BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 5 2004, pages 808–811
DOI: 10.1093/bioinformatics/bth030

# Database model and specification of GermOnline Release 2.0, a cross-species community annotation knowledgebase on germ cell differentiation

C. Wiederkehr[1,†], R. Basavaraj[1,†], C. Sarrauste de Menthière[2,†], R. Koch[1,†], U. Schlecht[1], L. Hermida[1], B. Masdoua[2], R. Ishii[3], V. Cassen[4], M. Yamamoto[3], C. Lane[5], M. Cherry[5], N. Lamb[2] and M. Primig[1,*]

[1]Biozentrum and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland, [2]Institute of Human Genetics, 141, rue de la Cardonille, 34396 Montpellier, France, [3]University of Tokyo, Department of Biophysics and Biochemistry, Hongo, Tokyo 113-0033, Japan, [4]University of Washington, Department of Genome Sciences, Seattle, WA, USA and [5]Stanford University, Stanford, CA 94305-3005, USA

Received on September 9, 2003; revised on November 17, 2003; accepted on November 17, 2003
Advance Access publication January 29, 2004

## ABSTRACT

**Summary:** GermOnline is a web-accessible relational database that enables life scientists to make a significant and sustained contribution to the annotation of genes relevant for the fields of mitosis, meiosis, germ line development and gametogenesis across species. This novel approach to genome annotation includes a platform for knowledge submission and curation as well as microarray data storage and visualization hosted by a global network of servers.

**Availability:** The database is accessible at http://www.germonline.org/. For convenient world-wide access we have set up a network of servers in Europe (http://germonline.unibas.ch/; http://germonline.igh.cnrs.fr/), Japan (http://germonline.biochem.s.u-tokyo.ac.jp/) and USA (http://germonline.yeastgenome.org/).

**Contact:** michael.primig@unibas.ch

**Supplementary information:** Extended documentation of the database is available through the link 'About GermOnline' at the websites.

## INTRODUCTION

Recent advances in genomics have spawned a huge amount of data in the form of DNA sequences, RNA concentrations and protein structures which need to be organized and provided for downloading through web-accessible databases (Baxevanis, 2003). A major challenge to those working in biology and bioinformatics is keeping pace with the genome sequencing facilities which produce DNA sequence data so efficiently that annotation has become the limiting factor. One approach to overcome this problem is to separate the genes into groups relevant for biological processes and to solicit life scientists to contribute their knowledge (Primig *et al*., 2003). Such an effort must be based upon a cross-species approach because fundamental biological processes, e.g. mitotic growth and sexual reproduction, involve conserved genes.

Here we describe the model and the specifications of GermOnline, a web-based platform for cross-species gene annotation and microarray data visualization. The platform is developed by life scientists in cooperation with curators and computer scientists and covers mitosis, meiosis and germ cell development. These processes have been studied for many years in a variety of organisms (Schlecht and Primig, 2003). A detailed description of the novel approach and how to contribute and retrieve information from GermOnline are published elsewhere (Primig *et al*., 2003; Wiederkehr *et al*., 2004). The database presented here will be an extremely useful tool because knowledge in the form of text, images, controlled vocabulary (Yeh *et al*., 2003) and original references (Wheeler *et al*., 2003), presented in the context of high-throughput genomics data, helps researchers interpret information and facilitates hypothesis building.

## SYSTEMS AND METHODS

The server network consists of one master at the Biozentrum (Basel) and three mirrors: Stanford University (Palo Alto),

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

**Table 1.** URLs of sources for locus lists, GeneOntology keywords and lists of potential homologs

| | |
|---|---|
| *Saccharomyces cerevisiae* | ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/orf_geneontology.tab |
| *S.pombe* | ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep |
| *Neurospora crassa* | ftp://ftp-genome.wi.mit.edu/pub/annotation/neurospora/assembly3/genes.csv |
| *Arabidopsis thaliana* | ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Genes/TAIR_sequenced_loci |
| *Z.mays* | ftp://ftp.expasy.org/databases/ |
| *Caenhorabditis elegans* | ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep.table |
| *Drosophila melanogaster* | http://www.flybase.org/allied-data/extdb/external-databases.txt |
| *Danio rerio* | ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL.out.gz |
| *X.laevis* | ftp://ftp.expasy.org/databases/ |
| *Mus musculus* | ftp://ftp.informatics.jax.org/pub/reports/MRK_Sequence.rpt |
| *Rattus norvegicus* | http://rgd.mcw.edu/pub/data_release/GENES |
| *Homo sapiens* | ftp://ftp.ncbi.nih.gov/refseq/LocusLink/LL.out.gz |
| GeneOntology | http://www.godatabase.org/dev/database/archive/latest/ |
| Homologs yeast | ftp://ftp.sanger.ac.uk/pub/yeast/pombe/S_pombe_S_cerevisiae_orthologs/README |
| Homologs NCBI | ftp://ftp.ncbi.nih.gov/pub/HomoloGene/hmlg.ftp |

the IGH (Montpellier) and the University of Tokyo. The database and its web interface are developed using MySQL, PHP, Perl and Javascript. The JpGraph library for PHP is used for graphical display of microarray expression data. The programs and the database are synchronized by the master/slave tool Rsync. Slave servers require Apache 1.3.23, MySQL 3.23, PHP 4.1 with the GD library 1.8.4, Rsync 2.5 as well as Perl 5.0 with the CGI module and the LWP::UserAgent. Standard PC servers equipped with RedHat 7.3 or Solaris 2.8 operating systems are used.

## The GermOnline data sources and interconnectivity

During submission authors choose a locus from lists established by species-specific reference databases or, in the case of *Zea mays* and *Xenopus laevis*, Swiss-Prot (Table 1). A locus is identified by the `orf_id` which is associated with external locus identification numbers, genetic names and aliases as published in the literature and provided by the reference databases and other sources. During the monthly locus updates newly defined genes are inserted and those that have become obsolete are deleted. Should the latter group include curated genes, the authors will be alerted to the conflicting annotation data before the locus is deleted from the database. GermOnline's report pages for conserved genes are directly connected between species; such links are currently available for the yeasts and the mammals.

We seek to establish reciprocal cross-references to other databases. Examples for deep links into GermOnline from external databases are *Schizosaccharomyces pombe* GeneDB [http://germonline.unibas.ch/gene_page.php?tax_id = 4896&sys_sp=SPAC8E11.02c] and Swiss-Prot [http://germonline.unibas.ch/gene_page.php?orf_id=139289].

## The submission/curation process

During submission and curation the author, the curator and GermOnline staff (GeOmaster) interact until the contribution is published (Fig. 1A); automatic e-mails are sent to inform the participants about the state of the process. First, an author selects a locus from a list provided in the pre-submission form. If no previously published contribution is on record an empty form is called up. Should an author update an existing entry, the form contains the corresponding text, keywords, images and references. After submission, the GeOmaster assigns the data to a curator who can accept or reject the submission or ask for revisions. Deletion requests are manually processed by the GeOmaster after consultation with the author.

## Database structure and implementation

The database model is designed to fulfil three major functions (Fig. 1B). First, it organizes information about species and loci. The *Species* table contains a numerical `species_id` that identifies each organism as well as information about the genome, transcriptome and proteome, when available. The *Orf* table contains the locus-specific numerical `orf_id` that can be associated with several external locus identifiers and genetic names as well as their aliases. These names are stored in the `orf_name` attribute field. To identify the source of locus lists, `nomenclature_id` which refers to the *Nomenclature* table, is used (Table 1). This table holds information about the species-specific reference databases, NCBIs Locuslink and Swiss-Prot as well as the respective source uniform resource locators (URLs). The `externalId` establishes a link between the unique identifier of the external database and the `orf_id`. To connect the locus pages of homologues within GermOnline, their corresponding `orf_ids` are stored in the *Ortholog* table. The *Link*, *LinkAssign* and *Layout* tables comprise a collection of links and deep links to individual locus report pages that point to external databases and to relevant studies about meiosis.

Second, the model covers information about the submission and curation process. Authors register and provide data (affiliation, address, contact, e-mail address) stored in tables termed *User*, *Lab*, *User_role*, *Title* and *Country*. Only one entry for each locus from a research group represented by the

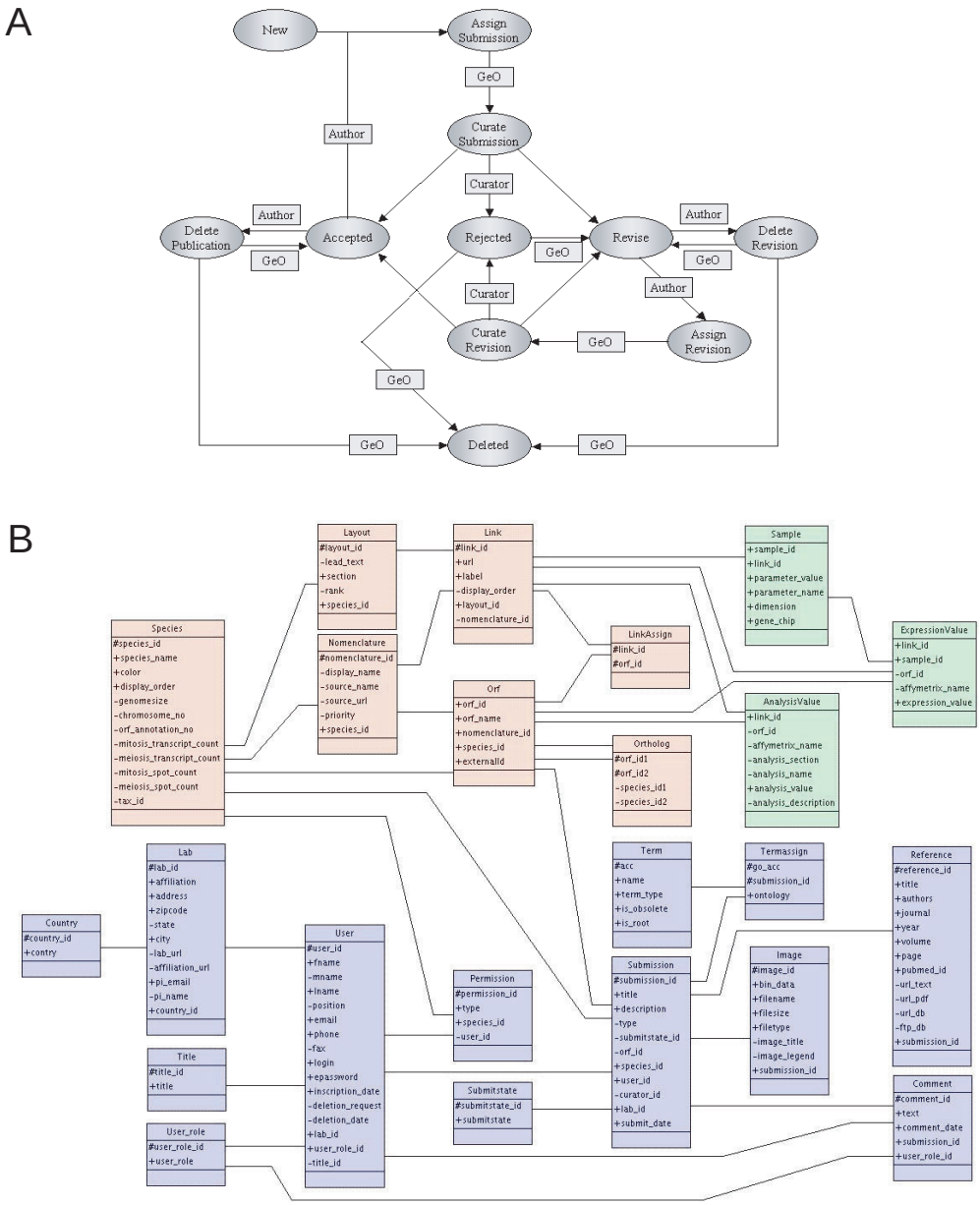


**Fig. 1.** The submission/curation system and the database model. (**A**) Summarizes the life cycle of a submission. (**B**) Shows the database model. GeO is GermOnline staff mediating the interaction between curators and authors. Tables that belong to functional groups are colour coded; species/loci are given in red, submission/curation in blue and expression data in green.

Principal Investigator is allowed. The content of a contribution is stored in tables called *Submission*, *Reference*, *Image*, *Term* and *Termassign*. A submission on a given locus consists of a title, a description, images, controlled vocabulary from the GeneOntology consortium and up to three of the most recent and relevant references (more are possible upon request by the authors). The table *Submitstate* describes the status of a contribution during curation (e.g. submitted, in revision, rejected or published). Request for revisions or other remarks are stored in the *Comments* table allowing authors and curators to communicate. Invited submissions about the genome, transcriptome or proteome of a species are contributed by authors or GeO staff. Since these contributions are not curated like submissions on gene, a special authorization is necessary; this information is stored in the *Permission* table.

Third, the model organizes data from microarray studies displayed as curves or bar diagrams and provides complex queries for genes that display a particular transcriptional pattern (Mata *et al*., 2002; Primig *et al*., 2000; Reinke *et al*., 2000; Schlecht *et al*., 2003; Williams *et al*., 2002). The

expression data is kept in the *Sample* and *ExpressionValue* tables. The lists of loci identified in various expression profiling experiments are stored in the *AnalysisValue* and can be called up through advanced search forms.

## CONCLUSION

We describe the first functional web-based platform for community annotation and microarray data visualization designed to organize data by biological subjects and across species. The project is based upon open source software and all scientists and computer programmers working in the field of database development are invited to contribute their ideas.

## ACKNOWLEDGEMENTS

We thank N.L. for initial help in setting up the database, R. Jenni, D. Flanders (FMI) and R. Poehlman for excellent IT infrastructure support within the framework of the Basel Computational Biology Center [$(BC)^2$] and M. Aslett for critical reading of the manuscript. We are indebted to the members of the board of scientists who oversee the project. GermOnline workshops were sponsored by the National Science Foundation (USA), the Swiss National Science Foundation (SNF), INSERM and CNRS. This project is funded by the Swiss Institute of Bioinformatics.

## REFERENCES

Baxevanis,A.D. (2003) The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res.*, **31**, 1–12.

Mata,J., Lyne,R., Burns,G. and Bahler,J. (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nat. Genet.*, **32**, 143–147.

Primig,M., Wiederkehr,C., Basavaraj,R., Sarrauste de Menthière,C., Hermida,L., Koch,R., Schlecht,U., Dickinson,H., Fellous,M., Grootegoed,A. (2003) GermOnline: a novel cross-species community annotation database on germ line development and gametogenesis. *Nat. Genet.*, **35**, 291–292.

Primig,M., Williams,R.M., Winzeler,E.A., Tevzadze,G.G., Conway,A.R., Hwang,S.Y., Davis,R.W. and Esposito,R.E. (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet.*, **26**, 415–423.

Reinke,V., Smith,H., Nance,J., Wang,J., Van Doren,C., Begley,R., Jones,S., Davis,E., Scherer,S., Ward,S. and Kim,S. (2000) A global profile of germline gene expression in *C.elegans*. *Mol. Cell.*, **6**, 605–616.

Schlecht,U., Demougin,P., Koch,R., Hermida,L., Wiederkehr,C., Descombes,P., Pineau,C., Jégou,B. and Primig,M. (2003) Expression profiling of mammalian male meiosis and gamete development identifies novel candidate genes for roles in the regulation of fertility. *Mol. Biol. Cell* (in press).

Schlecht,U. and Primig,M. (2003) Mining meiosis and gametogenesis with DNA microarrays. *Reproduction*, **125**, 447–456.

Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.

Wiederkehr,C., Basavaraj,R., Sarrauste de Menthiere,C., Hermida,L., Koch,R., Schlecht,U., Amon,A., Brachat,S., Breitenbach,M. and Briza,P. (2004) GermOnline, a cross-species community knowledgebase on germ cell growth and development. *Nucleic Acids Res.* **32**, 560–567.

Williams,R.M., Primig,M., Washburn,B.K., Winzeler,E.A., Bellis,M., Sarrauste de Menthiere,C., Davis,R.W. and Esposito,R.E. (2002) The Ume6 regulon coordinates metabolic and meiotic gene expression in yeast. *Proc. Natl Acad. Sci., USA*, **99**, 13431–13436.

Yeh,I., Karp,P.D., Noy,N.F. and Altman,R.B. (2003) Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO). *Bioinformatics*, **19**, 241–248.