GermOnline, a cross-species community knowledgebase on germ cell differentiation

C. Wiederkehr, R. Basavaraj, C. Sarrauste de Menthière¹, L. Hermida, R. Koch, U. Schlecht, A. Amon², S. Brachat, M. Breitenbach³, P. Briza³, S. Caburet⁴, M. Cherry⁵, R. Davis⁵, A. Deutschbauer⁵, H. G. Dickinson⁶, T. Dumitrescu⁷, M. Fellous⁴, A. Goldman⁸, J. A. Grootegoed⁹, R. Hawley¹⁰, R. Ishii¹¹, B. Jégou¹², R. J. Kaufman¹³, F. Klein¹⁴, N. Lamb¹, B. Maro¹⁵, K. Nasmyth¹⁶, A. Nicolas¹⁷, T. Orr-Weaver¹⁸, P. Philippsen, C. Pineau¹², K. P. Rabitsch¹⁶, V. Reinke¹⁹, H. Roest⁹, W. Saunders⁷, M. Schröder¹³, T. Schedl²⁰, M. Siep⁹, A. Villeneuve⁵, D. J. Wolgemuth²¹, M. Yamamoto¹¹, D. Zickler²², R. E. Esposito²³ and M. Primig^{*}

Biozentrum and Swiss Institute of Bioinformatics, Basel, Switzerland, ¹Institut de Génétique Humaine, Montpellier, France, ²Howard Hughes Medical Institute, MIT, Cambridge, MA, USA, ³University of Salzburg, Austria, ⁴Hôpital Cochin, Paris, France, ⁵Stanford University, Palo Alto, CA, USA, ⁶Oxford University, UK, ⁷University of Pittsburgh, PA, USA, ⁸University of Sheffield, UK, ⁹Erasmus University, Rotterdam, The Netherlands, ¹⁰Stowers Institute, Kansas City, KS, USA, ¹¹University of Tokyo, Japan, ¹²University of Rennes I, Bretagne, France, ¹³Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA, ¹⁴University of Vienna, Austria, ¹⁵Université Paris VI, France, ¹⁶Institute of Molecular Pathology, Vienna, Austria, ¹⁷Institut Curie, Paris, France, ¹⁸Whitehead Institute, MIT, Cambridge, MA, USA, ¹⁹Yale University, New Haven, CT, USA, ²⁰Washington University, St Louis, MO, USA, ²¹Columbia University, New York, NY, USA, ²²Université Paris Sud, Orsay, France and ²³University of Chicago, IL, USA

Received August 15, 2003; Revised and Accepted September 23, 2003

ABSTRACT

GermOnline provides information and microarray expression data for genes involved in mitosis and meiosis, gamete formation and germ line development across species. The database has been developed, and is being curated and updated, by life scientists in cooperation with bioinformaticists. Information is contributed through an online form using free text, images and the controlled vocabulary developed by the GeneOntology Consortium. Authors provide up to three references in support of their contribution. The database is governed by an international board of scientists to ensure a standardized data format and the highest quality of GermOnline's information content. Release 2.0 provides exclusive access to microarray expression data from Saccharomyces cerevisiae and Rattus norvegicus, as well as curated information on ~700 genes from various organisms. The locus report pages include links to external databases that contain relevant annotation, microarray expression and proteome data. Conversely, the Saccharomyces Genome Database (SGD), S.cerevisiae GeneDB and Swiss-Prot link to the budding yeast section of GermOnline from their respective locus pages. GermOnline, a fully operational prototype subjectoriented knowledgebase designed for community annotation and array data visualization, is accessible at http://www.germonline.org. The target audience includes researchers who work on mitotic cell division, meiosis, gametogenesis, germ line development, human reproductive health and comparative genomics.

INTRODUCTION

The recent developments of large-scale DNA sequencing techniques, microarray technology and bioinformatics enable scientists to identify potential genes, study their patterns of expression and analyse their promoters at the genomic level (1). Novel genetic and biochemical high-throughput approaches to determining the functions and interactions of gene products further contribute to the rapidly growing body of data that biologists and bioinformaticists need to process and interpret (2–5). To meet this challenge, a common biological language is being developed by the GeneOntology (GO) Consortium. This project is a collaborative effort of many major genome databases e.g. the *Saccharomyces* Genome Database (SGD), The Arabidopsis Information Resource (TAIR), Flybase, Wormbase and the Mouse

*To whom correspondence should be addressed. Tel: +41 61 267 2098; Fax: +41 61 267 3398; Email: michael.primig@unibas.ch

arch by Gene/ORF name Show exam ontains GO arch by Authors Show all Authors arch by Authors Show all Authors arch by available GeneOntology keywords Show all Authors arch by available GeneOntology keywords Show all Authors arch by available GeneOntology keywords Show all Authors lect Species A theliana Hold control key for multiple selection logical Process elosis So lect Species A theliana lect Species So of the selection lular Component So of the selection	Simple	Search			Advanced			
ortains Image: Control by Authons arch by Authons Image: Control by Available Control by keywords arch by available ConeOntology keywords Image: Control by Authons Image: Control by Authons Image: Control by Control key for multiple selection Control by Control by Control key for multiple selection Control by Control by Control key for multiple selection Control by Co	Search	by Gene	e/ORF i	name		Show exam		
arch by Authons Show all Aut ter name Image: Show all Aut arch by available GeneOntology keywords arch by available GeneOntology keywords Arch assa Hold control key for multiple selection obigical Process elocital Function Image: Show all Aut Interference Image: Show all Aut Provide Automatic A	Contair	ns 💌	í 🗌		GO			
arch by Authors Show all Aut ter name GO arch by available GeneOntology keywords arch by available GeneOntology keywords A thelan Hold control key for multiple selection ological Process eiosis eiosis Processing office set for any office set Control Ilular Component Control Ilular Component Control International Processing office set Control International Processing office set International Processing of International Processing of International Processing of Int								
ter name	Search	by Auth	ors			Show all Auth		
<pre>during</pre>	Entor n			GO	1			
arch by available GeneOntology keywords S. pombe S. pombe A crass a Hold control key for multiple selection logical Process S. pombe logical Function S. pombe	Luter n	ame						
And if if y available Concertioning Regivering S. pombe N. crassa A. thaliana Hold control key for multiple selection alogical Process elocities eloc	Search	hu avail	ahla G	anaOntal	ony kosword			
	Search	ny avan		eneontoi	ogy keyword	5		
			S. cere	evisiae				
			N. cras	ssa				
	Select !	Species	A. thal	iana	- Hold co	ntrol key for multiple selection		
	Biologi	cal Proc	229					
Initial initi								
	meiosi	s						
	Molecu	lar Func	tion					
Idea Composed Image:						- GO		
Fundame Components Image: Components Sciences Sciences <td>C-11 1</td> <td></td> <td></td> <td></td> <td></td> <td></td>	C-11 1							
Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences Sciences	Cellula	r Compo	nent					
						▼ GU		
Screwbirks Screwbirks Screwbirks Screwbirks								
Decempoind product of the second pro	_				C an	raulatan		
Depression profile south Depression profile south Choose name © systematic name © Standard name Induced in Net induced in © Standard name Induced in Net induced in © Standard name Induced in Stat Mat alphna © Standard name Induced in Net induced in © Standard name Induced in Net induced in © Standard name Induced in Net induced in © Standard name Induced in Not alphnabin W033 MAT alphnabin Induced in W033 MAT alphnabin W033 MAT alphnabin Indue Status Indue Status © Standard name Indue Status Indue Status Indue Status Indue Status			E-market	tion and the second	J. 60	Everante state		
	C	hoose name	Expres	sion profile search		Expression profile search Choose name		
Induced in Not modeled in Induced in Not modeled in Induced in Sk1 MAT alaphaa Sk1 MAT alaphaa Sk1 MAT alaphaa W 2003 MAT alaphaa Sk1 MAT alaphaa W 2003 MAT alaphaa W 2003 MAT alaphaa Bergession cluster Image alaphaa W 2003 MAT alaphaa W 2003 MAT alaphaa Bergession cluster Image alaphaa W 2003 MAT alaphaa Image alaphaalaphaa W 2003 MAT alap	1	🕫 Systematic nam	e	C Standard name		Systematic name Standard name		
Is in the response to standardipheremene (sanstein) Is in the response to standardipherem	In	Induced in		Not induced in		Choose gene list		
Wrdd3 MAT atsipha Wrd303 MAT atsipha Wrdd3 MAT atsipha Wrd303 MAT atsipha Regulatory motif Wrd303 MAT atsipha Regulatory motif Wrd303 MAT atsipha Regulatory motif Wrd303 MAT atsipha Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation (detayed) Brenes induced in response to starvation induced in response instarvation indu	i	C SKI MAT alpha/alpha		E SK1 MAT alpha/alpha		C Genes induced in response to starvation/pheromone (transient)		
Wy303 MAT w/s Wy303 MAT w/s Regulation (math Wy303 MAT w/s Regulation (math Wy303 MAT w/s Regulation (math Wy303 MAT w/s Beness induced in response to introgen starvation (delayed) Entry series 0 1 2 4 5 6 7 Correlation (cofficient between Correlation (cofficient between Correlation (cofficient between 1<0	r	W303 MAT alalp	ha	🗆 W303 MAT a/alpha		C Genes induced in response to starvation/pheromone (continuous)		
Regulatory motif URS1b URS1b URS1b Expression cluster Indiate genes Indiate genes 0 1 2 3 1 2 3 Indiate genes Correlation (coefficient batween Indiate genes Indiate genes 1 3 10 Indiate genes Correlation (coefficient batween Indiate genes Indiate genes 1 3 Indiate genes Indiate genes 1 3 Indiate genes Indiate genes Correlation (coefficient batween Indiate genes Indiate genes Indiate genes 1 3 Indiate genes Indiate genes Indiate genes Correlation (coefficient batween Indiate genes Indiate genes Indiate genes 1 3 Standard name Indiate genes Indiate genes Indiate genes Indiate genesis/tem-3 (spermatogenesis/tem-3 (spermatogenesis) Indiate genesis Indiate genes Indiate genesis <t< td=""><td>ſ</td><td>🗆 W303 MAT ala</td><td></td><td colspan="2">C W303 MAT ara</td><td>Genes induced in response to nitrogen starvation (delayed)</td></t<>	ſ	🗆 W303 MAT ala		C W303 MAT ara		Genes induced in response to nitrogen starvation (delayed)		
Expression cluster	R	egulatory motif	IT MRE			Middle genes		
Image: Standard name Image: Standard name Choose name Standard name Imm- Smalehim-S.glp-4 (no germine) male Immed testing test	E	apression cluster		· · ······		Late genes		
Image: Correlation coefficient bathween Image: Correlation coefficient bathween Image: Correlation coefficient bathween Image: Correlation coeffi	1	0 0	E1	E 2	E 3	In involve and potential transcriptional regulators induced during melosi sporulation		
Initian and Initian Initian and Initian Expression profile search	i c	4 Correlation coefficie	nt between	IT 6	10.7	Core meiotic transcriptome conserved in fission and budding yeasts		
C. clogans Expression profile search Choose name Choose pane list Choose pane l	ſ	-1.0 and 1.0 and						
Expression profile search Veloceme to the rat eXpression database © systematic name © Standard name © hoose name © Standard name © home gene list Image: Standard name © home Signer signer © standard name © with emaphrodite/lipo-1 (soggenerise) © standard name © with emaphrodite/lipo-3 (name © with emaphrodite/lipo-3 (soggenerise) © with emaphrodite/lipo-3 male © wi	1							
Expression profile search					~	elegane		
Veloceme to the rate Xpression database Choose gene list Welcome to the rate Xpression database Choose gene list The can search to reducid list anong the complex priser as the USA > B Own-Chap (all to a finite to a Mean of Mean) Choose gene list The can search to reducid list anong the complex priser as the USA > B Own-Chap (all to a finite to a Mean of Mean) Choose gene list The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce difference The can search to reduce di					0.1	GermOnline		
Wetcome to the rate Xpression database © Systematic name © Standard name Choose gene list Image: Standard name © fem-3 (spermatiogenesis) Image: Standard name © hm-5 mateRim-5,gip-4 (no germline) male Image: Standard name Image: Num-5 mateRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 mateRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 mateRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 maleRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 maleRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 maleRim-5,gip-4 (no germline) hermaphrodte Image: Standard name Image: Num-5 maleRim-5, male Image: Standard name Image: Num-5 male Image: Standard name Image: Num-5 maleRim Num-5 male Image: Standard name Image: Num-5 male Image: Standard name Image: Num-5 maleRim Num-5 male <td>3</td> <td>Choose name</td> <td>Expre</td> <td>ssion profile search</td> <td></td> <td>Welcome to the est of the second seco</td>	3	Choose name	Expre	ssion profile search		Welcome to the est of the second seco		
Choose gene list Choos		Systematic na	me	C Standa	d name	wercome to the rat expression database		
Important		Choose gene list				This can easish by individual lists arranging the complete set of panes present an the USAA + B-GeneChoo (with a a limited ant of panes (select e.g. Meinte cluster).		
C useden (C Lussee C Luss	fem-1 (oogenesis)/tem-3 (spermatogenesis) fem-3 (spermatogenesis)/tem-1 (oogenesis)			ermatogenesis)		Search By		
Imin-5 matehim-5.glp-4 (no germline) male Imin-5 male/imin-5 male Imin-5 male/imin-5 male Imin-5 male/imin-5 male Imin-5 male/imin-5 male Imin-5 male/imin-5 male				n-1 (oogenesis)		Bankai C Subban D C LouisLik D C Meditae		
Imin-5 male/wt hermaphrod/te Imin-5 male/wt hermaphrod/te/min-5 male Imin-5 male/wt hermaphrod/te/min-5 male Imin-5 male/mto/te/min-5 male		T him-5 mate/him-5;gtp-4 (no germline) male				1 (0.), C 38 OF H1.16514 OF 25678 OF speech of 4900002_ab		
Image: with thermaphrod targing 4 (no. germline) hermsphrodite Planetine or gradual displays Image: with hermaphrod targing 6 and a second displays Planetine or gradual displays Image: with hermaphrod targing 6 and a second displays Planetine or gradual displays Image: with hermaphrod targing 6 and a second displays Planetine or gradual displays Image: with hermaphrod targing 6 and a second displays Planetine or gradual displays Image: with hermaphrod targing 6 and a second displays Planetine or gradual displays and a second display and a second displays and a second display and a second display and a second displays	Inm-5 maleWet hermaphrodBe wit hermaphrodBe wit hermaphrodBe(gip-4 (no germline) hermaphrodBe wit hermaphrodBe(gip-4 (no germline)					You can also retrieve groups offics as defined in this study (e.g. select pool-Meistic Chutler to call up 115 probe		
Wet hermaphrod/fie/him-5 maie Conservation C					ite	Search for groups of genes		
C Break Cataly C Materia Cataly C Materia Cataly C particular Cataly C particular States C Pater State game and pages and pages C Materialy game and pages C Materialy game and pages C Materialy game and pages C Materialy game and pages C Material States C Material						P 1400 septembraneright		
 Mentin Custome Mentin Custome Providentification Providentification		r : va namahungnanus a maai				C Senate Custor		
C passimiliarité Counter C Passimiliarité passimiliarité passimilie funcionalité C separate 2015 v vert nous sandaute pas annotation protein						C Melats Cluster		
C Potentially generic cut typoloh transcriptio C regulated ESTs with new sensitivity to an according to an						C postamoistis Cluster		
 regulated ESTs with new strateging to an entropy of the strategy to an entropy of the strategy of						C Polostale gave call specific barrough		
Compared Birth with an annual to annual to an						C regulated EDTs with new similarity to an annotated protein C recorded eDTs with no similarity in any constraint an annotated proteins		

Figure 1. Simple and advanced search options. Users can search for individual genes and authors who contributed to the database as well as GO keywords that were associated with the gene in question during curation as shown in (A). Advanced search forms that enable the user to retrieve lists of genes that display a certain expression pattern or have other properties (e.g. the presence of known regulatory sites in their promoters) are summarized for several species as indicated in (B).

Genome Database (MGD) (6). The aim is to develop a coherent semantic framework for gene nomenclature and description across species. The controlled vocabulary describes the biological process genes are involved in, the molecular functions their products fulfil and the cellular component to which gene products localize (7).

A common feature of genome databases is that each contains general information on a particular species (or genus) and it is often laborious to compare evolutionarily conserved loci across organisms from within their locus pages. Genome databases are developed and maintained by curators and bioinformaticists who are responsible for defining and naming a locus through a unique identification code and gene annotation in cooperation with the members of sequencing consortia. While researchers provide the raw data they are usually not directly involved in the formulation of gene descriptions, the choice of GO keywords or the decision as to how much emphasis is given to models and theories discussed in the literature. Some databases have recently begun to solicit contributions from users for information content, e.g. SGD (8); however, this remains supplementary information and does not constitute the core knowledge provided by the database.

GermOnline is a new, unique comprehensive subjectoriented database containing information from diverse species specifically focused on mitotic growth and meiotic development. These important biological processes have been extensively studied in a variety of model organisms including *Homo sapiens* and a substantial amount of knowledge and genomics data is available that needs to be organized; the field is therefore particularly suitable for community-based annotation presented in the context of high-throughput data. Information about gene function in experimental systems used to study sexual reproduction will ultimately help to improve human reproductive health.

Scientists who publish their findings in peer-reviewed journals contribute the essence of their results using free text, images and/or figures (including legends), automatically updated GO keywords and original references. This information is presented within the context of relevant microarray and proteomics data as well as other external sources of gene annotation. Multiple contributions on a given locus from different groups are encouraged to ensure that gene function(s) are broadly discussed and different perspectives covered in depth. The present paper describes how scientists can retrieve data from and contribute information to GermOnline and outlines the curation procedure. The latter is overseen by the members of an international board of scientists who work in the fields of meiosis, germ line development and gametogenesis using a variety of model organisms.

Programmatic aspects of the database model, database interconnectivity and mirroring as well as automatic update of

locus lists and GO keywords will be published elsewhere. The database is accessible at http://www.germonline.org. To ensure round-the-clock availability and convenient access, a network of mirrors in Europe (http://germonline.igh. cnrs.fr), Japan (http://germonline.biochem.s.u-tokyo.ac.jp/) and the US (http://germonline.yeastgenome.org) has been installed.

THE SCOPE OF GermOnline

It is the aim of GermOnline to provide the most up-to-date expert information on all genes involved in germ cell growth and development across species. GermOnline currently covers 11 model organisms including yeasts (Saccharomyces cerevisiae, Schizosaccharomyces pombe, Neurospora crassa), plants (Aribidopsis thaliana, Zea mays), a nematode (Caenorhabditis elegans), an arthropod (Drosophila melanogaster), vertebrates (Danio rerio, Xenopus laevis), mammals (Mus musculus, Rattus norvegicus) and Homo sapiens. Approximately 190 000 locus pages and 475 000 locus names (including aliases) are available for queries. In over 700 cases of highly relevant genes, the locus reports contain curated information from one or several scientists in addition to links to useful internal and external information sources. S.cerevisiae is the first organism for which genes involved in meiosis, spore formation and post-germination vegetative growth are practically all represented in the database. Work on this species is therefore considered proof of principle since GermOnline provides microarray expression data for nearly all the 6000 known yeast genes and curated information for ~680 genes (many with multiple entries). About 30 genes are available as prototype contributions for other species at the time of writing this paper. Scientists will now be encouraged to participate in this global project and contribute their data and knowledge. Contact with researchers who work on the species represented in GermOnline will be established through a sustained information campaign that includes talks and poster presentations at scientific meetings, a 'call for contributions' advertisement in various scientific journals as well as solicitation by electronic and regular mail. Since the yeast community responded very positively to the project, we fully expect the same reaction from scientists who use higher eukaryotic model systems.

HOW TO RETRIEVE INFORMATION

Search options

There are several entry points into GermOnline. The welcome page contains simple search options to access the locus page of a particular gene using either the systematic or genetic name, gene symbol or any alias provided by the relevant speciesspecific databases. Users can also search for authors who have

Figure 2. The locus page and expression data samples. (A) The section of the locus report page that summarizes the GO keywords and the title of curated contributions from different authors on the SPO11 genes from yeast and worm. Also shown is the gene-info box, which contains the author's contact information, the respective gene descriptions and the original reference relevant for the contribution. (B) Examples of the graphical display of various expression data sets on yeast sporulation (left), transcriptional deregulation in the absence of the Ume6 transcription factor (top right) and gene expression in purified rat Sertoli cells and testicular germ cells (bottom right). (C) An example of how data sources in the locus page are referenced: the title of the most recent publication as well as links to PubMed and (whenever available) relevant websites, search forms and specific deep links to locus pages (more) are given.

and unes

770

610

510 4(1)

- YHL022C/SP011_at

X75785_at

				and the set of the set
		Manual of Street	haulture	Bacchardmands, such a principality particular particular Bacchardmands, such perincipality particular particular particular social sectors and sectors and sectors and sectors and sectors and a sector sector sectors and and and and and and and and
		Tables res	Mail and and	a construct industry leases input of the space promotion
Name		Opposition of the log characteristics	(look been in	nenn bele de plas - Managanh Tananan Fagilie en
YHL022C SPO11 P2	3179	Approximate manip strategy unputs to the	New Internet	and in a local sector of the s
GeneOntology Keywo	ds	print parties, particularly, m.		Single chard according to the shared local scale of the single of restriction of the state of the state state.
Biological process:	meiotic DNA double-strand break formation; double-strand break repair via single-strand annealing; double-strand break repair via homologous recombination; double-strand break repair; sporulation (sensu Saccharomyces); meiosis; meiotic recombination; germination (sensu Saccharomyces); spore germination (sensu Fung)	Part 1965 (Inner Der Hat 194, 2) P. Textus antique	College composition of the compo	
Molecular function:	endonuclease activity		Science of a	detair companies in the provide source in well-listed to standard over Targeting of fairs (11 is a specific source intergencements well-listed to standard over
Cellular component:	nucleus; meiotic chromosome		Amore and Amore	elle fante, de los plans, filial fante Asses Entren Lan Berlinder (anticipational de Referencia anticipational de altabate de la Secondor el 148 bases de anticipations
Curated Information			R. Sameta B.	constraint inspects the Booth spectrum inspects of providents. I Mills Bengio et al an Andrew Constraints and Annual Annua Annual Annual Annua Annual Annual Annua Annual Annual Annu
Spo11-induced double-strand breaks are required for normal regulation of 5' to 3' resection. Dr. Alastair Goldman 2003-05-01 [more]				Bernardson et al. Sector: Aller and the product of the sector. The sector differs result is an adverse of the sector and the sector adverse of the sect
SPO11 encodes a cor Adam Deutschbauer ;	served endonuclease required for spore germination [more] Ronald Davis 2003-03-21			particul fields pdf particul fig
recombination Alain Gerard Nicolas	2003-03-20 C. elegans			
				seed:: TRAVELI - CERENCE (per LT PERCEPT) - LEADNERVL U LEADNERVL U LEADNERVL UP - LEADNERVL UP - LEADNERVL UP - LEADNERVL mentale Medicale Leadnerve up - Leadnerve UP - Leadnerve Medicale Leadnerve UP - Leadnerve mentale UP - Leadnerve UP - Leadnerve mentale UP - Leadnerve UP - Leadnerve mentale UP - Leadnerve
Locus Information				
Locus Information Name				eller fersten als Martine andere elle Auflik besonse Carene, 14 (11) Martine Carene (14) Na Carene elle Auflik besonse de Carene (14) (11) Martine (14) (14) (14) Na Carene elle Auflik besonse de Carene (14) (14) (14) (14) (14) (14)
Locus Information Name 105E11.4 CE06363 s	po-11 Q22236 CAA92974.1			The second secon
Locus Information Name T05E11.4 CE06363 s GeneOntology Keywor	po-11 Q22236 CAA92974.1 IS			And the set of the set
Cocus Information Name 105E11.4 CE06363 s GeneOntology Keywor Biological process: Molecular function: Cellular component:	po-11 Q22236 CAA92974.1 ds meiotic DNA double-strand break formation; meiotic recombination; meiosis endonuclease activity meiotic chromosome; nucleus			The second secon
COUS Information Name 105E11.4 CE06363 s GeneOntology Keywor Biological process: Molecular function: Cellular component: Curated Information	ipo-11 Q22236 CAA92974.1 ds meiotic DNA double-strand break formation; meiotic recombination; meiosis endonuclease activity meiotic chromosome; nucleus			A set of
Cours Information Name 105E11.4 CE06363 s Biological process: Molecular function: Cellular component: Curated Information spo.11 is required for conserved mechaniss Anne Villeneuve 2003	po-11 Q22236 CAA92974.1 ds meiotic DNA double-strand break formation; meiotic recombination; meiosis endonuclease activity meiotic chromosome; nucleus initiation of recombination by double strand breaks via a n [more] -04-04			Bernstein Bernstein eine Bernstein eine Bernstein Be
Cours Information Name 105E11.4 CE06363 s GeneOntology Keywor Biological process: Molecular function: Cellular component: Curated Information spo.11 is required for conserved mechanisr Anne Villeneuve 2003	ppo-11 Q22236 CAA92974.1 ds meiotic DNA double-strand break formation; meiotic recombination; meiosis endonuclease activity meiotic chromosome; nucleus initiation of recombination by double strand breaks via a n .04-04			A statistical s
Cours Information Name 105E11.4 CE06363 s GeneOntology Keywor Biological process: Molecular function: Cellular component. Curated Information spo-11 is required for conserved mechanism Anne Villeneuve 2003	po-11 Q22236 CAA92974.1 ds meiotic DNA double-strand break formation; meiotic recombination; meiosis endonuclease actMty meiotic chromosome; nucleus initiation of recombination by double strand breaks via a n [more] -04-04 MAT a/alpha		2501	YHL622c/SP011_at.

0 1 2 3 4 5 6 7 8 9 10 11 12 time

0 1 2 3 4 5 6 7 8 9 10 11 12 time

2 000

100 Hote

1 ex

200 Internal

1303 MAT a/a

С

Reference database

관 80

te eoo

E 401

Portial

Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. Pubmed | website | search | more

- VHL022C/SP011_at

The Wellcome Trust Sanger Institute S. serevisiae GeneDB.

0 1 2 3 4 5 6 7 8 9 10 11 12 time

0 1 2 3 4 5 6 7 8 9 10 11 12 5 1 2 3 4 5 6 7 8 9 10 11 12

W303 MAT a/alpha

website | search | more

MIPS: a database for genomes and protein sequences. Pubmed | search | more

Genome annotati

Reinvestigation of the Saccharomyces cerevisiae genome annotation by comparison to the genome of a related fungus: Ashtya gossypii. Pubmed / website

A Re-annotation of the Saccharomyces cerevisiae Genome. Pubmed | website contributed information. Moreover, it is possible to choose one or more species and call up genes associated with particular GO keywords from the three categories, Biological Process, Molecular Function and Cellular Component. It should be emphasized that not all existing GO keywords can be selected to search the database but only those used during the curation process (this avoids a 'no entry' report); the available keywords are listed in alphabetical order in a pop-up menu (Fig. 1A). An advanced search option enables users to retrieve groups of genes according to their expression pattern or phenotype as determined in various high-throughput studies. Currently such query forms are available for S.cerevisiae, S.pombe, C.elegans and R.norvegicus (Fig. 1B). It is also possible to enter the section of a given species via the link in the navigation bar and directly access all information provided for the given organism.

The locus report

The locus report is separated into four sections, which contain links to internal and external data sources. A colour code marks out different species to facilitate navigation and to visually distinguish different sections of the database. The first section, Locus Information, displays the names, aliases and GO keywords chosen by one or more authors. Curated Information displays the title of a contribution followed by a link that leads to the complete description of a gene provided by the authors. Their names, contact information and their original references are shown in a separate window called the gene info box (Fig. 2A). If several contributions are available they are displayed in the locus page in chronological order based upon their date of publication in GermOnline with the most recent appearing at the top of the list. Furthermore, this section contains deep links to species-specific annotation databases (6), high-throughput loss-of-function studies (9–13) and, importantly, a direct connection to the locus report page of a potential homologue in another species represented in GermOnline. Note that deep links are convenient because they do not lead to the welcome or search page but directly to the locus report or data page of an external database.

Expression Information covers microarray expression data and is divided into links to relevant studies, consolidated expression databases and the US, European and Asian public array data repositories. The latter are displayed in the locus pages of all species. When following the links users can call up a graphical display of expression data externally to GermOnline, for e.g. *S.cerevisiae* undergoing the mitotic cell cycle and sporulation, sporulation in *S. pombe*, germline development in *C.elegans* and spermatogenesis in *R.norvegicus* [some examples of which are shown in Fig. 2B (14–19)]. External data sources are those that are either maintained by other laboratories or that contain expression data for poorly characterized loci that are not yet annotated and therefore not yet represented in GermOnline, such as numerous ESTs present on the U34 rat GeneChip (20).

Protein/Proteome Information is split into Protein/Proteome and Interactome sections, which include references with deep links to Swiss-Prot and data from biochemical and genetic high-throughput studies as well as deep links to GRID, a comprehensive database on protein–protein interaction data (21).

The *Toolbox* provides links to various web servers including some that enable researchers to predict the threedimensional structure of proteins that are similar to those for which the crystal structure is known (22), make multiple sequence alignments (23) or cluster their microarray expression data (24).

All sources of information accessible through the locus report are referenced by the title of the most recent relevant publication as well as links to PubMed (or a journal's web page), and, whenever possible, websites, search forms and deep links to the respective locus report pages (Fig. 2C). When information is available for only a subset of the genes in a genome, links are displayed only in those cases. For example, the annotation studies in *S.cerevisiae* mark out several hundred ORFs as spurious, hypothetical or very hypothetical (25,26).

Cross-references to other databases

GermOnline is highly cross-referenced and provides constantly updated deep links to more than 50 published external data sources including all relevant annotation and microarray expression databases. Conversely, SGD (8), S.cerevisiae and S.pombe GeneDB (Database Cross-Reference) and Swiss-Prot (27) connect to GermOnline from within their locus report pages ('Literature and Functional Analysis' in the case of SGD and 'Cross-References' for Swiss-Prot and GeneDB, respectively). Furthermore, reciprocal deep links are provided that lead to the Swiss-Model Repository, which contains automatically predicted putative protein structures that are similar to those for which the crystal structure is known (22) (28) and Mammalian Reproductive Genetics (MRG), which provides annotation and microarray expression data for mouse genes involved in spermatogenesis (B. Braun and V. Cassen; unpublished).

THE SUBMISSION AND CURATION PROCESS

Registration and curation

Scientists who wish to contribute information are asked to register. This simple and brief process ensures efficient communication between authors, curators and the GermOnline staff. Contact information is displayed in the personal gene info box (see Fig. 2A). While an author is generally considered to be the Principal Investigator (PI), other members of the laboratory who provide the email address of the registered PI are also encouraged to make contributions. An automatic email notifies authors (and PIs) of the state of curation once information is submitted. The flow of events is depicted in Figure 3A. An incoming submission is allocated to a curator who accesses the form and either publishes the content or asks for revisions. The curation process may be aborted and the contribution rejected should a submission be without merit or a hoax. When revisions are requested, authors are notified by email and in a text box visible below the submission form. This process may be repeated until the final revised version is published. Authors can make only one contribution to a given gene at a time. Multiple entries from one scientist on one gene are not possible. However, database entries on several different genes from one author are highly encouraged.

A	Author — Curator	Reject Accept Revise
В	Semicritice function: Logost Say profile Gene submission to Change login S. cerevisine Change password S. pombe Change submissions N. crease My gene submissions A. theilone To revise Z. revise	Orf name: YNL012W Trie: Inelg Description: Inelg
	Published submissions D. reno Liab game submissions D. reno Published submissions M. maculus Published submissions M. maculus Submissions to bit counted H. septens To curate : 0 In revision : 4 View prenting submissions Assign to curator. In curation In revision Revision Revisions	Upload Image: Secondative Intre:[help] Please select biological Process Event Biolocular Process Event Molocular
	Deletion requests Published submissions Ome submissions: 4 Gene submissions: 38	Function Image: Component Component Image: Component Meteoreent (hild) Image: Component Pubmed ID: Image: Component Trile: Image: Component Authors: Image: Component Journal: Image: Component
D	Sciencesbers help field (20 terms 10 - Netrosage Image: Sciencesbers help field 00 terms 10 Soatching for forms in Process Type a servedent of syour choice to search for corresponding CeneOmbility terms (e.g. sciencesber, search search for Corresponding CeneOmbility terms (e.g. sciencesber, search search for Spearch for Spe	Year: Volume: Pages: Full text URL: Vveside FTP site Pubmed ID: Tale: Authors: Journal: Year:
	Add and remove terms in Process Found I term certaining "In Process enbiling. Click on highlighted terms to add them to the submission form	Full tool URL: Full tool URL: Website: FTP site Bitercence Search Pubmed select Pubmed ID manual entry Pubmed ID: Table: Authors: Journal
	Add Remove	Volume Pages Fut ted URL. Weisste: FTP ate E-mail Principal Investigator. E-mail author. Reset Send

Figure 3. The submission and curation process. (A) Summary of the flow of events after an author has made a contribution. (B) A screen shot of the personal home page, which contains many functionalities (not all of which are available to all users) including a pop-up menu to choose the species. Once an organism and the gene/ORF of interest are selected, authors can call up the submission form shown in (C). To select GO keywords users follow the appropriate links in the submission form and then type a query in the search form displayed in (D). A list of hits is provided and the author can select and forward the keyword into the submission form by pressing the 'enter' button.

Making the initial contribution

Once an author has registered, immediate access to the submission form is given via the user's individualized home page after login (Fig. 3B). After a species and a gene/ORF are selected from the list of available organisms and loci, the submission form is displayed (Fig. 3C). It is recommended that the official locus ID or gene name/symbol as defined by the respective species databases be used as a particular alias may not always be available.

The submission form provides text fields for title and description. The title should be a concise summary of an author's findings on a gene function. The description briefly outlines the experimental approach and then summarizes facts concerning the biological process, the molecular function and the localization of the gene product in question. Authors are encouraged to include a brief comment on the next step(s) of their research. Important work from other groups is mentioned and cited in the text. To support the description, images and/or figures including titles and legends can also be uploaded. GO keywords can be selected by following the 'Select Process' link and typing a search term (e.g. cell cycle, meiosis, spermatogenesis, oogenesis) and options can be chosen to refine the search as indicated ('Contains' or 'Exact Match', Fig. 3D). Authors are able to select as many keywords as they deem appropriate. If a keyword is missing, it is recommended that it is submitted directly to GO where the updated keyword lists are retrieved regularly by GermOnline (7). Note that the names of keywords can change, they can become obsolete or even get deleted as the project evolves; any changes implemented by GO are valid for GermOnline as well. Finally, it is possible to provide up to three of the most recent PubMed references, which are automatically retrieved after entering the PubMed ID; publications that are not in PubMed can be entered manually (29).

Updating an existing database entry

Any update based upon the most recent publication replaces the previous database entry, thereby ensuring that the content does not become a historical overview of published work (like PubMed), or become outdated, but remains a concise, precise summary of the state-of-the-art views held by the author. Authors follow the link 'Published Submissions' in their home page and select one of their previous gene entries to update to directly access the current information in the submission form (Fig. 3B). After appropriate changes are made the submission is curated again. Updated information is not visible in the database during the curation process.

THE NEXT STEP

GermOnline is an ongoing project and therefore future directions and the development of functionalities will largely depend on community feedback. Improved services and options currently planned for releases higher than 2.0 include integration with the Ashbya Genome Database (AGD; S. Brachat, L. Hermida, R. Basavaraj, T. Schwede, P. Philippsen and M. Primig, manuscript in preparation), Flybase, Wormbase, MGD and Rat Genome Database (RGD). In addition, more data and links to potential homologues and experimentally verified orthologues will be integrated and tools for similarity searches and array data mining will be provided. More elaborate search options will include Boolean connection of the keywords from the GO categories. Finally, an online submission form for microarray data is envisaged.

CONCLUSION

GermOnline is a novel approach to biological information management and gene annotation. The idea is to meet the challenge of large-scale genome annotation by dividing the task into biological subjects and relevant species rather than having to cope with thousands of genes involved in different functions in all known species (30). This may help the global community of scientists who work on specific problems, often using only one or a few model organisms, to define groups of genes they can annotate in depth, collaborating with database developers and professional curators. Such expert information is all the more useful when placed within the context of relevant high-throughput data that provide insight into gene conservation and gene expression as well as protein localization and protein-protein interaction. This community-based approach suggested for data management in the field of germ cell growth and differentiation is applicable to all conserved biological processes studied in model organisms of various complexity and H.sapiens.

ACKNOWLEDGEMENTS

We thank A. Mizeracki for critical reading of the manuscript and B. Braun, B. Masdoua and V. Cassen for stimulating discussions and contributions to database programming. We acknowledge initial support of the project by N. Lamb and support of our IT infrastructure by Roger Jenni, F. Roesel and M. Jacquot. Funding for GermOnline Workshops was provided by the National Science Foundation (US), INSERM and CNRS. This project is funded by a grant from the Swiss Institute of Bioinformatics.

REFERENCES

- Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, 405, 827–836.
- Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
- Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.
- Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, 418, 387–391.
- Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
- 6. Baxevanis, A.D. (2003) The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res.*, **31**, 1–12.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, 25, 25–29.
- Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S.S., Engel, S., Fisk, D.G., Hong, E. et al. (2003)

Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.

- Colaiacovo, M.P., Stanfield, G.M., Reddy, K.C., Reinke, V., Kim, S.K. and Villeneuve, A.M. (2002) A targeted RNAi screen for genes involved in chromosome morphogenesis and nuclear organization in the *Caenorhabditis elegans* germline. *Genetics*, **162**, 113–128.
- Briza, P., Bogengruber, E., Thur, A., Rutzler, M., Munsterkotter, M., Dawes, I.W. and Breitenbach, M. (2002) Systematic analysis of sporulation phenotypes in 624 non-lethal homozygous deletion strains of *Saccharomyces cerevisiae*. *Yeast*, **19**, 403–422.
- 11. Enyenihi, A.H. and Saunders, W.S. (2003) Large-scale functional genomic analysis of sporulation and meiosis in *Saccharomyces cerevisiae*. *Genetics*, **163**, 47–54.
- Deutschbauer,A.M., Williams,R.M., Chu,A.M. and Davis,R.W. (2002) Parallel phenotypic analysis of sporulation and postgermination growth in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **99**, 15530–15535.
- Rabitsch,K.P., Toth,A., Galova,M., Schleiffer,A., Schaffner,G., Aigner,E., Rupp,C., Penkner,A.M., Moreno-Borchart,A.C., Primig,M. *et al.* (2001) A screen for genes required for meiosis and spore formation based on whole-genome expression. *Curr. Biol.*, **11**, 1001–1009.
- Mata, J., Lyne, R., Burns, G. and Bahler, J. (2002) The transcriptional program of meiosis and sporulation in fission yeast. *Nature Genet.*, 32, 143–147.
- Chu,S., DeRisi,J., Eisen,M., Mulholland,J., Botstein,D., Brown,P.O. and Herskowitz,I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, 282, 699–705.
- 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. *Nature Genet.*, 26, 415–423.
- 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.
- Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D.,

Lockhart, D.J. et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell, 2, 65–73.

- Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J., Davis, E.B., Scherer, S., Ward, S. et al. (2000) A global profile of germline gene expression in *C. elegans. Mol. Cell*, 6, 605–616.
- Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.
- Breitkreutz, B.J., Stark, C. and Tyers, M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biol.*, 4, R23.
- Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
- Poirot,O., O'Toole,E. and Notredame,C. (2003) Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res.*, 31, 3503–3506.
- Getz,G. and Domany,E. (2003) Coupled two-way clustering server. Bioinformatics, 19, 1153–1154.
- Wood, V., Rutherford, K., Ivens, A., Rajandream, M. and Barrell, B. (2001) A re-annotation of the Saccharomyces cerevisiae genome. Comp. Funct. Genet., 2, 143–154.
- Brachat,S., Dietrich,F.S., Voegeli,S., Zhang,Z., Stuart,L., Lerch,A., Gates,K., Gaffney,T. and Philippsen,P. (2003) Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii. Genome Biol.*, 4, R45.
- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Kopp,J. and Schwede,T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, 32, D230–D234.
- 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. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
- Stein,L. (2001) Genome annotation: from sequence to biology. Nature Rev. Genet., 2, 493–503.