Highly Variable Rates of Genome Rearrangements between Hemiascomycetous Yeast Lineages

Gilles Fischer[1*], Eduardo P. C. Rocha[2,3], Frédéric Brunet[4], Massimo Vergassola[5], Bernard Dujon[1]

1 Unité de Génétique Moléculaire des Levures (CNRS URA 2171, UFR927 Université Pierre et Marie Curie), Département de Structure et Dynamique des Génomes, Institut Pasteur, Paris, France, **2** Unité Génétique des Génomes Bactériens (CNRS URA 2171), Département de Structure et Dynamique des Génomes, Institut Pasteur, Paris, France, **3** Atelier de Bioinformatique, Université Pierre et Marie Curie, Paris, France, **4** Laboratoire de Biologie Moléculaire de la Cellule (CNRS UMR 5161, INRA LA 1237), Ecole Normale Supérieure de Lyon, Lyon, France, **5** Unité de Génétique in silico (CNRS URA 2171), Département de Structure et Dynamique des Génomes, Institut Pasteur, Paris, France

Hemiascomycete yeasts cover an evolutionary span comparable to that of the entire phylum of chordates. Since this group currently contains the largest number of complete genome sequences it presents unique opportunities to understand the evolution of genome organization in eukaryotes. We inferred rates of genome instability on all branches of a phylogenetic tree for 11 species and calculated species-specific rates of genome rearrangements. We characterized all inversion events that occurred within synteny blocks between six representatives of the different lineages. We show that the rates of macro- and microrearrangements of gene order are correlated within individual lineages but are highly variable across different lineages. The most unstable genomes correspond to the pathogenic yeasts *Candida albicans* and *Candida glabrata*. Chromosomal maps have been intensively shuffled by numerous interchromosomal rearrangements, even between species that have retained a very high physical fraction of their genomes within small synteny blocks. Despite this intensive reshuffling of gene positions, essential genes, which cluster in low recombination regions in the genome of *Saccharomyces cerevisiae,* tend to remain syntenic during evolution. This work reveals that the high plasticity of eukaryotic genomes results from rearrangement rates that vary between lineages but also at different evolutionary times of a given lineage.

Citation: Fischer G, Rocha EPC, Brunet F, Vergassola M, Dujon B (2006) Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genet 2(3): e32.

Introduction

The class of Hemiascomycete comprises several hundreds of simple fungi, the vast majority of which are yeasts. Among them, there are few opportunistic pathogens such as *Candida albicans* that are responsible for the majority of all forms of candidiasis [1]. *Debaryomyces hansenii*, a cryotolerant species that tolerates high salinity levels, is a close relative to *C. albicans* (Figure 1). Although considered a nonpathogenic yeast, *D. hansenii* and its anamorph *Candida famata* have been associated with one case of bone infection and several cases of superficial infections [2,3]. The second causative agent of human candidiasis is *Candida glabrata*. In spite of its genera name, this species is phylogenetically more closely related to *Saccharomyces cerevisiae* than to *C. albicans* (Figure 1). The level of genetic diversity between yeast species is often unsuspected. For instance, the average protein divergence of more than 50% found between *S. cerevisiae* and *Yarrowia lipolytica*, an alkane-using yeast, reveals that Hemiascomycetes are molecularly as diverse as the entire phylum of chordates [4]. The level of protein divergence within the *Saccharomyces* "sensu stricto" complex (see Figure 1), whose different members are thought to be in the early stages of the speciation process, already compares to the one found between mammals [4–7].

A high level of synteny conservation of more than 98% has been reported between the genomes of the *Saccharomyces* "sensu stricto" species [7–9]. The term synteny originally referred to gene loci that map on the same chromosome, but it is now commonly used to design chromosomal regions in different genomes that share a common evolutionary origin. In other words, two regions are named syntenic when multiple consecutive genes are found in a (nearly) conserved order between the two genomes considered. Homologous chromosomes between the *Saccharomyces* "sensu stricto" species are almost colinear, differing from each other by only few translocations and large inversions that cause macrosynteny breakpoints (i.e., the simultaneous relocalization or reorientation of many genes). It has also been shown that the rate of formation of translocations is not constant in this group of species, indicating that bursts of rearrangements might have occurred at some points of their evolutionary history [10]. In fact, the majority of gene order changes between these species corresponds to microsynteny breakpoints created by the alternative loss of duplicates in

Editor: James Haber, Brandeis University, United States of America

Received September 20, 2005; **Accepted** January 26, 2006; **Published** March 10, 2006

A previous version of this article appeared as an Early Online Release on January 26, 2006 (DOI: 10.1371/journal.pgen.0020032.eor).

DOI: 10.1371/journal.pgen.0020032

Abbreviations: GOC, gene order conservation; GOL, gene order loss

* To whom correspondence should be addressed. E-mail: fischer@pasteur.fr

Synopsis

The yeast *Saccharomyces cerevisiae* has proved to be a very powerful model organism for deciphering the molecular functioning of our cells. It also is the first eukaryote (the domain of life that includes human) whose genome has been completely sequenced in 1996. There are hundreds of species of yeast covering a tremendous genetic diversity. Almost ten years after the release of the first complete eukaryotic genome sequence, yeasts are still at the forefront of the field of genomics as they represent the monophyletic group of eukaryotes for which the largest number of complete genome sequences has been unveiled. The comparative analysis of their organization now provides an exquisite tool to dissect the mechanistic underpinnings of the process of genome evolution. This study reveals the extraordinary plasticity of the eukaryotic genomes. It also shows that genomes get rearranged at different rates both between the different lineages but also at the different evolutionary times of a given lineage. Finally, in spite of their distant phylogenetic relationship, pathogenic yeasts such as the two main causatives of human candidiasis, *Candida albicans* and *Candida glabrata* species, harbor the most unstable genomes of all lineages.

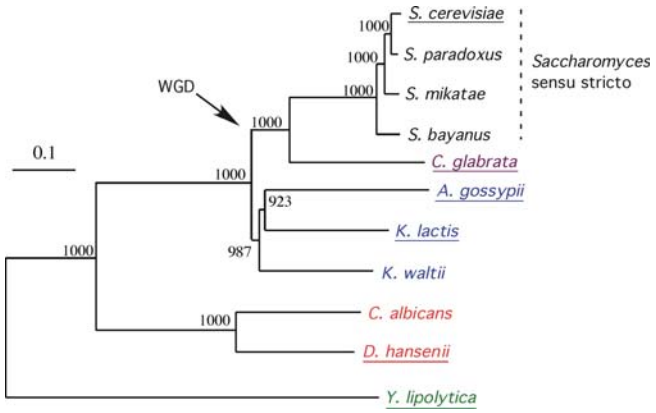

Figure 1. Phylogeny of Hemiascomycete Species

The phylogenetic tree was built on the concatenated sequences of 25 proteins having clear orthologs in all of the 11 studied species. Bootstraps of the tree are given at the branches (out of 1,000). Species whose names are underlined correspond to fully sequenced genomes for which the number of supercontigs is identical to the number of chromosomes. The arrow indicates the place where the whole genome duplication (WGD) event occurred in the tree.

DOI: 10.1371/journal.pgen.0020032.g001

the different species [9]. A whole genome duplication event, that occurred after the divergence of the *Saccharomyces* and *Kluyveromyces* lineages [4,11–13] (see Figure 1), resulted in the sudden doubling of all gene copies. The return to the diploid state was accompanied by a massive loss of nearly 90% of the duplicated gene copies, leaving only one copy of each gene in each genome. The differential loss of the two copies in two different species has lead to microsynteny breakpoints between their genomes.

At broader evolutionary distances, the coincidence between chromosome maps is blurred by the accumulation of numerous interchromosomal rearrangements [4]. However, little is known about the degree and the rate of chromosomal reorganization in the different lineages. Nearly complete genome sequences are now available for numerous yeast species [4,7,12–17] so we assessed the influence of both macro- and microrearrangements onto the evolution of the genomic architectures of representative yeast species covering the entire Hemiascomycete phylum. In this study we used the complete (or nearly complete) genome sequences of 11 yeast species to quantify the rates of macrorearrangements by measuring the level of gene order conservation between pairs of species. We also identified all inversion events that occurred within synteny regions shared between the genomes of six fully sequenced species. We discovered a tremendous level of chromosomal reorganization outside of the *Saccharomyces* "sensu stricto" species and showed that different rates of both macro- and microrearrangements applied in the different yeast lineages.

Results/Discussion

Rates of Genome Instability

To quantify the relative rates of rearrangements in the different lineages we first computed a gene order conservation (GOC) index [18,19] between the 11 yeast species for which the phylogeny is presented in Figure 1. Putative orthologs were identified for all pairs of genomes and two pairs of orthologs are in a "relation of conserved order" if they are separated by less than four intervening genes in both genomes (Materials and Methods). GOC is the proportion of such syntenic pairs among the total number of orthologs between the two compared genomes. Hence, GOC varies between 0 (no pair of syntenic orthologs) and 1 (complete GOC). GOCs were calculated for the 55 pairs of species ($[n(n-1)]/2$, with $n = 11$) and a phylogenetic tree derived from the concatenated sequences of 25 orthologous proteins in the 11 genomes was constructed (see Materials and Methods) to estimate the evolutionary distances between all pairs of species (Figure 1). The tree topology was further supported by using the concatenated sequences of 16 ribosomal proteins to construct a second tree. The resulting topology is completely identical to that described in Figure 1 (not shown). Naturally, GOCs arising from the comparisons of closely related species are closer to one than the ones between distant species (Table 1). Reciprocally, the proportions of gene order loss (GOL = 1 − GOC) increase along with the phylogenetic distances (Table 1). We reasoned that each of the 55 interspecies GOL values results from the sum of all events of genome rearrangements that occurred in the different branches separating two species from their last common ancestor on the phylogenetic tree. There are 19 branches in total on the phylogenetic tree for which branch-specific GOL can be estimated. We inferred them from the 55 interspecies GOL values presented in Table 1. Branch-specific GOLs were calculated by minimizing the sum, over the 55 pairwise comparisons, of the squared differences between the frequency of observed genome rearrangements (GOL) and the sum of the predicted branch-specific GOL values (Materials and Methods). The resulting branch-specific GOL values are presented in Table 2. We checked that the sum of the branch-specific GOL values between two species gave an estimation (GOL_{est}) close to the original GOL values obtained from the GOC analysis (Table 1). For instance, GOL_{est} between *Y. lipolytica* and *D. hansenii* (Table 1) is the sum of the three branch-specific GOL values x_1, x_2, and x_3 (Table 2). It differs from the original GOL value between *Y. lipolytica* and *D. hansenii* by 1.4% only (Table 1). For

Table 1. GOC and GOL between the 55 Pairwise Comparisons

Comparison	Species	Distance[a]	GOC	GOL	GOLest	Difference (%)
1	*S. paradoxus–S. cerevisiae*	0.029	0.9915	0.0085	0.0085	0.0
2	*S. mikatae–S. paradoxus*	0.046	0.9914	0.0086	0.0079	8.1
3	*S. mikatae–S. cerevisiae*	0.051	0.9916	0.0084	0.0091	8.3
4	*S. bayanus–S. paradoxus*	0.07	0.9906	0.0094	0.0102	8.5
5	*S. bayanus–S. cerevisiae*	0.074	0.9871	0.0129	0.0114	11.6
6	*S. bayanus–S. mikatae*	0.075	0.9899	0.0101	0.0101	0.0
7	*C. albicans–D. hansenii*	0.386	0.738	0.262	0.262	0.0
8	*S. paradoxus–C. glabrata*	0.388	0.8009	0.1991	0.2097	5.3
9	*S. bayanus–C. glabrata*	0.389	0.7855	0.2145	0.1995	7.0
10	*K. waltii–K. lactis*	0.39	0.8695	0.1305	0.1301	0.3
11	*C. glabrata–S. cerevisiae*	0.394	0.7908	0.2092	0.2109	0.8
12	*S. mikatae–C. glabrata*	0.396	0.7913	0.2087	0.2095	0.4
13	*K. waltii–S. mikatae*	0.421	0.8373	0.1627	0.1715	5.4
14	*K. waltii–S. paradoxus*	0.423	0.8285	0.1715	0.1717	0.1
15	*K. waltii–S. bayanus*	0.425	0.8516	0.1484	0.1615	8.8
16	*K. waltii–S. cerevisiae*	0.426	0.807	0.193	0.1729	10.4
17	*K. waltii–A. gossypii*	0.454	0.8905	0.1095	0.1097	0.2
18	*S. paradoxus–K. lactis*	0.454	0.7407	0.2593	0.254	2.0
19	*S. mikatae–K. lactis*	0.456	0.7389	0.2611	0.2538	2.8
20	*K. lactis–S. cerevisiae*	0.457	0.7361	0.2639	0.2552	3.3
21	*S. bayanus–K. lactis*	0.457	0.7666	0.2334	0.2438	4.5
22	*A. gossypii–K. lactis*	0.458	0.8538	0.1462	0.1462	0.0
23	*K. waltii–C. glabrata*	0.479	0.7561	0.2439	0.2346	3.8
24	*C. glabrata–K. lactis*	0.499	0.6862	0.3138	0.3169	1.0
25	*S. mikatae–A. gossypii*	0.509	0.7576	0.2424	0.2335	3.7
26	*S. paradoxus–A. gossypii*	0.513	0.7567	0.2433	0.2337	3.9
27	*S. bayanus–A. gossypii*	0.517	0.7859	0.2141	0.2234	4.3
28	*A. gossypii–S. cerevisiae*	0.52	0.7665	0.2335	0.2348	0.6
29	*A. gossypii–C. glabrata*	0.562	0.6986	0.3014	0.2965	1.6
30	*K. waltii–D. hansenii*	0.866	0.3814	0.6186	0.5897	4.7
31	*C. albicans–K. waltii*	0.87	0.3184	0.6816	0.6455	5.3
32	*D. hansenii–K. lactis*	0.88	0.3532	0.6468	0.6719	3.9
33	*C. albicans–K. lactis*	0.886	0.2832	0.7168	0.7277	1.5
34	*S. paradoxus–D. hansenii*	0.897	0.2789	0.7211	0.7136	1.0
35	*S. mikatae–D. hansenii*	0.899	0.2847	0.7153	0.7134	0.3
36	*D. hansenii–S. cerevisiae*	0.902	0.2684	0.7316	0.7148	2.3
37	*S. bayanus–D. hansenii*	0.906	0.2842	0.7158	0.7034	1.7
38	*C. albicans–S. paradoxus*	0.91	0.1932	0.8068	0.7694	4.6
39	*C. albicans–S. mikatae*	0.91	0.2105	0.7895	0.7692	2.6
40	*C. albicans–S. cerevisiae*	0.912	0.2076	0.7924	0.7706	2.8
41	*C. glabrata–D. hansenii*	0.916	0.2489	0.7511	0.7765	3.4
42	*C. albicans–S. bayanus*	0.919	0.2584	0.7416	0.7592	2.4
43	*C. albicans–C. glabrata*	0.945	0.2068	0.7932	0.8323	4.9
44	*A. gossypii–D. hansenii*	0.946	0.3725	0.6275	0.6516	3.8
45	*C. albicans–A. gossypii*	0.959	0.3247	0.6753	0.7074	4.8
46	*C. albicans–Y. lipolytica*	1.158	0.1195	0.8805	0.8934	1.5
47	*D. hansenii–Y. lipolytica*	1.161	0.1504	0.8496	0.8376	1.4
48	*K. waltii–Y. lipolytica*	1.175	0.2041	0.7959	0.7484	6.0
49	*K. lactis–Y. lipolytica*	1.187	0.1904	0.8096	0.8306	2.6
50	*C. glabrata–Y. lipolytica*	1.242	0.1201	0.8799	0.9352	6.3
51	*S. mikatae–Y. lipolytica*	1.245	0.1343	0.8657	0.8721	0.7
52	*S. bayanus–Y. lipolytica*	1.247	0.1304	0.8696	0.8621	0.9
53	*S. paradoxus–Y. lipolytica*	1.25	0.1172	0.8828	0.8723	1.2
54	*S. cerevisiae–Y. lipolytica*	1.253	0.1218	0.8782	0.8735	0.5
55	*A. gossypii–Y. lipolytica*	1.275	0.1847	0.8153	0.8103	0.6

Pairwise comparisons are ordered by increasing phylogenetic distances.
[a]Distances correspond to the sum of the branch lengths separating two species from their last common ancestor.
DOI: 10.1371/journal.pgen.0020032.t001

all 55 pairwise comparisons, differences between GOL_{est} and the original GOL values are very limited (average 3.2%, min = 0%, max = 11.6%, Table 1). Branch-specific rates of genome rearrangements were obtained by dividing the branch-specific GOL values by the length of their corresponding branches in the tree and centered around one by dividing them by the mean rate (Table 2). Normalized GOL rates for the 19 branches are presented on Figure 2A with a color code indicating rates of gene order rearrangements either higher (red) or lower (blue) than average. It clearly appears that rates of rearrangements remained high in all branches from node 1 to the two main causative agents of human candidiasis, *C.*

Table 2. Branch-Specific Rates of GOL

Branch	Branch-Specific GOL	Branch Length	GOL Rate	Normalized GOL Rate[a]
Node 1–*Y. lipolytica*	$x_1 = 0.4981$	0.743	0.67	1.2
Node 1–node 2	$x_2 = 0.2364$	0.223	1.06	1.89
Node 2–*D. hansenii*	$x_3 = 0.1031$	0.188	0.55	0.98
Node 2–*C. albicans*	$x_4 = 0.1589$	0.198	0.8	1.43
Node 1–node 3	$x_5 = 0.2263$	0.248	0.91	1.63
Node 3–node 4	$x_6 < 10^{-5}$	0.012	$< 10^{-3}$	$< 10^{-2}$
Node 4–*K. waltii*	$x_7 = 0.0239$	0.183	0.13	0.23
Node 4–node 5	$x_8 = 0.0229$	0.01	2.29	4.09
Node 5–*K. lactis*	$x_9 = 0.0833$	0.197	0.42	0.75
Node 5–*A. gossypii*	$x_{10} = 0.0629$	0.261	0.24	0.43
Node 3–node 6	$x_{11} = 0.0744$	0.061	1.22	2.18
Node 6–*C. glabrata*	$x_{12} = 0.1363$	0.216	0.63	1.13
Node 6–node 7	$x_{13} = 0.0632$	0.14	0.45	0.81
Node 7–*S. bayanus*	$x_{14} < 10^{-5}$	0.037	$< 10^{-3}$	$< 10^{-2}$
Node 7–node 8	$x_{15} = 0.0062$	0.011	0.56	1.00
node 8–*S. mikatae*	$x_{16} = 0.0039$	0.024	0.16	0.29
Node 8–node 9	$x_{17} = 0.0004$	0.01	0.04	0.08
Node 9–*S. paradoxus*	$x_{18} = 0.0036$	0.012	0.3	0.54
Node 9–*S. cerevisiae*	$x_{19} = 0.0048$	0.017	0.28	0.51

[a]GOL rates are normalized relatively to the mean GOL rate value.
DOI: 10.1371/journal.pgen.0020032.t002

albicans and *C. glabrata*. Given the external position of *Y. lipolytica* on the phylogenetic tree, only one branch covers the entire lineage from node 1 to the present-day species. Globally, deep branches of the tree that stem out from node 1 present high rearrangement rates. A general decrease of the rates is observed both in the *Saccharomyces* and in the *Kluyveromyces/Ashbya* lineages. Rearrangements have also slowed down in the *D. hansenii*–specific branch since it diverged from *C. albicans*. In addition, the concomitant presence of both stable (in the *Saccharomyces* species) and unstable (in *C. glabrata*) branches subsequent to the ancestral genome duplication suggests that rates of rearrangements have not been influenced by this event. Except for the external species, *Y. lipolytica,* rates of rearrangements can be compared either between species-specific terminal branches only, or globally over the whole evolutionary distance between node 1 and the present-day species. Rates of rearrangements on terminal branches give an estimate of the most recent level of genome instability. The most stable genome corresponds to that of *S. bayanus* followed by those of *K. waltii* and *S. mikatae,* while the most unstable ones correspond to those from the pathogenic yeasts, *C. albicans* and *C. glabrata* (Figure 2A). These two species also present the highest cumulated rates of genome instability when entire evolutionary distances since node 1 are taken into account (Figure 2B). *C. albicans* and *C. glabrata* yeasts occupy narrow ecological niches and in the process of becoming more specialized, their genomes may have accumulated more rearrangements because of selective sweeps or because of lower population sizes leading to less efficient selection onto gene order. It is also possible that the population structures of these pathogenic yeasts that are largely if not entirely clonal due to the lack of mating might contribute to the apparent genome plasticity. At the other end of the scale, the most stable genomes during evolution correspond to the *Kluyveromyces/Ashbya* lineage as well as that of *Y. lipolytica*.

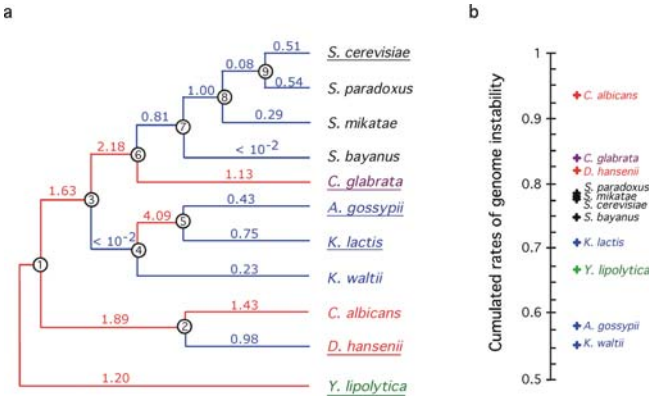

Figure 2. Rates of Genome Instability in Hemiascomycetes

(A) Branch-specific normalized rates of genome rearrangements are indicated either in red or in blue illustrating higher or lower rates than average, respectively. Nodes are numbered from 1 to 9.
(B) Cumulated rates of genome instability correspond to the ratios between the sum of the branch-specific GOL and the phylogenetic distance separating each species from node 1 (Table 2).
DOI: 10.1371/journal.pgen.0020032.g002

Small Inversions within Synteny Blocks

Variable rates of microrearrangements of gene order were also found between the different lineages. Microrearrangements were sought by characterizing all small inversions that occurred within the synteny blocks (Materials and Methods) shared between the six fully sequenced yeast genomes [4,12,14] (underlined species on Figure 1). The total number of inversions between two genomes varies by more than one order of magnitude depending on the species (Table 3). The mean number of inversions per synteny block as well as the mean frequency of gene inversion show that in spite of the largest synteny blocks, *A. gossypii* and *K. lactis* have undergone much fewer inversions than any other couple of species, even fewer than *S. cerevisiae* and *C. glabrata,* which are more closely related. Inversions are found in only 10% of the synteny blocks between *A. gossypii* and *K. lactis,* while there is on average more than one inversion per synteny block between *D. hansenii* and any of the other species. In all comparisons involving *D. hansenii,* the mean frequencies of gene inversion range between 0.42 and 0.65 (i.e., approximately half of the genes have been inverted). Despite a much larger evolutionary distance, the mean frequencies of gene inversions are twice as small for comparisons involving *Y. lipolytica*. Branch-specific expected numbers of inversion per gene were extracted from these pairwise comparisons by minimizing the sum of the relative errors (see Materials and Methods). The *D. hansenii* branch shows, by far, the highest rate of gene inversion (0.351, Figure 3). By comparison, a very low inversion rate applies in the *Y. lipolytica* branch (0.064). A global decrease in the inversion rates occurred in all branches leading to the four other species from their last common ancestor (i.e., from node 2 on Figure 3) and this trend is even more pronounced in the *A. gossypii*– and *K. lactis*–specific branches (originating from node 4) than in the *S. cerevisiae*– and *C. glabrata*–specific branches (originating from node 3). Note that the null value of the branch between nodes 2 and 3 is due to the fact that the numbers of gene inversions in pairwise comparisons involving *S. cerevisiae* are very close to the numbers of inversions in the corresponding pairwise

Table 3. Small Inversions within Synteny Blocks between Six Representative Species of Hemiascomycetes

Group	Species	Synteny Blocks		Edge Inversions			Internal Inversions			All Inversions			
		Total	Mean Number of Genes per Block	Total	Mean Number of Genes per Inversion	Mean Number of Genes per Block	Total	Mean Number of Genes per Inversion	Mean Number of Genes per Block	Total	Edge/ Internal	Mean Number of Inversions per Block	Mean Frequency of Gene Inversion[a]
1	*S. cerevisiae–C. glabrata*	509	7.6	74	2.5	9.9	30	3.7	14.7	104	2,5	0.2	0.08
	S. cerevisiae–K. lactis	608	6.0	90	1.5	8.4	21	2.4	11.8	111	4,3	0.2	0.05
	S. cerevisiae–A. gossypii	537	6.6	100	1.9	7.3	44	2.7	17.4	144	2,3	0.3	0.09
	C. glabrata–K. lactis	670	5.0	98	1.7	6.7	29	2.6	9.1	127	3,4	0.2	0.07
	C. glabrata–A. gossypii	642	5.0	100	1.9	6.8	32	2.0	10.6	132	3,1	0.2	0.08
	K. lactis–A. gossypii	427	8.8	44	1.7	13.8	15	1.6	21.7	59	2,9	0.1	0.03
2	*S. cerevisiae–D. hansenii*	572	2.6	585	1.3	2.8	72	1.4	4.1	657	8,1	1.1	0.57
	C. glabrata–D. hansenii	527	2.5	498	1.2	2.6	50	1.4	3.9	548	10,0	1	0.52
	K. lactis–D. hansenii	605	2.7	633	1.4	3.2	140	1.7	4.8	773	4,5	1.3	0.65
	A. gossypii–D. hansenii	586	2.9	634	1.4	3.2	129	1.7	4.6	763	4,9	1.3	0.65
3	*S. cerevisiae–Y. lipolytica*	256	2.2	148	1.2	2.3	8	1.4	4.0	156	18,5	0.6	0.32
	C. glabrata–Y. lipolytica	251	2.2	147	1.0	2.3	6	1.0	3.6	153	24,5	0.6	0.28
	K. lactis–Y. lipolytica	367	2.4	210	1.3	2.5	14	1.3	4.2	224	15,0	0.6	0.32
	A. gossypii–Y. lipolytica	351	2.3	177	1.2	2.5	12	1.2	4.2	189	14,8	0.5	0.28
	D. hansenii–Y. lipolytica	359	2.3	274	1.1	2.3	18	1.3	3.4	292	15,2	0.8	0.42

Pairwise comparisons are separated in three groups corresponding to increasing phylogenetic distances.
[a]Ratio between the number of inverted segments multiplied by the number of genes in the inverted segments divided by the total number of genes in the synteny blocks.
DOI: 10.1371/journal.pgen.0020032.t003

Figure 3. Branch-Specific Expected Number of Inversions per Gene
The tree topology is deduced from Figure 1. The nodes are numbered from 1 to 4. Calculated inversion numbers are indicated on each branch of the tree.
DOI: 10.1371/journal.pgen.0020032.g003

comparisons involving *C. glabrata* (see Table 3). These variable rates of microrearrangements of gene order are fully consistent with the relative rates of macrorearrangements characterized above (Figure 2). Previous works based on partial sequences of yeast genomes pointed out that the proportions of locally inverted genes remained rare over a relatively long evolutionary distance (less than 5% between the *Saccharomyces* and the *Kluyveromyces* spp.), [8] but become prominent over longer evolutionary distances (30% to 40% between *S. cerevisiae* and *C. albicans* or *D. hansenii*) [8,20]. Here we show that this difference is not solely due to the phylogenetic distances but relies on an accelerated rate of rearrangement in the *C. albicans/D. hansenii* lineage as compared to much slower rates in the *Kluyveromyces* and *Saccharomyces* groups of species (Figure 3).

Inversions are categorized as "internal" when the whole inverted segment is comprised within a synteny block or as "edge" when one end of the inverted segment coincides with the end of the block. In all 15 pairwise comparisons, edge inversions are far more frequent than internal ones and the ratio between these two categories increases with the phylogenetic distances (Table 3). In addition, the length of the synteny blocks that contain edge inversions is on average smaller than the length of blocks carrying internal inversions only. These observations suggest that edge inversions could correspond in fact to internal inversion events that were subsequently interrupted by a synteny breakpoint. Indeed, synteny breakpoints occurring within an inverted segment would not only produce two new edge inversions but would also result in a decrease of the size of the two resulting synteny blocks. The small size of the edge-containing blocks of synteny as well as the increasing proportions of edge inversions at larger phylogenetic distances are fully compatible with such a scenario of formation of edge inversions. This also implies that the original events of inversion were probably larger than the size of the remaining inverted segments at the edge of the synteny blocks. The average length of the original events of inversion was estimated from the sole internal events. It varies from one gene, for comparisons between distant species, to 3.7 genes between the two closest species (*S. cerevisiae* versus *C glabrata*, Table 3). Despite a possible underestimation of the inversion sizes

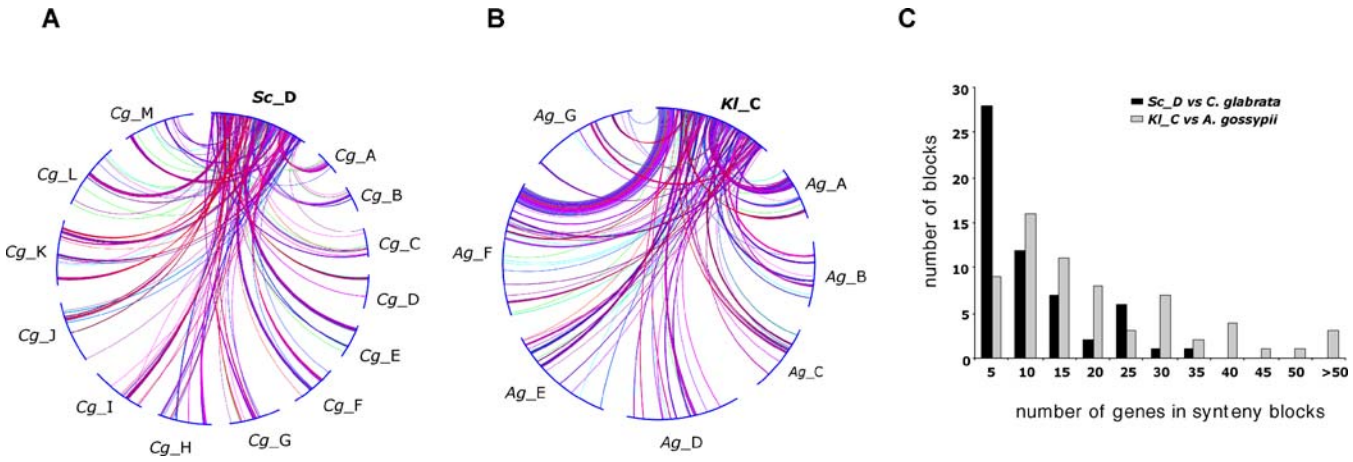

Figure 4. Chromosomal Map Reorganization between Related Species

The circular representation is adapted from [25].

(A) Chromosome D from *S. cerevisiae* (*Sc*_D) is represented in a circle with the 13 chromosomes from *C. glabrata* (*Cg*_A to M). Each line joins two orthologs and the color of the lines represents the percentage of similarity between orthologous gene products (green $\leq$ 50% $\leq$ cyan $\leq$ 60% $\leq$ blue $\leq$ 70% $\leq$ magenta $\leq$ 80% $\leq$ dark magenta $\leq$ 90% $\leq$ red).

(B) Same representation between chromosome C of *K. lactis* (*Kl*_C) and the seven chromosomes from *A. gossypii* (*Ag*_A to G). The diagram shows large uninterrupted regions of conserved synteny between *Kl*_C and *Ag*_F or *Ag*_A, while no such conservation is visible between *Sc*_D and any of the *C. glabrata* chromosomes.

(C) Distribution of the length of the corresponding syntenic blocks between *Sc*_D and *C. glabrata* (black) and *Kl*_C and *A. gossypii* (gray).

DOI: 10.1371/journal.pgen.0020032.g004

within the *S. cerevisiae/C. glabrata* group (due to the relatively small size of the synteny blocks), inversions appear to be significantly longer between these two species than between *K. lactis* and *A. gossypii* (mean length of 3.7 $\pm$ 0.5 genes versus 1.6 $\pm$ 0.4 genes, respectively).

Reorganization of the Chromosomal Maps

The higher genomic stability in the *K. lactis/A. gossypii* lineage as compared to the *S. cerevisiae/C. glabrata* group of species is directly observable at the chromosomal level. In spite of its genera name, *C. glabrata* is the closest relative to the *Saccharomyces* clade with a fully sequenced genome. A slightly larger phylogenetic distance separates *K. lactis* from *A. gossypii* than *S. cerevisiae* from *C. glabrata*. However, chromosomal colinearity is better preserved between the former pair. Large uninterrupted chromosomal regions are still recognizable between some of the *K. lactis* and *A. gossypii* chromosomes, while any individual chromosome of *S. cerevisiae* is scattered into small and intermingled pieces onto virtually all of the chromosomes of *C. glabrata* (Figures 4, S1, and S2). By contrast, very few macrorearrangements have disrupted the chromosomal colinearity between the genomes of the *Saccharomyces* "sensu stricto" species [7,10]. This underlines an important evolutionary leap in the level of chromosomal reorganization between the "sensu stricto" species and *C. glabrata*. Interestingly, in spite of the important level of chromosome map reshuffling, a very high fraction of the genomes of *S. cerevisiae* and *C. glabrata* are conserved within synteny blocks. The total length spanned by the synteny blocks between *S. cerevisiae* and *C. glabrata* represents 88% of the physical length of these genomes. This proportion rises to 93% when subtelomeric regions are excluded from the analysis, as no conservation of synteny was found between these regions. Although almost the entire genomes of these species are comprised within small synteny blocks, the global chromosomal colinearity has been completely destroyed by the accumulation of numerous and overlapping interchromosomal rearrangements. This clearly shows that loss of synteny primarily results from the accumulation of chromosomal rearrangements rather than from sequence divergence between orthologous regions that would impede recognition of their common evolutionary origin. It is also notable that, consistent with a higher level of chromosomal reorganization in the *S. cerevisiae/C. glabrata* than in the *K. lactis/A. gossypii* lineages, the size of the syntenic blocks is on average smaller in the former than in the latter (Table 3, Figure S3). The smaller size of the synteny blocks in *S. cerevisiae/C. glabrata* is also attributable to the massive gene loss that occurred subsequent to the whole genome duplication event [11], whereas the corresponding regions in the *K. lactis* genome have retained virtually all genes.

Constraints on Gene Order Changes

Genome dynamics results from the accumulation of both micro- and macrorearrangements and leads to an apparent randomization of gene order between distantly related yeast species. However, there is some evidence that gene order is under selection in eukaryotes [21]. In *S. cerevisiae,* essential genes tend to be clustered and these clusters are in regions of low recombination rates [22]. If gene order is constrained by natural selection, synteny breaks within such clusters would be counter-selected. We determined the proportions of genes in synteny blocks that are essential (i.e., those for which the knockout is lethal in *S. cerevisiae*) between representative species of the different lineages. This proportion increases with the phylogenetic distance between species (Figure 5A). This trend is even stronger for essential genes concomitantly conserved in synteny between three, four, or the five compared species. This suggests that essential genes tend to remain clustered within genomes during evolution. However, essential genes evolve more slowly than nonessential ones [23]. Therefore, this increase could be due, at least partly, to the better sequence conservation of essential genes that would result in a higher proportion of such genes among all

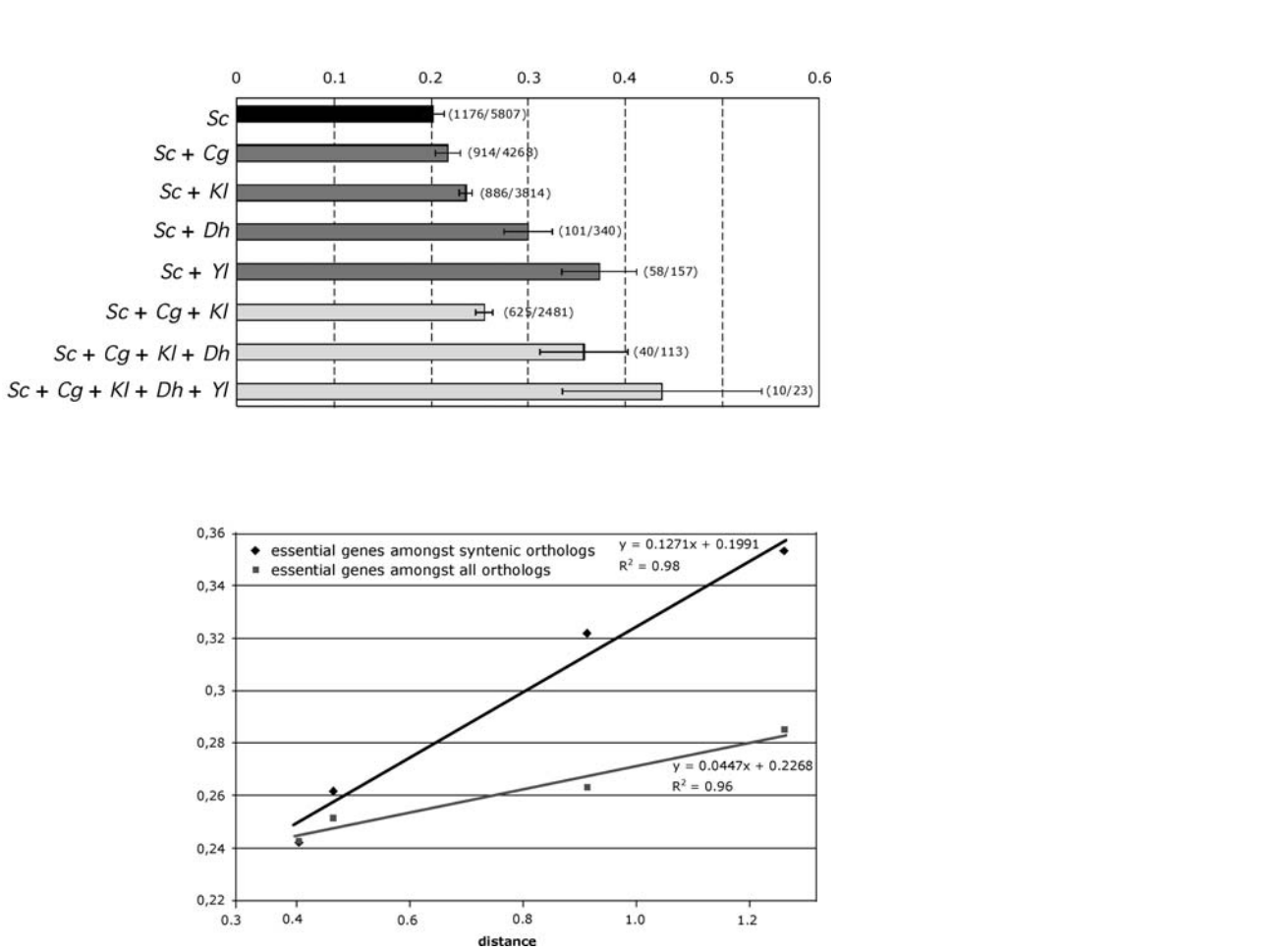

Figure 5. Proportion of Essential Genes within Conserved Synteny Blocks

(A) The black bar represents the proportion of essential genes in the genome of *S. cerevisiae (Sc),* as defined in the Comprehensive Yeast Genome Database (http://mips.gsf.de/genre/proj/yeast). The relative proportions of orthologs to these genes among the total number of genes comprised within the syntenic blocks with the genomes of *C. glabrata (Cg), K. lactis (Kl), D. hansenii (Dh),* and *Y. lipolytica (Yl)* are represented by dark gray bars. Proportions of essential genes concomitantly conserved within synteny blocks in three, four, and five species are indicated by light gray bars. Error bars represent two standard deviations, and the number of genes considered in each case is indicated in parentheses.

(B) Comparison of the proportions of essential genes among syntenic orthologs and among all orthologs at increasing phylogenetic distances. Phylogenetic distances between *S. cerevisiae* and *C. glabrata, S. cerevisiae* and *K. lactis, S. cerevisiae* and *D. hansenii,* and *S. cerevisiae* and *Y. lipolytica* are reported on the *x*-axis.

DOI: 10.1371/journal.pgen.0020032.g005

identified orthologs. We plotted the proportion of essential genes among all orthologs for the four pairwise comparisons and showed that it increases along with the phylogenetic distance but to a significantly lower rate than the proportion of essential genes in synteny (Figure 5B). Altogether, these results show that essential genes tend to remain adjacent during evolution, and this trend remains observable even at very large evolutionary distances where genomes have been massively reshuffled by chromosomal rearrangements.

Future Prospects

This work shows that genome dynamics varies very significantly between related yeast species. However, within each genome macro- and microrearrangements occur at similar relative rates. In higher eukaryotes, a slow rate of genome reorganization has been characterized in human compared to that of rodent, and an even slower rate has been characterized in chicken [24]. A very slow rate of interchromosomal rearrangements has also been described for the very compact genome of *Tetraodon* [25]. Rates of chromosome evolution have also been compared between eight mammalian species [32]. In addition to variations between the different orders, the authors characterized a global increase in breakage rates after the Cretaceous-Tertiary boundary. These results are fully consistent with our findings that rearrangement rates not only vary between different yeast lineages but also at different evolutionary times of a given lineage. It remains to be understood why such differences exist. One could invoke intrinsic reasons to explain why some genomes can be less stable than others (e.g., because they could contain a higher proportion of repetitive sequences [transposable elements, duplicated genes] and/or because DNA damages would be less efficiently repaired). Moreover, selection is likely to act differently in different genomes. In this case, rearrangements could be fixed more frequently in yeasts with smaller effective population sizes, as it is probably the case for the pathogenic ones.

Materials and Methods

Orthology searches and GOC. Genes were regarded as putative orthologs in pairwise comparisons if their products were reciprocal best-hits with at least 40% similarity in sequence and their sequences were less than 30% different in length, as in [18]. For the genomes of *S. mikatae, S. paradoxus, S. bayanus,* and *K. waltii,* where the annotations were not available, we mapped the genes within the contigs using FASTA searches [26]. We retained only the hits that were the best matches both in terms of score and E-value (and this smaller than E-10). Genome sequences were downloaded from http://www.yeastgenome.org for *S. cerevisiae,* http://cbi.labri.fr/Genolevures/index.php for *C. glabrata, K. lactis, D. hansenii,* and *Y. lipolytica,* http://agd.unibas.ch for *A. gossypii*, http://www.broad.mit.edu/seq/YeastDuplication for *K. waltii,* http://www.genome.wustl.edu/projects/yeast for *S. paradoxus, S. mikatae,* and *S. bayanus,* and http://www.candidagenome.org for *C. albicans.*

The GOC index was adapted from previous works [18,19] by allowing the presence of intervening genes between syntenic pairs of orthologs in order to recover most relations of GOC that would otherwise be lost due to the massive gene loss that occurred after the whole genome duplication event. After some experimentation, the upper limit was set to four intervening genes, as larger neighborhoods typically led to a GOC less than 1% higher but a smaller statistical confidence. Synteny blocks were defined as series of neighboring syntenic pairs of orthologs separated by less than ten intervening genes in both compared genomes.

Phylogenetic analysis. The distance matrix between the species was computed using maximum likelihood with Tree-Puzzle [27] (JTT + $\Gamma(8)$ model). We built two phylogenetic trees, one using 25 randomly chosen highly conserved ubiquitous genes (orthologs to *YAL044w-a, YAL016w, YBL057c, YBR025c, YBR127c, YCR009c, YCR011c, YDL031w, YDR140w, YDR449c, YER068w, YER110c, YER141w, YGR096w, YGR235c, YIL043c, YJR010w, YJR096w, YKL184w, YKL134c, YKL103c, YLR351c, YNL062c, YNR007c,* and *YPR051w*), and another using 16 ubiquitous ribosomal proteins. Both trees had exactly the same topology. Trees were built from the distance matrix using BIONJ [27] and the robustness of the branches was assessed with 1,000 bootstraps using BOOTSEQ and CONSENSE from the PHYLIP package [28].

Inversions. We searched for local inversions within synteny blocks using the DERANGE algorithm [29]. This program is intended to find the most economical number of moves (inversions, transpositions, and transversions) to transform an ordered and orientated sequence of n genes in the first genome to the actual order of the corresponding n orthologs in the second genome. When all types of move, inversions (e.g., a sequence of four genes, A B C D, becomes A – B C D, with "–" denoting a switch of coding strand for gene B), transpositions (e.g., A B C D becomes A C D B), or transversions (e.g., A B C D becomes A C D –B) are assigned the same weight, inversions appear to be far more frequent than transpositions or transversions (65% to 85% of the moves). Transversions and transpositions were massively penalized as all gene order/orientation changes observed within synteny blocks can easily be explained by the only mean of inversions, even if this tends to increase the total number of moves (from 10% to 25% depending on the compared species). The few remaining synteny blocks still containing transposition or transversion events were analyzed with the GRIMM-Synteny program [30] to reconstruct, by inversions only, the gene order/orientation changes that occurred in the corresponding regions.

Calculation of branch-specific values. Branch-specific GOL values, x_j, were calculated by minimizing the following equation

$$L = \sum_{i=1}^{55} \left(\sum_{j=1}^{19} b_{i,j} x_j - GOL_i \right)^2 \quad (1)$$

where $b_{i,j}$ is a Boolean variable indicating if the branch-specific GOL variable x_j (Table 2) contributes to the i-th interspecies comparison and GOL_i are the values measured in the interspecies comparisons ($GOL_i = 1 - GOC_i$). For example, in the 47th comparison between *Y. lipolytica* and *D. hansenii* (Table 1), $b_{47,1} = b_{47,2} = b_{47,3} = 1$ and all the others are zero. The resulting optimization problem is quadratic, with the constraints that all variables x_j must be positive. It is easy to verify the convexity of the quadratic form L to be minimized (positive Hessian), ensuring the uniqueness of the minimum, which is computed solving the linear Karush-Kuhn-Tucker optimality conditions by matrix inversion [31].

Expected numbers of inversions per gene on each of the nine branches of the phylogenetic tree on Figure 3 were inferred likewise by minimizing the sum, over the 15 pairwise comparisons, of the squared differences between the number of predicted inversions and the number of inversions observed in the pairwise comparisons. For each pair of species, the former is given by the sum of the expected number of inversions per gene along the branches separating the two species.

Supporting Information

Figure S1. Chromosomal Map Reorganization between *S. cerevisiae* and *C. glabrata*

Each chromosome from *S. cerevisiae* (SACE__A to P) is represented in a circle with the 13 chromosomes from *C. glabrata* (CAGL__A to M). Each line joins two orthologs and the color of the lines represents the percentage of similarity between orthologous gene products (green $\leq$ 50% $\leq$ cyan $\leq$ 60% $\leq$ blue $\leq$ 70% $\leq$ magenta $\leq$ 80% $\leq$ dark magenta $\leq$ 90% $\leq$ red).

Found at DOI: 10.1371/journal.pgen.0020032.sg001 (1.1 MB PDF).

Figure S2. Chromosomal Map Reorganization between *K. lactis* and *A. gossypii*

Each chromosome from *K. lactis* (KLLA__A to F) is represented in a circle with the seven chromosomes from *A. gossypii* (ASGO__A to G). Each line joins two orthologs and the color of the lines represents the percentage of similarity between orthologous gene products (green $\leq$ 50% $\leq$ cyan $\leq$ 60% $\leq$ blue $\leq$ 70% $\leq$ magenta $\leq$ 80% $\leq$ dark magenta $\leq$ 90% $\leq$ red).

Found at DOI: 10.1371/journal.pgen.0020032.sg002 (630 KB PDF).

Figure S3. Distribution of the Length of the Syntenic Blocks between *S. cerevisiae* and *C. glabrata* (Black Bars) and between *K. lactis* and *A. gossypi* (Gray Bars)

Found at DOI: 10.1371/journal.pgen.0020032.sg003 (23 KB PDF).

Acknowledgments

We thank Emmanuelle Fabre, Ingrid Lafontaine, Ed J. Louis, and Bertrand Llorrente for critical reading of the manuscript, and the members of the Génolevures network as well as our colleagues from the Unité de Génétique Moléculaire des Levures for fruitful discussions.

Author contributions. GF conceived and designed the experiments. GF, EPCR, FB, and MV performed the experiments. GF, EPCR, MV, and BD analyzed the data. BD contributed reagents/materials/analysis tools. GF wrote the paper.

Funding. This work was supported by a grant from the Association pour la Recherche sur le Cancer (3266) and by the CNRS (GDR2354 "Génolevures II"). BD is a member of the Institut Universitaire de France.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Pfaller MA, Jones RN, Doern GV, Sader HS, Messer SA, et al. (2000) Bloodstream infections due to *Candida* species: SENTRY antimicrobial surveillance program in North America and Latin America, 1997–1998. Antimicrob Agents Chemother 44: 747–751.
2. Nishikawa A, Tomomatsu H, Sugita T, Ikeda R, Shinoda T (1996) Taxonomic position of clinical isolates of *Candida famata.* J Med Vet Mycol 34: 411–419.
3. Wong B, Kiehn TE, Edwards F, Bernard EM, Marcove RC, et al. (1982) Bone infection caused by *Debaryomyces hansenii* in a normal host: A case report. J Clin Microbiol 16: 545–548.
4. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, et al. (2004) Genome evolution in yeasts. Nature 430: 35–44.
5. Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A 95: 9407–9412.
6. Malpertuy A, Tekaia F, Casaregola S, Aigle M, Artiguenave F, et al. (2000) Genomic exploration of the hemiascomycetous yeasts: 19. Ascomycetes—Specific genes. FEBS Lett 487: 113–121.
7. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254.
8. Llorente B, Malpertuy A, Neuveglise C, de Montigny J, Aigle M, et al. (2000)

Genomic exploration of the hemiascomycetous yeasts: 18. Comparative analysis of chromosome maps and synteny with *Saccharomyces cerevisiae*. FEBS Lett 487: 101–112.
9. Fischer G, Neuveglise C, Durrens P, Gaillardin C, Dujon B (2001) Evolution of gene order in the genomes of two related yeast species. Genome Res 11: 2009–2019.
10. Fischer G, James SA, Roberts IN, Oliver SG, Louis EJ (2000) Chromosomal evolution in *Saccharomyces*. Nature 405: 451–454.
11. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
12. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, et al. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304: 304–307.
13. Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428: 617–624.
14. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science 274: 546, 563–567.
15. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, et al. (2004) The diploid genome sequence of *Candida albicans*. Proc Natl Acad Sci U S A 101: 7329–7334.
16. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science 301: 71–76.
17. Liti G, Louis EJ (2005) Yeast evolution and comparative genomics. Annu Rev Microbiol 59: 135–153.
18. Rocha EP (2003) DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet 19: 600–603.
19. Rocha EP (2005) Inference and analysis of the relative stability of bacterial chromosomes. Mol Biol Evol. 23: 513–522.
20. Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, et al. (2000) Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci U S A 97: 14433–14437.

Note Added in Proof

Reference 32 was added while this paper was in proofs stage; as a result, it is cited out of order in the text.

21. Hurst LD, Pal C, Lercher MJ (2004) The evolutionary dynamics of eukaryotic gene order. Nat Rev Genet 5: 299–310.
22. Pal C, and Hurst LD (2003) Evidence for co-evolution of gene order and recombination rate. Nat Genet 33: 392–395.
23. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, et al. (2005) Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A 102: 5483–5488.
24. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G (2005) Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Res 15: 98–110.
25. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431: 946–957.
26. Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183: 63–98.
27. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.
28. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6 [computer program]. Distributed by the author. Seattle: Department of Genome Sciences, University of Washington.
29. Blanchette M, Kunisawa T, Sankoff D (1996) Parametric genome rearrangement. Gene 172: GC11–GC17.
30. Pevzner P, Tesler G (2003) Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. Genome Res 13: 37–45.
31. Bazaraa M, Sherali D, Shetty C (1992) Nonlinear programming: Theory and algorithms. New York: Wiley. 640 p.
32. Murphy WJ, Larkin DM, Everts-van der Wind A, Bourque G, Tesler G, et al. (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309: 613–617.