Population genomics of domestic and wild yeasts

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

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1HH, UK, [2]Institute of Genetics, Queen's Medical Centre, University of Nottingham, Nottingham NG7 2UH, UK, [3]National Collection of Yeast Cultures, Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK, [4]Department of Cell and Molecular Biology, Lundberg Laboratory, University of Gothenburg, Medicinaregatan 9c, 41390 Gothenburg, Sweden, [5]Division of Biology, Imperial College London, Silwood Park, Ascot, Berks., SL5 7PY, UK, [6]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK, [7]Department of Physics, Massachusetts Institute of Technology, Cambridge MA 02139, USA

**These authors contributed equally to this work*

+Present address: Department of Cell & Systems Biology, University of Toronto, Canada, M5S 2J4

‡ Present address: School of Biological Sciences, University of Liverpool, Liverpool, LG9 3BX

The natural genetics of an organism is determined by the distribution of sequences of its genome. Here we present one- to four-fold, with some deeper, coverage of the genome sequences of over seventy isolates of the domesticated baker's yeast,

***Saccharomyces cerevisiae*, and its closest relative, the wild *S. paradoxus*, which has never been associated with human activity. These were collected from numerous geographic locations and sources (including wild, clinical, baking, wine, laboratory and food spoilage). These sequences provide an unprecedented view of the population structure, natural (and artificial) selection and genome evolution in these species. Variation in gene content, SNPs, indels, copy numbers and transposable elements provide insights into the evolution of different lineages. Phenotypic variation broadly correlates with global genome-wide phylogenetic relationships however there is no correlation with source. *S. paradoxus* populations are well delineated along geographic boundaries while the variation among worldwide *S. cerevisiae* isolates show 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* shows a few well defined geographically isolated lineages and many different mosaics of these lineages, supporting the notion that human influence provided the opportunity for outbreeding and production of new combinations of pre-existing variation.**

Since the completion of the genome sequence of *Saccharomyces cerevisiae* in 1996, the first eukaryotic organism sequenced[1, 2], there has been an exponential increase in complete genome sequences accompanied by great advances in our understanding of genome evolution. This has been particularly evident from comparative genomics both of close relatives, e.g. for *Saccharomyces* yeasts[3, 4] and for *Drosophila* species[5, 6], and of more distant relatives, e.g. for hemiascomycetes yeast species[7, 8]. Less well understood at the whole genome level are the evolutionary processes acting within populations and species leading to adaptation to different environments, phenotypic differences and reproductive isolation. There has been increased recognition of the importance of understanding the genome sequence variation within a species in order to fully

understand the nature of a species and individual phenotypic variation, for example single nucleotide or copy number variation in relation to human disease[9]. A single genome sequence for a species is inadequate for answering many biological questions which require large-scale population sampling. The tools for large-scale comparisons of populations of whole genome sequences do not yet exist and require adequate and realistic datasets for their development and evaluation. The *Saccharomyces sensu stricto* yeasts, including the best understood model eukaryote *S. cerevisiae*, are ideal for population genomics studies, as there are many isolates available from several species from around the world, and many aspects of their biology are well characterised and amenable to study. 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[10, 11], population structure[12-15], and sexual versus asexual reproduction and gene flow[16, 17]. When combined with a population view of whole genome sequence variation, these approaches will increase our understanding of the ecology and evolutionary biology of these important organisms.

The baker's yeast, *S. cerevisiae*, has had a long association with human activity[18] leading to the idea that use in fermentation imposed selection on specific lineages leading to domestication. A survey of five loci in many isolates supports the idea that there were at least two independent domestication events arising from human activity, one for sake strains and one for wine which were derived from wild populations[19]. In contrast, the closest relative of *S. cerevisiae*, *S. paradoxus*, has never been associated with human activity and is found globally and in many cases in the same locations as *S. cerevisiae*[10, 11]. A preliminary comparison of a few genes from several isolates of each species within the *Saccharomyces sensu stricto* group exhibited extensive variation between three well defined geographic populations of *S. paradoxus* but limited variation among *S. cerevisiae* isolates and no correlation with geographic location[15]. *S. cerevisiae* is usually isolated from fermentation processes (beer, wine, sake, bread) but can also be

found in the wild (either associated with fermenting fruit or nectar or isolated from non-fermenting substrates such as soil or bark), on humans (either as non-pathogenic commensals or as disease causing clinical isolates), and in food spoilage situations. Isolates from different sources have phenotypes associated with adaptation to that source, so that baking strains, for example, are fundamentally different from wine strains in that they exhibit vigorous maltose fermentation[20], while clinical isolates exhibit the ability to grow at high temperatures, which is associated with virulence[21].

Here we present the results of the first large-scale multi-species population genomics analysis in eukaryotes using the genome sequences of 36 strains *S. cerevisiae* and 35 strains *S. paradoxus* isolated from several sources and locations. These sequences not only provide a dataset to aid development of new population genomics tools but will also help us understand the nature of *Saccharomyces* yeasts, the genetic variation underlying important phenotypic differences, and the effect of human activity on the evolution of one of the most important model and industrial organisms.

Strain selection

Strains of *S. cerevisiae* and *S. paradoxus* were chosen to maximize the potential variation across geography as well as environments (Tables 1 and S1). Strain selection was subject to certain constraints to avoid technical difficulties with sequencing as well as maximize future utility. Except for the laboratory strains S288c, W303 and the baking strains, the isolates selected were diploid. As nearly 10% of *Saccharomyces sensu stricto* isolates appear to be hybrids between different species[22], we selected strains that behaved as diploid non-hybrids in meiosis and test crosses. We obtained a single spore isolate from the strains in order to avoid the problem of heterozygosity within the diploid genome, which would cause problems with SNP calling and

assemblies. Each single spore isolate sequenced, as well as the other three spores from the same meiosis, have been stored and are available at NCYC[23].

The *S. cerevisiae* isolates include strains from a large variety of sources: laboratory, pathogenic, baking, wine, food spoilage, natural fermentation, sake, probiotic, plant, soil as well as S288c, the progenitor of the reference strain sequence. These strains were isolated from numerous locations around the world and were obtained primarily from various culture collections or other researchers (Table S1). The *S. paradoxus* isolates were mostly isolated from oak tree exudates and bark from various locations around the world and include the three major populations now recognised as European, Far Eastern and American[13, 15, 22]. The latter include two South American isolates originally designated as a separate species, *S. cariocanus*[24]. The European group also includes 18 strains isolated in the UK from the same geographic area. Two new locations, Siberia, which is geographically located halfway between European and Far East Asia, and the island of Hawaii, were also included. There is overlap in the general geographic sources of isolates from both species.

Sequencing and assembly

The majority of sequences were obtained using standard Sanger sequencing on ABI3730 sequencing machines. For some strains short sequencing-by-synthesis (SBS) reads were obtained using the Illumina Genetic Analyser. Most strains were covered to a depth of 1-3X with a few covered extensively with both ABI and Ilumina reads. The basic statistics are shown in Table 1 while details of number of reads, depth of coverage and description of libraries are given in supplementary information (Table S2). High levels of ABI coverage were obtained for the S288c version we sequenced as well as laboratory strains W303, SK1, Y55 and of our single spore isolate of the *S. paradoxus* type strain, CBS432. The sequence reads, assemblies, alignments, other data and a

genome browser are all publicly available[25]. Reads are also deposited at the international trace archive[26].

Reference-based genome assemblies were created for each strain in a series of steps. First, each read was aligned to the reference genome for the relevant species (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 assembling tool, PALAS (see Methods in SI), 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. Even using PALAS, some sequences such as the subtelomeric regions, which structurally are highly polymorphic[27], could not be reliably assembled. Because of their extreme AT-richness, the mitochondrial genome sequences are also incomplete. In all we identified 235127 high-quality SNPs and 14051 indels in the *S. cerevisiae* nuclear genome, and 623287 SNPs and 25267 indels in *S. paradoxus.* Further details are given in supplementary information (Table S2).

Nearly complete assemblies have been obtained for the four laboratory strains of *S. cerevisiae* including our version of S288c. Our S288c varies from the reference genome (downloaded from SGD 10-10-2007) by 659 high-quality SNPs, of which we estimate at least 33 (5%) are due to accumulated divergence since the two versions separated. Many of the differences found are in subtelomeric regions and there appears to be a higher rate of discrepancy in chromosomes 1 and 2 than in the rest of the genome (Fig. S1). Of the high-quality aligned positions, for 480 SNPs our S288c sequence is supported by other strains sequenced whereas the reference version has no support, while for 18 SNPs the reference genome sequence is supported by other strains

while our S288c has no support. Many of the former set are likely to represent errors in the reference genome sequence, which still implies a low error rate of 1 in 24kb (better than Q40). A list of likely errors is given in Table S3 and has been submitted to SGD.

The available sequence for the type strain of *S. paradoxus*[4] is not complete and so we sequenced a single spore isolate from our version of CBS432 to 4.3X coverage with ABI. Although the two sequences are closely related and are part of the same population, there are numerous SNP differences between them. This could be due to the fact that we used a single spore isolate which would not have any heterozygosity, while the assembly created by Kellis *et al*[4] has the other allele at some SNP locations. Alternatively, the difference may come from the strains actually not being the same as they were obtained from separate culture collections (Northern Regional Research Laboratory, NRRL, and Centraalbureau voor Schimmelcultures, CBS). Our version of CBS432 is identical to that in the CBS collection over specific regions of SNP differences confirmed by sequencing of an independent sample from CBS. In order to circumvent this issue and to produce a *S. paradoxus* reference sequence with the greatest contiguity we collected together all the reads from the UK isolates of *S. paradoxus*, the reads from the Broad project[4], and artificial reads created by shredding the Broad contigs. We assembled these reads using Phusion[28] into 608 contigs, which we then aligned to the *S. cerevisiae* reference to place them on chromosomes. We used the resulting sequence as the starting point for the reference-based assemblies for the individual *S. paradoxus* strains. We also aligned this sequence and the contigs for *S. bayanus, S. kudriavzevii* and *S. mikatae* (downloaded from SGD[29]) to create cross-species multiple alignments used in subsequent analyses. Further details of both procedures are in the supplementary information.

SNPs and the Population structures of the two species

We generated neighbour-joining (NJ) phylogenetic trees based on pairwise SNP differences in the strain-sequence alignments (Fig. 1 and Fig. S2) to obtain a global picture of relationships among the strains and species. As has been previously reported[13, 15] for a smaller gene sample, the *S. paradoxus* strains fall into three very well separated populations with one lineage outside these (the Hawaiian isolate discussed in more detail below). The *S. cerevisiae* isolates exhibit somewhat more variation species-wide than a single *S. paradoxus* population but much less than the overall species-wide variation seen in *S. paradoxus*. Most of the SNPs in *S. paradoxus* are private polymorphisms within each population while most of the SNPs in the *S. cerevisiae* samples are shared. There is some structure in the *S. cerevisiae* sequences, as will be discussed below, but there is evidence for extensive recombination between lineages.

S. paradoxus exhibits a clear picture of population structure using the software STRUCTURE[30] which indicates three populations plus the single Hawaiian isolate (Fig. 2A). Previous analysis with smaller segments sequenced resulted in the same picture of three well defined populations[13, 15]. The European population was sampled extensively which provided a picture of within population dynamics (Fig. 1B). The topology of the NJ tree of the European population correlates with the West to East geographic locations of isolation of the strains.

The *S. cerevisiae* population structure is more complicated. 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 strain sets from Malaysia, West Africa, sake and related fermentations, North American, and a large cluster of mixed sources that contains many European and wine making strains (Wine/European). The remaining strains are on long branches between the large Wine/European cluster and the other four clean lineages. There are some lineages that

correspond to geographic origin, such as the wild isolates from North America and Malaysia. However many of the closely related strains seem to be from widely separated locations. The lab strains SK1 and Y55, isolated in North America and Europe respectively, are on a branch with two West African strains; and one African strain (Y12) is on a branch with sake strains while another (DBVPG1853) is on a long branch unrelated to other strains. This mixed architecture could be due to human traffic in yeast strains and subsequent recombination between the lineages. Analysis with STRUCTURE is consistent with separate populations for the West African, Malaysian, sake lineages and the Wine/European cluster at the far right of the NJ tree (Fig. 2B). The wild North American isolates share some polymorphisms with all four separate populations, while all of the rest share polymorphisms with the European lineage and at least one other.

Trees constructed on a chromosome-by-chromosome basis or by smaller segments make the mosaic nature of these genomes clear, as do segmental similarity comparisons (Fig. 2C and Fig. S3). The laboratory strains SK1 and Y55 appear to be the result of recent crosses between the West African lineage and the European lineage, as can be clearly seen in pairwise comparisons along the chromosomes (Fig. S3B). Similarly W303 is a recent cross with the reference S288c lineage and one or more other lineages. Other strains appear to be more ancient combinations of the different lineages with no long segments (linkage disequilibrium blocks) retained from one or another lineage. Some chromosomes or segments fall into different locations in the NJ tree (Fig. 2C) while many stay on a long branch in roughly the same relationship to other strains. The recently sequenced clinical derivative YJM789[31], for example, is one of these long-branch mosaic strains. The wild North American isolates represent a clean lineage as most segments exhibit the same phylogenetic relationship with the other lineages across the genome. The shared polymorphisms with the other four clean lineages indicate a more ancient interaction. This complex population structure of *S. cerevisiae* is

consistent with previous analyses[19] and is similar to a recent study by Schacherer *et al* (this issue). However, rather than concluding that there were two independent domestication events in the history of this species[19], the picture presented here is more consistent with there being five well-delineated lineages, two of which contain multiple isolates used in fermentation industries. In many cases, the nearest relative to a fermentation strain over some portion of the genome is a wild or clinical strain. There is the added complication that many extant strains are recombinants between lineages, most but not all of which are also used for fermentation. Nevertheless, phenotypic profiling and analysis of rDNA repeat unit variation both produce results consistent with this overall picture of the *S. cerevisiae* population structure (see below).

Whether we have sampled the entire space of existing *S. cerevisiae* lineages is open to debate. It is clear that segments from many of the strains, such as W303 and the mosaic strains (Fig. S3), are not related to any of the lineages delimited here and are probably derived from yet to be determined or no longer existing lineages. Many of these strains are segmentally related to each other. Future work may be able to provide a partial reconstruction of the other lineages from the sequences that now exist.

We have previously identified in the left arm of chromosome XIV a 23 Kbp region of introgression from *S. cerevisiae* into the European population of *S. paradoxus*[15]. An analysis of the sequence reads in this region of the genome is consistent with a single event that spread through the population, as the end points of introgression are the same in all strains sequenced over this region (Fig. S4). We searched for a possible *S. cerevisiae* donor and found several smaller segments within the introgression that appeared similar to different clean lineages indicating the donor was one of the mosaic strains.

We also find a significant number of *S. cerevisiae*-like reads in the Hawaiian isolate of *S. paradoxus*. In many cases there are *S. paradoxus* reads covering the

homologous/allelic locations. At present this is not easily explained as the strain behaves as a diploid member of the *S. paradoxus* species by inherent fertility and meiotic viability in test crosses. Both sets of sequences exist in the strain as confirmed by PCR and independent isolation and DNA preparation. Further sequencing and analysis of paired reads will help to resolve the potential hybrid origin of the Hawaiian isolate of *S. paradoxus*.

Population genetics: recombination and selection

Sequence variability in populations is typically estimated in two ways, as the average pairwise divergence of sequences in a population (θ_π) and from the proportion of polymorphic or segregating sites (θ_S)[17]. Here we estimate these parameters for (i) the UK population of *S. paradoxus* (the closest we have to several isolates from a single natural population); (ii) a global sample of *S. cerevisiae*, excluding all isolates that appear to be clonemates; and (iii) the Wine/European cluster of *S. cerevisiae* (the best sampled of any of the clean lineages, again excluding likely clonemates; Table S4). Both θ_π and θ_S are about 0.001 in the UK population of *S. paradoxus*, indicating an average of 1 difference per 1000bp between two strains. This is very similar to the value previously found for chromosome III in the same population[17]. The Wine/European cluster of *S. cerevisiae* has approximately the same level of diversity as the UK *S. paradoxus* population, while the global sample of *S. cerevisiae* has >5 times more diversity, but is less diverse than the global sample of *S. paradoxus*. The average divergence of two *S. cerevisiae* strains is about 40% of that between European and Far East *S. paradoxus*, and about 15% of that between the American *S. paradoxus* and either of the other two populations[32]. In both the global and Wine/European samples of *S. cerevisiae* Tajima's D[33] (a standardised measure of the difference between θ_π and θ_S) is significantly negative, indicating an excess of singleton polymorphisms. This may be

a consequence of our sampling strategy for *S. cerevisiae*, which included single isolates from many sources. By contrast, the UK sample of *S. paradoxus* is from a single population, collected within a 10km^2 area, and Tajima's D is positive, indicating a relative abundance of mid-frequency polymorphisms, though the deviation from random expectation is small.

We have also analysed nucleotide diversity separately for each chromosome. In the global sample of *S. cerevisiae* there is a significant negative correlation between the length of the chromosome and the amount of variation, in particular for shorter chromosomes (Kendall's $\tau = -0.52$, $p=0.008$). This appears to be because there is more variation in the subtelomeric regions extending 30kb from each end of each chromosome, and these regions make up a larger proportion of shorter chromosomes. If these regions are excluded, the correlation is no longer significant.

The covariation of alleles at different sites, or linkage disequilibrium, also differs between samples (Fig. 3A). For *S. paradoxus* linkage disequilibrium declines smoothly as the distance between sites increases, decaying to half its maximum value at about 9kb, similar to that previously reported[17]. For both *S. cerevisiae* samples the linkage disequilibrium decays much faster, with a half life of 3kb or less. This implies more recombination in *S. cerevisiae*, perhaps due to selection for recombinants, or more opportunity for strains to mate and recombine.

Our genome-scale population variation dataset, in conjunction with the high-quality genome annotation available for *S. cerevisiae*[29], allows us to compare systematically the patterns of variation for mutations with respect to the functional genomic regions in which they occur. Although the low coverage, variable completeness and complex population structure in our data makes comparison of our observations to theoretical expectations difficult to interpret, it is possible for us to compare our data between

functional classes to obtain a picture of the relative effects of selection on new mutations.

As expected under a model of weakly deleterious effects of mutations, in coding regions we find a shift in the derived allele frequencies (see methods) for amino acid replacement polymorphism (Fig. 3B, a) towards lower frequencies compared to synonymous polymorphism (Fig. 3B, s). To quantify this, we computed the a/s ratio at different allele frequencies. For polymorphism with minor allele frequency (MAF) less than 20%, there were 0.86 amino acid changing polymorphisms for each silent polymorphism. In contrast, for those with MAF greater than 20% this ratio was 0.34, indicating that at least 61% (1-0.31/0.86) of the 24418 amino acid changing polymorphisms with minor allele frequency less than 20% are deleterious, and are destined to be removed from the population by natural selection.

The preceding calculations make the assumption that synonymous polymorphisms are neutral. In yeast, however, this is known not to be the case; genes with high expression levels exhibit high codon bias[34] and this can be measured by the codon-adaptation index (CAI)[35]. We computed the derived allele frequency spectrum for the silent polymorphisms in genes with high levels of codon bias (average CAI>0.6) and found an excess of polymorphism at both low and high frequency (Fig. 3B, s*). This suggests the action of both positive and negative selection on polymorphism at these sites. To test this we identified polymorphisms in which the derived allele was either a preferred or un-preferred codon, as defined by the CAI, and compared their allele frequencies in genes with high codon bias to those in the rest of the genome. Consistent with both positive and negative selection acting on synonymous sites in highly expressed genes, we observe both an excess of low-frequency (DAF<20%) SNPs that created un-preferred codons (38% vs. 51%, $p<10^{-6}$) and an excess of high-frequency (DAF>20%) SNPs that create preferred codons (67% vs. 54%, $p<0.01$ Fisher's Exact Test). These

results indicate that codon bias in *S. cerevisiae* is maintained by both purifying and positive selection, as suggested by the mutation-selection-drift model[36].

An unexpected finding in previous genome-wide studies of population variation[37] was the large number of seemingly highly deleterious alleles. Interestingly, we also found large numbers of mutations that were predicted to introduce stop codons in our data (Fig. 3B, 'create stop'), including 5 stop codons in essential genes. These mutations showed an extremely skewed allele frequency distribution, consistent with them being deleterious and likely to be removed from the population. One possible explanation is that the truncations they introduce are compatible with protein function. Indeed, we found that these stop mutations were significantly enriched in the C-termini (the final 5% of protein length, Fig. 3B inset) of proteins, suggesting that in some cases the stop codons cause relatively moderate consequences to protein function.

Our genome-wide variation data was derived from direct sequencing (rather than SNP genotyping as in previous studies[37]) and this afforded the opportunity to consider insertion and deletion mutations (indels) as well as single nucleotide polymorphisms (Fig. 3C). We identified 3377 indels segregating in the coding regions of the *S. cerevisiae* population (467 with minor allele frequency greater than 10%), including 622 (65 with minor allele frequency greater than 10%) in genes identified as essential in S288c (methods in SI). Of the indels with minor allele frequency greater than 10%, 241 of 467 are predicted to cause frame shift mutations. Once again, we found an enrichment for these mutations in the C-terminal 5% of the protein (Fig. 3C, inset), consistent with the indels attaining high frequency in the population being those with relatively mild effects. Nevertheless, we found that the proportion of frame-shift mutations relative to in-frame indels decreases strongly as a function of minor allele frequency, consistent with the action of purifying selection to remove these highly deleterious mutations (Fig. 3D). For example, at minor allele frequency >15% there are

0.69 out-of-frame indels for every in-frame polymorphism, compared to 17.2 at minor allele frequency <10%. Based on this, we estimate that 96% (1-0.69/17.2) of the 2750 out-of-frame indels segregating at minor allele frequency below 10% in the *S. cerevisiae* population are deleterious and will be removed by natural selection.

Non-coding regions contain many functional sequences including regulatory sequences and non-coding RNA genes. We computed the derived allele frequency spectrum for the SNPs that we identified in non-coding regions (Fig. 3E). We found that there is strong evidence for purifying selection in all classes of non-coding regions in yeast (compared to synonymous sites), and that tRNA genes show a skew in their allele frequency distribution comparable to amino acid altering mutations in protein coding regions.

One of the most exciting applications of population genomics is to systematically identify mutations that may have been the targets of positive natural selection and hence may underlie the adaptive differences between species. Comparisons of divergence to diversity in different classes of sites, such as the McDonald-Kreitman test (M-K[38]) provide a powerful means to do so, and are relatively robust to demographic complexity[39]. As noted above, there was a large excess of amino-acid replacement polymorphism at low allele frequencies (Fig. 3F inset). We therefore excluded SNPs with minor allele frequency less than 20%[40] and performed directional Fisher's exact tests on the 1105 genes for which there were at least 5 SNPs, to test the hypotheses that there was an excess of fixed amino acid differences between *S. cerevisiae* and *S. paradoxus*. Overall, we found a skew in the distribution of the M-K ratio towards values less than one, indicating either pervasive purifying selection on the differences between species, or reduction in the efficacy of selection within the current *S. cerevisiae* population[38]. In the positive tail (M-K ratio>1) of this distribution lie genes enriched for amino-acid changes between species, candidates for adaptive differences. However, none of these were significant after a multiple testing correction.

Repeated sequence families: Ty elements, rDNA and CNVs

We estimated the overall level of Ty element abundance for each strain directly from ABI sequencing reads, since a set of dispersed features like Ty elements is sampled proportionally with light shotgun sequencing coverage and because large insertions like Ty elements present a challenge for reference-based genome assembly. The proportion of Ty sequences is typically less than 3% in all strains in both species, with the highest abundance observed in the laboratory strain S288c (3.53%) (Fig. 4A). Globally, we find that the proportion of Ty sequence per strain is higher among strains in *S. cerevisiae* relative to *S. paradoxus* (Table S5; Wilcoxon Test, p = 0.02275). Interestingly, *S. paradoxus* strains from South America (UFRJ50791 and UFRJ50816), which are partially reproductively isolated from other *S. paradoxus* lineages and have been previously known as a separate species (*S. cariocanus*)[15, 24], have the highest Ty abundance for this species. This correlates well with the increased rate of rearrangement due to reciprocal translocations in the South America *S. paradoxus* lineage, and supports the hypothesis that a burst of rearrangements occurred in this lineage due to increased Ty activity [41]. We were able to map all 14 breakpoints of the large translocations in the South American and three additional *S. paradoxus* strains most of which involve Ty/LTRs (see Fig. S5). Levels of variation in Ty abundance are similar in the Wine/European *S. cerevisae* population and in the UK *S. paradoxus* population, and levels of variation in Ty abundance for the global *S. cerevisae* sample are substantially higher than the Wine/European *S. cerevisae* population (Table S5). Finally, we find that the mosaic strains of *S. cerevisae* have higher Ty abundance than the clean lineages (Table S5; Wilcoxon Test, p=0.006896), suggesting that hybridization may have led to an increase in Ty element activity in this species.

Sequence coverage was sufficient to ensure that each position in the 9.1kb rDNA repeat was covered many times in each strain (average 140). A complete analysis of this dataset is given elsewhere (O'Kelly *et al,* this issue) but we report here an interesting correlation between intragenomic rDNA sequence variation and genome mosaicism (Fig. 4B). By comparing the number of rDNA reads against total overall reads (Methods in SI) we estimated rDNA copy number for each *S. cerevisiae* strain to range between 54 (K11) and 511 (YJM981). The estimates were generally in good agreement with previous estimates [42, 43]. The exceptionally high count for strain YJM981 appears to be linked to an unusual karyotype resulting in a much larger than average overall genome size (Fig. S6). A link between rDNA copy number and genome size has been reported previously [44]. We also investigated the possibility of a link between rDNA copy number and genome mosaicism but found them to be unrelated. However, when polymorphic sites in the rDNA repeats from individual strains were enumerated (Methods in SI), mosaic genomes were found to have significantly more variable positions than the clean lineages ($p=1.3 \times 10^{-6}$ under Mann-Whitney U rank test). We further characterised this variation in terms of number of substitutions per position and found the vast majority of variant sites to be in the 0-10% minor allele frequency range (Fig. 4b). In contrast to Ganley and Kobayashi[45] who found only four polymorphic sites in strain RM11-1A, we identified 518 polymorphic sites of which 156 are resolved to a complete SNP (relative to the S288c rDNA consensus sequence) in at least one strain. A possible reason for the increased polymorphism within mosaic strains is that differences in the rDNA sequence between the parents of the mosaic may not have yet been resolved by gene conversion[46], suggesting the mosaics are relatively recent.

This sequence survey provides the opportunity to address issues of CNV including segmental duplications, as well as identification of sequences not found in reference genomes. Table S2 shows the proportion of reads for each strain that were unplaced in the assembly and hence potentially novel. Much of the unplaced material is likely to

originate from subtelomeric regions, which are repetitious and structurally variable and hence hard to assemble. A genome-wide analysis of copy number based on the numbers of reads of each strain aligning to each gene showed very little significant CNV outside the rDNA region with the exception of three strains[25]. Because of the number of gene-strain combinations, only copy numbers of 10 or more could be reliably detected, and only six instances of this were found. The clearest such example, the *CUP1* gene, is discussed below.

Correlation of phenotypes to genotypes

Finding the genetic basis for natural as well as artificially generated variation in phenotypes is a fundamental goal in biology, with implications for our understanding of evolution, disease and biotechnological traits. We found surprisingly little correlation between the phylogeny and the source environment (Fig. 1) among *S. cerevisiae* isolates. The baking strains loosely cluster together globally but fall in different regions of the tree when individual chromosomes and segments are analysed (not shown). Of the six clinical isolates, three are well inside the European cluster while the other three are well separated on individual long branches. Similarly, many of the wine and food spoilage strains are on individual long branches. In order to provide a phenotypic map of the sequenced strains, the entire set was subjected to high throughput phenotypic analysis under a multitude of conditions (Fig. 5A). Growth curves were exhaustively sampled (>250 time points) over several days and the physiologically relevant growth variables growth lag, growth rate (slope) and growth efficiency (maximum density) were extracted[47] to provide roughly 200 distinct phenotypic traits.

Overall there is enough variation in phenotype to allow clustering of strains into phenotypic clusters. Remarkably, there is a high qualitative overlap between the

phenotypic clustering and the phylogenies based on SNPs (Fig. 5A and Fig. 1). Also on a quantitative level the correlation between genotypic and phenotypic similarity within *S. cerevisiae* is surprisingly good (linear correlation r = 0.37, sequence similarity vs. Pearson correlation coefficients from phenotyping), given that conventional phenotypic taxonomy generally fails even to resolve the *Saccharomyces sensu stricto* species. Here, the *S. paradoxus* strains were well separated from the *S. cerevisiae* strains. The only exception was the Hawaiian isolate, which was discussed above. In addition, the *S. cerevisiae* isolates fell into two groups. The first contains most of the Wine/European lineage and most of the long-branch recombinants, while the second consists of the other lineages. The main phenotypic characteristic separating these two groups appears to be rapid growth (short lag and steep slope in rate) for the Wine/European and mosaics, which could be advantageous for the fermentation processes many of these strains are used for.

Some specific phenotypes were consistent with known genotypes or sequences determined here. For example, S288c showed only limited growth on galactose due to a known *gal2* mutation, while the reference strain BY4741 is Gal^+ as this mutation was corrected (all phenotypic data is displayed in relation to BY4741). Deleterious mutations in *GAL* genes, notably a frameshift in the *GAL2* gene in Y55 and DBVPG6044, and a premature stop-codon in the *GAL3* gene in Y55, DBVPG6044, and NCYC110 (West African lineage), are likely to cause the severe galactose growth defects observed in these strains (Fig. 5B, C). Complete absence of galactose growth was only observed for *S. cerevisiae* strains 273614N and YS2 which have no *GAL* genes despite good sequence coverage. Parallel inactivation of *GAL* genes has earlier been reported as a means of ecological diversification[48].

Another example of a stringent genotype - phenotype link is the ability to utilise melibiose. Unperturbed growth on melibiose was only observed in five *S. cerevisiae*

strains: the three Malaysian, the Hawaian and the West African strain NCYC110, which contain either one or two copies of the melibiose gene *MEL1*. Strains with two *MEL1* copies tend to grow better on melibiose (Fig. 5D). The only exception to the strict link between melibiose utilisation and *MEL1* is DBVPG6044, which contains two *MEL1* genes with no non-synonymous mutations but exhibited limited growth on melibiose. However, expression of *MEL1* is governed by the GAL regulatory network[49], suggesting that the above mentioned galactose negative phenotype and prevalent mutations in Gal regulatory proteins may explain its melibiose phenotype. *S. cerevisiae* isolates displayed clear phenotypes in relation to copper. Phenotypes can also be correlated with copy number (Fig. 5E). The gene dosage of *CUP1,* which in *S. cerevisiae* varies between 0 and up to 40 for DBVPG1373, is directly proportional to copper resistance. There is a clear distinction between the two groups of *S. cerevisiae* strains. The Wine/European and long branched mosaics strains tended to show higher copper resistance. This may be partially explained by the use of copper as a fungicide or copper containers in breweries, imposing a selective pressure for copper resistance and thus for *CUP1* copy number increase. The wild *S. paradoxus* strains contain only one *CUP1*copy and, with the exception of the Far Eastern strains, show high sensitivity to copper (Fig. S7). This could explain the lack of *S. paradoxus* in fermentation or the lack of use has not imposed selective pressure for increase copy number.

Discussion

We have presented here the first full-sequence population genomic survey of two closely related species, one associated with human activity and one a natural species with no human association. Both are globally distributed and can be found in similar niches, even from the same bark sample of an oak tree[10, 11]. Despite this ecological similarity, we find extensive differences in the population genomic history of these two species, mirrored by extensive phenotypic variation. This study offers an unprecedented

view of sequence diversity across the genome within closely related species, lays the background for functional phenotypic analysis, and gives us further insight into the processes underlying sequence and genome evolution in this important model organism.

With these powerful genomic resources, we can finally address the issue of what selective influence human activity has had on baker's yeast, if any, and start to determine the underlying genetics of important traits (for industrial purposes or for biological studies). Is there evidence of domestication in *S. cerevisiae* as previously debated[19, 50]? One could interpret the results of this analysis in two ways. One that there was a domestication of at least one group, the Wine/European strains, or perhaps two with the sake strains, with selection for improved fermentation properties. This domesticated group then gave rise to feral and clinical derivatives as well as being involved in the generation of outbred progeny found in all sources. Alternatively, human activity simply utilised existing strains from populations that have appropriate fermentation properties, and that all human activity had provided was the opportunity to outbreed through movement of strains. Using domestication to imply "species bred in captivity"[51] the strains that best fulfil this definition are the baking isolates as they have clearly arise from crosses between lineages, however further studies are necessary. Comparison to extant *S. paradoxus* should provide a picture of how *S. cerevisiae* looked like prior human activity. Any lineages that were selected from captively bred strains would be expected to have lower diversity than other lineages/populations not selected. This is not the case for the Wine/European or sake lineages, which have similar or greater levels of diversity compared to 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 the results and interpretation of analysis of over 600 strains[52].

In addition to providing a more complete picture of what the genome space of the species *S. cerevisiae* actually looks like, our results raise a large number of questions about genome evolution in general. For example, we have found a great deal of variation within protein coding genes, even essential genes, raising questions about the evolutionary mechanisms that maintain sequence variation in this species. Our strategy to shotgun sequence genomes from many isolates allows us to determine additional sequences as well as rearrangements, raising new questions that could not be found using array CGH. Many of these rearrangements are likely to be subtelomeric, and may explain adaptation to different environments including industrial fermentation properties. The assembly and analysis of these regions will be important for understanding genome evolution in these organisms. Tandem arrays of rDNAs can contain many SNPs within the array that may be useful in developing an understanding of tandem array dynamics and evolution in addition to their use in strain typing. Ty distributions and copy numbers correlate with genome rearrangements over evolutionary time scales, the analysis of which may add to our understanding of the forces governing genome evolution.

A major use of the sequences will be to support genetic analysis between strains, which given the high recombination rate in yeast will allow rapid fine mapping of the genetic determinants of many phenotypic differences between the strains. One approach that has already been taken is QTL analysis of crosses and backcrosses between pairs of founder strains [53, 54]. The near-complete genome sequences and phenotypes identified here provide a rich source of material for such studies. In the future, the recombinant nature of the mosaic *S. cerevisiae* lineages, together with the variation identified here, present an unprecedented opportunity to map to traits of phenotypic importance using naturally occurring recombinant inbred lines by genome wide association studies. The processes leading to these observations and their functional and evolutionary

consequences are ripe for future studies, and the results presented here are only the starting point for future studies.

Methods

Details of the methods mentioned above are provided in supplementary information.

References

1. Goffeau, A. et al. Life with 6000 genes. Science 274, 546, 563-7 (1996).

2. Mewes, H. W. et al. Overview of the yeast genome. Nature 387, 7-65 (1997).

3. Cliften, P. et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301, 71-6 (2003).

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, 241-54 (2003).

5. Begun, D. J. et al. Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol 5, e310 (2007).

6. Clark, A. G. et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450, 203-18 (2007).

7. Dietrich, F. S. et al. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304, 304-7 (2004).

8. Dujon, B. et al. Genome evolution in yeasts. Nature 430, 35-44 (2004).

9. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. Nat Genet 39, S37-42 (2007).

10. 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, 2144-52 (2008).

11. 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).

12. 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, 702-15 (2006).

13. Koufopanou, V., Hughes, J., Bell, G. & Burt, A. The spatial scale of genetic differentiation in a model organism: the wild yeast Saccharomyces paradoxus. Philos Trans R Soc Lond B Biol Sci (2006).

14. 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, 407-11 (2007).

15. Liti, G., Barton, D. B. & Louis, E. J. Sequence diversity, reproductive isolation and species concepts in Saccharomyces. Genetics 174, 839-50 (2006).

16. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. Nat Genet 38, 1077-81 (2006).

17. 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 U S A 105, 4957-62 (2008).

18. Pretorius, I. S. Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. Yeast 16, 675-729 (2000).

19. Fay, J. C. & Benavides, J. A. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. PLoS Genet 1, 66-71 (2005).

20. Bell, P. J., Higgins, V. J. & Attfield, P. V. Comparison of fermentative capacities of industrial baking and wild-type yeasts of the species Saccharomyces cerevisiae in different sugar media. Lett Appl Microbiol 32, 224-9 (2001).

21. McCusker, J. H., Clemons, K. V., Stevens, D. A. & Davis, R. W. Genetic characterization of pathogenic Saccharomyces cerevisiae isolates. Genetics 136, 1261-9 (1994).

22. 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, 177-92 (2005).

23. Pope, G. Saccharomyces Genome Resequencing Project Strains. (2008). http://www.ncyc.co.uk/sgrp.php

24. 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 Pt 5, 1931-42 (2000).

25. Carter, D. M. Saccharomyces Genome Resequencing Project. (2005). http://www.sanger.ac.uk/Teams/Team71/durbin/sgrp/index.shtml

26. The Sanger Institute/EMBL. Ensembl Trace Server. http://trace.ensembl.org/

27. Liti, G. & Louis, E. J. Yeast evolution and comparative genomics. Annu Rev Microbiol 59, 135-53 (2005).

28. Mullikin, J. C. & Ning, Z. The phusion assembler. Genome Res 13, 81-90 (2003).

29. Hong, E. L. et al. Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res 36, D577-81 (2008).

30. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945-59 (2000).

31. Wei, W. et al. Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. Proc Natl Acad Sci U S A 104, 12825-30 (2007).

32. Bensasson, D., Zarowiecki, M., Burt, A. & Koufopanou, V. Rapid evolution of yeast centromeres in the absence of drive. Genetics 178, 2161-7 (2008).

33. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-95 (1989).

34. Coghlan, A. & Wolfe, K. H. Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. Yeast 16, 1131-45 (2000).

35. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15, 1281-95 (1987).

36. Bulmer, M. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897-907 (1991).

37. Clark, R. M. et al. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317, 338-42 (2007).

38. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, 652-4 (1991).

39. Fay, J. C. & Wu, C. I. Sequence divergence, functional constraint, and selection in protein evolution. Annu Rev Genomics Hum Genet 4, 213-35 (2003).

40. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Positive and negative selection on the human genome. Genetics 158, 1227-34 (2001).

41. Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G. & Louis, E. J. Chromosomal evolution in Saccharomyces. Nature 405, 451-4 (2000).

42. Kobayashi, T., Heck, D. J., Nomura, M. & Horiuchi, T. Expansion and contraction of ribosomal DNA repeats in Saccharomyces cerevisiae: requirement of

replication fork blocking (Fob1) protein and the role of RNA polymerase I. Genes Dev 12, 3821-30 (1998).

43. Petes, T. D. Yeast ribosomal DNA genes are located on chromosome XII. Proc Natl Acad Sci U S A 76, 410-4 (1979).

44. Prokopowich, C. D., Gregory, T. R. & Crease, T. J. The correlation between rDNA copy number and genome size in eukaryotes. Genome 46, 48-50 (2003).

45. Ganley, A. R. & Kobayashi, T. Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res 17, 184-91 (2007).

46. Ohta, T. Some models of gene conversion for treating the evolution of multigene families. Genetics 106, 517-28 (1984).

47. Warringer, J. & Blomberg, A. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in Saccharomyces cerevisiae. Yeast 20, 53-67 (2003).

48. Hittinger, C. T., Rokas, A. & Carroll, S. B. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. Proc Natl Acad Sci U S A 101, 14144-9 (2004).

49. Johnston, M. A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. Microbiol Rev 51, 458-76 (1987).

50. Vaughan-Martini, A. & Martini, A. Facts, myths and legends on the prime industrial microorganism. J Ind Microbiol 14, 514-22 (1995).

51. Diamond, J. Evolution, consequences and future of plant and animal domestication. Nature 418, 700-7 (2002).

52. Legras, J. L., Merdinoglu, D., Cornuet, J. M. & Karst, F. Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. Mol Ecol 16, 2091-102 (2007).

53. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752-5 (2002).

54. Steinmetz, L. M. et al. Dissecting the architecture of a quantitative trait locus in yeast. Nature 416, 326-30 (2002).

55. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. Genome Res 8, 464-78 (1998).

56. Fingerman, E. G., Dombrowski, P. G., Francis, C. A. & Sniegowski, P. D. Distribution and sequence analysis of a novel Ty3-like element in natural Saccharomyces paradoxus isolates. Yeast 20, 761-70 (2003).

'**Supplementary Information** accompanies the paper on **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 for comments and suggestions and to L. Kruglyak and J. Schacherer for sharing their unpublished manuscript. We also thank the British Council and Chinese Academy of Sciences for providing the opportunity to conceive and develop this project. Research at The 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 and the 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., V.K. were supported by NERC and I.J.T. by The Wellcome Trust. D.B. was supported by NERC. M.J.T.O. and A.V. were supported by NSF, NIH and a Hertz fellowship.

Author Contributions R.D. and E.J.L. conceived and designed the project. G.L. selected and manipulated yeast strains and extracted DNA samples. M.J., M.A.Q., I.G., S.S., 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 rest. A.M.M. did the selection studies. E.J.L., G.L. D.M.C., L.B. did the population structure analysis. C.M.B. and D.B. performed the analysis of Ty elements abundance. S.A.J., R.P.D., M.J.T.O., A.V. and I.N.R. analysed the rDNA. A.B., V.K. and I.J.T. did the recombination analysis. J.W. and A.B. generated the phenomics data.. E.J.L. with G.L. wrote the paper, coordinating everyone's contributions.

Author Information Correspondence and requests for materials should be addressed to R. D. (rd@sanger.ac.uk) or E. J. L. (ed.louis@nottingham.ac.uk).

Fig.1. *Saccharomyces* phylogenomics

NJ trees based on SNP differences of **a**, *S. cerevisiae* and *S. paradoxus* strains sequenced in this project using *S. mikatae, S. kudriavzevii* and *S. bayanus* as outgroups; **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 name refers to source and colour dots refer to geographic origin.

Fig. 2. *Saccharomyces* population structure

a, Inference of population structure using STRUCTURE on *S. paradoxus*, (markers: 7544 SNPs with >30 strains passing neighbourhood quality standard, NQS), assuming K=6 subpopulations and correlated allele frequencies, linkage model based on marker distances in basepairs, 15000 iteration burn in, and 5000 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.

b, As **a**, but for *S. cerevisiae* (markers: 3413 SNPs with >30 strains passing NQS)

c, Changing topology of the NJ trees along chromosome VIII. The clean lineages exhibit the same NJ topology across the genome whereas mosaic strains exhibit different topologies for different segments. For example in strain UWOPS83.787.3 (green) the leftmost 80kb of chromosome VIII groups with the Wine/European cluster. In the interval from 240-320 kb the strain groups with the North American strains. For the same intervals the lab strain S288c (violet) is in a long branch followed by the Wine/European cluster.

Fig. 3. Population genomics: variation and selection

a, Linkage disequilibrium between pairs of sites as a function of the distance between them. Each point is the average for a 1kb bin. Insets show the decline in linkage disequilibrium over the first 10kb. Points in the *S. cerevisiae* Wine/European plot are more scattered due to the smaller number of strains analysed. Numbers of strains and sites shown in Table S4.

b, Derived allele frequencies in single nucleotide polymorphisms (SNPs) in coding regions. Amino acid changing SNPs (blue bars, 'a') show an excess of low frequencies compared to synonymous SNPs (unfilled bars, 's'). Synonymous SNPs in genes with strong codon bias ('s*', defined as CAI>0.6) show an excess of SNPs at both low and high frequency, suggesting the action of both positive and negative selection. SNPs that create stop codons (red bars, 'create stop') show a very strong skew to low frequencies, suggesting that they are on average more deleterious than amino acid changing polymorphisms. Inset is the number of mutations occurring over the length of the protein, which exceeds three standard deviations (dotted trace) from the mean (solid trace) in the extreme C-terminus (95% of length) indicating that stop codons that remain in the population tend to be those causing relatively mild effects on protein function.

c, Distribution of sizes of insertion/deletion polymorphisms (indels) in coding regions. High frequency indels (minor allele frequency greater than 10%, red bars) show a greater tendency to occur in multiples of 3 than low frequency indels (grey bars), indicating selection against indels that disrupt the open reading frame. Inset is the number of mutations occurring over the length of the

protein, which exceeds three standard deviations (dotted trace) from the mean (solid trace) in the extreme C-terminus (95% of length) indicating that segregating high frequency indels tend to be those causing relatively mild effects on protein function.

d, Minor allele frequency distribution of indels in coding regions. Out of frame indels (not multiples of 3, grey bars) show great excess at low frequencies when compared to in frame indels (unfilled bars). The proportion of out of frame indels (solid trace) decreases as minor allele frequency increases indicating the action of purifying selection to remove most out of frame indels. Error bars represent the standard error of the proportion.

e, Derived allele frequencies in single nucleotide polymorphisms (SNPs) in non-coding regions compared to those in synonymous sites (unfilled symbols). All categories of non-coding SNPs show an excess of low frequency alleles indicating the action of purifying selection, with SNPs in tRNAs (red bars) showing the strongest effect.

f, Distribution of McDonald-Kreitman (M-K) ratios for yeast genes. Dotted trace indicates the expectation in the absence of selection or changes in effective population size. Inset is the ration of amino acid changing to synonymous polymorphisms as a function of allele frequency. Because of the excess of amino acid polymorphism at low frequency, only SNPs with minor allele frequency >20% were used for M-K tests.

Fig. 4. Transposon abundance and rDNA variability

a, Abundance of Ty transposable element sequences across *Saccharomyces* strains. The proportion of Ty sequences in ABI shotgun sequencing reads identified by RepeatMasker is shown for each strain of *S. cerevisae* (red) and *S. paradoxus* (blue). Population are defined as in Fig. 1 and abbreviated as follows: WE - Wine/European, CL - Clean Lineage, MO - Mosaic, UK - United Kingdom, RU - Russia, FE - Far East, DK - Denmark, HA - Hawaii, NA - North America, SA - South America. The 6 strains identified as potential clonemates in Table S5 are also excluded from this analysis. Comparison of overall TE abundance in shotgun reads (3.53%) from *S. cerevisiae* strain S288c with the reference genome sequence from the same strain (3.35%) reveals that estimates of overall TE content from reads are similar to finished genome assemblies. Our estimate of Ty abundance for the finished S288c assembly is slightly higher than previous estimates (3.1%)[55] since we used an expanded RepeatMasker library with newly reported *Saccharomyces* Ty variants[56].

b, Ribosomal DNA polymorphism associated with genome mosaicism. Variable nucleotide positions (substitutions only) found in the 9.1 kb rDNA repeat are mapped for each *S. cerevisiae* strain. The proportion of sequencing reads showing base substitution at a given variable position is indicated by the colour of the bar. Fully resolved polymorphisms (i.e. SNPs) are excluded. Low frequency polymorphisms, indicated by red bars, predominate, with increasing within-strain variation particularly evident in the non-coding IGS1 and IGS2 regions. Strain names are given on the left, with structured genome strains shown in blue and mosaic strains in red. Ribosomal DNA copy number estimate (in green) and total number of polymorphic positions (in black) are shown for each strain on the right. Strains are sorted top to bottom by increasing number of polymorphic positions. Strains with mosaic genomes possess significantly greater numbers of polymorphisms. The average for mosaic strains was 98

polymorphic sites compared to the structured strain average of 34 sites ($p=1.3 \times 10^{-6}$ under Mann-Whitney U rank test). L-1528 was excluded from this analysis due to a high number of unaligned reads resulting in unreliable data.

Fig. 5. *Saccharomyces* phenotype variation

a, Phenotype variation among sequenced *S. cerevisiae* and *S. paradoxus* strains. Strain growth phenotypes in 67 environments were quantified using high resolution micro-cultivation, automated measurements of population density (OD) and calculation of strain doubling times, lags and maximum densities. See SI for details. A) Strain (n=2) doubling time phenotypes in relation to the S288c derivative BY4741 (n=20) are displayed (Logarithmic strain coefficients, LSC = LN (strain/BY4741)). Green = poor growth, red = good growth. Hierarchical clustering of strain doubling time, lag and maximum density phenotypes was performed using a centered pearson correlation metric and average linkage mapping. Blue lines = *S. paradoxus*, pink lines = *S. cerevisiae*, grey line = *S. bayanus* isolate CBS7001. The complete set of phenotypes, including lag and maximum density data, is displayed in Fig. S7.

b-c, Example of large effect polymorphisms (stop codons and indels) that correlated with a phenotype. **b**, shows 15 residues window around SNPs that create a stop codon or out of frame deletion (red boxes) in *Gal2* and *Gal3* respectively. **c**, strains harbouring large effect polymorphisms in genes in the gal pathway tend to show slow growth when galactose is the only carbon source. Red = West African lineage, Blue = reference BY4741.

d-e, Linking gene copy number to growth phenotypes. Best estimates of gene copy number were derived by dividing the number of BLAST matches of

sequence reads to a gene (including only matches starting in the first 10bp of the gene so as to avoid inflation for long genes) by the average aligned coverage depth for that strain. Growth phenotypes (LSC) are put in relation to the reference strain BY4741 (which has no *MEL1* genes and 10 *CUP1* genes). d) *MEL1* versus melibiose growth e) *CUP1* versus copper growth.

Table 1. Origin of strains and sequence coverage

	Source or location[a]	Strains	[b]ABI	[c]AGI
S. cerevisiae		36	43.3	37.3
	Fermentation	13	11.4	
	Clinical	6	5.6	
	Wild	9	10.0	
	Laboratory	4	10.6	37.3
	Baking	3	2.6	
	Unknown	1	3.1	
S. paradoxus		35	38.5	71.8
	England	18	10.7	71.8
	Continental Europe/Siberia	6	12.6	
	Far East Russia/Japan	4	6.8	
	North & South America	6	6.8	
	Hawaii	1	1.6	

[a]Geographic origin of *Saccharomyces* strains and more detailed information are given in Supplementary table S1. Number of [b]ABI and [c]Illumina GA (Solexa) nucleotides successfully aligned and divided by the genome size estimate to give mean coverage depth.

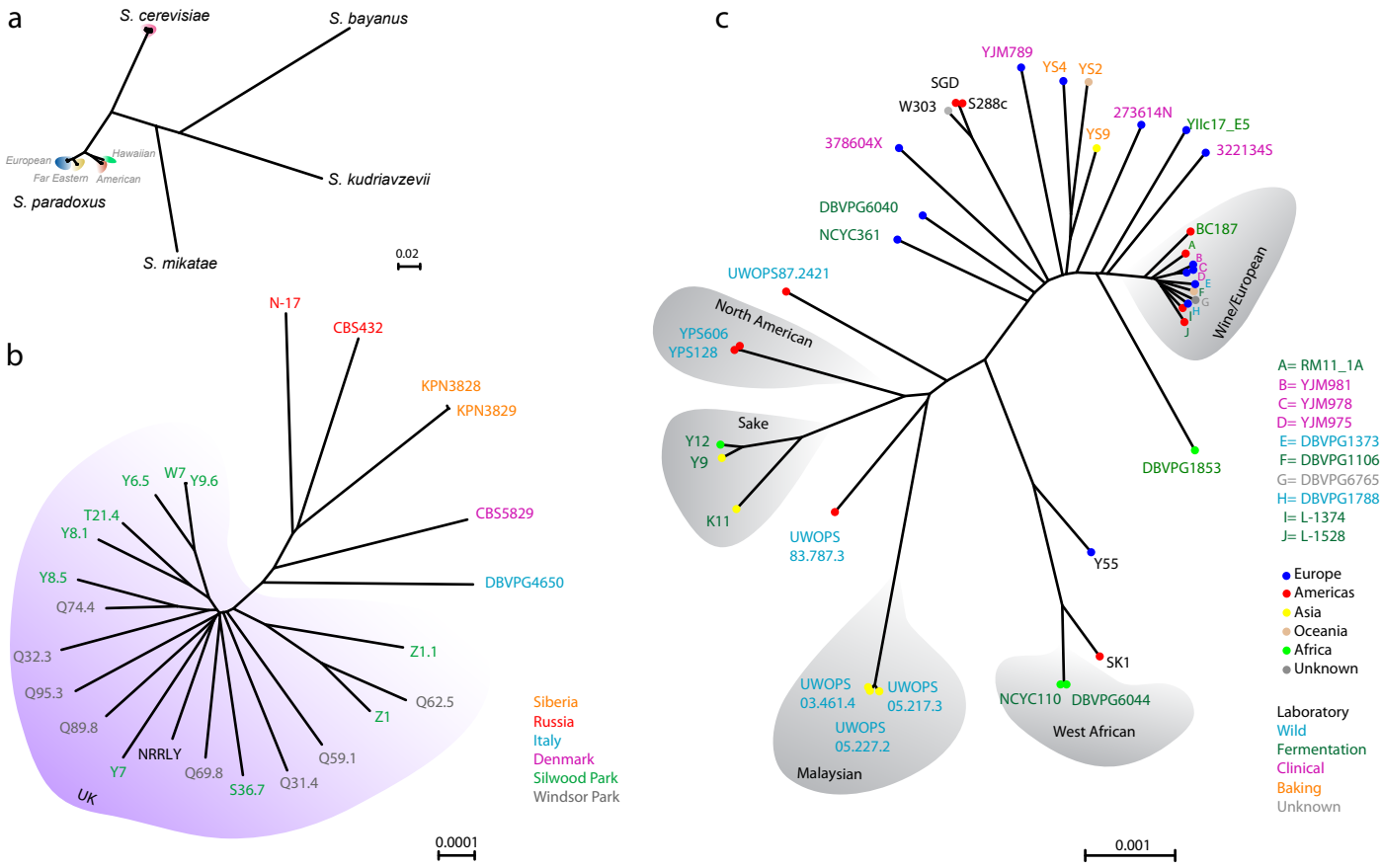
a
S. cerevisiae
S. bayanus
S. kudriavzevii
S. mikatae
S. paradoxus
European
Far Eastern
Hawaiian
American
0.02
b
N-17
CBS432
KPN3828
KPN3829
CBS5829
DBVPG4650
Z1.1
Z62.5
Z1
Q59.1
Q31.4
S36.7
Q69.8
NRRLY
Q89.8
Q95.3
Q32.3
Q74.4
Y8.5
Y8.1
T21.4
Y6.5
W7
Y9.6
Y7
Uk
Siberia
Russia
Italy
Denmark
Silwood Park
Windsor Park
0.0001
c
YJM789
YS4
YS2
SGD
W303
S288c
273614N
YIIc17_E5
YS9
322134S
378604X
DBVPG6040
NCYC361
BC187
A
B
C
D
E
F
G
H
I
J
Wine/European
UWOPS87.2421
North American
YPS606
YPS128
Sake
Y12
Y9
K11
UWOPS
83.787.3
DBVPG1853
Y55
SK1
UWOPS
03.461.4
UWOPS
05.217.3
UWOPS
05.227.2
Malaysian
NCYC110
DBVPG6044
West African
0.001
A= RM11_1A
B= YJM981
C= YJM978
D= YJM975
E= DBVPG1373
F= DBVPG1106
G= DBVPG6765
H= DBVPG1788
I= L-1374
J= L-1528
Europe
Americas
Asia
Oceania
Africa
Unknown
Laboratory
Wild
Fermentation
Clinical
Baking
Unknown

Figure 1

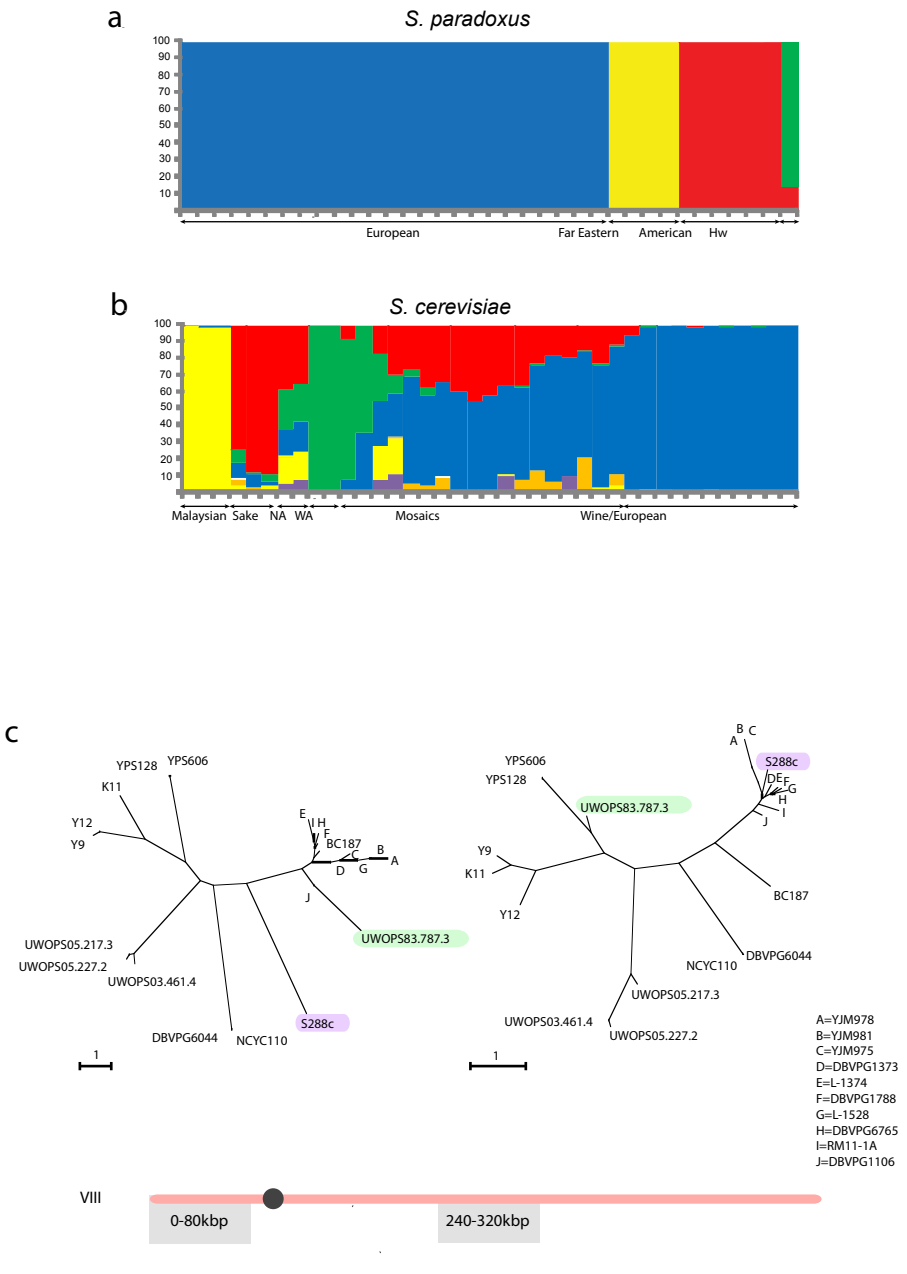
a
S. paradoxus
European
Far Eastern
American
Hw
b
S. cerevisiae
Malaysian
Sake
NA
WA
Mosaics
Wine/European
c
YPS128
YPS606
K11
Y12
Y9
BC187
UWOPS83.787.3
UWOPS05.217.3
UWOPS05.227.2
UWOPS03.461.4
DBVPG6044
NCYC110
S288c
A=YJM978
B=YJM981
C=YJM975
D=DBVPG1373
E=L-1374
F=DBVPG1788
G=L-1528
H=DBVPG6765
I=RM11-1A
J=DBVPG1106
VIII
0-80kbp
240-320kbp

Figure 2

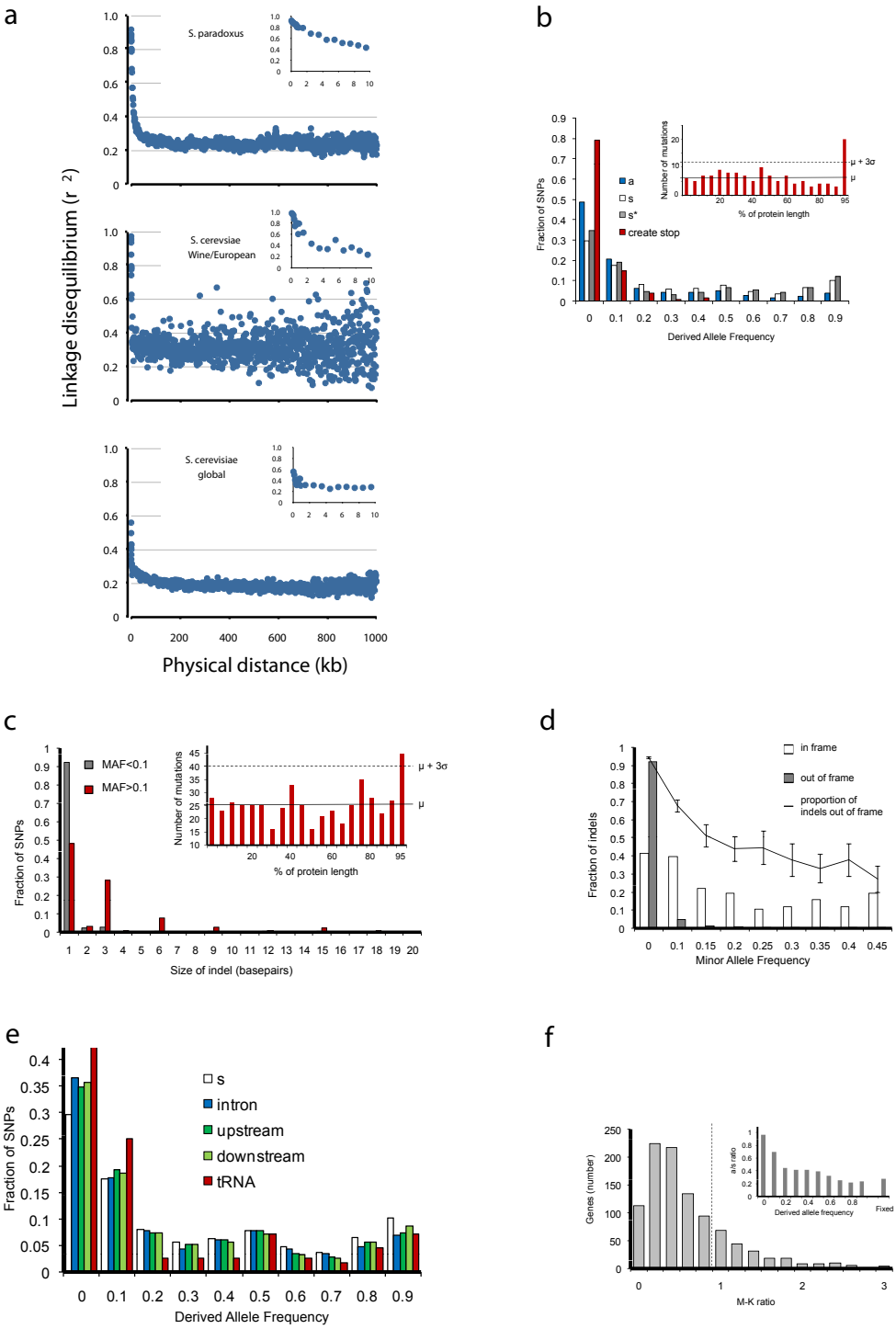
a
S. paradoxus
S. cerevsiae Wine/European
S. cerevisiae global
Linkage disequilibrium (r 2)
Physical distance (kb)
b
Fraction of SNPs
a
s
s*
create stop
Number of mutations
μ + 3σ
μ
% of protein length
Derived Allele Frequency
c
MAF<0.1
MAF>0.1
Fraction of SNPs
Number of mutations
μ + 3σ
μ
% of protein length
Size of indel (basepairs)
d
in frame
out of frame
proportion of indels out of frame
Fraction of indels
Minor Allele Frequency
e
s
intron
upstream
downstream
tRNA
Fraction of SNPs
Derived Allele Frequency
f
Genes (number)
a/s ratio
Derived allele frequency
Fixed
M-K ratio

Figure 3

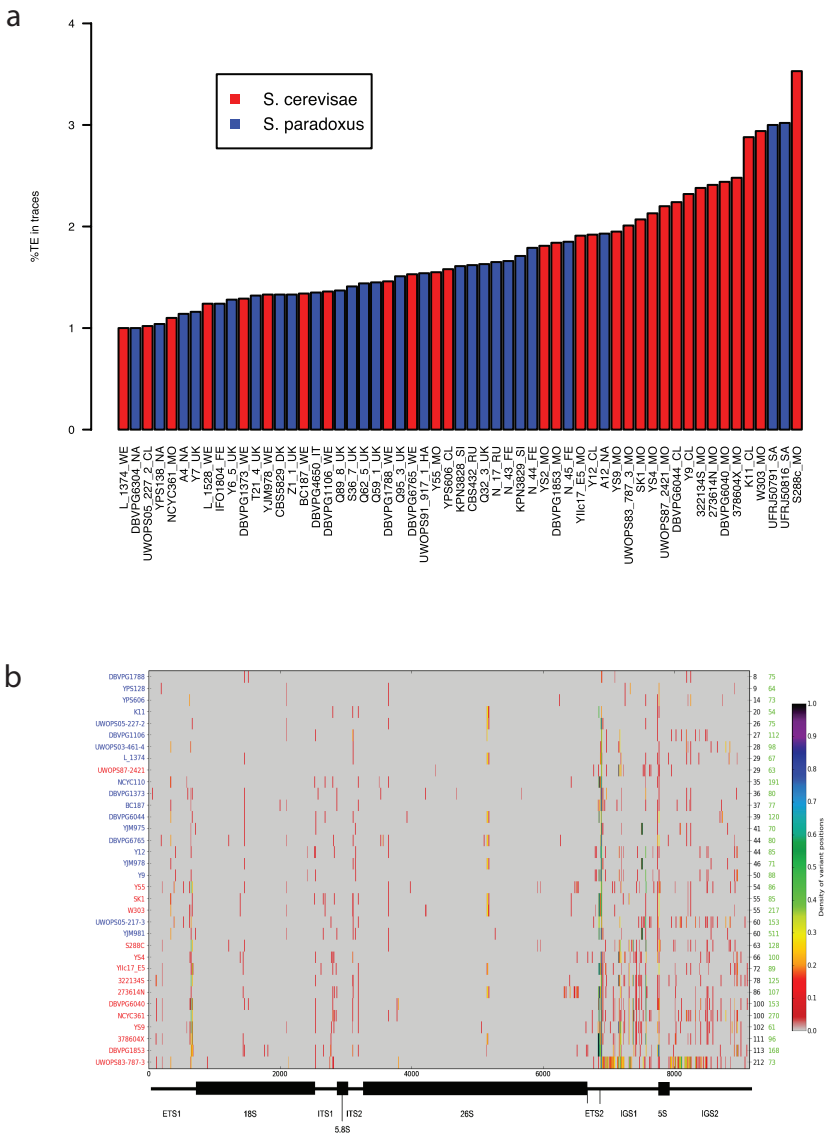
a
b
S. cerevisae
S. paradoxus
%TE in traces
0
1
2
3
4
L_1374_WE
DBVPG6304_NA
UWOPS05_227_2_CL
YPS138_NA
NCYC361_MO
A1_NA
Y7_UK
L_1528_WE
IFO1804_FE
Y6_5_UK
DBVPG1373_WE
I21_4_UK
YJM978_WE
CBS5829_DK
Z1_1_UK
BC187_WE
DBVPG4650_IT
DBVPG1106_WE
Q89_8_UK
S82_7_UK
Q82_5_UK
Q59_1_UK
DBVPG1788_WE
Q95_3_UK
DBVPG6765_WE
UWOPS91_917_1_HA
Y55_MO
YPS606_CL
KPN3828_SI
CBS432_RU
Q32_3_UK
N_17_RU
N_43_FE
KPN3824_SI
N_44_FE
YS2_MO
DBVPG1853_MO
N_45_FE
YIIc17_E5_MO
Y12_CL
A12_NA
YS9_MO
UWOPS83_787_3_MO
SK1_MO
YS4_MO
UWOPS87_2421_MO
DBVPG6044_CL
Y9_CL
322134S_MO
273614N_MO
DBVPG6040_MO
378604X_MO
K11_CL
W303_MO
UFRJ50791_SA
UFRJ50816_SA
S288c_MO
DBVPG1788
YPS128
YPS606
K11
UWOPS05-227-2
DBVPG1106
UWOPS03-461.4
L_1374
UWOPS87-2421
NCYC110
DBVPG1373
BC187
DBVPG6044
YJM975
DBVPG6765
Y12
YJM978
Y9
Y55
SK1
W303
UWOPS05-217-3
YJM981
S288C
Y54
YIIc17_E5
322134S
273614N
DBVPG6040
NCYC361
Y59
378604X
DBVPG1853
UWOPS83-787-3
0
2000
4000
6000
8000
ETS1
18S
ITS1
ITS2
5.8S
26S
ETS2
IGS1
5S
IGS2
Density of variant positions
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 4

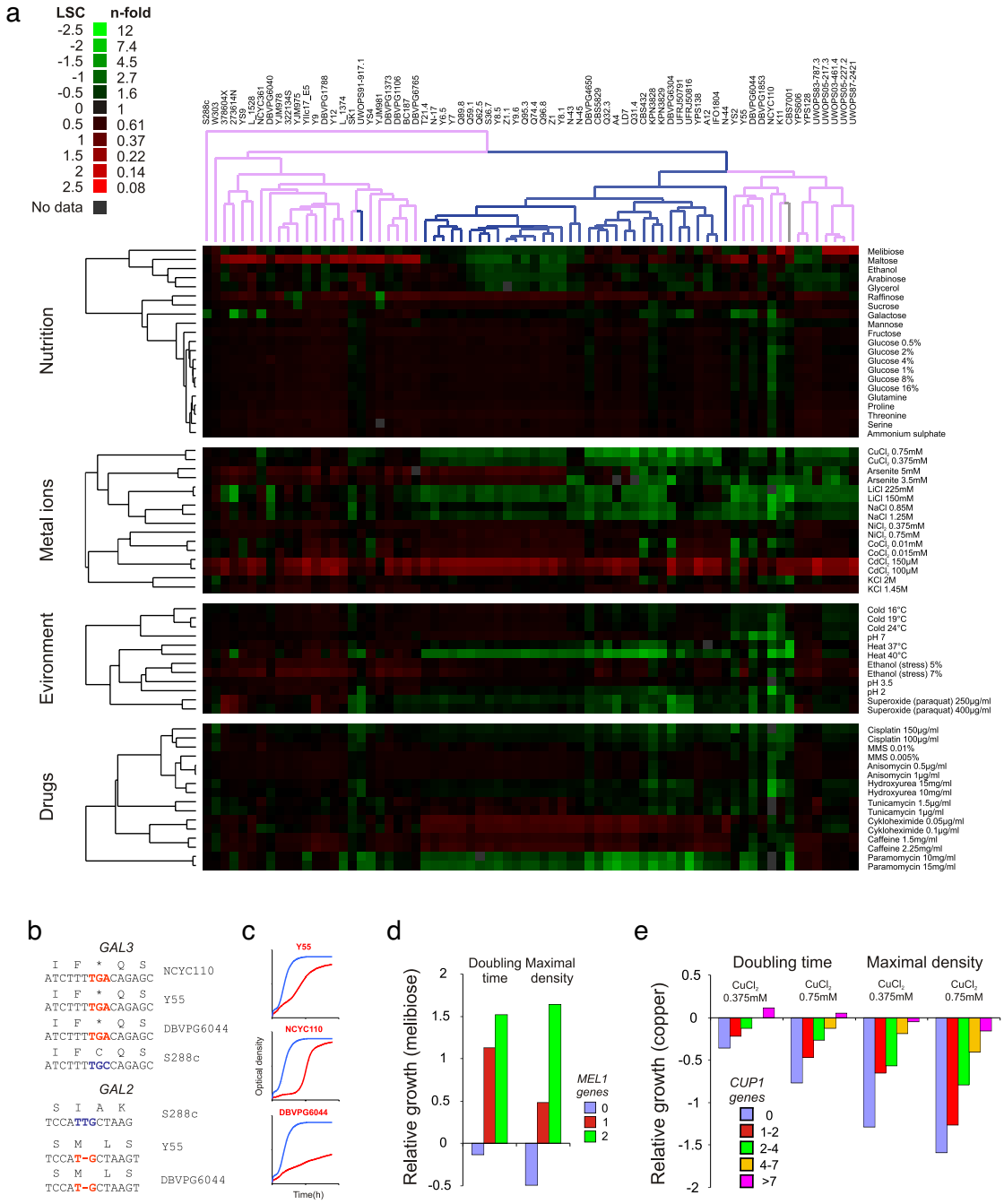
a
LSC
n-fold
-2.5 12
-2 7.4
-1.5 4.5
-1 2.7
-0.5 1.6
0 1
0.5 0.61
1 0.37
1.5 0.22
2 0.14
2.5 0.08
No data
Nutrition
Melibiose
Maltose
Ethanol
Arabinose
Glycerol
Raffinose
Sucrose
Galactose
Mannose
Fructose
Glucose 0.5%
Glucose 2%
Glucose 4%
Glucose 1%
Glucose 8%
Glucose 16%
Glutamine
Proline
Threonine
Serine
Ammonium sulphate
Metal ions
CuCl2 0.75mM
CuCl2 0.375mM
Arsenite 5mM
Arsenite 3.5mM
LiCl 225mM
LiCl 150mM
NaCl 0.85M
NaCl 1.25M
NiCl2 0.375mM
NiCl2 0.75mM
CoCl2 0.01mM
CoCl2 0.015mM
CdCl2 150µM
CdCl2 100µM
KCl 2M
KCl 1.45M
Evironment
Cold 16°C
Cold 19°C
Cold 24°C
pH 7
Heat 37°C
Heat 40°C
Ethanol (stress) 5%
Ethanol (stress) 7%
pH 3.5
pH 2
Superoxide (paraquat) 250µg/ml
Superoxide (paraquat) 400µg/ml
Drugs
Cisplatin 150µg/ml
Cisplatin 100µg/ml
MMS 0.01%
MMS 0.005%
Anisomycin 0.5µg/ml
Anisomycin 1µg/ml
Hydroxyurea 15mg/ml
Hydroxyurea 10mg/ml
Tunicamycin 1.5µg/ml
Tunicamycin 1µg/ml
Cykloheximide 0.05µg/ml
Cykloheximide 0.1µg/ml
Caffeine 1.5mg/ml
Caffeine 2.25mg/ml
Paramomycin 10mg/ml
Paramomycin 15mg/ml
b
GAL3
I F * Q S
ATCTTTTGACAGAGC
NCYC110
Y55
DBVPG6044
I F C Q S
ATCTTTTGCCAGAGC
S288c
GAL2
S I A K
TCCATTGCTAAG
S288c
S M L S
TCCAT-GCTAAGT
Y55
DBVPG6044
c
Y55
NCYC110
DBVPG6044
Optical density
Time(h)
d
Doubling time
Maximal density
Relative growth (melibiose)
MEL1 genes
0
1
2
e
Doubling time
Maximal density
CuCl2 0.375mM
CuCl2 0.75mM
Relative growth (copper)
CUP1 genes
0
1-2
2-4
4-7
>7

Figure 5