Vol 458|19 March 2009|doi:10.1038/nature07743

Population genomics of domestic and wild yeasts

Gianni Liti[1]*, David M. Carter[2]*, Alan M. Moses[2,3], Jonas Warringer[4], Leopold Parts[2], Stephen A. James[5], Robert P. Davey[5], Ian N. Roberts[5], Austin Burt[6], Vassiliki Koufopanou[6], Isheng J. Tsai[6], Casey M. Bergman[7], Douda Bensasson[7], Michael J. T. O'Kelly[8], Alexander van Oudenaarden[8], David B. H. Barton[1], Elizabeth Bailes[1], Alex N. Nguyen Ba[3], Matthew Jones[2], Michael A. Quail[2], Ian Goodhead[2]†, Sarah Sims[2], Frances Smith[2], Anders Blomberg[4], Richard Durbin[2]* & Edward J. Louis[1]*

Since the completion of the genome sequence of *Saccharomyces cerevisiae* in 1996 (refs 1, 2), there has been a large increase in complete genome sequences, accompanied by great advances in our understanding of genome evolution. Although little is known about the natural and life histories of yeasts in the wild, there are an increasing number of studies looking at ecological and geographic distributions[3,4], population structure[5–8] and sexual versus asexual reproduction[9,10]. Less well understood at the whole genome level are the evolutionary processes acting within populations and species that lead to adaptation to different environments, phenotypic differences and reproductive isolation. Here we present one- to fourfold or more coverage of the genome sequences of over seventy isolates of the baker's yeast *S. cerevisiae* and its closest relative, *Saccharomyces paradoxus*. We examine variation in gene content, single nucleotide polymorphisms, nucleotide insertions and deletions, copy numbers and transposable elements. We find that phenotypic variation broadly correlates with global genome-wide phylogenetic relationships. *S. paradoxus* populations are well delineated along geographic boundaries, whereas the variation among worldwide *S. cerevisiae* isolates shows less differentiation and is comparable to a single *S. paradoxus* population. Rather than one or two domestication events leading to the extant baker's yeasts, the population structure of *S. cerevisiae* consists of a few well-defined, geographically isolated lineages and many different mosaics of these lineages, supporting the idea that human influence provided the opportunity for cross-breeding and production of new combinations of pre-existing variations.

The baker's yeast *S. cerevisiae* has had a long association with human activity[11], leading to the idea that its use in fermentation lead to its domestication. Two domestication events have been suggested, one for sake strains and one for wine[12]. In contrast, its closest relative, *S. paradoxus*, has never been associated with human activity and is found globally, sometimes in the same locations as *S. cerevisiae*[3,4]. A preliminary comparison within the *Saccharomyces sensu stricto* group exhibited extensive variation between *S. paradoxus* populations on different continents but limited variation among *S. cerevisiae* isolates and no correlation with geographic location[8].

Here we report nearly complete genome sequences of *S. cerevisiae* and *S. paradoxus* from a large variety of sources and locations (Supplementary Tables 1 and 2). The *S. cerevisiae* strains included the reference strain S288c plus other lab, pathogenic, baking, wine, food spoilage, natural fermentation, sake, probiotic and plant isolates. The *S. paradoxus* isolates were mostly from oak tree bark from the three recognized populations[6,8,13] as well as Siberia, Hawaii and the previously designated *Saccharomyces cariocanus*[14]. There is overlap among the general geographic sources of isolates from both species. The majority of strains were sequenced using Sanger sequencing on ABI 3730 DNA sequencers (Applied Biosystems). For some strains, sequence was obtained using an Illumina Genome Analyzer. Most strains were covered to one- to fourfold depth with a few covered more extensively (Supplementary Table 3). The sequence reads, assemblies, alignments, a BLAST tool and a genome browser are all publicly available[15].

We identified 235,127 high-quality single nucleotide polymorphisms (SNPs) and 14,051 nucleotide insertions or deletions (indels) in the *S. cerevisiae* nuclear genome, and 623,287 SNPs and 25,267 indels in *S. paradoxus*. Our S288c sequence differs from the reference genome by 498 high-quality, unambiguous SNPs (Supplementary Fig. 1). For 480 SNPs our S288c sequence is supported by other strains whereas the reference has no support, and for 18 SNPs the reference sequence is supported by other strains whereas ours is not. Many of the 480 SNPs are likely to represent errors in the reference sequence (Supplementary Table 4). The reference sequence for the type strain of *S. paradoxus*[16] was not complete, so we sequenced the type strain CBS432 to 4.3-fold coverage with an ABI 3730 sequencer and to 80-fold coverage with the Illumina Genome Analyzer.

Sequence surveys allow novel sequences not found in the reference genome to be identified. The proportions of unplaced reads for each strain are shown in Supplementary Table 2. We found 38 new hypothetical open reading frames (ORFs) in these sequences that are likely to be real. These ORFs are present in more than one strain (Supplementary Fig. 2), with some specific to a single lineage, such as the hypothetical protein 5 (Supplementary Information) in the West African lineage, which contains a conserved methyltransferase domain. Much of the unplaced material is subtelomeric. This is in contrast to a genome-wide analysis of copy number based on the numbers of reads of each strain aligning to each gene in the reference sequence, which showed very little significant copy number variation outside the ribosomal DNA (rDNA) region (Supplementary Information).

Neighbour-joining phylogenetic trees based on pairwise SNP differences in the alignments were generated (Fig. 1 and Supplementary Fig. 3). The *S. paradoxus* strains fall into the three previously described populations, plus one isolate from Hawaii. Most of the SNPs in *S. paradoxus* are private polymorphisms within each population,

[1]Institute of Genetics, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK. [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1HH, UK. [3]Department of Cell & Systems Biology, University of Toronto, Ontario M5S 2J4, Canada. [4]Department of Cell and Molecular Biology, Lundberg Laboratory, University of Gothenburg, Medicinaregatan 9c, 41390 Gothenburg, Sweden. [5]National Collection of Yeast Cultures, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK. [6]Division of Biology, Imperial College London, Silwood Park, Ascot SL5 7PY, UK. [7]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK. [8]Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. †Present address: School of Biological Sciences, University of Liverpool, Liverpool LG9 3BX, UK.
*These authors contributed equally to this work.

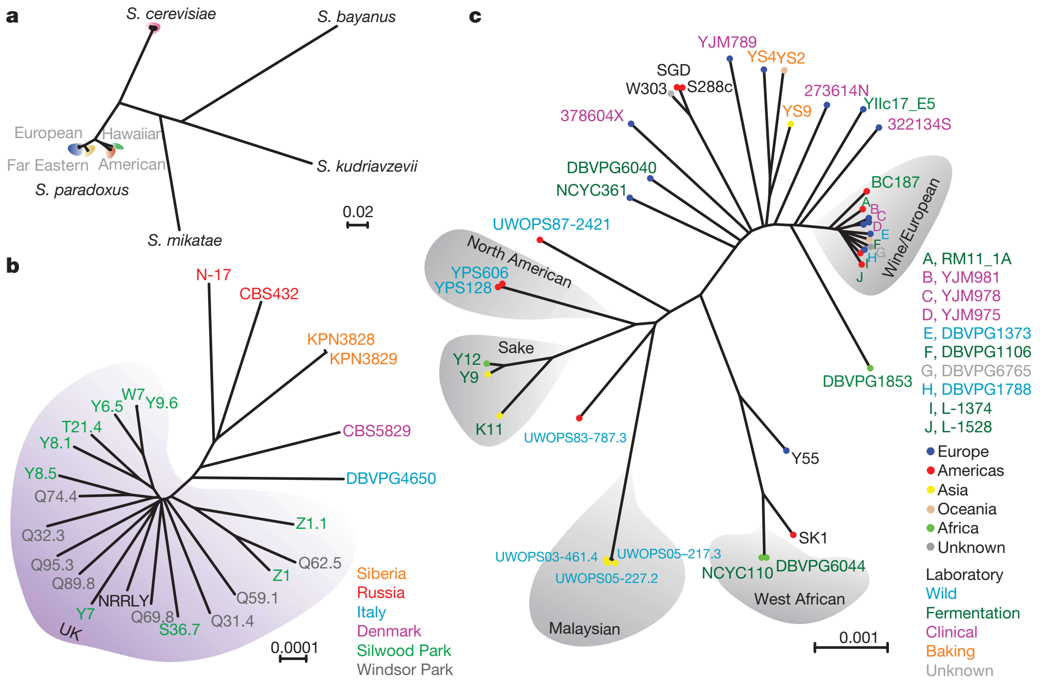

Figure 1 | *Saccharomyces* phylogenomics. **a**, Neighbour-joining trees based on SNP differences of *S. cerevisiae* and *S. paradoxus* strains sequenced in this project, using *Saccharomyces mikatae, Saccharomyces kudriavzevii* and *Saccharomyces bayanus* as out-groups. **b**, Close-up of the European *S. paradoxus*, with UK isolates highlighted in violet. **c**, *S. cerevisiae* strains with clean lineages highlighted in grey; colour indicates source (name) and geographic origin (dots). Scale bars indicate frequencies of base-pair differences.

resulting in a clear separation of the three populations[17] (Fig. 2a). The European population was sampled extensively, which provided a picture of within-population structure (Fig. 1b).

The *S. cerevisiae* population structure is more complex. There are five lineages that exhibit the same phylogenetic relationship across their entire genomes, which we consider to be 'clean' non-mosaic lineages (Fig. 1c). These are strains from Malaysia, West Africa, sake and related fermentations (labelled 'Sake' in Fig. 1c), North America, and a large cluster of mixed sources containing many European and wine strains ('Wine/European'). The remaining strains are on long branches between the Wine/European cluster and the other four clean lineages. Although some lineages correspond to geographic origin, such as those from North America and Malaysia, many closely related strains are from widely separated locations. This mixed architecture could be due to human traffic in yeast strains and subsequent recombination between them. Analysis with Structure is consistent with separate populations for the West African, Malaysian, Sake and Wine/European lineages (Fig. 2b). The North American isolates share some polymorphisms with all four separate populations, whereas the rest of the strains share polymorphisms with the European lineage and at least one other population. Analysis of SNP distributions (Supplementary Table 5) is consistent with the neighbour-joining tree phylogeny (Fig. 1c) and the Structure analysis (Fig. 2b). Each clean lineage is monomorphic for the majority of segregating sites, whereas the mosaics are polymorphic for the majority of sites.

Phylogenetic trees constructed for individual chromosomes or smaller segments (Supplementary Fig. 4) demonstrate the mosaic nature of these genomes, as do segmental comparisons (Supplementary Fig. 5). For example, the laboratory strains SK1 and Y55 appear to be the result of recent crosses between the West African lineage and the European lineage (Supplementary Fig. 5b). Similarly, W303 is a recent cross between the reference S288c lineage and one or more other lineages. Different segments of the mosaics fall into different locations in the neighbour-joining tree (Supplementary Fig. 4). The recently sequenced clinical derivative YJM789 (ref. 18) is another example. This complex population structure of *S. cerevisiae* is reported in a similar study[19] and is consistent with five well-delineated lineages, two of which contain isolates used in fermentation industries[12], plus a number of recombinant strains, many of which are also used for fermentation. Phenotypic profiling (see below), and analyses of rDNA repeat unit variation (Supplementary Fig. 6) and Ty element abundance (Supplementary Fig. 7 and Supplementary Table 6), produce results consistent with this overall picture of the *S. cerevisiae* population structure.

It is unlikely that the entire sequence space of *S. cerevisiae* has been sampled. It is clear that segments from many of the mosaic strains are

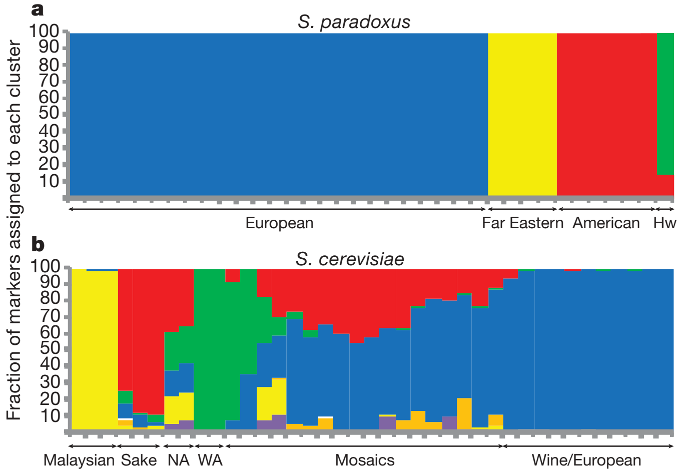

Figure 2 | *Saccharomyces* population structure. **a**, Inference of population structure using the program Structure (version 2.1) on *S. paradoxus* (markers: 7,544 SNPs with >30 strains passing neighbourhood quality standard), assuming $K = 6$ subpopulations and correlated allele frequencies, linkage model based on marker distances in base pairs, 15,000-iteration burn in, and 5,000 iterations of sampling. Each mark on the x axis represents one strain, and the blocks of colour represent the fraction of the genetic material in each strain assigned to each cluster. Hw, Hawaiian isolate. **b**, As in **a**, but for *S. cerevisiae* (markers: 3,413 SNPs with >30 strains passing neighbourhood quality standard). NA, North America; WA, West Africa.

not related to any of the five clean lineages and are probably derived from lineages that are yet to be determined or no longer exist. One-quarter (24%) of SNPs are found only in the mosaics (Supplementary Table 5), which provides a measure of the unsampled *S. cerevisiae* species space.

Sequence variability was quantified using the average pairwise divergence within a population (θ_π) and the proportion of polymorphic sites (θ_S)[10]. We estimated these parameters for various populations (Supplementary Table 7). Both θ_π and θ_S are about 0.001 in the UK population of *S. paradoxus*. The Wine/European cluster of *S. cerevisiae* has approximately the same level of diversity. In both the global and Wine/European samples of *S. cerevisiae*, Tajima's *D* (ref. 20) is significantly negative, indicating an excess of singleton polymorphisms, which may be a consequence of our sampling strategy. By contrast, the UK sample of *S. paradoxus* from a single population has a positive Tajima's *D*, although not significantly, indicating a relative abundance of mid-frequency polymorphisms. Linkage disequilibrium differs between samples (Fig. 3a). For *S. paradoxus*, linkage disequilibrium declines smoothly with distance, decaying to half its maximum value at about 9 kb, as previously reported[10]. For both *S. cerevisiae* samples, the linkage disequilibrium decays much faster, with a half maximum at 3 kb or less. This implies more recombination in *S. cerevisiae*, perhaps due to more opportunities for strains to mate and recombine.

Patterns of variation can reveal evidence of natural selection. As expected for weakly deleterious mutations, the derived allele frequencies (DAF; see Supplementary Information) for non-synonymous polymorphism are lower than synonymous polymorphism (Fig. 3b). For polymorphisms with DAF <20%, there were 0.86 amino-acid-changing polymorphisms for each silent one. In contrast, for those with DAF >20% this ratio was 0.34, indicating that at least 61% (1 − 0.34/0.86) of the 24,418 amino-acid-changing polymorphisms with DAF <20% are deleterious. Similar calculations (Supplementary Information) indicate that 27% of non-coding polymorphisms with DAF <20% are deleterious (Supplementary Fig. 8a). We also performed McDonald–Kreitman tests[21] on 1,105 genes for which we had enough statistical power. No evidence for positive selection after a multiple testing correction was found (Supplementary Fig. 8b). These analyses assumed that synonymous polymorphisms are neutral. However, we found an excess of polymorphism at both low and high frequency (Fig. 3b) in genes with high codon bias (codon adaptation index >0.6). Further analysis (Supplementary Information) indicates that codon bias in *S. cerevisiae* is maintained by both purifying and positive selection, as suggested by the mutation-selection-drift model[22].

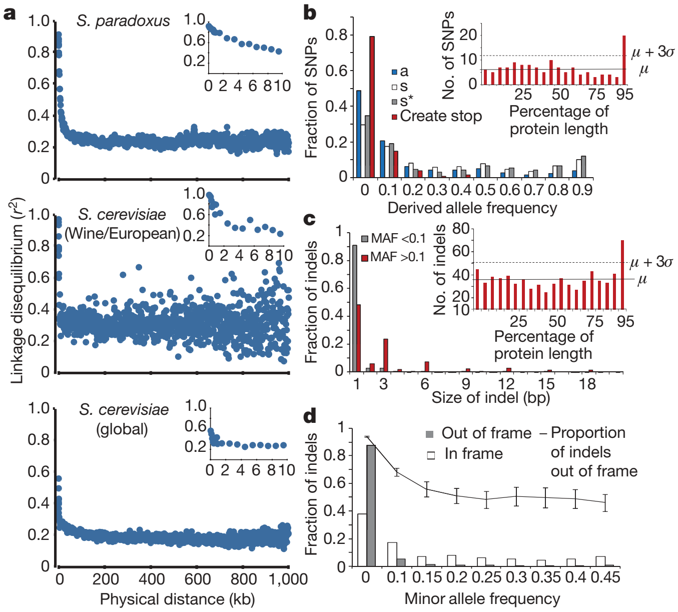

Figure 3 | Population genomics: variation and selection. **a**, Linkage disequilibrium as a function of distance averaged over one kilobase (kb), expressed in terms of correlation coefficient, r^2. Insets show the decline in linkage disequilibrium over the first 10 kb. Details are shown in Supplementary Table 7. **b**, Derived allele frequencies of SNPs in coding regions. Amino-acid-changing SNPs (labelled 'a') show an excess of low frequencies in comparison with synonymous SNPs ('s'). Synonymous SNPs in genes with strong codon bias ('s*') are in excess at low and high frequencies. SNPs that create stop codons ('create stop') show skew to low frequencies. Inset is the number of mutations occurring over the length of the protein, exceeding three standard deviations (σ) from the mean (μ) in the C terminus. **c**, Distribution of sizes of indel polymorphisms in coding regions. High-frequency indels (>10%, red) more often occur in multiples of three than do low-frequency indels (grey). Inset is as for **b**. **d**, Frequency distribution of indels in coding regions. Out-of-frame indels (grey) show excess at low frequencies relative to in-frame indels (open). The proportion of out-of-frame indels decreases as frequency increases. Error bars represent the standard error of the proportion. Numbers of observations for each bin: 0, $n = 2,910$; 0.1, $n = 184$; 0.15, $n = 68$; 0.2, $n = 52$; 0.25, $n = 29$; 0.3, $n = 29$; 0.35, $n = 36$; 0.4, $n = 29$; 0.45, $n = 40$.

A previous genome-wide study in *Arabidopsis*[23] reported a large number of seemingly highly deleterious alleles. We found 134 mutations that were predicted to introduce stop codons (Fig. 3b), including five in genes previously reported to be essential in S288c (Supplementary Information). These mutations showed a skewed frequency distribution and were enriched in the C termini (the final 5% of proteins; Fig. 3b inset).

This data set allowed the consideration of insertions and deletions (Fig. 3c). We identified 3,870 indels in the coding regions of the *S. cerevisiae* population. Of these, 731 had minor allele frequency (MAF) greater than 10%. We also found 657 indels (72 with MAF >10%) in genes identified as essential (Supplementary Information). Indels with MAF >10% predicted to cause frame shifts were enriched in the C-terminal 5% of the protein (Fig. 3C, inset). The proportion of frame-shift to in-frame indels decreases strongly as a function of MAF (Fig. 3d). For example, at MAF >15% there are 1.0 out-of-frame indels for every in-frame indel, in comparison with 15.5 at MAF <10%. We estimate that 93% (1 − 1.0/15.5) of the 2,949 out-of-frame indels with MAF <10% are deleterious.

All strains were subjected to high-throughput phenotypic analysis under multiple conditions (Fig. 4 and Supplementary Fig. 9). Growth curves were sampled (>250 time points) over three days and the relevant growth variables—lag (adaptation), rate (slope) and efficiency (maximum density)—were extracted[24], providing roughly 200 phenotypic traits. The phenotypic variation allowed clustering of strains. There is a high qualitative overlap between the phenotypic clustering and the phylogenies based on SNPs (Figs 1 and 4). Also, the correlation between genotypic and phenotypic similarity within *S. cerevisiae* is surprisingly good (Spearman's rank test: correlation coefficient, 0.30; $P = 10^{-26}$) given that conventional phenotypic taxonomy generally fails even to resolve the *Saccharomyces sensu stricto* species. No individual environment determined the overall correlation between genotype and phenotype.

The *S. paradoxus* strains were well separated from the *S. cerevisiae* strains (Fig. 4), except for the Hawaiian isolate. The phenotypes separating the two species most clearly ($P < 10^{-9}$) were strong *S. paradoxus* resistance to cycloheximide and sensitivity to paramomycin, heat and copper (Supplementary Fig. 10a). The *S. cerevisiae* isolates fell into two groups (Fig. 4). One contains most of the Wine/European and Sake lineages and most of the long-branch recombinants, whereas the other mainly consists of the North American, Malaysian and African lineages. The main phenotypic characteristic separating these groups is rapid growth (short lag and steep slope in rate, $P < 10^{-4}$) for the Wine/European lineages and the mosaics, which could be advantageous for the fermentation processes in which many of these strains are used (Supplementary Fig. 10b). Despite genomic variation, *S. paradoxus* strains (excluding the Hawaiian isolate) show 38% lower phenotypic variation than *S. cerevisiae* strains ($P = 0.002$). In *S. cerevisiae*, the phenotypic variance is as high among the clean lineages as among the

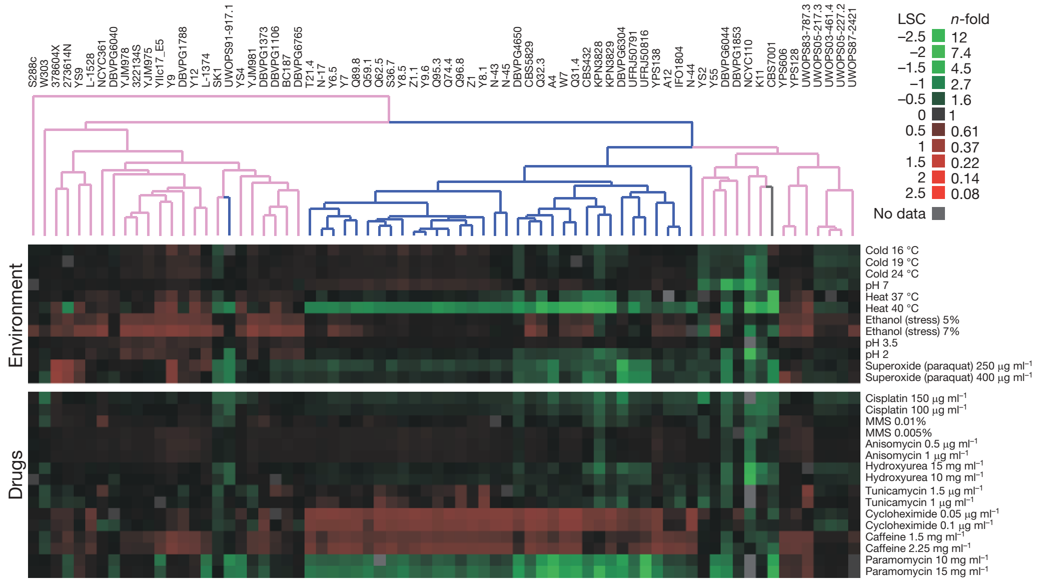

Figure 4 | *Saccharomyces* phenotype variation. A selection of growth phenotypes for *S. cerevisiae* and *S. paradoxus* strains in different environments and drugs. The complete set of lag, rate and density phenotypes in 67 environments is displayed in Supplementary Fig. 9. Phenotypes were quantified using high-resolution micro-cultivation measurements of population density. Strain ($n = 2$) doubling time (rate) phenotypes in relation to the S288c derivative BY4741 ($n = 20$) are displayed. Green, poor growth; red, good growth. Hierarchical clustering of phenotypes was performed using a centred Pearson correlation metric and average linkage mapping. Blue, *S. paradoxus*; pink, *S. cerevisiae*; grey, *S. bayanus* isolate CBS7001. LSC, logarithmic strain coefficient.

mosaic lineages ($P = 0.78$). Hence, the higher phenotypic variance in *S. cerevisiae* is not driven by out-breeding or domestication per se, but rather suggests that *S. cerevisiae* occupies a wider diversity of ecological niches than *S. paradoxus*.

This survey of *S. cerevisiae* and *S. paradoxus* population genomics reveals extensive differences in genomic and phenotypic variation despite ecological similarities and will allow rapid fine mapping of the genetic determinants. Domestication of *S. cerevisiae* has previously been debated[12]. Our results could be interpreted in two ways. One is as a domestication of one or two groups, the Wine/European and Sake strains, with selection for improved fermentation properties. These domesticated groups then gave rise to feral and clinical derivatives and were involved in the generation of out-crossed derivatives found in all sources. The alternative interpretation is that human activity simply may have used existing strains from populations that had appropriate fermentation properties providing the opportunity to out-breed through movement of strains and supplying a novel disturbed environment. Using domestication to imply 'species bred in captivity'[25], the strains that best fulfil this definition are the baking isolates, as they have clearly arisen from crosses between lineages. Lineages that were selected from captively bred strains would be expected to have lower diversity than other lineages. This is not the case for the Wine/European or Sake lineages, which have similar or greater levels of diversity in comparison with the other clean lineages or to *S. paradoxus* populations. This view of human activity simply moving yeast strains around without captive breeding is consistent with analysis of over 600 strains[26]. Recent findings in the Malaysian rainforest (from which our three Malaysian *S. cerevisiae* strains were isolated) of chronic intake of alcoholic nectar from Bertram palms by wild tree shrews suggest that the association of fermented beverages and primates is ancient and not exclusive to humans[27].

Beyond the analyses we have presented here, the sequence data we have obtained for these strains have many other applications, and have already been used both for global[28] and gene-specific[29] studies. With the advent of new sequencing technology, it is becoming possible to undertake similar population genomic studies for species with much larger genomes, including humans[30], enabling a new era of genome wide evolutionary and functional genetics.

METHODS SUMMARY

Strains to be sequenced were selected to maximise the variety of sources and locations of isolation. Except for laboratory strains, a single meiotic diploid spore was isolated from the original strain to remove any heterozygosity[8]. DNA was extracted from overnight cultures[8] for subsequent sequencing on ABI 3730 DNA sequencers and an Illumina Genome Analyzer[15]. Reference-based genome assemblies were created for each strain in a series of steps[15]. Each read was aligned to the reference genome (S288c or CBS432). As this approach cannot deal with large indels or with sequences not present in the reference genome, we developed an iterative parallel-alignment assembling tool, PALAS (Supplementary Methods), to introduce insertions that were allowed to share material between related strains. Two versions of each strain sequence were produced, a partial assembly derived just from data collected from that strain and a more complete assembly using an imputation process to infer the most likely sequence of the strain taking into account data from related strains. In both cases, confidence estimates are given for each base call. The SNPs obtained were used to generate neighbour-joining phylogenetic trees[15], infer population structure[17], estimate sequence divergence[10] and analyse polymorphisms[10]. Non-aligned reads (those missing in the reference genome) were searched for potential novel genes. Each strain isolate was subjected to precise phenotyping in 67 experimental conditions using a high-resolution micro-cultivation Bioscreen C (Oy Growth Curves, Finland)[24]. Two consecutive rounds of 48-h pre-cultivation in synthetic complete media were followed by a 72-h cultivation in stress media. Readings of optical density were taken every 20 min. Strains were tested as duplicates ($N = 2$). Growth variables were normalized to the behaviour of the 20 BY4741 replicates.

Details of the methods mentioned above are provided in Supplementary Information.

Received 18 June; accepted 22 December 2008.
Published online 11 February 2009.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
2. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387** (suppl.), 7–8 (1997).
3. Sampaio, J. P. & Goncalves, P. Natural populations of *Saccharomyces kudriavzevii* in Portugal are associated with oak bark and are sympatric with *S. cerevisiae* and *S. paradoxus*. *Appl. Environ. Microbiol.* **74,** 2144–2152 (2008).
4. Sniegowski, P. D., Dombrowski, P. G. & Fingerman, E. *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Res.* **1,** 299–306 (2002).
5. 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).

6. Koufopanou, V., Hughes, J., Bell, G. & Burt, A. The spatial scale of genetic differentiation in a model organism: the wild yeast *Saccharomyces paradoxus*. *Phil. Trans. R. Soc. Lond. B* **361,** 1941–1946 (2006).
7. Kuehne, H. A., Murphy, H. A., Francis, C. A. & Sniegowski, P. D. Allopatric divergence, secondary contact, and genetic isolation in wild yeast populations. *Curr. Biol.* **17,** 407–411 (2007).
8. Liti, G., Barton, D. B. & Louis, E. J. Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics* **174,** 839–850 (2006).
9. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nature Genet.* **38,** 1077–1081 (2006).
10. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc. Natl Acad. Sci. USA* **105,** 4957–4962 (2008).
11. Pretorius, I. S. Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast* **16,** 675–729 (2000).
12. Fay, J. C. & Benavides, J. A. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae. PLoS Genet.* **1,** 66–71 (2005).
13. 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).
14. Naumov, G. I., James, S. A., Naumova, E. S., Louis, E. J. & Roberts, I. N. Three new species in the *Saccharomyces sensu stricto* complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *Int. J. Syst. Evol. Microbiol.* **50,** 1931–1942 (2000).
15. Carter, D. M. *Saccharomyces* genome resequencing project. *Wellcome Trust Sanger Institute* ⟨http://www.sanger.ac.uk/Teams/Team118/sgrp/⟩ (2005).
16. 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).
17. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
18. Wei, W. *et al.* Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc. Natl Acad. Sci. USA* **104,** 12825–12830 (2007).
19. Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae. Nature* doi:10.1038/nature07670 (this issue).
20. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
21. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila. Nature* **351,** 652–654 (1991).
22. Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129,** 897–907 (1991).
23. Clark, R. M. *et al.* Common sequence polymorphisms shaping genetic diversity in *Arabidopsis* thaliana. *Science* **317,** 338–342 (2007).
24. Warringer, J. & Blomberg, A. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in *Saccharomyces cerevisiae. Yeast* **20,** 53–67 (2003).
25. Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature* **418,** 700–707 (2002).
26. Legras, J. L., Merdinoglu, D., Cornuet, J. M. & Karst, F. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol. Ecol.* **16,** 2091–2102 (2007).
27. Wiens, F. *et al.* Chronic intake of fermented floral nectar by wild treeshrews. *Proc. Natl Acad. Sci. USA* **105,** 10426–10431 (2008).
28. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454,** 479–485 (2008).
29. Demogines, A., Wong, A., Aquadro, C. & Alani, E. Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. *PLoS Genet.* **4,** e1000103 (2008).
30. Siva, N. 1000 Genomes project. *Nature Biotechnol.* **26,** 256 (2008).

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

Acknowledgements We thank all the members of the Sanger sequencing production teams for generating the sequence data. We thank members of the Durbin and Louis laboratories and C. Nieduszynski for comments and suggestions, L. Kruglyak and J. Schacherer for sharing their unpublished manuscript and R. Ames and S. Lovell for sharing unpublished results. We also thank the British Council and Chinese Academy of Sciences for providing the opportunity to design and develop this project. Research at the Wellcome Trust Sanger Institute (D.M.C., A.M.M., L.P, M.J., M.A.Q., I.G., S.S., F.S. and R.D.) is supported by The Wellcome Trust. G.L., D.B.H.B., E.B. and E.J.L. were supported by the Wellcome Trust, the Royal Society and the Biotechnology and Biological Sciences Research Council (BBSRC). S.A.J., R.P.D. and I.N.R. were supported by the BBSRC. A.B. and J.W. were supported by the Swedish Research Council and the Swedish Foundation for Strategic Research. A.B. and V.K. were supported by the National Environment Research Council (NERC) and I.J.T. was supported by the Wellcome Trust. D.B. was supported by NERC. A.M.M. was supported by the Canada Foundation for Innovation. M.J.T.O. and A.v.O. were supported by the National Science Foundation, the National Institutes of Health and a Hertz fellowship.

Author Contributions R.D. and E.J.L. designed the project. G.L. selected and manipulated yeast strains and extracted DNA samples. M.J., M.A.Q., I.G., S.S. and F.S. performed the subcloning and sequencing. D.M.C. did the reference comparison and assembly of the sequences. D.M.C. and G.L. coordinated the collection of data. D.M.C. and R.D. performed much of the global analysis, which was the basis for specific analyses performed by the other authors. A.M.M. did the selection studies. E.J.L., G.L., D.M.C. R.D., D.B.H.B., E.B. and L.P. did the population structure and analysis of new genes. C.M.B. and D.B. performed the analysis of Ty-element abundance. S.A.J., R.P.D., M.J.T.O., A.v.O. and I.N.R. analysed the rDNA. A.B., V.K. and I.J.T. did the sequence variation and recombination analyses. A.M.M. and A.N.N.B. created a BLAST server. J.W. and A.B. generated the phenomics data. E.J.L. and G.L. wrote the paper, coordinating the contributions of the other authors.

Author Information ABI data were submitted to the NCBI Trace Archive (accession numbers: TI1250539685–TI1250559725, TI1253231395–TI1253251114, TI1253289161–TI1253346699, TI1253412335–TI1253476299, TI1253519268–TI1253585710, TI1253601045–TI1253652620, TI1253786059–TI1253860555, TI1253882296–TI1253926729, TI1253930843–TI1253993129, TI1253998006–TI1254046664, TI1254080659–TI1254150525, TI1254155855–TI1254203444, TI1254229054–TI1254316243, TI1254369656–TI1254427407, TI1254441246–TI1254496019, TI1254545851–TI1254582436. The Solexa data were submitted to the European Read Archive (accession number ERA000011). Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.D. (rd@sanger.ac.uk) or E.J.L. (ed.louis@nottingham.ac.uk).