 'complement'), searching the database of human genes and genetic disorders, Online Mendelian Inheritance in Man (OMIM; https://www.omim.org; 8), with keywords ('yeast' and 'human' and 'complement'). The data collected at SGD was also mined by reviewing papers previously assigned by SGD to specific literature guide topics such as 'cross-species expression' and 'disease gene related.' A list was also generated based on information included in summary paragraphs and gene descriptions contained within SGD, in addition to papers internally tagged as containing information on 'Homology/Disease'. Another strategy involved generating a list of yeast genes with human homologs contained within OMIM, followed by extracting of the associated publications. Curators reviewed the associated scientific literature obtained using these strategies and extracted the yeast gene name, the HUGO Gene Nomenclature Committee approved symbol and ID for the human gene (HGNC; http://www.genenames.org), the direction of complementation, the PubMed identifier, and, when appropriate, notes describing details associated with the experimental setup.

In an independent approach, biocurators at the Princeton Protein Orthology Database (P-POD; http://ppod.princeton.edu; 9) manually curated experimentally-determined functional complementation data from the yeast literature for Ortholog Groups of Protein Sequences (OrthoMCL; http://orthomcl.org/orthomcl; 10), and predicted yeast-human ortholog pairs using the literature guide topics 'cross-species expression' and 'disease-gene related,' that were assigned at SGD to categorize the yeast literature. After removing redundant annotations curated by both groups, a total of 1014 experimentally-determined functional complementation pairs were identified by either SGD (674) or by P-POD (340). This set represents 569 unique yeast genes that when mutated can be partially or fully complemented by 615 unique human genes. Also included are a small number of cases in which a yeast gene can complement a defect in a human cell line or in an RNAi ablation experiment (Table 1).

FUNCTIONAL COMPLEMENTATION DATA RETRIEVAL

There are currently two places that this cross-species complementation data can be retrieved: the SGD Downloads site and the YeastMine data warehouse. A file named functional_complementation.tab, located in the Curated Data folder of the download site (https://downloads.yeastgenome.org/curation/literature), contains the experimentally-determined yeast-human functional complementation pairs described above. Each row in this file contains: the systematic and standard name of the yeast gene, the HGNC-approved human gene symbol and database identifier, the direction of complementation, the PubMed identifier for the reference from which the relationship was curated, the curation source (SGD or P-POD), and an optional free text note to describe additional details (Supplemental Table S1).

Complementation data can also be retrieved using YeastMine, SGD's data search and retrieval tool (Figure 1; https://yeastmine.yeastgenome.org). Two pre-composed queries, also known as templates, can be used to identify complementation data for specific yeast or human genes, or for a list of genes. Using the template 'Gene → Functional Complementation' (Figure 2; https://yeastmine.yeastgenome.org/yeastmine/template.do?name=Gene_Complements&scope=global), a yeast gene, or list of genes, can be entered to retrieve information on human genes that can complement yeast mutants, or yeast genes that can complement mutations in human genes expressed in cell lines. This template returns the same information as the downloadable file described above, as well as additional identifiers for the human genes in other databases such as NCBI Gene (https://www.ncbi.nlm.nih.gov/gene) and Ensembl (http://www.ensembl.org). For a complete list of complementation results, a pre-composed list of all yeast open reading frames (ORFs) is available for use.

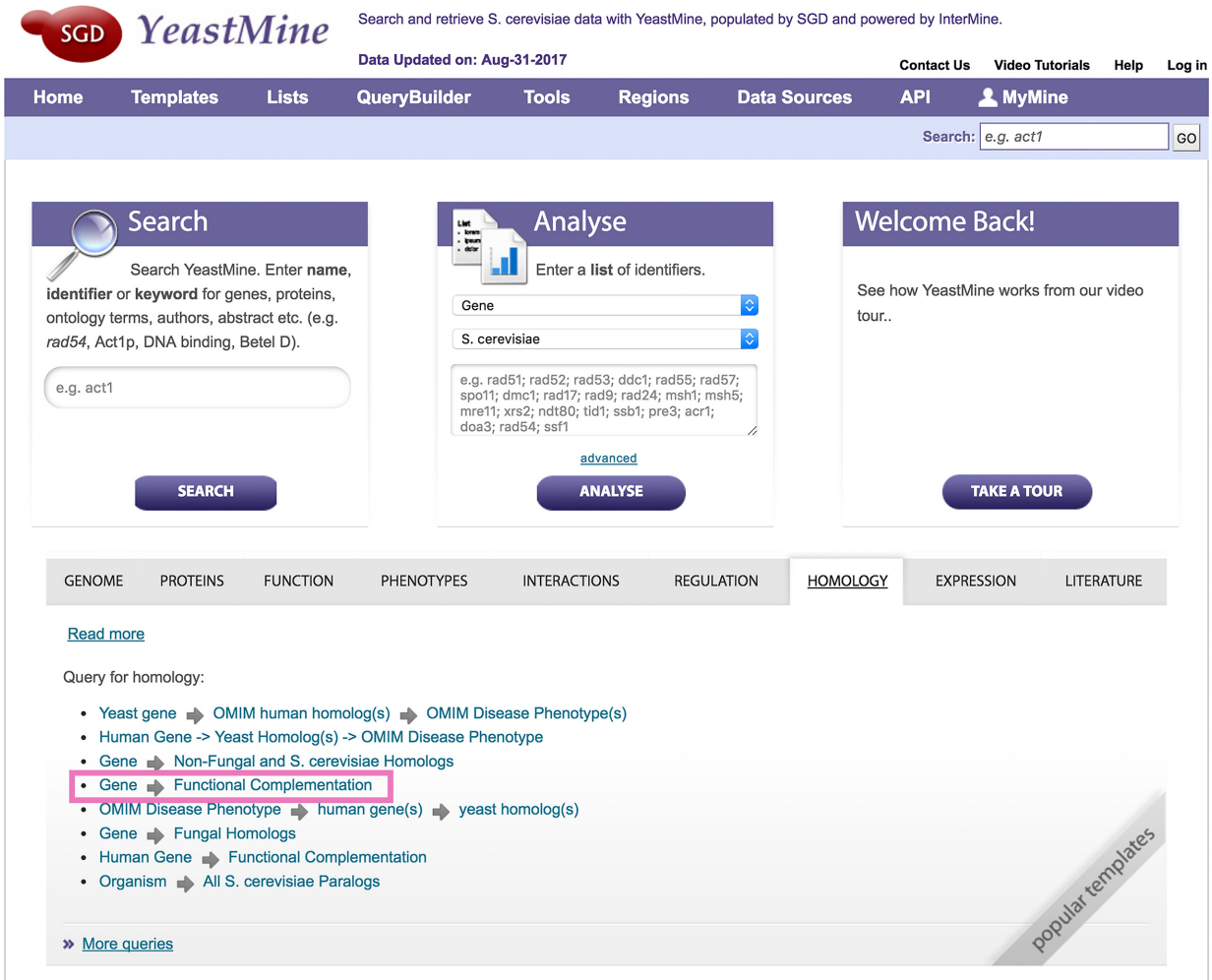

Figure 1. The top portion of the YeastMine homepage (https://yeastmine.yeastgenome.org/). The Homology tab is selected and the Gene → Functional Complementation template query is indicated by the pink rectangle. Selecting this query brings up the Gene → Functional Complementation query page shown in Figure 2.

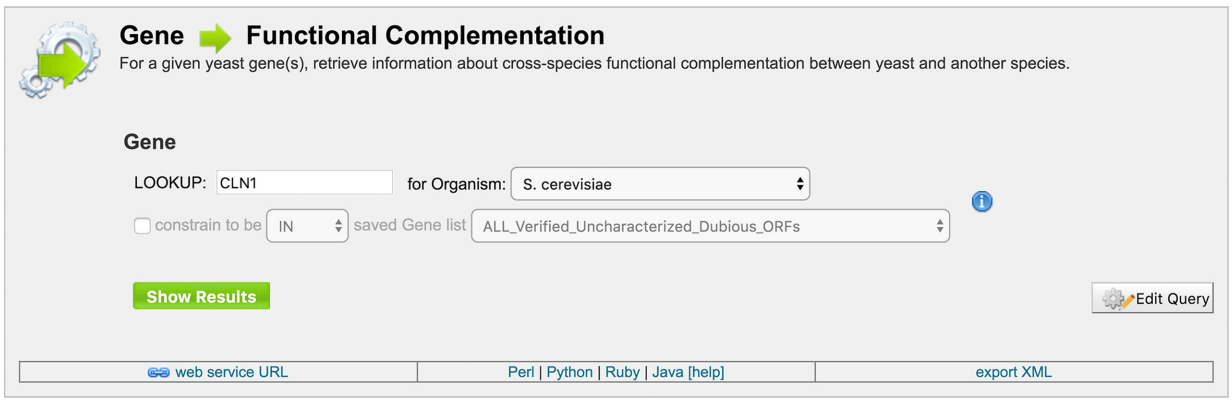

Figure 2. A portion of the Gene → Functional Complementation query page in YeastMine. Entering a yeast gene or a list of yeast genes retrieves human genes that show functional complementation. Currently, only data for yeast and human genes are available.

Using the 'Human Gene → Functional Complementation' template (https://yeastmine.yeastgenome.org/yeastmine/template.do?name=HumaGene_YeastGene_Complement&scope=global), a human gene, or list of genes, can be entered to retrieve information on yeast genes that can be complemented by the human gene and vice versa.

USING SGD TOOLS TO UNDERSTAND HUMAN GENES

The list of yeast genes that are functionally complemented by human genes by itself provides insights into conservation of biological processes across taxonomic kingdoms. This set of genes is enriched in genes involved in key cellular processes, such as nucleic acid and protein metabolism, including transcription, translation, and protein processing, as well as fundamental processes of energy generation and biosynthesis (Table 2). These processes are compromised in multiple genetic deficiencies in humans, thus this evolutionary conservation makes budding yeast an indispensable model for the study of genetic diseases (Figure 3;11,12).

Having identified a yeast gene capable of cross-species functional complementation with a human gene, the researcher would want to gather all available information about it, including its function, subcellular localization in organelles, membership in protein complexes, participation in metabolic or regulatory pathways and modes of regulation. Mutant phenotype data collected in SGD may provide a goldmine of information about the protein's role within the cell, as well as about its structure, domains, and active sites.

SGD provides tools to find all these types of data and more. Searching for a yeast gene name leads to the Locus Summary page (LSP), which shows an overview of available information. The LSP is divided into sections that outline specific aspects: genomic context, protein statistics, functional information captured as Gene Ontology (GO) annotations and mutant phenotypes. A new feature recently implemented on the LSPs is expertly composed free-text summaries for the Gene Ontology and Phenotype sections, which are intended to outline the most salient points. Each section provides links to more data made available on additional SGD webpages, also accessible via tabs that run across the top of the LSP. For example, the regulation pages contains transcription and protein regulatory information, including any predicted DNA binding site motifs as listed in YeTFaSCo (13), targets, regulators, and conditions under which the regulatory relationship is observed, all based on experimental evidence. For genes with regulatory targets, shared biological processes for these targets are listed. Other pages include those for expression, which present detailed information about the expression of a particular gene product, along with references for each observation. An expression overview includes a histogram, which graphically represents the distribution of conditions in which the gene's expression is up- or down-regulated. Interactions pages list curated genetic and physical interactions for a particular gene product, along with references, imported on a regular basis from BioGRID (14).

The search box present near the top of most pages in SGD has much more robust capabilities than searching for individual genes. SGD has recently implemented a faceted search, which uses categories to facilitate exploration and filtering through many different data types (Figure 4). For example, searching for 'mitotic cell cycle' produces thousands of results that are sorted into meaningful categories (facets) such as genes, biological processes, cellular components, phenotypes, chemicals and references, making it easy to select a desired subset.

SGD's YeastMine allows exploring a gene, or multiple genes, across multiple types of data. For example, given a list of human genes that are involved in cancer, a researcher could leverage the data in YeastMine to identify new potential chemotherapy targets. Recently, synthetic lethal interactions between BRCA1/2 and PARP have been exploited to kill cancer cells while leaving normal cells unaffected (15). First, using the YeastMine query: Human Gene → Yeast Homolog(s) → OMIM Disease Phenotype with the list of human genes, a list of the homologous yeast genes would be made. Then, with this second list of yeast genes, one would identify potential synthetic lethal interactions with other yeast genes using the YeastMine query: Gene → Interaction, and filtering for only 'Synthetic Lethal' interactions. If interested, any function or process annotations or pathways associated with the genes in one or both of these lists can also be explored with the Gene → GO Term and Gene→Pathway queries. Finally, the human homologs to the yeast genes that synthetically interact with the starting list of genes can be identified using the Yeast gene > OMIM human homolog(s) > OMIM Disease Phenotype(s) query. We have a detailed scenario using YeastMine as well as other tutorial videos under 'Playlists' on our YouTube channel (https://www.youtube.com/SaccharomycesGenomeDatabase).

At the end of data mining in SGD, any potentially interesting findings among yeast genes can be 'translated' back to human genes using YeastMine queries such as: Yeast gene > OMIM human homolog(s) > OMIM Disease Phenotype(s). Given a yeast gene or list of yeast genes, this query retrieves all human homologs and any associated OMIM disease phenotype. A query for the converse is also available: OMIM Disease Phenotype > human gene(s) > yeast homolog(s). This query, given a specific OMIM phenotype as input, will return all associated human genes, along with their yeast homologs.

SGD AND THE ALLIANCE OF GENOME RESOURCES

SGD, along with the five other major model organism database (MOD) groups (Mouse Genome Database (MGD; http://www.informatics.jax.org), Rat Genome Database (RGD; http://rgd.mcw.edu), Zebrafish Information Network (ZFIN; http://zfin.org), WormBase (http://wormbase.org), and FlyBase (http://flybase.org)) and the Gene Ontology (GO) Consortium (http://www.geneontology.org), recently founded the Alliance of Genome Resources, a new data resource that aims to integrate, develop and maintain genome information resources in order to facilitate the use of a variety of model organisms to understand the genetic

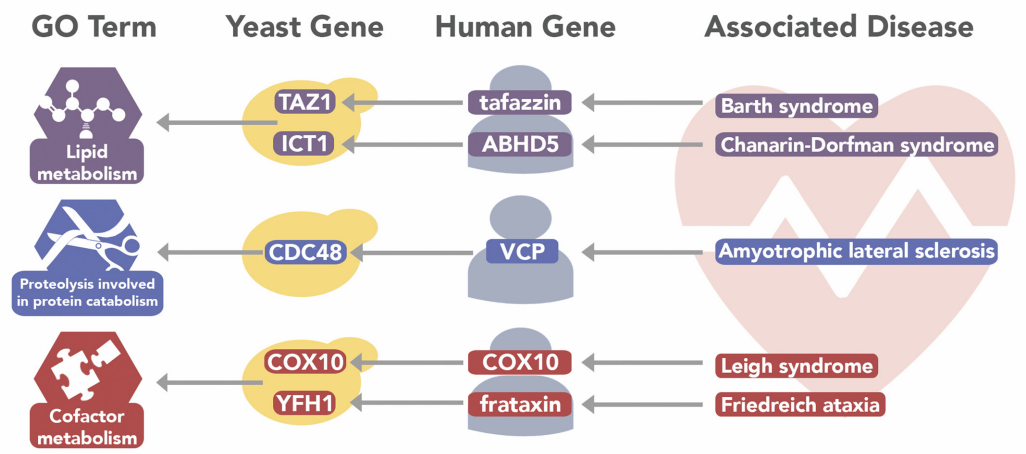

Figure 3. Sample of disease-relevant, functionally complemented yeast genes from the top 3 enriched GO terms in Table 2. Mutations in each of the yeast genes shown are either fully or partially complemented by their respective human homolog, indicated on the right. Diseases associations for the complementing human genes were taken from the Online Mendelian Inheritance in Man (OMIM; https://www.omim.org/; 8), a catalog of human genes and genetic disorders.

Table 2. Gene Ontology term enrichment for SGD's functional complementation dataset. SGD's GO Slim Mapper tool (https://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl) was used to analyze the yeast genes in the functional complementation dataset. The Frequency column shows what percentage of genes in the dataset is annotated to a given GO term, in comparison to the percentage of genes annotated to that term in the entire genome (Genome Frequency column). Only those GO terms that have two-fold or higher enrichment are shown

Gene Ontology Term	Frequency	Genome Frequency
lipid metabolic process	13.4%	4.6%
cofactor metabolic process	9.0%	2.8%
proteolysis involved in cellular protein catabolic process	7.0%	3.4%
nucleobase-containing small molecule metabolic process	6.9%	3.1%
generation of precursor metabolites and energy	5.8%	2.4%
RNA splicing	4.7%	2.2%
RNA catabolic process	4.4%	2.1%
tRNA processing	4.4%	1.7%
cellular ion homeostasis	4.0%	2.0%
cellular respiration	3.2%	1.3%
DNA-templated transcription, initiation	3.0%	1.1%
translational initiation	2.8%	0.9%
tRNA aminoacylation for protein translation	2.6%	0.6%
protein glycosylation	2.3%	1.0%
transcription from RNA polymerase III promoter	2.3%	0.6%
protein lipidation	2.1%	0.7%
ribosome assembly	2.1%	0.9%
ribosomal subunit export from nucleus	1.9%	0.7%
protein maturation	1.9%	0.8%
snoRNA processing	1.6%	0.7%

bases of human biology and disease. The establishment of the Alliance builds upon and expands the collaborations that have existed between the MODs and GO over the last two decades seeking to enhance data consolidation, dissemination, and the application of shared standards. These groups are merging in order to build unified data visualizations, and coordinate data retrieval and analysis, all in an effort to support comparative model organism genomics. This effort is aimed at serving human geneticists, basic science researchers, computational biologists, educators, and students. The first open-to-the-public release of the Alliance public web portal in October 2017 enables users to quickly retrieve functional details, orthology information, and disease associations for genes in human, mouse, rat, zebrafish, worm, fly, and yeast, all framed in a comparative perspective. As the integration of the MODs and GO moves forward, other model organism resources and bioinformatic tools will be integrated within this unique, powerful common data platform, further facilitating data analysis and discovery.

FUTURE DIRECTIONS

The model organism budding yeast is a well-established model for studying human biology (16–19). Moving forward, SGD will continue to incorporate newly published yeast-human functional complementation. We also plan to gather more information on the expression of human proteins in yeast (20,21). We are already capturing allele and phenotype information from 'humanized yeast' genes and proteins, in which sequence changes corresponding to human variants are introduced into yeast mutants for the hu-

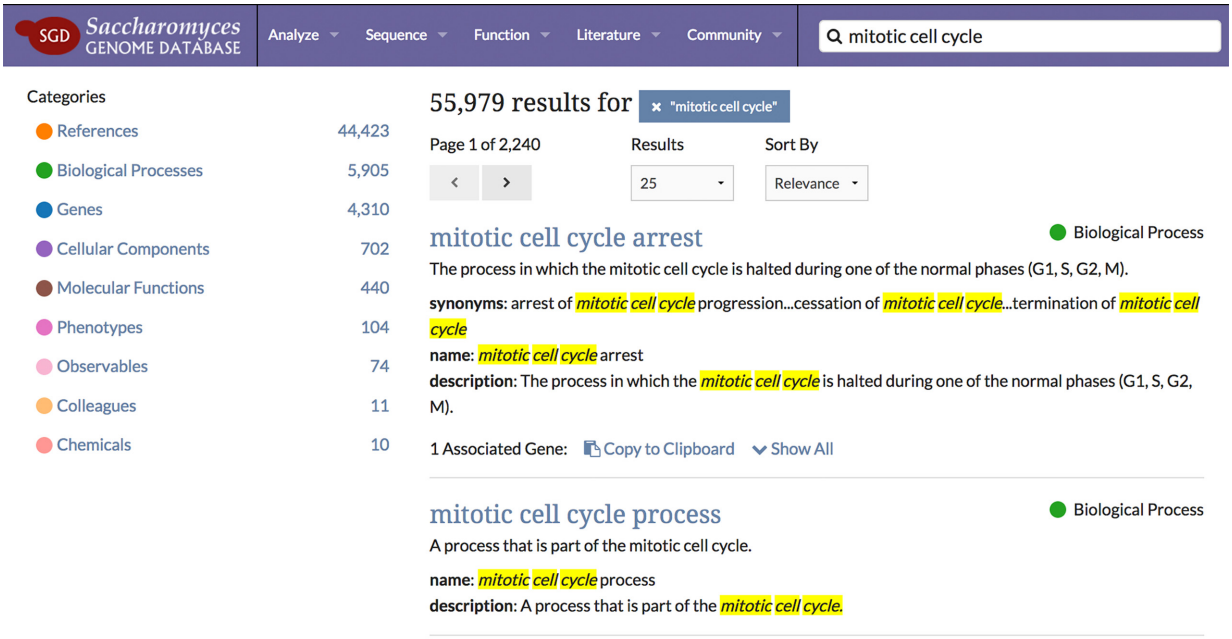

Figure 4. Example search results for 'mitotic cell cycle' at SGD (https://yeastgenome.org/). Only the first two results are shown. Multiple categories, or 'facets', are displayed on the left side of the page to enable filtering and refining of search results. Additional facets are available within each category and appear once a category is selected.

man orthologs (22,23), and will expand upon this curation. All these data types will be incorporated into SGD webpages and tools in an effort to support the scientific community as they continue to tease apart the complexities of human biology and health.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank the P-POD team at Princeton University for their generous sharing of their data with us. We thank the yeast research community for their input and support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health. The funders had no role in design, data processing, implementation, decision to publish, or preparation of the manuscript.

FUNDING

National Human Genome Research Institute at the United States National Institutes of Health [U41 HG001315]. Funding for open access charge: U41 grant from the National Human Genome Research Institute at the United States National Institutes of Health [HG001315].
Conflict of interest statement. None declared.

REFERENCES

1. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546, 563–567.
2. Gabaldón,T. and Koonin,E.V. (2013) Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.*, **14**, 360–366.
3. Douzery,E.J., Snell,E.A., Bapteste,E., Delsuc,F. and Philippe,H. (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 15386–15391.
4. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
5. Kachroo,A.H., Laurent,J.M., Yellman,C.M., Meyer,A.G., Wilke,C.O. and Marcotte,E.M. (2015) Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, **348**, 921–925.
6. Sun,S., Yang,F., Tan,G., Costanzo,M., Oughtred,R., Hirschman,J., Theesfeld,C.L., Bansal,P., Sahni,N., Yi,S. *et al.* (2016) An extended set of yeast-based functional assays accurately identifies human disease mutations. *Genome Res.*, **26**, 670–680.
7. 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.
8. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
9. Heinicke,S., Livstone,M.S., Lu,C., Oughtred,R., Kang,F., Angiuoli,S.V., White,O., Botstein,D. and Dolinski,K. (2007) The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. *PLoS One*, **2**, e766.
10. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
11. Papur,O.S., Terzioglu,O. and Koc,A. (2015) Functional characterization of new mutations in Wilson disease gene (ATP7B) using the yeast model. *J. Trace Elem. Med. Biol.*, **31**, 33–36.
12. Whited,K., Baile,M.G., Currier,P. and Claypool,S.M. (2013) Seven functional classes of Barth syndrome mutation. *Hum. Mol. Genet.*, **22**, 483–492.
13. de Boer,C.G. and Hughes,T.R. (2012) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res.*, **40**, D169–D179.

14. Chatr-Aryamontri,A., Oughtred,R., Boucher,L., Rust,J., Chang,C., Kolas,N.K., O'Donnell,L., Oster,S., Theesfeld,C., Sellam,A. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, **45**, D369–D379.
15. Konecny,G.E. and Kristeleit,R.S. (2016) PARP inhibitors for BRCA1/2-mutated and sporadic ovarian cancer: current practice and future directions. *Br. J. Cancer*, **115**, 1157–1173.
16. Tenreiro,S., Franssens,V., Winderickx,J. and Outeiro,T.F. (2017) Yeast models of Parkinson's disease-associated molecular pathologies. *Curr. Opin. Genet. Dev.*, **44**, 74–83.
17. Yilmaz,L.S. and Walhout,A.J. (2017) Metabolic network modeling with model organisms. *Curr. Opin. Chem. Biol.*, **36**, 32–39.
18. Lees,H., Walters,H. and Cox,L.S. (2016) Animal and human models to understand ageing. *Maturitas*, **93**, 18–27.
19. McIsaac,R.S., Lewis,K.N., Gibney,P.A. and Buffenstein,R. (2016) From yeast to human: exploring the comparative biology of methionine restriction in extending eukaryotic life span. *Ann. N. Y. Acad. Sci.*, **1363**, 155–170.
20. Synoradzki,K. and Bieganowski,P. (2015) Middle domain of human Hsp90 isoforms differentially binds Aha1 in human cells and alters Hsp90 activity in yeast. *Biochim. Biophys. Acta*, **1853**, 445–452.
21. Saleh,A.A., Bhadra,A.K. and Roy,I. (2014) Cytotoxicity of mutant huntingtin fragment in yeast can be modulated by the expression level of wild type huntingtin fragment. *ACS Chem. Neurosci.*, **5**, 205–215.
22. Baruffini,E., Ferrari,J., Dallabona,C., Donnini,C. and Lodi,T. (2015) Polymorphisms in DNA polymerase γ affect the mtDNA stability and the NRTI-induced mitochondrial toxicity in *Saccharomyces cerevisiae*. *Mitochondrion*, **20**, 52–63.
23. Mayfield,J.A., Davies,M.W., Dimster-Denk,D., Pleskac,N, McCarthy,S., Boydston,E.A., Fink,L., Lin,X.X., Narain,A.S., Meighan,M. *et al.* (2012) Surrogate genetics and metabolic profiling for characterization of human disease alleles. *Genetics*, **190**, 1309–1323.