MBE Advance Access published October 27, 2014

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 Univ Paris 06, UMR 7238, Biologie Computationnelle et Quantitative, Paris, France

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

* Corresponding authors:

JS - schacherer@unistra.fr,

GF - gilles.fischer@upmc.fr

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

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 *S. cerevisiae*: *Lachancea kluyveri* (formerly *Saccharomyces kluyveri*). We identified 6.5 million SNPs 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.

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Introduction

Detailed examination of the patterns of genetic variation is the first step towards 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* (Andersen et al. 2012; Cao et al. 2011; The 1000 Genomes Project Consortium 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 (Schacherer et al. 2009; Liti et al. 2009; Hittinger et al. 2010; Wang et al. 2012; Skelly et al. 2013; Cromie 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 *S. cerevisiae* lineage prior to its ancestral whole genome duplication, more than 100 million years ago (Wolfe and Shields 1997), however they both share the same life cycle (McCullogh and Herskowitz 1979). *L. 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 to 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.

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Results and discussion

Samples for resequencing were comprised of a collection of nearly all the currently available natural isolates of *L. kluyveri* originating from diverse geographical and ecological niches (Supplementary table 1). 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 2). Across our samples, we identified a total of 6,515,704 high-quality 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 3). Among the latter, we detected 2,668,367 synonymous (69.9%) and 1,150,061 nonsynonymous (30.1%) SNPs (Supplementary table 4). 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. 1). 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 4 main clusters, which diverged from each other (fig. 1a). 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. By 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 5). 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. 1b). Population structure inference clearly indicates the presence of two clean lineages

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

comprised mostly by the North American and European isolates. By 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. 2), 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 8 chromosomes, we divided the genome into equally sized, non-overlapping windows of 50 kb, and examined levels of nucleotide diversity as measured by θ_w and π (fig. 2a and Supplementary fig. 3). 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. 3a). Nevertheless, estimates of π were generally ~1.2x 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. 3b). In addition, Sakl0C-left has a lower Tajima's *D* value than the rest of the genome (Supplementary fig. 3b). 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 *Lachancea cidri*, the most closely related species, to infer ancestral and derived alleles (see 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 1.6x higher on average for the Sakl0C-left than for the rest of the genome at the 3rd 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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

strongly biased towards G/C bases on Sakl0C-left with greater substitution rates of A:T→G:C as well as A:T→C:G compared to the rest of the genome (fig. 3a). This difference is probably the signature of the effect of the non-adaptive process of biased gene conversion associated with recombination on this chromosomal arm (Lesecque et al. 2013).

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. 4a). In *L. kluyveri*, average LD decayed relatively quickly compared to *S. cerevisiae* and *S. paradoxus* where LD decays to half of its maximum value at about 3 kb and 9 kb, respectively (Schacherer et al. 2009; Tsai et al. 2008). We then looked at the rate of decay of LD at the level of individual chromosomes (fig. 4b). 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 ~0.3 kb (fig. 4b). This observation confirms that the recombination rate was probably higher on this chromosomal arm compared to 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 per 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 non-overlapping 50-kb windows to obtain a finer-scale genetic map for each of the 8 chromosomes (fig. 2c). 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 2.7x 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 to 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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

compared to 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 Methods). Basically, we tested for statistically significant increases in ρ compared to 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. 4 and Supplementary table 6). We looked for 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 transposons, we found an enrichment of hotspots in transposable elements and tRNA (24 out of the 49 hotspots) (Supplementary table 7).

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 Sakl0C-left has greater A:T→G:C as well as A:T→C:G substitution rates compared to the rest of the genome (fig. 3a). However, these substitution rates mask the discrepancy in terms of number of AT and GC sites involved. As a consequence, even if it seems counter-intuitive, a higher substitution rate towards GC bases does not preclude a higher number of substitutions towards AT bases. Indeed, Sakl0C-left exhibits more GC than AT sites in particular at 3rd codon positions where GC-content reaches 71% on average (Supplementary fig. 5). As a consequence, the total number of substitution towards AT is higher than towards GC (fig. 3b) but the substitution rates towards GC are higher than towards AT (fig. 3a). This observation is simply due to the fact that the total numbers of GC mutable sites is much higher than the total number of AT mutable sites.

In the figure 3b, we displayed the fraction of nucleotide substitutions (not the substitution rate), meaning the ratio between the number of substitution (AT→GC or GC→AT) and the total number of substitutions. Across the whole genome, with the exception of Sakl0C-left, we found that the total number of substitutions towards AT bases (153,386 substitutions) equals the number of substitutions towards GC bases (154,034 substitutions) at 3rd codon positions (fig. 3b) and the site frequency spectrum of these two types of substitutions are equivalent except at Sakl0C-left (Supplementary fig. 6). This strongly suggests that base composition is at mutational equilibrium apart from the Sakl0C-left, which has not yet reached this balance. Interestingly, substitutions towards AT (27,375 substitutions) largely outnumber substitutions towards GC (18,334 substitutions) (fig. 3b). This observation suggests that GC-content is actually decreasing on Sakl0C-left even if the AT→GC

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

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. 6).

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 3rd codon positions of the constant sites (i.e. the non-polymorphic and therefore ancestral sites) and compared it to all the sites (including the polymorphic sites) across the 28 sequenced genomes (Supplementary fig. 5). Interestingly, the GC-content is higher for the constant sites compared to 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 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. 3c). 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 x 10^6 generations since the last common ancestor (See Methods). If we consider that cell division of yeast in the wild ranges from 1 to 8 generations

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

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 ~500 generations of outcrossing occurred during the evolutionary history of *L. kluyveri* (See 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* (Tsai et al. 2008; Fay and Benavides 2005).

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 (θ_w = 0.021) whereas the level of LD is lower ($LD_{1/2}$ = ~1.5 kb) in *L. kluyveri* compared to *S. cerevisiae*. Notably, our population genomic analysis strongly suggest that an ~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.

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Material 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 1). 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 3 Illumina HiSeq2000 lanes and subjected to paired-end sequencing with read lengths of 102 and 104 bp, 6 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 130X 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/) was used to clean the reads with options "-t 20 -l 50". The clean reads were than mapped with the Burrows-Wheeler aligner (BWA, version 0.5.8) to the CBS 3082 reference genome, allowing 8 mismatches and 2 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 2).

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

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 (Guidon 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 4 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 non-overlapping sliding window approach along all chromosomes, with a window size of 50 kb. To compare the nucleotide diversity between coding and non-coding regions, θ_w, π and Tajima's *D* were also estimated along 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

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).

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Substitution rates

The genome of *L. cidri,* the closest relative to *L. kluyveri*, was completely sequenced and annotated (C. Neuvéglise, unpublished) and used as the outgroup species to root a phylogenetic tree (Supplementary fig. 7). We identified 885 syntenic homologs between the 2 species (Supplementary Material) 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 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 1000 bootstrap replicates in PhyML (Supplementary fig. 7). Ancestral sequences were inferred for the 338,378 3rd 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 3rd 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).

Linkage disequilibrium

Linkage disequilibrium was assessed by generating an r^2 value with the Plink package (Purcell 2007), both for the whole population and for each subpopulation. SNP data excluding singletons were used in these studies. LD decay plot were 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 (rjMCMC) 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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

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/

Accession codes

All reads are archived in the NCBI Sequence Read Archive under the accession number PRJEB5130 - http://www.ebi.ac.uk/ena/data/view/PRJEB5130

Acknowledgements

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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

References

Almeida P, Gonçalves C, Teixeira S, Libkind D, Bontrager M, et al. A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. Nat Commun. 2014; 5: 4044.

Andersen EC, Gerke JP, Shapiro JA, Crissman JR, Ghosh R, et al. Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat Genet. 2012; 44:285-90.

Baack EJ, Rieseberg LH. A genomic view of introgression and hybrid speciation. Curr Opin Genet Dev. 2007; 17:513-518.

Boyle EI1, Weng S, Gollub J, Jin H, Botstein D, et al. GO::TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 2004; 20:3710-3715.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011; 43:956-63.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17:540-552.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 1997; 387:67-73.

Cromie GA, Hyma KE, Ludlow CL, Garmendia-Torres C, Gilbert TL, et al. Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. G3 (Bethesda) 2013; 3:2163-2171.

Darriba Da, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 2012; 9:772.

Drillon G, Carbone A, Fischer G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 2014; 9:e92621.

Dujon B. Yeast evolutionary genomics. Nat Rev Genet. 2010; 11:512-524.

Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009; 10:285-311.

Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004; 5:113.

Fay JC, Benavides JA. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. 2005; 1:e5.

Fearnhead P. SequenceLDhot: detecting recombination hotspots. Bioinformatics 2006; 22:3061-3066.

Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003; 52:696-704.

Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, Rokas A. Remarkably ancient balanced polymorphisms in a multi-locus gene network. Nature. 2010; 464:54-58.

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006; 23:254-267.

Hutter S, Vilella AJ, Rozas J. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics 2006; 7:409.

Jung PP, Fridedrich A, Reisser C, Hou J, Schacherer J. Mitochondrial genome evolution in a single protoploid species. G3 (Bethesda) 2012; 2:1113-1127.

Lesecque Y, Mouchiroud D, Duret L. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol Biol Evol. 2013; 30:1409-19.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25:1754-1760.

Li H, et al. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 2009; 25:2078-2079.

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

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

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

Lynch M. The Origins of Genome Architecture. Sinauer Associates Inc. 2007.

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

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

Morales L, Dujon B. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. Microbiol Mol Biol Rev. 2012; 76:721-39.

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

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

Purcell S1, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559-575.

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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

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

Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, et al. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Res. 2013; 9:1496-504.

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

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

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

The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491:56-65.

Tsai IJ, Bensasson D, Burt A, Koufopanou V. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. Proc Natl Acad Sci USA 2008; 105:4957-4962.

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

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

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

Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Legends

Figure 1. Phylogenetic relationship and population structure of the 28 *L. 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.

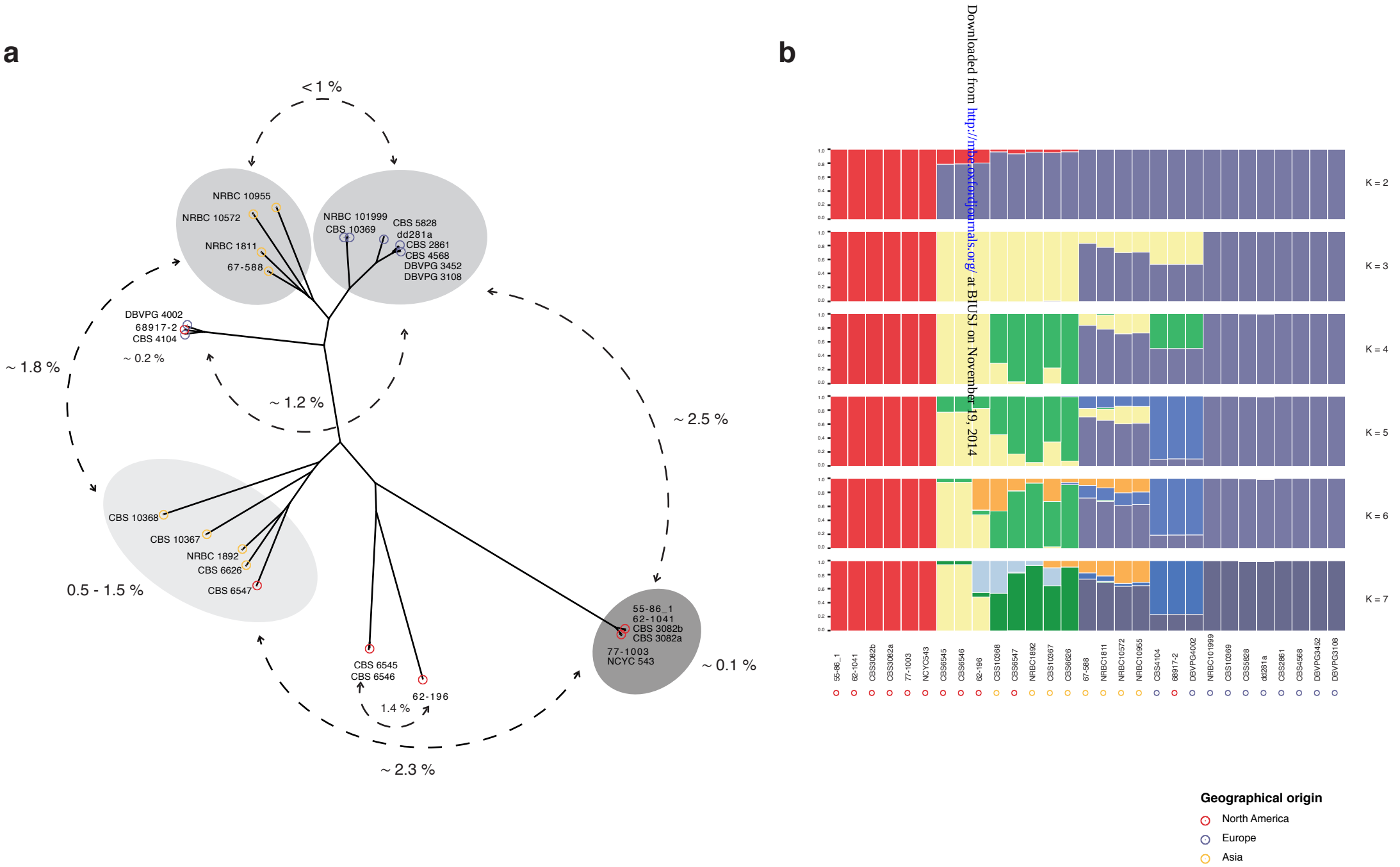
Figure 2. Variation of genetic metrics along chromosomes within the *L. kluyveri* population
Metrics were computed within 50 kb non-overlapping sliding windows. Grey shading delimits the left arm of chromosome C. (**a**) Proportion of polymorphic sites θ_w. (**b**) Population-scale recombination rate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

Figure 3. Mutational spectrum
(**a**) Substitution rates of the six different types of polymorphisms at 3rd codon positions in the *L. kluyveri* genome, polarized against *L. cidri*. Error bars represent the confidence intervals at 95%. (**b**) Fraction of GC→AT and AT→GC substitutions polarized against *L. cidri* in Sakl0C-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.

Figure 4. Decay of linkage disequilibrium 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.

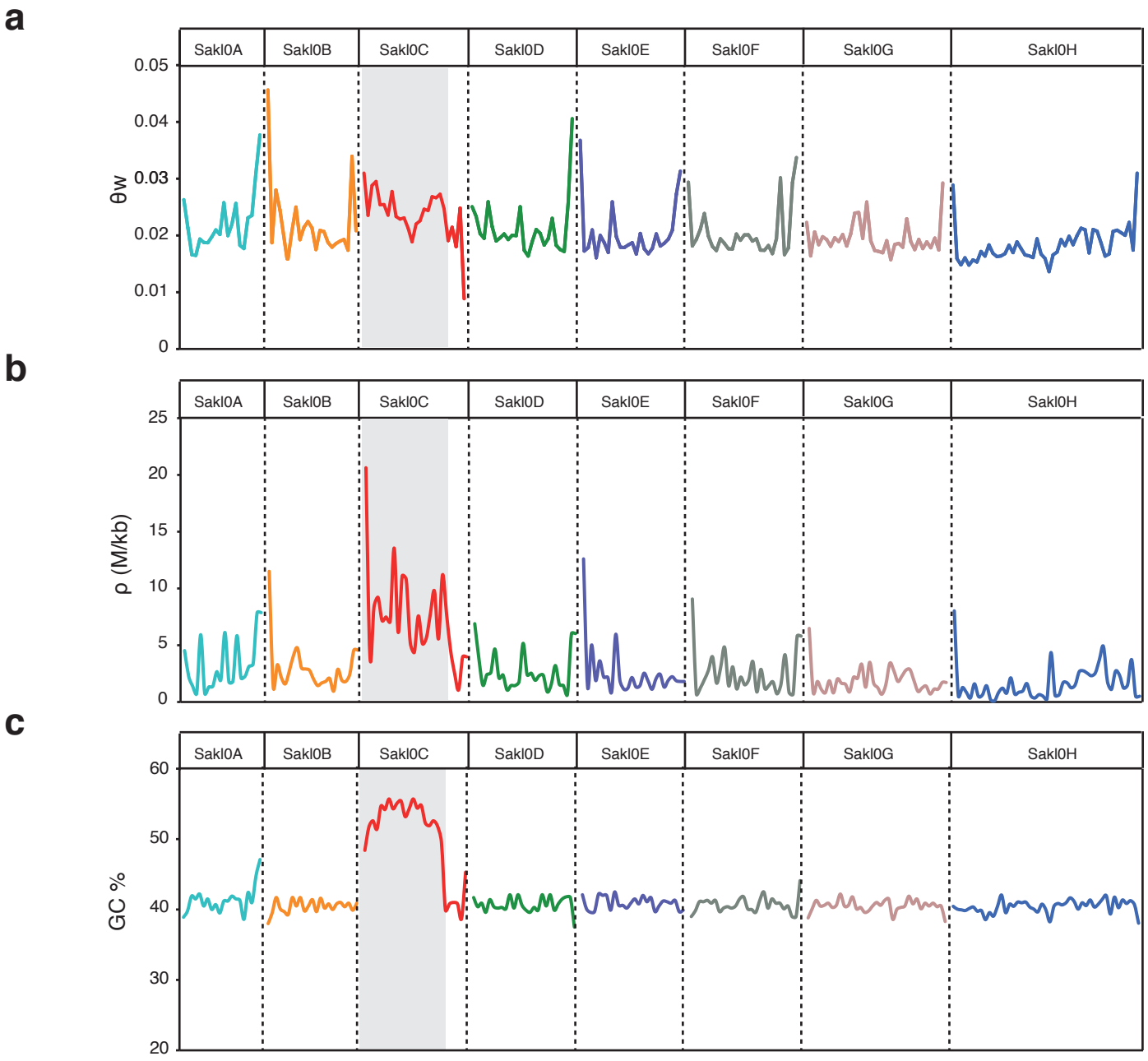
Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Figure 1.



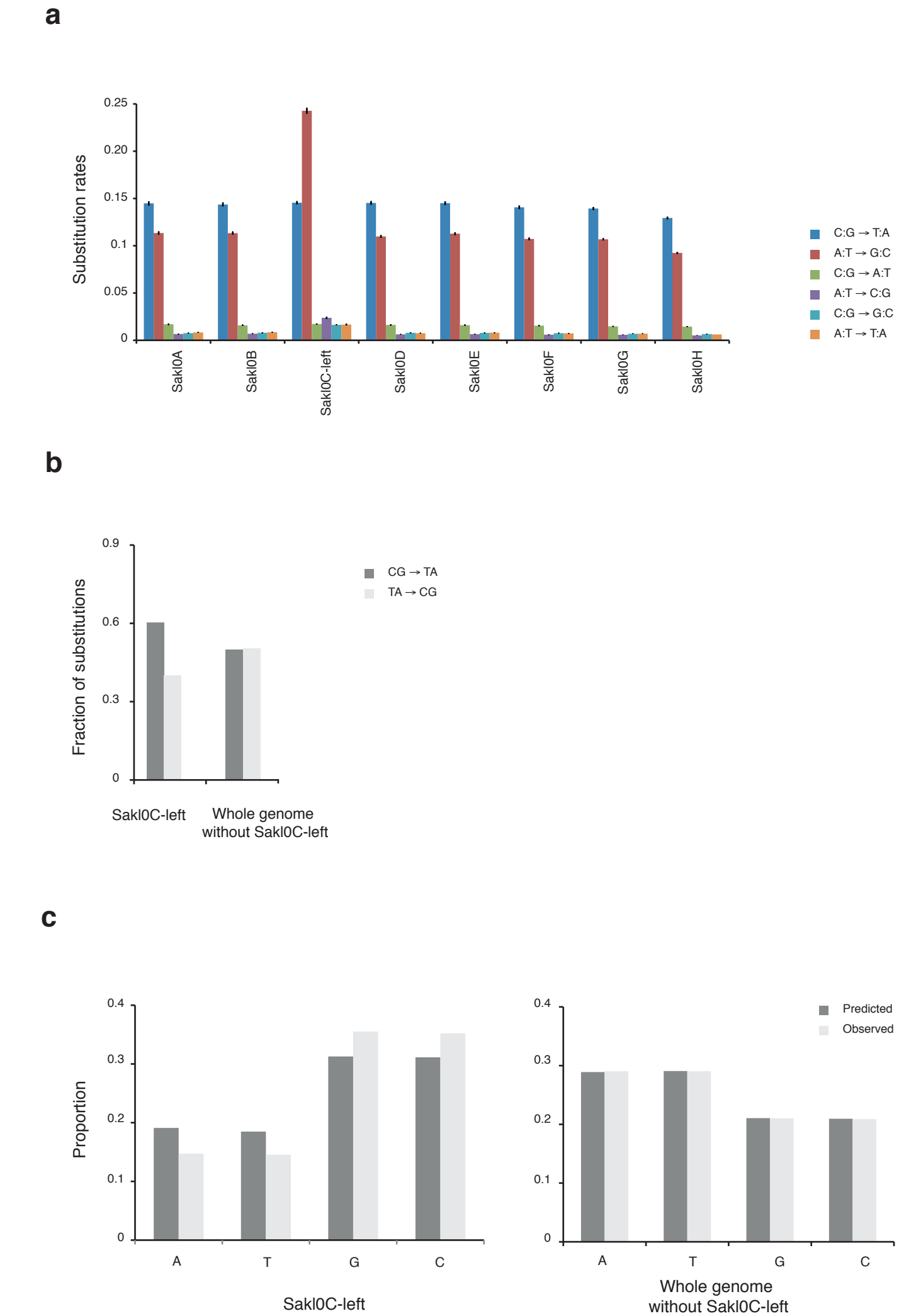
Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Figure 2.



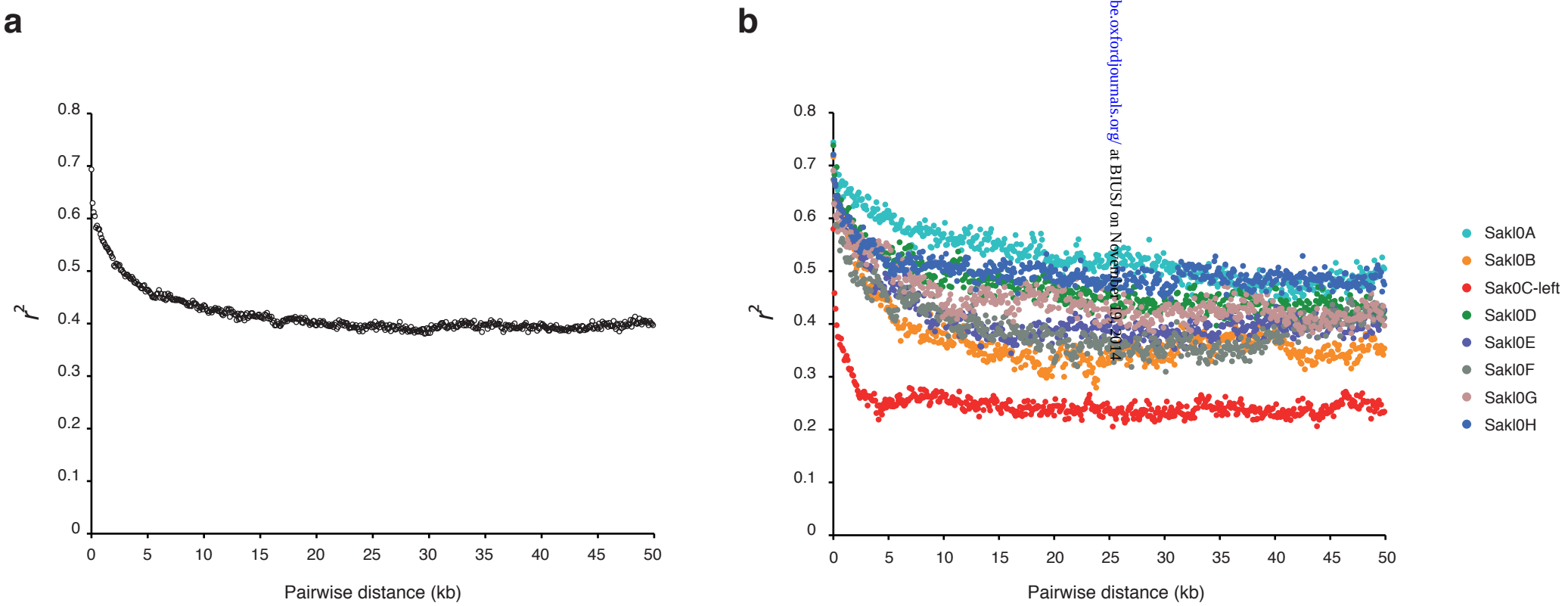
Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Figure 3.



Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014

Figure 4.



Downloaded from http://mbe.oxfordjournals.org/ at BIUSJ on November 19, 2014