Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*

Odile Ozier-Kalogeropoulos*, Alain Malpertuy, Jeanne Boyer, Fredj Tekaia and Bernard Dujon

Unité de Génétique Moléculaire des Levures (URA 1300, CNRS and UFR 927, Université Pierre et Marie Curie, Paris, France) Institut Pasteur, 25 rue du Dr Roux, F-75724 Paris Cedex 15, France

Received June 11, 1998; Revised and Accepted September 29, 1998 DDBJ/EMBL/GenBank accession nos AJ229366–AJ230023

ABSTRACT

The genome of the yeast *Kluyveromyces lactis* was explored by sequencing 588 short tags from two random genomic libraries (random sequenced tags, or RSTs), representing altogether 1.3% of the *K.lactis* genome. After systematic translation of the RSTs in all six possible frames and comparison with the complete set of proteins predicted from the *Saccharomyces cerevisiae* genomic sequence using an internally standardized threshold, 296 *K.lactis* genes were identified of which 292 are new. This corresponds to ~5% of the estimated genes of this organism and triples the total number of identified genes in this species. Of the novel *K.lactis* genes, 169 (58%) are homologous to *S.cerevisiae* genes of known or assigned functions, allowing tentative functional assignment, but 59 others (20%) correspond to *S.cerevisiae* genes of unknown function and previously without homolog among all completely sequenced genomes. Interestingly, a lower degree of sequence conservation is observed in this latter class. In nearly all instances in which the novel *K.lactis* genes have homologs in different species, sequence conservation is higher with their *S.cerevisiae* counterparts than with any of the other organisms examined. Conserved gene order relationships (synteny) between the two yeast species are also observed for half of the cases studied.

INTRODUCTION

Comparative genomics is a rapidly growing field of investigation (1,2). Initially limited to the analysis of ESTs (3–6), or to the comparison of chromosomal regions (7), it can now be extended to complete genomes, thanks to the recent release of several microbial genomic sequences (8–21). If such comparisons are highly informative to identify or classify the many novel genes issued from large systematic sequencing programs, their results are of more limited significance to describe the mechanisms of molecular evolution because, with the sole exception of the two *Mycoplasma* species (22), the presently sequenced microorganisms are only distantly related to one another. Hence the fact that in nearly all organisms sequenced so far, sizeable fractions of the genes remain without homolog. Many such genes, even in the case of well studied organisms such as *Escherichia coli* (14), *Bacillus subtilis* (18) or the yeast *Saccharomyces cerevisiae* (23), have not been functionally characterized. It is therefore impossible to distinguish between the possible existence of functional orthologs whose sequence may have diverged beyond recognition, and genes that may be truly specific to a given phylogenetic group of organisms.

Of the organisms whose genomes have been completely sequenced, *S.cerevisiae* so far stands alone among the Eucaryotes (23). Comparisons of yeast sequences with the numerous human ESTs (24) or with the partial genomic sequence of *Caenorhabditis elegans* (25) confirm the phylogenetic relationship of *S.cerevisiae* with the animal kingdom, but evolutionary distances are too large to help us describe mechanisms of molecular evolution within Eucaryotes. For this goal, comparison of *S.cerevisiae* with closely related yeasts is the most appropriate approach. The yeasts represent a large and diverse taxonomic group and, at present, a significant number of sequences are available only for two species, the pathogenic yeast *Candida albicans* (http://candida.stanford.edu) and the fission yeast *Schizosaccharomyces pombe* (ftp://ftp.sanger.ac.uk/pub/PomBase/). In fact, although the fission yeast belongs to the yeast phylum, this species is very distantly related to *S.cerevisiae* and sequence comparisons give results not very different from the comparison of *S.cerevisiae* with *C.elegans.*

Using a conservative threshold, ~36% of the protein-coding genes predicted from the genome of *S.cerevisiae* remain without structural homologs in other organisms. Among these, there exist genes that have been functionally characterized in yeast (11% of total), genes whose function can be tentatively assigned from their structural homologs in yeast itself (3% of total), and genes (22% of the total) that remain of unpredictable function because none of their homologs in yeast is functionally characterized or because they have no structural homolog in yeast (26). To help us understand the origin and nature of the latter genes (called ‘orphans’; 27), we decided to explore the genome of a yeast

*To whom correspondence should be addressed. Tel: +33 1 40 61 30 59; Fax: +33 1 40 61 34 56; Email: odozier@pasteur.fr species closely related to *S.cerevisiae* (28). *Kluyveromyces lactis* was selected because, from the limited number of genes previously characterized in this species (82 protein-coding genes were found in public sequence databases at the start of this work; 29), it appeared a limited sequence divergence from *S.cerevisiae* (mean amino acid identity, 83%; standard deviation, 19%), and because the two genomes share a number of similarities such as the rare occurrence of introns, comparable gene density, and short intergenic regions. In addition, and despite the fact that *K.lactis* has only six chromosomes, its genome size and total gene number are similar to that of *S.cerevisiae*, and even the centromeres are of comparable structure (30). Like *S.cerevisiae*, *K.lactis* is also a microorganism of biotechnological interest (31,32) that has been studied for many years, notably by Louis Pasteur (33).

Most of the *K.lactis* genes characterized so far were isolated on the basis of their structural or functional similarity with *S.cerevisiae*. In order to allow a significant comparison between the two genomes, and at the same time to identify possible homologs to *S.cerevisiae* orphans (none exists in the limited list of *K.lactis* genes available), we therefore decided to explore the genome of *K.lactis* using a totally random approach by sequencing inserts from a genomic library. A similar approach has been adopted on a limited scale for the filamentous fungus *Ashbya gossypii* (34), and for *Drosophila melanogaster* (35), although in the latter case no genomic sequence of closely related organisms is available yet. This approach, applied to only 1.3% of the *K.lactis* genome, proved not only useful for the rapid identification of many novel *K.lactis* genes (5% of total), but also informative for the identification of homologs to *S.cerevisiae* orphan genes.

MATERIALS AND METHODS

Construction of the *K.lactis* random genomic library

DNA preparation. Cells from a 500 ml YPglu (2% glucose, 1% bactopeptone, 1% yeast extract) culture of *K.lactis* strain CBS2359 grown overnight at 30°C were harvested by centrifugation, rinsed in water, resuspended in 50 ml of spheroplasting buffer [1 M Sorbitol, 50 mM Na–K phosphate buffer pH 7.5, 25 mM EDTA, 1% (v/v) β-mercaptoethanol] containing 60 mg of Zymolyase (20 000 U; Seikagaku Kogyo), and incubated for 1 h at 30°C with gentle agitation. Spheroplasts were collected by low speed centrifugation (3000 *g* for 10 min) and resuspended in 20 ml of lysis buffer [TE buffer pH 8.0 with 1% (w/v) sodium dodecyl sulfate] containing 2 mg of Proteinase K (Boehringer Mannheim), and incubated for 2 h at 50°C, followed by 30 min at 65°C. Two phenol–chloroform extractions were performed and the aqueous phase was precipitated by addition of 0.1 vol of 3 M NaCl and 0.8 vol of isopropanol. Precipitated DNA was taken out of the solution with a sterile loop, washed with a 70% (v/v) ethanol solution, air-dried and redissolved in 1 ml of TE buffer pH 8.0. Eight microliters of a 10 mg/ml solution of RNase A (Boehringer Mannheim) were added, and the solution was incubated for 30 min at 37°C, followed by a second ethanol–0.3 M NaCl precipitation. The DNA was washed again with a 70% ethanol solution, air-dried and finally dissolved in 500 µl of TE (pH 8.0).

DNA fragmentation. Two *K.lactis* genomic libraries were constructed in parallel, a 'long-fragment' library containing 2–3 kb inserts and a 'short-fragment' library of 0.8–1.2 kb inserts. *Kluyveromyces lactis* DNA was randomly fragmented by nebulization (DNA Nebulizer, GATC GmbH, Germany). For the long-fragment library, 100 µg of the DNA (1 µg/µl solution) were added to 1.9 ml of TE pH 7.5 and the solution was nebulized for 45 s using pressurized argon (10^5 Pa). For the short-fragment library, 50 µg of DNA were added to 750 µl of 80% glycerol (v/v) and 1.2 ml of TE (pH 7.5) and the solution was nebulized for 90 s under the same conditions. Each nebulized DNA solution was aliquoted into six microcentrifuge tubes, precipitated by addition of a solution of ethanol–3 M sodium acetate. Pellets were rinsed with a 70% ethanol solution, dried and redissolved in 50 µl of TE (pH 7.5). The contents of the six tubes were pooled for a second precipitation, and the recovered DNA (~80 µg) was finally redissolved in 20 µl TE (pH 7.5). DNA preparations were end-filled (30 min at 15°C) using T4 DNA Polymerase (N.E. Biolabs) and the four deoxyribonucleotide triphosphates, and loaded on a preparative low-melting agarose gel. After electrophoresis, fragments corresponding in size to ~1 kb (short-fragment library), or ~3 kb (long-fragment library) were excised from the gels and extracted using QIAquick columns, following the recommendations of the manufacturer (Qiagen Inc.). Size-fractionated DNA was eluted in 30 µl of water.

Vector preparation and ligation. Aliquots of 10 µg of pBluescript SK+ vector were digested by *Eco*RV and dephosphorylated using Calf Intestinal Phosphatase (N.E. Biolabs) following the manufacturer's protocol. After phenol extraction and precipitation, the DNA was redissolved in 20 µl of water and purified by low-melting agarose gel electrophoresis. Linearized vector (500 ng) and *K.lactis* DNA fragments (1 µg) were ligated overnight at 16°C (T4 DNA ligase; N.E. Biolabs). After phenol–chloroform extraction and precipitation, the DNA was recovered in 20 µl of water. One-tenth of each ligation mix was used to transform *E.coli* DH5α cells by electroporation (36). Bacteria were plated on LB medium containing ampicillin (100 mg/ml). Each ligation mix had the potential to yield 20–30 000 primary clones with inserts (white colonies).

Sequencing strategy and sequence quality

Inserts from the large-fragment library were sequenced from both ends, using direct and reverse end-labelled primers, on double-stranded DNA prepared with QIAGEN columns (37). Inserts from the short-fragment library were sequenced from one end, using direct end-labelled primer and single-stranded PCR-amplified DNA, prepared using magnetic beads with covalently coupled streptavidin (38), Dynabeads M-280 (Dynal AS, Norway). The capture of the biotin-labelled single-stranded DNA was automated on a Biomek 2000 workstation (Beckman), coupled with a magnetic robot (Polyseq, PolyGen GMBH). Sequencing reactions were analyzed on automatic fluorescent DNA sequencers (ALF and ALFexpress, Pharmacia). To ensure high and even sequence quality for the entire set of genomic tags (designated here random sequenced tags, or RSTs), each sequencing profile was inspected on a Sparc II workstation using the Alfsplit and Ted programs of the Staden package (39) and sequences containing any base-calling ambiguity, or <100 nt were eliminated. Putative frameshifts issued from errors in our single-read sequences detected by BLASTX comparisons (see below), or by using DNA-Strider dot plot matrices (40), were corrected according to DNA–DNA alignments. The average error rate of our RSTs is 0.5% for nucleotide substitution and 0.3% for base addition/omission, as estimated from fragments of the pBluescript vector that were resequenced in a few empty clones.

Analysis of the *K.lactis* RSTs

Each of the 658 RSTs was first compared with all others to detect duplicates or partially overlapping sequences. Seven were found to be partially overlapping and merged into three distinct contigs. The resulting 654 RSTs were then compared with rDNA, tRNA, mitochondrial DNA and intergenic sequences of *S.cerevisiae* to identify *K.lactis* homologs. Sixty-six such sequences were found. All above comparisons were carried out using the BLASTN program (version 1.4) (41).

Each of the 588 remaining RSTs was systematically compared with a complete, non-redundant database of *S.cerevisiae* protein sequences (compiled in our laboratory and containing 6182 predicted protein products; unpublished data), using the BLASTX program (version 1.4) (41) with the PAM100 substitution matrix. This program compares all possible translation products of a nucleotide query sequence (all six frames) against a protein sequence database. In order to determine the significance of BLASTX probability scores, two sets of random sequences, identical in number and in size distribution to the actual *K.lactis* RSTs, were extracted from the *S.cerevisiae* genomic sequence using a Perl script written for this purpose (Randomseq). Sequences from these two control sets (called random extracted sequences, or RESs) were systematically compared with the complete, non-redundant database of *S.cerevisiae* protein sequences, as above. Sequence divergence was calculated from amino acid sequence alignments produced by the ALIGN program using the PAM100 substitution matrix (42,43).

Accession numbers and access to annotated sequences

The 658 sequences have been deposited in EMBL and are accessible with the accession nos AJ229366–AJ230023. Annotated sequences are accessible at the MIPS site (http://www.mips.biochem.mpg.de/mips/yeast/)

RESULTS

Analysis of RSTs from the *K.lactis* genome

We have rapidly explored the genome of *K.lactis* by single-pass sequencing of 658 inserts from randomly picked clones of two independent genomic libraries (Table 1). After elimination of sequences corresponding to mtDNA, rRNA or tRNA (Materials and Methods), each sequence that read longer than 100 nt, called RST (a total of 588), was systematically translated in the six possible frames and compared with: (i) a complete, non-redundant database of *S.cerevisiae* protein sequences; (ii) the predicted translation products of 13 fully sequenced microorganism genomes (8–12, 14–21) and of 81% of the *C.elegans* genome (http://www.sanger.ac.uk/Projects/C_elegans/) and (iii) all publically available protein sequences (44).

The fact that RSTs are randomly distributed with respect to *K.lactis* ORFs, and are of limited average size (267 bp) with a large range of variation (from 100 nt, our lower limit, to 579 nt), makes BLASTX probability scores [sum(*P*) values] complex to interpret. In order to select the best possible limit to distinguish the *K.lactis* RSTs having *S.cerevisiae* homolog(s) from those that do not, we devised an internal control consisting of two sets of RESs from the *S.cerevisiae* genome, each identical in number and size distribution to the actual *K.lactis* RST set. Figure 1 shows the distribution of the best *P*-value scores obtained for the two sets of RESs when compared with the *S.cerevisiae* complete database, together with actual results for the *K.lactis* RSTs. Best *P*-value scores range from 10^{-126} to 1 for the *K.lactis* RSTs and from 10^{-154} to 1 for the *S.cerevisiae* RESs. Note that each RES containing a long enough segment of a *S.cerevisiae* ORF possesses a self-score against the *S.cerevisiae* database, whereas RSTs do not, resulting in the excess of RES scores over RST scores for *P*-values $<10^{-20}$. Examination of sequence alignments indicate that 84% of the RESs contain ORF fragments giving such self-scores, while the remaining 16% of the RESs only contain intergenic sequences or very short, unrecognizable ORF fragments. These figures are consistent with the known distribution of *S.cerevisiae* ORF sizes. If we assume (Discussion) that a similar distribution exists for the RSTs with respect to the *K.lactis* ORFs, a total of 469 RSTs should contain an ORF, and only 119 should represent intergenic sequences or short ORF fragments. Since all RESs corresponding to intergenic sequences gave best *P*-values $\geq 10^{-4}$, we consider this limit to represent non-significant matches with the *S.cerevisiae* translation products. Using the same limit for RSTs (a conservative limit because it ignores sequence divergence between the two species), we found instead that 309 *K.lactis* RSTs have at least one significant homolog among the *S.cerevisiae* genes, and the remaining 279 do not. We ruled out the possibility that the latter might represent contaminating DNA (e.g. *E.coli* or vectors) by analyzing 10 of them taken at random, and showing that they specifically hybridize with purified

Table 1. Overall characteristics of *K.lactis* RSTs

A. Breakdown of *K. lactis* RSTs*	**Total number**
based on their origin	
Short library fragments	105
Long library fragments:	
- two ends	332
- one end	221
Total	658
based on their comparison with *S. cerevisiae* and other genomes	
- ORFs whose translation products show:	
- homology with at least one *S. cerevisiae* protein **	309
- closer homology with proteins of other species than *S. cerevisiae*	5
- Intergenic regions of *K. lactis* or ORFs whose translation products show no significant similarity with any organism	274
- Ribosomal DNA	29
- tRNA genes	7
- Mitochondrial DNA or tracts of nucleotide repeats	33
- Centromere (CEN3***)	1
B. Size and composition of *K. lactis* RSTs	
Average length (nucleotides)	267
Standard deviation (nucleotides)	103
Average G+C content (%):	
- global	38.6
- in ORFs	40.2

*Only RSTs with at least 100 nt of unambiguous base-calling were retained (Materials and Methods).

**Amino acid identity: mean, 63.6%; standard deviation, 49.7%. Nucleotide identity: mean, 66.6%; standard deviation, 13.3%.

***(ref. 30).

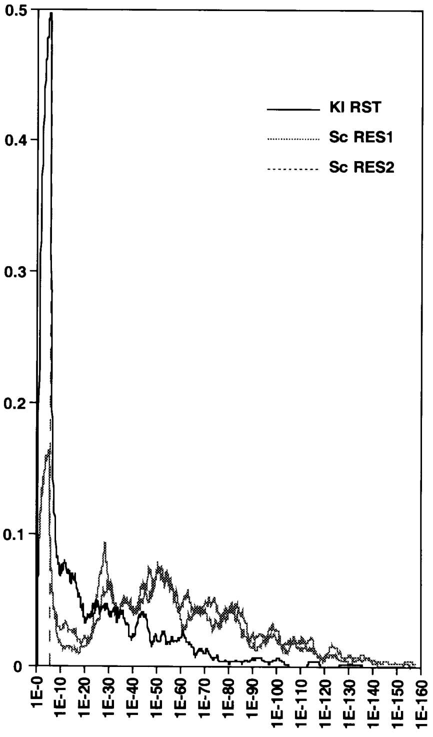


Figure 1. Frequency distribution of best *P*-value scores in BLASTX comparisons of *K.lactis* RST and *S.cerevisiae* RES nucleotide sequences with all predicted *S.cerevisiae* gene product sequences. The 588 *K.lactis* RSTs (solid line), or the same number of *S.cerevisiae* RESs used as standards (dotted lines), were ranked by increasing best *P*-value scores after comparisons with the complete set of *S.cerevisiae* translation products (Materials and Methods), and the frequency distributions were calculated using logarithmic windows of 1E-0.5 and sliding steps of 1E-0.1. *x*-axis, best *P*-value scores; *y*-axis, frequency of RSTs or RES in windows relative to total. Chosen threshold for homology significance (see text) is 1E-4 (vertical dashed line).

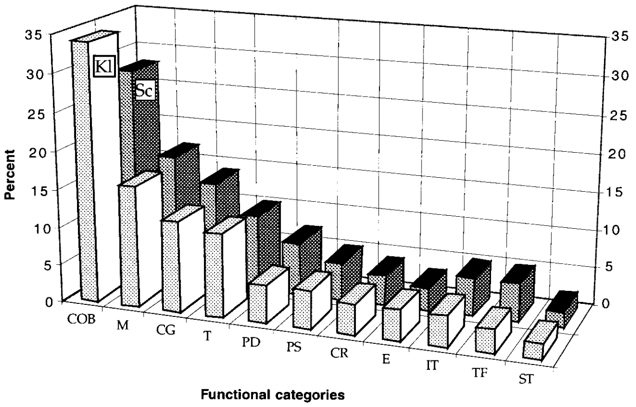


Figure 2. Breakdown of *K.lactis* RSTs into functional categories as defined by their *S.cerevisiae* orthologs. Functional categories are as defined in the MIPS Gazetteer (26): cellular organization and biogenesis (COB); metabolism (M); cell growth, cell division and DNA synthesis (CG); transcription (T); protein destination (PD); protein synthesis (PS); cell rescue (CR); energy (E); intracellular transport (IT); transport facilitation (TF); and signal transduction (ST). Bars represent the relative proportion (%) of each category when the complete list of functionally characterized proteins of *S.cerevisiae* is considered (Sc row), and when the subset of orthologs to the *K.lactis* RSTs is considered (Kl row). In both distributions, a single gene product can be assigned to more than one category (26). The two distributions are not significantly different, (χ^2 = 12.27, df =10, $P < 0.20$).

K.lactis DNA (data not shown). By careful examination of all alignments, we have also ruled out the possibility that the 279 *K.lactis* RSTs might contain genes with weak homologies to *S.cerevisiae* that would have been eliminated by too stringent a threshold. Furthermore, no significant sequence homology was found when comparing those *K.lactis* RSTs with all intergenic sequences of *S.cerevisiae* using BLASTN.

New *K.lactis* genes identified from their *S.cerevisiae* homologs

In cases in which a single *S.cerevisiae* gene product was found to give a significant BLASTX *P*-value score ($<10^{-4}$) with a *K.lactis* RST translation product, the corresponding gene was considered to be the ortholog of the *K.lactis* RST (Discussion). In cases in which several distinct *S.cerevisiae* gene products gave significant *P*-value scores with a *K.lactis* RST, we considered, arbitrarily, that a ratio >200 between the best *P*-value score and the next one was sufficient to define the ortholog. In the few cases in which this ratio was <200, protein and nucleic acid alignments were used to help us to identify the likely orthologs. In total, 297 RSTs corresponding to 284 different *K.lactis* genes fall in one of the above classes (Table 2A). Twelve other RSTs remained, representing 12 different *K.lactis* genes having two or more possible homologs in *S.cerevisiae*. In such cases, all possible homologs are listed (Table 2B). Two *K.lactis* RSTs share homology with *S.cerevisiae* ORFs considered as questionable or even disregarded (YDR445c and A-A110) in the MIPS classification (http://www.mips.biochem.mpg.de/mips/yeast/). In these two cases, partially overlapping ORFs were retained as more likely candidates for actual genes (YDR444w and YAR045w, respectively). The present finding of *K.lactis* homologs suggests the opposite.

Note that among the 296 *K.lactis* genes identified here, only four were previously described (Table 2), and 292 are novel. This is 3.5 times more than the total number of *K.lactis* genes previously identified at the molecular level (29).

Functional classification of the novel *K.lactis* genes based on their *S.cerevisiae* orthologs

Computation of the results shown in Table 2 shows that 58% of the identified *K.lactis* RSTs correspond to functionally known genes of *S.cerevisiae*, 22% to *S.cerevisiae* genes whose functions were assigned based on structural homology of their products with other species, and 20% correspond to genes of *S.cerevisiae* previously without homologs. Since our collection of RSTs represents a random sample of the *K.lactis* genome, it is interesting to examine and compare the distribution of their *S.cerevisiae* orthologs relative to the previously defined functional categories (26). Representatives of all these categories are found among the structural homologs of the *K.lactis* genes identified in this work, with no obvious bias in favour of any category (Fig. 2). As in *S.cerevisiae*, the most frequently observed *K.lactis* genes are involved in cell organization and biogenesis. Interestingly, many of the novel *K.lactis* genes discovered here correspond to functions that have been poorly studied, or not studied at all, in this organism. This is the case, for example, for genes involved in the cell cycle (SSN3, CDC37, CDC4, CDC33 and CDC53), chromosome segregation (SMC1), RNA splicing (PRP28 and

Table 2. List of the *K.lactis* RSTs having *S.cerevisiae* homologs: *K.lactis* RSTs were classified by increasing *P*-values and are indicated with their nomenclature, accession number and size (in bp). Homologous *S.cerevisiae* genes are identified by the systematic nomenclature and the corresponding gene name when available

(A) *K.lactis* RSTs having a single *S.cerevisiae* homolog

K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession n°	Size sequence (nt)	BLASTX Smallest Sum Probability P(N)	Alignment length (nt)	% identity (aa)	Systematic nomenclature	Gene name	Brief identification
okam3b02r	AJ229626	562	1.4 E-126	468	92	YNR001c	*CIT1*	citrate synthase, mitochondrial
okam3a09r	AJ229617	469	7.6 E-114	468	86	YPL042c	*SSN3*	cyclin-dependent ser/thr prt kinase
okam3g09r	AJ229715	496	3.0 E-113	495	84	YBR038w	*CHS2*	chitin synthase II
am1c01d	AJ229375	489	3.0 E-100	471	77	YMR229c	*RRP5*	processing of pre-ribosomal RNA
okam5a05d	AJ229837	464	7.1 E-99	393	89	YLL008w	*DRS1*	RNA helicase of the DEAD box family
okam3d04r	AJ229665	561	6.5 E-98	507	70	YPR176c	*BET2*	geranylgeranyltransferase type II beta subunit
okam5g05r	AJ229937	405	1.1 E-96	396	96	YOR117w	*YTA1*	26S proteasome subunit
okam5b02r	AJ229856	332	4.9 E-92	330	97	YER133w	*GLC7*	ser/thr phosphoprt phosphatase 1, catalytic chain
okam3a09d	AJ229616	523	1.1 E-91	459	77	YPL043w	*NOP4*	nucleolar prt
okam4e07d	AJ229788	394	3.7 E-90	357	92	YLR397c	*AFG2*	member of the Sec18p, Pas1p, Cdc48p, TBP-1 family of ATPases
okam3d06r	AJ229667	464	2.8 E-89	447	77	YLR069c	*MEF1*	translation elongation factor G, mitochondrial
okam4a09d	AJ229741	448	4.4 E-87	441	72	YPL215w	*CBP3*	required for assembly of cytochrome bc1 complex
okam1c06d	AJ229567	382	3.7 E-86	381	86	YGL026c	*TRP5*	tryptophan synthase
okam3b04d	AJ229628	390	3.9 E-83	387	81	YEL031w	*SPF1*	P-type ATPase
am1f06d	AJ229395	377	7.6 E-81	366	82	YPL116w	*HOS3*	prt with simil. to Hda1p, Rpd3p, Hos2p, and Hos1p
okam1d04r	AJ229570	562	2.6 E-75	180	78	YDR168w	*CDC37*	cell division control prt
okam5d04r	AJ229887	369	2.0 E-74	363	79	YOL058w	*ARG1*	argininosuccinate synthetase
okam2a08r	AJ229600	388	2.6 E-71	159	87	YPR201w		simil. to B.subtilis hypo. prt
okam3a06r	AJ229613	347	4.9 E-71	333	82	YDR388w	*RVS167*	reduced viability upon starvation prt
okam3d10d	AJ229673	330	6.2 E-71	246	95	YLR175w	*CBF5*	centromere/microtubule binding prt
okam4b08d	AJ229753	393	9.7 E-69	351	74	YDL132w	*CDC53*	controls G1/S transition
okam3b06d	AJ229632	384	6.3 E-68	264	82	YPL111w	*CAR1*	arginase
okam4b04d	AJ229748	377	2.3 E-67	273	73	YOR067c	*ALG8*	glucosyltransferase
okam3f08r	AJ229700	360	2.7 E-67	354	66	YJL165c	*HAL5*	ser/thr prt kinase
okam6a01r	AJ229963	333	3.7 E-66	327	75	YKR031c	*SPO14*	phospholipase D
okam3f02r	AJ229690	313	1.0 E-64	300	81	YGL256w	*ADH4*	alcohol dehydrogenase IV
okam6d05r	AJ230017	346	3.6 E-64	327	69	YBR260c		simil. to C.elegans GTPase-activating prt
am2g03r	AJ229457	321	2.1 E-63	309	79	YJL079c	*PRY1*	homology to the plant PR-1 class of pathogen related prts
okam11g05d	AJ229544	417	3.9 E-63	309	75	YML093w		simil. to P.falciparum liver stage antigen LSA-1
okam4f07r	AJ229800	271	2.0 E-61	179	86	YOR294w		simil. to human hypo. prt
okam5c08r	AJ229878	261	6.6 E-61	258	99	YDR394w	*YTA2*	26S proteasome subunit
okam3f02d	AJ229689	323	6.7 E-61	162	78	YKL216w	*URA1*	dihydroorotate dehydrogenase
okam5d02r	AJ229885	324	7.8 E-61	318	72	YKR018c		strong simil. to hypo. prt YJL082w
okam6b06d	AJ229987	282	1.1 E-60	279	82	YBR114w	*RAD16*	nucleotide excision repair prt
okam11b01d	AJ229484	314	4.9 E-60	312	76	YHR064c		simil. to heat shock prts
okam5h01r	AJ229951	317	1.2 E-59	282	77	YER091c	*MET6*	5-m.tetrahydropteroyltriglutamate-h.cysteine m.transferase
okam3d09r	AJ229672	410	2.0 E-59	360	59	YNR003c	*RPC34*	DNA-directed RNA pol. III, 34 KD subunit
okam4g03r	AJ229812	278	5.1 E-59	276	86	YBR221c	*PDB1*	pyruvate dehydrogenase (lipoamide) beta chain precursor
okam1f02d	AJ229581	450	2.1 E-58	300	63	YMR231w	*PEP5*	vacuolar biogenesis prt
okam5c07r	AJ229877	263	8.4 E-58	261	82	YHR030c	*SLT2*	ser/thr prt kinase of MAP kinase family
okam4a02r	AJ229737	448	1.6 E-57	147	65	YOR093c		simil. to S.pombe hypo. prt SPAC22F3.04
okam4f05r	AJ229796	282	5.1 E-57	282	82	YGR144w	*THI4*	thiamine-repressed prt
okam1b02r	AJ229559	291	7.2 E-57	267	74	YIL035c	*CKA1*	casein kinase II, catalytic alpha chain
am1c08d	AJ229377	288	1.5 E-56	285	88	YDL102w	*CDC2*	DNA-directed DNA pol. delta, catalytic 125 KD subunit
okam5f05d	AJ229921	266	2.5 E-56	252	81	YOR378w		strong simil. to aminotriazole resistance prt
okam4f12r	AJ229806	288	3.1 E-56	429	79	YJR109c	*CPA2*	arginine-specific carbamoylphosphate synthase, large chain
okam3h06r	AJ229729	307	3.6 E-56	282	79	YOR317w	*FAA1*	long-chain-fatty-acid-CoA ligase
am1d06d	AJ229382	302	5.1 E-55	300	76	YBR208c	*DUR1,2*	urea amidolyase
okam3f07d	AJ229697	481	1.2 E-54	294	68	YPL105c		simil. to Smy2p
okam4d12r	AJ229782	295	1.3 E-53	219	96	YGR152c	*RSR1*	GTP-binding prt
okam5f04d	AJ229919	313	2.0 E-53	285	75	YOR367w		simil. to mammalian smooth muscle prt SM22
okam5a12r	AJ229852	288	9.3 E-53	288	73	YOL006c	*TOP1*	DNA topoisomerase I
okam4e03r	AJ229785	235	2.1 E-52	231	87	YBL015w	*ACH1*	acetyl-CoA hydrolase
okam5c12d	AJ229882	441	3.4 E-52	396	74	YPL110c		simil. to C.elegans hypo. prt, weak simil. to Pho81p
okam3a02d	AJ229606	473	5.5 E-52	306	59	YOR361c	*PRT1*	translation initiation factor eIF3 subunit
okam6b06r	AJ229988	253	7.3 E-52	252	80	YBR115c	*LYS2*	L-aminoadipate-semialdehyde dehydrogenase, large subunit
okam5a03r	AJ229836	324	4.3 E-51	179	59	YLR386w		hypo. prt
okam11a05d	AJ229478	226	7.4 E-50	225	88	YBR245c		strong simil. to SNF2/SWI2 DNA binding regulatory prt
okam5b03d	AJ229857	256	1.7 E-49	246	82	YBR236c	*ABD1*	methyltransferase
okam3h02r	AJ229722	275	1.9 E-49	210	90	YJR007w	*SUI2*	translation initiation factor eIF2, alpha chain
okam4b06r	AJ229752	209	2.4 E-49	207	90	YOR168w	*GLN4*	glutaminyl-tRNA synthetase
am2d03d	AJ229436	340	4.1 E-49	339	70	YNL064c	*YDJ1*	mitochondrial and ER import prt

Table 2. (A) (*continued*)

K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession n°	Size sequence (nt)	BLASTX Smallest Sum Probability P(N)	Alignment length (nt)	% identity (aa)	Systematic nomenclature	Gene name	Brief identification
am1g02d	AJ229400	374	4.3 E-49	306	67	YMR239c	*RNT1*	double-stranded ribonuclease
okam3a06d	AJ229612	415	5.8 E-49	264	66	YDR387c		simil. to ltr1p and ltr2p and E.coli araE
okam6c04r	AJ230001	385	2.4 E-48	300	67	YGR046w		hypo. prt
okam3b12r	AJ229643	449	2.5 E-48	234	80	YOL018c		simil. to Pep12p
okam5b03r	AJ229858	356	4.5 E-47	180	75	YBR237w	*PRP5*	pre-mRNA processing RNA-helicase
okam4d02r	AJ229768	208	3.8 E-46	207	80	YKR096w		simil. to mitochondrial aldehyde dehydrogenase Ald1p
okam5c06r	AJ229876	306	5.3 E-46	303	65	YGR244c		strong simil. to rumen fungus beta-succinyl CoA synthetase
okam4a06d	AJ229740	389	1.7 E-45	387	50	YHR052w		weak simil. to P.yoelii rhoptry prt
okam3f10r	AJ229703	341	3.9 E-45	264	72	YIR004w		simil. to Caj1p, Ydj1p and to DNAJ-like prts
okam1e09r	AJ229579	410	2.9 E-44	231	82	YKL190w	*CNB1*	calcineurin B, regulatory subunit
okam5g10d	AJ229944	313	9.1 E-44	306	60	YDR430c		simil. to C.perfringens hypo. hypA prt
okam6b09r	AJ229992	213	3.5 E-43	210	94	YPR173c	*VPS4*	vacuolar sorting prt
okam4d03r	AJ229770	207	5.3 E-43	207	83	YLL018c	*DPS1*	aspartyl-tRNA synthetase, cytosolic
okam3d02r	AJ229661	564	5.9 E-43	183	49	YDR407c		weak simil. to Myo1p
okam11b04d	AJ229487	177	7.3 E-43	177	97	YPR181c	*SEC23*	component of COPII coat of ER-golgi vesicles
okam6a03r	AJ229965	297	1.0 E-42	288	59	YLR389c	*STE23*	protease involved in a-factor processing
okam3a10d	AJ229618	328	1.3 E-42	177	70	YLR345w		simil. to Pfk26p and other 6-phosphofructo-2-kinases
am1b11d	AJ229373	327	2.1 E-42	264	71	YDL213c		RNA recognition domain in the N-terminal region
okam6c08r	AJ230006	333	2.1 E-42	177	95	YKL081w	*TEF4*	translation elongation factor eEF1, gamma chain
okam3d08r	AJ229671	421	5.4 E-42	180	72	YGR187c	*HGH1*	weak simil. to human Hmg1p and Hmg2p
okam5c12r	AJ229883	300	7.7 E-42	132	61	YPL110c		simil. to C.elegans hypo. prt, weak simil. to Pho81p
okam4e05r	AJ229787	208	8.7 E-42	204	82	YNL006w		simil. to Met30p
okam5a03d	AJ229835	414	2.2 E-41	213	75	YLR387c		simil. to YBR267w
okam4g07d	AJ229817	218	2.6 E-41	216	79	YDR234w	*LYS4*	homoaconitase
am2c02d	AJ229427	500	3.3 E-41	180	75	YDL130w	*RPLA3*	acidic ribosomal prt L44prime
am1b06d	AJ229372	410	4.7 E-41	285	61	YGR208w	*SER2*	phosphoserine phosphatase
am1b04d	AJ229371	314	7.0 E-41	312	57	YFL036w	*RPO41*	DNA-directed RNA pol., mitochondrial
okam5b07d	AJ229865	205	9.0 E-41	204	82	YMR015c	*ERG5*	C-22 sterol desaturase
am1e12r	AJ229390	207	2.6 E-40	204	81	YML008c	*ERG6*	S-adenosyl-methionine delta-24-sterol-c-methyltransferase
okam5d06d	AJ229888	299	3.0 E-40	180	65	YOR274w	*MOD5*	tRNA isopentenyltransferase
okam3e08d	AJ229681	270	5.4 E-40	84	79	YDR081c	*PDC2*	pyruvate decarboxylase regulatory prt
okam5f08d	AJ229925	234	5.7 E-40	147	80	YJL005w	*CYR1*	adenylate cyclase
okam11g08d	AJ229545	255	7.9 E-40	252	61	YFL008w	*SMC1*	chromosome segregation prt
okam3f12r	AJ229707	415	3.3 E-39	114	63	YBR239c		weak simil. to transcription factor Put3p
okam4g01d	AJ229807	225	3.0 E-38	216	72	YOL139c	*CDC33*	translation initiation factor eIF4E
okam1d05r	AJ229571	396	3.6 E-38	375	50	YDL202w		ribosomal prt, mitochondrial
okam11d06d	AJ229511	166	1.1 E-37	162	96	YLR058c	*SHM2*	serine hydroxymethyltransferase, cytoplasmic
okam3a01d	AJ229604	413	1.6 E-37	78	58	YPL226w		simil. to translation elongation factor eEF3
am1f12r	AJ229398	206	4.8 E-37	204	82	YNL329c	*PAS8*	peroxisomal assembly prt
okam4f08r	AJ229802	198	8.4 E-37	110	81	YOR093c		simil. to S.pombe hypo. prt SPAC22F3.04
okam11c08d	AJ229501	210	3.4 E-36	210	70	YPL169c	*MEX67*	factor for nuclear mRNA export
okam5a09d	AJ229845	232	4.4 E-36	210	77	YLR086w		simil. to chromosome condensation prts
okam5a06r	AJ229840	232	6.2 E-36	177	78	YHR023w	*MYO1*	myosin-1 isoform (type II myosin) heavy chain
okam5h03r	AJ229955	302	9.0 E-36	222	64	YMR321c		strong simil. to hypo. prts YPL273w and YLL062c
okam5h12r	AJ229962	294	9.7 E-36	291	57	YIL112w		simil. to ankyrin and coiled-coil prts
okam5f04r	AJ229920	206	3.5 E-35	204	81	YBR287w		hypo. prt
okam6d06r	AJ230019	280	7.4 E-35	251	61	YER172c	*BRR2*	RNA helicase-related prt
am2b06d	AJ229422	253	1.4 E-34	198	76	YLR303w	*MET25*	O-acetylhomoserine sulfhydrylase
okam4a11d	AJ229743	359	3.5 E-34	174	59	YKL205w	*LOS1*	pre-tRNA splicing prt
okam6d10r	AJ230023	255	5.2 E-34	249	60	YGL195w	*GCN1*	translational activator
okam3g03r	AJ229710	466	6.2 E-34	72	58	YNL243w	*SLA2*	cytoskeleton assembly control prt
okam3d01r	AJ229659	521	7.1 E-34	216	49	YDL234c	*GYP7*	prt of unknown function
okam5g05d	AJ229936	222	9.1 E-34	141	92	YOR116c	*RPO31*	DNA-directed RNA pol. III, 160 KD subunit
okam11d04d	AJ229509	262	1.2 E-33	180	73	YGR012w		simil. to E.nidulans cysteine synthase
okam3a02r	AJ229607	538	2.3 E-33	147	59	A_A110		
okam11b05d	AJ229488	215	3.0 E-33	225	65	YDR243c	*PRP28*	pre-mRNA splicing factor RNA helicase of DEAD box family
okam11d02d	AJ229507	237	4.2 E-33	234	56	YPL259c	*APM1*	clathrin-associated prt
okam1e04d	AJ229578	419	1.4 E-32	207	46	YIL002c	*SJH1*	synaptojanin homolog 1
okam4a05r	AJ229739	336	2.3 E-32	231	64	YIR029w	*DAL2*	allantoinase
okam5a02d	AJ229833	405	2.5 E-32	375	68	YPL008w	*CHL1*	prt of the DEAH box family
okam6d08r	AJ230022	234	4.5 E-32	234	63	YFR010w		simil. to C.elegans tRNA-guanine transglycosylase
okam3c08r	AJ229650	349	5.0 E-32	252	63	YNL308c		simil. to S.pombe and C.elegans hypo. prts
okam6c04d	AJ230000	201	8.7 E-32	198	70	YGR047c	*TFC4*	TFIIIC (transcription initiation factor) subunit, 131 KD
okam6a11d	AJ229977	270	1.1 E-31	267	55	YNL279w		hypo. prt
okam6d04d	AJ230014	218	1.1 E-31	135	76	YPR159w	*KRE6*	glucan synthase subunit
okam3a07r	AJ229614	538	4.0 E-31	108	72	YKR076w		strong simil. to YMR251w and YGR154c
okam3d03r	AJ229663	579	7.5 E-31	171	72	YLR218c		hypo. prt

Table 2. (A) (*continued*)

K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession	Size	BLASTX	Alignment	%	Systematic	Gene	Brief identification
	n°	sequence	Smallest Sum	length	identity	nomenclature	name	
		(nt)	Probability	(nt)	(aa)			
			P(N)					
okam6b08d	AJ229989	284	8.3 E-31	288	52	YGR111w		weak simil. to mosquito carboxylesterase
okam2a12d	AJ229603	200	2.8 E-30	195	72	YOR095c	*RKI1*	D-ribose-5-phosphate ketol-isomerase
okam5a06d	AJ229839	393	9.8 E-30	180	52	YHR023w	*MYO1*	myosin-1 isoform (type II myosin) heavy chain
okam1d08r	AJ229573	366	1.6 E-29	324	43	YDL028c	*MPS1*	serine/threonine/tyrosine prt kinase
okam3b03d	AJ229627	411	1.8 E-29	258	47	YOR228c		weak simil. to YNR013c
okam1b07d	AJ229561	326	2.1 E-29	225	53	YHR063c		weak simil. to translational activator CBS2
okam3a04r	AJ229610	412	2.4 E-29	180	67	YCR019w	*MAK32*	necessary for struct. stability of L-A dsRNA-cont. particles
okam3h08d	AJ229730	300	4.2 E-29	150	72	YHL019c	*APM2*	involved in clathrin-dependent transport processes
okam5a02r	AJ229834	421	5.6 E-29	155	73	YPL008w	*CHL1*	prt of the DEAH box family
okam5b06r	AJ229864	249	1.5 E-28	99	88	YNL237w	*YTP1*	weak simil. to mitochondrial electron transport prts
okam6a07r	AJ229972	262	2.1 E-28	234	55	YKL221w		weak simil. to human X-linked PEST-containing transporter
okam6b02r	AJ229982	373	2.7 E-28	42	71	YLR151c		hypo. prt
okam6b09d	AJ229991	277	3.3 E-28	267	46	YLR455w		weak simil. to human G/T mismatch binding prt
am2g11r	AJ229462	284	7.6 E-28	186	65	YLR304c	*ACO1*	aconitate hydratase
okam6c05r	AJ230002	321	8.5 E-28	75	56	YJL056c		simil. to developmental control zinc finger prts
okam11b12d	AJ229494	173	1.8 E-27	168	80	YPR145w	*ASN1*	asparagine synthetase
okam5g11d	AJ229946	299	2.0 E-27	120	74	YPL138c		weak simil. to fruit fly polycomblike nuclear prt
okam5b05r	AJ229862	328	4.4 E-27	171	56	YJR013w		simil. to C.elegans B0491.1 prt
okam5d08d	AJ229892	269	1.1 E-26	192	56	YHR078w		hypo. prt
okam4f04r	AJ229794	282	2.1 E-26	144	50	YGL036w	*MTC2*	Mtf1 Two hybrid Clone 2
am2g02r	AJ229456	317	2.5 E-26	105	80	YNL267w	*PIK1*	phosphatidylinositol 4-kinase
okam11b06d	AJ229489	245	9.4 E-26	297	42	YNL025c	*SSN8*	DNA-directed RNA pol. II holoenzyme
okam6c12r	AJ230011	174	1.1 E-25	171	61	YDR408c	*ADE8*	phosphoribosylglycinamide formyltransferase (GART)
okam5g11r	AJ229947	386	2.1 E-25	318	47	YOR229w	*WTM2*	transcriptional modulator
am1f01d#	AJ229391	150	2.4 E-25	147	76	YBR011c	*IPP1*	inorganic pyrophosphatase, cytoplasmic
okam3c08d	AJ229649	428	3.0 E-25	213	54	YNL312w	*RFA2*	DNA replication factor A, 36 kDa subunit
okam11a11d	AJ229482	262	3.4 E-25	195	55	YOL027c		simil. to YPR125w
okam1c05r	AJ229566	197	4.8 E-25	195	57	YGL246c		weak simil. to C.elegans dom-3 prt
okam3g08r	AJ229714	513	7.4 E-25	117	58	YMR156c		hypo. prt
okam3d07d	AJ229668	141	1.1 E-24	132	89	YDR148c	*KGD2*	2-oxoglutarate dehydrogenase complex E2 component
okam2a11r	AJ229602	217	2.3 E-24	192	56	YCL036w		simil. to hypo. prt YDR514c
okam4d06r	AJ229776	159	2.5 E-24	153	69	YHR197w		hypo. prt
okam5e12d	AJ229913	411	6.0 E-24	156	62	YOR258w		hypo. prt
okam11d11d	AJ229516	223	6.8 E-24	222	58	YPL110c		simil. to C.elegans hypo. prt, weak simil. to Pho81p
okam5b04d	AJ229859	305	9.6 E-24	207	42	YAL043c	*PTA1*	pre-tRNA processing prt
okam5a09r	AJ229846	340	1.2 E-23	114	55	YLR087c		hypo. prt
okam1a09d	AJ229555	282	1.3 E-23	165	60	YLL031c		simil. to hypo. prt YJL062w
okam4h12d	AJ229830	271	1.5 E-23	156	54	YLR414c		weak simil. to YLR413w
am2f12d	AJ229454	186	2.6 E-23	183	59	YGR270w	*YTA7*	26S proteasome subunit
okam11c05d	AJ229498	199	3.2 E-23	141	70	YLR168c	*MSF1*	probably involved in intramitochondrial prt sorting
okam3f03d	AJ229691	498	3.5 E-23	228	50	YGR117c		hypo. prt
okam5h06d	AJ229958	124	8.7 E-23	120	93	YGR083c	*GCD2*	translation initiation factor eIF2B, 71 kDa (delta) subunit
am2d06d	AJ229437	223	1.1 E-22	225	53	YIL126w	*STH1*	subunit of the RSC complex
okam6b08r	AJ229990	259	8.6 E-22	249	40	YPR122w	*AXL1*	protease
okam6a06r	AJ229970	254	1.6 E-21	102	74	YML104c	*MDM1*	intermediate filament prt
okam3a11r	AJ229620	432	2.1 E-21	98	72	YMR306w		simil. to 1,3-beta-glucan synthases
okam6b01d	AJ229979	273	2.1 E-21	192	52	YPL162c		hypo. prt
am1f02d	AJ229392	127	2.3 E-21	122	85	YHR190w	*ERG9*	farnesyl-diphosphate farnesyltransferase
okam11c09d	AJ229502	173	2.5 E-21	153	65	YJL112w		simil. to Met30p and N.crassa sulfur controller-2
okam3b11r	AJ229641	410	2.6 E-21	111	68	YJL095w	*BCK1*	ser/thr prt kinase of the MEKK family
okam5d07d	AJ229890	238	2.8 E-21	111	84	YPR119w	*CLB2*	cyclin, G2/M-specific
okam6c02d	AJ229998	144	2.8 E-21	114	79	YMR266w		simil. to A.thaliana hyp1 prt
okam11d09d	AJ229514	209	3.0 E-21	203	52	YKL221w		weak simil. to human X-linked PEST-containing transporter
okam11e09r	AJ229529	200	3.6 E-21	186	52	YJL080c	*SCP160*	histone-like prt
okam1c03d	AJ229563	408	8.7 E-21	108	67	YJR033c		hypo. prt
okam11a10d	AJ229481	205	1.1 E-20	126	69	YMR146c	*TIF34*	translation initiation factor eIF3, P39 subunit
okam3b01r#	AJ229624	351	1.8 E-20	111	78	YLR081w	*GAL2*	galactose (and glucose) permease
okam4g04r	AJ229814	139	6.5 E-20	111	78	YPR084w		hypo. prt
okam4d04d	AJ229771	181	9.2 E-20	183	73	YLR340w	*RPL10E*	acidic ribosomal prt L10.e
okam5a01r	AJ229832	199	1.3 E-19	98	49	YNL247w		simil. to cysteinyl-tRNA synthetases
okam5b06d	AJ229863	247	1.4 E-19	147	49	YNL236w	*SIN4*	global regulator prt
am2a08d	AJ229417	149	1.6 E-19	147	76	YGL206c	*CHC1*	clathrin heavy chain
okam6d07r	AJ230020	242	3.7 E-19	207	44	YMR176w		hypo. prt
okam4f12d	AJ229805	192	4.3 E-19	140	68	YJR109c	*CPA2*	arginine-specific carbamoylphosphate synthase, large chain
okam4g11d	AJ229821	178	1.4 E-18	177	48	YIL137c		simil. to M.musculus aminopeptidase
okam3c11d	AJ229654	400	3.4 E-18	210	36	YDL099w		weak simil. to myosin heavy chain prts
okam3f10d	AJ229702	512	4.2 E-18	93	65	YIR003w		weak simil. to mammalian neurofilament triplet H prts

Table 2. (A) (*continued*)

K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession n°	Size sequence (nt)	BLASTX Smallest Sum Probability P(N)	Alignment length (nt)	% identity (aa)	Systematic nomenclature	Gene name	Brief identification
okam6c10d	AJ230009	202	5.3 E-18	197	48	YLR389c	*STE23*	protease involved in a-factor processing
okam1a07r	AJ229553	103	1.3 E-17	99	85	YDR091c		strong simil. to hum. RNase L inhibitor and M.jan. ABC transp. pr
okam5d11r	AJ229898	178	7.4 E-17	90	67	YML006c		hypo. prt
okam3b07r	AJ229635	417	2.0 E-16	174	57	YFL009w	*CDC4*	cell division control prt
okam3c12d	AJ229656	398	2.2 E-16	117	74	YLR200w	*YKE2*	strong simil. to mouse KE2 prt
okam3f11r	AJ229705	436	2.3 E-16	123	68	YJR122w	*CAF17*	CCR4 associated factor
okam2a06r	AJ229596	388	2.6 E-16	126	60	YKL212w	*SAC1*	recessive suppressor of secretory defect
okam11b11d	AJ229493	151	2.8 E-16	135	56	YIL129c		hypo. prt
okam5d10d	AJ229895	221	4.5 E-16	87	69	YHR172w	*SPC97*	spindle pole body component
okam5f01d	AJ229914	273	5.7 E-16	171	52	YDL057w		hypo. prt
okam11f10d	AJ229540	215	7.7 E-16	99	73	YKL171w		ser/thr prt kinase
am1f09d	AJ229396	237	1.3 E-15	99	82	YHR025w	*THR1*	homoserine kinase
okam4g02d	AJ229809	204	1.7 E-15	198	46	YDL080c	*THI3*	positive regulation factor of thiamin metabolism
okam4h02d	AJ229823	172	2.8 E-15	111	63	YLR147c	*SMD3*	strong simil. to small nuclear ribonucleoprt D3
okam4d03d	AJ229769	299	3.0 E-15	141	60	YGR075c	*PRP38*	pre-mRNA splicing factor
okam5b10d	AJ229869	471	4.6 E-15	144	50	YOR298w		hypo. prt
okam2a11d	AJ229601	257	4.9 E-15	126	64	YCL035c		strong simil. to glutaredoxin
okam5c02r	AJ229872	304	5.5 E-15	180	55	YML076c		weak simil. to transcription factor
okam11a12d	AJ229483	237	5.6 E-15	84	53	YFR019w	*FAB1*	probable PIP 5-kinase
am1e05r	AJ229385	220	6.6 E-15	114	66	YMR287c	*MSU1*	3'-5' exonuclease for RNA 3' ss-tail, mitochondrial
okam3g12r	AJ229718	191	8.9 E-15	120	50	YOR374w		strong simil. to aldehyde dehydrogenase
okam3f12d	AJ229706	200	1.0 E-14	123	59	YBR238c		strong simil. to general chromatin factor Spt16p
okam3c09d	AJ229651	123	1.2 E-14	78	85	YLR249w	*YEF3*	translation elongation factor eEF3
okam4h05d	AJ229825	293	2.1 E-14	159	54	YKL068w	*NUP100*	nuclear pore prt
okam5g10r	AJ229945	384	3.3 E-14	183	57	YML011c		hypo. prt
okam3a12d	AJ229621	148	4.9 E-14	102	74	YGL010w		hypo. prt
okam11a02d	AJ229476	187	9.4 E-14	195	48	YPL109c		weak simil. to hypo. prt YLR253w
okam6b11r	AJ229994	264	1.0 E-13	144	50	YIL130w		simil. to Put3p and to hypo. prt YJL206c
okam1g02d	AJ229585	286	1.1 E-13	99	52	YIL154c	*IMP2*	sugar utilization regulatory prt
okam4d10d	AJ229778	234	1.1 E-13	180	38	YLR181c		hypo. prt
okam11e03d	AJ229520	234	1.4 E-13	102	53	YLR084c		hypo. prt
okam4f04d	AJ229793	251	1.5 E-13	165	53	YLL019c	*KNS1*	ser/thr prt kinase
okam6d06d	AJ230018	127	2.1 E-13	378	57	YER172c	*BRR2*	RNA helicase-related prt
okam1e01r	AJ229576	300	2.7 E-13	168	48	YDR476c		hypo. prt
okam3b04r	AJ229629	546	9.6 E-13	105	60	YJR046w		weak simil. to Xenopus vimentin 4
okam1c06r	AJ229568	151	1.0 E-12	90	60	YBL014c	*RRN6*	RNA pol. I specific transcription initiation factor
okam5h01d	AJ229950	257	1.1 E-12	45	60	YER093c		weak simil. to S.epidermidis PepB prt
okam3c07r	AJ229648	115	1.4 E-12	108	64	YBR187w		simil. to mouse putative transmembrane prt FT27
okam5d02d	AJ229884	329	4.7 E-12	95	59	YJL083w		simil. to hypo. prt YKR019c
am2d11d	AJ229441	156	6.0 E-12	105	54	YLR430w	*SEN1*	positive effector of tRNA-splicing endonuclease
okam11e01r	AJ229519	246	7.4 E-12	108	56	YLL034c		simil. to mammalian valosin
am1h04d	AJ229410	201	7.5 E-12	105	67	YMR089c	*YTA12*	protease of the SEC18/CDC48/PAS1 family of ATPases (AAA)
okam5e08r	AJ229909	140	1.0 E-11	123	59	YOR260w	*GCD1*	translation initiation factor eIF2bgamma subunit
okam11c11d	AJ229504	245	1.2 E-11	129	42	YIL030c	*SSM4*	involved in mRNA turnover
okam4g08d	AJ229818	212	1.3 E-11	183	54	YIR007w		hypo. prt
am2e04d	AJ229444	202	1.4 E-11	60	80	YOL107w		weak simil. to human PL6 prt
okam4b11d	AJ229756	234	2.0 E-11	171	42	YNL262w	*POL2*	DNA-directed DNA pol. epsilon, catalytic subunit A
okam5b04r	AJ229860	330	3.0 E-11	135	51	YOR359w		hypo. prt
okam5b05d	AJ229861	246	4.1 E-11	105	63	YGR196c		weak simil. to Tetrahymena acidic repetitive prt arp1
am2c06d	AJ229429	126	5.5 E-11	123	49	YOR160w	*MTR10*	involved in mRNA transport
okam11g09d	AJ229546	218	9.8 E-11	90	80	YLR162w		hypo. prt
okam3f07r	AJ229698	374	1.2 E-10	111	37	YPL105c		simil. to Smy2p
okam1a06r	AJ229552	444	1.3 E-10	171	50	YDL153c	*SAS10*	involved in silencing
okam11e01d	AJ229518	332	1.4 E-10	98	57	YLL035w		hypo. prt
am2h05r	AJ229466	214	2.3 E-10	93	55	YFL024c		weak simil. to YMR164c and Gal11p
am2c07d#	AJ229430	134	2.4 E-10	108	64	YDR007w	*TRP1*	phosphoribosylanthranilate isomerase
am2b11d	AJ229426	113	2.7 E-10	108	67	YJR062c	*NTA1*	amino-terminal amidase
okam3c12r	AJ229657	342	2.9 E-10	93	68	YDL168w	*SFA1*	long-chain alcohol dehydrogenase
okam11e10r	AJ229531	154	5.1 E-10	141	49	YBR060c	*RRR1*	origin recognition complex, 72 kDa subunit
okam3e03d	AJ229677	109	6.8 E-10	66	64	YKL213c	*DOA1*	involved in ubiquitin-dependent proteolysis
okam3b06r	AJ229633	449	1.4 E-09	139	45	YPL112c		weak simil. to YOR193w
okam1g08d	AJ229588	363	1.6 E-09	110	45	YJR062c	*NTA1*	amino-terminal amidase
okam3e12d	AJ229686	320	1.7 E-09	116	46	YJL066c		hypo. prt
okam6c09r	AJ230008	259	2.3 E-09	105	49	YGR040w	*KSS1*	ser/thr prt kinase of the MAP kinase family
am2f10d	AJ229453	151	2.7 E-09	41	78	YBR105c		simil. to hypo. prt YGR066c
okam4b09d	AJ229755	213	2.7 E-09	78	73	YML119w		hypo. prt
okam6c11r	AJ230010	294	2.8 E-09	158	41	YLR423c		hypo. prt

Table 2. (A) (*continued*)

K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession n°	Size sequence (nt)	BLASTX Smallest Sum Probability P(N)	Alignment length (nt)	% identity (aa)	Systematic nomenclature	Gene name	Brief identification
okam11c10d	AJ229503	222	4.7 E-09	117	53	YOR346w	*REV1*	DNA repair prt
okam3c01d	AJ229644	234	5.2 E-09	105	51	YDR160w		simil. to amino acid permeases Lyp1p and Dip5p
am1h09r	AJ229413	221	7.9 E-09	93	58	YGR044c	*RME1*	zinc finger transcription factor
okam3d02d	AJ229660	142	8.1 E-09	78	77	YPL082c	*MOT1*	transcriptional accessory prt
okam5a10r	AJ229848	244	8.5 E-09	66	73	YNL294c		hypo. prt
okam3f09d	AJ229701	423	9.0 E-09	69	87	YBR103w		weak simil. to Dip2p, Pwp2p and Msi1p
okam6c07r	AJ230004	333	2.4 E-08	219	33	YLR119w	*SRN2*	suppressor of rna1-1 mutation
am2d09d	AJ229440	358	2.9 E-08	47	62	YPR049c		simil. to Uso1p
okam5f02d	AJ229915	222	5.3 E-08	171	42	YGR222w	*PET54*	splicing prt and translational activator, mitochondrial
okam0b08r	AJ229474	125	5.6 E-08	90	60	YDR389w	*SAC7*	suppressor of actin mutation
okam4f07d	AJ229799	312	5.9 E-08	267	85	YOR296w		hypo. prt
okam3h04d	AJ229724	310	7.9 E-08	165	32	YLR039c	*RIC1*	involved in transcription of ribosomal prts and ribosomal RNA
okam3d03d	AJ229662	183	8.2 E-08	108	56	YDL195w	*SEC31*	component of the COPII coat of ER-golgi vesicles
okam11e05r	AJ229522	149	9.0 E-08	114	45	YNL172w	*APC1*	subunit of anaphase-promoting complex (cyclosome)
okam3g11r	AJ229717	465	1.1 E-07	96	66	YDR445c		questionable ORF
okam3d08d	AJ229670	121	1.7 E-07	23	100	YLL013c		simil. to Drosophila pumilio prt
okam5g12r	AJ229949	275	1.7 E-07	90	57	YMR027w		hypo. prt
okam4b06d	AJ229751	221	2.4 E-07	78	42	YLR263w	*RED1*	meiosis-specific prt
okam3a04d	AJ229609	393	4.3 E-07	183	65	YEL013w		simil. to intracellular attachement prts
am2h09r	AJ229469	184	4.5 E-07	47	93	YEL055c	*POL5*	DNA pol. V
am2e07d	AJ229446	119	6.7 E-07	87	48	YLR292c	*SEC72*	ER prt-translocation complex subunit
am2g09r	AJ229461	125	6.9 E-07	84	50	YGR098c	*ESP1*	required for normal spindle structure
okam6d04r	AJ230015	253	1.2 E-06	84	64	YDR464w	*SPP41*	negative regulator of PRP3 and PRP4 gene expression
okam11b09d	AJ229491	275	1.6 E-06	95	31	YML104c	*MDM1*	intermediate filament prt
am1e10r	AJ229388	138	1.7 E-06	92	45	YDL153c	*SAS10*	involved in silencing
okam1d03r	AJ229569	494	7.2 E-06	158	35	YOR371c		simil. to YAL056w
okam5a08r	AJ229844	293	8.4 E-06	53	72	YDR541c		simil. to dihydroflavonol-4-reductases
okam11c02d	AJ229496	210	1.0 E-05	96	59	YPL060w		strong simil. to Mrs2p
okam3h08r	AJ229731	211	1.4 E-05	53	94	YKL040c		weak simil. to nitrogen fixation prt nifU
okam5d06r	AJ229889	220	1.9 E-05	66	63	YMR212c		weak simil. to myosins
okam5f10r	AJ229928	140	3.4 E-05	38	61	YLR305c	*STT4*	phosphatidylinositol-4-kinase
okam5d11d	AJ229897	161	5.5 E-05	63	66	YDR421w		hypo. prt
okam5h02d	AJ229952	192	6.3 E-05	87	38	YNR035c		hypo. prt
okam4a03r	AJ229738	132	6.8 E-05	38	61	YDR141c		hypo. prt

PRP5), nucleic acid synthesis (URA1) or amino acid metabolism (TRP5, ARG1, MET6 and LYS2). Other *K.lactis* genes identified here (such as IMP2, a putative sugar utilization regulatory protein) are novel, although belonging to a functional category well studied in this organism (45). We also discovered a *K.lactis* homolog to the alcohol dehydrogenase gene ADH4 of *S.cerevisiae* (46), bringing the total number of alcohol dehydrogenases in *K.lactis* to five, instead of four as previously thought (47). Four other RSTs were found corresponding to unknown *K.lactis* genes encoding ribosomal proteins. Two of them (am2c02d and okam3f03r) overlap *S.cerevisiae* genes containing introns. The nucleotide sequence of am2c02d corresponds to that of the *S.cerevisiae* gene YDL130w without the presence of an intron. The nucleotide sequence of okam3f02r corresponds to the 3′ parts of the *S.cerevisiae* genes YGR118w and YPR132w, and shows a high level of conservation (89 and 91% identity, respectively) with the exons of these two genes but not with the introns. However, the existence of the 5′ and 3′ splice sites and of the branchpoint (48,49) in the am2c02d sequence indicates the presence of an intron in the *K.lactis* gene at a conserved position relative to the *S.cerevisiae* gene.

New *K.lactis* genes absent in *S.cerevisiae*

As expected, for each *K.lactis* translation product having homologs in several organisms, the closest sequence similarity is generally observed with *S.cerevisiae*, but in a few specific cases the situation is reversed or no homolog at all is found in *S.cerevisiae*. A notable example of the latter case is the β-galactosidase gene of *K.lactis* (50) which is absent in *S.cerevisiae*.

Five *K.lactis* RST translation products have homologs in other species but not in *S.cerevisiae* (Table 3). One such case concerns a protein of the yeast *C.albicans*. The other examples are more surprising because the *K.lactis* products show a high degree of sequence conservation with distantly related species (*S.pombe*, *Aspergillus nidulans* and *Cuphea lanceolata*) but not with the *S.cerevisiae* counterparts.

'Universal', 'yeast-specific' and 'species-specific' genes

Systematic sequence comparisons of *S.cerevisiae* with all other organisms presently sequenced (Bacteria, Archea or Eucaryotes) indicate that 64% of its predicted proteins have a homolog in at least one of the other species while the remaining 36% do not. While the first category may be regarded as 'universal genes' because they are present in species of different domains of life, the latter may be hypothesized to be specific *S.cerevisiae* genes or, alternatively, be limited to its phylogenetic group, the hemiascomycetes. The present sequences of *K.lactis* genes offer a means to differentiate between these two possibilities. A total of 85 novel genes of *K.lactis* (30% of the identified RSTs) have homologs in *S.cerevisiae* that themselves had no other homolog before. Such

Table 2. (B) *K.lactis* RSTs having several *S.cerevisiae* homologs

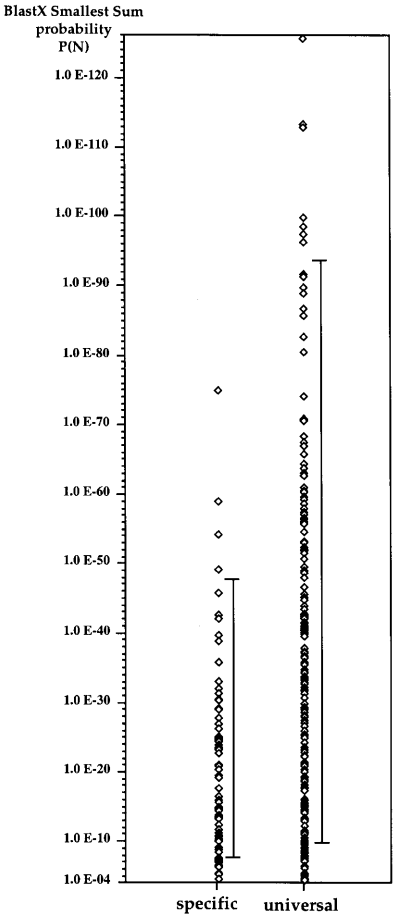
K. lactis RST			*Kl* and *Sc* comparison			*S.cerevisiae*		
Nomenclature	Accession n°	Size sequence (nt)	BLASTX Smallest Sum Probability P(N)	Alignment length (nt)	% identity (aa)	Systematic nomenclature	Gene name	Brief identification
okam3e05r	AJ229679	466	8.1 E-78	312	82	YMR199w	*CLN1*	cyclin, G1/S-specific
			1.0 E-78	315	79	YPL256c	*CLN2*	cyclin, G1/S-specific
okam3f03r	AJ229691	403	1.4 E-66	279	94	YGR118w	*RPS28A*	ribosomal prt S23.e
			1.4 E-66	279	94	YPR132w	*RPS28B*	ribosomal prt S23.e
okam4g06d	AJ229815	255	2.5 E-61	252	89	YOR096w	*RP30*	ribosomal prt
			1.8 E-59	252	89	YNL096c		strong simil. to ribosomal prt S7
okam6d01r	AJ230013	291	5.4 E-52	282	76	YPL089c	*RLM1*	transcription factor of the MADS box family
			6.8 E-51	282	76	YBR182c	*SMP1*	MADS-box transcription factor
okam4d10r	AJ229779	195	4.9 E-44	195	89	YDR502c	*SAM2*	S-adenosylmethionine synthetase 2
			6.8 E-44	195	89	YLR180w	*SAM1*	S-adenosylmethionine synthetase 1
okam5d04d	AJ229886	225	1.6 E-41	222	77	YOL059w	*GPD3*	glycerol-3-phosphate dehydrogenase (NAD+), mitochondrial
			7.9 E-40	222	75	YDL022w	*GPD1*	glycerol-3-phosphate dehydrogenase (NAD+), cytoplasmic
okam5g04r	AJ229935	142	8.8 E-32	140	91	YNR016c	*ACC1*	acetyl-CoA carboxylase
			5.6 E-30	140	85	YMR207c	*HFA1*	simil. to acetyl-CoA carboxylase
okam11d10d	AJ229515	213	1.9 E-28	173	67	YBL017c	*PEP1*	vacuolar prt sorting/targeting prt
			3.5 E-26	173	63	YIL173w		strong simil. to Pep1p
			3.5 E-26	173	63	YJL222w		strong simil. to Pep1p
			1.3 E-25	173	63	YNR065c		strong simil. to YJL222w, YIL173w and Pep1p
okam3e12r	AJ229687	491	1.1 E-26	275	36	YOL153c		strong simil. to Cps1p, two in-frame stop codons
			4.5 E-25	275	44	YJL172w	*CPS1*	Gly-X carboxypeptidase YSCS precursor
am2c10d	AJ229433	122	7.3 E-14	122	63	YKL129c	*MYO3*	myosin type I
			1.3 E-13	111	70	YMR109w	*MYO5*	myosin I
okam5a11d	AJ229849	223	1.1 E-09	110	62	YDR134c		strong simil. to Flo1p, Flo5p, Flo9p and YLR110c
			1.3 E-08	110	54	YLR110c		simil. to Flo1p
okam5c11r#	AJ229881	190	9.9 E-06	47	68	YBR019c	*GAL10*	UDP-glucose 4-epimerase
			1.0 E-05	47	68	YNR071c		strong simil. to UDP-glucose 4-epimerase Gal10p

# indicates RSTs corresponding to previously known *K.lactis* genes: *IPP1* (am1f01d), *GAL2* (okam3b01r), *TRP1* (am2c07d) and *GAL10* (okam5c11r), corresponding to EMBL accession nos X14230, X53752, X14230 and X07039, respectively.

Table 3. List of the *K.lactis* RSTs closer to proteins of other species than to *S.cerevisiae*

K. lactis RST			*K. lactis* and other genome comparison		
Nomenclature	Accession n°	Size sequence (nuc.)	BLASTX Smallest Sum Probability P(N)	Brief identification	Species
am1a04d	AJ229366	409	3.0 E-41	D-arabinitol dehydrogenase (EC 1.1.1-)	*Candida albicans*
			3.4 E-41	D-arabinitol dehydrogenase (EC 1.1.1-)	*Candida tropicalis*
			5.8 E-41	D-arabinitol dehydrogenase (EC 1.1.1-)	*Pichia stipitis*
am2f05d	AJ229451	261	6.0 E-18	Putative agmatinase precursor (EC 3.5.3.11)	*Schizosaccharomyces pombe*
			2.0 E-08	Possible agmatinase (EC 3.5.3.11)	*Streptomyces clavuligerus*
			4.9 E-06	Hypothetical agmatinase (EC 3.5.3.11)	*Methanothermus fervidus*
okam5d07r	AJ229891	230	0.1 E-10	Beta-ketoacyl-ACP reductase (EC 1.1.1.100)	*Cuphea lanceolata*
			0.2 E-10	Hypothetical protein 5	*Xanthobacter sp.*
			6.3 E-10	Hypothetical oxidoreductase	*Bacillus subtilis*
am2d01d	AJ229435	151	2.9 E-07	YOL5_CAEEL hypothetical 45.3 KD protein	*Cænorhabditis elegans*
okam1d10r	AJ229574	268	0.2 E-06	Acetamidase (EC 3.5.1.4)	*Aspergillus nidulans*

SWISS-PROT (release 34) (44) and laboratory database were used for comparison.
The latter is composed of 46 630 translation products of complete sequenced genomes of 10 bacteria. (8,9,11,12,14,16,18–21), three archaea (10,15,17) and 81% of the *C.elegans* genome (http://www.sanger.ac.uk/Projects/C_elegans/).



class	yeast-specific protein	universal protein
n	85	212
median value	2.0 10^{-16}	2.0 10^{-33}

Figure 3. Range of BLASTX *P*-values among yeast-specific and universal genes. Best *P*-values between each *K.lactis* RST translation product and the complete set of *S.cerevisiae* proteins were listed separately in two columns: one for yeast-specific genes (yeast-specific column) and the other for functionally characterized genes or homologs to functionally characterized proteins of other organisms (universal column). Bars indicate 95% confidence limits.

genes should now be regarded as 'yeast-specific' rather than *S.cerevisiae*-specific (Discussion). Closer examination of the distribution of sequence similarities between the universal genes and the yeast-specific genes reveals an additional interesting observation (Fig. 3). Whereas *P*-value scores between *K.lactis* and *S.cerevisiae* gene products range from 6.8×10^{-5} (close to our selected limit) to 1.4×10^{-126} for genes of *S.cerevisiae* having homologs in other organisms, the same distribution only reaches 1.0×10^{-75} for genes of *S.cerevisiae* devoid of previous homologs (the median values of the two distributions are 2.0×10^{-33} and 2.0×10^{-16}, respectively). In other words, the subclass of *S.cerevisiae* genes that were previously without structural homologs show greater sequence divergence than average when compared with their *K.lactis* counterparts, suggesting that part of the yeast-specific genes correspond to sequences that diverge more rapidly than others.

Conserved gene order relationships (synteny)

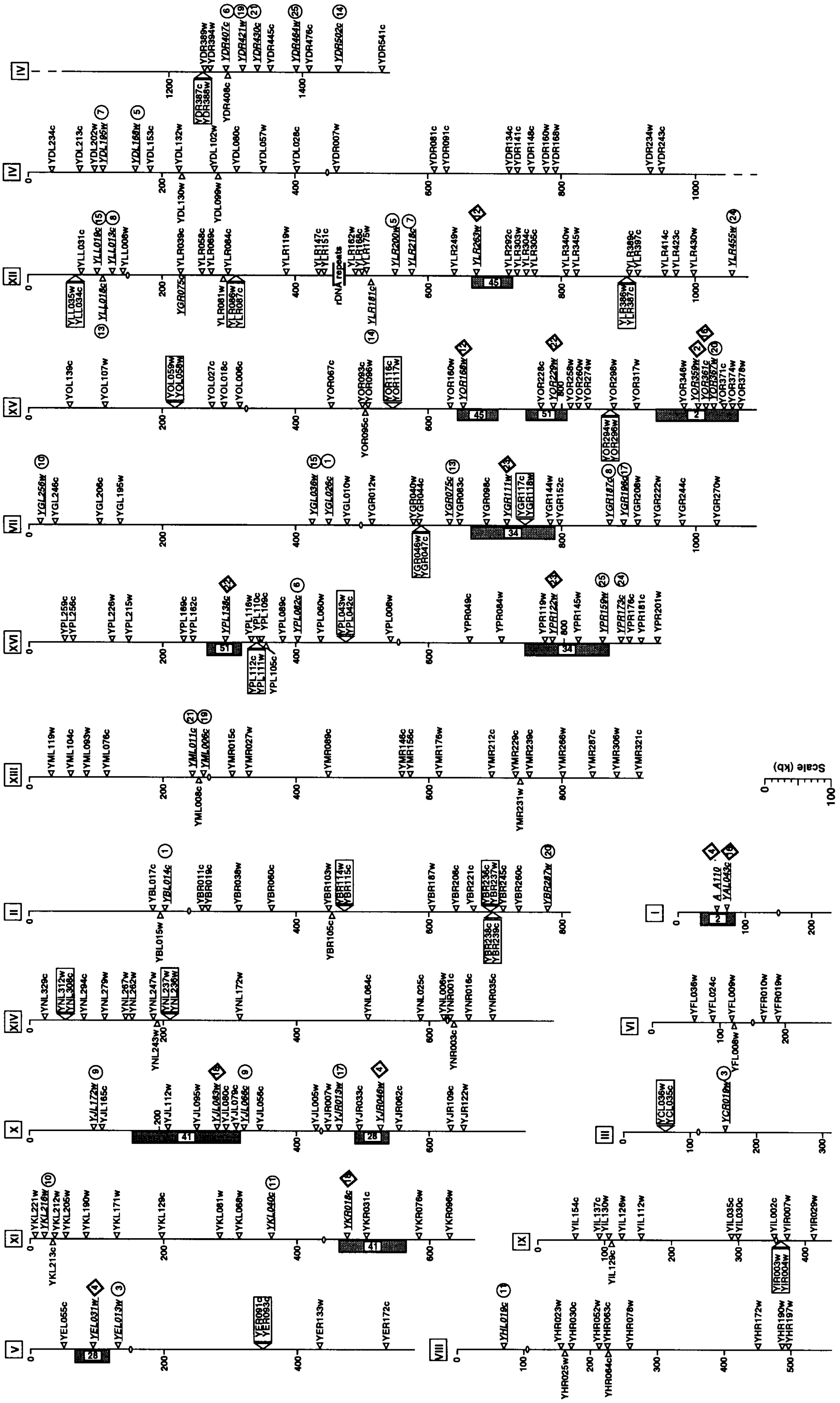
As expected, the *S.cerevisiae* homologs to the *K.lactis* genes identified in this work are scattered throughout the 16 chomosomes (Fig. 4). Sequencing the two ends of inserts from clones of the long-fragment library allowed us to identify 45 cases in which two distinct neighbouring genes were present on the same insert (90 genes in total), and hence to examine the synteny between the two genomes. We found that 42 genes have the same neighbour in the two species (24 pairs) whereas 48 genes do not. In one case (the gene pair Kl-YNL308c and Kl-YNL312w), a short local inversion explains the difference between the two species, bringing the general synteny to ~50% of the cases examined (44/90). In all other cases, the two genes that are neighbours in *K.lactis* are separated in *S.cerevisiae*. Interestingly, in seven cases the two neighbouring genes in *K.lactis* have two homologs, each located in one of the two regions believed to represent an ancient chromosomal duplication in *S.cerevisiae* (26,51).

DISCUSSION

In this work, a random collection of *K.lactis* short genomic sequences was constituted to enable comparisons between the genome of this yeast and that of *S.cerevisiae*. Beside the discovery and identification of novel *K.lactis* genes, this work was primarily aimed at examining the conservation, loss or divergence of the various classes of *S.cerevisiae* genes, and in particular the most intriguing one, the orphans. It was, therefore, essential that the collection of sequences determined represent a random sample of the *K.lactis* genome. The randomness of our libraries can be estimated from the fact that only seven RSTs were found to partially overlap one another, a figure which is almost exactly as expected considering the fact that the 658 RSTs cover 1.3% of the *K.lactis* genome. Randomness of our *K.lactis* RSTs is also supported by the scattered distribution of the homologs in the *S.cerevisiae* map (Fig. 4).

Exploration of a novel genome by RSTs is a quick and efficient procedure, provided a closely related organism whose genome has been entirely sequenced is available to serve as a reference. At the beginning of such a project, the number of genes identified exceeds the fraction of the genome covered [here, 296 new *K.lactis* genes (5% of the genes) were identified by sequencing only 1.3% of the genome]. Assuming that gene size distribution and gene density are similar in *K.lactis* and *S.cerevisiae* (all our results support this idea), only 5–6000 RSTs of ~300 nucleotides, representing 15% of the genome, would allow the identification of nearly half of all *K.lactis* genes. Longer sequences would obviously increase this number as well as the number of sequences that are identifiable. The drawback of RSTs compared with systematic genome sequencing is, of course, that only parts of the gene sequences are available and that the quality of the sequence is that of single reads, but the information obtained can be rapidly used for functional studies by setting up hybridization matrices, for example.

Compared with the 85 genes identified previously (29), this work more than tripled the number of *K.lactis* genes identified at the molecular level. But more importantly, it shows that the previous set of *K.lactis* genes, essentially isolated by functional complementation of *S.cerevisiae* mutants or by heterologous hybridization, was partially biased in favour of the most conserved genes, while that actual sequence divergence between


IV
1200
YDR387c
YDR388w
YDR389w
YDR394w
YDR407c
6
YDR408c
YDR421w
19
YDR430c
21
YDR445c
YDR464w
25
YDR476c
1400
YDR502c
14
YDR541c
IV
0
YDL234c
YDL213c
YDL202w
YDL195w
7
YDL158w
5
YDL153c
YDL132w
200
YDL130w
YDL102w
YDL099w
YDL080c
YDL057w
400
YDL028c
YDR007w
600
YDR081c
YDR091c
YDR134c
YDR141c
YDR148c
800
YDR160w
YDR168w
1000
YDR234w
YDR243c
XII
YLL035w
YLL034c
YLL031c
YLL019c
15
YLL018c
13
YLL013c
8
YLL006w
200
YGR075c
YLR039c
YLR058c
YLR069c
YLR081w
YLR084c
YLR086w
YLR087c
400
YLR119w
YLR147c
YLR151c
rDNA repeats
YLR162w
YLR166c
YLR175w
YLR181c
14
YLR200w
5
YLR218c
7
600
YLR249w
YLR263w
45
YLR292c
YLR303w
YLR304c
YLR305c
800
YLR340w
YLR345w
YLR386w
YLR387c
YLR389c
YLR397c
YLR414c
YLR423c
1000
YLR430w
YLR455w
24
XV
YOL139c
YOL107w
YOL059w
YOL055c
200
YOL027c
YOL018c
YOL006c
400
YOR067c
YOR093c
YOR096w
YOR095c
YOR116c
YOR117w
600
YOR160w
YOR168w
45
YOR228c
YOR229w
800
YOR258w
YOR260w
YOR274w
51
YOR294w
YOR296w
YOR298w
YOR317w
YOR346w
YOR359w
YOR361c
YOR362w
YOR371c
YOR374w
YOR376w
2
VII
YGL256w
10
YGL248c
YGL206c
YGL195w
200
400
YGL036w
15
YGL026c
1
YGL010w
YGR012w
YGR040w
YGR044c
YGR046w
YGR047c
YGR075c
13
YGR083c
YGR098c
YGR111w
YGR117c
YGR118w
34
YGR144w
YGR152c
800
YGR187c
8
YGR196c
17
YGR208w
YGR222w
YGR244c
1000
YGR270w
XVI
YPL259c
YPL256c
YPL226w
YPL215w
200
YPL169c
YPL162c
YPL138c
51
YPL116w
YPL110c
YPL109c
YPL112c
YPL111w
YPL105c
YPL089c
400
YPL082c
6
YPL060w
YPL043w
YPL042c
YPL006w
600
YPR049c
YPR084w
YPR119w
YPR122w
800
YPR145w
YPR159w
25
YPR173c
24
YPR176c
YPR181c
YPR201w
34
XIII
YML119w
YML104c
YML093w
YML076c
200
YML011c
21
YML008c
19
YMR015c
YMR027w
400
YMR089c
YMR148c
YMR156c
600
YMR176w
YMR212c
YMR229c
YMR231w
YMR239c
800
YMR266w
YMR287c
YMR306w
YMR321c
Scale (kb)
0
100
II
YBL017c
YBL014c
1
YBL015w
YBR011c
YBR018c
YBR036w
YBR060c
400
YBR103w
YBR114w
YBR115c
YBR105c
600
YBR187w
YBR206c
YBR221c
YBR236c
YBR237w
YBR245c
YBR238c
YBR239c
YBR260c
800
YBR287w
20
I
A A110
YAL043c
2
100
200
XIV
YNL329c
YNL312w
YNL308c
YNL294c
YNL279w
YNL267w
YNL262w
YNL247w
YNL237w
YNL236w
YNL243w
200
YNL172w
400
YNL064c
600
YNL025c
YNL006w
YNR001c
YNR003c
YNR016c
YNR035c
VI
YFL036w
YFL024c
100
YFL009w
YFL008w
YFR010w
200
YFR019w
X
YJL172w
9
YJL165c
200
YJL112w
YJL095w
YJL083w
YJL080c
YJL079c
YJL066c
9
YJL056c
41
400
YJL005w
YJR007w
YJR013w
17
YJR033c
YJR046w
28
YJR062c
600
YJR109c
YJR122w
III
YCL036w
YCL035c
100
YCR019w
3
200
300
XI
YKL221w
YKL216w
10
YKL212w
YKL205w
YKL213c
YKL190w
YKL171w
200
YKL129c
YKL081w
YKL066w
YKL040c
11
400
YKR018c
YKR031c
41
600
YKR076w
YKR096w
IX
YIL154c
YIL137c
100
YIL130w
YIL129c
YIL126w
YIL112w
200
YIL035c
YIL030c
300
YIL002c
YIR007w
YIR003w
YIR004w
400
YIR029w
V
YEL055c
YEL031w
YEL013w
3
28
200
YER091c
YER093c
400
YER133w
YER172c
VIII
YHL019c
1
100
YHR023w
YHR030c
YHR025w
200
YHR052w
YHR063c
YHR064c
YHR078w
300
400
YHR172w
YHR190w
YHR197w
500

the two yeast species is more heterogeneous than previously thought (the mean amino acid identity with *S.cerevisiae* for the 85 previously identified *K.lactis* genes was 83.5% with a standard deviation of 19.1%, compared with a mean of 63.6% with a standard deviation of 49.7% for the present set of random sequences).

In this work, *K.lactis* genes were identified solely on the basis of sequence similarity without experimental data concerning their actual functions. Genes were therefore classified as orthologs on the basis of structural relationships. Yet the high levels of similarity observed in a number of cases makes the functional conservation between the two species a tempting hypothesis (Table 2). In other cases, however, prediction of function is less reliable, especially when the function of the *S.cerevisiae* homolog is itself only tentative.

The very existence of *K.lactis* homologs to *S.cerevisiae* orphan genes confirms that such genes are actual functional genes that have previously remained without homologs because no other yeast sequence was available. However, their under-representation in our RSTs (85 were found when 107 were predicted) suggests that a fraction of the orphan *S.cerevisiae* genes will remain orphans even if the genome of *K.lactis* was entirely sequenced. This may be partly due to the fact that the degree of sequence conservation between translated products of yeast-specific genes is generally lower than the average sequence conservation (Fig. 3), hence lowering the number of recognizable matches in this category. Using the same rationale and the same limit of significance as for *K.lactis* (Materials and Methods), we found similar results for the recently released set of ~25 000 genomic short sequences of the pathogenic yeast *C.albicans* (http://www.candida.stanford.edu). In this yeast, believed to be less closely related to *S.cerevisiae* than *K.lactis* (28), the proportion of sequences giving a significant match with *S.cerevisiae* translation products is only of the order of 30% compared with ~55% for *K.lactis*. Yet, as for *K.lactis*, the average degree of similarity of level of *C.albicans* sequences with *S.cerevisiae* is lower for the yeast-specific genes than for universal genes. A lower sequence conservation between genes of unknown functions, as compared with the functionally assigned genes, has also been observed for the two related bacterial species *Mycoplasma pneunoniae* and *Mycoplasma genitalium* (22).

The fact that some gene sequences may diverge more quickly than others during evolution is not new (52), but the lower sequence conservation in genes of unknown function may indicate a lower functional constraint on them, or a higher flexibility of primary sequences to sustain function. In any case, orphan genes of *S.cerevisiae* that now have homologs in another yeast species should no longer be regarded as orphans, and their existence in *K.lactis* may help identify their function. Perhaps a number of such genes are specific to the *Saccharomyces–Kluyveromyces* lineages, or to the ascomycetous yeasts at large, or even to the fungal kingdom.

The degree of synteny between related species is a most important parameter to consider for genome evolution, as it emphasizes divergence created at the chromosome level rather than at the gene sequence level. Prior to this work, synteny between *K.lactis* and *S.cerevisiae* was estimated for only a few clusters of two and three genes (53). Nearly all of the genes located inside those clusters were found contiguous in the two species. Our present results, including 90 genes, indicate an average synteny of ~50% (Fig. 4), a figure comparable with that calculated in a recent work from the more limited number of publicly available sequences (54). Beside the practical benefit that can be derived from the consideration that the neighbours of our *K.lactis* genes can be predicted with 50% confidence from the genome of *S.cerevisiae*, the observed synteny facilitates interpretation of some aspects of the evolutionary history of the *Kluyveromyces* genus, which remains poorly characterized. The chromosome rearrangements that led to non-synteny must have occurred after the separation of the *Saccharomyces* and *Kluyveromyces* lineages or, alternatively, duplication of chromosomal domains followed by random loss of genes may have taken place. This latter hypothesis has recently been clearly stated (54). We observed cases here in which two neighbouring genes in *K.lactis* do indeed find their homologs in duplicated chromosome blocks in *S.cerevisiae*, but we also observed cases of neighbouring *K.lactis* genes having their homologs dispersed in *S.cerevisiae* and falling outside of duplicated chromosome blocks.

ACKNOWLEDGEMENTS

This work is dedicated to the memory of Angelos Kalogeropoulos. We thank Arnaud Perrin for the maintenance of a comprehensive *S.cerevisiae* database used in this work and our colleagues of the ‘Unité de Génétique Moléculaire des Levures’ for fruitful discussions. This work was supported in part by EUROFAN (BIO4-CT95-0080) from DGXII of the European Commission and by a grant (ACC-SV1) from MENESR. B.D. is Professor of Molecular Genetics at University Pierre et Marie Curie, Paris and a member of the Institut Universitaire de France.

REFERENCES

1 Lander,E.S. (1996) *Science*, **274**, 536–539.
2 Doolittle,R.F. (1998) *Nature*, **392**, 339–342.
3 Green,P., Lipman,D., Hillier,L., Waterston,R., States,D. and Claverie,J.M. (1993) *Science*, **259**, 1711–1716.
4 Adams,M.D., Kerlavage,A.R., Fleischmann,R.D., Fuldner,R.A., Bult,C.J., Lee,N.H., Kirkness,E.F., Weinstock,K.G., Gocayne,J.D., White,O. *et al.* (1995) *Nature*, **377** (Suppl.), 3–174.
5 Banfi,S., Borsani,G., Rossi,E., Bernard,L., Guffanti,A., Rubboli,F., Marchitiello,A., Giglio,S., Coluccia,E., Zollo,M. *et al.* (1996) *Nature Genet*., **13**, 167–174.
6 Makalowski,W., Zhang,J. and Boguski,M.S. (1996) *Genome Res*., **6**, 846–857.
7 Stubbs,L., Carver,E.A., Shannon,M.E., Kim,J., Geisler,J., Generoso,E.E., Stanford,B.G., Dunn,W.C., Mohrenweiser,H., Zimmermann,W. *et al.* (1996) *Genomics*, **35**, 499–508.

Figure 4. (Opposite) Map location of *S.cerevisiae* orthologs to *K.lactis* RSTs and synteny between the two yeast species. Map of the 16 *S.cerevisiae* chromosomes, classified by decreasing size order from right to left (top) and from left to right (bottom), with indication of the homologs to *K.lactis* RSTs. Boxes indicate pairs of genes that remain neighbours in the two species. Pairs of neighbouring genes in *K.lactis* and dispersed in *S.cerevisiae* (in italics and underlined) have been arbitrarily numbered. Diamonds indicate pairs corresponding to two genes which are located in duplicated chromosomal regions [numbered as by Wolfe and Shields (51); only those duplicated regions where pairs were found are indicated on the present map]. Circles indicate pairs composed of one or two genes outside of such duplications.

8 Fleischman,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlevage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.

9 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleishman,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G., Kelley,J.M. *et al.* (1995) *Science*, **270**, 397–403.

10 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleishmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. *et al.* (1996) *Science*, **273**, 1058–1073.

11 Himmelreich,H.R., Plagens,H., Hilbert,H. and Hermann,R. (1996) *Nucleic Acids Res*., **24**, 4420–4449.

12 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) *DNA Res*., **3**, 109–136.

13 Goffeau,A., Aert,R., Agostini-Carbone,M.L., Ahmed,A., Aigle,M., Alberghina,L., Albermann,K., Albers,M., Aldea,M., Alexandraki,D. *et al.* (1997) *Nature*, **387** (Suppl.), 5–105.

14 Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1462.

15 Klenk,H.-P., Clayton,R.A., Tomb,J.-F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D. *et al.* (1997) *Nature*, **390**, 364–370.

16 Tomb,J.-F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleishmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al.* (1997) *Nature*, **388**, 539–547.

17 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K. *et al.* (1997) *J. Bacteriol*., **179**, 7135–7155.

18 Kunst,F., Ogasawara,N., Moszer,I., Albertin,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) *Nature*, **390**, 249–256.

19 Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R., White,O., Ketchum,K.A., Dodson,R. *et al.* (1997) *Nature*, **390**, 580–586.

20 Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al.* (1998) *Nature*, **392**, 353–358.

21 Cole,S.T., Brosh,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al.* (1998) *Nature*, **393**, 537–544.

22 Himmelreich,H.R., Plagens,H., Hilbert,H., Beiner,B. and Hermann,R. (1997) *Nucleic Acids Res*., **25**, 701–712.

23 Goffeau,A., Barrel,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1997) *Science*, **274**, 546–567.

24 Boguski,M.S. and Schuler,G.D. (1995) *Nature Genet*., **10**, 369–371.

25 Hodgkin,J., Plasterk,R.H.A. and Waterston,R.H. (1995) *Science*, **270**, 410–414.

26 Mewes,H.W., Albermann,K., Bähr,M., Gleissner,G., Hani,J., Heumann,K., Kleine,K., Maierl,A., Oliver,G. and Zollner,A. (1997) *Nature*, **387** (Suppl.), 7–8.

27 Dujon,B. (1996) *Trends Genet*., **12**, 263–270.

28 Wilmotte,A., Van De Peer,Y., Goris,A., Chapelle,S., De Baere,R., Nelissen,B., Neefs,J.M., Hennebert,G.L. and De Wachter,R. (1993) *System. Appl. Microbiol*., **16**, 436–444.

29 Wesolowski-Louvel,M., Breunig,K.D. and Fukuhara,H. (1995) In Wolf,K. (ed.), *Non-Conventional Yeasts in Biotechnology.* Springer, Berlin, pp. 140–199.

30 Heus,J.J., Zonneveld,B.J., de Steensma,H.Y. and van den Berg,J.A. (1993) *Mol. Gen. Genet*., **236**, 355–362.

31 Yeh,P., Landais,D., Lemaitre,M., Maury,I., Crenne,J.Y., Becquart,J., Murry-Brelier,A., Boucher,F., Montay,G., Fleer,R. *et al.* (1992) *Proc. Natl Acad. Sci. USA*, **89**, 1904–1908.

32 Fleer,R., Chen,X.J., Amellal,N., Yeh,P., Fournier,A., Guinet,F., Gault,N., Faucher,D., Folliard,F., Fukuhara,H. *et al.* (1991) *Gene*, **107**, 285–295.

33 Pasteur,L. (1857) C.R. *Acad. des Sciences*, **XLV**, 913–916.

34 Altmann-Jöhl,R. and Philippsen,P. (1996) *Mol. Gen. Genet*., **250**, 69–80.

35 Louis,C., Madueno,E., Modolee,J., Mahmoud,M.O., Papagiannakis,G., Saunders,R.D.C., Savakis,C., Sidén-Kiamos,I., Spanos,L., Topalis,P. *et al.* (1997) *Gene*, **195**, 187–193.

36 Dower,W.J., Miller,J.F. and Ragsdale,C.W. (1988) *Nucleic Acids Res*., **16**, 6127–6145.

37 Ansorge,W., Voss,H., Wiemann,S., Schwager,C., Sproat,B., Zimmermann,J., Stegemann,J., Erfle,H., Hewitt,N. and Rupp,T. (1992), *Electrophoresis*, **13**, 616–619.

38 Hultman,T., Stahl,S., Hornes,E. and Uhlen,M. (1989) *Nucleic Acids Res*., **17**, 4937–4946.

39 Staden,R. (1994) In Griffin,A.M. (ed.), *Methods in Molecular Biology.* Humana Press Inc., pp. 9–170.

40 Marck,C. (1988) *Nucleic Acids Res*., **16**, 1829–1836.

41 Altschul,S.F., Gish,W., Miller,W., Myers,M. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.

42 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

43 Myers,M. and Miller,W. (1988) *CABIOS*, **4**, 11–17.

44 Bairoch,A. and Apweiler,R., (1998) *Nucleic Acids Res*., **26**, 38–42.

45 Zenke,F.T., Engles,R., Vollenbroich,V., Meyer,J., Hollenberg,C.P. and Breunig,K.D. (1996) *Science*, **272**, 1662–1665.

46 Williamson,V.M. and Paquin,C.E. (1987) *Mol. Gen. Genet*., **209**, 374–381.

47 Shain,D.H., Salvadore,C. and Denis,C.L. (1992) *Mol. Gen. Genet*., **232**, 479–488.

48 Eng,F.J. and Warner,J.R. (1991) *Cell*, **65**, 797–804.

49 Bergkamp-Steffens,G.K., Hoekstra,R. and Planta,R.J. (1992) *Yeast*, **8**, 903–922.

50 Poch,O., L'Hôte,H., Dallery,V., Debeaux,F., Fleer,R. and Sodoyer,R. (1992) *Gene*, **118**, 55–63.

51 Wolfe,K.H. and Shields,D.C. (1997) *Nature*, **387**, 708–712.

52 Doolittle,R.F. (1992) *Prot. Sci*., **1**, 191–200.

53 Mulder,W., Scholten,I.H.J.M., de Boer,R.W. and Grivell,L.A. (1994) *Mol. Gen. Genet*., **245**, 96–106.

54 Keogh,R.S., Seoighe,C. and Wolfe,K.H. (1998) *Yeast*, **14**, 443–457.