Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*

Joseph Schacherer[1]*†, Joshua A. Shapiro[1]*, Douglas M. Ruderfer[1] & Leonid Kruglyak[1]

Comprehensive identification of polymorphisms among individuals within a species is essential both for studying the genetic basis of phenotypic differences and for elucidating the evolutionary history of the species. Large-scale polymorphism surveys have recently been reported for human[1], mouse[2] and *Arabidopsis thaliana*[3]. Here we report a nucleotide-level survey of genomic variation in a diverse collection of 63 *Saccharomyces cerevisiae* strains sampled from different ecological niches (beer, bread, vineyards, immunocompromised individuals, various fermentations and nature) and from locations on different continents. We hybridized genomic DNA from each strain to whole-genome tiling microarrays and detected 1.89 million single nucleotide polymorphisms, which were grouped into 101,343 distinct segregating sites. We also identified 3,985 deletion events of length >200 base pairs among the surveyed strains. We analysed the genome-wide patterns of nucleotide polymorphism and deletion variants, and measured the extent of linkage disequilibrium in *S. cerevisiae.* These results and the polymorphism resource we have generated lay the foundation for genome-wide association studies in yeast. We also examined the population structure of *S. cerevisiae*, providing support for multiple domestication events as well as insight into the origins of pathogenic strains.

With their small and compact genomes, the hemiascomycetes (the group of fungi that includes *S. cerevisiae*) represent a powerful model for comparative genomics and studies of genome evolution[4–6]. As a result, more than 18 hemiascomycetes species are either completely or partially sequenced. The availability of the sequence data has presented an unprecedented opportunity to evaluate DNA sequence variation and genome evolution in a phylum spanning a broad evolutionary range[7]. This wealth of data on interspecific sequence differences stands in contrast to our limited knowledge of sequence variation within *S. cerevisiae.* Because of its importance both to human activities and as a model system, we sought to generate a comprehensive view of sequence polymorphism in *S. cerevisiae.* To determine sequence variation at the nucleotide level, we hybridized genomic DNA from 63 ecologically and geographically diverse strains (Supplementary Table 1) to a high-density Affymetrix Yeast Tiling Microarray (YTM) and identified positions likely to differ from the reference sequence with the software package SNPscanner[8]. We detected a total of 1,896,131 single nucleotide polymorphisms (SNPs) in non-repetitive regions of the genome (Supplementary Table 1). Because of variation of up to a few base pairs (bp) in the location of SNPs detected by SNPscanner, we used a grouping procedure (see Methods) to identify the sites of polymorphic variation across strains. We also removed all singletons (SNPs called in only one strain) to reduce false positives further. This approach detected a total of 1,299,811 individual SNP calls, which were grouped into 101,343 distinct segregating sites. At each of these sites, every strain was classified as having either the same or different nucleotide relative to the reference strain (S288c).

We evaluated the coverage and accuracy of our polymorphism survey by comparing our data with the low-coverage sequence generated by ref. 9; 13 strains are shared between the two data sets. Most of the array-called SNPs with sequence data in the region had corresponding polymorphisms in the sequence data (median of 92% per strain), showing that our data have a low false-positive rate. Array-based polymorphism calls captured most (median of 73% per strain) of the high-quality (quality score >30), independent (>25 bp from the next closest polymorphism) SNPs present in the sequence data, showing that our data have high coverage. Discrepancies between array-based and sequence-based polymorphism calls probably reflect false positives and false negatives in each type of data, and may also derive from genuine sequence differences between strains with the same name but obtained from different sources by the two studies.

We detected an average of 30,097 SNPs per strain (Supplementary Table 1). Excluding laboratory strains, most of which are closely related to the reference strain, the frequency of polymorphisms varied between 0.0011 to 0.0041 per bp (0.0028 on average), representing an average density of 2.8 SNPs per kilobase (kb). Across all strains, we observed 8.35 non-singleton segregating sites per kb ($\theta_W \,\mathrm{kb}^{-1} = 2.26$). The frequency spectrum of the observed polymorphisms is highly skewed towards an excess of low-frequency alleles, even after corrections for the grouping procedure and genotyping errors (Supplementary Fig. 1). This excess of rare alleles resulted in a lowered value for the frequency-weighted measure of nucleotide diversity ($\pi \,\mathrm{kb}^{-1} = 1.92$). Some of the excess of low-frequency alleles can be attributed to the presence of slightly deleterious variants, which are kept at low frequency by negative selection but have not yet been purged from the population. We expect that mutations in coding regions are more likely to be deleterious than those in noncoding regions, resulting in a lower overall level of polymorphism in coding regions, and we do observe that coding regions are approximately 17% less polymorphic than noncoding regions (Table 1). The coding regions also show a larger excess of low-frequency polymorphism in their frequency spectrum (Supplementary Fig. 2). These trends are further emphasized in the 1,114 genes known to be essential in the reference strain S288c, which show both a lower overall level of polymorphism and a greater skew in the frequency spectrum. Noncoding regions are subject to selection on regulatory elements. Short intergenic regions should carry a higher proportion of functional regulatory sequences than longer noncoding regions, and we observe that intergenic regions shorter than 300 bp have significantly lower rates of polymorphism than longer regions (Supplementary Fig. 3). We found a markedly nonrandom distribution of polymorphism levels across the genome. We observed a decrease in SNP density within 25 kb of centromeres (Supplementary Fig. 4a). This observation is consistent with the lack of

[1]Lewis-Sigler Institute for Integrative Genomics, Department of Ecology and Evolutionary Biology and Howard Hughes Medical Institute, Princeton University, Princeton, New Jersey 08544, USA. †Present address: Department of Molecular Genetics, Genomics and Microbiology, Louis-Pasteur University and CNRS, UMR7156, Strasbourg 67083, France.
*These authors contributed equally to this work.

Table 1 | Functional variation in levels of polymorphism

Genomic region	Number of segregating sites	θ_W kb^{-1}	π kb^{-1}
Noncoding regions	33,465	2.69	2.33
Coding regions	71,420	2.23	1.89
Paralogues	14,774	2.28	1.93
Essential genes	13,107	2.08	1.77

In a standard neutral model $4N\mu$ is the population mutation rate, where N is the effective population size and μ the rate of mutation. θ_W is Watterson's theta: an estimate of $4N\mu$ based on the number of segregating sites. π is an estimate of $4N\mu$ equal to the mean pairwise difference among individuals.

DNA double-strand breaks (that is, the presence of meiotic recombination cold spots) near the centromeres[10]. By contrast, subtelomeric regions, which undergo frequent recombination[11], show higher variation at the sequence level in the regions 15–45 kb from telomeres (Supplementary Fig. 4b).

The genomic extent of linkage disequilibrium—nonrandom association of alleles at different polymorphic sites—provides information about recombination and population structure, and is also a critical parameter for population studies of association between genotype and phenotype. Our data provided the first opportunity to measure genome-wide properties of linkage disequilibrium across a large collection of diverse strains. We examined pairwise linkage disequilibrium for the 101,343 segregating sites and found that linkage disequilibrium falls to one-half of its maximum value at about 11 kb (Fig. 1). Because the yeast genome is physically compact (12 Mb), the 101,343 segregating sites reported here (nearly a site every 100 bp, of which close to one-half have a minor allele frequency $>$10%) provide a high-density polymorphism resource for *S. cerevisiae* from which an optimized panel of sites sufficient for whole-genome association studies in yeast can be chosen. To characterize further the architecture of linkage disequilibrium, we examined each of the sampling groups that contained at least 10 strains (wine, clinical, distillery and laboratory strains; Supplementary Fig. 5). In the wine strains, linkage disequilibrium falls to half of its maximum value at $\sim$2.5 kb, but is more extensive in clinical ($\sim$7 kb), distillery ($\sim$9.5 kb) and laboratory ($\sim$23.8 kb) strains. Because most of the laboratory strains are recently derived from the same founder strain S288c[12], linkage disequilibrium is expected to be greater than in the other groups. By contrast, the low level of linkage disequilibrium in the wine strains probably reflects a long time since the most recent common ancestor of these strains, and perhaps a higher frequency of outcrossing events.

To examine structural variation, we identified all deletion events $>$200 bp in the 63 strains (Supplementary Tables 1 and 2). We observed 3,985 deletions (an average of 63 per strain). The number of deletions varied from one in BY4716 (which is isogenic to the reference but carries an engineered deletion of *LYS2*) to 106 in YJM320. The deletions ranged in size from 200 bp to 13.8 kb, with nearly half falling between 200 bp and 400 bp (Supplementary Fig. 6).

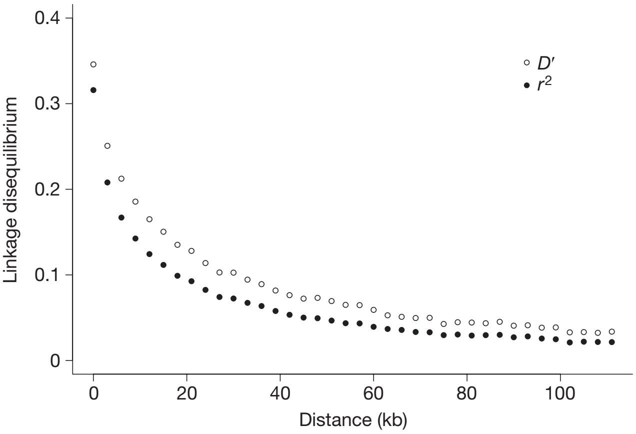

Figure 1 | Decay of linkage disequilibrium as a function of distance. Averages of pairwise linkage disequilibrium measures r^2 (black circles) and D' (open circles) are plotted for each bin of distances between pairs of SNPs. The linkage disequilibrium values were corrected for finite-size effects by subtracting the average value computed for a random subset of pairs of SNPs located on different chromosomes.

The deletions are unevenly distributed across the genome (Supplementary Fig. 7), with enrichment in subtelomeric regions (45.4% of events in $<$10% of the genome; Supplementary Fig. 7b) and a deficit near the centromeres (Supplementary Fig. 7a). These patterns are consistent with SNP rates and may similarly be explained by variation in recombination rates. A total of 254 genes contained a whole (119 genes) or partial (135 genes) deletion in at least one strain (Supplementary Table 3). Most were deleted in one to four strains, but some were deleted in many strains (Supplementary Fig. 8). For example, the gene YAR047C is deleted in 59 of the 63 surveyed strains. This gene is annotated as a dubious open reading frame (ORF) unlikely to encode a protein, on the basis of comparative sequence data of *Saccharomyces sensu stricto* species[4]. Our observation within the *S. cerevisiae* species strongly supports this hypothesis. Dubious ORFs accounted for 37 of the gene deletions. The set of deleted genes is enriched for those with known functions in transport, and in particular for sugar and hexose transporters (Supplementary Table 4). Most of these deleted genes are located in the subtelomeric regions. These results provide clear evidence of the importance of variation at subtelomeric regions in adaptation of strains to different carbon sources, as previously suggested[12,13].

We looked for deletions in genes known to be essential in the S288c strains[14]. We observed partial deletions in only four of the 1,114 essential genes (*KRS1*, *PGS1*, *SMT3* and *ERG20*), many fewer than the 49.6 genes that would be expected from the overall deletion frequency ($\chi^2 = 52$; $P < 0.0001$), which shows that the vast majority of the genes defined to be essential in the S288c background are also essential in all other genetic backgrounds of *S. cerevisiae*. With the exception of *KRS1*, these deletions were observed in only a few strains (Supplementary Table 3). Moreover, the deletions observed in the four essential genes affect a small fraction of the ORF, and the genes may still be functional. We examined more closely the partial deletion in *KRS1*, which encodes the lysyl-tRNA synthetase. We looked at the spore viability from crosses between the S288c reference strain and several of the strains (K1, CLIB219, K12 and Y9) in which the *KRS1* gene is partially deleted (Supplementary Fig. 9). We observed a high spore viability of around 90% in each cross, which shows that the *KRS1* gene is still functional in these strains. We also observed a reduced deletion rate in duplicated gene pairs derived from the whole genome duplication event (20 observed compared with 49.4 expected; $\chi^2 = 21.7$; $P < 0.0001$; Supplementary Table 3).

We sought to use the genome-wide genotypes at the 101,343 polymorphic sites across our diverse collection of strains to elucidate the phylogenetic relationships among strains and to evaluate the effects of ecological factors and geographical locations on strain diversity. We used standard neighbour-joining methods to build a majority-rule consensus tree of the surveyed strains (Fig. 2), and also analysed the data with the model-based clustering algorithm implemented in the program Structure[15] (Fig. 3). Both analyses showed at least three distinct subgroups based on the source from which the strains were isolated. Most of the wine strains (with the exception of CLIB219, which was isolated in Russia) are members of a single well-defined subpopulation. Because these wine strains were collected from dispersed locations, this observation provides strong evidence of a single domestication event of yeast for winemaking, followed by human-associated migration of wine yeast all over the world. The wine strains show the lowest level of polymorphism among the groups (Table 2), as well as an excess of low-frequency SNPs, consistent with a bottleneck during domestication. This subpopulation also includes a number of strains collected from distilleries, nature (soil, cocoa beans, prickly pear and *Tuber magnatum*) and clinical sources, indicating that these strains derived from domesticated wine strains, which transited out of

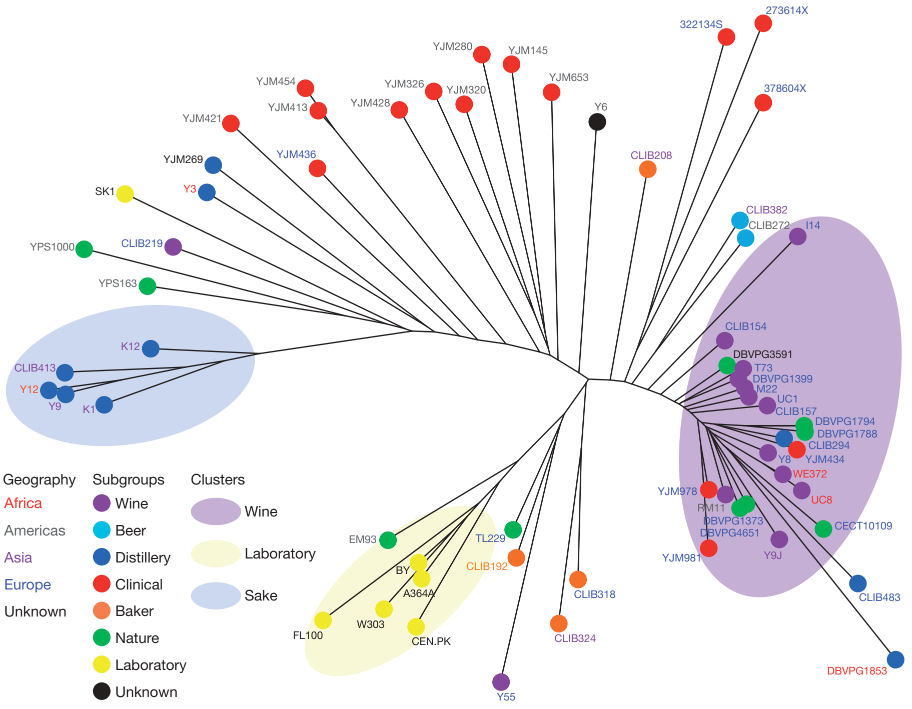

Figure 2 | Neighbour-joining tree of 63 *S. cerevisiae* strains. The tree was constructed on the basis of the 101,343 segregating sites identified in the surveyed strains. Branch lengths are proportional to the number of the 101,343 segregating sites that differentiate each pair of strains. Font colour of strain name denotes geographical origin and circle colour denotes ecological niche as specified in the key.

this group to other human-associated fermentations as well as back into nature and therefore escaped their human-manufactured environment. The second major population group contains the strains used for sake production and provides strong evidence for a second and independent domestication event, as hypothesized previously[16]. The laboratory strains, with the exception of SK1, form a third clear group, a consequence of the fact that most of the commonly used *S. cerevisiae* strains, with the exception of SK1, are derived from the S288c genetic background[12]. It is worth noting that the EM93 strain, the progenitor of S288c originally isolated from a rotting fig[17], is seen to be closely related to the laboratory strains. A number of strains did not fall into clear groups on the tree and did not cluster into coherent groups in the Structure analysis; their genomes seem to be mosaics of contributions from the three genetically distinct subgroups.

Although *S. cerevisiae* is usually considered to be a benign organism, there is a growing recognition that it can be a cause of opportunistic pathogenic fungal infection, typically, but not exclusively, in immunocompromised individuals[18]. To investigate the origin of these strains, we examined 16 strains isolated from different clinical sources (for example, blood, mouth, sputum) in Europe and the Americas (Supplementary Table 1). The clinical isolates were broadly distributed across the tree, and did not cluster with each other or with any one subgroup of strains in the Structure analysis (Figs 2 and 3). Three European clinical strains (YJM434, YJM978 and YJM981) were closely related to wine strains. Three other European strains from the same geographical origin (Newcastle, UK) were closely related to each other, and had some similarity to beer and baker strains. The remaining ten strains (nine American, one European) branched from

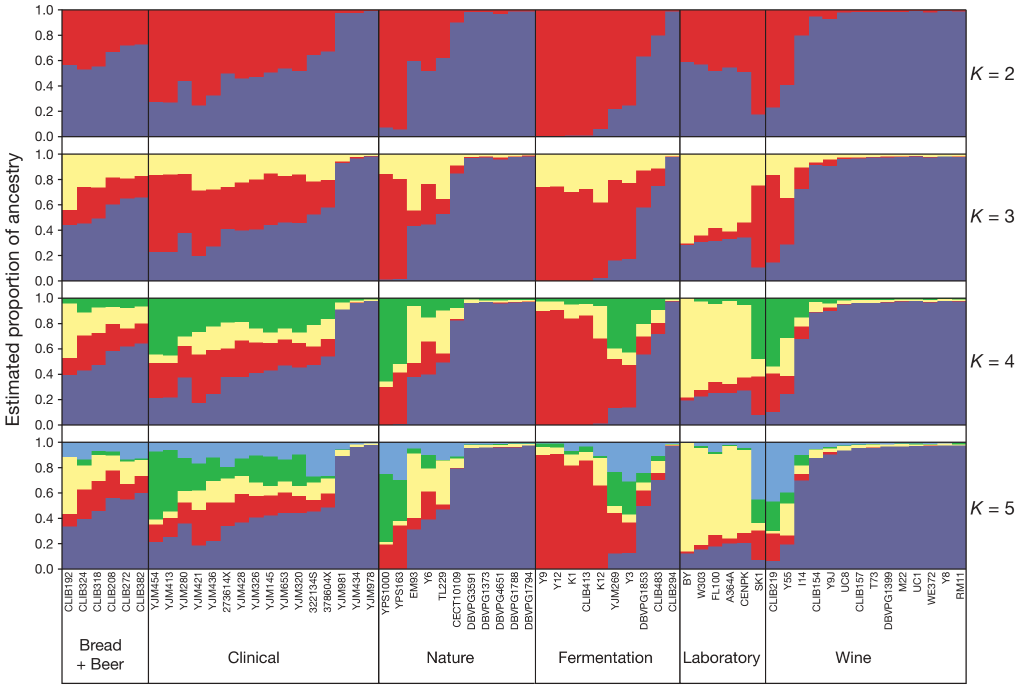

Figure 3 | Population structure of 63 *S. cerevisiae* strains. Cluster results from a structure analysis on 101,343 segregating sites identified in the 63 surveyed strains. The Structure program implements a Bayesian model-based clustering algorithm that attempts to identify genetically distinct subpopulations on the basis of patterns of allele frequencies[15]. Each strain is represented by a single vertical bar, which is partitioned into *K* coloured segments that represent the strain's estimated ancestry proportion in each of the *K* clusters.

Table 2 | Polymorphism and SNP frequency within groups

Strains	θ_W kb^{-1}	π kb^{-1}
All	2.26	1.92
Wine	1.38	1.25
Clinical	2.53	2.54
Distillery	2.54	2.60
Nature	2.03	1.97

See legend to Table 1.

a similar part of the tree, but did not appear to be closely related to each other or to any other coherent group of strain. Our interpretation of these results is that clinical isolates do not derive from a common ancestor or any one type of strain, but rather represent multiple events in which strains present in the environment opportunistically colonize human tissues. Our data provide strong evidence that wine strains are capable of such colonization, and suggest that strains from other sources (beer, bakery, laboratory, nature) can also do this. These results are consistent with clinical reports of patients infected with *S. cerevisiae* baker's strains and with the strain *Saccharomyces boulardii*, which is used therapeutically to treat diarrhoea and is also sold as a probiotic nutritional supplement[19]. Because the main environmental niches for *S. cerevisiae* in nature are not known, clinical strains might represent the best approximation of the overall species diversity of *S. cerevisiae*.

The polymorphism resource we generated, made freely available in the Yeast SNPs Browser database (http://gbrowse.princeton.edu/yeastSNP), enables genome-wide association studies of the phenotypic differences among these and other yeast strains. Phenotypic diversity among yeast isolates is significant, and variation is apparent among the surveyed strains at different levels. The genetic basis of a number of interesting phenotypes can be studied in yeast, including growth at high temperature, sporulation efficiency, telomere length, gene expression and response to drugs[20–24]; these studies can now move from linkage in crosses between two strains to the population level. *S. cerevisiae* provides a powerful model system for studies of complex traits because of the ease with which genetic analyses and phenotyping can be carried out and the ability to engineer and test the effects of individual polymorphisms and their combinations on different genetic backgrounds.

Our analysis also provides insight into the population structure of this yeast species. We show evidence for genetic differentiation of three distinct subgroups based on the source from which the strains were isolated: vineyards, sake and related fermentations, and laboratory strains. Thus, population structure at least partly reflects different ecological niches. Surveys of additional strains are needed to resolve fully the roles of ecology versus geography in the genetic differentiation of this species. Our data strongly support the hypothesis that these three groups represent separate domestication events, and that *S. cerevisiae* as a whole is not domesticated. Finally, our results suggest that *S. cerevisiae* strains from a range of environments are capable of opportunistic colonization of human tissues.

METHODS SUMMARY

Genomic DNA was extracted from 63 yeast strains (listed in Supplementary Table 1) and hybridized to Affymetrix Yeast Tiling Arrays. We used SNPscanner[8] to identify putative SNPs in each of the 63 strains on the basis of the hybridization intensity at each probe. Because there is error in the precise location of SNP calls made by SNPscanner, we used a grouping procedure (described in Methods) in order to integrate SNP calls across strains and minimize the effects of erroneous positive and negative calls.

We constructed a neighbour-joining tree of the 63 strains from the SNP data using Splitstree[25], with branch lengths proportional to the number of segregating sites that differentiate each node. To infer the population ancestry of the strains we used Structure[15], with ancestral population numbers between two and six. We calculated linkage disequilibrium across the genome using two standard metrics: D' and r^2, both for the whole genome and for each subpopulation. We calculated other population genetic summary statistics using code based on the libsequence package[26], and performed coalescent simulations of genome evolution using FastCoal[27], with corrections for expected error rates and our grouping procedure.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 June; accepted 25 November 2008.
Published online 11 February 2009.

1. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
2. Frazer, K. A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).
3. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342 (2007).
4. 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).
5. Cliften, P. *et al.* Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301**, 71–76 (2003).
6. Dujon, B. *et al.* Genome evolution in yeasts. *Nature* **430**, 35–44 (2004).
7. Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.* **22**, 375–387 (2006).
8. Gresham, D. *et al.* Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. *Science* **311**, 1932–1936 (2006).
9. Carter, D. M. *et al.* Population genomics of domestic and wild yeasts. *Nature* (submitted).
10. Gerton, J. L. *et al.* Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae. Proc. Natl Acad. Sci. USA* **97**, 11383–11390 (2000).
11. Pryde, F. E., Gorham, H. C. & Louis, E. J. Chromosome ends: all the same under their caps. *Curr. Opin. Genet. Dev.* **7**, 822–828 (1997).
12. Schacherer, J. *et al.* Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. *PLoS ONE* **2**, e322 (2007).
13. Winzeler, E. A. *et al.* Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics* **163**, 79–89 (2003).
14. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
15. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
16. Fay, J. C. & Benavides, J. A. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae. PLoS Genet.* **1**, 66–71 (2005).
17. Mortimer, R. K. & Johnston, J. R. Genealogy of principal strains of the yeast genetic stock center. *Genetics* **113**, 35–43 (1986).
18. Enache-Angoulvant, A. & Hennequin, C. Invasive *Saccharomyces* infection: a comprehensive review. *Clin. Infect. Dis.* **41**, 1559–1568 (2005).
19. de Llanos, R., Querol, A., Peman, J., Gobernado, M. & Fernandez-Espinar, M. T. Food and probiotic strains from the *Saccharomyces cerevisiae* species as a possible origin of human systemic infections. *Int. J. Food Microbiol.* **110**, 286–290 (2006).
20. Steinmetz, L. M. *et al.* Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).
21. Deutschbauer, A. M. & Davis, R. W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nature Genet.* **37**, 1333–1340 (2005).
22. Gatbonton, T. *et al.* Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.* **2**, e35 (2006).
23. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
24. Perlstein, E. O., Ruderfer, D. M., Roberts, D. C., Schreiber, S. L. & Kruglyak, L. Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nature Genet.* **39**, 496–502 (2007).
25. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
26. Thornton, K. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**, 2325–2327 (2003).
27. Marjoram, P. & Wall, J. D. Fast ''coalescent'' simulation. *BMC Genet.* **7**, 16 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to all the researchers and institutions, and especially J. Fay, for sharing yeast strains. We thank K. Dolinski and J. Matese for technical support and D. Gresham for comments on the manuscript. The authors acknowledge discussions with members of the Kruglyak and Botstein laboratories. This work was supported by NIH grant R37 MH059520 and a James S. McDonnell Foundation Centennial Fellowship to L.K., and NIH grant GM071508 to the Lewis-Sigler Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.K. (leonid@genomics.princeton.edu).

METHODS

Yeast strains. Yeast strains were obtained from a number of laboratories: J. Fay (Washington University), J. Perez-Ortin (University of Valencia), G. Liti and E. Louis (The University of Nottingham), J. McCusker (Duke University) and Jean-Luc Souciet (Louis-Pasteur University). We also purchased strains from different yeast culture collections: CLIB (Collection de Levures d'Intérêt Biotechnologique), CBS (Centraalbureau voor Schimmelcultures), DBVPG (Dipartimento di Biologia Vegetale e Agroambientale of the University of Perugia) and CECT (Coleccion Espanola de Cultivos Tipo). Strains used in this study are listed in Supplementary Table 1.

SNP identification. Yeast strains were grown in yeast extract, peptone and dextrose (YPD) medium. Total genomic DNA was purified from 30 ml YPD culture using Qiagen Genomic-Tips 100/G and Genomic DNA buffers as per the manufacturer's instructions. Genomic DNA was digested with DNaseI, labelled and hybridized to Affymetrix Yeast Tiling Arrays (YTMs) as described previously[8]

We used SNPscanner[8] to identify putative SNPs in each of the 63 strains on the basis of the hybridization intensity of DNA at each probe. SNPs from each strain were independently called against the reference FY3 genome using the following parameters: lod score >2, number of probes covering a base >1, and positive region length >6. These parameters are further described in ref. 8 and in the SNPscanner documentation (http://genomics-pubs.princeton.edu/SNPscanner/). With these parameters, we previously showed, using the complete genome sequence of strain YJM789, that 90.1% of true SNPs were detected, with only 49 false-positive SNP calls over the entire genome (a false-positive rate of 4×10^{-6} per bp).

Owing to the 4-bp resolution of the YTMs and the variance associated with DNA hybridization intensities, the SNP position predicted by SNPscanner may fall at varying positions surrounding the actual site of the SNP. This variance required us to perform a grouping procedure, combining all the calls within 6 bp of each other into a single segregating site. As the average density of putative SNPs is 1 per 6.3 bp, the probability of grouping two distinct sites is nontrivial. To reduce this probability, we implemented several heuristic filters: first, to reduce false positives, we required that at least one of the called SNPs in each grouping have a lod score >6. Second, we eliminated possible deletion events by removing putative SNPs with large prediction regions (>100 bp). Finally, we required at least 9 bp between each SNP in a group and the next closest call in the genome. We performed this grouping procedure in a top-down manner, by first grouping the SNPs with the most calls at a given position.

We tested the accuracy of this grouping procedure using a set of known high-confidence SNPs from the completely sequenced genomes of the strains S288c, RM11-1a and YJM145. Specifically, we examined 13,839 SNPs for which YJM145 and RM11-1a had the same allele and differed from the reference sequence. For this set, 12,578 (91%) and 11,518 (83%) SNPs were detected before grouping in YJM145 and RM11-1a, respectively. After grouping, 9,119 SNPs were detected in at least one strain, and 8,086 were correctly called in both strains, from which we infer a false-negative rate of 5.7% per strain, given detection in at least one strain. The grouping procedure almost never separated the same site into multiple sites (one case across the genome), and rarely combined two distinct sites (394 cases; $<5\%$ of sites after grouping). These cases are typically SNPs that are located within 4 bp of each other, closer than the theoretical resolution of the YTMs. We also removed all singletons (SNPs called in only one strain) to reduce false positives further.

Tree building and Structure analysis. We constructed a neighbour-joining tree of the 63 strains from the SNP data using the software package Splitstree[25], with branch lengths proportional to the number of segregating sites that differentiate each node. We ran Structure using the linkage model with the population number parameter, K, set from 2 to 6, for 100,000 iterations after a burn-in of 100,000 iterations, the first 50,000 of which were run under the free-recombination model[15].

Linkage disequilibrium. We calculated linkage disequilibrium across the genome using two standard metrics: D' and r^2. We computed these statistics for all pairs of sites located within a given distance, both for all the strains and within each predefined subpopulation. To correct for finite-size effects and differences in sample size among the subpopulations, we subtracted from each statistic the average value for a random subset of SNP pairs located on different chromosomes (which should not show linkage disequilibrium).

Polymorphism and divergence statistics. We calculated population genetic summary statistics of polymorphism using code based on the libsequence package[26]. To correct for the removal of singleton SNPs in the data set, modified estimators of the population mutation parameters θ_W and π were used[28]. An analogue of Tajima's D was calculated as the difference between these modified estimates[29]. To obtain significance values, we simulated under a modified coalescent model as described below, conditioning on the observed number of segregating sites and the approximate length of the sequence. The significance of an observed statistic was then taken to be the probability of observing a more extreme value in at least 10,000 simulations. Divergence rates were calculated from the multiple species alignments of ref. 4. We used PAML[30] to obtain maximum likelihood estimates of the rate of evolution along the *S. cerevisiae* branch after divergence from *Saccharomyces paradoxus*.

Coalescent simulations of genome evolution were performed using FastCoal[27]. Output from each coalescent simulation was run through a series of steps to mirror the sources of error inherent in the SNPscanner data. First, called SNPs were randomly removed with a probability of 5%. The addition of randomly missed calls creates a characteristic dearth of high-frequency SNPs in the data set; simulations under a 5% false-negative rate fit very closely with the observed pattern of polymorphism at high frequency. To correct for incorrectly grouped SNPs, we performed the previously described grouping procedure on all simulated data.

28. Fu, Y. X. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics* **138**, 1375–1386 (1994).
29. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
30. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).