ELSEVIER

Gene 238 (1999) 253–261

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Updated map of duplicated regions in the yeast genome

Cathal Seoighe, Kenneth H. Wolfe *

Department of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland

Received 26 March 1999; received in revised form 30 June 1999; accepted 13 July 1999; Received by G. Bernardi

Abstract

We have updated the map of duplicated chromosomal segments in the *Saccharomyces cerevisiae* genome originally published by Wolfe and Shields in 1997 (Nature 387, 708–713). The new analysis is based on the more sensitive Smith–Waterman search method instead of BLAST. The parameters used to identify duplicated chromosomal regions were optimized such as to maximize the amount of the genome placed into paired regions, under the assumption that the hypothesis that the entire genome was duplicated in a single event is correct. The core of the new map, with 52 pairs of regions containing three or more duplicated genes, is largely unchanged from our original map. 39 tRNA gene pairs and one snRNA pair have been added. To find additional pairs of genes that may have been formed by whole genome duplication, we searched through the parts of the genome that are not covered by this core map, looking for putative duplicated chromosomal regions containing only two duplicate genes instead of three, or having lower-scoring gene pairs. This approach identified a further 32 candidate paired regions, bringing the total number of protein-coding genes on the duplication map to 905 (16% of the proteome). The updated map suggests that a second copy of the ribosomal DNA array has been deleted from chromosome IV. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Gene duplication; Gene order; *Kluyveromyces lactis*; Molecular evolution; Polyploidy; *Saccharomyces cerevisiae*

1. Introduction

The genome of the yeast *Saccharomyces cerevisiae* contains many large paired chromosomal regions, consisting of duplicated gene pairs arranged in the same order on two chromosomes, interspersed with many unique genes (e.g. Lalo et al., 1993; Melnick and Sherman, 1993; Goffeau et al., 1996; Coissac et al., 1997; Mewes et al., 1997; Philippsen et al., 1997; Wolfe and Shields, 1997). Our laboratory has proposed that these regions are the result of a single, ancient, duplication of the entire genome (which was subsequently fragmented by reciprocal translocations among chromosomes) rather than numerous successive independent duplication events (Wolfe and Shields, 1997). The evidence to support this interpretation is (i) that the transcriptional orientation of duplicated gene pairs in yeast is almost always the same, either towards or away from the centromere; (ii) that the large 'sister' duplicated sections of chromosome do not overlap with one another; and (iii) that gene order in the related species *Kluyveromyces lactis* is the same as what would be expected for a species that diverged from *S. cerevisiae* before genome duplication occurred in the *S. cerevisiae* lineage. These observations are not compatible with the alternative hypothesis of multiple independent duplications of sections of chromosome. Our hypothesis is also strongly supported by recent extensive gene mapping data from the ascomycete *Ashbya gossypii* (see Dietrich et al., 1999).

Our model of yeast chromosome evolution is shown explicitly in Keogh et al. (1998), and the extent of genomic rearrangement subsequent to this event was estimated by Seoighe and Wolfe (1998). In brief, we hypothesize that the entire genome was duplicated, increasing the number of genes to 200% of its original value, but then that numerous deletions of redundant duplicate copies of genes reduced this figure to 108% (i.e. $2 \times 8\%$ in pairs and 92% unique). Thus, the 'duplicated chromosomal regions' that have been described consist of duplicated genes separated by numerous unique genes that were returned to a single-copy state by the deletion of a homolog. The duplicated genes formed by the genome duplication are only a minor fraction of all the gene families in yeast, but are one of the most striking features of its genome organization.

Abbreviations: BLAST, basic local alignment search tool; SGD, Saccharomyces Genome Database; snRNA, small nuclear RNA; YPD, Yeast Protein Database.

* Corresponding author. Tel.: +353-1-608-1253; fax: +353-1-679-8558.

E-mail address: khwolfe@tcd.ie (K.H. Wolfe)

0378-1119/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0378-1119(99)00319-4

In principle, under the genome duplication/reciprocal translocation hypothesis, each point on every yeast chromosome should have a 'sister' point elsewhere in the genome. However, the low proportion of retained duplicated genes, as well as approximately 10^8 years of sequence divergence since duplication, means that it is impossible to assign the whole of the genome into sister regions using data from *S. cerevisiae* alone (even though its complete genome sequence is known) and instead only a patchwork of duplicated chromosomal regions can be detected. In our original study only 50% of the genome length could be paired up.

In making the map of duplications in yeast, Wolfe and Shields (1997) deliberately chose very conservative search criteria to define chromosomal regions that are unarguably duplicated. We did this because our aim was to show that these regions have properties that are characteristic of what is predicted by the genome duplication/reciprocal translocation model [i.e. properties (i) and (ii) above]. Consequently, the map published by Wolfe and Shields (1997) does not show some gene pairs that may have been formed by the same genome duplication event, but for which the evidence is weaker.

The aim of the present paper is to try to maximize the amount of the yeast genome that is mapped into sister chromosomal regions, working under the *assumption* that the hypothesis of simultaneous whole-genome duplication is correct. Because the hypothesis was proposed based on the existence of the duplications, this might sound like circular reasoning, but it is not. We are not trying to test the hypothesis in this paper, but to explore its consequences. Obviously, this is only useful if the hypothesis is correct, but no other credible explanation for observations (i)–(iii) above has been put forward in the two years since we made our proposal. Here, we are using the genome duplication/reciprocal translocation hypothesis to predict which gene pairs in yeast may have been formed by polyploidy. It is of interest to identify these gene pairs because they are expected to be equivalent to single genes in other species of fungi such as *K. lactis* (Keogh et al., 1998). Our approach has been to construct a core map of 'probable' sister regions, and then to overlay this map with 'possible' regions that may also be sisters, but for which the evidence is less convincing. In doing this we have taken a more methodical approach than was used in our earlier study, or by other groups who identified duplicate regions in yeast (Coissac et al., 1997; Mewes et al., 1997). Lastly, we integrated the available gene order data from *K. lactis* with the map of *S. cerevisiae* duplications.

2. Data and methods

The sequences used were the same 5790 proteins as in Wolfe and Shields (1997) and are available on our website (http://acer.gen.tcd.ie/~khwolfe/yeast). Subtelomeric repeat regions were excluded as in Seoighe and Wolfe (1998). Gene names were updated to those in version 7.1 of the Yeast Protein Database (YPD; http://www.proteome.com). The tRNA and snRNA genes analyzed were those listed by the Saccharomyces Genome Database (SGD; http://genome-www.stanford.edu). All-against-all Smith–Waterman searches (Smith and Waterman, 1981) were done using the SSEARCH program in the FASTA package (Pearson and Lipman, 1988), using the BLOSUM62 matrix (Henikoff and Henikoff, 1992) and the **seg** filter (Wootton and Federhen, 1996). Computation time for these searches on a high performance parallel computer (DEC Alphastation 8400 with eight processors) was generously provided by Compaq Computer Corporation. Duplicated chromosomal regions were identified by analyzing these results using computer programs written in C and Perl languages. The map in Fig. 2 was produced by a program written in Microsoft Visual Basic, and the version shown on our website was produced by the **gd** package (http://www.boutell.com).

3. Results and discussion

3.1. Optimizing the parameters for defining duplicated chromosomal blocks

In our previous version of the map of sister chromosomal regions, pairs of homologs with BLASTP scores (Altschul et al., 1990) in excess of 200 were included. The Smith–Waterman algorithm (Smith and Waterman, 1981) has been used instead of BLASTP for the revised map. Much work has been done on the relative merits of different algorithms and techniques for searching databases to find homologs of a query sequence. Smith–Waterman is generally accepted as the best method currently available in terms of sensitivity and specificity (Shpaer et al., 1996), but requires much more computer time than does BLAST. We used the SSEARCH Smith–Waterman program (Pearson and Lipman, 1988) with log-length normalization following Shpaer et al. (1996). Raw scores from the Smith–Waterman algorithm are dependent upon the lengths of the sequences being compared, but dividing by the product of the logarithms of the sequence lengths removes this dependence and greatly improves selectivity.

When searching for sister chromosomal regions we are not interested in all duplicated proteins, but only those proteins that were duplicated as part of the whole-genome duplication. Paralogs that existed before that time, or that were formed more recently, are of no use in determining the map of sister regions. We did not consider it feasible to use either a molecular clock approach or a phylogenetic approach (Yuan et al., 1998)

to identify the set of paralogs that were duplicated simultaneously, because (i) there are no closely related outgroup sequences for many of the yeast gene pairs, and (ii) molecular clock analysis of a small number of tetraploidy-derived paralogs yielded a considerable range of date estimates, possibly due to gene conversion (Wolfe and Shields, 1997; see also Skrabanek and Wolfe, 1998).

Instead, we followed the logic that under the hypothesis of genome duplication, followed predominantly by reciprocal translocation, there should be no overlapping blocks (sister chromosomal regions). The fraction of the genome placed in overlapping blocks (with each block containing three or more duplicated genes, as in Wolfe and Shields, 1997) was plotted for different cut-off values of similarity score (Fig. 1a). Very high cut-offs do not yield any duplicated blocks, whereas very low cut-offs generate many overlapping blocks. A cut-off of 17.5 (log-length normalized Smith–Waterman score) was chosen as the lowest similarity score that did not produce overlapping blocks.

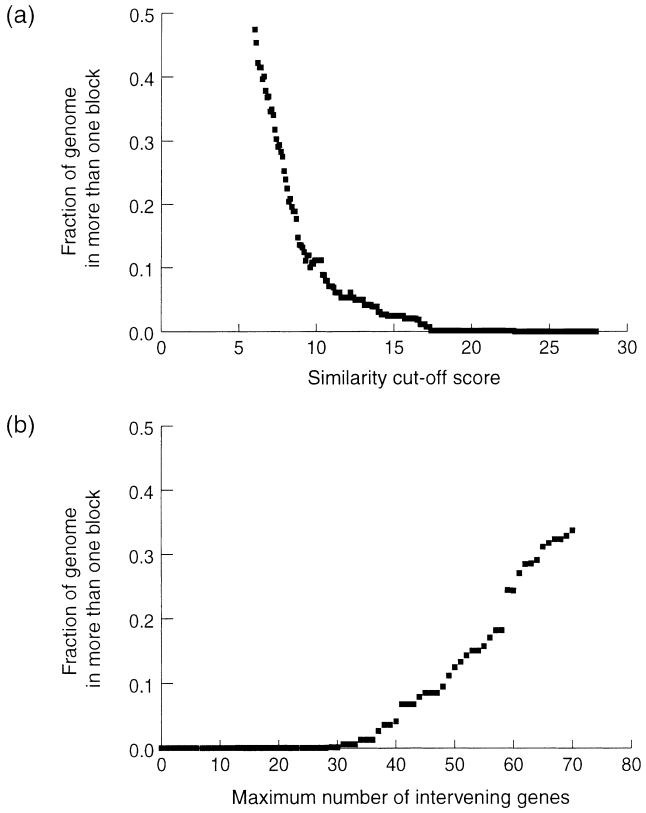

Fig. 1. Optimization of parameters used to construct the duplication map. (a) Fraction of the yeast genome simultaneously paired with more than one sister block (each block having three or more paralogs), plotted as a function of the sequence similarity cut-off score (log-length normalized Smith–Waterman score) used to define paralogs. (b) Fraction of the yeast genome simultaneously paired with more than one sister block, as a function of the maximum physical distance allowed (number of intervening non-duplicated genes) between successive paralogs making up a block.

We previously used an arbitrary limit of 50 kilobases (kb) as the maximum permitted gap between duplicated genes making up a block; this corresponds to approximately 25 genes (Wolfe and Shields, 1997). In Fig. 1b the fraction of the genome assigned to overlapping blocks is plotted against the maximum number of intervening genes allowed between neighboring paralogs. From this result we chose a cut-off distance of 30 intervening genes.

3.2. Construction of the updated map

The updated map (Fig. 2) is organized into two levels: a core framework of duplicated chromosomal blocks that are 'probable' products of genome duplication, and a second level of 'possible' paralogs and regions for which the evidence is weaker. The map was constructed by first identifying the 'probable' regions using stringent criteria, and then relaxing the criteria both to add extra 'possible' genes to the blocks already identified, and to find additional 'possible' blocks. These 'possible' genes and blocks were only added to the map where they were not in conflict with the 'probable' framework. The 'possible' genes shown in Fig. 2 are thus a selective representation of the data, and we emphasize again that our aim is to maximize the biological information that can be extracted from the map when the genome duplication hypothesis is assumed to be correct.

The paralogous gene pairs that form the 'probable' duplicated blocks are shown as thick colored bars with gene names written to the right of chromosomes in Fig. 2. There are 52 'probable' blocks and 45.5% of the genes in the genome are located inside them. These blocks contain 655 'probable' paralogs (this is not an even number because, as well as simple gene pairs, it includes a few cases where a gene in a block has two tandemly duplicated paralogs in the sister block). For only 11 pairs among these, the transcriptional orientation of one gene appears inverted as compared to the other (relative to the rest of the block that contains them), indicating a DNA inversion that occurred after the whole genome duplication. These inverted genes are marked with '@' symbols and named to the left of the chromosomes in Fig. 2. Seven of these inverted genes result from three multi-gene inversions in blocks 27, 37 and 41.

A further 34 pairs of paralogs are included as 'possible' additional genes within the 'probable' blocks. These do not have similarity scores greater than the cut-off value but they are otherwise consistent with the rest of the map. These 'possibles' are named to the left in Fig. 2, marked '(L)' for low-scoring. Transcriptional orientation, relative to the rest of the block, is conserved for 31 of these 34 pairs, which indicates that the majority of these are true paralogs. The ends of some of the

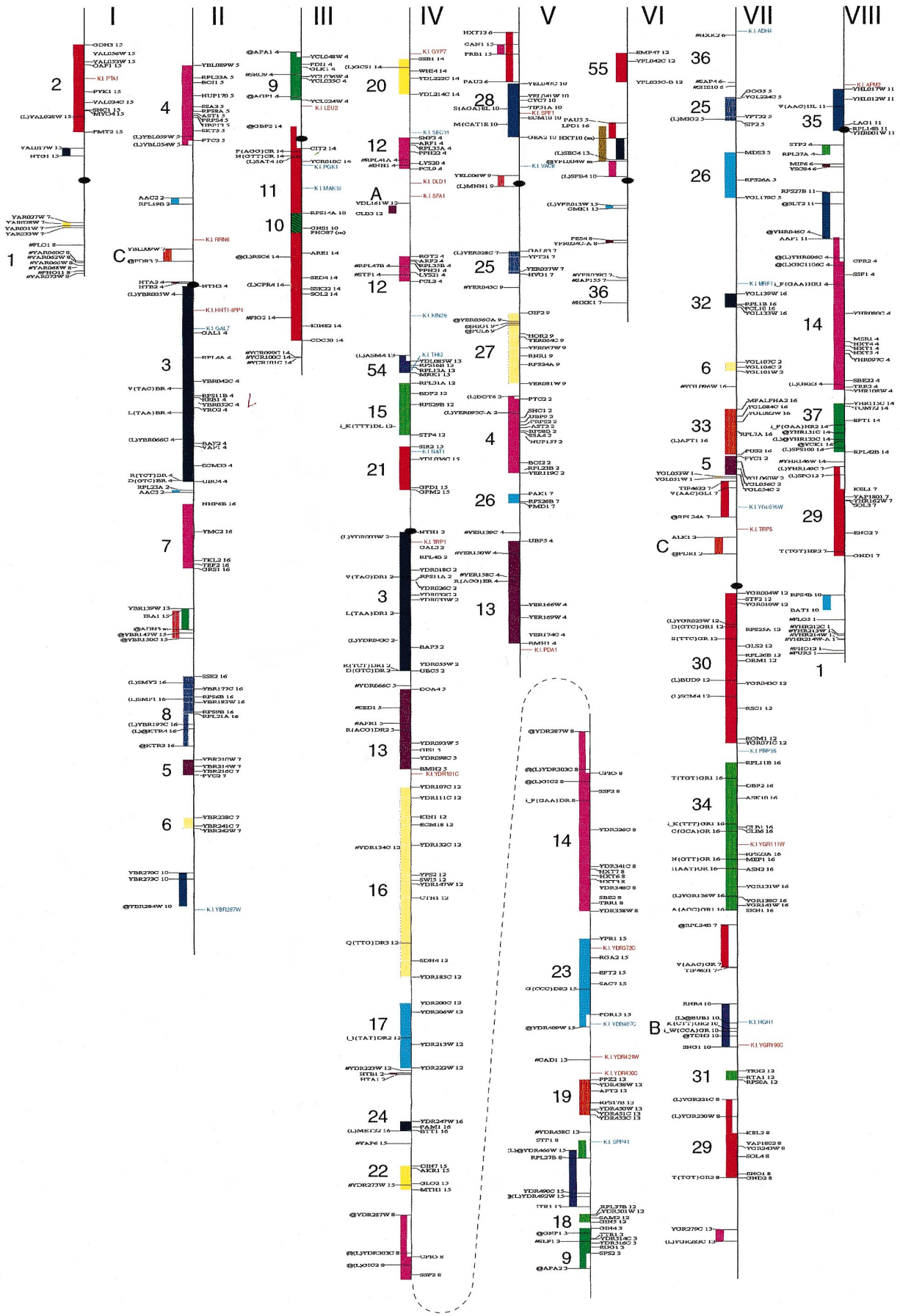

Fig. 2.

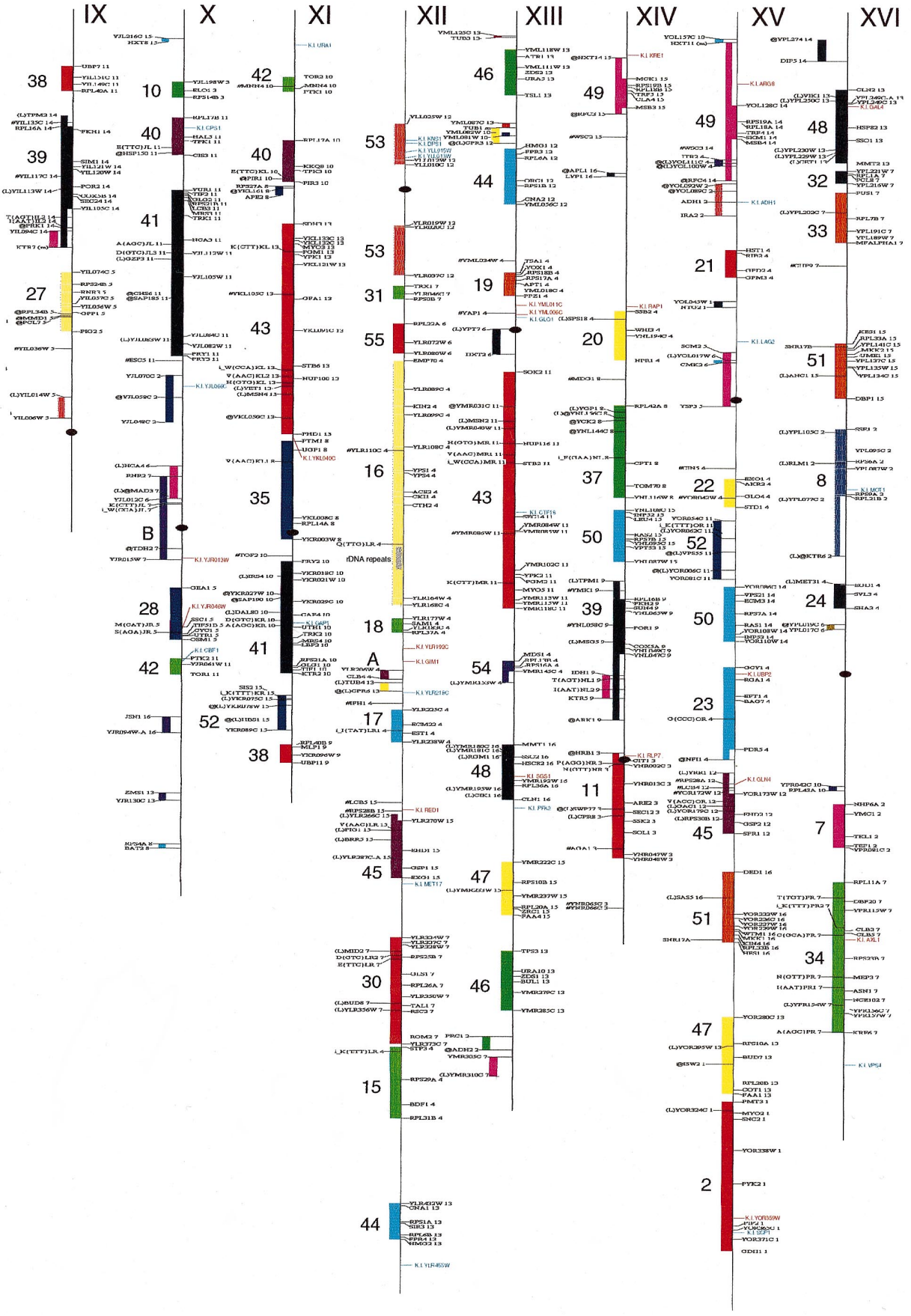

Fig. 2. (*continued*)

'probable' blocks can be extended by including 'possible' paralogs (i.e. gene pairs that are either inverted or low-scoring), and these extensions are shown as narrower colored bars on the map (Fig. 2).

There are 117 additional smaller 'possible' blocks. Of these, 32 have both copies in genomic regions outside the 'probable' blocks (excluding any extensions as described above), while 11 have both copies completely inside 'probable' blocks. This indicates that approximately 21 of the 32 two-membered blocks are genuine sister regions (the other 11 being artefacts), which is in good agreement with the theoretical prediction for the number of two-membered blocks in yeast (Seoighe and Wolfe, 1998). Only the 32 two-membered blocks that are outside the 'probable' blocks in both copies are shown in Fig. 2. It should be noted that approximately 11 of these are expected to be artefactual.

The revised map includes 39 tRNA gene pairs as well as one snRNA gene pair (*SNR17A/SNR17B*; Hughes et al., 1987). A tRNA gene was included in the map if it occurred within a block and had a homolog located in the sister block, in the equivalent interval between protein paralogs. RNA genes are named on the left of the map in Fig. 2. We used a BLASTN score ≥ 200 as the cut-off for identifying tRNA genes as homologs. This is not entirely satisfactory since it is a length-insensitive cut-off, but in the majority of cases tRNA BLASTN scores were clearly separated into high and low scoring groups. tRNAs and snRNAs could not be used to construct blocks because most of the tRNAs had too many BLASTN hits.

3.3. Comparison with the original map

52 of the 55 blocks on our earlier map appear as 'probable' blocks in Fig. 2, where they are numbered using the same scheme as in Wolfe and Shields (1997). Blocks 1 and 36 were rejected because they are very close to telomeres (on chromosomes I/VIII and VI/VII, respectively). Block 52 (on chromosomes XI/XV) is reduced to 'possible' status because the three pairs of paralogs in the center of the block are low-scoring. To facilitate comparison with the earlier map, all genes that were on that map but which would not otherwise have been included in the revised map, are shown to the left in Fig. 2 marked by hash symbols ('#'). The total numbers of genes marked in Fig. 2 are: 655 'probable', 250 'possible', 78 tRNA and two snRNA, as well as 71 withdrawn ('#' symbols). This compares to 743 protein genes in Wolfe and Shields (1997). The fraction of the proteome involved in the whole-genome duplication is approximately 16% (905 proteins on the updated map/5523 proteins encoded by non-telomeric regions of the genome).

The most remarkable change in the updated map is that block 16 has been extended so that it spans the ribosomal DNA array on chromosome XII, pairing it with part of chromosome IV. On chromosome IV, *SDH4* and *Q(TTG)DR3* (a glutamine tRNA gene) are about 15 kb apart, but their paralogs on chromosome XII [*YLR164W* and *Q(TTG)LR*] are separated by approximately 1 megabase (100–200 copies of the 9137 base-pair ribosomal DNA repeat; Johnston et al., 1997). A second copy of the rDNA array seems to have been deleted without trace from this section of chromosome IV. A similar deletion of an rDNA array may have occurred during the formation of the allopolyploid species *S. pastorianus*, which is a hybrid between *S. cerevisiae* and an *S. bayanus*-like species, but which contains only *S. bayanus*-like rDNA (James et al., 1997; Kurtzman and Robnett, 1998; McCullough et al., 1998).

A large new 'possible' duplicated block was discovered between chromosomes VII and X (labeled as block B in Fig. 2). It includes *RNR4/RNR2* (encoding a ribonucleotide reductase subunit), *BUB1/MAD3* (spindle-assembly checkpoint kinases), *TDH3/TDH2* (glyceraldehyde-3-phosphate dehydrogenase), *SNG1/YJR015W* (transport proteins), and two tRNA genes. Curiously, this block spans the centromere of chromosome X but not chromosome VII.

The updated map includes several well-known duplicated gene pairs that did not appear in the previous map. These include *PDR1/PDR3* (transcription factors), *IRA1/IRA2* (GTPase activating proteins), *HTA1/HTA2* and *HTB1/HTB2* (histones), *CLB3/CLB4* (cyclins), and *NTG1/NTG2* (glycosylases). Some other gene families

Fig. 2. Updated map of duplicated regions in the yeast genome. A web version of this map with links to information about each gene is at http://acer.gen.tcd.ie/~khwolfe/yeast. Colored rectangles adjacent to the vertical chromosome lines are 'probable' duplicated regions associated with genome duplication, containing three or more duplicated genes. Gene names written to the right of the chromosome lines indicate the genes making up these 'probable' blocks. Colored rectangles displaced to the left are 'possible' additional or alternative duplicated regions. Large numerals (1–55) show block numbers from Wolfe and Shields (1997) and large letters (A–C) show new blocks that are supported by *K. lactis* information. Numbers after gene names indicate the chromosome on which the duplicate copy is located; 'm' indicates genes with paralogs on multiple other chromosomes. '@' symbols before gene names indicate that the orientations of a pair of genes are not consistent with the orientations of the rest of the genes in the blocks in which they lie. '(L)' symbols before gene names indicate low-scoring matches (log-length normalized Smith–Waterman score between 15 and 17.5). '#' symbols before gene names indicate genes that appeared on the original map (Wolfe and Shields, 1997) but which would not otherwise appear on the updated map using the current criteria. tRNA genes are indicated by names such as P{AGG}CR (indicating a proline tRNA with anticodon AGG on the right arm of chromosome III). *K. lactis* gene order information from Table 1 is shown in red or blue lettering (with the prefix *K.l.*). Red lettering indicates *K. lactis* neighboring pairs that support the block structure; blue lettering indicates those that are either neutral or conflict with the block structure. Cases of complete gene order conservation between *K. lactis* and *S. cerevisiae* (left-hand column in Table 1) are not shown.

Table 1
Gene order comparison between *K. lactis* and *S. cerevisiae*

Gene pairs adjacent in both species	Gene pairs conserved between duplicated blocks[a]	Gene pairs adjacent in *K. lactis* but not conserved in *S. cerevisiae*
Observed: 55% (46 pairs)	*Observed: 23% (19 pairs)*	*Observed: 23% (19 pairs)*
Predicted[b]*: 59%*	*Predicted*[b]*: 22%*	*Predicted*[b]*: 19%*
RPL32–RPL24A[b]	PTA1–YOR359W[c] block 2 {16}	CTF18–CBF1[b]
RFT1–HAP3[b]	HHT1–TRP1[b] block 3	GAL7–NAT1[b]
GAL1–GAL10[b]	TRP1–IPP1[b] block 3	GAP1–ADH1[b]
GAL10–GAL7[b]	RLP7–LEU2[b] block 11	GLO1–PFK2[b]
RAD16–LYS2[c]	PDA1–YDR101C[b] block 13	KIN28–MRF1[b]
LYS2–TKL2[b]	YDR421W–YML006C[c,m] block 19 {19}	LAG2–PGK1[b]
ABD1–PRP5[c]	YDR430C–YML011C[c,m] block 19 {21}	MET17–YLL015W[b]
YBR238C–YBR239C[c]	RAP1–GYP7[b] block 20	THI3–CYC1[l,m]
YCL036W–YCL035C[c]	UBP2–YDR372C[b] block 23	MAK32–VAC8[c] {3}
MRK1-THI3[g,m]	SPF1–YJR046W[c] block 28 {4}	YDR407C–MOT1[c] {6}
PEX3–SKP1[e]	YGR111W–AXL1[c] block 34 {23}	SEC31–YLR218C[c] {7}
YDR387C–RVS167[c]	APM2–YKL040C[c,m] block 35 {11}	HGH1–YLL013C[c] {8}
ERD1–YDR412W[b]	RED1–GLN4[c] block 45 {12}	CPS1–YJL066C[c] {9}
APA2–QCR7[b]	GAL4–SGS1[b] block 48	ADH4–URA1[c] {10}
MET6–YER093C[c,h]	ARG8–KRE1[b] block 49	PRP38–DPS1[c] {13}
YGR046W–TFC4[c]	SFA1–GIM1[c] block A {5}	YGL036W–KNS1[c] {15}
YGR117C–RPS23A[c]	DLD1–YLR192C[k] block A	YBR287W–SCP1[c] {20}
CDC68–CHC1[b]	YGR196C–YJR013W[c] block B {17}	YLR455W–VPS4[c] {24}
SPT4–COX18[d]	RRN6–TRP5[c] block C {1}	SPP41–KRE6[c] {25}
YIR003W–DJP1[c]		
ERG20–QCR8[b]		
YJL082W–YJL083W[c,m] {18}		
SDH3–CTK1[h,j]		
YKL006CA-CAP1[f]		
YLL035W–YLL034C[c]		
SMC4–YLR087C[c]		
SAM1–YLR181C[c,m] {14}		
YLR181C–SWI6[b]		
YLR386W–YLR387C[c]		
URA5–SEC65[b]		
GAL80–YML050W[b]		
RPL41A–YNL161W[b]		
YNL217W–RAP1[b]		
ZWF1–YNL240C[b]		
YNL240C–KEX2[b,h]		
KEX2–YTP1[b]		
YTP1–SIN4[c]		
RFA2–YNL308C[c,h,i]		
YOL119C–RPL18A[b]		
GPD2–ARG1[c]		
GAL11–GSH2[b]		
RPO31–RPT5[c]		
YOR294W–YOR296W[c,h]		
UME1–YPL138C[c,m] {22}		
YPL112C–CAR1[c]		
NOP4–SSN3[c]		

[a] Blocks (duplicated chromosomal regions) are numbered or lettered as in Fig. 2. Numbers in braces correspond to numbered features in Fig. 4 of Ozier-Kalogeropoulos et al. (1998).
[b] See Keogh et al. (1998).
[c] From Ozier-Kalogeropoulos et al. (1998), based on clone-end sequencing.
[d] Hikkel et al. (1998).
[e] Winkler et al. (1997).
[f] Banfield (1998).
[g] Rodriguez-Belmonte et al. (1998).
[h] The *S. cerevisiae* genes are not immediately adjacent.
[i] Orientation of one gene is inverted between the species.
[j] Lee and Greenleaf (1995).
[k] Lodi et al. (1998).
[l] Ramil et al. (1998).
[m] Our interpretation differs from that of the original authors.

are not resolved into pairs and remain in competing alternative 'possible' blocks, for example *ADH1/ADH2/ADH5* (alcohol dehydrogenases) and *TUB1/TUB3/TUB4* (tubulins).

3.4. Comparison with Kluyveromyces lactis

The limited gene order information that is available from related species can provide useful information about the location of new sister regions, as well as serving as a check on existing regions. In a previous study we looked at gene pairs that were adjacent in the yeast *Kluyveromyces lactis*, and compared the locations of their orthologs in *S. cerevisiae* (Keogh et al., 1998). The *K. lactis* genome appears not to be duplicated, based on gene order data, number of chromosomes, and phylogenetic analysis of duplicated gene sequences (Wolfe and Shields, 1997; Keogh et al., 1998). With extensive additional data from *K. lactis* (Ozier-Kalogeropoulos et al., 1998) and a revised map of the duplicated regions in *S. cerevisiae*, it is worth re-examining adjacent gene pairs in *K. lactis*.

Table 1 lists 84 pairs of adjacent *K. lactis* genes and groups them into three categories of gene order conservation, as in Keogh et al. (1998). The genes listed in the middle column of Table 1 ('conserved between blocks') are labelled in red in Fig. 2; these are 19 cases where gene order in *K. lactis* resembles the gene order that existed in an ancestor of *S. cerevisiae* prior to genome duplication and gene deletion. The gene pairs listed in the right-hand column in Table 1 are labelled in blue in Fig. 2; these are 19 cases where the gene order in *K. lactis* does not appear related to the known block structure in *S. cerevisiae*. Where both of these blue labels occur in unpaired parts of the genome, they may indicate previously undetected (highly fragmented) blocks, for example the genes *ADH4* and *URA1* which are adjacent in *K. lactis* and near the telomeres of chromosomes VII and XI in *S. cerevisiae*. Other blue labels conflict with the 'probable' framework and indicate either interspecies rearrangements (translocations in *K. lactis* or transpositions in either species) or mistakes in the map. Four of these each involve a gene located in duplicated block 53 on chromosome XII (Fig. 2). Four rearrangements in the small region occupied by block 53 seems unlikely, so this block is probably spurious. It contained the minimum number of paralogs (just three) for inclusion in the original map, and two paralogs (*YLL025W/YLR037C*) are members of the large PAU multigene family.

Adjacent *K. lactis* genes that map to locations near three 'possible' sister regions add weight to these new candidate blocks (blocks A, B and C; Table 1 and red labels in Fig. 2). These examples illustrate how complete mapping of *K. lactis* (or *A. gossypii*) would provide a much clearer picture of the sister regions in *S. cerevisiae* and of the evolution of gene order after genome duplication. Another example of the utility of *K. lactis* information is the relationship between block 49 (chromosomes XIV and XV) and the genes *KRE1* and *ARG8* which are adjacent in *K. lactis*. The positions of *KRE1* and *ARG8* in *S. cerevisiae* are incompatible with the possible extension of block 49 to include the gene pair *HXT14/HXT11*, so the *HXT* pair is probably artefactual.

In Table 1, the 'predicted' values for the percentage of gene pairs in three columns are taken directly from our previous study (Keogh et al., 1998), which used the original map of duplicated regions (Wolfe and Shields, 1997). They were not updated because it is not clear how to include uncertain ('possible') regions in the analysis. Also, the results of Ozier-Kalogeropoulos et al. (1998) are based on 'genome survey' sequencing of both ends of plasmid clones, and in some cases their paired *K. lactis* sequences correspond to *S. cerevisiae* genes that are separated by a small number of intervening genes; this data is awkward to analyze. However, the difference between the maps is not significant and the observations from *K. lactis* (Table 1) remain close to the predictions in Keogh et al. (1998).

Acknowledgements

We thank Compaq Computer Corporation's software engineering group in Galway for access to computers, and Mike McLean for help with making Fig. 2. Supported by the European Communities 4th Framework Biotechnology Program (BIO4-CT95-0130).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Banfield, D.K., 1998. DNA sequence of the SFT1 gene from *Kluyveromyces lactis*, GenBank/EMBL/DDBJ database accession number AF072674.

Coissac, E., Maillier, E., Netter, P., 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol. Biol. Evol. 14, 1062–1074.

Dietrich, F.S., Voegeli, S., Gaffney, T., Mohr, C., Rebischung, C., Wing, R., Choi, S., Goff, S., Philippsen, P., 1999. Gene map of chromosome I of *Ashbya gossypii*. Curr. Genet. 35, 233

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., et al., 1996. Life with 6000 genes. Science 274, 546–547.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA 89, 10915–10919.

Hikkel, I., Gbelska, Y., Subik, J., 1998. Identification and functional analysis of a *Kluyveromyces lactis* homologue of the *SPT4* gene of *Saccharomyces cerevisiae*. Curr. Genet. 34, 375–378.

Hughes, J.M., Konings, D.A., Cesareni, G., 1987. The yeast homologue of U3 snRNA. EMBO J. 6, 2145–2155.

James, S.A., Cai, J., Roberts, I.N., Collins, M.D., 1997. A phylogenetic

analysis of the genus *Saccharomyces* based on 18S rRNA gene sequences: description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martiniae* sp. nov.. Int. J. Syst. Bacteriol. 47, 453–460.

Johnston, M., Hillier, L., Riles, L., Albermann, K., Andre, B., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. Nature 387, Suppl., 87–90.

Keogh, R.S., Seoighe, C., Wolfe, K.H., 1998. Evolution of gene order and chromosome number in *Saccharomyces*, *Kluyveromyces* and related fungi. Yeast 14, 443–457.

Kurtzman, C.P., Robnett, C.J., 1998. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie Van Leeuwenhoek 73, 331–371.

Lalo, D., Stettler, S., Mariotte, S., Slonimski, P.P., Thuriaux, P., 1993. Une duplication fossile entre les régions centromériques de deux chromosomes chez la levure. C.R. Acad. Sci. Paris 316, 367–373.

Lee, J.M., Greenleaf, A.L., 1995. *Kluyveromyces lactis* CTD kinase largest subunit (*CTK1*) gene, GenBank/EMBL/DDBJ database accession number U24219.

Lodi, T., Goffrini, P., Bolondi, I., Ferrero, I., 1998. Transcriptional regulation of the *KlDLD* gene, encoding the mitochondrial enzyme D-lactate ferricytochrome c oxidoreductase in *Kluyveromyces lactis*: effect of *Klhap2* and *fog* mutations. Curr. Genet. 34, 12–20.

McCullough, M.J., Clemons, K.V., McCusker, J.H., Stevens, D.A., 1998. Intergenic transcribed spacer PCR ribotyping for differentiation of *Saccharomyces* species and interspecific hybrids. J. Clin. Microbiol. 36, 1035–1038.

Melnick, L., Sherman, F., 1993. The gene clusters *ARC* and *COR* on chromosomes 5 and 10, respectively, of *Saccharomyces cerevisiae* share a common ancestry. J. Mol. Biol. 233, 372–388.

Mewes, H.W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., et al., 1997. Overview of the yeast genome. Nature 387, Suppl., 7–65.

Ozier-Kalogeropoulos, O., Malpertuy, A., Boyer, J., Tekaia, F., Dujon, B., 1998. Random exploration of the *Kluyveromyces lactis* genome and comparison with that of *Saccharomyces cerevisiae*. Nucleic Acids Res. 26, 5511–5524.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Philippsen, P., Kleine, K., Pohlmann, R., Dusterhoft, A., Hamberg, K., et al., 1997. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIV and its evolutionary implications. Nature 387, Suppl., 93–98.

Ramil, E., Freire-Picos, M.A., Cerdan, M.E., 1998. Characterization of promoter regions involved in high expression of *KlCYC1*. Eur. J. Biochem. 256, 67–74.

Rodriguez-Belmonte, E., Gonzalez-Siso, I., Cerdan, E., 1998. The *Kluyveromyces lactis* gene *KLGSK-3* combines functions which in *Saccharomyces cerevisiae* are performed by *MCK1* and *MSD1*. Curr. Genet. 33, 262–267.

Seoighe, C., Wolfe, K.H., 1998. Extent of genomic rearrangement after genome duplication in yeast. Proc. Natl. Acad. Sci. USA 95, 4447–4452.

Shpaer, E.G., Robinson, M., Yee, D., Candlin, J.D., Mines, R., Hunkapiller, T., 1996. Sensitivity and selectivity in protein similarity searches: a comparison of Smith–Waterman in hardware to BLAST and FASTA. Genomics 38, 179–191.

Skrabanek, L., Wolfe, K.H., 1998. Eukaryote genome duplication — where's the evidence? Curr. Opin. Genet. Devel. 8, 694–700.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Winkler, A., Goedegebure, R., Zonneveld, B.J.M., Steensma, H.Y., Hooykaas, P.J.J., 1997. *Kluyveromyces lactis SKP1* can complement a mutation in *CTF13*, a gene coding for a centromeric protein of *Saccharomyces cerevisiae*, GenBank/EMBL/DDBJ database accession number AF012338.

Wolfe, K.H., Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708–713.

Wootton, J.C., Federhen, S., 1996. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266, 554–571.

Yuan, Y.P., Eulenstein, O., Vingron, M., Bork, P., 1998. Towards detection of orthologues in sequence databases. Bioinformatics 14, 285–289.