Genélevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes

David J. Sherman¹,², Tiphaine Martin¹, Macha Nikolski¹, Cyril Cayla¹, Jean-Luc Souciet³ and Pascal Durrens¹ for the Genélevures Consortium

¹LaBRI, Laboratoire Bordelais de Recherche en Informatique, UMR CNRS 5800, ²INRIA, Institut National de Recherche en Informatique, Centre de Recherche Bordeaux Sud-Ouest, 351 Cours de la Libération, 33405 Talence cedex and ³Université Louis Pasteur, CNRS, UMR 7156, GDR 2354, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg, France

Received September 15, 2008; Revised October 15, 2008; Accepted October 16, 2008

ABSTRACT

The Genélevures online database (http://cbi.labri.fr/Genolevures/ and http://genolevures.org/) provides exploratory tools and curated data sets relative to nine complete and seven partial genome sequences determined and manually annotated by the Genélevures Consortium, to facilitate comparative genomic studies of Hemiascomycete yeasts. The 2008 update to the Genélevures database provides four new genomes in complete (subtelomere to subtelomere) chromosome sequences, 50 000 protein-coding and tRNA genes, and in silico analyses for each gene element. A key element is a novel classification of conserved multi-species protein families and their use in detecting synteny, gene fusions and other aspects of genome remodeling in evolution. Our purpose is to release high-quality curated data from complete genomes, with a focus on the relations between genes, genomes and proteins.

INTRODUCTION

Since 1999, the Genélevures Consortium explores eukaryote genome evolution through the large-scale comparison of manually annotated yeast genomes. The Genélevures on-line database undergoes a major update with every significant data release: in 2004 with 13 partial genomes (1,2), in 2006 with four complete genomes (3,4) and in 2008 with four complete genomes (this work). Our purpose is to release high-quality data sets produced by the 13 partners in the Consortium, rather than to provide a site for integration of data available elsewhere, although we do integrate external data for comparison.

Yeasts are small eukaryotes covering an evolutionary range comparable to the entire Chordate phylum (5), and the unique combination of genetic and genomic tools available for yeasts make them ideal candidates for experimental study of metabolism, genetic engineering and molecular genetics. All of the many yeasts sequenced so far have small genomes (10–20 Mb) which allows for detailed comparative genomics for a relatively modest price. The economic impact of yeasts is widespread: different species are used for the production of beer, wine and bread and more recently of various metabolic products, such as vitamins, ethanol, citric acid, lipids, etc. Yeasts can degrade hydrocarbons (genera Candida, Yarrowia and Debaryomyces), metabolize xylose (Pichia stipitis), depolymerise tannin extracts (Zygosaccharomyces rouxii), and produce hormones and vaccines in industrial quantities through heterologous gene expression (6). Several human diseases are due to yeast species among them the Hemiascomycetes Candida albicans, Candida glabrata, Candida tropicalis and even Saccharomyces cerevisiae in immunocompromised patients (7). The biology of S. cerevisiae has been extensively studied for decades as a model organism for molecular genetics and cell biology studies, and as a cell factory. Its genome (8) is the most thoroughly annotated among eukaryotes, and is a common reference for the annotation of other species. Genélevures focuses on the Hemiascomycete yeasts, a homogeneous phylogenetic group which nonetheless covers a broad range of physiological and ecological lifestyles. Hemiascomycete yeast genes contain introns and alternative splicing is observed (9). Comparative genomics studies in this group have proven informative (10–16); see (5) for review.

The Genélevures Consortium sequences, annotates and analyzes complete genomes from various branches of the Hemiascomycete class, and subjects them to large-scale...
in silico and experimental comparisons (Table 1). From these comparisons, we produce classifications of genes, proteins and sequences to address questions of molecular evolution, such as gene conservation, specific genes, function conservation and genome remodeling. We do not provide detailed annotations of individual genes and proteins of *S. cerevisiae*, which are already carefully maintained by MIPS and CYGD (http://mips.gsf.de/projects/fungi) (17) and SGD (http://www.yeastgenome.org/) (18) as well as in general purpose databases such as UniProt (19) and EMBL (20).

NEW GENOMES IN 2008

The Génolevures Consortium has released four new genomes in 2008, sequenced and assembled at high coverage and completely annotated by a team of 20 curators (Génolevures Consortium, submitted for publication). These genomes are those of *Zygosaccharomyces rouxii*, *Saccharomyces kluveri*, *Klyveromyces thermotolerans*, previously only sequenced to low or partial coverage, and *Debaryomyces hanseni*, partly resequenced to improve coverage. The former three, plus *K. lactis* (3) and *Ashbya gossypii* (21), are members of the *Saccharomyces* clades that are not descended from the ancestor that is thought to have undergone a whole genome duplication (unlike *C. glabrata* and *Saccharomyces cerevisiae*). The availability in the Génolevures of complete genomes for these five unduplicated genomes will allow close study of the core repertoire of protein families and functions and of the dynamics of genome remodeling.

PROTEIN FAMILIES

Extensive map reshuffling and a wide range of GC compositions from one species to another present a challenge for genome comparison in the yeasts. An essential tool for addressing this challenge is provided by ‘protein families’, a classification of protein-coding gene sequences into phylogenetic groups; members of a family are homologous and in many cases this homology is suggestive of functional similarity. The 48 889 proteins in the predicted proteomes of the nine complete genomes are classified into 7927 families. From these families, other sub-classifications are made, such as the identification of syntenic orthologs in the hemiascomycete yeasts (29,3,30), within phylogenetically delimited groups pairwise coverage is high and can be informative for studying genome dynamics. Pairs of homologous chromosomal regions between two species can be identified by comparing gene content and order, using protein families as an indication of gene-level homology within the regions. We identified these ‘pairwise syntenic blocks’ using the i-ADHoRE method (31) on gene homology relations identified using protein families. The approximately 19 000 pairwise synteny blocks obtained in the fashion can be interactively examined using the genome browser, as described in the Datasets area of the web site, and can be downloaded.

SYNTENY

Although synteny is in general poorly conserved in the Hemiascomycete yeasts (29,3,30), within phylogenetically delimited groups pairwise coverage is high and can be informative for studying genome dynamics. Pairs of homologous chromosomal regions between two species can be identified by comparing gene content and order, using protein families as an indication of gene-level homology within the regions. We identified these ‘pairwise syntenic blocks’ using the i-ADHoRE method (31) on gene homology relations identified using protein families. The approximately 19 000 pairwise synteny blocks obtained in the fashion can be interactively examined using the genome browser, as described in the Datasets area of the web site, and can be downloaded.

FUSIONS

Chromosomal rearrangements and segmental duplications may result in the creation or destruction of genes, when the breakpoint falls within the boundary of existing genes. ‘Gene fission’ occurs when a gene is broken in this way, and the 5' (or even the 3') fragment continues to be transcribed—creating a new, albeit truncated, gene. ‘Gene fusion’ occurs when the chromosomal rearrangement or duplication event leads to a fortuitous combination of previously existing genes, creating a new, longer gene. Fusions and fissions are mechanisms by which new genes can be formed in radical steps that do not obey the tree-like phylogenetic relations between species, and which can lead to significant changes in protein function. These events have to date been rarely taken into account in genome comparisons. Using a new method for detection of genes involved in gene fusion and fission events that makes its computations at the level of paralogous groups (32), we computed a catalogue of such events for complete genomes from the Hemiascomycete yeasts as well as other fungi. Both the paralogous groups of proteins used in the computation, and complete lists of identified fusion/fission events, are
now available in Généolevures and are downloadable in the Datasets area.

OTHER DATA SETS

Tandem gene repeats are another means of gene formation. Unlike gene copies resulting from segmental duplications or retrotransposons, truncated or chimeric genes are not observed at the boundaries of tandem gene arrays. Data on tandem gene repeats are available in Datasets area.

The Génosplicing database of spliceosomal introns and intron motifs in Hemiascomycete yeasts (C. Neuveglise, unpublished data) is available from the Datasets area, and can be used for both the study of splicing patterns and for the development of methods for predicting gene architectures from genomic data.

The YETI classification of yeast membrane and transporter proteins (33) defined using evolutionary relationships traced using non-ambiguous functional and phylogenetic criteria derived from the TCDB (34) classification system is available in the Datasets area.

Other analyses, such as (35) considering coverage of KEGG pathways (36), are available.

EXPLORING GÉNOLEVURES DATA

The design of the Généolevures on-line database has been revamped to provide improved tools for gaining insight into the mechanisms of eukaryotic molecular evolution. The key questions in the Généolevures use cases are:

(1) What genes exist, as orthologs for my favorite gene or as members of a functional class (keyword search, alignment, homologs)?

(2) What is known about a given chromosomal element (chromosomal elements)?

(3) What relations exist in a protein family (protein families)?

(4) How are the individual genomes organized (maps and genome browser, synteny)?

(5) How are genes and proteins classified (data sets for fusions, tandem repeats, introns, pathways, YETI)?

Additionally, the web site provides a query system that simultaneously searches for and can return: genes that have or may have a translation product, RNA and other genes that may have a transcription product only, cis-active elements and cross-genome protein families.

AVAILABILITY OF DATA AND URL CONSTRUCTION RULES

All data from the Généolevures web site are freely available, and instructions for proper citation are included in each section. The Généolevures web site is developed using a ‘representational state transfer’ architecture (37) and URLs for individual identified resources built from the database can be constructed systematically. Identified resources include chromosomal elements, such as genes (prefix/elt/Abbrev/Element_identifier), protein families (prefix/fam/Family_identifier), DNA sequences (prefix/seq/Sequence_identifier) and biomolecular pathways (prefix/pathway/Pathway_identifier). Nomenclatures for genome abbreviations, systematic chromosomal element names, and protein family names are described in the Documentation area of the web site.

Généolevures uses a bespoke object model mapped to a relational database, and uses the SOFA (38) and GO (39)
ontologies extensively. The web interface is developed in Mason. Key software components are provided by the NCBI Blast program suite (22) and the Stein Lab’s Generic Genome Browser (40).

ONGOING DEVELOPMENTS

The Génolevures Consortium continues its effort to sequence and annotate other complete genomes of Hemiascomycete yeasts. These genomes as well as the incorporation of other genomes annotated by third parties will help to refine the classifications presented above.

SUPPLEMENTARY DATA

Supplementary Data are available Online at http://genolevures.org/.

FUNDING

French National center for Scientific Research (CNRS) (GDR 2354, partial); French National Research Agency (ANR) (ANR-05-BLAN-0331; GENARISE, partial); Region Aquitaine (‘Pôle de Recherche en Informatique’ (2005-1306001AB, partial) and ACI IMPBIO (IMPB114, partial) (‘Génolevures En Ligne’). Funding for open access charter: CNRS GDR 2354.

Conflict of interest statement. Large-scale Smith-Waterman alignments used in constructing protein families were performed using GenCore 6 software licensed from Bioceleration Inc., 77 Milltown Road, Suite B4, East Brunswick, NJ 08816, USA.

REFERENCES

1. Souciet,J.L. and Génolevures Consortium. (2000) Special issue: Génolevures, FEBS Lett., 487, 1–149.
2. Sherman,D.J., Durrens,P., Beyne,E., Nikolski,M. and Souciet,J.L. (2004) Génolevures: comparative genomics and molecular evolution of hemiascomyeotic yeasts. Nucleic Acids Res., 32, D315–D318.
3. Dujon,B., Sherman,D.J., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., De Montigny,J., March,C., Neuvéglise,C., Tailla,E. (2004) Genome evolution in yeasts. Nature, 430, 35–44.
4. Sherman,D., Durrens,P., Iragne,F., Beyne,E., Nikolski,M. and Souciet,J.L. (2006) Génolevures complete genomes provide data and tools for comparative genomics of hemiascomyceteous yeasts. Nucleic Acids Res., 34, D432–D435.
5. Dujon,B. (2006) Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet., 22, 375–387.
6. Kurtzmann,C.P. and Fell,J.W. (eds) (1998) The Yeasts, A Taxonomic Study, 4th edn. Elsevier, Amsterdam.
7. Tawfik,O.W., Papasian,C.J., Dixon,A.Y. and Potter,L.M. (1989) Saccharomyces cerevisiae cervisiae in a patient with acquired immune deficiency syndrome. J. Clin. Microbiol., 27, 1689–1691.
8. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. et al. (1996) Life with 6000 genes. Science, 274, 546, 563–567.
9. Neuvéglise,C., Chalvet,F., Wincker,P., Gaillardin,C. and Casaregola,S. (2005) Mutator-like element in the yeast Yarrowia lipolytica displays multiple alternative spliccings. Eukaryot. Cell, 4, 615–624.
10. Clifton,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science, 301, 71–76.
11. Kelis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423, 241–254.
12. Richard,G.F., Kerrest,A., Lafontaine,I. and Dujon,B. (2005) Mol. Biol. Evol., 22, 1011–1023.
13. Fabre,E., Müller,H., Therizols,P., Lafontaine,I., Dujon,B. and Fairhead,C. (2005) Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination. Mol. Biol. Evol., 22, 856–873.
14. Knop,M. (2006) Evolution of the hemiascomycete yeasts: on life styles and the importance of inbreeding. Bioessays, 28, 696–708.
15. Rokas,A. and Carroll,S.B. (2005) More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol., 22, 1337–1344.
16. Blank,L.M., Lehmbrock,F. and Sauer,U. (2005) Metabolic-flux and network analysis in fourteen hemiascomyceteous yeasts. FEMS Yeast Res., 5, 545–558.
17. Mewes,H.W., Dietmann,S., Frishman,D., Gregory,R., Mannhaupt,G., Mayer,K., Muensterkötter,M., Ruepp,A., Spannagl,M., Stuempflen,V. et al. (2008) MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res., 36, D196–D201.
18. Nash,R., Weng,S., Hitz,B., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E. et al. (2007) Expanded protein information at SGD: new pages and proteome browser. Nucleic Acids Res., 35, D468–D471.
19. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res., 33, D154–D159.
20. Kanz,C., Aldebert,P., Althorpe,N., Barker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. et al. (2005) The EMBL Nucleotide Sequence Database. Nucleic Acids Res., 33, D29–D33.
21. Dietrich,F.S., Voegeli,S., Brachat,S., Lerch,A., Gates,K., Steiner,S., Mohr,C., Pohlmann,R., Luedi,P., Cho,S. et al. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science, 304, 304–307.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.
23. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol., 147, 195–197.
24. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P., Ledley,R.S. et al. (2004) PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res., 32, D201–D205.
25. Van Dongen,S. (2000) Graph clustering by flow simulation. PhD Thesis, University of Utrecht, The Netherlands.
26. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res., 30, 1575–1584.
27. Nikolski,M. and Sherman,D.J. (2007) Family relationships: should consensus reign?: consensus clustering for protein families. Bioinformatics, 23, e71–e76.
28. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhodov,S.L., Nikolskaya,A.N. et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 4, 41.
29. Llorente,B., Malpertuy,A., Neuvéglise,C., de Montigny,J., Aigle,M., Artiguenave,F., Blandin,G., Bolotin-Fukuhara,M., Bon,E., Brottier,P. et al. (2000) Genomic exploration of the hemiascomyceteous yeasts: 18. Comparative analysis of chromosome maps and synteny with Saccharomyces cerevisiae. FEBS Lett., 487, 101–112.
30. Fischer,G., Rocha,E.P., Brunet,F., Vergassola,M. and Dujon,B. (2006) Highly variable rates of genome rearrangements between hemiascomyceteous yeast lineages. PLoS Genet., 2, e32.
31. Shimillon,C., Vandepoele,K., Sacys,Y. and Van de Peer,Y. (2004) Building genomic profiles for uncovering segmental homology in the twilight zone. Genome Res., 14, 1095–1106.
APPENDIX

The Génotextes Consortium is coordinated by J.L. Souciet and is composed of laboratories from Institut de Biologie et de Technologies de Saclay (iBiTec-S), CEA, F-91191 Gif-sur-Yvette CEDEX, France (C. Marck); Laboratoire de Chimie Bactérienne, CNRS-UPR9043, Université de la Méditerranée, 31 Chemin Joseph Aiguier 13402 Marseille Cedex 20 (E. Talla); Laboratoire Microbiologie, Adaptation et Pathogénie, Université Claude Bernard - Lyon 1, CNRS/UMR 5240, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne, France. (M. Lemaire); Génoscope (CEA), 2 rue Gaston Crémieux, BP 191, F-91057 Evry Cedex, France (J. Weissend); Laboratoire Bordelais de Recherche en Informatique, LaBRI, UMR 5800 CNRS, INRIA Bordeaux Sud-Ouest (MAGNUM), 351 cours de la Libération, 33405 Talence Cedex, France (D. Sherman); Laboratoire de Microbiologie et Génétique Moleculaire, UMR 1238 INRA/2585 CNRS/Institut National Agronomique Paris-Grignon, AgroParisTech, F-78850 Thivarval-Grignon, France (C. Gaillardin); Institut Pasteur/URA2171 CNRS and Université Pierre et Marie Curie UFR927, 25 rue du Docteur Roux, F-75724 Paris Cedex 15, France (B. Dujon); Laboratoire Génétique Moléculaire, Génomique, Microbiologie, UMR 7156, Université Louis Pasteur, Institut de Botanique, 28 rue Goethe, F-67000 Strasbourg, France (S. Potier, J.-L. Souciet); Washington University School of Medicine, Department of Genetics, Campus Box 8510 4566 Scott Ave., St Louis, MO 63110 USA (M. Johnston); Washington University School of Medicine, Genome Sequencing Center, Campus Box 8501, 4444 Forest Park Avenue, St Louis, MO 63108 USA; (R. Fulton); Pasteur Genopole: Intégration et Analyse Génomique, 28 rue du Docteur Roux, F-75724 Paris Cedex 15, France (L. Frangeul); Unité de Génétique, Université Catholique de Louvain, Croix du Sud 2 bte 14, 1348 Louvain-la-Neuve, Belgique (P. Baret); Institut de Biologie Moléculaire et Cellulaire, Architecture et Réactivité de L’ARN, IBMC (UPR 9002 CNRS), 15 rue René Descartes, 67084 Strasbourg, France (E. Westhof).