



GermOnline, a new cross-species community annotation database on germ-line development and gametogenesis

To the editor:

We report the development of a unique knowledge base of phenotypic and gene expression data relevant to germ-line development and gametogenesis. GermOnline includes an interactive platform for cross-species gene annotation by research scientists. This new approach to knowledge management partitions genomes of key model systems into groups of genes involved in important conserved biological processes.

There is now extensive information about genes involved in sexual reproduction in many species, and a platform that allows comparison of data from different organisms is extremely useful. Insights emerging from comparative analysis across species should enhance our understanding of the underlying mechanisms that cause errors

resulting in human infertility and birth defects¹ and may help develop new approaches to contraception.

Current knowledge bases like Swiss-Prot and various organism-specific databases (e.g., Saccharomyces Genome Database, Flybase, Wormbase and Mouse Genome Database) are maintained by professional curators who must read and interpret a rapidly increasing volume of scientific literature to create database entries on individual loci². GermOnline includes a different approach to gene annotation based on direct interaction with the scientific community. It allows researchers to directly contribute and update knowledge about genes of interest, in cooperation with a team of database developers and scientific curators. This process permits up-to-date information to be readily available and

facilitates species comparisons in a wide variety of areas that are now laborious to execute. The database enables authors to provide their published knowledge using an online form that includes text, images, GeneOntology keywords³ and automatic retrieval of original references. Scientists summarize their published findings on a specific gene in a single updatable entry. Multiple entries on a given gene by different scientists are encouraged to obtain coverage of all aspects of a gene's functions. These contributions are read by a board of scientific curators to ensure a uniform data format and the highest quality information content. The utility of expert knowledge is increased by inclusion of relevant high-throughput expression profiling and proteome data. The locus report pages for homologous genes from different species are

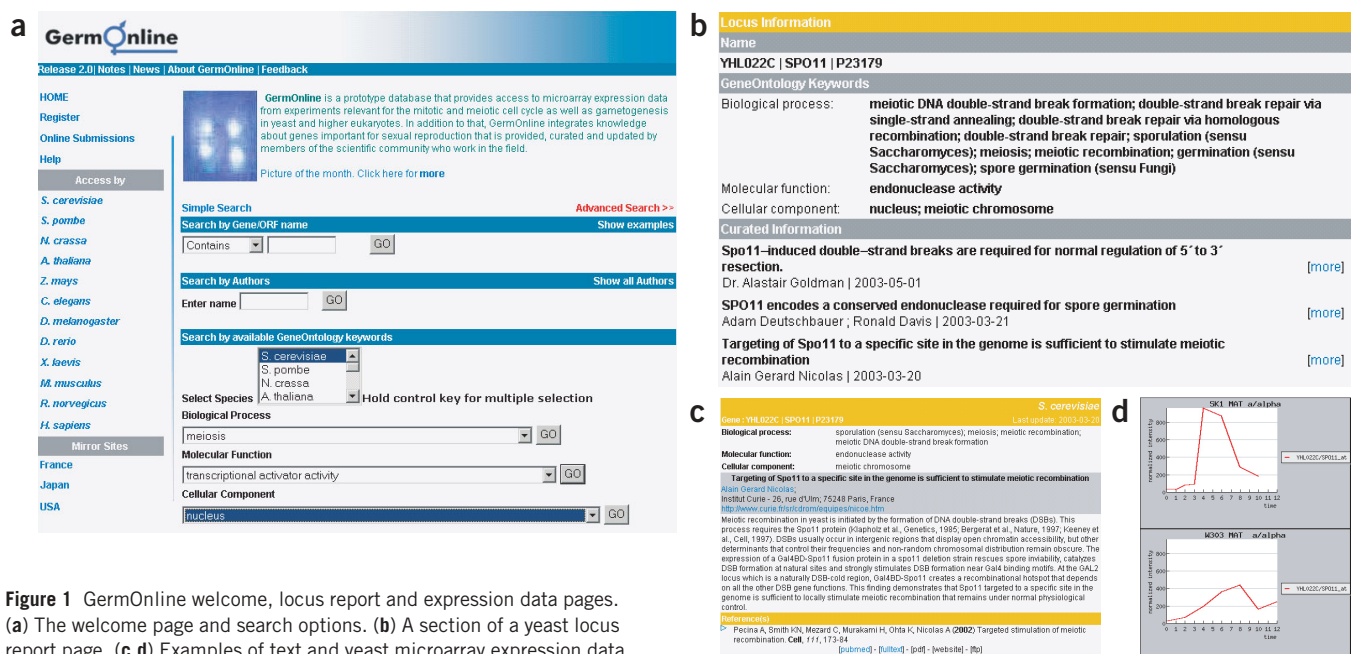


Figure 1 GermOnline welcome, locus report and expression data pages. (a) The welcome page and search options. (b) A section of a yeast locus report page. (c,d) Examples of text and yeast microarray expression data.

directly linked based on large-scale similarity searches retrieved from HomoloGene (National Center for Biotechnology Information) and GeneDB (Sanger Centre)^{4,5}. GermOnline can be searched by author name, gene name, keyword, expression profile and phenotype⁶ (Fig. 1a). The locus report pages provide curated information (Fig. 1b) and links to relevant external data sources. Likewise, major databases, such as Swiss-Prot, Saccharomyces Genome Database and GeneDB, provide links to our database. GermOnline is especially useful for genome biologists, who often need to process and interpret data on large numbers of genes for which extensive literature is available.

GermOnline Release 2.0 contains information on 684 genes from *Saccharomyces cerevisiae* involved in meiosis (Fig. 1c), spore formation and germination, as well as about 30 prototype contributions from other species. Microarray expression data covering the yeast cell cycle⁷, yeast sporulation^{8,9} (Fig. 1d) and spermatogenesis in the rat (U.S. *et al.*, unpublished data) are provided, as well as external links to relevant studies using *S. cerevisiae*,

Schizosaccharomyces pombe and *Caenorhabditis elegans*. Release 2.0 is accessible at <http://www.germonline.org>. Detailed descriptions of how to retrieve and contribute information, as well as the database model and specifications, will be published elsewhere¹⁰. The approach described is applicable to a wide variety of conserved biological processes studied in different species, including *Homo sapiens*.

ACKNOWLEDGMENTS

We thank B. Braun for discussions, B. Masdoua and V. Cassen for contributing to programme development and M. Cherry and M. Yamamoto for hosting the database on local servers. We acknowledge funding for Workshops from the SNF, CNRS, INSERM and the US National Science Foundation. The initial stages of this project were supported by an EU grant awarded to N.L. Continuing support is provided by a grant from the Swiss Institute of Bioinformatics awarded to M.P. Software and database developers are C.W., R.B., C.S.d.M., L.H., R.K. and U.S.

M Primig¹, C Wiederkehr¹, R Basavaraj¹, C Sarrauste de Menthère², L Hermida¹, R Koch¹, U Schlecht¹, H G Dickinson³, M Fellous⁴, J A Grootegoed⁵, R S Hawley⁶, B Jégou⁷, B Maro⁸, A Nicolas⁹, T Orr-Weaver¹⁰, T Schedl¹¹, A Villeneuve¹², D J Wolgemuth¹³, M Yamamoto¹⁴, D Zickler¹⁵, N Lamb² & R E Esposito¹⁶

¹Biozentrum & Swiss Institute of Bioinformatics, Basel, Switzerland. ²Institut de Génétique Humaine, Montpellier, France. ³Oxford University, UK. ⁴Hopital Cochin, Paris, France. ⁵Erasmus MC, Rotterdam, The Netherlands. ⁶The Stowers Institute, Kansas City, Missouri, USA. ⁷The University of Rennes I, Bretagne, France. ⁸Université Paris VI, France. ⁹Institut Curie, Paris, France. ¹⁰The Whitehead Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. ¹¹Washington University, St. Louis, Missouri, USA. ¹²Stanford University, Palo Alto, California, USA. ¹³Columbia University, New York, New York, USA. ¹⁴The University of Tokyo, Japan. ¹⁵Université Paris XI, Orsay, France. ¹⁶The University of Chicago, Chicago, Illinois, USA. Correspondence should be addressed to M.P. (michael.primig@unibas.ch).

1. Nasmyth, K. *Annu. Rev. Genet.* **35**, 673–745 (2001).
2. Baxevanis, A.D. *Nucleic Acids Res.* **31**, 1–12 (2003).
3. Yeh, I., Karp, P.D., Noy, N.F. & Altman, R.B. *Bioinformatics* **19**, 241–248 (2003).
4. Wheeler, D.L. *et al. Nucleic Acids Res.* **31**, 28–33 (2003).
5. Wood, V. *et al. Nature* **415**, 871–880 (2002).
6. Schlecht, U. & Primig, M. *Reproduction* **125**, 447–456 (2003).
7. Cho, R.J. *et al. Mol. Cell* **2**, 65–73 (1998).
8. Primig, M. *et al. Nat. Genet.* **26**, 415–423 (2000).
9. Williams, R.M. *et al. Proc. Natl. Acad. Sci. USA* **99**, 13431–13436 (2002).
10. Wiederkehr, C. *et al. Nucleic Acids Res.* (in the press).

Chipping away at the chip bias: RNA degradation in microarray analysis

To the editor:

Measurement of gene expression is based on the assumption that an analyzed RNA sample closely represents the amount of transcripts *in vivo*. Transcripts show stability differences of up to two orders of magnitude *in vivo*¹, raising the possibility that partial degradation during cell lysis could cause a variable extent of bias in quantification of different transcripts. One of the most effective tools for characterizing RNA integrity is capillary electrophoresis, in which RNA degradation is indicated by an altered 28S/18S rRNA signal ratio². In the software of the commonly used system (Bioanalyzer 2100, Agilent), quantification of 18S and 28S rRNA is compromised by the fact that this calculation is based on area measurements that are heavily dependent on definitions of start and end points of peaks (Fig. 1a). Even accurate determination of this ratio is not sufficient to detect degradation efficiently (Fig. 1b). We developed a mathematical model that results in an objective number for quantitative characterization of RNA degradation. Aside

from three prominent peaks (small RNAs, 18S and 28S rRNA), a chromatogram of the size distribution of cellular RNAs shows a broad range of molecular weights with much weaker signals. With increasing degradation, heights of 18S and 28S peaks gradually decrease and additional 'degradation peak signals' appear in a molecular weight range between small RNAs and the 18S peak (Fig. 1b). The ratio of the average degradation peak signal to the 18S peak signal multiplied by 100 will hereafter be referred to as the degradation factor. This analysis has been tested on 19 tissues of seven organisms, and it is a reproducible parameter for degradation of mammalian RNA (Supplementary Table 1 online). As an example, 12 repeated measurements of the same sample yielded an average degradation factor of 27.14 with a standard deviation of 1.06. Degradometer software for calculation of the degradation factor can be downloaded from <http://www.dnaarrays.org>.

If one RNA sample was intact and the other was degraded during isolation, up to

three-quarters of the differential gene expression measured was due solely to differences in RNA integrity between two samples (Fig. 1d). Supplementary Figure 1 online shows changes in mRNA levels caused by alteration of RNA integrity. This effect was independent of the algorithm applied to raw data analysis (Supplementary Tables 2, 3 and 4 online).

For *GAPD* and *ACTB*, two transcripts for which signal intensities from 3' and 5' portions are frequently measured in microarray analysis, there is a positive correlation between the 3'/5' ratio and the degradation factor of samples (Fig. 1c). This correlation is tissue-dependent (Supplementary Fig. 2 online). The smaller the difference in degradation factors between samples, the more closely the measured expression differences reflect biological differences (Fig. 1d).

Aside from general RNase activity by members of the RNase A family³, RNase L, an enzyme activated in apoptotic