Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies[1]

Jean-Luc Souciet[a,*], Michel Aigle[b], François Artiguenave[c], Gaëlle Blandin[d], Monique Bolotin-Fukuhara[e], Elisabeth Bon[f], Philippe Brottier[c], Serge Casaregola[f], Jacky de Montigny[a], Bernard Dujon[d], Pascal Durrens[b], Claude Gaillardin[f], Andrée Lépingle[f], Bertrand Llorente[d], Alain Malpertuy[d], Cécile Neuvéglise[f], Odile Ozier-Kalogéropoulos[d], Serge Potier[a], William Saurin[c], Fredj Tekaia[d], Claire Toffano-Nioche[g], Micheline Wésolowski-Louvel[h], Patrick Wincker[c], Jean Weissenbach[c]

[a] *Laboratoire de Génétique et Microbiologie, UPRES-A 7010 ULP/CNRS, Institut de Botanique, 28 rue Goethe, 67000 Strasbourg, France*

[b] *Laboratoire de Biologie Cellulaire de la Levure, IBGC, 1 rue Camille Saint-Säens, 33077 Bordeaux Cedex, France*

[c] *Génoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, P.O. Box 191, Evry Cedex, France*

[d] *Unité de Génétique Moléculaire des Levures, Institut Pasteur/URA 2171 CNRS and UFR 927 Université Pierre et Marie Curie, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*

[e] *Institut de Génétique Moléculaire, IGM-bat 400, Université d'Orsay, 91405 Orsay Cedex, France*

[f] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire, INRA UMR 216, CNRS URA 1925, INA-PG, P.O. Box 01, 78850 Thiverval-Grignon, France*

[g] *Laboratoire de Bioinformatique des Génomes, IGM-bat 400, Université d'Orsay, 91405 Orsay Cedex, France*

[h] *Microbiologie et Génétique ERS 2009, CNRS/UCB/INSA, bat. 405 R2, Université Lyon I, 69622 Villeurbanne Cedex, France*

Received 9 November 2000; accepted 11 November 2000

First published online 27 November 2000

Edited by Horst Feldmann

Abstract   The identification of molecular evolutionary mechanisms in eukaryotes is approached by a comparative genomics study of a homogeneous group of species classified as Hemiascomycetes. This group includes *Saccharomyces cerevisiae*, the first eukaryotic genome entirely sequenced, back in 1996. A random sequencing analysis has been performed on 13 different species sharing a small genome size and a low frequency of introns. Detailed information is provided in the 20 following papers. Additional tables available on websites describe the ca. 20 000 newly identified genes. This wealth of data, so far unique among eukaryotes, allowed us to examine the conservation of chromosome maps, to identify the 'yeast-specific' genes, and to review the distribution of gene families into functional classes. This project conducted by a network of seven French laboratories has been designated 'Génolevures'.   © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Random sequencing; Phylogenetics; Bioinformatics; Comparative genomics; *Saccharomyces cerevisiae*

1. Introduction

Fungi, and yeasts in particular, have a long standing role in the development of genetics and molecular biology. *Saccharomyces cerevisiae*, the first eukaryotic organism to be sequenced [1], contributed significantly to the recent and rapid emergence of genomics. Comparative genomics is a new field which provides insights into the understanding of molecular evolution. It helps to interpret the complex genomes of higher eukaryotes as recently exemplified by the impact of the work on *Tetraodon nigroviridis*, a compact genome, for the human genome analysis [2,3]. The rapidly growing number of fully sequenced prokaryotic genomes, allows comparison between related species such as Mycoplasma [4–6] or Chlamydia [7]. For eukaryotic organisms, only few genomes have been completely sequenced so far: *S. cerevisiae* [1], *Caenorhabditis elegans* [8], and, more recently, *Drosophila melanogaster* [9]. Others are close to completion: *Schizosaccharomyces pombe* (http://www.sanger.ac.uk/Projects/S_pombe/genomic_sequence.shtml), *Arabidopsis thaliana* [10,11] and *Homo sapiens* in its first draft version [12,13]. The comparison *of S. cerevisiae* with *C. elegans* confirms its phylogenic relationship with the animal kingdom, but the evolutionary distance between these two species remains large, thus limiting our understanding of molecular evolution among eukaryotes. Even though *S. pombe* is an ascomycete, preliminary sequence comparisons between *S. pombe* and *S. cerevisiae* provide similar results indicating a very long evolutionary distance between the two yeasts [14]. It appears to us that the comparative analysis of closely related organisms is needed to answer basic questions concerning molecular evolution, such as: what are the species-specific genes and how numerous are they for speciation? how are genes distributed among functional families? what is the rate of amino acid divergence among the proteins encoded by these genes? what are the general mechanisms involved in reshaping of chromosomal maps? With their relatively small genome size, yeasts offer a unique opportunity to explore eukaryotic genome evolution by the comparative analysis of numerous species.

*Corresponding author. Fax: (33)-3-88 35 84 84. E-mail: souciet@gem.u-strasbg.fr

[1] Sequences and annotations are accessible at: Génoscope (http://www.genoscope.cns.fr), FEBS Letters Website (http://www.elsevier.nl/febs/show/), Bordeaux (http://cbi.genopole-bordeaux.fr/Genolevures) and were deposited into the EMBL database (accession number from AL392203 to AL441602).

0014-5793/00/$20.00
PII: S0014-5793(00)02272-9

Ascomycetes have been the object of numerous taxonomic studies which have led to successive modifications of classification [15]. The recent use of rDNA sequence comparisons, has longly improved this classification [3,16]. The group of Ascomycetes now comprises three classes: the Archiascomycetes (including the orders Schizosaccharomycetales, Taphrinales, Protomycetales, and Pneumocystidales); the Euascomycetes (with the subdivisions orders Pyrenomycetes, Discomycetes, Loculoascomycetes, and Plectomycetes) and the Hemiascomycetes with only one order, the Saccharomycetales. This classification corresponds to the introduction of the new class of the Archiascomycetes in 1993 [17]. The Hemiascomycetes encompasses the budding yeasts and the yeast-like genera Ascoidea and Cephaloascus. Family assignment within the Saccharomycetales order remains uncertain, and the classification presented by Kurtzman [16,18], is expected to be modified when novel molecular data will become available. The work presented here, based on the partial sequencing of 13 species selected across the entire Hemiascomycete realm, may contribute to sharpen this classification at the same time as to provide new data on molecular evolution.

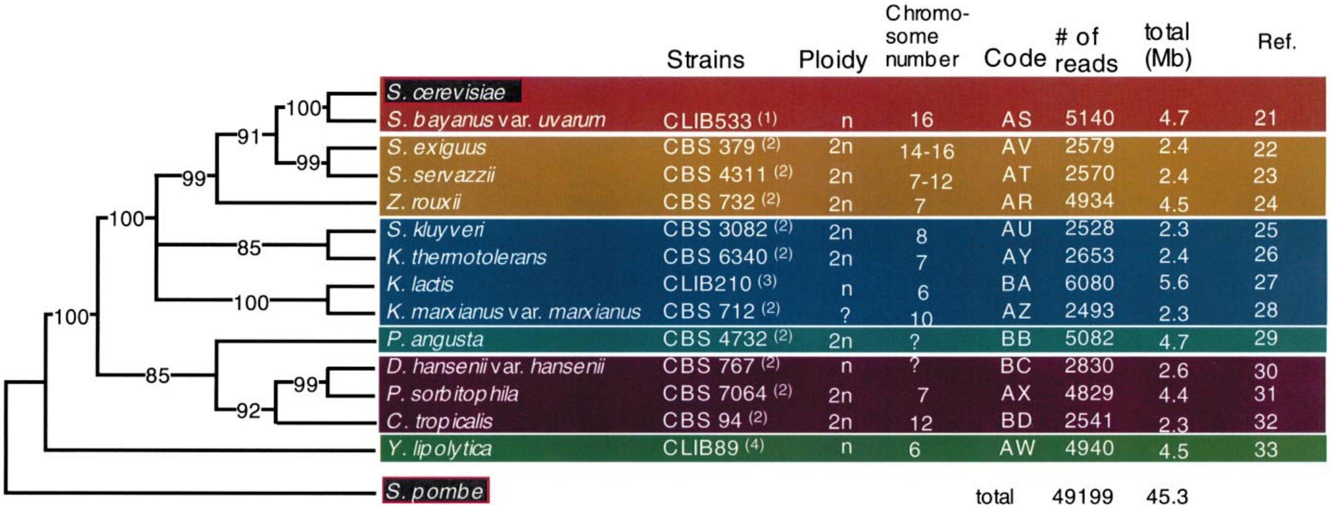
2. Materials and methods

For each species, a random genomic DNA library was prepared by generating fragments ranging in size from 4 to 5 kb. This size was selected as to offer the maximum probability of finding two neighbouring genes. Single-pass sequencing (up to 1.0 kb on average [19]) of both extremities of each insert was performed such that each individual insert is characterised by two random sequence tags (RSTs). The analysis of the sequences of each of the 13 selected species, using a manual validation step [20], is reported in the following papers [21–33]. Since the studied species show gene sizes comparable to *S. cerevisiae*, and similarly rare occurrence of introns, it was often possible to identify the presence of more than one ORF in each RST. In some favourable situations, thanks to the exceptional quality and length of the sequencing reads, five ORFs have been detected in the insert RST [21–33]. Previous sequences from *Kluyveromyces lactis* and *Candida albicans*, have shown an average gene density of 0.72 and a protein sequence identity with *S. cerevisiae* orthologues of 60%. Thus about 75% of these sequences data can be readily identified by comparison with *S. cerevisiae*. Assessing the proper number of RSTs which is required for the best characterisation of one yeast genome was a priori difficult. We have therefore conducted these experiments by analyzing some species with 5000 RSTs and other ones with only 2500 RSTs. Others genetic elements, not sharing identity to the *S. cerevisiae* genome, were identified by comparison with all predicted protein sequences from other completely sequenced organisms as well as with SwissProt [20]. The genome coverage of 0.2 for species studied by 2500 RSTs or 0.4 for four species studied by 5000 RSTs permitted the assembly of few RSTs in contigs where repetitive genetic elements, as for example Ty, rDNA, or plasmids, where further identified.

3. Results and discussion

3.1. Selection of a set of yeast species representative of the Hemiascomycete class

Limited comparative genomics information between related yeast species was provided by the studies of Ozier-Kalogéropoulos et al. [14] comparing *K. lactis* and *S. cerevisiae*, that of Hartung [34] on *C. albicans* and that of Prillinger et al. [35] on *Ashbya gossipyi*. In the present work, we decided to generate an important set of novel sequence data by focusing on a defined taxonomic group: the yeasts from the Hemiascomycete class. In setting up this list (Fig. 1) our primary concern was to select, as far as could be inferred from 18S rDNA phylogeny [16,36], a set of species representing the various branches of the Hemiascomycete class. A second criterion was within each branch to give preference to species of industrial or biomedical interest, which could be used for subsequent functional studies or have already been the object of extensive genetic studies. Following these criteria, we have retained the following species *Pichia angusta* (also called *Hansenula polymorpha*); *K. lactis*; *Yarrowia lipolytica* and *Zygosaccharomyces rouxii*; because they have been subjected to numerous genetic and molecular studies during the last 15 years and, as a direct consequence, are increasingly used for industrial purposes [37–43]. *Saccharomyces bayanus* (var. *uva-*



	Strains	Ploidy	Chromosome number	Code	# of reads	total (Mb)	Ref.
S. cerevisiae							
S. bayanus var. uvarum	CLIB533 (1)	n	16	AS	5140	4.7	21
S. exiguus	CBS 379 (2)	2n	14-16	AV	2579	2.4	22
S. servazzii	CBS 4311 (2)	2n	7-12	AT	2570	2.4	23
Z. rouxii	CBS 732 (2)	2n	7	AR	4934	4.5	24
S. kluyveri	CBS 3082 (2)	2n	8	AU	2528	2.3	25
K. thermotolerans	CBS 6340 (2)	2n	7	AY	2653	2.4	26
K. lactis	CLIB210 (3)	n	6	BA	6080	5.6	27
K. marxianus var. marxianus	CBS 712 (2)	?	10	AZ	2493	2.3	28
P. angusta	CBS 4732 (2)	2n	?	BB	5082	4.7	29
D. hansenii var. hansenii	CBS 767 (2)	n	?	BC	2830	2.6	30
P. sorbitophila	CBS 7064 (2)	2n	7	AX	4829	4.4	31
C. tropicalis	CBS 94 (2)	2n	12	BD	2541	2.3	32
Y. lipolytica	CLIB89 (4)	n	6	AW	4940	4.5	33
S. pombe				total	49199	45.3	

Fig. 1. Synopsis of the 13 hemiascomycetous yeast species analysed in the Génolevures programme. The left part of the figure corresponds to a cladogram that was constructed from the rDNA sequence encoding the 25S rRNAs by using the maximum parsimony (PAUP 3.1.1.), *S. pombe* was used as an outgroup. Values on branches represent bootstrap values on 500 replicates. The species name is followed by the reference of the strain analysed in this work, plus when available from the literature the level of ploidy and the chromosome number. For each species is indicated by two letters the code recovered at the beginning of each sequence, the number of different reads, the total length of DNA sequenced in this work and the reference of the corresponding article in this issue. (1) other strain named MCYC 623-6C ura3; (2) denotes type strain; (3) other strain named 2359/152; (4) other strain named W29, is a natural isolate.

Table 1A
Homologues to each protein-coding *S. cerevisiae* gene found in the 13 other hemiascomycetous species

ORF	Old	New	Sb	Se	Ss	Zr	Sk	Kt	Kl	Km	Pa	Dh	Ps	Ct	Yl	*T*
Chromosome 1																
YAL068c	3	2	·79	−	−	−	−	−	−	−	−	−	−	−	−	1
YAL067c	1	1	77	−	−	64	−	75	80	·41	50	·26	−	−	54	8
YAL066w	3	3	−	−	−	−	−	−	−	−	−	−	−	−	−	0
YAL065c	3	3	−	−	−	−	−	−	−	−	−	−	−	−	−	0
Chromosome 16																
YPR198w	1	1	−	·39	−	−	−	−	−	−	−	−	−	·37	−	2
YPR199c	3	3	−	−	−	−	−	−	−	−	−	−	−	−	−	0
YPR200c	3	3	−	−	−	−	−	−	−	−	−	−	−	−	−	0
YPR201w	1	1	−	−	−	−	−	−	67	−	−	−	−	−	−	1

The table indicates, for each of the 6263 ORFs of *S. cerevisiae* listed according to their map location (gene designation refers to [54]), the percentage of amino acid identity with the homologous gene in each of the 13 yeast species determined from blastx alignments (− no homologue identified in this species). In cases where several genes of a given yeast species are homologous to a given gene of *S. cerevisiae*, only the closest homologue is represented. · indicates that the ORF of *S. cerevisiae* is one of several possible homologues to the gene of the other yeast species, as result of the existence of gene families in *S. cerevisiae*. Old and New, class to which belong the ORF of *S. cerevisiae*, respectively, before and after this project: (1) ORF conserved in non-Ascomycetes, (2) ORF conserved in Ascomycetes only, (3) ORF without homology out of the *S. cerevisiae* genome, (4) questionable ORF (see [52] for details), *T*: total number of yeast species in which at least one homologue of the *S. cerevisiae* ORF has been identified. Sb, *S. bayanus* var. *uvarum*; Se, *Saccharomyces exiguus*; Ss, *Saccharomyces servazzii*; Zr, *Z. rouxii*; Sk, *Saccharomyces kluyveri*; Kt, *K. thermotolerans*; Kl, *K. lactis*; Km, *Kluyveromyces marxianus* var. *marxianus*; Pa, *P. angusta*; Dh, *D. hansenii* var. *hansenii*; Ps, *P. sorbitophila*; Ct, *C. tropicalis*; Yl, *Y. lipolytica.*

Table 1B
Homologues to each of the 52 *S. cerevisiae* tRNA genes species found in the 13 other hemiascomycetous species

	Sc	Sb	Se	Ss	Zr	Sk	Kt	Kl	Km	Pa	Dh	Ps	Ct	Yl	*T*
A (GCA) tRNA	5	−	−	−	−	+	−	−	−	−	−	−	−	−	1
A (GCT) tRNA	11	+	+	+	+	+	+	+	−	−	−	−	+	+	9
C (TGC) tRNA	4	−	+	+	+	+	+	−	−	+	−	+	−	−	7
D (GAC) tRNA	15	−	+	+	+	−	−	−	+	−	+	+	+	−	7
E (GAA) tRNA	14	+	+	+	+	+	+	−	+	−	−	+	+	+	10
E (GAG) tRNA	1	−	−	−	−	−	+	−	−	+	−	−	−	−	2
F (TTC) tRNA	10	+	+	+	+	+	+	−	+	+	+	−	+	+	11
G (GGA) tRNA	3	−	−	−	+	+	+	−	−	−	−	+	−	−	4
G (GGC) tRNA	16	+	−	+	+	+	+	−	−	+	−	+	+	+	9
G (GGG) tRNA	2	−	+	−	−	−	−	−	−	−	−	−	−	−	1
H (CAC) tRNA	7	−	+	−	+	+	−	−	−	−	+	+	−	−	5
I (ATA) tRNA	2	−	+	+	+	−	+	−	−	−	+	−	+	+	7
I (ATT) tRNA	13	−	+	−	+	+	+	−	+	+	+	−	+	+	9
K (AAA) tRNA	7	−	−	+	+	+	+	−	−	−	−	+	−	−	5
K (AAG) tRNA	14	−	−	+	+	+	+	−	+	−	−	−	−	−	5
L (CTA) tRNA	3	−	+	+	+	−	+	−	−	−	−	−	−	−	4
L (CTC) tRNA	1	−	−	−	−	−	−	−	−	−	−	+	−	−	1
L (TTA) tRNA	7	+	+	+	+	+	−	−	+	+	+	+	−	−	9
L (TTG) tRNA	10	+	+	+	+	+	+	−	−	+	−	−	−	−	7
M (ATG) tRNA	10	−	+	+	+	+	−	−	+	+	−	−	+	+	8
N (AAC) tRNA	10	−	+	+	+	−	+	−	−	+	+	+	−	−	7
P (CCA) tRNA	10	+	+	+	+	−	+	−	+	+	+	−	+	−	9
P (CCT) tRNA	2	−	−	−	+	−	−	−	−	−	−	−	−	+	2
Q (CAA) tRNA	9	−	−	+	+	+	+	−	+	+	+	+	+	+	10
Q (CAG) tRNA	1	−	−	−	−	−	−	−	−	+	−	−	−	−	1
R (AGA) tRNA	11	−	+	+	+	−	+	−	+	−	+	−	+	−	7
R (AGG) tRNA	1	−	−	−	−	−	−	−	−	−	−	−	−	−	0
R (CGG) tRNA	1	−	−	−	+	−	−	−	−	−	−	−	−	−	1
R (CGT) tRNA	6	+	+	−	+	−	+	−	+	−	−	−	−	−	5
S (AGC) tRNA	4	+	−	−	+	−	+	−	+	+	−	−	−	+	6
S (TCA) tRNA	3	+	+	+	−	−	+	−	−	+	−	−	+	−	6
S (TCG) tRNA	1	−	−	+	+	−	+	−	−	−	−	−	−	−	3
S (TCT) tRNA	11	+	+	+	+	+	+	−	+	+	+	+	−	+	11
T (ACA) tRNA	4	−	+	−	+	−	+	−	+	−	+	+	−	−	6
T (ACG) tRNA	1	−	−	+	+	+	−	−	−	−	−	−	−	−	3
T (ACT) tRNA	11	−	−	+	+	+	+	−	+	+	+	+	+	+	10
V (GTA) tRNA	2	−	−	−	−	−	+	−	−	−	−	−	−	−	1
V (GTG) tRNA	2	−	−	−	+	+	+	−	−	+	+	−	−	−	5
V (GTT) tRNA	14	−	−	+	+	+	+	−	+	+	+	+	+	+	10
W (TGG) tRNA	6	+	+	+	+	+	+	−	+	+	+	+	−	+	11
Y (TAC) tRNA	8	−	+	−	+	+	+	+	−	−	+	+	−	+	8
Total by species	274	12	22	24	33	22	29	2	17	19	17	17	14	15	

The first column indicates the total number of *S. cerevisiae* tRNA genes belonging to a given species. Abbreviations as in Table 1A.

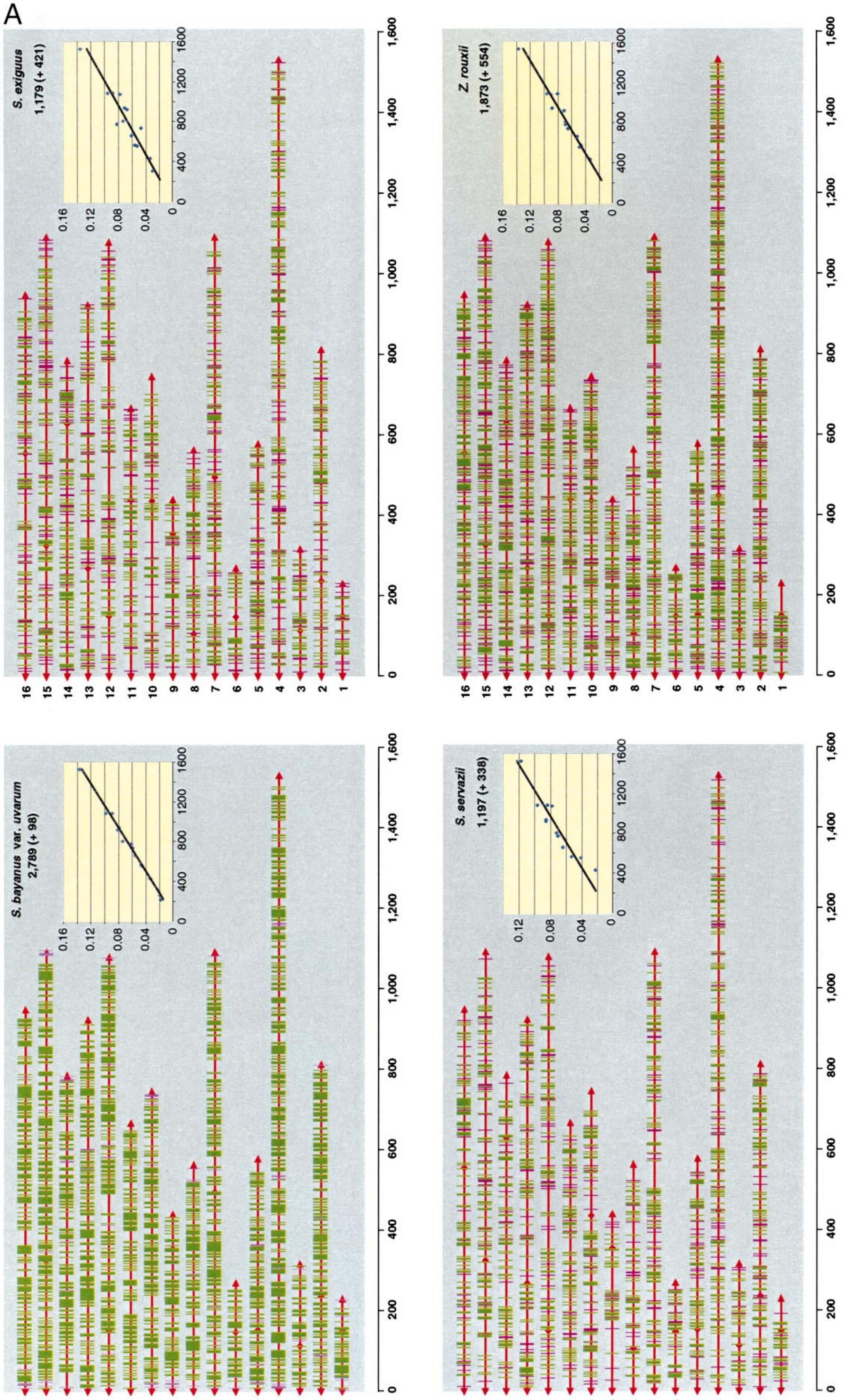

A
S. exiguus
1,179 (+ 421)
Z. rouxii
1,873 (+ 554)
S. bayanus var. uvarum
2,789 (+ 98)
S. servazii
1,197 (+ 338)
0.16
0.12
0.08
0.04
0
400
800
1200
1600
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
200
400
600
800
1,000
1,200
1,400
1,600

Fig. 2.

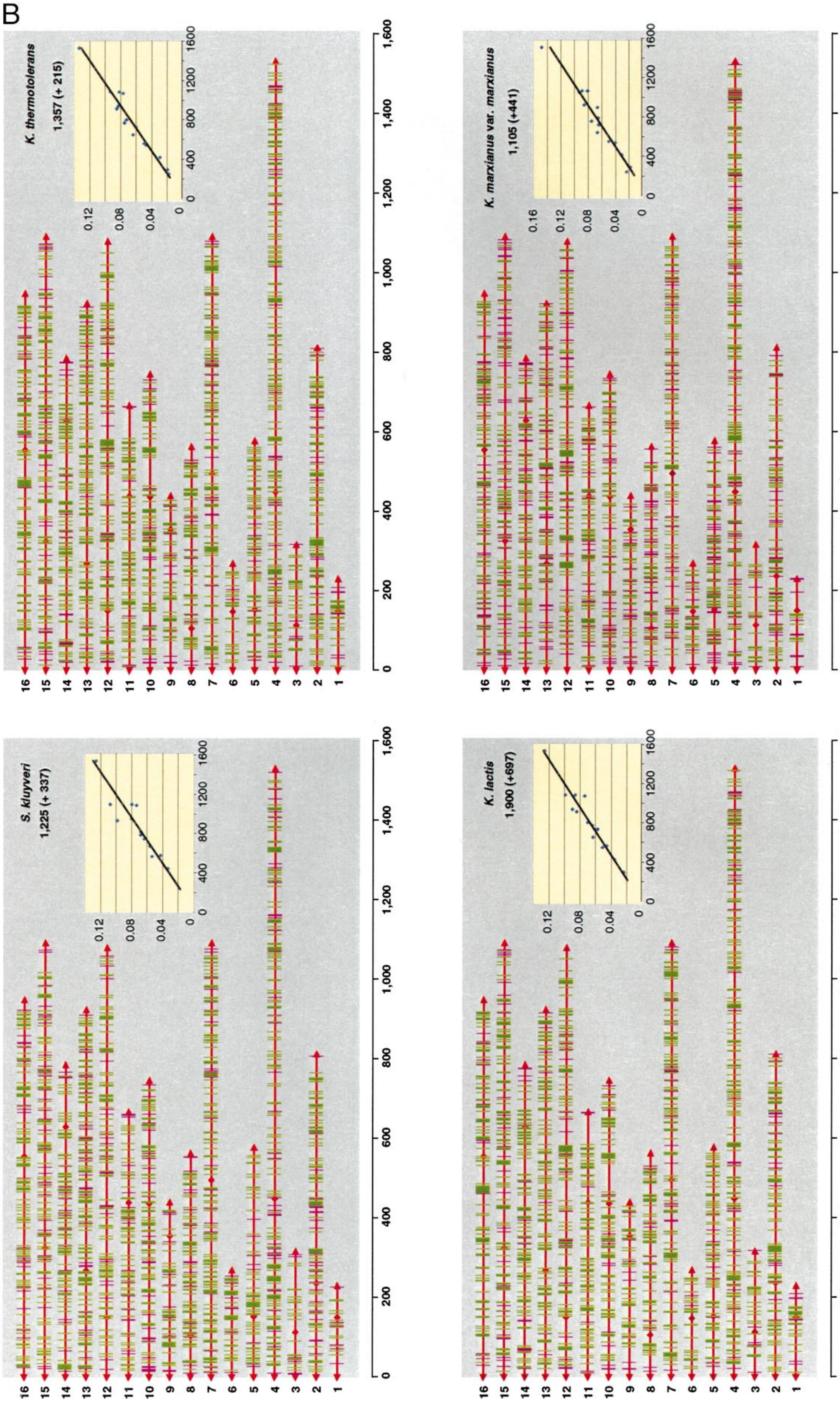


Fig. 2 (*continued*).

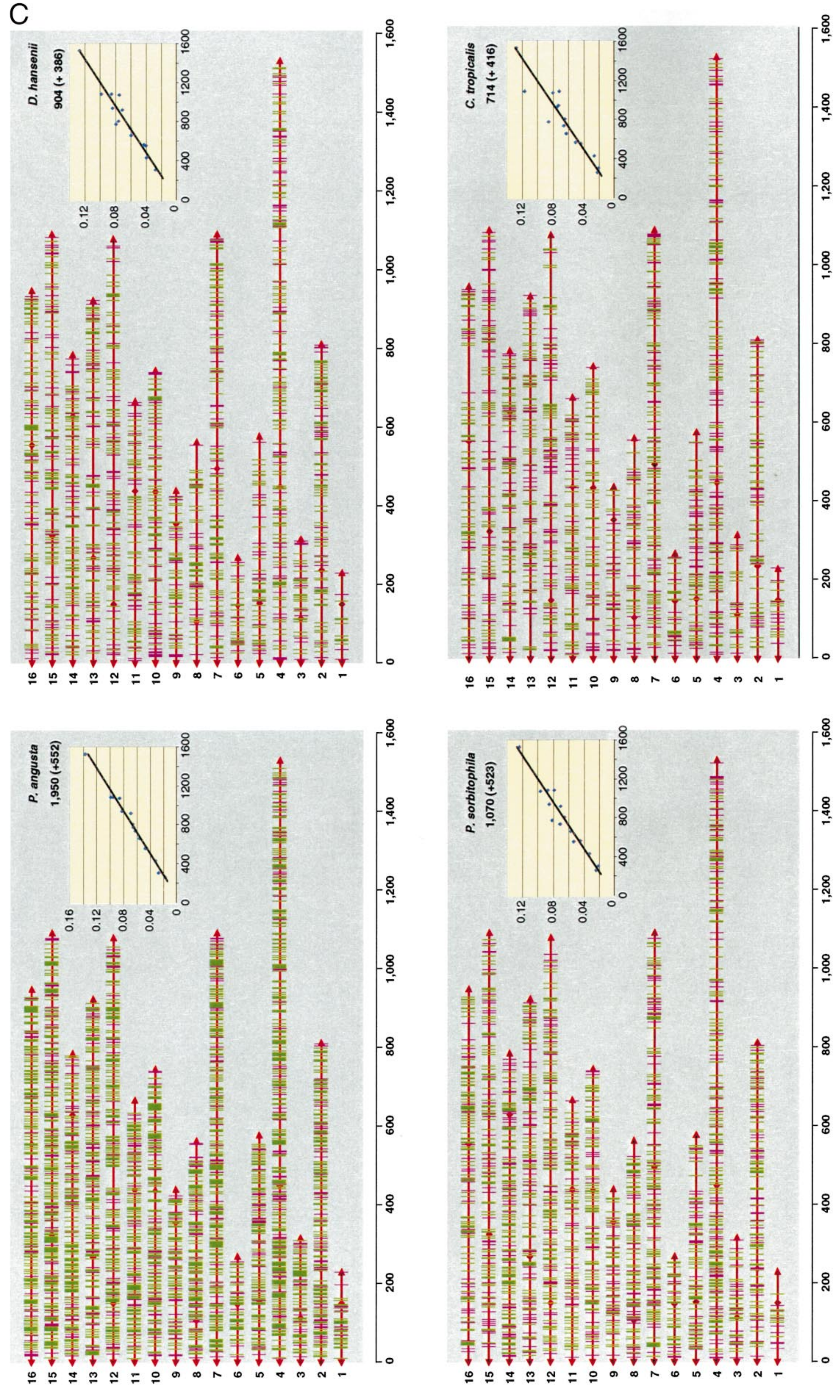


Fig. 2 (*continued*).

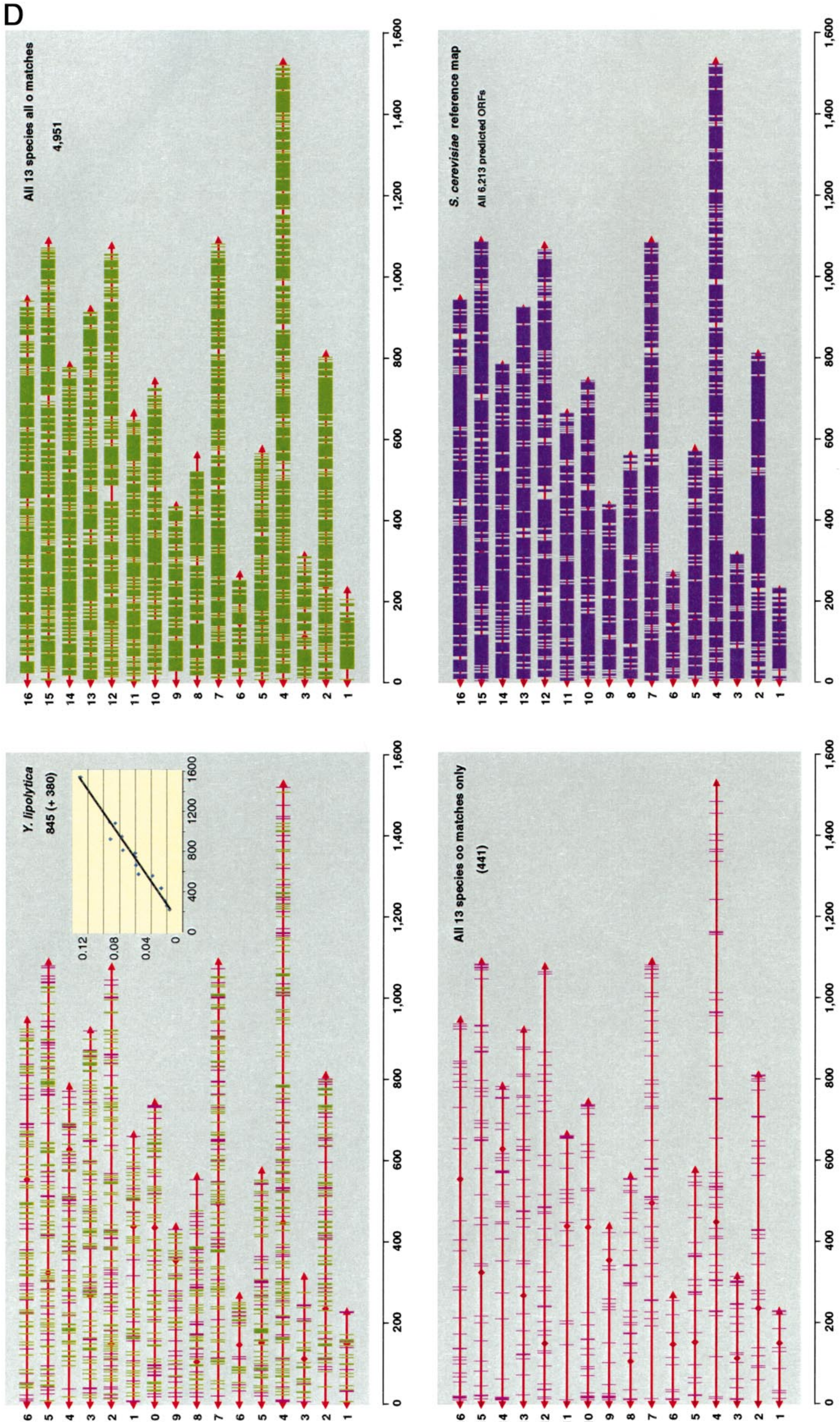


Fig. 2 (*continued*).

Fig. 2. Map location of the *S. cerevisiae* homologues of the yeast species sequenced. Each *S. cerevisiae* chromosome (designated by its number 1–16) is drawn to scale as a horizontal red line (scale in kb) with its telomeres and centromere represented by red void triangles and diamonds, respectively. For each of the 13 yeast species studied in this work (name in the upper right corner of each panel), the unambiguous *S. cerevisiae* homologues (alignments validated as 'o') are represented by green vertical bars (total number below the yeast name). All other possible homologues (alignments validated as 'oo') are represented by purple vertical bars (total number in brackets below the yeast name). Panels 14 and 15 show the location of the *S. cerevisiae* homologues to all 13 species, altogether. Panel 16 shows the location of all predicted *S. cerevisiae* ORFs using the same representation for reference of gene density variation along chromosomes. Inserts on upper right corner indicate the total number of genes on each *S. cerevisiae* chromosome having unambiguous homologues (ordinate) relative to the chromosome size (abscissa). Note that the size of chromosome 12 excludes rDNA.
←

rum) is close to *S. cerevisiae* and represents an important economical species due to its oenological properties. The list of the remaining species fill the gap between the members of the Hemiascomycete class. Some of them are important human pathogens such as *Candida tropicalis*, or are potentially useful for industrial purposes such as *Pichia sorbitophila*. All species were obtained from the CBS (Centraal bureau voor Schimmelcultures) or the CLIB (Collections des Levures d'Intérêt Biotechnologique) collections.

3.2. *Yeasts as cell factories*

Most of the arguments listed in this paragraph, on the impact of the hemiascomycetous yeasts on human life, were taken from the review of Demain et al. [44]. Yeasts are linked to human life for the production of beer (whose first records could be traced back to Sumeria around 7000 BC). Yeasts also produce wine, bread and various metabolic products such as vitamins, ethanol, citric acid, lipids, etc. Enzymes such as α-galactosidase used for the crystallisation of sugar from sugar beet is produced by *Saccharomyces pastorianus*. Species of the genus *Kluyveromyces* are lactose fermenters producing ethanol from by-products of the milk industry. *Candida utilis* is grown on cheap substrates unfavourable for other yeasts and is used for the production of animal nutriment but also for the production of some flavours such as ethyl acetate and acetaldehyde. Species from the genera *Candida*, *Yarrowia* and *Debaryomyces* assimilate hydrocarbons. Yet, their use to clean oil spill was unsuccessful so far. Citric acid is produced from *n*-alkanes, vegetable oils or glucose under aerobic conditions by *Y. lipolytica*. Several species of the genus *Pichia* are able to produce large quantities of extracellular phosphomannans from glucose under aerobic conditions, which can be used as food additives. *Z. rouxii* is able to depolymerise tannin extracts. More recently, recombinant DNA technology in yeast has allowed the expression of heterologous gene products such as hormones or vaccine [45]. Good knowledge of the yeast secretion pathway lead to efficient recovery of the recombinant proteins in the media. In addition, these products are glycosylated (though not exactly as in the human cells). The range of yeasts species used as hosts is growing. Most of the earlier applications of this technology used *S. cerevisiae* as was the case, for example, for the production of the surface antigen of hepatitis B virus [46]. Now the use of alternative yeasts is increasing. Heterologous expression is now also performed in *P. angusta, Y. lipolytica*, and *K. lactis*. The aspartyl protease chymosin, for example, which is the active constituent of the cheese rennet, is overproduced and secreted from *K. lactis* [47]. The number of sophisticated vectors specifically designed for appropriate expression in various yeast backgrounds is rapidly extending [48].

3.3. *Yeasts as human pathogens*

Despite their general usefulness, several yeast species are pathogenic for humans. Among the most frequent disease agents are the Hemiascomycetes *C. albicans*, *C. tropicalis* and *Candida glabrata* and the Basidiomycete, *Cryptococcus neoformans*. But a variety of other yeast species are occasionally found in patients. Even *S. cerevisiae* though extensively used in classical diet, may produce fatal complications in immunocompromised patients [49,50].

3.4. *A random sequencing approach for the exploration of the 13 yeast genomes*

We reasoned that a large scale comparative analysis between *S. cerevisiae* and the 13 selected species, all belonging to the Hemiascomycete class, should improve our knowledge on the structural, functional and evolutionary properties of yeast genomes. Because complete genome sequencing of 13 species would have been a major endeavour, we rationalised that a random sequencing approach should allow a quicker exploration upon which subsequent analysis could be performed. The *S. cerevisiae* genome sequence provided us with a unique reference to characterise each of the 13 other yeast species by sequence comparison, then allowing us to compare the structure of the specific genetic element discovered among the 13 species. We also rationalised that the identification by sequence comparisons of chromosomal rearrangements could help reconstitute part of the chromosome evolution history within that specific phylum.

To maximise our chances of detecting such rearrangements, the sequencing of 1000 bp from the two ends of 4–5 kb long inserts was most appropriate, because their distance corresponds to the average distance between neighbouring genes in *S. cerevisiae*. In addition, the number of RSTs should reveal a significant number of such rearrangements. Moreover, sequence divergence and other components of speciation, could be inferred from the same set of experiments.

3.5. *Phylogenetic relationships between the 13 selected yeast species as revealed by the sequences of rRNAs*

The sequences of the rDNA of the 13 yeast species, encoding 18S and 25S rRNA, were determined in this work thanks to their redundancy. These sequences were used to construct a cladogram (Fig. 1). A synoptic view of some characteristics pertinent to the 13 yeast species is available from this figure. The phylogenetic relationships among species determined from this cladogram nicely correspond to the distances deduced by using criteria such as the decrease of synteny conservation [51], the amino acid sequence divergence [52], and the distribution of the 'Ascomycetes-specific' genes [52]. These data permit to propose the phylogenetic position of *P. sorbitophila* close to that of the *Candida* species.

3.6. Genomic analysis and major achievements

The wealth of novel genomic sequences deciphered in the Génolevures programme on comparative genomics, ca. 45 000 000 of nucleotides distributed among 13 related species, provides a set of data to examine two questions: first, the identification of the set of genes specific to one phylum and the distribution of their products into distinct protein families; second, the distinction between mechanisms frequently proposed to play a major role in the genome reshaping and their implication on the fluctuation of gene family sizes.

With the random sequencing approach, about 20 000 new genes (18 109 genes with the 'o' notation or 22 844 genes with the 'o' plus 'oo' notations) were identified. They are described in Table 1 (accessible from websites) which lists the 6264 *S. cerevisiae* genes and indicates the corresponding orthologue in each species. Some of these new genes are not present in the *S. cerevisiae* reference genome. This could be explained by a low but significant level of species-specific genes [53] even if the polymorphism among *S. cerevisiae* strains could be invoked for a few cases.

The Génolevures project allowed the distinction of two sets of genes in the genome of *S. cerevisiae*: genes commonly conserved through different phyla during evolution, 'conserved-genes', and genes which have no homologues in a phylum other than the Ascomycetes, 'Ascomycete-specific genes'. The vast majority of this latter set, 70%, are genes that have an equivalent in the other yeast species, the remaining 30% being mostly composed of questionable ORFs [52]. It has to be noted that the sequences of the 'Ascomycete-specific genes' tend to diverge more rapidly in evolution than the sequences of the 'conserved-genes' [52].

The analysis of the statistical distribution of the number of yeast species providing homologues to each of the predicted *S. cerevisiae* ORFs leads to the estimation of 5651 coding genes in the *S. cerevisiae* genome [52]. These comparison studies also offer the opportunity to revisit the *S. cerevisiae* genome, as they reveal the presence of 50 novel genes in DNA segments previously considered as 'intergenic' [54].

The gene redundancy observed in *S. cerevisiae* [55], as in nearly all genomes sequenced until now, is recovered in the 13 other species. The classification in gene families as described by Blandin et al. [54] was used to conduct a statistical analysis of gene redundancy among the 13 yeast genomes. From this analysis it is concluded that even if the distribution of genes among families could be subjected to variations, gene family sizes are mostly conserved. The overall degree of gene redundancy seems comparable across all the Hemiascomycete species analysed [55].

A comparative functional classification of genes was conducted to link the physiological diversity observed between species to the variation in their respective gene contents. This analysis revealed that biological diversity result, mostly, from the variations of the distributions of rapidly evolving 'Ascomycete-specific genes' among functional classes. On the other hand, the number of species-specific genes seems very low [53]. Moreover there is no strong evidence of gene acquisition through horizontal gene transfer [53].

Unraveling the global gene content of each species was followed by the analysis of their arrangements in their respective genomes. A comparative view of the chromosomal gene distribution in each species is reported in Fig. 2, where the presence of the corresponding orthologues of *S. cerevisiae* is indicated per chromosome of *S. cerevisiae*. All chromosomal arms are homogeneously covered by orthologous genes indicating that each chromosomal fragment of *S. cerevisiae* possesses one equivalent in the 13 studied species.

The complete genome sequence of *S. cerevisiae* revealed the presence of 55 chromosomal fragments that have resulted from ancestral duplication events. However duplication could proceed from multiple repetitive duplication blocks or from a complete duplication of an ancestral genome. The latter hypothesis was retained by Wolfe [56], who sustained that the *S. cerevisiae* genome may result from a tetraploidisation event, occurring after the separation of the *Kluyveromyces* and *Saccharomyces* genus. Recently, a study on chromosomal rearrangements in the *Saccharomyces* sensu stricto conclude to the lack of correlation between these events and speciation [57]. The present work, with a broad range of species encompassing the Hemiascomycete class, offers an opportunity to reconsider this question. The analysis of the conservation of synteny between *S. cerevisiae* and each of the 13 other yeast species was approached by studying 9000 couples of genes. The set of yeast species chosen appears, a priori, appropriate for the study performed because the overall degree of synteny conservation with *S. cerevisiae* varies from as high as 98% for *S. bayanus* to only 10% for *Y. lipolytica* with a significant number of intermediate values. Thus, all phenomena of map evolution (inversion of gene orientation, frequency of gene deletion) could be quantified over a broad evolutionary scale.

4. Conclusions

Altogether, our results on the conservation of synteny [51], on genetic redundancy [54], and on functional classification of genes [53] led to the conclusion that speciation could result from a limited reshaping of the genetic repertoire [53] and that a whole genome duplication is not a prerequisite to explain the structure of the *S. cerevisiae* genome and other species of the *Saccharomyces* sensu stricto group.

The detailed analysis of the genome of the studied species is reported in 13 papers (7–19). Plasmids, Ty and related elements have been identified in numerous species, and in some cases as for *Debaryomyces hansenii*, the mitochondrial DNA has been entirely sequenced. Interestingly, when nuclear introns are identified they are still shorter than those observed in the *S. cerevisiae* genome and with the exception of *Kluyveromyces thermotolerans*, there is a deficit in genes encoding tRNAs and rDNA.

Acknowledgements: This work was supported in part by a BRG Grant (ressources génétiques des microorganismes No 11-0926-99). We thank H. Feldmann and A. Goffeau for careful reading of the manuscript. The yeast genetic department of the UPRES-A 7010 is a member of the 'Génopole Alsace-Lorraine'. B.D. is a member of Institut Universitaire de France.

References

[1] Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W. and Dujon, B. et al. (1996) Science 274, 563–567.

[2] Fischer, C., Ozouf-Costaz, C., Roest-Crollius, H., Dasilva, C., Jaillon, O., Bouneau, L., Bonillo, C., Weissenbach, J. and Bernot, A. (2000) Cytogenet. Cell Genet. 88, 50–55.

[3] Roest-Crollius, H., Jaillon, O., Bernot, A., Dasilva, C. and Bouneau, L. et al. (2000) Nature Genet. 25, 235–238.

[4] Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B. and Herrmann, R. (1997) Nucleic Acids Res. 25, 701–712.

[5] Alm, R.A. and Trust, T.J. (1999) J. Mol. Med. 77, 834–846.

[6] Razin, S., Yogev, D. and Naot, Y. (1998) Microbiol. Mol. Biol. Rev. 62, 1094–1156.

[7] Read, T.D., Brunham, R.C., Shen, C., Gill, S.R. and Heidelberg, J.F. et al. (2000) Nucleic. Acids. Res. 15, 1397–1406.

[8] *C. elegans* Sequencing Consortium, (1998) Science 282, 2012–2018.

[9] *D. melanogaster* Sequencing Consortium, (2000) Science 2897, 2181–2274.

[10] Lin, X., Kaul, S., Rounsley, S., Shea, T.P. and Benito, M.I. et al. (1999) Nature 402, 761–768.

[11] *A. thaliana* Sequencing Consortium, (1999) Nature 402, 769–777.

[12] Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S. and Hunt, A.R. et al. (1999) Nature 402, 489–495.

[13] Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H. and Yada, T. et al. (2000) Nature 405, 311–319.

[14] Ozier-Kalogéropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F. and Dujon, B. (1998) Nucleic Acids Res. 26, 5511–5524.

[15] Lodder, J. (1970) in: The Yeasts, A Taxonomic Study, 2nd edn. (Lodder, J., Ed.), pp. 1–33, North-Holland, Amsterdam.

[16] Kurtzman, C.P. (1998) in: The Yeasts, A Taxonomic Study, 4th edn. (Kurtzman, C.P. and Fell, J.W. Eds.), pp. 111–121, Elsevier, Amsterdam.

[17] Nishida, H. and Sugiyama, J. (1993) Mol. Biol. Evol. 10, 431–436.

[18] Kurtzman, C.P. and Blanz P.A. (1998) in: The Yeasts, A Taxonomic Study, 4th edn. (Kurtzman, C.P. and Fell, J.W. Eds.), pp. 69–74, Elsevier, Amsterdam.

[19] Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F. et al., (2000) FEBS Lett. 487, 13–16 (this issue).

[20] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P. et al., (2000) FEBS Lett. 487, 17–30 (this issue).

[21] Bon, E., Neuvéglise, C., Casaregola, S., Artiguenave, F., Wincker, P. et al., (2000) FEBS Lett. 487, 37–41 (this issue).

[22] Bon, E., Neuvéglise, C., Lépingle, A., Wincker, P., Artiguenave, F. et al., (2000) FEBS Lett. 487, 42–46 (this issue).

[23] Casaregola, S., Lépingle, A., Neuvéglise, C., Bon, E., Vang Nguyen, H. et al., (2000) FEBS Lett. 487, 47–51 (this issue).

[24] de Montigny, J., Straub, M.L., Potier, S., Tekaia, F., Dujon, B. et al., (2000) FEBS Lett. 487, 52–55 (this issue).

[25] Neuvéglise, C., Bon, E., Lépingle, A., Wincker, P., Artiguenave, F. et al. (2000) FEBS Lett. 487, 56–60 (this issue).

[26] Malpertuy, A., Llorente, B., Blandin, G., Artiguenave, F., Wincker, P. et al., (2000) FEBS Lett. 487, 61–65 (this issue).

[27] Bolotin-Fukuhara, M., Lemaire, M., Marmeisse, R., Montrocher, R., Termier, M. et al., (2000) FEBS Lett. 487, 66–70 (this issue).

[28] Llorente, B., Malpertuy, A., Blandin, G., Wincker, P., Artiguenave, F. et al., (2000) FEBS Lett. 487, 71–75 (this issue).

[29] Blandin, G., Llorente, B., Malpertuy, A., Wincker, P., Artiguenave, F. et al., (2000) FEBS Lett. 487, 76–81 (this issue).

[30] Lépingle, A., Casaregola, S., Bon, E., Neuvéglise, C., Vang Nguyen, H et al., (2000) FEBS Lett. 487, 82–86 (this issue).

[31] de Montigny, J., Spehner, C., Souciet, J.L., Tekaia, F., Dujon, B. et al., (2000) FEBS Lett. 487, 87–90 (this issue).

[32] Blandin, G., Ozier-Kalogéropoulos, O., Wincker, P., Artiguenave, F. and Dujon, B. (2000) FEBS Lett. 487, 91–94 (this issue).

[33] Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C. et al., (2000) FEBS Lett. 487, 95–100 (this issue).

[34] Hartung, K., Frishman, D., Hinnen, A. and Wolfl, S. (1998) Yeast 14, 1327–1332.

[35] Prillinger, H., Schweigkofler, W., Breitenbach, M., Briza, P. and Staudacher, E. et al. (1997) Yeast 13, 945–960.

[36] Wilmotte, A., van de Peer, Y., Goris, A., Chapelle, S. and de Baere, R. et al. (1993) Syst. Appl. Microbiol. 16, 436–444.

[37] Hamilton, G.E., Morton, P.H. and Young, T.W. (1999) Biotechnol. Bioeng. 64, 310–321.

[38] Chang, C.C., Ryu, D.D., Park, C.S. and Kim, J.Y. (1998) Biotechnol. Bioeng. 59, 379–385.

[39] Kabani, M., Boisrame, A., Beckerich, J.M. and Gaillardin, C. (2000) Gene 241, 309–315.

[40] Morlino, G.B., Tizzani, L., Fleer, R., Frontali, L. and Bianchi, M.M. (1999) Appl. Environ. Microbiol. 65, 4808–4813.

[41] Billard, P., Dumond, H. and Bolotin-Fukuhara, M. (1997) Mol. Gen. Genet. 257, 62–70.

[42] Mayer, A.F., Hellmuth, K., Schlieker, H., Lopez-Ulibarri, R., Oertel, S., Dahlems, U., Strasser, A.W. and van Loon, A.P. (1999) Biotechnol. Bioeng. 63, 373–381.

[43] Yoshikawa, S., Oguri, I., Kondo, K., Fukuzawa, M., Shimosaka, M. and Okazaki, M. (1995) FEMS Microbiol. Lett. 127, 39–143.

[44] Demain, A.L., Phaff, H.J. and Kurtzmann C.P. (1998) in: The Yeasts, A Taxonomic Study, 4th edn. (Kurtzman, C.P. and Fell, J.W. Eds.), pp. 13–19, Elsevier, Amsterdam.

[45] Kjeldsen, T., Pettersson, A.F., Hach, M., Diers, I., Havelund, S., Hansen, P.H. and Andersen, A.S. (1997) Protein Expr. Purif. 9, 331–336.

[46] Valenzuela, P., Medina, A., Rutter, W.J., Ammerer, G. and Hall, B.D. (1982) Nature 298, 347–350.

[47] van den Berg, J.A., van der Laken, K.J., van Ooyen, A.J., Renniers, T.C. and Rietveld, K. et al. (1990) Biotechnology 8, 135–139.

[48] Saliola, M., Mazzoni, C., Solimando, N., Crisa, A., Jung, G. and Fleer, R. (1999) Appl. Environ. Microbiol. 65, 53–60.

[49] Murphy, A. and Kavanagh, K. (1999) Enzyme Microb. Technol. 25, 551–557.

[50] Hennequin, C., Kauffmann-Lacroix, C., Jobert, A., Viard, J.P. and Ricour, C. et al. (2000) Eur. J. Clin. Microbiol. Infect. Dis. 19, 16–20.

[51] Llorente, B., Malpertuy, A., Neuvéglise, C., de Montigny, J., Aigle, M. et al. (2000) FEBS Lett. 487, 101–112 (this issue).

[52] Malpertuy, A., Tekaia, F., Casaregola, S., Aigle, M., Artiguenave, F. et al. (2000) FEBS Lett. 487, 113–121 (this issue).

[53] Gaillardin, C., Duchateau-Nguyen, G., Tekaia, F., Llorente, B., Casaregola, S. et al. (2000) FEBS Lett. 487, 134–149 (this issue).

[54] Blandin, G., Durrens, P., Tekaia, F., Aigle, A., Bolotin-Fukuhara, M. et al. (2000) FEBS Lett. 487, 31–36 (this issue).

[55] Llorente, B., Durrens, P., Malpertuy, A., Tekaia, F., Aigle, M. et al. (2000) FEBS Lett. 487, 76–81 (this issue).

[56] Wolfe, K.H. and Shields, D.C. (1997) Nature 387, 708–713.

[57] Fischer, G., James, S.A., Roberts, I.M., Oliver, S.G. and Louis, E.J. (2000) Nature 405, 451–454.