nature genetics

Population genomic analysis of outcrossing and recombination in yeast

Douglas M Ruderfer, Stephen C Pratt, Hannah S Seidel & Leonid Kruglyak

The budding yeast *Saccharomyces cerevisiae* has been used by humans for millennia to make wine, beer and bread[1]. More recently, it became a key model organism for studies of eukaryotic biology and for genomic analysis[2]. However, relatively little is known about the natural lifestyle and population genetics of yeast[3]. One major question is whether genetically diverse yeast strains mate and recombine in the wild. We developed a method to infer the evolutionary history of a species from genome sequences of multiple individuals and applied it to whole-genome sequence data from three strains of *Saccharomyces cerevisiae*[2,4] and the sister species *Saccharomyces paradoxus*[5]. We observed a pattern of sequence variation among yeast strains in which ancestral recombination events lead to a mosaic of segments with shared genealogy. Based on sequence divergence and the inferred median size of shared segments (~2,000 bp), we estimated that although any two strains have undergone approximately 16 million cell divisions since their last common ancestor, only 314 outcrossing events have occurred during this time (roughly one every 50,000 divisions). Local correlations in polymorphism rates indicate that linkage disequilibrium in yeast should extend over kilobases. Our results provide the initial foundation for population studies of association between genotype and phenotype in *S. cerevisiae.*

S. cerevisiae has a complex life cycle[3]. It can reproduce through asexual cell division, either as a diploid or as a haploid. The haploid spores produced when a diploid undergoes meiosis can mate with each other in the ascus. In homothallic strains, the haploids can also undergo mating type switching after cell division, and the mother and daughter cells can then mate to form a diploid. In principle, a haploid cell can also mate with a haploid cell of opposite mating type from a different, genetically divergent strain. Such outcrossing would have important consequences for the population structure of the species and the patterns of sequence variation across the genome. Although there is limited recent evidence that outcrossing takes place in *S. cerevisiae*[6,7] and in its closest relative *S. paradoxus*[8], its frequency in nature is unknown. Traditional approaches for estimating rates of outcrossing rely on the frequency of heterozygous genotypes across several loci and require sampling a large number of individuals[9–12]. We developed a method that estimates rates of outcrossing from sequence data for only three individuals of the same species and an outgroup but that requires extensive contiguous sequence. Availability of whole-genome sequence for three independent strains of *S. cerevisiae*[2,4] and for *S. paradoxus*[5] allowed us to estimate the rate of outcrossing in *S. cerevisiae.*

We sought evidence for outcrossing in *S. cerevisiae* by independently assessing the genealogies of a dense set of polymorphic loci. Under the null hypothesis of zero recombination among strains (that is, no outcrossing since their last common ancestor), all polymorphisms should support the same genealogical tree. Recombination disrupts this uniformity by juxtaposing chromosomal segments from different lineages. The greater the frequency of recombination, the shorter the distance over which loci will share a particular tree. Thus, we can estimate the frequency of outcrossing in *S. cerevisiae* by computing the distance over which inter-strain polymorphisms share genealogies. To implement this approach, we identified a large set of SNPs in a multiple alignment of whole-genome sequences for three *S. cerevisiae* strains: the common laboratory strain S288C, the vineyard isolate RM11-1a, and the clinical isolate YJM789 (ref. 4). Despite its long time in the laboratory, S288C does not seem to show an accelerated rate of evolution[13]. To root the trees inferred at each SNP, we also aligned the sequence of *S. paradoxus*, the closest known relative of *S. cerevisiae*[5]. The alignment successfully covered ~87% of the fully sequenced S288C genome, with the unaligned portion reflecting incomplete sequence, insertions, deletions and chromosomal rearrangements.

The alignments uncovered a total of 77,776 SNPs, from which we identified a subset of 25,254 that parsimoniously supported a single genealogy for the three strains. These tree-diagnostic SNPs have two alleles, one of which is shared between *S. paradoxus* and a single *S. cerevisiae* strain. At these loci, only one phylogenetic tree is consistent with the minimum possible number of mutations—one (**Fig. 1**). We inferred the location of ancestral recombination events by determining where a run of tree-diagnostic SNPs supporting a given tree gives way to one supporting an alternative tree (see Methods). The commonly inherited chromosome segments between these inferred ancestral recombination events had a median length of 943 bp and

Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA. Correspondence should be addressed to L.K. (leonid@genomics.princeton.edu).

Received 3 May; accepted 10 July; published online 6 August 2006; doi:10.1038/ng1859

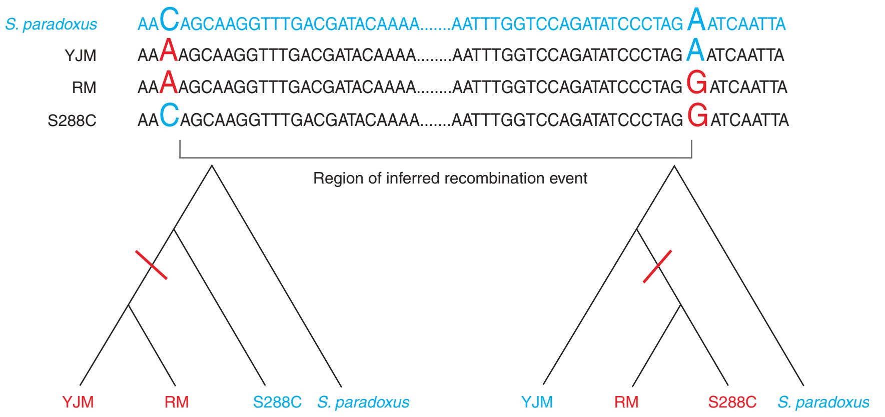
S. paradoxus AACAGCAAGGTTTGACGATACAAAA.......AATTTGGTCCAGATATCCCTAG A ATCAATTA
YJM AAAAGCAAGGTTTGACGATACAAAA........AATTTGGTCCAGATATCCCTAG A ATCAATTA
RM AAAAGCAAGGTTTGACGATACAAAA........AATTTGGTCCAGATATCCCTAG G ATCAATTA
S288C AACAGCAAGGTTTGACGATACAAAA.......AATTTGGTCCAGATATCCCTAG G ATCAATTA
Region of inferred recombination event
YJM
RM
S288C
S. paradoxus
YJM
RM
S288C
S. paradoxus

Figure 1 Inference of ancestral recombination from tree-diagnostic SNPs. Under the assumption of maximum parsimony (a single mutation per site), we can unambiguously determine that strains YJM and RM share the most recent common ancestor at the polymorphic site on the left, and strains RM and S288C share the most recent common ancestor at the polymorphic site on the right. Because the two polymorphic sites have different genealogies, we can infer that an ancestral recombination event occurred between them. Blue denotes ancestral sequence; derived alleles are shown in red, with red lines marking the branches on which the single mutation producing each derived allele is inferred to have occurred.

a mean length of 2,724 bp, with 4,473 inferred ancestral recombination events across the genome (**Fig. 2**). This provides clear evidence of historical outcrossing since the common ancestor of these strains.

To improve this estimate of ancestral recombination, we considered several potential sources of error in defining commonly inherited segments. Errors in sequencing or alignment and violations of the assumption of parsimony (that is, occurrence of multiple mutations at a given site) can artificially shorten or lengthen segment lengths by creating spurious tree-diagnostic SNPs or concealing true ones. In order to account for these errors, we built a hidden Markov model (HMM)[14] that incorporated both the tree-diagnostic SNPs and an additional 45,176 SNPs that do not parsimoniously specify a single genealogical tree but nonetheless carry probabilistic phylogenetic information. This additional set includes two types of loci: those in which no *S. cerevisiae* strain carries the *S. paradoxus* allele and those where two of the *S. cerevisiae* strains carry the *S. paradoxus* allele. The HMM infers the true genealogy at each locus by considering both the genealogy at neighboring loci and the observed alleles at the locus itself (Methods). This allows the partially informative SNPs to resolve ambiguities caused by multiple mutations at a site or by errors in sequencing or alignment. In addition, the HMM explicitly takes into account the greater probability for a change in genealogy with increasing distance between SNPs, thus minimizing the distorting effects of sequence gaps and regions of low polymorphism. With this model, we inferred a median interval between ancestral recombination breakpoints of 2,002 bp and a mean interval of 4,910 bp, with 2,506 inferred ancestral recombination events across the genome (**Fig. 2**). The nearly 50% decrease in ancestral recombination frequency relative to the raw estimate suggests that the HMM acted mainly to eliminate false tree-changing events. Nonetheless, the refined estimate still shows that outcrossing has occurred among these strains.

We used the observed frequency of ancestral recombination events to estimate the effective population recombination rate R by performing coalescent simulations of the genealogies of the three strains under the standard neutral model of population genetics[15] for a range of R values (see Methods). The simulations were constrained to have the same polymorphism rate as that observed in the actual data. We applied the HMM to the simulated data to infer recombination breakpoints and compute the expected length of segments between them as a function of R (**Fig. 3**). The observed median segment length of 2,002 bp was matched most closely by simulated data with $R = 2.2 \times 10^{-3}$, which had a median segment length of 1,992 bp, and the range of R values consistent with the data could be confined to 2×10^{-3} to 2.4×10^{-3} (**Fig. 3**). Similar results were obtained using the mean segment length after correction for gaps in the sequence alignment (see Methods). The distribution of shared segment lengths in simulations with $R = 2.2 \times 10^{-3}$ closely matched the distribution inferred from the data with the HMM (**Fig. 2**).

In the standard neutral model[15], R is given by $2G_Or$, where G_O is the number of generations of outcrossing since the most recent common ancestor of any pair of strains, and r is the recombination rate per nucleotide per generation. In yeast, the average recombination rate is 0.35 cM/kb (ref. 16) or $r = 3.5 \times 10^{-6}$. Thus, an estimate of $R = 2.2 \times 10^{-3}$ translates into an estimate of the number of outcrossing events of $G_O = 314$. An estimate of the total number of generations (G_T) since the most recent common ancestor of any pair of strains (that is, the number of cell divisions) can be obtained from the observed population mutation parameter $\theta = 2G_T\mu$, which, under the standard neutral model, is equal to the average pairwise sequence divergence, observed from the alignments to be 5.8×10^{-3}. A mutation rate $\mu = 1.84 \times 10^{-10}$ mutations per nucleotide per generation[7] then corresponds to $G_T = 1.58 \times 10^7$. Combining this estimate of the total number of generations with the estimate of the number of generations of outcrossing, we obtain the result that a true

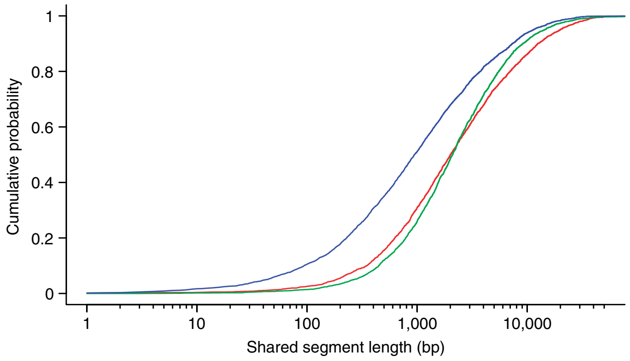

Figure 2 Cumulative distribution of the length of shared ancestral segments. Blue line indicates lengths inferred based on maximum parsimony, red line indicates lengths inferred with a hidden Markov model (HMM), and green line indicates lengths from a coalescent simulation with values of population mutation parameter $\theta = 0.0058$ and population recombination parameter $R = 0.0022$, as inferred from the data. Note that the effect of the HMM is to produce longer segments by eliminating spurious ancestral recombination events. The distribution of lengths inferred with the HMM is in close agreement with expectation from coalescent theory but has a slight excess of both short and long segments relative to expectation, possibly as a result of variable recombination rates.

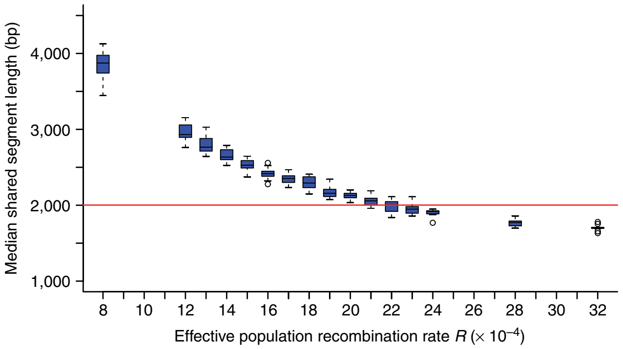

Figure 3 Coalescent simulations of median shared segment lengths as a function of population recombination parameter *R*. Each value of *R* represents 10–20 simulations, with 20 simulations performed for *R* values producing shared segment lengths most closely resembling the data (0.0012–0.0024). In box plots, the ends of each box mark the upper and lower quartiles, the line inside the box gives the median, and the brackets extend to the furthest data point no more than 1.5 times the interquartile range from the box, with more distant outliers shown as individual circles. Red line indicates the median length inferred from the data with the HMM. Note that this length was never seen in simulations with $R \leq 0.0020$ or $R \geq 0.0024$.

outcrossing event is quite rare, occurring on the order of once every 50,000 generations. The rate of cell division of yeast in the wild is not well characterized, but it may fall in the range of 1–8 divisions per day[7]. The generation estimates above then correspond to approximately 5,400–43,000 years since the most recent common ancestor of any two strains, with outcrossing occurring once every 17–137 years.

This estimate of the frequency of outcrossing depends on several assumptions, including random mating among unrelated strains, constant population size and no selection[15]. All methods for estimating recombination, including ours, can detect only a fraction of all ancestral recombination events—those that change the genealogy[11]. If the assumptions above are satisfied, our method correctly estimates the true underlying rate of recombination by using coalescent simulations to determine the correspondence between it and the frequency of inferred ancestral recombination events. Deviations from these assumptions, such as population subdivision, natural selection and changes in population size, can alter this correspondence, typically by increasing the fraction of undetectable recombination events and thus leading to an underestimate of the true frequency of outcrossing[9,10]. Because our results are based on analysis of the entire genome, are not altered by removal of nonsynonymous SNPs and are not different for coding regions (see Methods), the assumption of neutral evolution seems reasonable. Although the precise degree of population subdivision in *S. cerevisiae* is not well characterized, it is not very high based on surveys to date, and it seems to be due to ecological rather than geographic factors[6,7,17]. For moderate levels of subdivision, we would expect at most a several-fold increase in the frequency of outcrossing (see Methods), corresponding to more frequent mating between more closely related strains.

We observed a pattern of sequence variation among yeast strains in which ancestral recombination events lead to a mosaic of commonly inherited segments with a typical size of ~2,000 bp. This pattern should lead to correlation in polymorphism rates at nearby regions of the genome, with this correlation declining with distance but persisting over distances on the order of typical shared segment size[18]. To test this prediction, we examined variation in polymorphism rates along chromosomes (see Methods). Polymorphism rates of nearby regions were strongly correlated, with the correlation decaying to half-maximum for regions separated by 1,000 bp (**Fig. 4**). Residual correlation persisted out to distances of 10,000 bp. In order to show that this correlation was determined by locally shared genealogy rather than local variation in mutation rate, we compared the correlation of sequence divergence across each pair of strains with the same value measured across two distinct pairs of strains (Methods)[18]. As expected in the case of shared genealogies, the former measure (correlation in *cis*) but not the latter (correlation in *trans*) showed a high value at short distances and a decline with distance (**Fig. 4**). Correlation in *trans* was indistinguishable from zero at all distances, indicating little or no local variation in mutation rate. To directly estimate this variation, we analyzed the correlation of sequence divergence at nearby regions of the genome from pairwise alignments of the lab strain S288C and *S. paradoxus*. This correlation should largely reflect variation in mutation rate, as divergence times should not be locally variable. We observed appreciable correlation in divergence only over distances <500 bp (data not shown).

The dependence of linkage disequilibrium with distance tracks the correlation in polymorphism rates[18,19]; thus, substantial linkage disequilibrium in yeast should extend over kilobases. Notably, the extent of correlation in polymorphism rates in yeast is not very different from that in humans[18]. This similarity arises because although the ratio of genetic to physical distance is approximately 300 times higher in yeast than in humans, the number of generations of outcrossing since the most recent common ancestor, based on our results, is roughly 100 times fewer in yeast. The level of nucleotide diversity in yeast is several-fold higher than in human, but both the diversity and the extent of linkage disequilibrium are similar enough to suggest that studies of population genetics in yeast may provide useful insights for human studies. Our results provide the initial foundation for whole-genome studies of association between genotype and phenotype in large collections of *S. cerevisiae* strains.

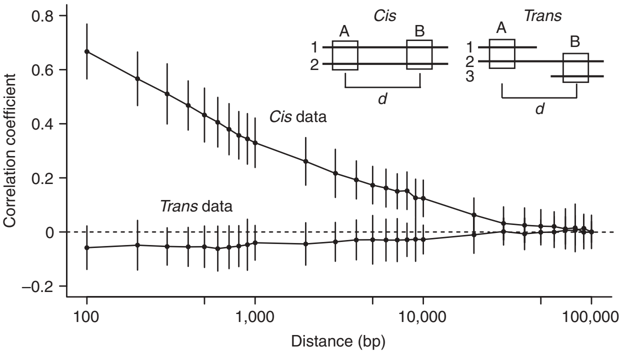

Figure 4 Correlation in polymorphism as a function of distance, corrected for stochastic variance (Methods). The lines indicate *cis* and *trans* correlations as labeled. The three pairwise strain comparisons showed very similar correlation values at each distance and were averaged to produce the plot. Error bars correspond to one standard deviation among correlations computed separately for each of the 16 yeast chromosomes and each pairwise comparison (a total of 48 values at each distance). *Cis* correlations are computed from pairwise polymorphism rates between sequences of two strains (1,2) over two regions of one hundred bases each (A,B) separated by a distance *d*. *Trans* correlations are computed from polymorphism rates between sequences of strains (1,2) for region A and between sequences of strains (2,3) for region B.

Yeast have a complex life cycle, having the ability to reproduce vegetatively as haploids or diploids and to mate with clonal cells as well as to outcross, and little has been known about the frequency and significance of outcrossing in nature. Our results show that outcrossing does take place and that although it is relatively infrequent, it is sufficient to result in mixing of genomes through recombination such that ancestral segments remain unbroken only over kilobases. Environments that provide opportunities for outcrossing may not be uncommon. For example, grapes have been found to carry 10^5–10^6 yeast cells on them, and several yeast strains have been isolated from a single grape, probably transported there by insects[20]. Our finding that yeast engage in true sexual reproduction, but that they do so infrequently, makes this species an interesting test case for theories of evolution and maintenance of sex[21]. Similar population-genomic techniques may be used to draw inferences about the evolutionary history of other species from genome sequences of multiple individuals.

METHODS

Strains and sequence assembly. The sequence of S288C was obtained from the Saccharomyces Genome Database (http://www.yeastgenome.org), those of RM11-1a and *S. paradoxus* from the Broad Institute (http://www.broad.mit.edu/annotation/fgi/), and that of YJM789 from the Stanford Genome Technology Center (http://www-sequence.stanford.edu/yjm789_public/). The S288C genome is fully sequenced and assembled. RM11-1a has been sequenced to 8× coverage, and the sequence is assembled into 17 super-contigs, which we matched to the corresponding S288C chromosomes by length. The YJM789 genome has been sequenced to 10× coverage and assembled into 295 contigs assigned to chromosomes. We assembled these contigs by aligning them with MUMmer to the corresponding S288C chromosome[22]. Our assembly included the largest non-overlapping significant matches. The *S. paradoxus* genome has been sequenced to 7× coverage and assembled into 832 contigs. We assembled these using the synteny map from the Saccharomyces Genome Database and again included the largest non-overlapping contigs.

Parsimonious phylogenetic analysis. All four genomes were simultaneously aligned using the software package MAVID with default settings[23]. From the alignment, we identified 77,776 SNPs in *S. cerevisiae* in which the allelic state for *S. paradoxus* was also known. Of these, 25,254 SNPs could be parsimoniously assigned a single phylogenetic tree (**Fig. 1**). These tree-diagnostic SNPs were biallelic, with a single *S. cerevisiae* strain sharing the *S. paradoxus* allele. Another 45,176 SNPs were partially informative but did not parsimoniously support a single tree, either because two *S. cerevisiae* strains shared the *S. paradoxus* allele, or because the *S. paradoxus* allele was distinct from either of the *S. cerevisiae* alleles. Another 192 SNPs had a different allele in each *S. cerevisiae* strain and thus provided no phylogenetic information.

We estimated the location of recombination breakpoints by scanning the alignments for tree-diagnostic SNPs that supported a different genealogy from the previous tree-diagnostic SNP. The distances between ancestral recombination events were estimated as the distances between these tree-changing SNPs.

In order to determine if our results were affected by selection on coding regions, we computed the distributions of shared ancestral segment lengths in coding regions and in noncoding regions, including only those segments that were entirely contained within a coding or noncoding region. Both distributions were shifted toward shorter segment lengths relative to the distribution for the entire genome, because longer segments typically cross the boundaries between coding and noncoding regions, and the average segment length was greater within coding regions because in yeast these are longer than noncoding regions. However, each distribution was the same as the equivalent distribution for regions of the same length obtained by shifting all coding and noncoding boundaries by 5,000 bp to randomize the occurrence of coding and noncoding sequences within a region. We also looked for an effect of selection by removing all nonsynonymous SNPs, computing the distribution of shared ancestral segment lengths and then randomly removing an equal number of either noncoding SNPs or synonymous SNPs and recomputing the distribution of segment lengths. Again, the distributions were not different.

Probabilistic phylogenetic inference. We probabilistically inferred the genealogy at each of the 77,584 biallelic SNPs (both tree-diagnostic and partially informative) using a hidden Markov model (HMM). The HMM included SNPs for which the ancestral allele was unknown (that is, where the *S. paradoxus* allele differed from either *S. cerevisiae* allele), but excluded tri-allelic SNPs, as these loci contain no phylogenetic information. An HMM is a probabilistic framework used to infer a hidden state (in our case the genealogy of a given SNP), given a set of observed states (in our case the alleles of each strain at a given SNP). The unknown probabilities (prior, emission, transition) needed to perform this inference are learned iteratively by maximizing the likelihood of the observed states as a function of these parameters[14]. Prior probabilities can be estimated by counting, over all states, the number of times each tree is most probable. Our estimates for the three possible trees (0.35, 0.33, 0.32) showed that each genealogy is essentially equally frequent across the genome. The emission probabilities specify the probabilities of observed allelic configurations at SNP sites conditional on each of the three genealogies. These differ between different types of SNPs. For the tree-diagnostic SNPs, where a single *S. cerevisiae* allele is the same as the *S. paradoxus* allele, we expect a single genealogy to be highly probable, with the low probabilities of the other two genealogies corresponding to errors or non-parsimonious events (that is, multiple mutations at a site). This is what we observe. For example, in the case in which only YJM carries the *S. paradoxus* allele, the emission probabilities were estimated to be 0.93, 0.02 and 0.04 for the genealogies in which the strain to branch off first is YJM, S288C or RM, respectively. For the ambiguous case in which the *S. paradoxus* allele is different from both *S. cerevisiae* alleles, we expect from coalescent theory that the genealogy in which the strain with the allele observed once branches off first occurs with probability 7/9, with the other two genealogies each having probability 1/9. Consistent with this expectation, in the case in which YJM carries the singleton allele, the genealogy probabilities, ordered as above, were 0.85, 0.06 and 0.08. Because the probability of a recombination event increases with distance, the likelihood of a change in genealogy between adjacent SNPs must be estimated as a function of intra-SNP distance. To account for this, the probabilities of all pairs of observed trees at adjacent sites were grouped by the distance between the sites and estimated independently. We grouped distances between SNPs into 25-bp bins up to distances of 2,500 bp, with all distances greater than 2,475 bp included in the last bin owing to the small number of such longer intervals. For each bin, we estimated the transition probabilities, which ranged from 1% for transitions to a different genealogy at very short distances (fewer than 25 bp) to ∼20% at the longest distances.

Using the Viterbi algorithm[24], we inferred the most likely sequence of genealogies across each chromosome and used the distances between changes in genealogies to estimate the number and distribution of shared ancestral segments.

The HMM software is freely available from the authors.

Neutral model simulation. We sought to compare our data with what would be expected under the standard neutral model, simulated with the software package MS[25]. We performed a series of 10 or 20 simulations for each value of R in the range 0.0008–0.0032. We performed 20 simulations for R values that produced shared segment sizes closest to our empirical results (0.0012–0.0024) (**Fig. 3**). For each simulation, we constrained the polymorphism rate to the observed rate of $\theta = 5.8 \times 10^{-3}$. We created sequences from the resulting segregating site information, aligned those sequences and performed the HMM analysis in the same manner as with the actual data. We simulated the effect of gaps in the real data by artificially inserting those gaps into the same genomic positions for the simulated data and excluding any simulated SNPs in those regions. A single simulation of the genome with a value of $R = 0.0022$ was used to plot the cumulative distribution of segment sizes in **Figure 2**. We modeled population subdivision with an island model with three demes of equal size and equal migration rates among the demes. In this model, a population migration parameter $M = 1$ corresponds to an F_{ST} of 0.4 (refs. 26,27). If the three strains are drawn from three different demes, then for this value of M, the actual population recombination parameter R is approximately 1.67-fold higher than our estimate based on random mating. Lower values of M would lead to greater underestimation of R, but much lower values would produce unrealistically high values of F_{ST}.

Sequence variation and autocorrelation. We computed variation across the genome by first creating overlapping 100-bp bins starting at every nucleotide position. For each bin with completely aligned sequence data, we calculated the polymorphism rate as the fraction of variable sites. We eliminated bins with excessively high rates (>10 variable sites), which are likely to be alignment or sequence artifacts. We calculated the spatial autocorrelation with stochastic variance correction[18] along the chromosomes over all non-overlapping bins that had complete sequence information for each pairwise strain comparison over distances ranging from 100–100,000 bp. The analysis was performed in both *cis* and *trans* as defined in ref. 18.

ACKNOWLEDGMENTS

We thank D. Botstein, R. Brem, H. Coller, J. Ronald and especially M. Rockman for helpful discussions and comments on the manuscript. We thank the Broad Institute Fungal Genome Initiative, led by B. Birren, for generating the sequences of RM11-1a and *S. paradoxus*, and the Stanford Genome Technology Center, especially L. Steinmetz and R. Davis, for generating the sequence of YJM789. This work was supported by US National Institute of Mental Health grant R37 MH059520 and a James S. McDonnell Foundation Centennial Fellowship. SCP is supported by a grant from the Pew Charitable Trusts (award 2000-002558).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Mortimer, R.K. Evolution and variation of the yeast (*Saccharomyces*) genome. *Genome Res.* **10**, 403–409 (2000).
2. Goffeau, A. *et al.* Life with 6,000 genes. *Science* **274**, 546–563–7 (1996).
3. Landry, C.R., Townsend, J.P., Hartl, D.L. & Cavalieri, D. Ecological and evolutionary genomics of *Saccharomyces cerevisiae*. *Mol. Ecol.* **15**, 575–591 (2006).
4. Gu, Z. *et al.* Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **102**, 1092–1097 (2005).
5. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
6. Aa, E., Townsend, J.P., Adams, R.I., Nielsen, K.M. & Taylor, J.W. Population structure and gene evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **6**, 702–715 (2006).
7. Fay, J.C. & Benavides, J.A. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.* **1**, 66–71 (2005).
8. Johnson, L.J. *et al.* Population genetics of the wild yeast *Saccharomyces paradoxus*. *Genetics* **166**, 43–52 (2004).
9. Carvajal-Rodriguez, A., Crandall, K.A. & Posada, D. Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. *Mol. Biol. Evol.* **23**, 817–827 (2006).
10. Hey, J. & Wakeley, J. A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846 (1997).
11. Hudson, R.R. & Kaplan, N.L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164 (1985).
12. Enjalbert, J. & David, J.L. Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics* **156**, 1973–1982 (2000).
13. Ronald, J., Tang, H. & Brem, R.B. Genome-wide evolutionary rates in laboratory and wild yeast. *Genetics*, published online 2 July 2006 (doi:10.1534/genetics.106.060863).
14. Baum, L.E. & Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37**, 1554–1563 (1966).
15. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302 (2000).
16. Cherry, J.M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67–73 (1997).
17. Liti, G., Peruffo, A., James, S.A., Roberts, I.N. & Louis, E.J. Inferences of evolutionary relationships from a population survey of LTR-retrotransposons and telomeric-associated sequences in the *Saccharomyces* sensu stricto complex. *Yeast* **22**, 177–192 (2005).
18. Reich, D.E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**, 135–142 (2002).
19. McVean, G.A. A genealogical interpretation of linkage disequilibrium. *Genetics* **162**, 987–991 (2002).
20. Mortimer, R. & Polsinelli, M. On the origins of wine yeast. *Res. Microbiol.* **150**, 199–204 (1999).
21. Otto, S.P. & Lenormand, T. Resolving the paradox of sex and recombination. *Nat. Rev. Genet.* **3**, 252–261 (2002).
22. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
23. Bray, N. & Pachter, L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* **14**, 693–699 (2004).
24. Viterbi, A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory* **13**, 260–269 (1967).
25. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
26. Hein, J., Schierup, M.H. & Carsten, W. *Gene Genealogies, Variation and Evolution* (Oxford Univ. Press, New York, 2005).
27. Hudson, R.R., Slatkin, M. & Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583–589 (1992).