Population Genomics Reveals Chromosome-Scale Heterogeneous Evolution in a Protoploid Yeast

Anne Friedrich,[1] Paul Jung,[1] Cyrielle Reisser,[1] Gilles Fischer,*[,2,3] and Joseph Schacherer*[,1]

[1]Department of Genetics, Genomics and Microbiology, Université de Strasbourg/CNRS, UMR7156, Strasbourg, France

[2]Sorbonne Universités, UPMC Université Paris 06, UMR 7238, Biologie Computationnelle et Quantitative, Paris, France

[3]CNRS, UMR7238, Biologie Computationnelle et Quantitative, Paris, France

***Corresponding author:** E-mail: gilles.fischer@upmc.fr; schacherer@unistra.fr.

Associate editor: Jianzhi Zhang

Abstract

Yeast species represent an ideal model system for population genomic studies but large-scale polymorphism surveys have only been reported for species of the *Saccharomyces* genus so far. Hence, little is known about intraspecific diversity and evolution in yeast. To obtain a new insight into the evolutionary forces shaping natural populations, we sequenced the genomes of an expansive worldwide collection of isolates from a species distantly related to *Saccharomyces cerevisiae*: *Lachancea kluyveri* (formerly *S. kluyveri*). We identified 6.5 million single nucleotide polymorphisms and showed that a large introgression event of 1 Mb of GC-rich sequence in the chromosomal arm probably occurred in the last common ancestor of all *L. kluyveri* strains. Our population genomic data clearly revealed that this 1-Mb region underwent a molecular evolution pattern very different from the rest of the genome. It is characterized by a higher recombination rate, with a dramatically elevated A:T → G:C substitution rate, which is the signature of an increased GC-biased gene conversion. In addition, the predicted base composition at equilibrium demonstrates that the chromosome-scale compositional heterogeneity will persist after the genome has reached mutational equilibrium. Altogether, the data presented herein clearly show that distinct recombination and substitution regimes can coexist and lead to different evolutionary patterns within a single genome.

Key words: **population genomics, chromosome evolution, yeast.**

Introduction

Detailed examination of the patterns of genetic variation is the first step toward a broader understanding of the forces that shape genomic architecture and evolution. With the advent of high-throughput technologies for sequencing, the complete description of genetic variation that occurs in populations is foreseeable but yet far from being reached. Large-scale polymorphism surveys were reported for different model organisms including *Caenorhabditis elegans*, *Arabidopsis thaliana*, and *Homo sapiens* (Cao et al. 2011; 1000 Genomes Project Consortium 2012; Andersen et al. 2012). Spanning a broad evolutionary distance, *Saccharomycotina* yeasts with their compact and small genomes represent an ideal phylum for parallel intraspecific genetic diversity explorations (Dujon 2010; Liti and Schacherer 2011). To date, only the evolution within the *Saccharomyces* genus has been investigated (Liti et al. 2009; Schacherer et al. 2009; Hittinger et al. 2010; Wang et al. 2012; Cromie et al. 2013; Skelly et al. 2013; Almeida et al. 2014). We therefore decided to undertake the population genomic analysis of an unexplored yeast species: *Lachancea kluyveri*. This species diverged from the *Saccharomyces cerevisiae* lineage prior to its ancestral whole-genome duplication, more than 100 Ma (Wolfe and Shields 1997); however, they both share the same life cycle (McCullough and Herskowitz 1979). *Lachancea kluyveri* possesses many characteristics, which make it a powerful model organism for population genomics and quantitative genetics. Additionally, the genome of a reference strain from this species (CBS 3082) has already been completely sequenced and annotated (Souciet et al. 2009; Jung et al. 2012). Interestingly, the sequenced genome displays an intriguing compositional heterogeneity: A region of approximately 1 Mb, covering the whole left arm of chromosome C (hereafter called Sakl0C-left) and containing the *MAT* locus, has an average GC-content which is significantly higher than the rest of the genome (52.9% compared with 40.4%) (Payen et al. 2009). The phylogenetic relationships, as well as the gene content and synteny conservation between *L. kuyveri* and closely related species suggested that this region could be the result of a hybridization event between two *Lachancea* species (Payen et al. 2009). Previous to our research project, the origin of this compositional heterogeneity was poorly understood. Our analysis strongly favors the hypothesis that the Sakl0C-left region is a relic of an introgression event, which occurred in the last common ancestor of the species. Introgression is a key process in evolution, as it may contribute to speciation and adaptation to new environments (Baack and Rieseberg 2007) and its prevalence has been well established in yeast (Morales and Dujon 2012). Our genomic data provide new insight into the potential evolutionary fate of a large-scale introgression event, leading to chromosome-scale heterogeneous evolution through different recombination and substitution rates over an extended period of time.

Article

Results and Discussion

Samples for resequencing comprised a collection of nearly all the currently available natural isolates of *L. kluyveri* originating from diverse geographical and ecological niches (supplementary table S1, Supplementary Material online). Strains of this species have been isolated worldwide in association with plants, insect guts, soil, and trees. For each isolate, we generated an average 130-fold coverage with 100-bp paired-end reads on the Illumina HiSeq 2000 (supplementary table S2, Supplementary Material online). Across our samples, we identified a total of 6,515,704 high-quality single nucleotide polymorphisms (SNPs), which are distributed over 881,427 polymorphic sites. These genomes are highly polymorphic, with an average density of 28 SNPs/kb in intergenic regions, and 17 SNPs/kb in coding regions (supplementary table S3, Supplementary Material online). Among the latter, we detected 2,668,367 synonymous (69.9%) and 1,150,061 nonsynonymous (30.1%) SNPs (supplementary table S4, Supplementary Material online). The global *L. kluyveri* genetic diversity, estimated by the average pairwise divergence ($\pi = 0.017$) and the proportion of polymorphic sites per base ($\theta_w = 0.021$), is much higher than in the *S. cerevisiae* species ($\pi = 0.00192$ and $\theta_w = 0.00226$) (Schacherer et al. 2009).

Genetic Relationship among Strains and Population Structure

In addition, we examined the GC-content across the genomes and found that the GC-rich chromosomal region, Sakl0C-left, was present in all of the *L. kluyveri* isolates, suggesting that this region predated the diversification of the species (supplementary fig. S1, Supplementary Material online). To gather clues about the origin and the evolution of Sakl0C-left, we carried out phylogenetic inferences of strain relationships for this region as well as the rest of the genome. We first focused on the global phylogenetic relationships between the different isolates by using standard neighbor-joining methods to build a majority-rule consensus tree (fig. 1). Phylogenetic analysis of the isolates revealed four main clusters, which diverged from each other (fig. 1*a*). Most of the European isolates formed a tight cluster and most of the North American isolates were grouped in another cluster composed of strains closely related to the reference strain (CBS 3082) with low genetic diversity ($\pi = 0.0006$), indicating a recent common ancestry between these isolates. In contrast, strains isolated from Asia fell into two extremely distant and distinct regions of the tree, each harboring a level of genetic variation much higher than in the other groups ($\pi = 0.0075$ and 0.0107) (supplementary table S5, Supplementary Material online). Results from the model-based clustering algorithm implemented in the program "Structure" (Pritchard et al. 2000) based on the 881,427 polymorphic sites were globally consistent with the neighbor-joining tree (fig. 1*b*). Population structure inference clearly indicates the presence of two clean lineages comprised mostly by the North American and European isolates. In contrast, the Asian isolates are not members of any well-defined lineages and it is possible to infer that most of the strains have mixed ancestry. The phylogenetic tree based on polymorphic positions located in Sakl0C-left shows the same topology as the phylogeny of the strains based on the rest of the genome (supplementary fig. S2, Supplementary Material online), confirming that this GC-rich region was present in the last common ancestor of the species and that it followed the same evolutionary trajectory as the rest of the genome.

Global Patterns of Polymorphisms

To determine the extent to which genetic diversity varied across the eight chromosomes, we divided the genome into equally sized, nonoverlapping windows of 50 kb, and examined levels of nucleotide diversity as measured by θ_w and π (fig. 2*a* and supplementary fig. S3, Supplementary Material online). The left arm of chromosome C, showed a higher genetic diversity ($\pi = 0.019$ and $\theta_w = 0.025$) than the other chromosomes. Across the genome, variation in pairwise diversity π follows the same general pattern as that of θ_w (supplementary fig. S3*a*, Supplementary Material online). Nevertheless, estimates of π were generally approximately 1.2× lower than those of θ_w. This difference results in extremely negative values of Tajima's *D* and indicates an excess of low-frequency polymorphisms relative to those under the neutral expectation model (supplementary fig. S3*b*, Supplementary Material online). In addition, Sakl0C-left has a lower Tajima's *D* value than the rest of the genome (supplementary fig. S3*b*, Supplementary Material online). This observation is related to differences in the accumulation of new mutations and thus part of the variation in genetic diversity between Sakl0C-left and the rest of the genome might be due to an uneven mutation rate.

Comparison of the Substitution Rates

To have a better insight into this heterogeneity, we then focused on the spectrum of polymorphisms. The mutational profile results from multiple processes such as mutation, selection, and genetic drift as well as recombination, in particular through the effect of GC-biased gene conversion, which favors the transmission of G/C over A/T bases (Lynch 2007; Duret and Galtier 2009). To characterize the patterns of polymorphisms, we used the genome sequence of *L. cidri*, the most closely related species, to infer ancestral and derived alleles (see Materials and Methods). We only focused on the more neutrally evolving sites: The third codon positions, representing a total of 338,356 polymorphic sites. First, we found that the substitution rate is 1.6× higher on average for the Sakl0C-left than for the rest of the genome at the third codon positions (0.21 substitutions per site vs. 0.13, respectively), confirming the higher genetic diversity previously observed. Second, we found that the mutation spectrum is strongly biased toward G/C bases on Sakl0C-left with greater substitution rates of A:T → G:C as well as A:T → C:G compared with the rest of the genome (fig. 3*a*). This difference is probably the signature of the effect of the nonadaptive process of biased gene conversion associated with recombination on this chromosomal arm (Lesecque et al. 2013).

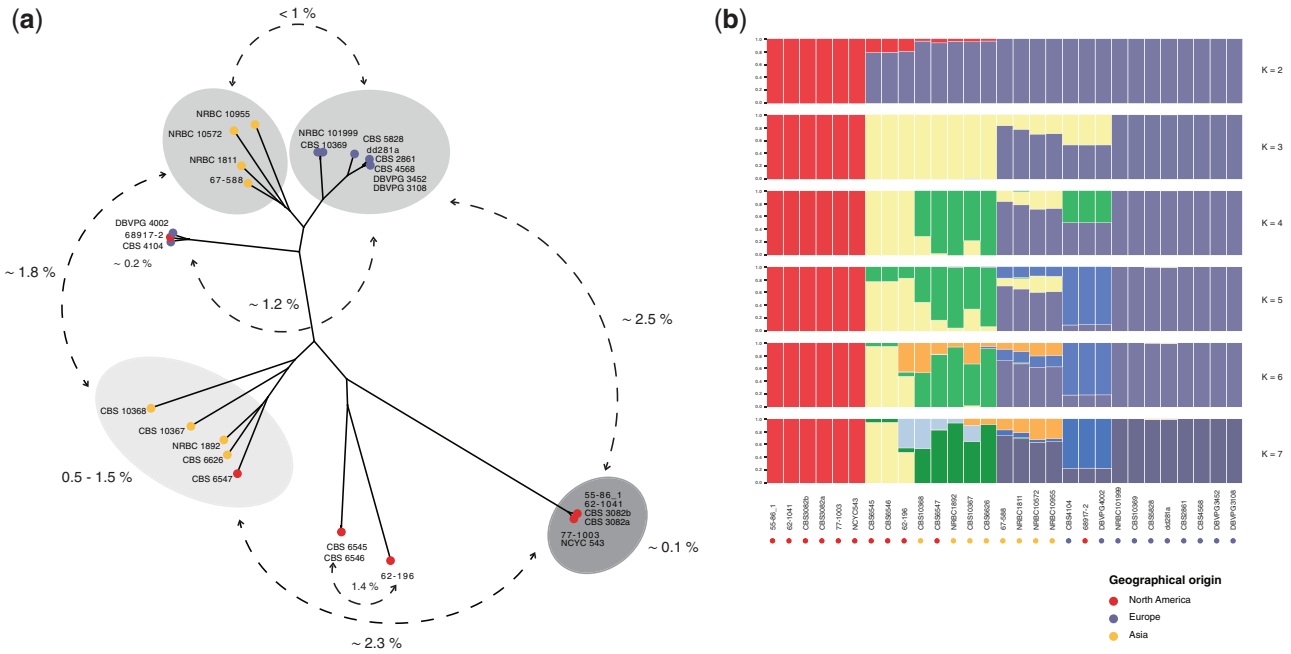

Fig. 1. Phylogenetic relationship and population structure of the 28 *Lachancea kluyveri* strains. (*a*) Neighbor-joining tree of the 28 *L. kluyveri* strains, constructed on the basis of the 881,427 polymorphic sites identified in the surveyed strains. Branch lengths are proportional to the number of sites that discriminate each pair of strain. (*b*) Model-based clustering analysis of the population with Structure. The number of populations (*K*) was predefined from 2 to 7. Each strain is represented by a single vertical bar, which is partitioned into *K* colored segments that represent the strain's estimated ancestry proportion in each of the *K* clusters. The circle colors denote the geographical origins of the strains.

Pattern of Linkage Disequilibrium

Such a regional difference should also have had an impact and been apparent in the linkage disequilibrium (LD) patterns. LD is a major aspect of the organization of genetic variation in natural populations. Thus, we calculated r^2, a measure of association for LD, for all pairs of polymorphic sites. Our data also provided the opportunity to measure genome-wide properties of LD within the *L. kluyveri* species. On average, LD decays within 25 kb, reaching 50% of its maximal value at about 1.5 kb (fig. 4*a*). In *L. kluyveri*, average LD decayed relatively quickly compared with *S. cerevisiae* and *S. paradoxus* where LD decays to half of its maximum value at about 3 and 9 kb, respectively (Tsai et al. 2008; Schacherer et al. 2009). We then looked at the rate of decay of LD at the level of individual chromosomes (fig. 4*b*). Most chromosomes exhibited rates of decay of LD that were similar to genome-wide values ($LD_{1/2}$ = ~ 1.5 kb), whereas Sakl0C-left had a much lower level of LD, with it falling to half of its maximum value at approximately 0.3 kb (fig. 4*b*). This observation confirms that the recombination rate was probably higher on this chromosomal arm compared with the rest of the genome.

Pattern of Population Recombination Rate

To evaluate the historical pattern of recombination, we estimated the population-scale recombination rate (ρ) across the genome. The value for ρ was calculated between neighboring pairs of SNPs using the program LDhat (McVean et al. 2004). The genome-wide average estimate of ρ was found to be 3.17 Morgans/kb, which is lower than the estimate for *S. cerevisiae* (ρ = 5.06 Morgans/kb) we determined using recently published sequence data (Skelly et al. 2013). To assess recombination rate variability across the genome, we averaged these ρ estimates in nonoverlapping 50-kb windows to obtain a finer-scale genetic map for each of the eight chromosomes (fig. 2*c*). The results show heterogeneity in recombination rate along the genome, particularly on Sakl0C-left, which has an increased rate. Recombination in this 1-Mb region is about 2.7× higher than that determined for the rest of the genome (ρ = 8.5 vs. 3.2 Morgans/kb). Therefore, Sakl0C-left is characterized by a higher GC-content and increased genetic diversity as well as a higher ancestral recombination rate as compared with the rest of the genome (fig. 2). All of these observations support the hypothesis that there is a stronger effect of GC-biased gene conversion on Sakl0C-left compared with the rest of the genome. We also scanned the genome for hotspots of recombination using stringent criteria to avoid the detection of false positives (see Materials and Methods). Basically, we tested for statistically significant increases in ρ compared with flanking regions, using the program SequenceLDhot (Fearnhead 2006). A total of 98 recombination hotspots, with an average length of 2 kb, were found in the whole population and half were located in the Sakl0C-left region (supplementary fig. S4 and table S6, Supplementary Material online). We looked for Gene Ontology (GO)-term enrichment among the genes located in hotspot regions but we did not find any convincing pattern. Nevertheless by excluding the Sakl0C-left region, which is totally devoid of

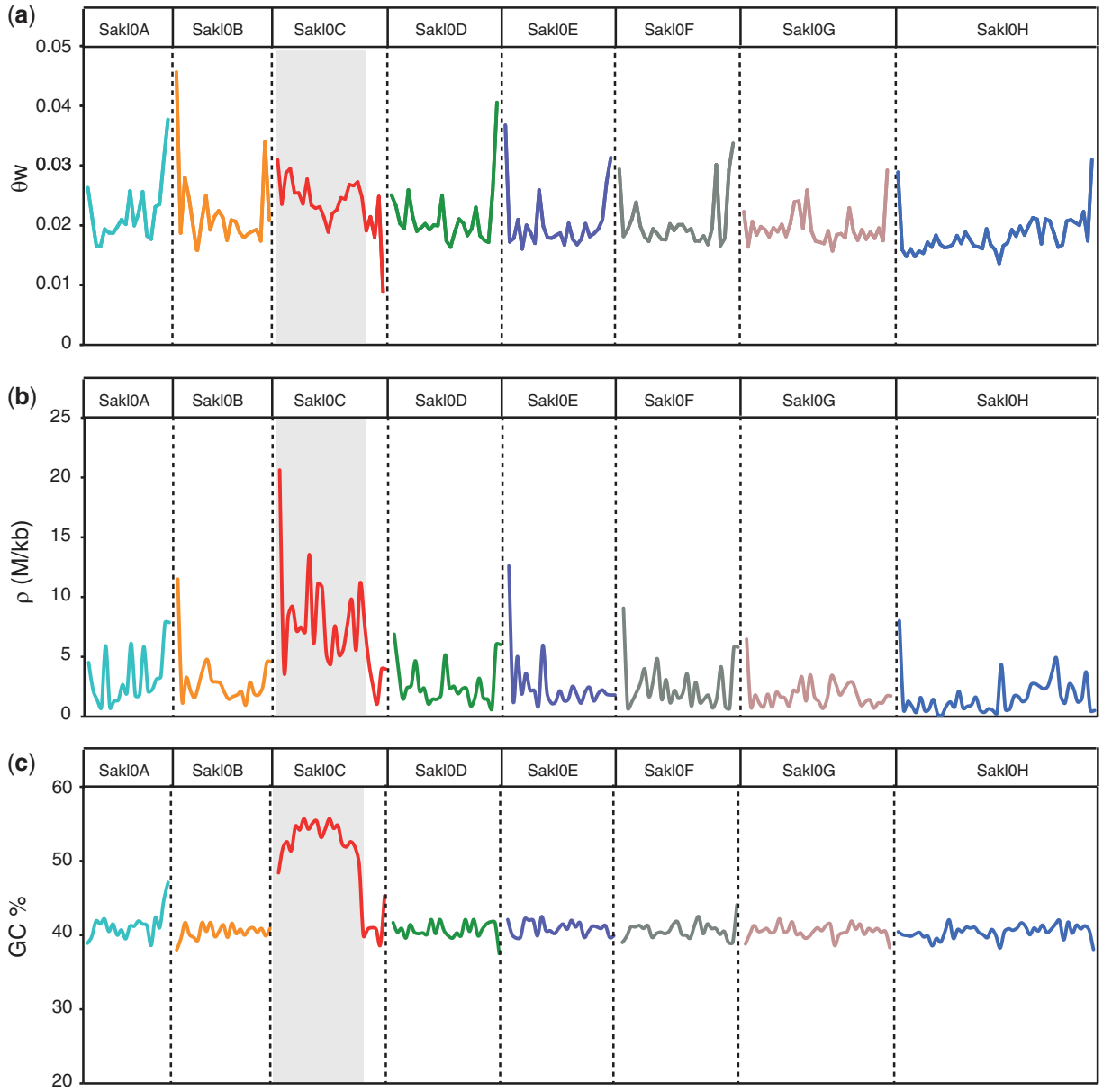

Fig. 2. Variation of genetic metrics along chromosomes within the *Lachancea kluyveri* population Metrics were computed within 50-kb nonoverlapping sliding windows. Gray shading delimits the left arm of chromosome C. (*a*) Proportion of polymorphic sites θ_w. (*b*) Population-scale recombination rate ρ. (*c*) Mean GC-content within the 28 *L. kluyveri* strains.

transposons, we found an enrichment of hotspots in transposable elements and tRNA (24 of the 49 hotspots) (supplementary table S7, Supplementary Material online).

Comparison between Substitution Rates and Number of Substitutions

Given the heterogeneity in the nucleotide composition in *L. kluyveri*, we also decided to compare the substitution rates and the number of substitutions. Previously, we have shown that SakI0C-left has greater A:T → G:C as well as A:T → C:G substitution rates compared with the rest of the genome (fig. 3*a*). However, these substitution rates mask the discrepancy in terms of number of AT and GC sites involved. As a consequence, even if it seems counterintuitive, a higher substitution rate toward GC bases does not preclude a higher number of substitutions toward AT bases. Indeed, SakI0C-left exhibits more GC than AT sites in particular at third codon positions where GC-content reaches 71% on average (supplementary fig. S5, Supplementary Material online). As a consequence, the total number of substitution toward AT is higher than toward GC (fig. 3*b*) but the substitution rates toward GC are higher than toward AT (fig. 3*a*). This observation is simply due to the fact that the total number of GC mutable sites is much higher than the total number of AT mutable sites.

In the figure 3*b*, we displayed the fraction of nucleotide substitutions (not the substitution rate), meaning the ratio between the number of substitution (AT → GC or

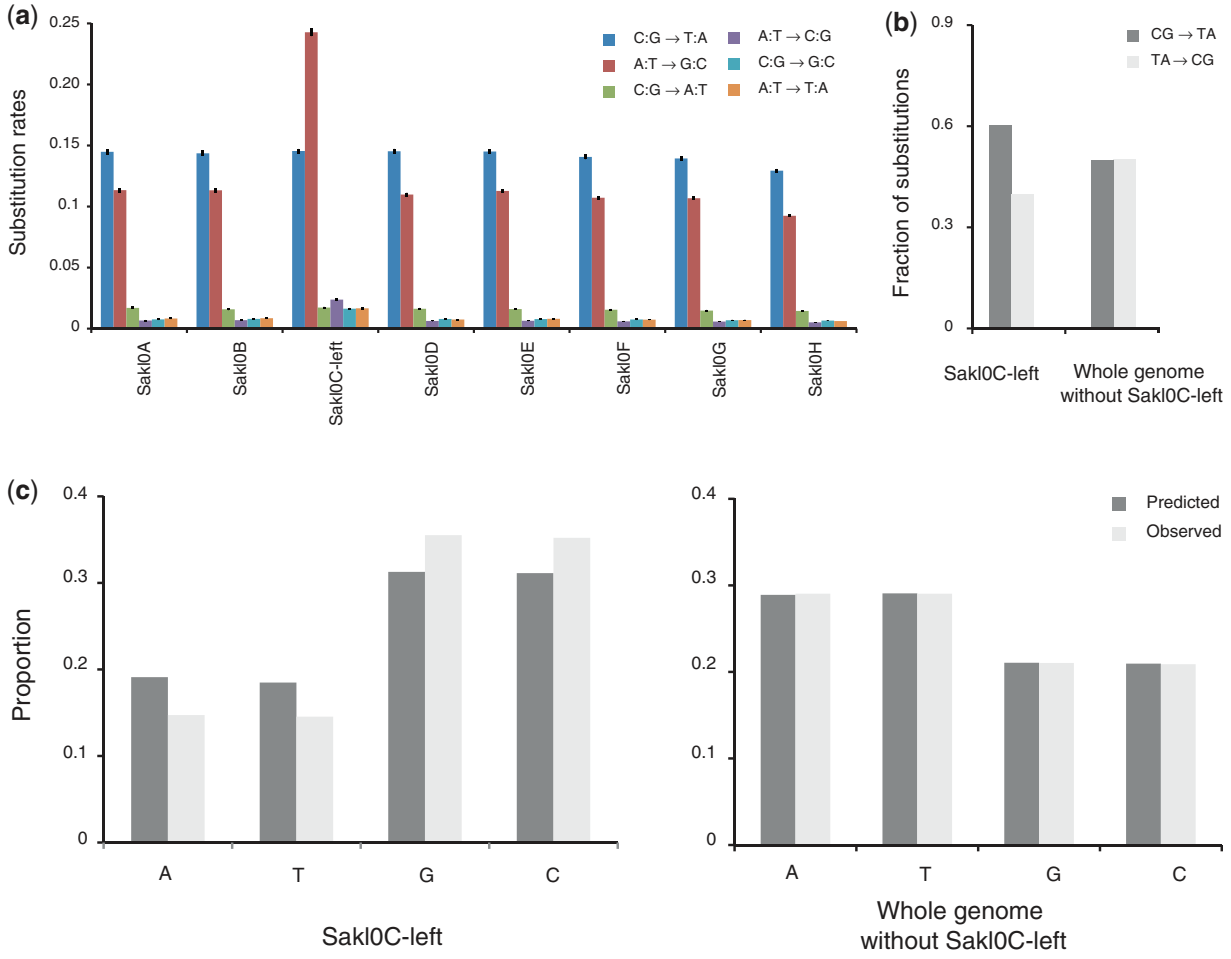

Fig. 3. Mutational spectrum. (*a*) Substitution rates of the six different types of polymorphisms at third codon positions in the *Lachancea kluyveri* genome, polarized against *Lachancea cidri*. Error bars represent the confidence intervals at 95%. (*b*) Fraction of GC → AT and AT → GC substitutions polarized against *L. cidri* in SakI0C-left and in the rest of the genome. (*c*) Comparisons between the observed base composition in the genome and the predicted base composition at mutational equilibrium.

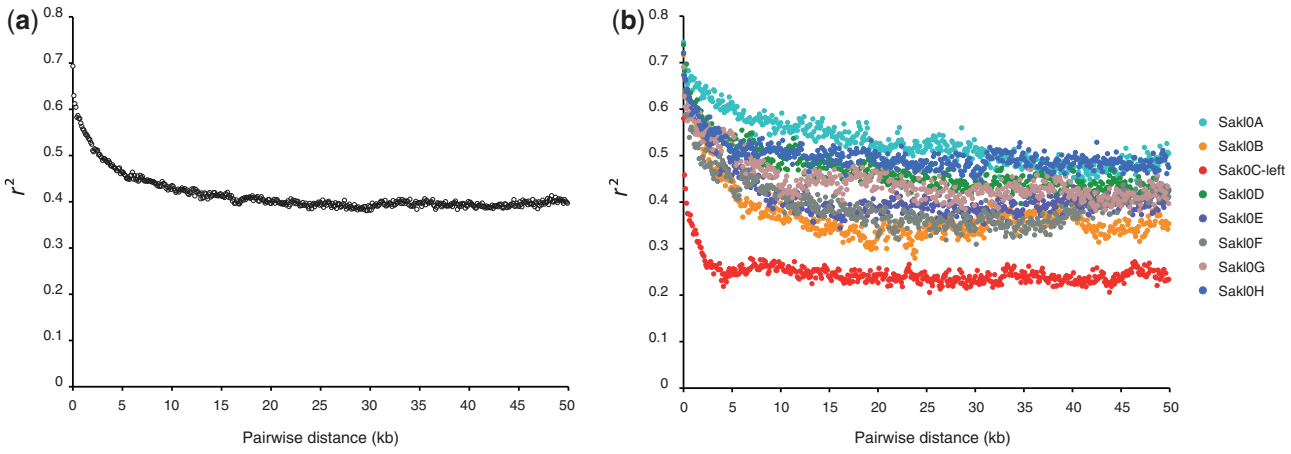

Fig. 4. Decay of LD as a function of distance. Squared correlations of allele frequencies (r^2) are plotted for each bin of distances between pairs of polymorphic sites. (*a*) Considering the whole genome. (*b*) Considering each chromosome individually.

GC → AT) and the total number of substitutions. Across the whole genome, with the exception of SakI0C-left, we found that the total number of substitutions toward AT bases (153,386 substitutions) equals the number of substitutions toward GC bases (154,034 substitutions) at third codon positions (fig. 3*b*) and the site frequency spectrum of these two types of substitutions is equivalent except at SakI0C-left (supplementary fig. S6, Supplementary Material online).

This strongly suggests that base composition is at mutational equilibrium apart from the Sakl0C-left, which has not yet reached this balance. Interestingly, substitutions toward AT (27,375 substitutions) largely outnumber substitutions toward GC (18,334 substitutions) (fig. 3*b*). This observation suggests that GC-content is actually decreasing on Sakl0C-left even if the AT → GC substitution rate is higher. The excess of substitutions from G or C to A or T is seen only in derived alleles of low frequency on Sakl0C-left, most of which were likely the result of relatively recent mutations (supplementary fig. S6, Supplementary Material online).

The fact that the GC-content is decreasing in Sakl0C-left implies that it was probably higher in the common ancestor than what it is currently observed. To obtain a better insight into this hypothesis, we looked at the GC-content at the third codon positions of the constant sites (i.e., the nonpolymorphic and therefore ancestral sites) and compared it with all the sites (including the polymorphic sites) across the 28 sequenced genomes (supplementary fig. S5, Supplementary Material online). Interestingly, the GC-content is higher for the constant sites compared with all the sites, clearly demonstrating that the CG-content was higher in the common ancestor and is currently decreasing.

Genome Composition at Mutational Equilibrium

At mutational equilibrium, the predicted nucleotide composition at the third codon position was then calculated (see Materials and Methods). As expected, the predicted nucleotide composition at equilibrium is very close to the observed composition for the entire genome except for Sakl0C-left (fig. 3*c*). The predicted base composition at equilibrium for Sakl0C-left also confirms that GC-content is actually decreasing in this region. It is also noteworthy that the predicted base composition at mutational equilibrium is radically different for the two regions, suggesting that a compositional heterogeneity will still be maintained once Sakl0C-left reaches the mutational equilibrium in the *L. kluyveri* species.

All of these observations clearly indicate that the GC-content of the C-left was higher in the common ancestor and that it is currently decreasing in this region of the genome. This trend is not easily reconcilable with the hypothesis of an intrinsic mutational mechanism at the origin of the GC heterogeneity because such mechanism would be predicted to actually increase the GC-content. Instead, our results show that the GC-content is decreasing, strongly supporting the hypothesis of an introgression of a large GC-rich region.

Estimation of the Timing of the Introgression Event

Finally, we used the population genomic data obtained to estimate the minimum number of generations that have passed since the last common ancestor, allowing for us to estimate the length of time necessary for the evolution of the introgressed region to reach its current state. Following the same strategy as previously used for *S. cerevisiae* (Ruderfer et al. 2006), our data lead to an estimation of 55.5×10^6 generations since the last common ancestor (see Materials and Methods). If we consider that cell division of yeast in the wild ranges from 1 to 8 generations per day (Fay and Benavides 2005), the number estimated above corresponds to approximately 19,000–150,000 years since the most recent common ancestor. In addition, we estimated that approximately 500 generations of outcrossing occurred during the evolutionary history of *L. kluyveri* (see Materials and Methods). Thus, it goes through a sexual cycle with outcrossing approximately once every 110,000 generations, which is equally as rare as previously observed in *S. cerevisiae* and *S. paradoxus* (Fay and Benavides 2005; Tsai et al. 2008).

Conclusion

In this study, we provide a comprehensive description and analysis of genome-wide variation in a protoploid yeast species. Our results indicate that the polymorphism rate is higher ($\theta_w = 0.021$) whereas the level of LD is lower ($LD_{1/2} = \sim 1.5$ kb) in *L. kluyveri* compared with *S. cerevisiae*. Notably, our population genomic analysis strongly suggests that an approximately 1-Mb GC-rich region was the result of an introgression within this species. Obviously, definitive evidence for such an introgression event would be the identification of the donor species, which is expected to be closely related to *L. kluyveri* with a genome showing a conserved synteny, a high GC-content and lacking the mating cassettes. Unfortunately, none of the genomes from the *Lachancea* clade sequenced so far meet these criteria. Interestingly, the evolution pattern of this region is characterized by a higher recombination rate as revealed by a higher LD as well as a higher mutation rate as indicated by a lower Tajima's *D* value than the rest of the genome. In addition, this introgression shapes the evolution of the mating-type chromosome. As previously observed, introgressions do not occur randomly across genomes. Interestingly, the silent mating-type cassettes (*HML* and *HMR*) are absent from the genome of *L. kluyveri* but present in the subtelomeres of the chromosomal arms orthologous to Sakl0C-left in closely related *Lachancea* species (Payen et al. 2009). Because of the loss of these cassettes, *L. kluyveri* is heterothallic (self-incompatible) preventing autodiploidization. In this context, introgression might have provided an adaptive advantage through the modification of the life cycle. Altogether, the results presented here, using an alternative yeast model organism, contribute to a better understanding of the evolutionary significance of introgression in natural populations. Hybridization events are recognized as an important and widespread process in yeast (Morales and Dujon 2012), and therefore chromosome-scale mutational heterogeneity might be a key factor of yeast genome evolution.

Materials and Methods

Strains

A collection of 28 strains isolated from diverse ecological (tree exudate, soil, insects) and geographical (Europe, Asia, and America) origins was compiled for this study (supplementary table S1, Supplementary Material online). The strains, selected to maximize the range of sources and location, were purchased from various yeast culture collections: CBS

(Centraalbureau voor Schimmelcultures), DBVPG (Dipartimento di Biologia Vegetale e Agroambientale of the University of Perugia), NBRC (NITE Biological Resource Center), and NCYC (National Collection of Yeast Cultures). dd281 was kindly provided by Michael Knop from the Center for Molecular Biology of the University of Heidelberg.

Sequencing and Polymorphism Detection

A single haploid clone was isolated from each strain for sequencing. Yeast cell cultures were grown overnight at 30 °C in 20 ml of YPD medium to early stationary phase. Total genomic DNA was subsequently extracted using the QIAGEN Genomic-tip 100/G according to the manufacturer's instructions.

Genomic Illumina sequencing libraries were prepared with a mean insert size of 280 bp. The 28 libraries were multiplexed in three Illumina HiSeq2000 lanes and subjected to paired-end sequencing with read lengths of 102 and 104 bp, six of them being dedicated to the multiplex barcode. A total of 52.8 Gb of high-quality genomic sequence was generated for a mean coverage of 130× per strain. All Illumina sequencing reads generated in this study have been deposited in the European Nucleotide Archive under the accession numbers PRJEB5130.

FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/, last accessed November 4, 2014) was used to clean the reads with options "-t 20 -l 50." The clean reads were than mapped with the Burrows–Wheeler aligner (version 0.5.8) to the CBS 3082 reference genome, allowing eight mismatches and two gaps (Li and Durbin 2009). After mapping, SNPs and short indels were identified on the basis of the pileup files generated by SAMtools (version 0.1.8) (Li et al. 2009).

In order to minimize false-positive SNPs calling, we considered only SNPs at positions covered by more than 20 reads and for which 90% of the reads were in accordance with the variation. A total of 881,427 polymorphic positions were highlighted (supplementary table S2, Supplementary Material online).

Tree Building and Structure Analysis

To obtain a view of the phylogenetic relationships between the different isolates, we constructed a neighbor-joining tree of the 28 strains from the 881,427 polymorphic positions detected in the whole population, using the software package Splitstree (Huson and Bryant 2006). Branch lengths are proportional to the number of segregating sites that differentiate each node. To compare the evolution of the Sakl0C-left to the rest of the genome, two phylogenetic trees were independently generated with PhyML (Guindon and Gascuel 2003) from 92,616 and 787,645 polymorphic sites located in Sakl0Cleft and in the rest of the genome, respectively, with the HKY85 substitution matrix and a discrete gamma model with four categories. A global tree based on all polymorphic sites was also generated with the same parameters for downstream analyses with the PAML software.

The estimation of the number of population clusters was performed with the same SNP data with the Structure program, version 2.3.1 (Pritchard et al. 2000). We ran the Structure program using the admixture model with the population number parameter K set from two to seven, on 20,000 replicates after a burn-in of 10,000 iterations.

Calculation of Population Genetic Statistics

Two standard estimates of the scaled mutation rate, θ_w, the proportion of segregating sites, and π, the average pairwise nucleotide diversity, as well as Tajima's D (Tajima 1989), the difference between π and θ, were used to characterize nucleotide diversity among the populations. These metrics were calculated with Variscan (Hutter et al. 2006), for the whole population, using a nonoverlapping sliding window approach along all chromosomes, with a window size of 50 kb. To compare the nucleotide diversity between coding and noncoding regions, θ_w, π, and Tajima's D were also estimated along coding sequences (CDS) and intergenic regions. For this purpose, Variscan was launched with BDF files associated. These latter files report the CDS/intergenic region coordinates, respectively, as found in the reference genome annotation files: http://www.genolevures.org/download.html#sakl (last accessed November 4, 2014).

We have estimated the number of generations since the most common ancestor of any pair of strains using the observed population mutation parameter ($\theta = 2N_\theta\mu$). The number of generations of outcrossing events that have contributed to the sample can be obtained based on the population recombination rate ($\rho = 2N_\rho r$). We used the mutation and recombination rates (μ and r) from laboratory estimates on *S. cerevisiae* (Cherry et al. 1997; Lynch et al. 2008).

Substitution Rates

The genome of *L. cidri*, the closest relative to *L. kluyveri*, was completely sequenced and annotated (Neuvéglise C, unpublished data) and used as the outgroup species to root a phylogenetic tree (supplementary fig. S7, Supplementary Material online). We identified 885 syntenic homologs between the two species (supplementary material, Supplementary Material online) sharing more than 85% of similarity using SynChro (Drillon et al. 2014). Homologous proteins were aligned with MUSCLE (Edgar 2004) and alignments were cleaned with Gblocks (Castresana 2000). Cleaned concatenated alignments were then analyzed with PhyML, which was run with the general time reversible (GTR) amino-acid substitution model. The root of the tree, inferred from the position of *L. cidri*, is located in the branch separating the North American group of strains from all the other isolates. Confidence scores were assessed by performing 1,000 bootstrap replicates in PhyML (supplementary fig. S7, Supplementary Material online). Ancestral sequences were inferred for the 338,378 third codon polymorphic positions at each node of the global phylogenetic tree with PAML 4.4 (Yang 2007) using the baseml program and the REV (GTR) matrix, defined as the best model by jModelTest-2.1.2 (Darriba et al. 2012). A total of 374,512 substitutions could be oriented by rooting the tree with the outgroup species *L. cidri*. The substitution rates were calculated from the subset

of third codon position SNPs by dividing the number of substitutions of a given type by the number of potentially mutable sites of the same type with the use of the AMADEA Biopack platform (Isoft, http://www.isoft.fr/bio/biopack_en.htm, last accessed November 4, 2014).

Linkage Disequilibrium

LD was assessed by generating an r^2 value with the Plink package (Purcell et al. 2007), both for the whole population and for each subpopulation. SNP data excluding singletons were used in these studies. LD decay plot was generated with a custom R script.

Rates of Recombination and Hotspot Identification

The population recombination rate ρ was calculated for consecutive pairs of SNPs using a penalized likelihood within a Bayesian reversible-jump Markov-chain Monte Carlo scheme implemented in the Interval program of the LDhat package (version 2.2) (McVean et al. 2004). Interval was run with 2,000,000 iterations, a block penalty of 10 and with samples taken every 5,000th iteration. As they are not informative in the context of recombination studies, singleton SNPs were also excluded here. Recombination hotspots were identified using sequenceLDhot (Fearnhead 2006). We tested for statistically significant increases in ρ in 2-kb window (every 1 kb) using a 3.5 rho driving and background value, and setting theta per site at 0.02. Genetic elements located in these hotspots regions were parsed with custom python scripts and their gene content tested for GO enrichment with GO::TermFinder (Boyle et al. 2004).

Nucleotide Composition at Mutational Equilibrium

Nucleotide frequencies at mutational equilibrium were computed from the 12 individual rates of substitution at third codon positions by solving simultaneously the four equations presented in Sueoka (1995), using the program Maxima 5.25.1: http://sourceforge.net/projects/maxima/files/Maxima-source/ (last accessed November 4, 2014).

Accession Codes

All reads have been deposited in the European Nucleotide Archive under the accession numbers PRJEB5130.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S7 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Joshua Shapiro, Gwenael Piganeau, Guillaume Achaz, and Kelle Freel for their invaluable advice and their colleagues from the GB-3 G project for helpful discussions. They also thank Sophie Siguenza for bioinformatics assistance. They are most grateful to the GeneCore sequencing team (EMBL, Heidelberg, Germany). This work was supported by an ANR grant (2010-BLAN-1606), a grant from CNRS and Région Alsace to C.R., an ATIP/avenir Plus grant from the CNRS to G.F., and an ANR grant (2011-JSV6-004-01) to J.S.

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
Almeida P, Gonçalves C, Teixeira S, Libkind D, Bontrager M, Masneuf-Pomarède I, Albertin W, Durrens P, Sherman DJ, Marullo P, et al. 2014. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat Commun.* 5: 4044.
Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, et al. 2012. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat Genet.* 44:285–290.
Baack EJ, Rieseberg LH. 2007. A genomic view of introgression and hybrid speciation. *Curr Opin Genet Dev.* 17:513–518.
Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. 2004. GO::TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 20:3710–3715.
Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet.* 43:956–963.
Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17: 540–552.
Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. 1997. Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* 387:67–73.
Cromie GA, Hyma KE, Ludlow CL, Garmendia-Torres C, Gilbert TL, et al. 2013. Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. *G3 (Bethesda)* 3: 2163–2171.
Darriba DA, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9:772.
Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* 9:e92621.
Dujon B. 2010. Yeast evolutionary genomics. *Nat Rev Genet.* 11:512–524.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10:285–311.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
Fay JC, Benavides JA. 2005. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.* 1:e5.
Fearnhead P. 2006. SequenceLDhot: detecting recombination hotspots. *Bioinformatics* 22:3061–3066.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52: 696–704.
Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, Rokas A. 2010. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* 464:54–58.
Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.
Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* 7:409.
Jung PP, Friedrich A, Reisser C, Hou J, Schacherer J. 2012. Mitochondrial genome evolution in a single protoploid species. *G3 (Bethesda)* 2: 1113–1127.
Lesecque Y, Mouchiroud D, Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol Biol Evol.* 30:1409–1419.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
Li H, et al. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078–2079.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. 2009. Population genomics of domestic and wild yeasts. *Nature* 458:337–341.

Liti G, Schacherer J. 2011. The rise of yeast population genomics. *C R Biol.* 334:612–619.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates Inc.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A.* 105:9272–9277.

McCullough J, Herskowitz I. 1979. Mating pheromones of *Saccharomyces kluyveri*: pheromone interactions between *Saccharomyces kluyveri* and *Saccharomyces cerevisiae*. *J Bacteriol* 138:146–154.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science* 304:581–584.

Morales L, Dujon B. 2012. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol Mol Biol Rev* 76: 721–739.

Payen C, Fischer G, Marck C, Proux C, Sherman DJ, et al. 2009. Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Res.* 19:1710–1721.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 81:559–575.

Ruderfer DM, Pratt SC, Seidel HS, Kruglyak L. 2006. Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet.* 38: 1077–1081.

Schacherer J, Shapiro J, Ruderfer DM, Kruglyak L. 2009. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 458:342–345.

Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, et al. 2013. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* 9:1496–1504.

Souciet JL, Dujon B, Gaillardin C, Johnston M, Baret PV, et al. 2009. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.* 19:1696–1709.

Sueoka N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol.* 40:318–325.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. 2008. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc Natl Acad Sci U S A.* 105:4957–4962.

Wang QM, Liu WQ, Liti G, Wang SA, Bai FY. 2012. Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol Ecol* 21:5404–5417.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.