Genomic Exploration of the Hemiascomycetous Yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*

Bertrand Llorente[a,1], Alain Malpertuy[a,1], Cécile Neuvéglise[b], Jacky de Montigny[c], Michel Aigle[d], François Artiguenave[e], Gaëlle Blandin[a], Monique Bolotin-Fukuhara[f], Elisabeth Bon[b], Philippe Brottier[e], Serge Casaregola[b], Pascal Durrens[d], Claude Gaillardin[b], Andrée Lépingle[b], Odile Ozier-Kalogéropoulos[a], Serge Potier[c], William Saurin[e], Fredj Tekaia[a], Claire Toffano-Nioche[f], Micheline Wésolowski-Louvel[g], Patrick Wincker[e], Jean Weissenbach[e], Jean-Luc Souciet[c], Bernard Dujon[a,*]

[a] *Unité de Génétique Moléculaire des Levures (URA 2171 CNRS and UFR927 Univ. P. and M. Curie, Paris), Institut Pasteur, 25 Rue du Dr Roux, F-75724 Paris Cedex 15, France*

[b] *Collection de Levures d'Intérêt Biotechnologique, Laboratoire de Génétique Moléculaire et Cellulaire (INRA UMR216, CNRS URA1925) INA-PG, P.O. Box 01, F-78850 Thiverval-Grignon, France*

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

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

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

[f] *Institut de Génétique Moléculaire (CNRS/UPS UMR 8621), Bâtiment 400, Université de Paris Sud, F-91405 Orsay, France*

[g] *Microbiologie et Génétique (CNRS/UCB/INSA ERS 2009), Bâtiment 405 R2, Université Lyon I, F-69622 Villeurbanne Cedex, France*

Received 13 November 2000; accepted 15 November 2000

First published online 29 November 2000

Edited by Horst Feldmann

Abstract We have analyzed the evolution of chromosome maps of Hemiascomycetes by comparing gene order and orientation of the 13 yeast species partially sequenced in this program with the genome map of *Saccharomyces cerevisiae*. From the analysis of nearly 8000 situations in which two distinct genes having homologs in *S. cerevisiae* could be identified on the sequenced inserts of another yeast species, we have quantified the loss of synteny, the frequency of single gene deletion and the occurrence of gene inversion. Traces of ancestral duplications in the genome of *S. cerevisiae* could be identified from the comparison with the other species that do not entirely coincide with those identified from the comparison of *S. cerevisiae* with itself. From such duplications and from the correlation observed between gene inversion and loss of synteny, a model is proposed for the molecular evolution of Hemiascomycetes. This model, which can possibly be extended to other eukaryotes, is based on the reiteration of events of duplication of chromosome segments, creating transient merodiploids that are subsequently resolved by single gene deletion events. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Duplication; Inversion; Deletion; Translocation; Redundancy; Loss

1. Introduction

Chromosome map conservation is an important criterion to judge about the phylogenetic relationship between living organisms. In return, this conservation is useful to identify genes of interest in a species whose genetic map is poor if there exists a related species with a higher resolution genetic map or, better, with a complete genome sequence. The problem of map conservation, however, is complicated by the degree of internal duplication within a genome. All genomes sequenced so far, even those of the simplest organisms, show traces of ancestral gene duplications with various degrees of subsequent sequence divergence, forming sets of paralogous genes. In the genomes of *Caenorhabditis elegans* or *Drosophila melanogaster* for example, half of the genes are not unique [1]. In *Bacillus subtilis* or *Escherichia coli*, the degree of genetic redundancy is higher than in *Saccharomyces cerevisiae*, despite their smaller genomes and lower gene numbers [2,3]. Part of the redundancy is attributable to functional selection. In *B. subtilis*, for example, the number of genes encoding demonstrated or putative ATP-binding transporters is very high [2], consistent with its habitat. It is much lower in an intracellular pathogen such as *Mycobacterium tuberculosis* [4]. But another part of the phenomenon, at least in eukaryotes, may simply result from basic mechanisms of chromosome dynamics such as non-reciprocal translocations, accidental non-disjunction at mitosis or meiosis creating aneuploids, or sequence healing by gap-repair or break-induced replication. Such mechanisms duplicate entire chromosomes or segments of chromosomes carrying several genes at the same time as they reshuffle the genetic map.

The genomic sequence of the yeast, *S. cerevisiae* [5,6], reveals a large number of chromosomal segments that appears as the likely result of ancestral duplications, followed by massive gene loss and various degrees of sequence divergence of the remaining genes [7–9]. Consequently, pairs of homologous genes are observed, forming series along two distinct chromosome segments, that are interspersed by unique genes. Over

*Corresponding author. Fax: (33)-1-40 61 34 56.
E-mail: bdujon@pasteur.fr

[1] These authors contributed equally to this work.

PII: S0014-5793(00)02289-4

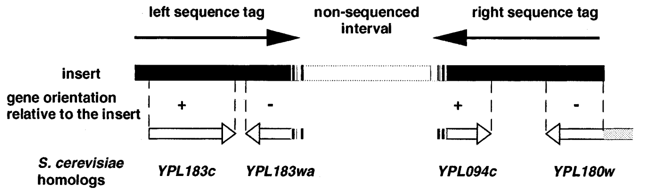


Fig. 1. Order and orientation of genes in inserts: Shown is an example of an insert from *P. angusta* in which four genes were recognized as having homologs in *S. cerevisiae*. The left sequence tag contains a first gene, homologous to *YPL183c*, whose coding strand is in direct orientation relative to the sequence of the insert (+ sign), and part of a second gene, homologous to *YPL183wa*, in opposite orientation (− sign). The right sequence tag contains part of a first gene, homologous to *YPL180w*, whose coding strand is in direct orientation relative to the sequence, and part of a second gene, homologous to *YPL094c*, whose coding strand is antiparallel relative to the sequence of the tag. After inversion of the right sequence tag to built the insert map, the last two genes receive a − sign and a + sign, respectively. For calculation, the order and orientation of genes in this insert is written as follows: *PaYPL183c* ($f1=+1$) *PaYPL183wa* ($f2=-1$) *PaYPL094c* ($f3=+1$) *PaYPL180w* ($f4=-1$), the fi values representing the orientation signs. All inserts from the present sequencing program were treated accordingly.

55 different chromosome blocks issued from ancestral duplications have been recognized in the yeast genome. Altogether, they cover nearly half of the yeast genome sequence, such that the hypothesis of a complete genome duplication, followed by massive gene loss, was proposed [8]. This duplication was even precisely placed in the phylogenetic tree of the yeasts [10].

The wealth of novel sequence information provided by the present program allows us to examine map conservation between *S. cerevisiae* and a variety of other yeast species selected, on purpose, across the entire hemiascomycete realm.

We could estimate the global degree of conservation of synteny between *S. cerevisiae* and each of the other yeast species studied from the conservation, or non-conservation, of neighboring gene-couples on the *S. cerevisiae* map. We could also estimate the frequency of single gene deletion, and of inversion of gene orientation. Finally, by examining the location of the homologs of the non-syntenic gene-couples on the *S. cerevisiae* map, we have discovered *trans*-chromosomal series that reveal ancestral segments of chromosome duplications. Interestingly, such series do not entirely coincide with the ancestral chromosome duplication blocks observed in the genome of *S. cerevisiae* when compared to itself. We conclude that a major driving force of molecular evolution, at least in eukaryotes, is the continuous duplication of chromosome segments, each encompassing a few genes, that are transposed to ectopic chromosomal locations in direct or inverted orientation relative to centromeres. The resulting increase in genome size and redundancy is counterbalanced by a high frequency of single gene deletion. With this hypothesis, there is no need to believe that *S. cerevisiae* would result from an ancestral whole-genome duplication as was proposed in [8].

2. Materials and methods

2.1. Data sets

For each yeast species, we first established the list of all sequenced inserts in which two or more protein-coding genes were recognized as having homologs in *S. cerevisiae*. Most inserts are composed of two non-overlapping sequenced tags (Fig. 1). Inserts with single tags [11] were also taken into consideration, provided they contain two or more genes. For each insert, the order and orientation of the genes were then computed. The left–right orientation of each sequenced insert being arbitrary, the orientation of each gene in the insert is designated arbitrarily with a + or − sign, as shown in Fig. 1. Note that the insert maps take into consideration genes whose sequences are entirely included within the insert as well as genes only partially included, provided their homology to a *S. cerevisiae* gene is unambiguous (to eliminate possible ambiguity, we only considered genes having a single clear-cut homolog in *S. cerevisiae*, notation 'o', see [12]). It follows that all tRNA genes were ignored because most of them are members of large gene families in *S. cerevisiae*. In some cases, a same gene may have a part of its sequence in the left-hand sequence tag and the other in the right-hand sequence tag of the same insert. Such genes were considered only once. Most inserts contain only two genes, but a significant number contain three genes, and a few inserts were found to contain four or even five genes (Table 1).

2.2. Computing gene-couples from insert maps

Cases in which three or more genes were present in the same insert were treated as several gene-couples according to the principle illustrated by Fig. 2 and explained in text. For each yeast species, a list of gene-couples was then established, taking into consideration data from all inserts. Due to the sequencing strategy used, this list contains genes that are immediate neighbors on the chromosome map of the yeast species of interest as well as genes that may be separated by one (or a few) other gene(s) that could be present in the unsequenced central part of each insert (see Fig. 1). The total number of gene-couples obtained for each yeast species is given by Table 1.

Table 1
The complete set of sequence data used for comparative mapping with *S. cerevisiae*

Yeast species	Total number of inserts with *n* genes				Deduced nb of gene-couples		Data taken from
	$n=2$	$n=3$	$n=4$	$n=5$	total	N.R. set	
S. bayanus var. *uvarum*	1273	378	41	4	2167	1787	[13]
S. exiguus	347	70	3		500	471	[14]
S. servazzii	424	85	11		639	561	[15]
Z. rouxii	386	178	12		819	719	[16]
S. kluyveri	381	106	12		651	579	[17]
K. thermotolerans	412	117	16		719	682	[18]
K. lactis	847	135	13		1188	980	[19]
K. marxianus var. *marxianus*	356	52	3	1	481	454	[20]
P. angusta	704	176	14		1112	995	[21]
D. hansenii var. *hansenii*	235	44			329	303	[22]
P. sorbitophila	83	14			112	98	[23]
C. tropicalis	157	12	3		193	188	[24]
Y. lipolytica	148	5			158	149	[25]
Total	5753	1372	128	5	9068	7966	

The table gives, for each yeast species, the total number of sequenced inserts in which two or more protein-coding genes were identified as having non-ambiguous homologs in *S. cerevisiae*, and the deduced number of gene-couples subsequently used for analysis (see Section 2).
N.R. set: Non-redundant set.

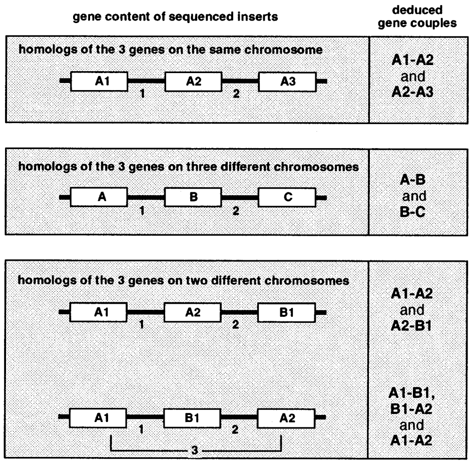

gene content of sequenced inserts
deduced gene couples
homologs of the 3 genes on the same chromosome
A1
A2
A3
1
2
A1-A2 and A2-A3
homologs of the 3 genes on three different chromosomes
A
B
C
1
2
A-B and B-C
homologs of the 3 genes on two different chromosomes
A1
A2
B1
1
2
A1-A2 and A2-B1
A1
B1
A2
1
2
3
A1-B1, B1-A2 and A1-A2

Fig. 2. Rationale for the computation of gene-couples from inserts having more than two genes with homologs in *S. cerevisiae*. For inserts containing three genes, only two intervals were taken into consideration when (i) the three genes have their homologs on the same *S. cerevisiae* chromosome, (ii) the three genes have their homologs on three different *S. cerevisiae* chromosomes, or (iii) the three genes have their homologs on two different *S. cerevisiae* chromosomes, the two falling on the same chromosome being contiguous. Three intervals were taken into consideration when the three genes have their homologs on two different *S. cerevisiae* chromosomes, the two falling on the same chromosome being separated from each other by the third gene. In this latter topology, cases in which the two external genes are syntenic with *S. cerevisiae* are designated as 'intermingled triples', see text and Fig. 8). Inserts containing four or five genes were treated accordingly.

2.3. Computing the S. cerevisiae reference map

In order to facilitate comparisons of gene orders, intervals and orientations, all 6213 predicted protein-coding genes of *S. cerevisiae* [12] were attributed an 'absolute number' (*N*) according to their order of appearance along the complete genome sequence starting at the left end of chromosome 1 and ending at the right end of chromosome 16 (Fig. 3). Partially overlapping genes were included in the numbering. The orientation of each gene relative to its centromere was also recorded using the designation 'To' and 'Fro' as explained in Fig. 3. Finally, the location of each gene on its chromosome (C) was defined by the coordinate of its center on the chromosome sequence.

2.4. Analysis of gene-couples relative to the S. cerevisiae map

For each yeast species, the set of gene-couples was treated as follows.

2.4.1. Step 1: reorientation of gene-couples. Because the orientation of the sequenced inserts is arbitrary, gene-couples were first reoriented from left to right such that $N2-N1>0$, if $N1$ and $N2$ are the two *S. cerevisiae* homologs of the left and right genes of the couple, respectively. This step is necessary to reduce the comparisons to an upper-diagonal half matrix. In the couples that had to be inverted, the signs of the genes (+ or −) were also inverted.

2.4.2. Step 2: chromosome classification. Gene-couples were then separated into two sets according to whether their *S. cerevisiae* homologs fall on the same chromosome ($X2-X1=0$) or not ($X2-X1 \geq 1$). The first were designated *cis*-couples, the second *trans*-couples.

2.4.3. Step 3: redundancy filtration. Despite our low genome coverage sequencing strategy [26], a number of redundant gene-couples were found. All such cases were eliminated before any further calculation, giving rise to the non-redundant sets of *cis*-couples and *trans*-couples (see Table 1).

2.4.4. Step 4: definition of the syntenic couples. For each *cis*-couple of the non-redundant list, the physical distance ($C2-C1$) separating the two *S. cerevisiae* homologs was calculated along with the number of intervening genes ($N2-N1-1$). The distributions of these two parameters were examined for each yeast species to establish experimentally the limit beyond which couples should be considered as non-syntenic even though their homologs fall on the same chromosome of *S. cerevisiae* (see Fig. 4 and text for additional explanations). All couples whose distances are within the accepted limit were regarded as syntenic with *S. cerevisiae* and listed as *syn*-couples. All couples whose homologs are more distant than the accepted limit were merged with the *trans*-couple list to constitute the *nonsyn*-couple list. The overall synteny of the sequenced yeast species to *S. cerevisiae* is given by the ratio S/(S+NS) where S is the total number of *syn*-couples and NS the total number of *nonsyn*-couples.

3. Results and discussion

3.1. Rationale for data analysis

The low-coverage random sequencing strategy used in this program does not allow complete map comparisons but it provides a wealth of data about local gene order relationships because the average size of the sequenced inserts (3–5 kb) was selected such as to coincide with the average distance between two neighboring protein-coding genes of *S. cerevisiae* [27] and because the sequences read by the LiCor sequencing machines are exceptionally long [11]. This strategy imposed the development of methods for the analysis of data that may prove useful for other comparative mapping studies in similar situations. Compared to classical works on synteny that usually consider large chromosomal blocks or even entire chromosomes, the basic idea of this work is to treat the problem by a bottom-up approach, starting from elementary gene-couples and building larger segments only latter.

Key points of our rationale are as follows. First, we have examined the gene contents of the sequenced inserts rather

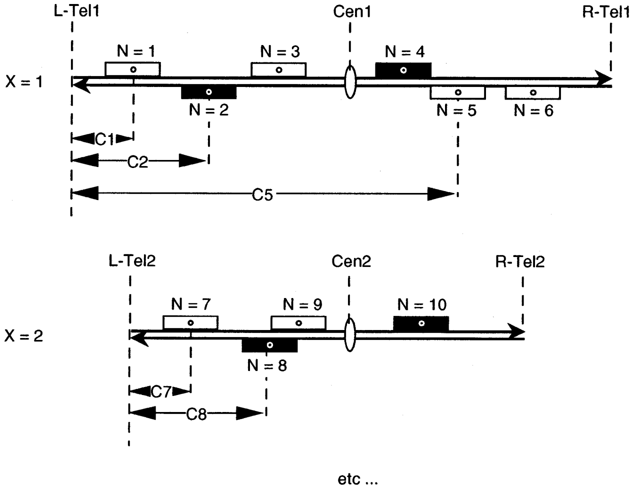
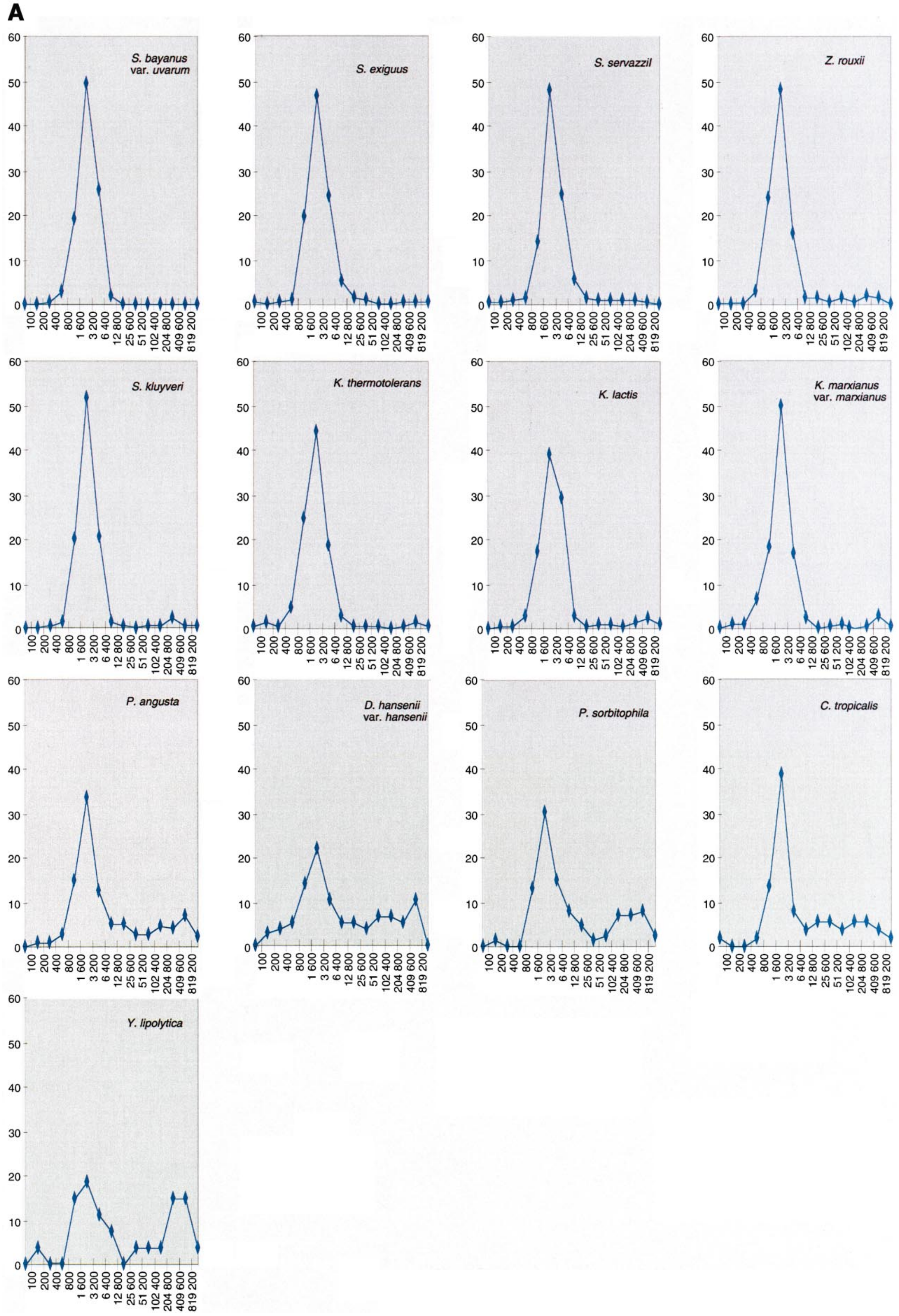


Fig. 3. Computing the *S. cerevisiae* reference map. The figure schematizes the mathematical parameters used for computation of synteny. All protein-coding genes of *S. cerevisiae* are described by four parameters (1) the number of their chromosome ($X=1$–16), (2) an absolute number starting from the left end of chromosome 1 end ending at the right end of chromosome 16 ($N=1$–6213), (3) the distance of the center of the gene from the left end of the chromosome (C), and (4) the orientation of the gene with respect to its centromere (Fro or To). *Fro* genes (filled boxes) are oriented away from centromeres (genes on the Watson strand on the right chromosome arms and genes on the Crick strand on the left arms). *To* genes (void boxes) are oriented towards the centromeres (genes on the Watson strand on the left chromosome arms and genes on the Crick strand on the right arms). Partially overlapping genes, as symbolized by N8 and N9, are numbered sequentially like all other genes. The four parameters were used to compute the gene-couples of the sequenced inserts of the other yeast species as explained in text (Section 2.4).



A
S. bayanus var. uvarum
S. exiguus
S. servazzil
Z. rouxii
S. kluyveri
K. thermotolerans
K. lactis
K. marxianus var. marxianus
P. angusta
D. hansenii var. hansenii
P. sorbitophila
C. tropicalis
Y. lipolytica
0
10
20
30
40
50
60
100
200
400
800
1 600
3 200
6 400
12 800
25 600
51 200
102 400
204 800
409 600
819 200

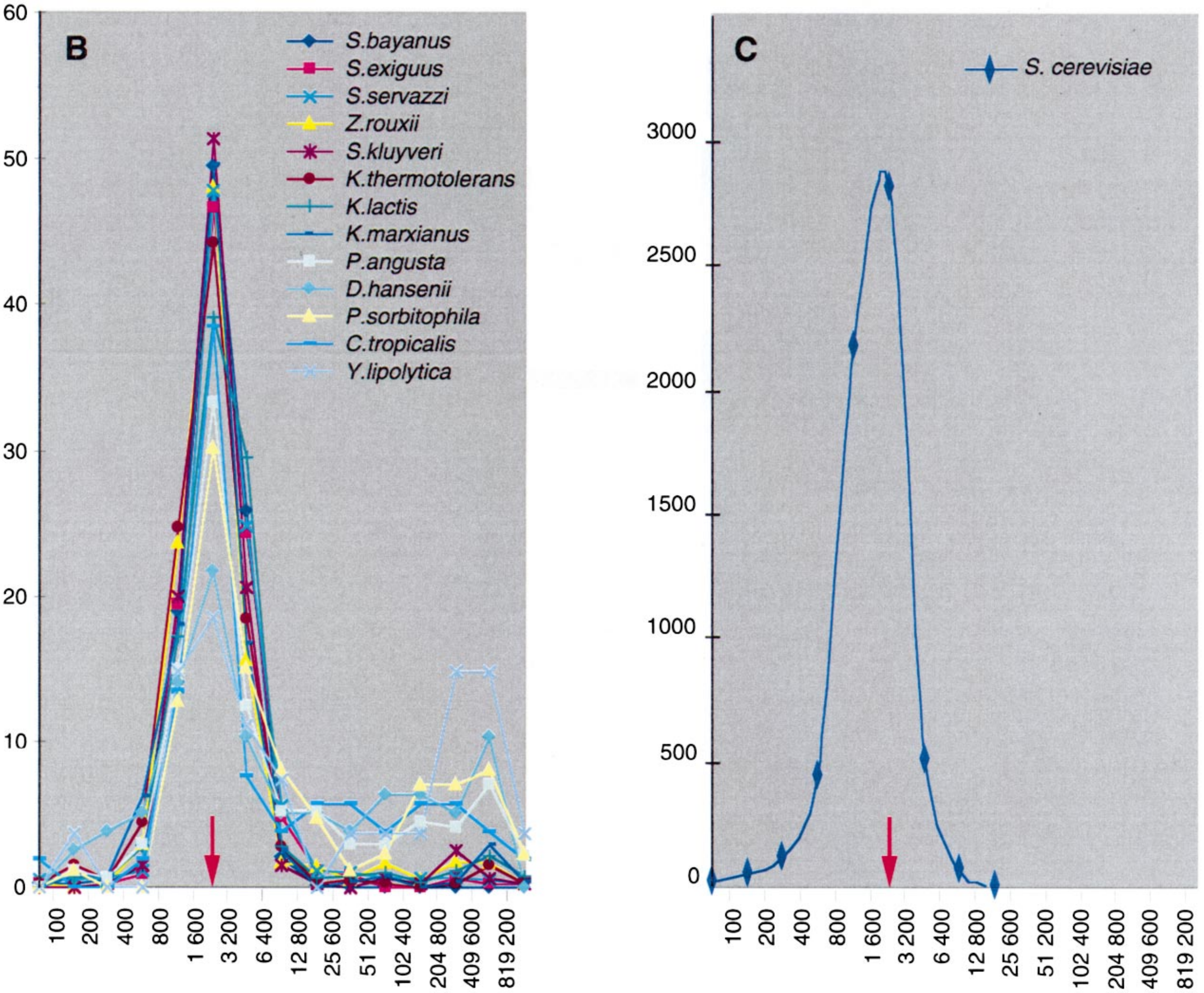
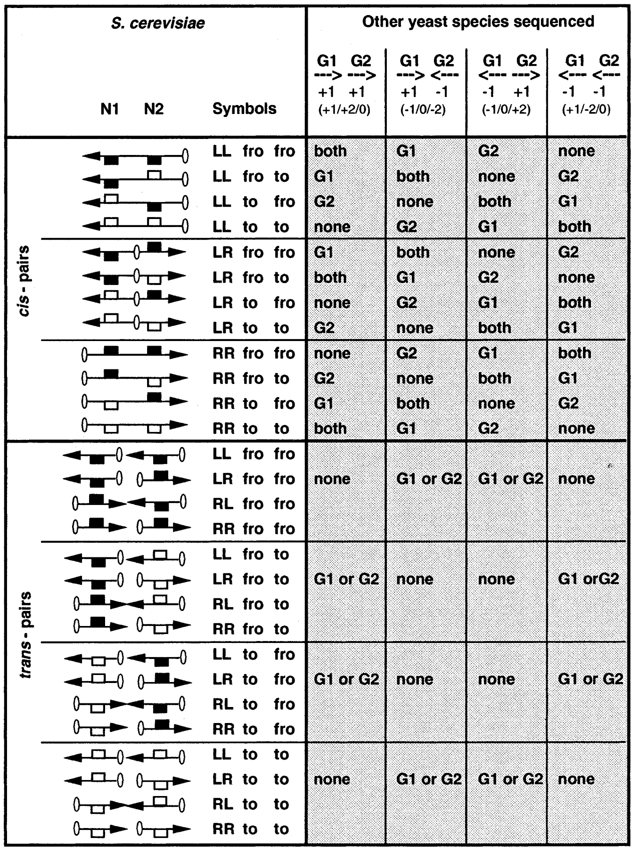


Fig. 4. Distribution of physical distances in intervals between the two *S. cerevisiae* homologs of *cis*-couples. The physical distance D (in bp) between the two *S. cerevisiae* homologs to each *cis*-couple was calculated ($D = C2 - C1$, see Fig. 3) and the distribution of such distances for a given species was computed for the entire non-redundant set of *cis*-couples. Classes of distances were defined as exponentially increasing, starting from 50 bp, using the equation: $D' = 50 \times 2^{\mathrm{int}\,[\log_2(D/50)]}$ where D' is the lower limit of the corresponding distance class (abscissa). To allow direct comparison between species, the frequency of each distance class (ordinate) was expressed as the percent of the total number of *cis*-couples in each species. Part A allows detailed visualization of the frequency distribution for each individual yeast species. Part B shows the superposition of all curves for the 13 yeast species to illustrate the remarkable coincidence of the major peak (centered between 1600 and 3200 bp, red arrow) irrespective of the divergence of frequencies in the larger distance classes (above 25 600 bp). For comparison, part C shows the distribution of distances between successive genes on the *S. cerevisiae* map, using the same distance classes. We consider that *cis*-couples having two *S. cerevisiae* homologs whose distance falls within the major peak correspond to conserved synteny despite the possible presence of few intervening genes on the sequenced insert or on the *S. cerevisiae* map. We consider that the *cis*-couples having their two *S. cerevisiae* homologs more distant than the limit of the major peak correspond to loss of synteny despite the fact that the two homologs are on the same chromosome (see text).

than that of single tags (see Fig. 1). Thus, our rationale can be applied to any organism, provided the genomic insert size is commensurate with the average distance between genes. Second, we have only considered sequenced inserts individually because at the low genome coverage at which we have worked, contigs are too rare to be informative. Third, we have compared each gene-couple of the sequenced inserts relative to a unique reference map (that of *S. cerevisiae*) that was mathematically parameterized for the purpose. We have not tried multidimensional species comparisons. Fourth, we have considered all genes identified in the sequenced inserts, whether completely or only partially sequenced, provided the homology with the *S. cerevisiae* genes was unambiguous.

3.2. *Definition of syntenic couples and conservation of synteny between S. cerevisiae and each of the 13 other yeast species*

For each of the 13 yeast species partially sequenced in this program, the complete set of gene-couples was analyzed with respect to the map location and orientation of their *S. cerevisiae* homologs. In a first step, couples were separated between those (designated *cis*-couples) having their two homologs on the same chromosome of the *S. cerevisiae* map and those (designated *trans*-couples) having their two homologs on distinct chromosomes (see Section 2). In the latter case, synteny of the two genes of the couple has necessarily been lost between *S. cerevisiae* and the yeast species considered. But the former case is more complex because we observed a number of instances in which the two *S. cerevisiae* homologs



	S. cerevisiae Symbols	Other yeast species sequenced G1 G2 ---> ---> +1 +1 (+1/+2/0)	G1 G2 ---> <--- +1 -1 (-1/0/-2)	G1 G2 <--- ---> -1 +1 (-1/0/+2)	G1 G2 <--- <--- -1 -1 (+1/-2/0)
cis - pairs	LL fro fro	both	G1	G2	none
	LL fro to	G1	both	none	G2
	LL to fro	G2	none	both	G1
	LL to to	none	G2	G1	both
	LR fro fro	G1	both	none	G2
	LR fro to	both	G1	G2	none
	LR to fro	none	G2	G1	both
	LR to to	G2	none	both	G1
	RR fro fro	none	G2	G1	both
	RR fro to	G2	none	both	G1
	RR to fro	G1	both	none	G2
	RR to to	both	G1	G2	none
trans - pairs	LL fro fro LR fro fro RL fro fro RR fro fro	none	G1 or G2	G1 or G2	none
	LL fro to LR fro to RL fro to RR fro to	G1 or G2	none	none	G1 orG2
	LL to fro LR to fro RL to fro RR to fro	G1 or G2	none	none	G1 or G2
	LL to to LR to to RL to to RR to to	none	G1 or G2	G1 or G2	none

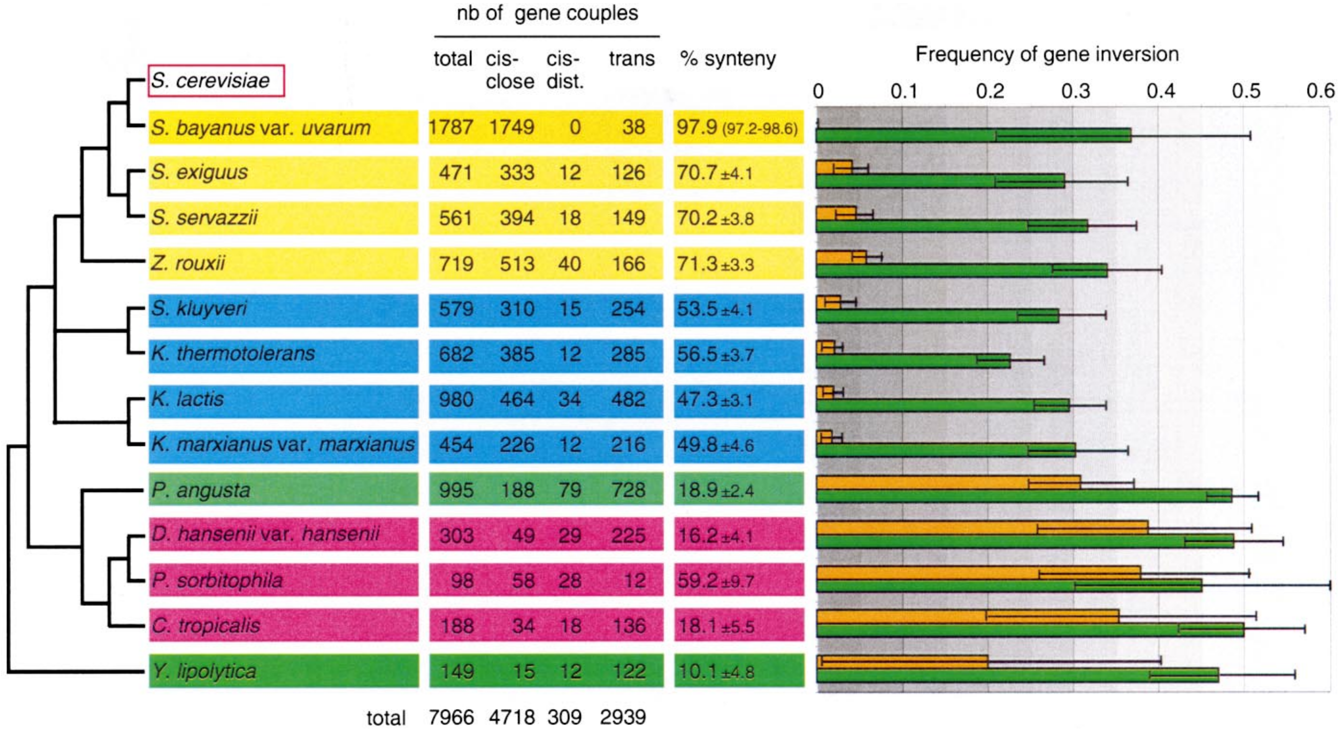
Fig. 5. Rationale for the determination of relative gene orientation. The double entry table compares every possible map location and orientation of two *S. cerevisiae* genes designated N1 and N2 (lines) with the four possible orientations of the gene-couples (G1 and G2) sequenced in the other yeast species (columns). G1 is homologous to N1, G2 is homologous to N2. The left and right arms of *S. cerevisiae* chromosomes are symbolized by arrows (and designated L or R), with an oval to identify the centromere. Genes are symbolized by rectangles above the arrow (Watson strand) or below the arrow (Crick strand). Full rectangles: genes oriented away from centromere (Fro), void rectangles: genes oriented towards centromere (To). The orientation of G1 and G2 in the sequenced gene-couples is visualized by the arrows. To facilitate computation of results a value ($f=+1$ or $f=-1$) was attributed to each orientation. Each of the four possibilities could then be identified mathematically by the product, sum and difference of the two f values of G1 and G2 (figures in brackets). For each logical combination, the figure indicates which of the two genes, G1 or G2, is inverted with respect to the *S. cerevisiae* genes (see text for additional explanations).

were distant from each other on the chromosome map. For all *cis*-couples, we have, therefore, examined the distribution of the physical distances between the two *S. cerevisiae* homologs (Fig. 4A). For all species, a major peak is observed corresponding to distances ranging from ca. 1600 to 6400 nucleotides (for *Debaryomyces hansenii* var. *hansenii* and *Yarrowia lipolytica*, this peak is not as precisely defined due to the small number of *cis*-pairs). But the remaining part of the distributions differs between the species. For the first eight species (*Saccharomyces bayanus* var. *uvarum*, *Saccharomyces exiguus*, *Zygosaccharomyces rouxii*, *Saccharomyces servazzii*, *Saccharomyces kluyveri*, *Kluyveromyces thermotolerans*, *Kluyveromyces lactis* and *Kluyveromyces marxianus* var. *marxianus*), only few gene-couples are observed outside of the major peak (less than 8% of the total, see data on Fig. 6). For the five other species more distantly related to *S. cerevisiae* (*Candida tropicalis*, *D. hansenii* var. *hansenii*, *Pichia angusta*, *Pichia sorbitophila*, and *Y. lipolytica*) a significantly larger proportion of the gene-couples (30–45%) have their homologs distant or even very distant on the *S. cerevisiae* chromosomes. Because the major peak perfectly coincides for all 13 species (Fig. 4B) and corresponds to the average distance between two neighboring genes in *S. cerevisiae* (Fig. 4C), we consider that *cis*-couples having two *S. cerevisiae* homologs whose distance falls within the major peak correspond to a conservation of synteny. On the contrary, we consider that the distant *cis*-couples correspond to loss of synteny, the two genes occurring by chance on the same *S. cerevisiae* chromosome after multiple events of rupture of synteny. This view is supported by the relative proportion of the distant *cis*-couples compared to the number of *trans*-couples in each species (see data on Fig. 6). For four species (*S. kluyveri*, *K. thermotolerans*, *K. lactis* and *K. marxianus* var. *marxianus*), the distant *cis*-couples represent ca. 4–7% of all non-syntenic couples in perfect agreement with the probability of any two genes to fall randomly on the same *S. cerevisiae* chromosome considering the total number of chromosomes and their relative sizes (the value of $\Sigma(\mathrm{pi})^2$, where pi represents the size of chromosome i relative to the total genome size, is 7.5%). For other species, that proportion varies between ca. 9 and 12%, and for *Z. rouxii* it reaches ca. 20%. But the most remarkable exception is represented by *P. sorbitophila* in which 28 cases of distant *cis*-couples were observed for only 12 *trans*-couples, placing the ratio at 70%.

Following above considerations, the conservation of synteny between *S. cerevisiae* and each of the thirteen other yeast species studied ranges from 98% for *S. bayanus* var. *uvarum* to only 10% for *Y. lipolytica* (see Fig. 6). The decrease of synteny conservation is in good agreement with phylogenetic distances estimated from rDNA sequences [26] and from amino acid sequence divergence [28]. Three species, *S. exiguus*, *Z. rouxii* and *S. servazzii* show ca. 70% of conservation of synteny with *S. cerevisiae*, while the figures slowly decrease from 56 to 47% for *K. thermotolerans*, *S. kluyveri*, *K. marxianus* var. *marxianus* and *K. lactis*, and then drops from 19 to 10% for the more distant species. In this latter group, *P. sorbitophila* appears as an exception with a conserved synteny of ca. 60% with *S. cerevisiae* due to the fact that very few *trans*-couples were observed in this species. The explanation of this phenomenon, as well as the frequent occurrence of distant *cis*-couples in the five species most distantly distantly related to *S. cerevisiae* is unclear. Consistent with the mechanism proposed below, it is possible that, as soon as two neighboring genes become separated by the insertion of an intervening chromosome segment, their chance to be further separated by subsequent insertions of other segments increases.

3.3. Gene inversion in syntenic and non-syntenic couples

Classical genetics textbooks teach us that reciprocal translocations between segments of chromosomes tend to conserve the orientation of the translocated genes with respect to centromeres because the opposite case generates acentric and dicentric chromosomes often leading to inviable descents. Thus, if the loss of synteny between two genes results from the presence of a translocation point between them, their orientation with respect to centromeres should be preserved. Classical genetics textbooks also teach us that inversions of gene orientation with respect to centromeres could result from inversions of chromosome segments encompassing one (single



	nb of gene couples				
	total	cis-close	cis-dist.	trans	% synteny
S. bayanus var. *uvarum*	1787	1749	0	38	97.9 (97.2-98.6)
S. exiguus	471	333	12	126	70.7 ±4.1
S. servazzii	561	394	18	149	70.2 ±3.8
Z. rouxii	719	513	40	166	71.3 ±3.3
S. kluyveri	579	310	15	254	53.5 ±4.1
K. thermotolerans	682	385	12	285	56.5 ±3.7
K. lactis	980	464	34	482	47.3 ±3.1
K. marxianus var. *marxianus*	454	226	12	216	49.8 ±4.6
P. angusta	995	188	79	728	18.9 ±2.4
D. hansenii var. *hansenii*	303	49	29	225	16.2 ±4.1
P. sorbitophila	98	58	28	12	59.2 ±9.7
C. tropicalis	188	34	18	136	18.1 ±5.5
Y. lipolytica	149	15	12	122	10.1 ±4.8
total	7966	4718	309	2939	

Fig. 6. Conservation of synteny and gene orientation. The figure shows for each yeast species, the total number of gene-couples falling in the three categories defined in text, and the overall percent of conservation of synteny with *S. cerevisiae*. The cladogram is taken from [28]. Shown on the right part of the figure is the frequency of gene inversion in the syntenic couples (orange bars) and the non-syntenic couples (green bars). Error bars represent the 95% confidence limits of percentages calculated according to the Gaussian distribution ($1.96\times\sqrt{[p(1-p)/n]}$, where n is the sample size) except for extreme values ($<10\%$ or $>90\%$) where the Poisson distribution was applied.

gene inversion) or several genes. In the latter case, the orientation of the genes contained within the inverted segment remains unchanged relative to one-another. Thus, if the loss of synteny between two genes results from the presence of the end of an inverted segment between them, one of the two genes should be inverted with respect to the centromere, while the series of genes contained in the inverted segment keep their local syntenic relationship and their relative orientation. Finally, the loss of synteny between two genes may also result from their simultaneous duplication followed by the loss of one gene in one of the two copies and the loss of the other gene in the other copy. In such a case, the orientation of the genes of the non-syntenic couple relative to the centromeres may be inverted or not depending upon the size of the duplicated segment. If the duplicated segment is very large relative to the chromosome, or if the duplication concerns the entire chromosome (abnormal disjunction at mitosis or meiosis are not rare in yeast), then gene orientation should be preserved relative to the centromeres. If the duplicated segment is small relative to the chromosome, gene orientation may not be conserved.

It was particularly interesting to examine the orientation of genes of the syntenic and non-syntenic couples in the various yeast species because this sequencing program offers us an unprecedented number of cases from a single eukaryotic phylum. Now, in the absence of detailed physical maps of the yeast species partially sequenced, the orientation of each sequenced insert relative to its centromere cannot be determined. Yet, the inversion or non-inversion of genes compared to *S. cerevisiae* can be determined in the gene-couples defined above. The rationale is illustrated by Fig. 5 for both *cis*- and *trans*-couples.

For *trans*-couples, only four logical possibilities exist for the *S. cerevisiae* homologs, depending solely upon their orientation relative to their centromeres (To or Fro). Because the gene order in *S. cerevisiae* is irrelevant in *trans*-couples and because the sequenced inserts of the other yeasts are not oriented with respect to their centromeres, the logical matrix is further simplified, leaving only two possibilities. For two *S. cerevisiae* genes of the same orientation relative to their centromeres (Fro–Fro couples or To–To couples), the mathematical product of the f values of the genes in the sequenced couples ($f1\times f2$) is positive if none of the two genes has been inverted (or if both genes have been inverted), and is negative if one of the two genes in the sequenced couples has been inverted. The opposite relationship applies when the two *S. cerevisiae* genes have opposite orientations relative to their centromeres (Fro–To couples or To–Fro couples).

For *cis*-couples, the situation is more complex because the order of the two genes of the couple now informs us about which of them has been inverted. Three logical possibilities exists for the *S. cerevisiae* homologs, depending upon the chromosome arm, each of which being further subdivided into four cases depending on the orientation of the genes relative to the centromere. When the two *S. cerevisiae* homologs fall on the same chromosome arm, inversion of a single gene of the couple in the other yeast is monitored by a negative value of the mathematical product $f1\times f2$ if the two *S. cerevisiae* homologs are co-oriented (Fro–Fro or To–To couples), and by a positive value of the same product if they are in opposite orientation (To–Fro and Fro–To couples). The converse is true for the cases in which the two *S. cerevisiae* homologs fall on two different arms of the chromosome. Note that this case is rare for close *cis*-couples but is frequent for the distant *cis*-couples. Because the gene-couples of the sequenced inserts are oriented according to increasing *N* values of their *S. cerevisiae* homologs (see Section 2), the four logical cases distinguish an inversion of the first gene, from an inversion of the second gene and from a double inversion, as explained by Fig. 5.

From above considerations, we have calculated the exact proportion of gene inversion for all syntenic or non-syntenic couples for each of the 13 yeast species studied. Results are given by Fig. 6. Two groups of species clearly appear. In the first group, composed of *S. bayanus* var. *uvarum*, *S. exiguus*, *Z. rouxii*, *S. servazzii*, *S. kluyveri*, *K. thermotolerans*, *K. lactis* and *K. marxianus* var. *marxianus*, the proportion of gene inversion among syntenic couples is low (0.2–5.8%) while this value fluctuates between ca. 22 and 36% for the non-syntenic couples. In the second group of species, composed of *C. tropicalis, D. hansenii* var. *hansenii, P. angusta, P. sorbitophila* and *Y. lipolytica*, the proportion of inversion among syntenic couples varies from ca. 20 to 38%, while it is very close to 50% among the non-syntenic couples.

These results call for several conclusions of primary importance for understanding the mechanisms involved. First, genes can be inverted without losing synteny with their neighbors. This phenomenon remains rare over relatively long evolutionary distances (from *S. cerevisiae* to the three *Kluyveromyces* studied, average amino acid identity is less than 60% [28] and overall conservation of synteny is around 50%). However, gene inversion without rupture of synteny becomes prominent over longer evolutionary distances. Among the five species concerned, we have identified a total of 114 such cases. Some of them involve inversion of a single gene. Others involve complex situations including inversion within inversion (Fig. 7). Second, loss of synteny is accompanied by a very significant increase in gene inversion frequency that plateaus at exactly 50%, the expected value if orientation becomes random. This will be discussed further in Section 3.7 after examination of other properties.

3.4. Deletion of intervening genes within syntenic couples: intermingled triples and quadruples

We have observed a significant number of gene-couples whose synteny is conserved with *S. cerevisiae* but in which the two homologous gene pairs are interspersed by one (or a few) intervening gene whose homolog lies on another *S. cerevisiae* chromosome. An example of such intermingled triples is shown by Fig. 8A. We have also observed intermingled quadruples (Fig. 8B) among the rarer cases in which four genes could be identified on the sequenced inserts. Such situations suggest that one (or a few) gene(s) may be deleted during evolution while the syntenic relationship of their neighbors is preserved. A precise estimation of the frequency of this phenomenon is of primary importance to characterize the molecular evolution of the hemiascomycete genomes. Fig. 9 shows the total number of intermingled triples observed for each yeast species and their proportion relative to the total number of syntenic couples. It can be seen that, within sampling errors due to the limited number of cases available, this proportion does not significantly differ for all species studied (average 3%). This situation is as expected if one considers that the intermingled triples result from the duplication of chromosome segments in the ancestry of *S. cerevisiae* followed by deletion of single genes in the duplicated copies. If we consider the topology of genes in the sequenced inserts of the other yeasts as the ancestral form, it follows that a single gene deletion after duplication of chromosomal segments in *S.*

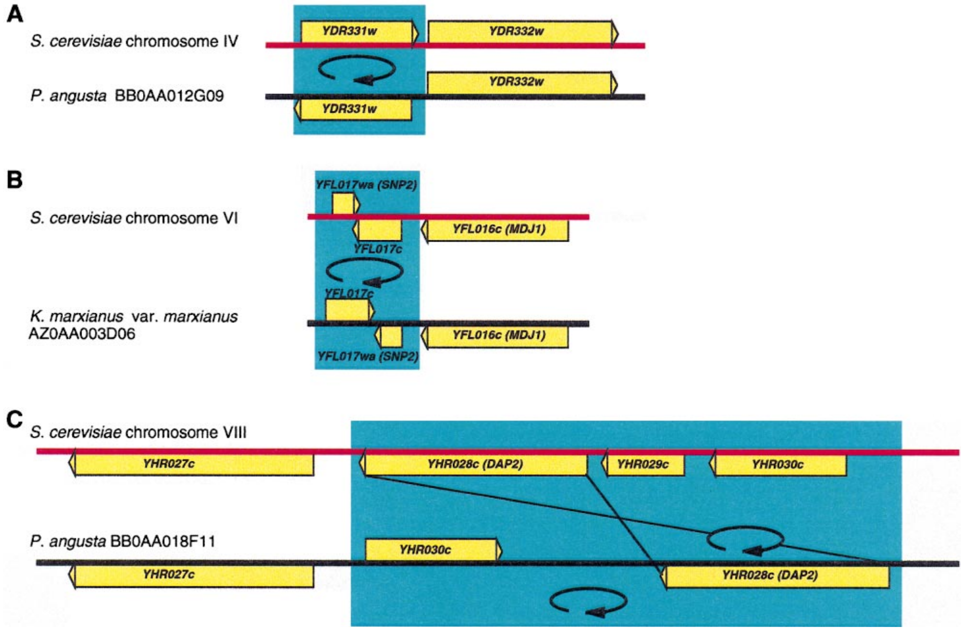


Fig. 7. Examples of gene inversions without rupture of synteny. Shown as red lines are parts of *S. cerevisiae* chromosomes with their genes (yellow rectangles). Shown as black lines are sequenced inserts of other yeast species with homologous genes (yellow rectangles). Inserts are designated with their code after the species name, see [26]. Gene orientation is indicated by the arrow and by the location of the rectangle above (rightward orientation) or below (leftward orientation) the DNA lines. Inverted segments are visualized by the blue background and identified by the curved arrow. Note the double inversion in example C.

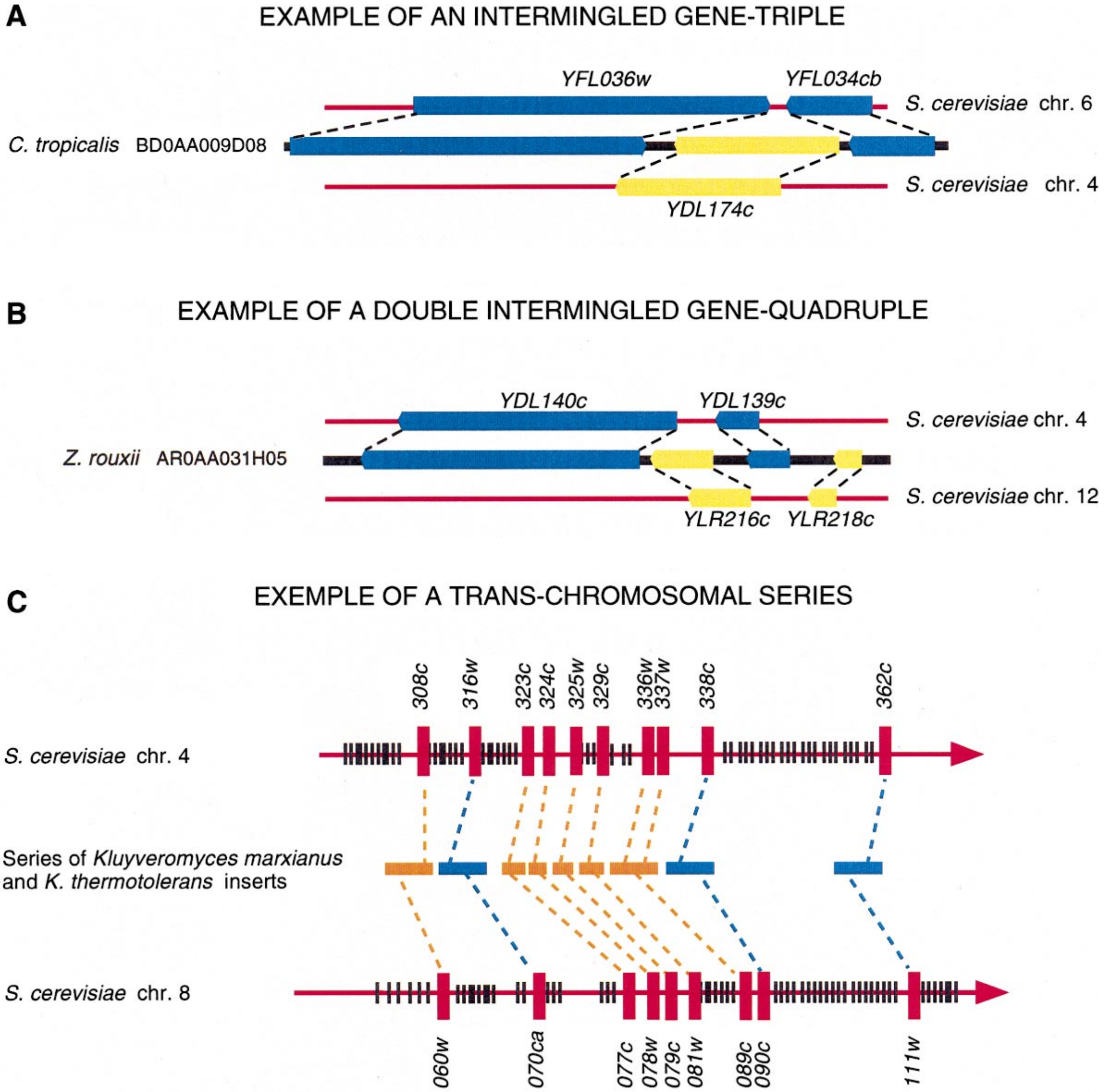

A
EXAMPLE OF AN INTERMINGLED GENE-TRIPLE
YFL036w
YFL034cb
S. cerevisiae chr. 6
C. tropicalis BD0AA009D08
S. cerevisiae chr. 4
YDL174c
B
EXAMPLE OF A DOUBLE INTERMINGLED GENE-QUADRUPLE
YDL140c
YDL139c
S. cerevisiae chr. 4
Z. rouxii AR0AA031H05
S. cerevisiae chr. 12
YLR216c YLR218c
C
EXEMPLE OF A TRANS-CHROMOSOMAL SERIES
308c 316w 323c 324c 325w 329c 336w 337w 338c 362c
S. cerevisiae chr. 4
Series of Kluyveromyces marxianus and K. thermotolerans inserts
S. cerevisiae chr. 8
060w 070ca 077c 078w 079c 081w 089c 090c 111w

Fig. 8. Examples of three characteristic topological relationships frequently observed in this work. The figure shows representative examples of the intermingled triples (A) or quadruples (B) observed as well as an example of a *trans*-chromosomal series of non-syntenic gene-couples (C). 'Intermingled triples' are gene-triples from sequenced inserts of a given yeast species in which the two external genes show conserved synteny with *S. cerevisiae* but are separated from each other by another gene whose homolog lies on a different *S. cerevisiae* chromosome (the most frequent case) or on the same *S. cerevisiae* chromosome but outside of the two syntenic pairs. For statistical analysis, the 31 gene-quadruples forming 'double intermingled structures' observed in this program were treated as successive triples using the same principles as explained in Fig. 2.

cerevisiae concerns at least 9% of the duplicated genes (the two flanking genes of each intermingled triple are also deleted in one of the two *S. cerevisiae* copies). Note that this is a global average value only. It is not possible to estimate individual cases with the set of data presently available.

Note that the gene deletion described above should not be confused with gene loss from a given yeast species because only cases of unambiguous homologs between *S. cerevisiae* and the other yeast species are considered here. The extent of gene loss can be estimated from the number of genes found in each other species that have no homolog in *S. cerevisiae* [13–25].

Orientation of the three genes in the intermingled triples was examined using the same principles as for the gene-couples. Interestingly, the two external genes show frequencies of inversion not different from the ones deduced, for each species, from the total number of syntenic gene-couples (11 cases out of 139, in total). In other words, deletion of a single gene in the *S. cerevisiae* genome is not accompanied by inversion of the flanking two genes. On the opposite, the central gene of the intermingled triples is often inverted (59 cases out of 139), suggesting that the ancestral duplication of chromosome segments of *S. cerevisiae* recognized from comparisons with other species, do not preserve orientation with respect to centromeres. This is opposite to the conclusion of Wolfe and Shields [8] concerning ancestral chromosomal duplications recognized from comparison of *S. cerevisiae* with itself.

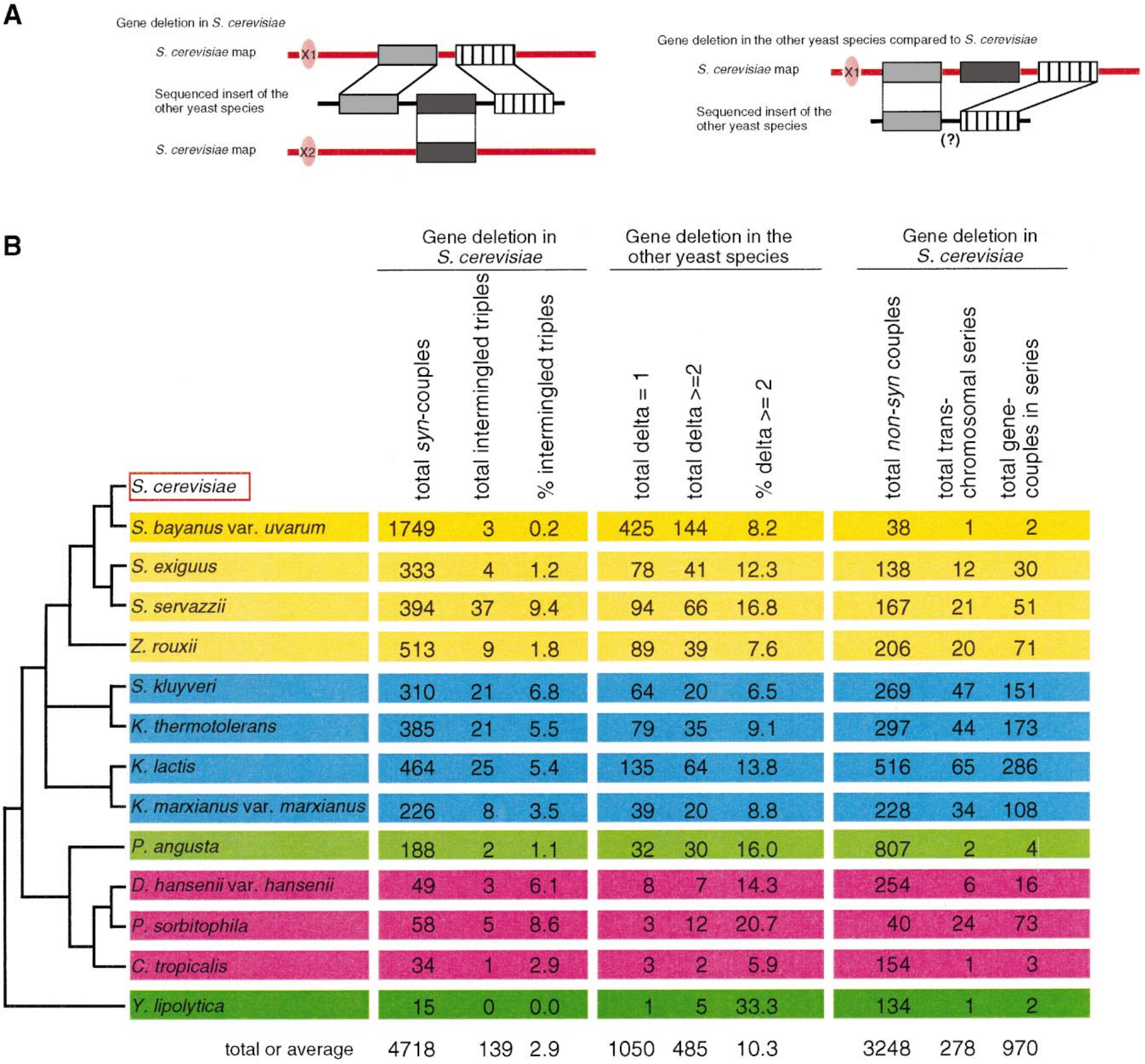
The location of intermingled triples on the *S. cerevisiae* map is informative of the ancestral duplications recognized from comparisons with other species (Fig. 10). As can be seen, they are distributed throughout the maps of the 16 chromosomes. Some of them coincide with the ancestral chromosomal duplications recognized from comparison of *S. cerevisiae* with itself, but not others.

3.5. Deletion of intervening genes within syntenic couples: non-*S. cerevisiae* yeasts

Reciprocally, estimation of the frequency of single gene deletion in each of the other yeast species should be possible from the cases in which syntenic couples are separated by one (or a few) intervening gene on the *S. cerevisiae* map (Fig. 9). This estimation is not as precise, however, because of the possible existence of genes in the central unsequenced part of most of our inserts (see Fig. 1). Yet, we have found a total of 1050 cases in which a syntenic couple is separated by one intervening gene on the *S. cerevisiae* map (Fig. 9), and 335, 91, 30, 23 and 6 cases in which they are separated by two, three, four, five or even six genes on the *S. cerevisiae* map, respectively (note that such cases should not be confused with the distant *cis*-couples, see Section 3.2). If we consider only the latter cases as an estimate of the frequency of gene deletion, the average value is close to 10% with significant variations between species (Fig. 9). If we consider the topology of genes in *S. cerevisiae* as the ancestral form, it follows that single gene deletion after duplication of chromosomal segments has occurred in all other yeast species studied with frequencies not very different from what is observed in *S. cerevisiae*.

3.6. Trans-chromosomal series

As mentioned above, among the limited number of inserts


A
Gene deletion in S. cerevisiae
S. cerevisiae map
X1
Sequenced insert of the other yeast species
S. cerevisiae map
X2
Gene deletion in the other yeast species compared to S. cerevisiae
S. cerevisiae map
X1
Sequenced insert of the other yeast species
(?)
B

	Gene deletion in *S. cerevisiae*			Gene deletion in the other yeast species			Gene deletion in *S. cerevisiae*		
	total *syn*-couples	total intermingled triples	% intermingled triples	total delta = 1	total delta >=2	% delta >= 2	total *non-syn* couples	total trans-chromosomal series	total gene-couples in series
S. cerevisiae									
S. bayanus var. *uvarum*	1749	3	0.2	425	144	8.2	38	1	2
S. exiguus	333	4	1.2	78	41	12.3	138	12	30
S. servazzii	394	37	9.4	94	66	16.8	167	21	51
Z. rouxii	513	9	1.8	89	39	7.6	206	20	71
S. kluyveri	310	21	6.8	64	20	6.5	269	47	151
K. thermotolerans	385	21	5.5	79	35	9.1	297	44	173
K. lactis	464	25	5.4	135	64	13.8	516	65	286
K. marxianus var. *marxianus*	226	8	3.5	39	20	8.8	228	34	108
P. angusta	188	2	1.1	32	30	16.0	807	2	4
D. hansenii var. *hansenii*	49	3	6.1	8	7	14.3	254	6	16
P. sorbitophila	58	5	8.6	3	12	20.7	40	24	73
C. tropicalis	34	1	2.9	3	2	5.9	154	1	3
Y. lipolytica	15	0	0.0	1	5	33.3	134	1	2
total or average	4718	139	2.9	1050	485	10.3	3248	278	970

Fig. 9. Gene deletions within conserved syntenic couples. The figure shows the possible cases of internal deletions in gene-couples whose synteny is conserved between *S. cerevisiae* and the yeast species of interest. The total number of cases observed for each yeast species is given. To analyze gene deletion in *S. cerevisiae* from trans-chromosomal series (right part of B), we have computed, for every non-syntenic gene-couple from each yeast species, the absolute numbers of the two *S. cerevisiae* homologs (*N*1 and *N*2) and examined the succession of the *N*2 incremental values when the couples are sorted according to increasing *N*1 values. Series of *trans*-chromosomal gene-couples thus appear for all yeast species considered.

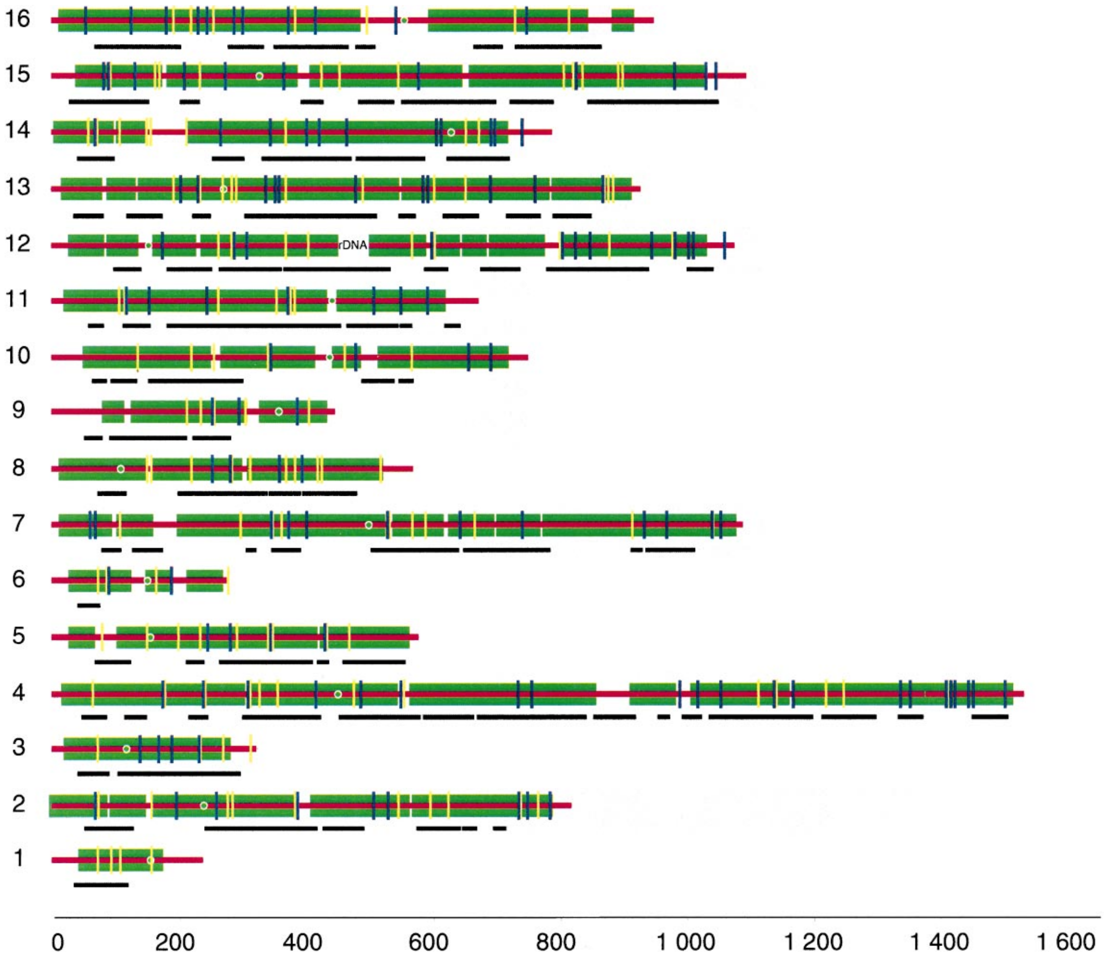


Fig. 10. Localization of intermingled triples and *trans*-chromosomal series on the *S. cerevisiae* map. The 16 *S. cerevisiae* chromosomes are drawn to scale (red lines) with a green dot to indicates centromeres. Vertical bars indicate the positions of intermingled triples (blue = external genes, yellow = internal gene). The green bars along chromosomes indicates maximal extensions of the *trans*-chromosomal series recognized in this work. Ancestral duplication blocks described in [10] are indicated as black lines below each chromosome for comparison.

having four or five genes with homologs in *S. cerevisiae*, a few appear as a mosaic of two *S. cerevisiae* chromosomes (see Fig. 8). Such intermingled series are formally equivalent to three successive non-syntenic gene-couples, involving the same two chromosomes of *S. cerevisiae*. Careful examination of the location of all non-syntenic couples on the *S. cerevisiae* map reveals that this phenomenon is much more general than originally observed from the limited number of inserts with four genes. Shown in Fig. 8C is an example of a series of nine *trans*-couples from *K. thermotolerans* and *K. marxianus* var. *marxianus* that contain genes whose homologs lie alternatively on chromosomes 4 and 8 of *S. cerevisiae*. It is tempting to assume that such *trans*-chromosomal series correspond to a chromosomal duplication in the ancestry of *S. cerevisiae*, compensated by massive gene deletion, leaving about half of the genes in each copy.

We have, therefore, tried to establish the list of *trans*-chromosomal series from the non-syntenic gene-couples of each of the 13 yeast species studied. A total of 278 series were recognized that include 970 gene-couples (Fig. 9). *Trans*-chromosomal series can be recognized in all species, but with significantly different frequencies. Relatively to the total number of non-syntenic gene-couples, they are particularly abundant in *P. sorbitophila*, *S. kluyveri*, *K. thermotolerans*, *K. lactis* and *K. marxianus* var. *marxianus*, but they are also frequently observed in *Z. rouxii*, *S. servazzii* and *S. exiguus*. They are very rare in the other five species. Thus there appears no clear-cut correlation between the phylogenetic distance of a given yeast species to *S. cerevisiae* and the number of series observed but this has to be taken with caution because it is easier to identify series from species sequenced at 0.4×g enome coverage than from species sequenced at 0.2×genome coverage and because the limit of series is imprecise. The longest series recognized encompasses 14 gene-couples (*K. lactis*) but there are 16 series encompassing 10 gene-couples or more and 142 encompassing five gene-couples or more.

The location of the *trans*-chromosomal series on the *S. cerevisiae* map is shown in Fig. 10. Interestingly, the series found in different yeast species tend to coincide on the map. A series observed in one species may prolong a series from another species or result in the fusion between two neighboring series. Like the situation of intermingled triples, some *trans*-chromosomal series coincide with the ancestral chromosomal duplications recognized from comparison of *S. cerevisiae* with itself, but not others.

3.7. Concluding remarks: a mechanism for eukaryotic genome evolution

From the very large set of novel sequence data obtained from this program, comparisons to *S. cerevisiae* could be performed at an unprecedented scale. Two major features emerge. First, there exists a correlation between loss of synteny and inversion of gene orientation. Second, numerous duplications of chromosomal segments compensated by gene deletions were identified. From these features, we propose that the molecular evolution of yeast genomes proceeds primarily by the duplications of chromosomal segments encompassing a few genes rather than by the duplication of entire chromosomes or by other chromosomal rearrangements that do not involve duplication of sequences. During this process, we propose that the duplicated copy of a given chromosome segment is inserted on a different chromosome, or at an ectopic position on the initial chromosome, in a random orientation relative to centromeres, hence creating series of genes that conserve their relative orientation between one another but have 50% chances to be inverted relative to the genes flanking the insertion point. Such events should not be significantly counterselected so long as the duplicated segments remain relatively short. Similarly, single gene deletions must occur at a high rate in the duplicated copies of the merodiploid which results from the duplication/insertion event without affecting cell viability, hence creating intermingled series of genes. Reiteration of this mechanism over evolutionary time scale is probably the major driving force responsible for the existence of gene families in eukaryotic genomes. It is sufficient to account for the partially duplicated structure of the *S. cerevisiae* genome [7–9], although additional mechanisms such as entire chromosome duplication (aneuploidy) are not excluded. In this view, genomic redundancy is a dynamic equilibrium between merodiploid formation and gene deletion, consistent with the distribution of gene families observed in the 13 yeast species studied [29]. Loss of synteny between neighboring genes results from the above mechanism as well as from reciprocal translocations which preserve gene orientation. However, over evolutionary time scales, the number of events of merodiploidization rapidly exceeds the number of translocations. It can be calculated that, within an evolutionary distance such as the one that separates *S. cerevisiae* from *Y. lipolytica* (the most distant species studied here), the number of merodiploidization events becomes similar to the total number of genes, resulting in a complete reshuffling of genomes. The proposed mechanism is consistent with the observation of 'segmental aneuploidy' recently reported to occur at an elevated rate in the genome of *S. cerevisiae* [30]. The frequency of the phenomenon is such that the distinction between orthologous genes and paralogous genes promptly vanishes over increasing evolutionary distances even if a single functional copy of the gene is present in the species examined.

Acknowledgements: We thank our colleagues from the Unité de Génétique moléculaire des levures for fruitful discussions and H. Feldmann and A. Goffeau for careful reading of the manuscript. B.D. is a member of Institut Universitaire de France.

References

[1] Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L. and Nelson, C.R. et al. (2000) Science 287, 2204–2215.
[2] Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M. and Alloni, G. et al. (1997) Nature 390, 249–256.
[3] Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T. and Burland, V. et al. (1997) Science 277, 1453–1462.
[4] Cole, S.T., Brosch, R., Parkhill, J., Garnier, T. and Chrurcher, C. et al. (1998) Nature 393, 537–544.
[5] Goffeau, A., Barrell, B., Bussey, H., Davis, R.W. and Dujon, B. et al. (1996) Science 274, 546–567.
[6] Goffeau, A., Aert, R., Agostini-Carbone, M.L., Ahmed, A. and Aigle, M. et al. (1997) Nature 387 (Suppl.), 5–105.
[7] Mewes, H.W., Albermann, K., Bähr, M., Frishman, D. and Gleissner, A. et al. (1997) Nature 387 (Suppl.), 7–65.
[8] Wolfe, K.H. and Shields, D.C. (1997) Nature 387, 708–713.
[9] Coissac, E., Maillier, E. and Netter, P. (1997) Mol. Biol. Evol. 14, 1062–1074.
[10] Seoighe, C. and Wolfe, K.H. (1999) Gene 238, 253–261.
[11] Artiguenave, F., Wincker, P., Brottier, P., Duprat, S., Jovelin, F., et al. FEBS Lett. 487, 13–16 (this issue).
[12] Tekaia, F., Blandin, G., Malpertuy, A., Llorente, B., Durrens, P., et al. FEBS Lett. 487, 17–30 (this issue).
[13] Bon et al., FEBS Lett. 487, 37–41 (this issue).
[14] Bon et al., FEBS Lett. 487, 42–46 (this issue).
[15] Casaregola et al., FEBS Lett. 487, 47–51 (this issue).
[16] de Montigny et al., FEBS Lett. 487, 52–55 (this issue).
[17] Neuvéglize et al., FEBS Lett. 487, 56–60 (this issue).
[18] Malpertuy et al., FEBS Lett. 487, 61–65 (this issue).
[19] Bolotin-Fukuhara et al., FEBS Lett. 487, 66–70 (this issue).
[20] Llorente et al., FEBS Lett. 487, 71–75 (this issue).
[21] Blandin et al., FEBS Lett. 487, 76–81 (this issue).
[22] Lépingle et al., FEBS Lett. 487, 82–86 (this issue).
[23] de Montigny et al., FEBS Lett. 487, 87–90 (this issue).
[24] Blandin et al., FEBS Lett. 487, 91–94 (this issue).
[25] Casaregola et al., FEBS Lett. 487, 95–100 (this issue).
[26] Dujon, B. (1996) Trends Genet. 12, 263–270.
[27] Souciet, J.-L., Aigle, M., Artiguenave, F., Blandin, G., Bolotin-Fukuhara, M., et al., FEBS Lett. 487, 3–12 (this issue).
[28] Malpertuy A. et al., FEBS Lett. 487, 113–121 (this issue).
[29] Llorente et al., FEBS Lett. 487, 122–133 (this issue).
[30] Hughes, T.R., Roberts, C.J., Dai, H., Jones, A.R. and Meyer, M.R.M. et al. (2000) Nature Genet. 25, 333–337.