Published online 28 October 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D695–D699***
doi:10.1093/nar/gkr853

PomBase: a comprehensive online resource for fission yeast

Valerie Wood[1,2,3,*], Midori A. Harris[1,2,*], Mark D. McDowall[4], Kim Rutherford[1,2], Brendan W. Vaughan[4], Daniel M. Staines[4], Martin Aslett[5], Antonia Lock[6], Jürg Bähler[6], Paul J. Kersey[4] and Stephen G. Oliver[1,2,*]

[1]Cambridge Systems Biology Centre, [2]Department of Biochemistry, University of Cambridge, Sanger Building, 80 Tennis Court Road, Cambridge CB2 1GA, [3]Cell Cycle Laboratory, Cancer Research UK, London Research Institute, 44 Lincoln's Inn Fields, London UK WC2A 3LY, [4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA and [6]Department of Genetics, Evolution and Environment, and UCL Cancer Institute, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

Received September 7, 2011; Accepted September 24, 2011

ABSTRACT

PomBase (www.pombase.org) is a new model organism database established to provide access to comprehensive, accurate, and up-to-date molecular data and biological information for the fission yeast *Schizosaccharomyces pombe* to effectively support both exploratory and hypothesis-driven research. PomBase encompasses annotation of genomic sequence and features, comprehensive manual literature curation and genome-wide data sets, and supports sophisticated user-defined queries. The implementation of PomBase integrates a Chado relational database that houses manually curated data with Ensembl software that supports sequence-based annotation and web access. PomBase will provide user-friendly tools to promote curation by experts within the fission yeast community. This will make a key contribution to shaping its content and ensuring its comprehensiveness and long-term relevance.

INTRODUCTION

The fission yeast *Schizosaccharomyces pombe* is a well-studied eukaryotic model organism that has been used since the 1950s to obtain valuable insights into diverse eukaryotic biological processes including the cell growth and division cycle, genome organization and maintenance, cell morphology and cytokinesis, signaling and stress responses, chromatin and gene regulation and meiotic differentiation (1). Moreover, since the completion of its genome sequence in 2002 (2), fission yeast has emerged as a prime model for the characterization of processes relevant to human disease and cell biology. A large and active community engages in biological and biomedical research using this model system, routinely applying molecular genetic, cell biological and biochemical techniques. Data from small- and large-scale projects are accumulating rapidly and will increase substantially over the next few years; the literature corpus currently exceeds 9000 publications and grows by about 500 publications per year.

PomBase (http://www.pombase.org) has recently been established to provide user-friendly and standardized access to genomic features and annotations, enabling scientists to assimilate novel findings into their research programs, improve experimental design, support the interpretation of genetic screens, and facilitate the interpretation of functional genomics and systems biology experiments. An accurate and comprehensive set of manual annotations of gene products based on published data lies at the centre of this database, and is supplemented by automatic annotation and information about non-genic features. The PomBase project aims to provide three key resources to the fission yeast community:

- Comprehensive and deep curation of the scientific literature;
- A software infrastructure to support curation activities; and

*To whom correspondence should be addressed. Tel: +44 1223 746961; Fax: +44 1223 766002; Email: vw253@cam.ac.uk
Correspondence may also be addressed to Midori A. Harris. Tel: +44 1223 761211; Fax: +44 1223 766002; Email: mah79@cam.ac.uk
Correspondence may also be addressed to Stephen G. Oliver. Tel: +44 1223 333667; Fax: +44 1223 766002; Email: sgo24@cam.ac.uk

- A computational infrastructure to integrate data curated from the literature with the genomic sequence, high-throughput data sets, and data from other fungal genomes.

DATA TYPES AND SOURCES

Historically, functional information about the genome and biology of *S. pombe* has been maintained in a repository hosted by the GeneDB project at the Wellcome Trust Sanger Institute (WTSI) (http://old.genedb.org/genedb/pombe/). This resource is now superseded by PomBase, which has not only inherited the data from GeneDB, but also includes additional curated data types and high-throughput data sets. The major data types available in PomBase are listed in Table 1.

Sequence feature annotation

DNA and protein features are annotated using the Sequence Ontology (SO) (4). Currently, 31 DNA feature terms are used in 25 650 annotations (examples include gene, exon, tRNA, centromere) and 22 protein feature terms for 919 annotations (examples include nuclear localization signal, ER retention signal, DDB box).

Gene Ontology annotation

Gene Ontology (GO) (5,6) terms are assigned to gene products to represent their molecular functions, cellular components (including complexes), and biological processes. PomBase GO annotation data comprises over 33 500 manual annotations and approximately 3500 automatically assigned annotations, using 3800 unique GO terms.

Phenotype annotation

To support the comprehensive and detailed representation of phenotypes, we are developing the Fission Yeast Phenotype Ontology (FYPO), a formal ontology of phenotypes observed in fission yeast. FYPO is a modular ontology that uses several existing ontologies from the Open Biological and Biomedical Ontologies (OBO) collection (7) as building blocks, including the phenotypic quality ontology PATO (8), GO, and Chemical Entities of Biological Interest (ChEBI) (9). Over 7000 existing annotations have been converted from the legacy GeneDB controlled vocabulary to FYPO terms; these annotations will support sophisticated querying, computational analysis, and comparison between different experiments and even between different species.

Genetic and physical interactions

Annotation of genetic and physical interactions are supported using the BioGRID (http://thebiogrid.org) (10,11) annotation format. Existing annotations curated by BioGRID are imported into PomBase, and newly created annotations will be exchanged with BioGRID.

Genome-scale data

PomBase incorporates a wide variety of data sets that can be mapped to the genome, obtained either from internal sources or via externally loaded URLs or data files. Examples of supported data sets include whole genome re-sequencing data, RNA-seq and ChIP-seq data, various microarray data and other high-throughput data types.

USER ACCESS TO POMBASE

A web portal has been developed for access to the PomBase data, which provides pages describing the current state of annotation of the *S. pombe* genome, items of interest to the community, and, most importantly, a 'Gene Overview Page' that summarizes key information about each gene. Embedded in this portal is a genome browser, providing access to genomic context, sequence-based analyses and high-throughput data.

Table 1. Summary of data types available in PomBase

Data type		Distinct descriptors	Total annotations
Sequence Ontology	DNA features	31	25 650
	Protein features	22	919
Gene Ontology	Biological process	1861	14 501
	Molecular function	1405	9149
	Cellular component	543	16 101
Fission yeast phenotype ontology		342	7043
Gene product descriptions		4331	7048
Disease associations		134	378
PSI-MOD(3) protein modification		22	1861
EC numbers		520	837
Name descriptions		186	513
Annotation extensions		254	340
Annotation status		7	5142
BioGRID	Genetic interactions	N/A	13 147
	Physical interactions	N/A	5131
Curated fission–budding yeast orthologs		N/A	5210

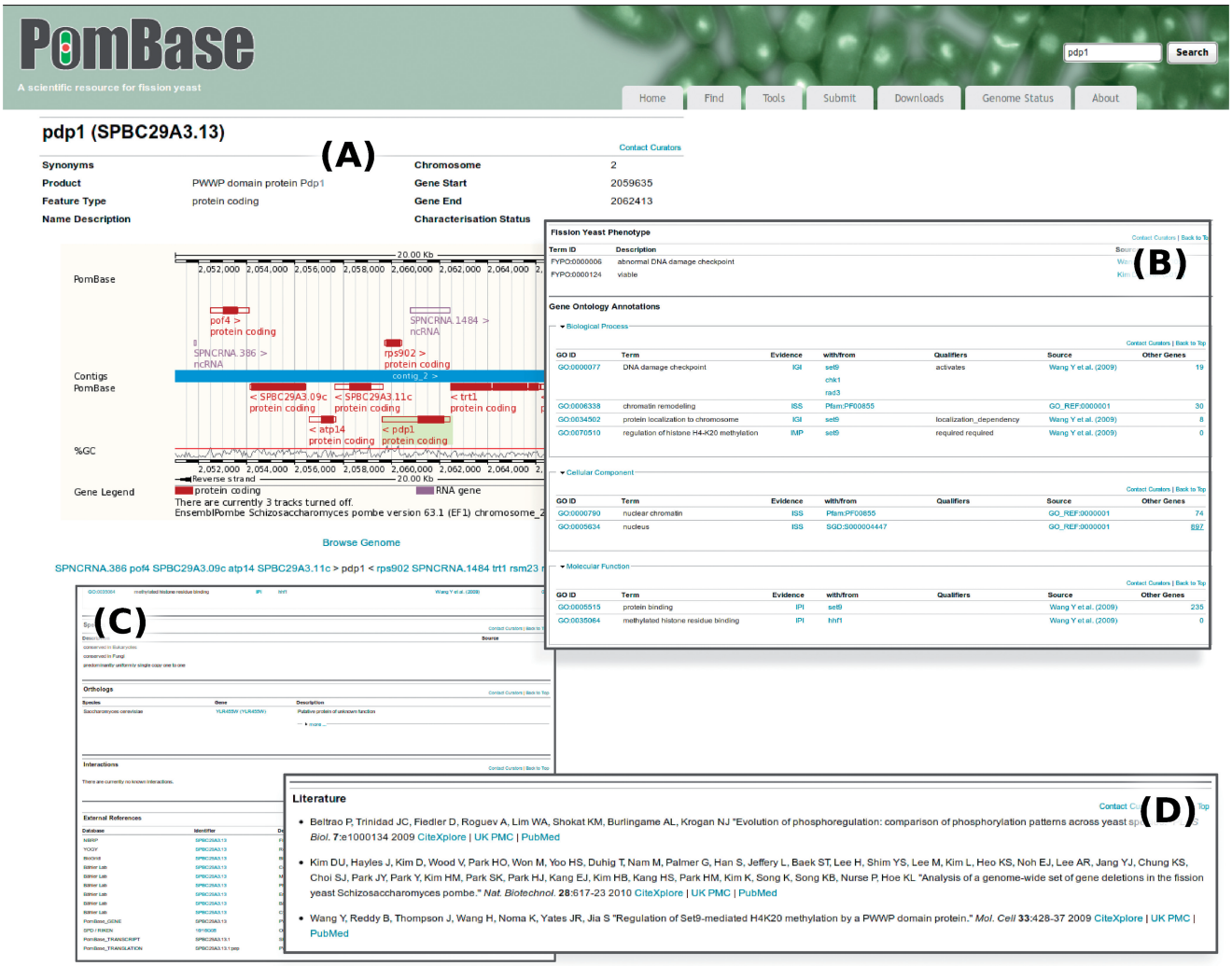

Figure 1. PomBase Gene Overview Page highlights. (**A**) Top of page showing essential information and genomic region. (**B**) Phenotype and GO annotation. (**C**) Additional curated data, including orthologs, interactions and external references. (**D**) Literature pertinent to the gene.

Gene Overview Pages

Gene Overview pages organize gene-specific information, including the gene type, product description, sequence features, phenotypes, Gene Ontology annotation and protein modifications as well as physical and genetic interactions. These pages are central to PomBase (see Figure 1).

Searching

A simple search is available on every PomBase page. The Advanced Search allows users to perform queries on multiple feature types including GO annotation, protein domain, characterization status, species distribution, protein length, etc. A query history summarizes queries and allows them to be edited or combined using union or intersection.

Genome browser

The PomBase genome browser has been implemented using software developed by the Ensembl project (12). The Ensembl genome browser is a powerful system, offering support for the visualization of sequence, functional annotations, alignments, comparative data and polymorphisms. Many of these features are already exploited by PomBase, and others will be used as and when required for the incorporation of new data types. The use of standard technology readily supports comparative analyses with the genomes of other species that are accessible via Ensembl (also see 'Implementation' section, below). Comparative analyses with other fungal genomes and with a range of taxonomically diverse genomes are provided using data generated by the Ensembl Genomes project (13). The Ensembl API enables users interested in comparative analyses to retrieve all data of interest from multiple species present in Ensembl.

IMPLEMENTATION

The implementation of PomBase harnesses and integrates three complementary, well supported and mature technologies, Chado (14), Ensembl and Drupal (http://

www.drupal.org). Chado provides an environment for manual curation and the management of curated data while Ensembl provides end-user access via the web portal and display of sequence-based features.

Chado

Curated PomBase data are stored in a PostgreSQL database using the Generic Model Organism Database (GMOD)-compliant Chado schema. Chado supports the management and storage of sequence annotation and literature curation using any combination of available ontologies, and is therefore easily extensible to new data types. Sequence features are curated in Chado using Artemis (15,16). New annotations produced by the PomBase curation team and the fission yeast community, and external data from BioGRID and UniProt/GOA (17), are loaded at regular intervals.

Ensembl

Ensembl is a generic software platform for the automatic annotation, analysis and display of genomes, in use for over a decade. In PomBase, Ensembl provides public access to the integrated fission yeast data, and has been extended to display the deep literature curation managed in Chado. Gene model and annotation data are retrieved from the Chado database using the Bio::Chado::Schema Perl API, and loaded into an Ensembl MySQL database using the Ensembl Perl API. Data in the Ensembl MySQL database are accessible directly or via the Ensembl Perl API, and provide the content served both in the Genome Browser and in the Gene Overview pages. Data from the Ensembl database are also loaded into a BioMart data warehouse (18,19) which supports data mining via web and programmatic interfaces and also provides support for the advanced search (see below).

Drupal

Drupal is a content management system that has been used to provide the web-based portal for PomBase. The use of Drupal has allowed the creation of a clean and intuitive user interface to access information about *S. pombe* and the PomBase project, and to support community-based functionality including wikis and discussion forums. To support the PomBase interface, two custom Drupal modules have been developed. The Gene Overview module is responsible for generating the Gene Overview pages by retrieving data about specific genes from a custom web service running on the Ensembl web server, which in turn uses the Ensembl Perl API to query the Ensembl MySQL databases. The Query Builder module supports the advanced search interface, and generates and submits custom queries to the BioMart web service to find genes matching specified criteria.

INTEGRATION OF USER DATA

PomBase welcomes contributions from the community to improve the coverage and accuracy of its data. Users wanting to add or modify data in PomBase can directly contact the curation staff (E-mail: helpdesk@pombase.org).

Following a successful pilot project conducted in 2008, a generic web-based curation environment is being developed to support the launch of a comprehensive community curation initiative early in 2012 (Rutherford *et al.*, in preparation). This will allow expert users to directly contribute annotations based on their publications, and will enhance the efforts of core curation staff and contribute to the sustainability of the curation effort in the face of increasing volumes of highly specialized published data.

Additionally, there are many ways that users can directly visualize sequence-based data within the context of the web browser, including access to a Distributed Annotation System, data upload to a private area of the site, and dynamic integration of locally stored BAM files (using standard protocols such as HTTP, allowing users to directly visualize large-scale experimental results in the context of the reference annotation). Producers of mature data sets of any scale that are ready for full integration in PomBase and public dissemination should contact PomBase at the address above.

AVAILABILITY AND DATA PROPAGATION

All of the tools, protocols and workflows developed by PomBase are publicly available (http://www.pombase.org/downloads) and can be implemented by other research communities to create analogous organism-specific databases either in collaboration with the Ensembl Genomes project or independently. Sequence, features and other annotation are available for bulk download via FTP, while subsets of data can be selectively downloaded using the PomBase BioMart.

FUTURE DIRECTIONS

PomBase will continue to incorporate large-scale data sets and curate new data types. We will also incorporate sequence data, automatic annotation, and high-throughput data sets available for other species in the *Schizosaccharomyces* genus (20).

We anticipate that usage of PomBase will extend beyond the *S. pombe* community to encompass evolutionary biologists studying genome variations and the evolution of yeasts, fungi, and the eukaryota in general; researchers seeking well-studied orthologs of genes of interest in human and other species; curators from other databases; and bioinformaticians and theoretical biologists requiring programmatic access to fission yeast data in order to construct and test novel hypotheses.

ACKNOWLEDGEMENTS

The authors thank members of the Ensembl and Ensembl Genomes teams for contributions to the PomBase data pipelines. We also thank Chris Mungall for helpful discussions on Chado and GO, and we thank members of the *S. pombe* research community whose feedback has helped establish priorities for PomBase development.

FUNDING

Wellcome Trust [WT090548MA to SGO]. Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Egel,R. (ed.), (2004) *The Molecular Biology of Schizosaccharomyces pombe*. Springer, Berlin, Germany.
2. Wood,V., Gwilliam,R., Rajandream,M.-A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al*. (2002) The genome sequence of Schizosaccharomyces pombe. *Nature*, **415**, 871–880.
3. Montecchi-Palazzi,L., Beavis,R., Binz,P.-A., Chalkley,R.J., Cottrell,J., Creasy,D., Shofstahl,J., Seymour,S.L. and Garavelli,J.S. (2008) The PSI-MOD community standard for representation of protein modification data. *Nat. Biotechnol.*, **26**, 864–866.
4. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
5. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
6. The Gene Ontology Consortium (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
7. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al*. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.
8. Gkoutos,G.V., Green,E.C.J., Mallon,A.-M., Hancock,J.M. and Davidson,D. (2005) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.
9. deMatos,P., Alcántara,R., Dekker,A., Ennis,M., Hastings,J., Haug,K., Spiteri,I., Turner,S. and Steinbeck,C. (2010) Chemical Entities of Biological Interest: an update. *Nucleic Acids Res.*, **38**, D249–D254.
10. Stark,C., Breitkreutz,B.-J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
11. Stark,C., Breitkreutz,B.-J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Auken,K.V., Wang,X., Shi,X. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
12. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al*. (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
13. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., Kähäri,A. *et al*. (2010) Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res.*, **38**, D563–D569.
14. Mungall,C.J., Emmert,D.B. and FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.
15. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
16. Carver,T., Berriman,M., Tivey,A., Patel,C., Böhme,U., Barrell,B.G., Parkhill,J. and Rajandream,M.-A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.
17. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
18. Smedley,D., Haider,S., Ballester,B., Holland,R., London,D., Thorisson,G. and Kasprzyk,A. (2009) BioMart–biological queries made easy. *BMC Genomics*, **10**, 22.
19. Kinsella,R.J., Kähäri,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al*. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011** (doi: 10.1093/database/bar030; epub ahead of print).
20. Rhind,N., Chen,Z., Yassour,M., Thompson,D.A., Haas,B.J., Habib,N., Wapinski,I., Roy,S., Lin,M.F., Heiman,D.I. *et al*. (2011) Comparative functional genomics of the fission yeasts. *Science*, **332**, 930–936.