Supplemental Material

The 100-Genomes Strains, an *S. cerevisiae* Resource that Illuminates Its Natural Phenotypic and Genotypic Variation and Emergence as an Opportunistic Pathogen

Pooja K. Strope, Daniel A. Skelly, Stanislav G. Kozmin, Gayathri Mahadevan, Eric A. Stone, Paul M. Magwene, Fred S. Dietrich, and John H. McCusker

Supplemental Text

Strain Rationales

S. cerevisiae isolates, S. cerevisiae strain construction, and rationales for S. cerevisiae strain choices

As the term "isolates" implies, *S. cerevisiae* isolates are isolated from environmental sources. Many *S. cerevisiae* isolates are multiply heterozygous (McCusker et al. 1994; Muller and McCusker 2009; Esberg et al. 2011), with ploidies ranging from diploid, to triploid, to tetraploid (Muller and McCusker 2009). Many *S. cerevisiae* isolates do not sporulate, or sporulate very poorly, and many isolates that do sporulate produce no or very few viable spores (McCusker et al. 1994; Muller and McCusker 2009). For these reasons, most *S. cerevisiae* isolates are unsuitable for many types of genetic analysis.

Some *S. cerevisiae* isolates sporulate and produce viable spores (segregants); we defined these mostly homothallic (self-diploidized) segregants as "strains." In contrast to typically multiply heterozygous *S. cerevisiae* isolates (McCusker et al. 1994; Muller and McCusker 2009; Esberg et al. 2011), the 100-genomes strains are homozygous diploids, which greatly facilitated our *de novo*, high quality, and extensively manually edited assembly and annotation of their genomes, as well as simplifying association mapping. A specific criterion in the choice of the 100-genomes strains was that they all sporulate and produce viable spores, making the 100-genomes strains suitable not only for population and quantitative genetics studies but also for the classical (e.g. crossing isogenic strains, making mutations) genetic studies that have previously been limited to laboratory strains. Relative to isolates, the 100-genomes strains are more suitable for quantitative genetic studies that analyze segregants from controlled crosses between non-isogenic strains. For example, as described below, we cross haploid spores from the 100-genomes strains with a haploid, canonical S288c background strain to assess chromosomal rearrangements, which are relevant to quantitative genetic and other studies. Finally, of the 100

strains, 43 are of clinical origin to provide insight into the emergence of *S. cerevisiae* as an opportunistic pathogen.

S. cerevisiae isolates from different geographic locations and environmental sources were obtained from multiple colleagues and from culture collections (Table S1). Isolates that did not sporulate or produced few or no viable spores (segregants) were set aside, as they would not be useful for many types of genetic analyses. For those *S. cerevisiae* isolates that were able to sporulate and produced viable spores, tetrads were dissected. Many *S. cerevisiae* isolates that sporulated and produced viable spores appeared to be heterozygous at multiple loci, as judged by heterozygosity for microsatellites. In addition, amongst segregants of isolates there was often segregation for variable colony sizes and colony morphologies as well as segregation for growth on standard dextrose-containing minimal defined medium (prototrophy vs. auxotrophy for amino acids, purines, or pyrimidines) and/or on the ability to utilize different sugars (McCusker et al. 1994; Clemons et al. 1997; Muller and McCusker 2009); such results are consistent with many of the isolates being heterozygous at multiple loci. The heterozygosities of isolates would prohibit many types of genetic analysis and greatly complicate genome assembly. Therefore, rather than isolates, we focused on haploid (heterothallic (*ho*), which we diploidized) and self-diploidized (homothallic (*HO*)) segregants as strains to sequence and phenotype in this study.

We imposed four criteria on isolate segregants (strains) before establishing final choices on strains to be sequenced. **First**, we analyzed only segregants from tetrads with four viable spores. We imposed this **first** criterion to reduce the likelihood of aneuploid segregants. Second, we excluded self-diploidized (*HO*/*HO*, homothallic) segregants that were unable to sporulate (tested at 25ºC and 30ºC) or had low spore viability. For example, we excluded the self-diploidized homothallic strains YJM280 and YJM339 because of their low spore viability (McCusker et al. 1994). We imposed this second criterion because a fully functional sexual cycle (the ability to sporulate and to produce viable spores/segregants) is important for many types of genetic analysis.

Third, with one exception, we excluded isolate segregants (strains) that were auxotrophic for amino acids, purines, or pyrimidines; that is, were unable to grow on standard dextrose-containing minimal defined medium. We imposed this third criterion because auxotrophies for amino acids, purines, or pyrimidines are deleterious and environmentally limiting. The one exception was YJM1433. YJM1433 is a segregant of the partially sequenced YIIc17_E5 that was previously shown to be auxotrophic (Liti et al. 2009). We found YJM1433 to be His^- and, by complementation testing, *his3*. While there was no obvious inactivating polymorphism, such as a premature stop or frameshift, in the YJM1433 *his3* ORF, there was a unique K216N polymorphism, predicted to be deleterious by PROVEAN (Choi et al. 2012). We introduced a HIS3 ORF PCR product from S288c into YJM1433 to generate a *HIS3* derivative; the His^+ *HIS3*/*HIS3* derivative YJM1869 (isogenic with YJM1433) was used for all phenotypic analyses.

Similar to auxotrophies, in the case of YJM1615 (Pet^-; *HO*/*HO cox15*-ochre/*cox15*-ochre *sup7*/*sup7*), which is isogenic with YJM421 (Pet^+; *HO*/*HO cox15*-ochre/*cox15*-ochre *sup7*/*SUP7*-ochre) (McCusker et al. 1994; Ito-Harashima et al. 2002), we engineered the isogenic Cox^+, Pet^+, sup^+ *HO*/*HO COX15*/*COX15 sup7*/*sup7* derivative YJM1628 to eliminate the deleterious effects of respiratory deficiency (*cox15*-ochre) and ochre suppression (*SUP7*-ochre). YJM1628 (isogenic with YJM1615) was used for all phenotypic analyses.

Finally, we performed microarray analysis on many strains (isolate segregants) (Muller and McCusker 2011) and excluded closely related strains. We imposed this final criterion to increase the genetic diversity of the strains to be sequenced and phenotyped. From these many strains, we chose to sequence the genomes of 93 strains (Table S1, Table S19). With three exceptions (YJM1419, YJM1250, YJM1388), all of the strains sequenced in this study were diploids, either self-diploidized *HO*/*HO* or *ho*/*ho* that we diploidized with a CEN HO plasmid. YJM1419, YJM1250, and YJM1388 were diploidized with a CEN HO plasmid to generate the isogenic strains YJM1847, YJM1870, and YJM1846, respectively, that were used for all phenotypic analyses.

Among the 93 strains that we sequenced to high coverage, assembled, and annotated in this study are single spore clones of 14 strains that had been previously sequenced to low coverage (Liti et al. 2009): Y55 (single spore clone: YJM627), YJM975, YJM978, YJM981, YIIc17_E5 (single spore clone: YJM1433), YPS606 (single spore clone: YJM1434), NCYC110 (single spore clone: YJM1439), UWOPS83-787.3 (single spore clone: YJM1443), UWOPS87-2421 (single spore clone: YJM1444), UWOPS05-227.2 (single spore clone: YJM1447), 273614N (single spore clone: YJM1450), Y12 (single spore clone: YJM1460), DBVPG1853 (single spore clone: YJM1463), and DBVPG6040 (single spore clone: YJM1549).

In addition to the 93 strains we sequenced to high coverage, we also included for phenotyping, association mapping, etc. purposes, isogenic, diploid, prototrophic derivatives, or the parents, of seven commonly used strains that had been sequenced to high, or relatively high, coverage in other studies: These strains are YJM1552 (isogenic with S288c) (Goffeau et al. 1996); YJM145 (the prototrophic *HO* parent of YJM789 (*ho*::hisG *MATα lys2*)) (Wei et al. 2007); SK1 (single spore clone: YJM1077) (Nishant et al. 2010); Σ1278b (diploidized to generate YJM1290) (Dowell et al. 2010); RM11 (YJM1293, the prototrophic *HO* parent of RM11-1a) (RM11 2004); M22 (single spore clone: YJM1529) (Doniger et al. 2008); and YPS163 (single spore clone: YJM1281) (Doniger et al. 2008).

In conclusion, in addition to their lack of heterozygosities facilitating genome assembly, these 100 genetically diverse sequenced strains (93 sequenced in this study, 7 sequenced in other studies) are suitable for association mapping studies, as described here, as well as both classical and quantitative genetic studies.

Sequence Assembly, Annotation, and Analysis

Genome sequencing

DNA was isolated from 40 ml saturated, overnight, 30ºC YPD cultures of the 93 strains using a standard Zymolyase protocol (Burke et al. 2000; Amberg et al. 2005). DNA from the 93 strains was sequenced using the Illumina Hiseq 2000, multiplexed with 12 strains per lane with bar codes using Illumina TrueSeq kits and protocols (www.illumina.com), producing paired end reads of 101 x 101 bases and insert sizes averaging 300 bases. Sequence coverage ranged from 22- to 650-fold per strain.

Genome assembly

After investigating three *de novo* genome assembly programs, Velvet (Zerbino and Birney 2008), SOAPdenovo (Li et al. 2010), and ABySS (Simpson et al. 2009), the assembly of the read pairs was performed using ABySS (v. 1.3.4), as it gave the longest contigs with the fewest assembly errors for this particular data set. The parameters '*k*', the k-mer length, and '*n*', the minimum number of pairs needed to join contigs, were optimized for each strain. The resulting contigs were then assembled into chromosomes using synteny with S288c using BLAST (Altschul et al. 1990). Additional assemblies generated with Velvet were used to identify possible problem regions in the chromosomes. In some cases regions that assembled poorly with ABySS were replaced with corresponding regions from Velvet assemblies. The rDNA units were assembled in a similar way with an additional parameter '*c*', the minimum mean k-mer coverage of a unitig, which was optimized for each strain. rDNA units were inserted into chromosome XII based on depth of coverage of the rDNA sequence.

Using the Illumina data, perl scripts, FASTA (Pearson and Lipman 1988) and tools from the EMBOSS set of programs (Rice et al. 2000), the chromosomes were edited in multiple iterations:

1) Based on depth of coverage analysis, some regions present in single copy in the genome had been mistakenly inserted twice; these problems were resolved.

2) Occurrences of N's or ambiguity codons in unique protein coding genes and in unique regions of the genome were examined by alignment of the raw data and most N's were resolved.

3) Most occurrences of N's or ambiguity codons in regions of low complexity were resolved. However, in cases of homopolymer runs of more than ~18 bases or 2-base repeats of more than ~30 bases, there was often ambiguity as to the exact length of the microsatellite, in which case the most prevalent length was used.

4) Most ambiguities in non-unique genes could be resolved from the available sequence data. Of the ~14,000 ambiguities in non-unique regions of the genome, read pairing identified ~4000 that could likely be resolved of which ~3635 have been edited and resolved and the remainder were not resolved on the first round of editing.

5) Tandem and non-tandem repeated regions, such as *CUP1* and *MAT/HMR/HML*, were checked and edited as necessary using depth of coverage to estimate copy number.

6) Additional errors were identified and corrected by performing a reference based genome assembly using BWA (Li and Durbin 2009) and samtools (Li et al. 2009). This analysis identified ~800 regions where ABySS or Velvet had inserted fewer bases into a homopolymer region than the data supported, and a small number of other errors.

7) Velvet assembly of reads identified ~200 regions of length ~400-20,000 bases that were not included in the initial assembly, in most cases due to lack of similarity to strain S288C. These regions were incorporated into the final assemblies.

Most of the sequence ambiguity problems that could be resolved by editing of the sequence were the result of multiple low quality reads at a specific location, slight cross contamination of sequence data from adjacent Illumina lanes, and the failure of the assemblers used to fully use the read pairing information. Most remaining sequence ambiguities are either telomeric, sub-telomeric, or in other repetitive regions, particularly Ty elements and in the repeated protein coding genes, such as the *PAU, FLO*, and hexose transporter gene families, most of which are sub-telomeric.

To insure that the sequences generated correlated with the initial strains, validation was performed using Restriction Fragment Length Polymorphisms (RFLP's). The PCR primer pairs used for the RFLP are listed in Table S3.

Sequence coverage of each of the 93 strains sequenced in this study is shown in Table S2. Sequence accuracy was estimated by measuring several aspects of the assembled sequences after the manual editing of more than 5000 sequence problems across the 93 assemblies Table S2. A blastp analysis of the ten longest minisatellite-free, single copy gene open reading frames identified no cases of frame shifting or erroneous stop codons. These 10 genes are: *YLR106C* (*MDN1*; 14732 bp), *YKR054C* (*DYN1*; 12278 bp), *YHR099W* (*TRA1*; 11234 bp), *YDR457W* (*TOM1*; 9806 bp), *YLL040C* (*VPS13*; 9434 bp), *YLR087C* (*CSF1*; 8876 bp), *YBL088C* (*TEL1*; 8363 bp), *YGL195W* (*GCN1*; 8018 bp), *YLR454W* (*FMP27*; 7886 bp), *YBL004WA* (*UTP20*; 7481 bp). Note that two genes were excluded from the list of 10 longest genes used in this analysis: *YBR140C* (*IRA1*) and *YOL081W* (*IRA2*), which are duplicate genes originating from the whole genome duplication (Wolfe and Shields 1997) that share 48% protein identity and 63% DNA sequence identity.

The manual editing process also identified that nearly all assembly errors were in telomeric and sub-telomeric sequences and genes (e.g. *MAL*, *COS*, *PAU*, *YRF*); transposable elements; repetitive genes, in particular the hexose transporters; protein coding genes with intragenic repeats (Genes listed in Fig. 1 of (Verstrepen et al. 2005)); nearly identical duplicate genes; and other repetitive sequences. The only other errors observed in any significant numbers were in microsatellites. Runs of more than 6 C's or G's and runs of more than 8 A's or T's were often not properly assembled, apparently due to the large number of low quality score sequences found at these locations. Analysis using BWA to identify sequence contigs missing from the current assemblies, as well as blast analysis of the current telomeric regions with the telomeric regions of S288c identifies that between 17 and 114 kb of genomic sequence that has not yet been included in the assemblies. By blast, this unincorporated sequence contains numerous copies of

MAL, *COS*, *PAU*, *YRF*, telomeric Y' helicases, and Ty encoded genes, as well as matching known telomerase-generated telomeric sequence. Thus it appears that nearly all of the sequence present in the raw data but missing from the current assemblies are repetitive Ty, sub-telomeric, and telomeric sequences. In some cases as much as 20 kb of sequence at a telomere is not assembled, though on average less than 5 kb is not assembled at each telomere end. Approximately 720 of the 2976 telomeres currently end with the telomerase-generated terminal telomeric sequence.

We analyzed the assembled genomes using Pilon ver. 1.8 (Walker et al. 2014). Combining the errors found by Pilon with the missing telomeres, N's, and ambiguity codons, between 144 and 440 identified errors are present per genome. Pilon identifies errors and proposes corrections. In examination of over 400 error/corrections generated by Pilon, nearly all of the errors identified appear to be actual assembly errors. In most but not all cases the proposed corrections appear to be correct. Other than the errors identified at N's and ambiguity codons in the sequence, and at the telomeric ends, most of the errors appear to be errors at microsatellite sequences, and in misassembly of nearly identical repeat sequences, particularly Ty1 delta sequences, though in some cases duplicate genes. A summary of the number of known errors is shown in Table S2. The number of Pilon errors does not include corrections proposed at N's, telomeres, or ambiguity codons, to prevent double counting. In addition, multiple errors reported by Pilon with 100 bases are counted as a single error, as typically the represent a single assembly problem. While there are quite possibly other unidentified errors within the sequence, it appears that the total number of errors per genome is most likely less than 1000, with few of these errors occurring in protein coding genes, other than the specific gene categories mentioned above.

Data Deposition

These 93 fully annotated genome sequences are available via the *Saccharomyces* Genome Database (Cherry et al. 2012) and GenBank (Benson et al. 2013); see Table S19 for accession

numbers. The short reads are available from the Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Tracecs/sra) and 100 sequenced strains are available from the Fungal Genetics Stock Center (McCluskey et al. 2010).

Extraction of sequence regions for phylogenetic and principal component analyses

Once the chromosomes were assembled and annotated, regions (that did not contain translocations, introgressions or large indels with respect to S288c) from each chromosome were extracted. The total length of these regions was 218 kb and included 124 protein-coding genes and intergenic regions. Within these 218 kb, SNP and indel polymorphisms relative to S288c were identified using lagan (Brudno et al. 2003) and perl (see Table S16). Phylogenetic analysis of the 218 kb regions was carried out using ClustalW v2.1 (Larkin et al. 2007) or Mafft v6.864b (Katoh and Toh 2008), and principal component analysis was carried out using the pairwise sequence identity across all strains employing the prcomp function in the R package (R Core Team 2014).

Annotation

The assembled chromosomes were aligned to S288C chromosomes using the software lagan (Brudno et al. 2003) to extrapolate the coordinates of the annotated sequences, and a table file was created using perl. The table file was then utilized by the NCBI tool tbl2asn (http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/) to create the annotation for each chromosome of each strain. Annotated chromosomes were deposited in GenBank, and the raw data in the Short Read Archive. The accession numbers for the 1648 GenBank files for the 93 strains are listed in Table S19.

Introgression analysis

The S288C gene set was used to search against each of the assembled genomes to find regions of low homology (75% to 96% identity) using the program ssearch36 (Smith and Waterman 1981) and BLAST (Altschul et al. 1990). While in most of the cases the sequence identity between any two *S. cerevisiae* strains is between 98-100%, there are regions that have less than 96% identity and in some cases these regions have 98-99% identity with one of the sibling species, indicating that DNA transfer has taken place between these two species by introgression. These regions were BLASTed back to the S288c genome to make sure that the best hit was still the query gene with which we started. These regions were then compared using BLAST with 10 sibling species genomes (*S. paradoxus* (4 strains), *S. kudriavzevii* (2 strains), *S. mikatae, S. bayanus, S. eubayanus,* and *S. arboricolus*) to find any highly identical regions in the sibling species genomes. This allowed identification of introgressed regions with respect to S288c in our 93 genomes. In cases where introgression of a specific gene was found in more than one strain, those sequences were then compared to each other to determine if the strains carried the same or different introgressions. Genes identified as introgressions are listed in Table S5.

Deleted Genes

We identified genes from S288c that were missing in one or more of the 93 strains. Genes were considered deleted when there were no hits (using BLAST and ssearch36) to an annotated S288c protein-coding gene with a sequence identity ≥75%. Genes identified as absent from one or more strains are listed in Table S5.

Novel Genes

From each of the genomes, the regions that match the S288c ORFs (≥80% id), and those that match the known genes not in S288c were taken out. From the remaining intergenic regions, ORFs were derived, with a minimum ORF size of 150. All of these novel ORFs were then clustered to get a unique set of novel ORFs using the program usearch –cluster_fast (Edgar 2010)

with a minimum id of 90% amino acid identity. The unique set of ORFs was then searched in the genomes of the 93 strains, s288c, and the NCBI nr database. For several, we have a close hit in NCBI from other *S. cerevisiae* strains, or other species. From the 93 strains, hits with more than 80% identity (tfastx) are listed in Table S5.

Genes with frameshift and or premature stop codon polymorphisms

For each of the protein-coding gene (excluding the duplicated genes, and those with introns) from S288c, its corresponding hits were gathered from each of the 93 strains. Only those hits that were present as a single copy and without large indels (which would disrupt a whole gene alignment) in the 93 strains were analyzed. Pairwise alignments were generated using lagan for each gene pair using S288c gene as a reference. Using perl scripts, the position of polymorphism between a pair of sequences was identified. Only those polymorphisms that cause frameshift or premature stop are listed in Table S12.

Non-introgression sequence variation and associations in the 100-genomes strains

Protein-encoding genes with frameshift and/or premature stop codon polymorphisms

Focusing on 4,522 annotated (in S288c), single copy genes without introns and without large indels, we identified 576 genes in the 93 strains with ORF length polymorphisms due to frameshift and/or premature stop codon polymorphisms (Table S12). For 57 of these 576 genes with ORF length polymorphisms, there were 49 genes where the S288c ORF length was present in a minority of the 93 strains and eight genes where the S288c ORF length was private to S288c (Table S12). Of the 57 genes where the S288c ORF length is a minority, 49 genes have majority frameshifts (present in 47 to 93 of the strains) and six genes have majority premature stop (present in 56 to 93 of the strains). The remaining two genes have a mixture of frameshifts and premature stops: *FYV12* (31 premature stop and 28 frameshift) and the essential gene *LTO1* (39 premature stop and 36 frameshift).

In 49 of these 576 genes, the ORF length polymorphisms had significant population-associated variation (Fisher's exact test, Bonferroni correction). In two of these genes, the ORF length polymorphisms had significant clinical origin-associated variation (Fisher's exact test, Bonferroni correction, no population structure correction). The *BIO5* ORF length polymorphism had both significant clinical origin- and population-associated variation (Table S7).

Of these 576 genes with ORF length polymorphisms, 35 genes are annotated (*Saccharomyces* Genome Database) as essential. In the 576 genes, 25-30% of the frameshift and/or premature stop codon polymorphisms, respectively, are in the last 10% of the ORFs; due to their ORF locations, these polymorphisms may cause complete, partial, or no loss of gene product functions. However, the remaining 70-75% of these premature stop codon and/or frameshift polymorphisms, respectively, that are distributed throughout the first 90% of the ORFs (Figure S10) seem likely to inactivate gene product functions. Similar to the introgressed and/or present/absent gene sets, genes with the GO terms transmembrane transporter activity and plasma membrane were among those significantly enriched within the frameshift and/or premature stop codon polymorphism gene set (Table S6).

Genes present in some but not all strains

Most of the 5,241 single copy genes present in S288c are also present in all 93 strains. However, 61 genes present in S288c are present in only a subset of the 93 strains (Table S5); for three of these genes, the presence/absence polymorphism frequency differed significantly between populations (Fisher's exact test, Bonferroni correction) (Table S7). Similarly, 219 genes not present or not annotated in S288c are present in at least one of the 93 strains (Table S5); for 18 of these genes, the presence/absence polymorphism frequency differed significantly between populations (Fisher's exact test, Bonferroni correction) (Table S7). While some of the present/absent genes, such as *RTM1*, *BIO1*, and *BIO6*, have been previously described (Ness and Aigle 1995; Hall and Dietrich 2007), many other genes, such as a block of 16 genes from a

species related to *Torulaspora delbrueckii*, are novel (Table S5). Similar to the introgressed gene set, genes with the GO terms transmembrane transporter activity and cell wall were among those significantly enriched within the present/absent gene set (Table S6).

Transposable element (Ty) copy numbers and LTR location polymorphisms

All Ty elements in these 93 strains were of types Ty1-Ty5; no novel transposable elements were found. Based on sequence coverage, and similar to Liti, et al. (Liti et al. 2009), the numbers of each Ty element in the 93 strains varied across a wide range: Ty1 (0 - 34 per strain; 31 in S288c); Ty2 (1 - 33 per strain; 13 in S288c); Ty3 (0 - 9 per strain; 2 in S288c); Ty4 (0 - 7 per strain; 3 in S288c); and Ty5 (0 - 5 per strain; 1 in S288c). Based on sequence coverage, the number of Ty1-Ty5 elements in the 93 strains ranged from 2-62 per strain (Table S9), with a median and average of 30 and 28 per strain, respectively.

Because of the limitations of the short read sequencing used in this project, it was not possible in most cases to distinguish between full-length Ty elements, which are flanked on both sides by Long Terminal Repeats (LTRs), and solo LTRs. Similarly, it was not possible to determine the locations of full-length Ty elements or solo LTRs integrated into repetitive sub-telomeric genes. However, it was possible to determine the locations of non-sub-telomeric LTRs (some of which may be full-length Ty elements) in the 93 strains (Table S10). Finally, the locations of non-sub-telomeric LTRs (some of which may be full-length Ty elements) inserted into the ORFs of single copy genes, which are likely to be phenotypically relevant, are listed in Table S11.

Genes with previously identified, phenotypically relevant polymorphisms and other likely inactivating polymorphisms

We compiled a list of 32 *S. cerevisiae* genes with previously identified, naturally occurring, phenotypically relevant polymorphisms (e.g. non-synonymous SNPs, premature stop

codons, indels, Ty insertions) and determined the presence of these polymorphisms in the 100 *S. cerevisiae* strains, as well as W303, a close relative of S288c, and JAY291. While many of these previously identified, phenotypically relevant polymorphisms showed a wide strain distribution, many others were found in very few strains or were private to one strain (Table S13). We also searched these 32 genes for other likely inactivating polymorphisms (i.e. frameshift, premature stop, deletions) and found six genes (*NCS2*, *AQY1*, *AQY2*, *RSF1*, *RME1*, and *TAO3*) with such polymorphisms in some of the 100-genomes strains (Table S13).

Phenotype rationales and phenotyping methods

Phenotype rationales

We determined multiple phenotypes of the 100-genomes strains (Table S17) for strain characterization purposes; to identify strains with highly divergent phenotypes for future quantitative genetic analyses; and to test for significant population, clinical/non-clinical origin, and genotype associations. Because temperature is a natural environmental variable, we assessed growth in four different media at low and high temperatures. We assessed two related forms of nutrient limitation-induced, cellular differentiation, the dimorphism traits of flocculation and biofilm formation, the latter being a fungal virulence trait (Douglas 2003; Ramage et al. 2009). We determined sporulation phenotypes in six environmental conditions to examine the impact of multiple environments on an important phenotype. In addition to strain characterization purposes, etc., overall sporulation efficiency and the production of 4-spored asci are critical for many types of genetic analysis and thus the usefulness of the 100-genomes strains as a resource.

There is evidence for the variable presence of vitamin biosynthetic pathways in eukaryotes (Helliwell et al. 2013), including in *S. cerevisiae* (Hall and Dietrich 2007), and for vitamin auxotrophy affecting fungal virulence (Sandhu et al. 1976; Domergue et al. 2005). For these reasons, we assessed environment-dependent, vitamin-remediable growth phenotypes (i.e. vitamin auxotrophy) by assaying growth in the absence of each of eight vitamins: nicotinamide

(niacin), p-aminobenzoic acid/folate, pantothenate, pyridoxine, biotin, riboflavin, thiamine, and inositol.

We measured resistance to cycloheximide, which is commonly used to assess drug export and cell membrane permeability defects; the clinically used antifungal drug ketoconazole and the clinically used, natural product polyenes amphotericin B and natamycin, all of which affect the cell membrane; and sulfite, which is both produced by *S. cerevisiae* and used in wine production as a sterilization agent (Fleet and Heard 1993; Romano and Suzzi 1993). We assessed Li^+, Na^+, and pH 8.0 resistance, which requires the variable copy number *ENA* P-type ATPase Li^+/Na^+ pump gene (Goto et al. 1991; Martinez et al. 1991; Wieland et al. 1995; Daran-Lapujade et al. 2009; Warringer et al. 2011). Finally, we assessed resistance to copper. In addition to copper being an essential trace element, copper is also an environmental heavy metal toxicant due, for example, to its use to kill downy mildew in vineyards (Mackie et al. 2012), one environment for *S. cerevisiae*. Copper toxicity also plays a key role in host innate defense against pathogens (Hodgkinson and Petris 2012; Samanovic et al. 2012; Ding et al. 2013).

Measurement of biofilm formation

The ability of the yeast strains to form biofilm was assayed according to Reynolds and Fink procedure (Reynolds and Fink 2001), with slight modifications. Yeast strains were grown in liquid (3 ml) synthetic dextrose SD minimal medium (0.67% yeast nitrogen base without amino acids, 2% D-glucose) in a roller drum at 30^{o}C to saturation (3 days). 1-ml aliquots were transferred to a 96-deep-well microplate (Genesee Scientific, San Diego, USA, cat. No 22-484S). Cells were collected by centrifugation, washed with 0.8 ml of deionized water, and re-suspended in 0.8 ml of SD medium containing 0.1% D-glucose ($SD_{0.1}$). The optical densities of the resulting suspensions were determined by measuring the absorbance at 600 nm (A_{600}) of 20-folds diluted samples using Tecan microplate reader (Tecan Group Ltd., Männedorf, Switzerland). Based on these measurements, yeast suspensions were diluted with appropriate volumes of $SD_{0.1}$ to yield

A_{600} ≈ 1.0, in a new 96-deep-well plate. 100-µl aliquots were transferred to several non-treated polystyrene 96-well microplates (Genesee Scientific, cat. No 25-104). The plates were incubated 6 h at 30°C; after that, 100 µl of 1% Crystal Violet (CV) were added to each well of each plate. After 15-min incubation at room temperature, the staining solution was removed, attached cells were washed 5 times with deionized water and re-suspended in 100 µl of 10% sodium dodecyl sulphate. After 30 min of incubation at room temperature, 100 µl of phosphate-buffered saline (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4; pH 7.4) were added into each well, the resulting solutions were mixed by pipetting, and A_{600} absorbance (since the λ_{max} of CV is 590 nm) of 50-µl aliquots was measured in a Tecan microplate reader. For most of the tested strains, the biofilm-formation data were obtained in 16 replicates (for each strain, 5 independent SD cultures were assayed 3-4 times); for the strains YJM1433, YJM1419, YJM1578, YJM1615 only 6 replicates were performed (for each strain, 2 independent SD cultures were assayed in triplicate).

Measurement of flocculation

Flocculation was assayed using 96-well microplate-adopted version of the Bony et al. procedure (Bony et al. 1998). Yeast strains were grown in liquid SD medium (3-ml cultures) in a roller drum at 30°C to saturation (3 days). 1-ml aliquots were transferred to a 96-deep-well microplate (Genesee Scientific, cat. No 22-484S). Cells were collected by centrifugation, washed twice with citrate/EDTA buffer (0.8 ml of 50 mM-Na-citrate, 5 mM EDTA; pH 3.0), and re-suspended in 0.8 ml of the same buffer. The optical densities of the resulting suspensions were determined by measuring the A_{600} of 20-folds diluted samples on Tecan microplate reader, and the suspensions were diluted with appropriate volumes of citrate/EDTA buffer to yield A_{600} ≈ 1.0 in a new 96-deep-well plate. 210-µl samples were then transferred to a 96-well plate (Genesee Scientific, cat. No 25-104). To determine more precisely the A_{600} values of these completely deflocculated cell suspensions ("D_1"), 10-µl aliquots were removed, mixed with 90 µl of 0.25 M

EDTA (pH 8.0), and analyzed on a Tecan microplate reader. Flocculation was initiated by addition of 4 µl of 1M $CaCl_2$ to the remained 200 µl of each cell suspension. The plate was incubated (periodically inverting) at room temperature for 10 min and finally left standing for 5 min. 10-µl aliquots were removed just below the meniscuses, mixed with 90 µl of 0.25 M EDTA (pH 8.0) and analyzed on a Tecan microplate reader (to obtain "D_2" A_{600}-absorbance values). The percent of flocculation was calculated as $(1-(D_2/D_1))\times100$. The experiment was repeated 3 times.

Sporulation phenotypes

For all sporulation conditions, strains were freshly revived from frozen cultures on YPD plates (1% Yeast Extract, 2% Bacto Peptone, 2% dextrose, 2% Bacto Agar) and grown overnight at 30ºC. Freshly grown cells from these YPD plates were inoculated into 2 ml of liquid YPD (2% dextrose), as the first step in determining liquid 1% KAc (1% potassium acetate) sporulation efficiencies, and onto YPD with 6% dextrose (YPD (6%)) plates, for pre-sporulation growth for determining plate sporulation efficiencies. Both the emphasis on the use of freshly grown cells and the use of YPD (6%) for liquid and plate pre-sporulation growth is as per (Elrod et al. 2009). All liquid growth (30ºC) and sporulation (25ºC and/or 30ºC) were carried out in 2 ml of media in 15 ml Falcon Tubes on roller drums.

Liquid YPD sporulation protocol: To determine YPD sporulation efficiencies, each strain was inoculated into YPD (2% dextrose), which was incubated for one week at 30ºC, at which point sporulation efficiencies were determined.

Liquid 1% KAc sporulation protocol: To determine liquid 1% KAc sporulation efficiencies, cells from liquid YPD (2% dextrose) 30ºC overnight cultures were inoculated into 2 ml of YPD (6%) and grown for ~ 18-24 hours at 30ºC. Cells were harvested at ~ 10^8 cells/ml, washed twice in sterile deionized H_2O, and suspended in 2 ml of 1% KAc to a concentration of ~ 10^7 cells/ml.

1% KAc cultures were incubated (separately) at 25ºC and 30ºC for four days, at which point sporulation efficiencies were determined.

Plate sporulation protocol: Freshly revived strains growing on YPD (2%) at 30ºC were patched (~ 2 cm × 1 cm patches) onto two YPD plates (6%); after 24 hours growth at 30ºC, one toothpick full of cells of each strain were patched (~ 2 cm × 1 cm patches) onto each of two "diet" KAc plates (diet KAc plates = 1% KAc, 2% Bacto Agar). Strains grown on these two YPD (6%) plates for 24 hours at 30ºC were also replica plated to two "regular" KAc plates (regular, or semi-defined, KAc plates: 20 g potassium acetate, 2.2 g Yeast Extract, 0.5 g dextrose, 460 mg COM mixture, 20 g Bacto Agar, 1 l deionized H_2O; COM (Complete) mixture = 800 mg adenine; 800 mg arginine; 800 mg histidine; 1200 mg leucine; 1200 mg lysine; 800 mg methionine; 2000 mg phenylalanine; 8000 mg threonine; 800 mg tryptophan; 1200 mg tyrosine; 800 mg uracil). Each YPD (6%) plate was replica plated to one regular KAc plate to ensure approximately equivalent cell densities of the same strain on KAc plates incubated at 25ºC and at 30ºC. Both diet and regular KAc plates were incubated (separately) at 25ºC and 30ºC for four days, at which point sporulation efficiencies were determined

Determining sporulation efficiency: For each sporulating culture-condition, the numbers of unsporulated cells and of asci with 2, 3, or 4 spores were counted (≥ 300 cells + asci). In addition, for cultures with ≥ 1% 4-spored ascus formation, the numbers of tetrahedral, diamond, and linear 4-spored asci were determined. Percentages were determined as below:

- % sporulation efficiency: number of asci (2, 3, or 4 spores)/(number of asci (2, 3, or 4 spores) + number of unsporulated cells) × 100
- % of 4-spored ascus formation: number of 4-spored asci/(number of asci (2, 3, or 4 spores) + number of unsporulated cells) × 100

High-throughput analysis of yeast phenotypes

Yeast were grown in 80 µl of liquid YPD_{10} (1% yeast extract, 2% peptone, 10% D-glucose) in a 384-well microplate (Greiner Bio-One, cat. No 781186). Cells were then transferred robotically onto rectangular agar plates (Greiner Bio-One, cat. No 781186) at a density of 1536 dots per plate using a 384-pin replicating tool with a BM5 robot (S&P Robotics Inc., Ontario, Canada). To minimize colony position effects (i.e. edge of the plate *versus* internal area; colony neighborhood), for each tested strain, 24 replicates were plated as 6 × 4 blocks, and only the 8 colonies internal to each block were scored. The plates were incubated at 30°C (except for temperature sensitivity tests, see below) for 1-3 days and imaged every 24 h using digital camera of the BM5 robot. The colony areas were quantified using ImageJ 1.47v program (available at http://imagej.nih.gov/ij/index.html) and Patch Detector Plus plug-in (available at University of Graz Microscopy Facility website: http://microscopy.uni-graz.at/index.php?item=new1). Thus, for each strain, we calculated the ratio of the median colony size (obtained in 8 internal colony replicates) observed on experimental phenotype-tester plates to the average colony size (obtained in 8 internal colony replicates) observed on control plates. These ratios were used to quantify and compare the phenotypes among individual strains. In addition, two initial tests (sulfite- and copper-sensitivity) were performed with 3 different random arrangements of 6 × 4 blocks of colonies of each strain (see above); however, based on the results obtained with these arrangements being indistinguishable (data not shown), all other tests were performed using one fixed arrangement of strains. With the exceptions of copper, lithium, and sodium resistance, all phenotypes were determined in media containing 0.2% weight/volume of water-soluble nigrosin (Sigma).

Sulfite sensitivity was measured on acidified YPD medium (1% yeast extract, 2% peptone, 2% D-glucose, 2% agar) containing 9 g/L citric acid and 4 g/L sodium citrate

monohydrate (Casalone et al. 1989) and supplemented (or not) with 3.0 mM and 6.0 mM sodium sulfite.

Copper sensitivity was tested on SD plates (0.67% yeast nitrogen base without amino acids, 2% D-glucose, 2% agar) supplemented (or not) with $CuSO_4$ at 0.075 mM, 0.1 mM, and 0.25 mM concentration.

Resistance to Li^+ (50mM LiCl), Na^+ (1M NaCl), and alkaline pH (pH8.0, 50mM MOPS) was determined on synthetic minimal medium (SD), similar to (Goto et al. 1991; Wieland et al. 1995; Daran-Lapujade et al. 2009; Warringer et al. 2011).

Sensitivity to several fungicides was scored on YPD plates supplemented (or not) with one of the following fungicide: cycloheximide (at concentrations of 0.25 mg/L and 0.5 mg/L); ketoconazole (at concentrations of 10 mg/L and 20 mg/L); amphotericin B (at concentration of 15 mg/L); and natamycin (at concentration of 3 mg/L).

Growth at different temperatures was tested on YPD, YPEG (1% yeast extract, 2% peptone, 2% ethanol, 2% glycerol, 2% agar), SD, and SEG (0.67% yeast nitrogen base without amino acids, 2% ethanol, 2% glycerol, 2% agar) plates incubated at 15°C, 37°C, 39°C, and 30°C (control).

Vitamin-prototrophy was scored on SD plates containing 0.17% vitamins-, amino acids-, and ammonium sulfate-free yeast nitrogen base (American Biorganics, Inc., Niagara Falls, USA, cat. No A25-9685), 0.5% ammonium sulfate, 2% D-glucose, 2% Difco Noble agar, and supplemented with all (control) or lacking one (or two, in case of folic acid/p-aminobenzoic acid omitted plate) of the following vitamins: 2 mg/L of D-biotin, 400 mg/L of calcium pantothenate, 2 mg/L of folic acid, 200 mg/L of p-aminobenzoic acid, 2000 mg/L of inositol, 400 mg/L of niacin, 400 mg/L pyridoxine hydrochloride, 200 mg/L of riboflavin, 400 mg/L of thiamine hydrochloride.

Polymorphisms and Principal Component Analysis of the 93 strains and S288c

Non-sub-telomeric regions (i.e. regions proximal to repetitive sub-telomeric genes) with only single-copy genes and no introgressions were extracted from each of the 16 chromosomes of each strain to provide an unbiased view of variation in unique sequence (total = 218 kb; 124 protein-coding genes; Supplemental Material, Table S16). In these regions, the percentage identity per strain relative to S288c ranged from 99.29% - 99.6%, with an average of 99.51%. The genic and intergenic polymorphisms consisted of approximately 93% SNPs and 7% indels. Within the 124 protein-coding genes, 67% of the polymorphisms were synonymous SNPs. The SNP frequency in the 218 kb of these 93 strains relative to S288c is approximately one per 150-250 bases. Comparison of polymorphisms between strains suggests extensive assortment has occurred (data not shown). Despite this assortment, principal component analysis of the sequence identity values of the 218 kb region (Figure S8), and a phylogeny (Figure 4) using the SNPs in the same region, show the clustering of the six populations described in the population structure analysis and clusters of clinically- and non-clinically-derived strains.

Population Structure and Association Testing

Obtaining genetic variants for population structure analysis

To facilitate the analysis of population structure in previously sequenced genomes along with the genomes we sequenced for this study, we called genetic variants using short reads and a common reference genome (S288c, SGD release 64). For the 93 strains sequenced in this study, 101 base pair Illumina paired-end reads were aligned to the yeast reference genome using the mem algorithm implemented in the program BWA version 0.7.4 (Li and Durbin 2009). For the six previously sequenced strains in the 100 genomes project (excluding the reference S288c), we used genome assemblies to construct files of short reads that could be mapped to the reference genome as above. Specifically, for two previously sequenced strains (M22 and YPS163; (Doniger et al. 2008)) with relatively short contig sizes, we simulated 101 base pair single-end reads to achieve 60× coverage. For the remaining four strains (RM11, Sigma1278b, SK1, YJM789;

http://www.broadinstitute.org, (Wei et al. 2007), (Dowell et al. 2010), (Nishant et al. 2010)) we simulated 101 base pair paired-end reads to achieve 60× coverage.

Picard version 1.101(http://picard.sourceforge.net) and SAMtools version 0.1.18 (Li et al. 2009) were used for sorting and merging of BAM output files from BWA. Freebayes 0.9.9 (Garrison 2012) and SAMtools (Li et al. 2009) were used for polymorphism discovery and genotyping. Freebayes genotype calls with quality greater than 10 were retained for further analysis. We found that this quality threshold resulted in reliable genotypes, as compared to genic regions from the manually edited *de novo* assemblies. Specifically, we used *de novo* assemblies of the 93 strains sequenced in this study to compile alignments of 5,088 genes (requiring a single best blat hit, >95% sequence identity, >80% coverage of the full gene). Comparing variants called using sequences from *de novo* assemblies and short reads, we found high concordance for both SNP and indel genotypes (mean 99.96% and 98.6% concordance, respectively). As expected, we found that most SNPs (98.9%) were detected by both methods, but that indels were undercalled in the short-read data (493 indels called by both methods as compared to 2,617 called only using *de novo* assembly data).

A variety of other *S. cerevisiae* strains have been previously sequenced to varying levels of coverage. For analyses of population structure, we included data from 21 strains sequenced by (Liti et al. 2009) and 23 strains available at *Saccharomyces* Genome Database (http://www.yeastgenome.org). We did not include duplicates of strains sequenced in multiple studies, or strains isogenic to strains sequenced in this study. Genotypes were obtained by aligning raw Sanger reads (for genomes from (Liti et al. 2009)) or simulated 101 base pair single-end reads (for the remaining assembled genomes) to the reference S288c using BWA as above, and calling genotypes using base quality values at each site.

Population structure analysis

For investigations of population structure, we focused on biallelic single nucleotide polymorphisms (SNPs). *structure* version 2.3.4 was used to infer population structure and assign individual strains to populations (Pritchard et al. 2000). *structure* assumes that loci are independent within populations, which is not a valid assumption for complete sequence data. Thus, PLINK (Purcell et al. 2007) was used to prune out SNPs in high linkage disequilibrium ($R^2 > 0.4$). This reduced set of SNPs (N = 24,360) was used as input to *structure*. Since *structure* is very slow with a set of SNPs this large, we sampled SNPs to create four largely independent smaller datasets consisting of N = 1,210 SNPs (approximately one SNP per 10 kb). We binned SNPs by minor allele frequency (MAF; ten equally spaced bins for $0 < MAF < 0.5$). SNP sampling probability was inversely proportional to the number of SNPs in the corresponding MAF bin. This procedure resulted in datasets with an approximately uniform distribution of MAFs. The procedure was motivated by the observation that most SNP variation in the dataset is rare. Common variation is likely to be useful for elucidating broad-scale patterns of population structure, while rare variation can provide finer-scale information (Henn et al. 2010). Although common variation would be suitable for our goal of ascertaining broad-scale patterns of population structure, we theorized that rarer variation might also be useful for distinguishing populations with limited sampling in our dataset. Thus, our procedure for sampling SNPs allowed for the inclusion of some rare variation while ensuring that the dataset was not dominated by rare SNPs.

structure runs used the linkage model, with a burn-in period of at least 200,000 iterations and a minimum of 1,000,000 iterations of MCMC after the burn-in. The best-fitting value of K, the number of populations, was investigated in two stages. First, at least three preliminary *structure* runs for K = 3 through K = 15 were carried out on one dataset for at least 250,000 iterations. Values K = 3 and $K > 10$ were significantly less likely (as measured by the estimated log probability of the data) than $3 < K < 11$. Subsequently, longer runs of at least 1,000,000 iterations post-burn-in were carried out for K = 4 through K = 10 on all four datasets, with three

independent runs per *K* value per dataset. Results presented in the text were obtained by averaging results from the three independent runs of the program for each of the four subsampled datasets described above. Results for independent runs were compared using CLUMPP version 1.1.2 (Jakobsson and Rosenberg 2007) to manage label switching and multimodality. We assigned individual yeast strains to one of six populations using results from *structure*. Specifically, as described in the legend to Figure 3 in the main text, a threshold of 60% ancestry from any single population was used to assign strains to populations, except for mosaic strains that have less than 60% ancestry from any of the other five populations.

Association testing

We tested for association between phenotypes and several types of genomic variation (restricting to a minor allele frequency of at least 5% for all classes of variation):

(1) all biallelic SNPs

(2) presence/absence of genes lost in a subset of strains

(3) presence/absence of novel genes not present in the reference strain S288c

(4) *S. cerevisiae* vs. other sequence in introgressed regions

We tested for association at all biallelic SNPs in order to maximize the opportunity to test the causal allele itself. In addition, we searched for local peaks of high association signal among SNPs, which would be expected to be especially pronounced (when testing all SNPs) if linkage disequilibrium is extended by recent selection acting on the causal variant.

The program GEMMA version 0.94beta (Zhou and Stephens 2012) was used to conduct association tests. This program takes a linear mixed model approach to controlling population structure using a relatedness matrix estimated using data at all loci. As noted by Zhou and Stephens (2012), mixed linear models employ this single genome-wide relatedness matrix to account for both relatedness among samples and for population stratification. For this exploratory analysis, we conducted association tests at all SNPs and examined in detail any hit with a highly significant *p*-value ($p < 10^{-6}$).

In order to examine the power and false positive rate for tests of genotype-phenotype association, we used our actual genotype data and simulated phenotype data for causal alleles with a range of effect sizes. Specifically, for N=1,000 simulations per effect size, we chose a single biallelic SNP at random from the genome and assigned it as the causal SNP. The genotypes for this SNP were then used to simulate phenotypes with effect size (proportion of phenotypic variance explained by the causal SNP) ranging from 5% to 90% (in increments of 5%) according to formulas presented in (Long and Langley 1999). The initial genotype data consisted of a set of 10,938 SNPs with minor allele frequency at least 5% and genotype calls for at least 95/100 strains. This set of SNPs resulted from pruning a larger dataset to remove one of each pair of SNPs with high linkage disequilibrium (LD); this pruning was performed using plink (Purcell et al. 2007) in order to remove SNPs in high LD ($R^2 > 0.5$). Figure S11A indicates that the power to detect loci of large effect (>35%) is relatively high (>80% at uncorrected $p < 1\times10^{-7}$). However, it is possible for false positives to arise even at very low significance thresholds, as illustrated in Figure S11B.

Genetic and Molecular Methods

Genetic and molecular testing for chromosome co-linearity in the 100 strains

Crosses of the canonical *S. cerevisiae* S288c background with the 100 sequenced strains, followed by sporulation and tetrad dissection, are a genetic species test (Naumov 1986; Naumov 1987; McCusker et al. 1994; Naumov 1996; Naumov et al. 2000; Naumov et al. 2006; Naumov et al. 2010) and test for chromosome co-linearity, at least outside of sub-telomeric regions that lack distal essential genes. Chromosome co-linearity with S288c greatly facilitates genome assembly. Chromosome co-linearity, and correspondingly high spore viability in crosses, also facilitates genetic analysis, particularly quantitative genetic analysis.

Deviations from chromosome co-linearity are chromosomal rearrangements that can have major effects on phenotypes (Sherman and Helms 1978; Perez-Ortin et al. 2002; Zimmer et al.

2014). Chromosomal rearrangements include chromosomal inversions and reciprocal chromosomal translocations that will only be identified by short read sequencing if their breakpoints do not involve large repeated sequences, such as Ty elements. Because of the limitations of short read sequencing in assessing chromosome co-linearity and identifying chromosomal rearrangements, we performed crosses and determined percent spore viabilities and spore viability patterns to identify strains with chromosomal rearrangements.

A haploid S288c background strain (YJM1617: *ho*Δ::kanMX4 *MATα* ρ^0) was crossed with haploid spores of each of the 100 strains (Table S4), including a positive (S288c background) control. Diploids were selected on YP(Ethanol + Glycerol) + G418 and sporulated at 30ºC. Tetrads were then dissected and spore viability was determined (Table S4). In addition to being genetic species tests, these crosses assessed the chromosome co-linearity (outside of sub-telomeric regions) of chromosomes. High spore viability crosses demonstrate co-linearity (outside of sub-telomeric regions) of all chromosomes of that strain with S288c. Crosses of 79 of the sequenced strains, including the positive control S288c × S288c cross (YJM1617 × YJM1552), had high percent spore viability with a high proportion of tetrads with four viable spores, consistent with these strains being members of the species *S. cerevisiae* with chromosomes co-linear with S288c. However, crosses of 21 strains with S288c showed lower spore viabilities with, in many cases, informative spore viability patterns.

YJM1250: While the known approximately 32.5 kb inversion in YJM789 (isogenic with YJM145) (Wei et al. 2007) has no obvious effect on spore viability, larger inversions would be expected to reduce spore viability, with a large excess of tetrads with two viable spores that would result from an odd number of recombination events within the inversion. Therefore, one hypothesis for the observed spore viability and spore viability pattern of YJM1617 × YJM1250 is heterozygosity for a large inversion.

YJM1447: We find that YJM1447, a single spore clone of UWOPS05-227.2, has high sequence similarity to *S. cerevisiae*; SGRP sequence data of the Malaysian isolate UWOPS05-227.2 also shows high sequence similarity to *S. cerevisiae* (Liti et al. 2009). Despite the high sequence similarity, S288c × YJM1447 had very low spore viability; other workers have shown crosses of Malaysian *S. cerevisiae* strains with other *S. cerevisiae* strains have similarly low spore viability (Naumov et al. 2006; Cubillos et al. 2011). Such low spore viability is consistent with heterozygosity for multiple reciprocal translocations. A diploid heterozygous for four reciprocal translocations with essential genes distal to all breakpoints would have a theoretical maximum of approximately 6.25% spore viability, which is similar to what we observe. Therefore, one hypothesis for the observed low spore viability of YJM1617 × YJM1447 is that YJM1447, and possibly other Malaysian *S. cerevisiae* strains, have multiple (likely four) reciprocal chromosomal translocations relative to S288c. PCR analysis showed that YJM1447 does not have a previously described (Zimmer et al. 2014) chromosome 15-16 translocation (data not shown).

YJM195: Sequence analysis showed that YJM195 has a 258 kb pericentric inversion on chromosome 9 at a 5 base pair microhomology, AGTAG, located between RPI1 and RHO3 at 138633 and on the minus strand between YIR020C and MRS1 at 396233, with essential genes being distal to both breakpoints. Consistent with YJM195 having a pericentric inversion, a cross of the haploid S288c background strain YJM1617 with haploid spores of YJM195 had 45.8% spore viability with few tetrads with four viable spores and an excess of tetrads with two viable spores.

YJM456, YJM1342, YJM1387, YJM1443, and YJM1592: Crosses of the haploid S288c background strain YJM1617 with haploid spores of YJM456, 1342, 1387, 1443, and 1592 had spore viabilities of 40.9 to 50.5% with similar spore viability patterns. A diploid heterozygous

for one reciprocal translocation with essential genes distal to both breakpoints would have few tetrads with four viable spores and an excess of tetrads with two viable spores. The observed spore viability patterns and percent spore viabilities are consistent with these diploids (YJM1617 × YJM456, 1342, 1387, 1443, and 1592) being heterozygous for a reciprocal translocation with essential genes being distal to both breakpoints. PCR analysis showed that none of these strains has a previously described (Zimmer et al. 2014) chromosome 15-16 translocation (data not shown). Sequence analysis showed that none of these strains has the pericentric inversion on chromosome 9 found in YJM195.

YJM189, YJM969, YJM972, YJM978, YJM981, YJM987, YJM996, YJM1129, YJM1433, YJM1526, YJM1529, YJM1549, and YJM1574: Crosses of the S288c background strain YJM1617 with haploid spores of YJM189, 969, 972, 978, 981, 987, 996, 1129, 1443, 1526, 1529, 1549, and 1574 had spore viabilities of 51.6 to 71.3% with similar spore viability patterns. A diploid heterozygous for a reciprocal translocation with no essential genes distal to one breakpoint (likely in a sub-telomeric region) and essential genes distal to the second breakpoint would have few tetrads with four viable spores and excesses of tetrads with three and two viable spores. The observed spore viability patterns and percent spore viabilities suggest that these diploids (YJM1617 × YJM189, 969, 972, 978, 981, 987, 996, 1129, 1433, 1526, 1529, 1549, and 1574) are heterozygous for a reciprocal translocation with essential genes distal to only one breakpoint.

The previously described *ECM34-SSU1* reciprocal translocation (Perez-Ortin et al. 2002) has no essential genes distal to the sub-telomeric *ECM34* breakpoint and multiple essential genes distal to the *SSU1* breakpoint. We used PCR analysis of all 100 strains to determine whether strains had the canonical S288c-like *ECM34* (sub-telomeric chromosome 8) and *SSU1* (chromosome 16L) chromosome structures or the *ECM34-SSU1* reciprocal translocation. While

88 strains, including YJM1433, had the canonical S288c-like *ECM34* and *SSU1* chromosome structures, YJM189, 969, 972, 978, 981, 987, 996, 1129, 1526, 1529, 1549, and 1574 all had the *ECM34-SSU1* reciprocal translocation (Table S4). Genome sequencing and assembly further confirmed these chromosome structures. We describe the contribution of the *ECM34-SSU1* reciprocal translocation to sulfite resistance in this study.

In conclusion, 21 of the 100 strains show evidence (percent spore viability, spore viability pattern) consistent with chromosomal rearrangements. For 12 of these 21 strains, we identified the relevant chromosomal rearrangement.

Yeast genomic DNA isolation used for PCR analyses

Yeast genomic DNA was isolated from 40-ml saturated YPD cultures (grown overnight at 30^{o}C with agitation) by Zymolyase-mediated protocol, described in (Burke et al. 2000).

Detection of the ECM34-SSU1 translocation and restriction fragment length polymorphism (RFLP) analysis of the 100 strains

The presence or absence of the *ECM34-SSU1* translocation was tested by PCR using primers ECM34D (5'-tcg aac atc gag cat gca-3'), ECM34R (5'-cca tat ttg tga tga tat cg-3'), SSU1MD (5'-acc tat cga gtc tcc cac-3'), and SSU1R (5'-gac acc cat gac cat cac-3') (Perez-Ortin et al. 2002). Genome sequences were used to design primers that, when digested by the appropriate restriction enzymes, would distinguish all 100 strains; the PCR products for RFLP analysis were generated using primer pairs listed in Table S3. In all cases, PCR amplification was performed with ~ 0.5-1 µg of the purified genomic DNA (as described in previous paragraph) as a template using Platinum Taq DNA polymerase (Invitrogen). The PCR conditions were: 1 min at 94^{o}C, 30 × (30 sec at 94^{o}C, 30 sec at 50^{o}C, 1 min at 72^{o}C), 5 min at 72^{o}C. Amplification of ≈ 0.6 kb and ≈ 0.5 kb PCR-products with the primer pairs ECM34D+SSU1R and SSU1MD+ECM34R,

respectively, and lack of the DNA amplification with primer pairs SSU1MD+SSU1R and ECM34D+ECM34R indicated the presence of *ECM34-SSU1* translocation (Table S4). In contrast, amplification of approximately 0.6 kb and approximately 0.2 kb PCR-products with the primer pairs SSU1MD+SSU1R and ECM34D+ECM34R, respectively, and lack of the DNA amplification with the primer pairs ECM34D+SSU1R and SSU1MD+ECM34R indicated absence of the *ECM34-SSU1* translocation (Table S4).

PCR-products amplified for the RFLP analysis were purified using QIAquick 96 Multiwell PCR Purification Kit (Qiagen). 25-µl aliquots of the purified PCR products were then hydrolyzed with 0.8-3.0 U of the corresponding restriction endonuclease (all restriction endonucleases were from New England BioLabs) (Table S3) and whole reaction mixtures were analyzed by agarose gel electrophoresis. In all cases, size of the intact PCR fragments was ≈ 0.5 kb, whereas size of hydrolyzed fragments was reduced to ≈ 0.25 kb.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Amberg DC, Burke DJ, Strathern JN. 2005. *Methods in Yeast Genetics: A Cold Spring Harbor Laboratory Course Manual*. Cold Spring Harbor Laboratory Press.

Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OV, Missawa SK, Galzerani F, Costa GG, Vidal RO, Noronha MF et al. 2009. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. *Genome research* **19**(12): 2258-2270.

Baruffini E, Lodi T, Dallabona C, Foury F. 2007. A single nucleotide polymorphism in the DNA polymerase gamma gene of *Saccharomyces cerevisiae* laboratory strains is responsible for increased mitochondrial DNA mutability. *Genetics* **177**(2): 1227-1231.

Ben-Ari G, Zenvirth D, Sherman A, David L, Klutstein M, Lavi U, Hillel J, Simchen G. 2006. Four linked genes participate in controlling sporulation efficiency in budding yeast. *PLoS genetics* **2**(11): e195.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. *Nucleic acids research* **41**(Database issue): D36-42.

Bony M, Barre P, Blondin B. 1998. Distribution of the flocculation protein, flop, at the cell surface during yeast growth: the availability of flop determines the flocculation level. *Yeast* **14**(1): 25-35.

Brown KM, Landry CR, Hartl DL, Cavalieri D. 2008. Cascading transcriptional effects of a naturally occurring frameshift mutation in *Saccharomyces cerevisiae*. *Mol Ecol* **17**(12): 2985-2997.

Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, Batzoglou S. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome research* **13**(4): 721-731.

Burke DJ, Dawson D, Stearns T. 2000. *Methods in Yeast Genetics*. Cold Spring Harbor Laboratory Press.

Casalone E, Colella CM, Ricci F, Polsinelli M. 1989. Isolation and characterization of Saccharomyces cerevisiae mutants resistant to sulphite. *Yeast* **5 Spec No**: S287-291.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40**(Database issue): D700-705.

Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**(10): e46688.

Clemons KV, Park P, McCusker JH, McCullough MJ, Davis RW, Stevens DA. 1997. Application of DNA typing methods and genetic analysis to epidemiology and taxonomy of Saccharomyces isolates. *J Clin Microbiol* **35**(7): 1822-1828.

Cubillos FA, Billi E, Zorgo E, Parts L, Fargier P, Omholt S, Blomberg A, Warringer J, Louis EJ, Liti G. 2011. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular Ecology* **20**(7): 1401-1413.

Daran-Lapujade P, Daran JM, Luttik MA, Almering MJ, Pronk JT, Kotter P. 2009. An atypical PMR2 locus is responsible for hypersensitivity to sodium and lithium cations in the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D. *FEMS yeast research* **9**(5): 789-792.

Demogines A, Smith E, Kruglyak L, Alani E. 2008a. Identification and dissection of a complex DNA repair sensitivity phenotype in Baker's yeast. *PLoS genetics* **4**(7): e1000123.

Demogines A, Wong A, Aquadro C, Alani E. 2008b. Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. *PLoS genetics* **4**(6): e1000103.

Deutschbauer AM, Davis RW. 2005. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat Genet* **37**(12): 1333-1340.
Diezmann S, Dietrich FS. 2011. Oxidative stress survival in a clinical *Saccharomyces cerevisiae* isolate is influenced by a major quantitative trait nucleotide. *Genetics* **188**(3): 709-722.
Dimitrov LN, Brem RB, Kruglyak L, Gottschling DE. 2009. Polymorphisms in multiple genes contribute to the spontaneous mitochondrial genome instability of *Saccharomyces cerevisiae* S288C strains. *Genetics* **183**(1): 365-383.
Ding C, Festa RA, Chen YL, Espart A, Palacios O, Espin J, Capdevila M, Atrian S, Heitman J, Thiele DJ. 2013. *Cryptococcus neoformans* copper detoxification machinery is critical for fungal virulence. *Cell host & microbe* **13**(3): 265-276.
Domergue R, Castano I, De Las Penas A, Zupancic M, Lockatell V, Hebel JR, Johnson D, Cormack BP. 2005. Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. *Science* **308**(5723): 866-870.
Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PLoS genetics* **4**(8): e1000183.
Douglas LJ. 2003. Candida biofilms and their role in infection. *Trends Microbiol* **11**(1): 30-36.
Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, Bernstein DA, Rolfe PA, Heisler LE, Chin B et al. 2010. Genotype to phenotype: a complex problem. *Science* **328**(5977): 469.
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**(19): 2460-2461.
Ekino K, Kwon I, Goto M, Yoshino S, Furukawa K. 1999. Functional analysis of HO gene in delayed homothallism in *Saccharomyces cerevisiae* wy2. *Yeast* **15**(6): 451-458.
Elrod SL, Chen SM, Schwartz K, Shuster EO. 2009. Optimizing sporulation conditions for different Saccharomyces cerevisiae strain backgrounds. *Methods in molecular biology* **557**: 21-26.
Esberg A, Muller LA, McCusker JH. 2011. Genomic structure of and genome-wide recombination in the *Saccharomyces cerevisiae* S288C progenitor isolate EM93. *PLoS One* **6**(9): e25211.
Fan HY, Cheng KK, Klein HL. 1996. Mutations in the RNA polymerase II transcription machinery suppress the hyperrecombination mutant hpr1 delta of *Saccharomyces cerevisiae*. *Genetics* **142**(3): 749-759.
Fleet GH, Heard GM. 1993. Chapter 2: Yeasts – Growth during fermentation. In *Wine Microbiology and Biotechnology*, (ed. GH Fleet), pp. 27 - 54. Harwood Academic Publishers, Canberra, Australia.
Fogel S, Welch JW, Maloney DH. 1988. The molecular genetics of copper resistance in *Saccharomyces cerevisiae*--a paradigm for non-conventional yeasts. *Journal of basic microbiology* **28**(3): 147-160.
Gaisne M, Becam AM, Verdiere J, Herbert CJ. 1999. A 'natural' mutation in *Saccharomyces cerevisiae* strains derived from S288c affects the complex regulatory gene HAP1 (CYP1). *Current genetics* **36**(4): 195-200.
Garrison EM, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*: 1207.3907.
Gerke J, Lorenz K, Cohen B. 2009. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323**(5913): 498-501.
Gimble FS, Thorner J. 1992. Homing of a DNA endonuclease gene by meiotic gene conversion in *Saccharomyces cerevisiae*. *Nature* **357**(6376): 301-306.
Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al. 1996. Life with 6000 genes. *Science* **274**(5287): 563-567.
Goto K, Fukuda H, Kichise K, Kitano K, Hara S. 1991. Cloning and nucleotide sequence of the KHS killer gene of *Saccharomyces cerevisiae*. *Agric Biol Chem* **55**(8): 1953-1958.

Hall C, Dietrich FS. 2007. The reacquisition of biotin prototrophy in Saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* **177**(4): 2293-2307.
Helliwell KE, Wheeler GL, Smith AG. 2013. Widespread decay of vitamin-related pathways: coincidence or consequence? *Trends in genetics : TIG* **29**(8): 469-478.
Henn BM, Gravel S, Moreno-Estrada A, Acevedo-Acevedo S, Bustamante CD. 2010. Fine-scale population structure and the era of next-generation sequencing. *Human molecular genetics* **19**(R2): R221-226.
Hodgkinson V, Petris MJ. 2012. Copper homeostasis at the host-pathogen interface. *The Journal of biological chemistry* **287**(17): 13549-13555.
Ito-Harashima S, Hartzog PE, Sinha H, McCusker JH. 2002. The tRNA-Tyr gene family of *Saccharomyces cerevisiae*. agents of phenotypic variation and position effects on mutation frequency. *Genetics* **161**(4): 1395-1410.
Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**(14): 1801-1806.
Jorgensen P, Nelson B, Robinson MD, Chen Y, Andrews B, Tyers M, Boone C. 2002. High-resolution genetic mapping with ordered arrays of *Saccharomyces cerevisiae* deletion mutants. *Genetics* **162**(3): 1091-1099.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics* **9**(4): 286-298.
Kim HS, Fay JC. 2007. Genetic variation in the cysteine biosynthesis pathway causes sensitivity to pharmacological compounds. *Proc Natl Acad Sci U S A* **104**(49): 19387-19391.
Knight SA, Labbe S, Kwon LF, Kosman DJ, Thiele DJ. 1996. A widespread transposable element masks expression of a yeast copper transport gene. *Genes & development* **10**(15): 1917-1929.
Kwan EX, Foss E, Kruglyak L, Bedalov A. 2011. Natural polymorphism in BUL2 links cellular amino acid availability with chronological aging and telomere maintenance in yeast. *PLoS genetics* **7**(8): e1002250.
Lang GI, Murray AW, Botstein D. 2009. The cost of gene expression underlies a fitness trade-off in yeast. *Proc Natl Acad Sci U S A* **106**(14): 5755-5760.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**(21): 2947-2948.
Lee HN, Magwene PM, Brem RB. 2011. Natural variation in CDC28 underlies morphological phenotypes in an environmental yeast isolate. *Genetics* **188**(3): 723-730.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.
Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**(2): 265-272.
Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**(7236): 337-341.
Liu H, Styles CA, Fink GR. 1996. *Saccharomyces cerevisiae* S288C has a mutation in FLO8, a gene required for filamentous growth. *Genetics* **144**(3): 967-978.
Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome research* **9**(8): 720-731.

Mackie KA, Muller T, Kandeler E. 2012. Remediation of copper in vineyards--a mini review. *Environmental pollution* **167**: 16-26.
Martinez R, Latreille MT, Mirande M. 1991. A PMR2 tandem repeat with a modified C-terminus is located downstream from the KRS1 gene encoding lysyl-tRNA synthetase in *Saccharomyces cerevisiae*. *Molecular & general genetics : MGG* **227**(1): 149-154.
McCluskey K, Wiest A, Plamann M. 2010. The Fungal Genetics Stock Center: a repository for 50 years of fungal genetics research. *Journal of biosciences* **35**(1): 119-126.
McCusker JH, Clemons KV, Stevens DA, Davis RW. 1994. Genetic characterization of pathogenic *Saccharomyces cerevisiae* isolates. *Genetics* **136**: 1261-1269.
Meiron H, Nahon E, Raveh D. 1995. Identification of the heterothallic mutation in HO-endonuclease of *S. cerevisiae* using HO/ho chimeric genes. *Current genetics* **28**(4): 367-373.
Muller LA, McCusker JH. 2009. Microsatellite analysis of genetic diversity among clinical and nonclinical *Saccharomyces cerevisiae* isolates suggests heterozygote advantage in clinical environments. *Mol Ecol* **18**(13): 2779-2786.
-. 2011. Nature and distribution of large sequence polymorphisms in *Saccharomyces cerevisiae*. *FEMS yeast research* **11**(7): 587-594.
Naumov GI. 1986. Genetic differentiation and ecology of the yeast *Saccharomyces paradoxus* Batschinskaia. *Dokl Botan Sciences* **289-291**: 213-216.
-. 1987. Genetic basis for classification and identification of the ascomycetous yeasts. *Studies in Mycology* **30**: 469-475.
-. 1996. Genetic identification of biological species in the Saccharomyces sensu stricto complex. *Journal of Industrial Microbiology* **17**(3-4): 295-302.
Naumov GI, James SA, Naumova ES, Louis EJ, Roberts IN. 2000. Three new species in the *Saccharomyces sensu stricto* complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *Int J Syst Evol Microbiol* **50 Pt 5**: 1931-1942.
Naumov GI, Naumova ES, Masneuf-Pomarede I. 2010. Genetic identification of new biological species *Saccharomyces arboricolus* Wang et Bai. *Antonie Van Leeuwenhoek* **98**(1): 1-7.
Naumov GI, Serpova EV, Naumova ES. 2006. [A genetically isolated population of Saccharomyces cerevisiae in Malaysia]. *Mikrobiologiia* **75**(2): 245-249.
Ness F, Aigle M. 1995. RTM1: a member of a new family of telomeric repeated genes in yeast. *Genetics* **140**(3): 945-956.
Nishant KT, Wei W, Mancera E, Argueso JL, Schlattl A, Delhomme N, Ma X, Bustamante CD, Korbel JO, Gu Z et al. 2010. The baker's yeast diploid genome is remarkably stable in vegetative growth and meiosis. *PLoS genetics* **6**(9): e1001109.
Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**(8): 2444-2448.
Perez-Ortin JE, Querol A, Puig S, Barrio E. 2002. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome research* **12**(10): 1533-1539.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**(3): 559-575.
R Core Team. 2014. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*.
Ramage G, Mowat E, Jones B, Williams C, Lopez-Ribot J. 2009. Our current understanding of fungal biofilms. *Crit Rev Microbiol* **35**(4): 340-355.
Reynolds TB, Fink GR. 2001. Bakers' yeast, a model for fungal biofilm formation. *Science* **291**(5505): 878-881.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG* **16**(6): 276-277.
RM11. 2004. *Saccharomyces cerevisiae* RM11-1a Sequencing Project. Broad Institute of Harvard and MIT

http://www.broadinstitute.org/annotation/genome/saccharomyces_cerevisiae/Home.html.
Romano P, Suzzi G. 1993. Chapter 13: Sulfur dioxide and wine microorganisms. In *Wine Microbiology and Biotechnology*, (ed. GH Fleet), pp. 373 - 393. Harwood Academic Publishers, Canberra, Australia.
Ronald J, Brem RB, Whittle J, Kruglyak L. 2005. Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS genetics* **1**(2): e25.
Samanovic MI, Ding C, Thiele DJ, Darwin KH. 2012. Copper in microbial pathogenesis: meddling with the metal. *Cell host & microbe* **11**(2): 106-115.
Sandhu DK, Sandhu RS, Khan ZU, Damodaran VN. 1976. Conditional virulence of a p-aminobenzoic acid-requiring mutant of *Aspergillus fumigatus*. *Infec Immun* **13**: 527-532.
Sherman F, Helms C. 1978. A chromosomal translocation causing overproduction of iso-2-cytochrome c in yeast. *Genetics* **88**(4 Pt 1): 689-707.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6): 1117-1123.
Sinha H, David L, Pascon RC, Clauder-Munster S, Krishnakumar S, Nguyen M, Shi G, Dean J, Davis RW, Oefner PJ et al. 2008. Sequential elimination of major-effect contributors identifies additional quantitative trait loci conditioning high-temperature growth in yeast. *Genetics* **180**(3): 1661-1670.
Sinha H, Nicholson BP, Steinmetz LM, McCusker JH. 2006. Complex genetic interactions in a quantitative trait locus. *PLoS genetics* **2**(2): e13.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *Journal of molecular biology* **147**(1): 195-197.
Torabi N, Kruglyak L. 2011. Variants in SUP45 and TRM10 underlie natural variation in translation termination efficiency in *Saccharomyces cerevisiae*. *PLoS genetics* **7**(7): e1002211.
Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat Genet*.
Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**(11): e112963.
Warringer J, Zorgo E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ et al. 2011. Trait variation in yeast is defined by population history. *PLoS genetics* **7**(6): e1002111.
Wei W, McCusker JH, Hyman RW, Jones T, Ning Y, Cao Z, Gu Z, Bruno D, Miranda M, Nguyen M et al. 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc Natl Acad Sci U S A* **104**(31): 12825-12830.
Wieland J, Nitsche AM, Strayle J, Steiner H, Rudolph HK. 1995. The PMR2 gene cluster encodes functionally distinct isoforms of a putative Na+ pump in the yeast plasma membrane. *The EMBO journal* **14**(16): 3870-3882.
Will JL, Kim HS, Clarke J, Painter JC, Fay JC, Gasch AP. 2010. Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations. *PLoS genetics* **6**(4): e1000893.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**(6634): 708-713.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**(5): 821-829.

Zhao Y, Strope PK, Kozmin SG, McCusker JH, Dietrich FS, Kokoska RJ, Petes TD. 2014. Structures of Naturally Evolved CUP1 Tandem Arrays in Yeast Indicate That These Arrays Are Generated by Unequal Nonhomologous Recombination. *G3* **4**(11): 2259-2269.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**(7): 821-824.

Zimmer A, Durand C, Loira N, Durrens P, Sherman DJ, Marullo P. 2014. QTL Dissection of Lag Phase in Wine Fermentation Reveals a New Translocation Responsible for Saccharomyces cerevisiae Adaptation to Sulfite. *PLoS One* **9**(1): e86298.

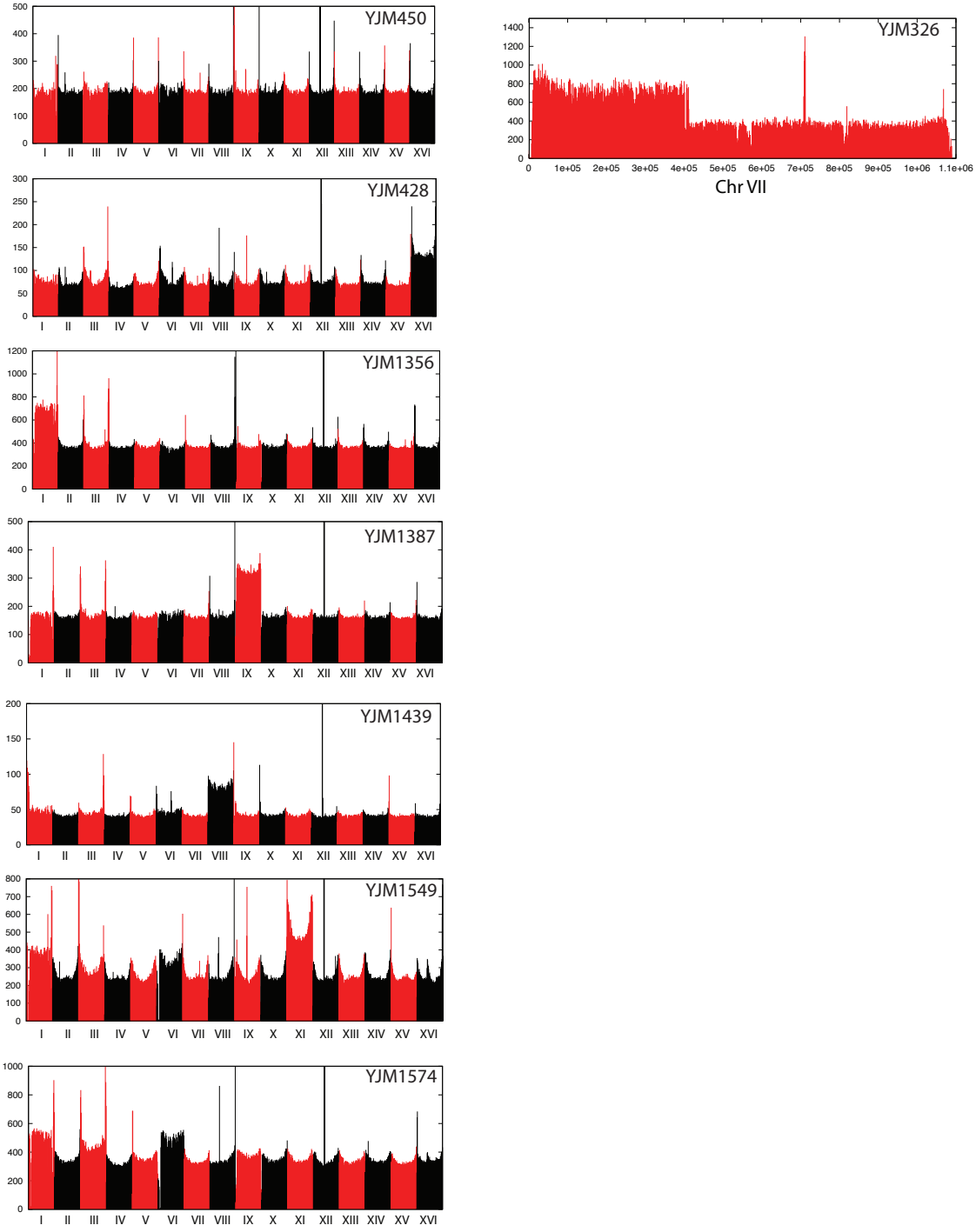


Figure S1: Strains with whole chromosome aneuploidies and a large segmental duplication. Based on sequence coverage, YJM450 is euploid and is shown here for comparison. Based on sequence coverage, YJM428, YJM1356, YJM1387, and YJM1439 are each aneuploid (2N+2) for one chromosome, while YJM1549 is 2N+2+1+1 and YJM1574 is 2N+1+1. YJM326 has a large segmental duplication of the left arm of Chromosome 7 (approximately 1 – 411,000).

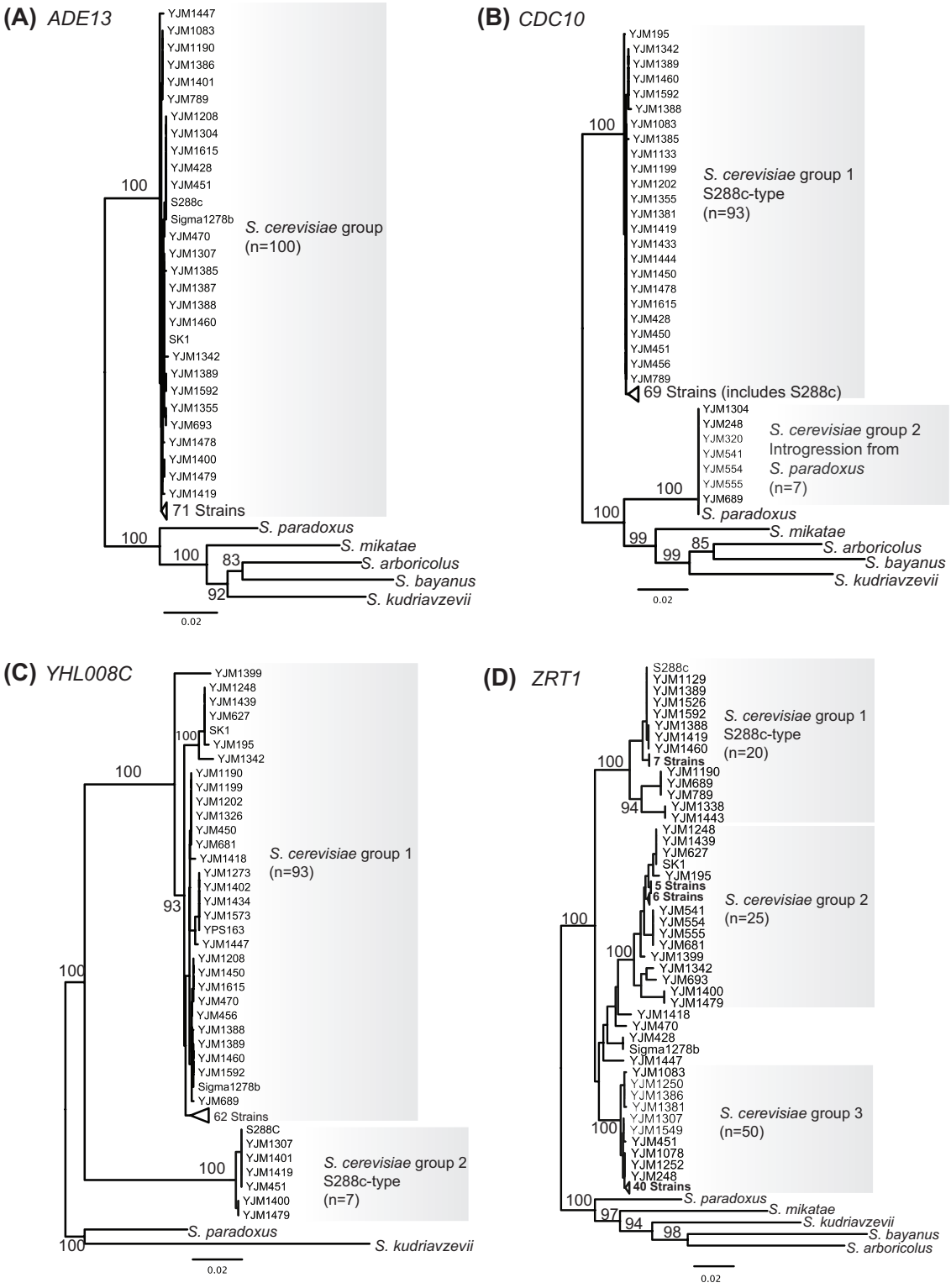

(A) ADE13
YJM1447
YJM1083
YJM1190
YJM1386
YJM1401
YJM789
YJM1208
YJM1304
YJM1615
YJM428
YJM451
S288c
Sigma1278b
YJM470
YJM1307
YJM1385
YJM1387
YJM1388
YJM1460
SK1
YJM1342
YJM1389
YJM1592
YJM1355
YJM693
YJM1478
YJM1400
YJM1479
YJM1419
71 Strains
S. cerevisiae group
(n=100)
S. paradoxus
S. mikatae
S. arboricolus
S. bayanus
S. kudriavzevii
100
100
100
83
92
0.02
(B) CDC10
YJM195
YJM1342
YJM1389
YJM1460
YJM1592
YJM1388
YJM1083
YJM1385
YJM1133
YJM1199
YJM1202
YJM1355
YJM1381
YJM1419
YJM1433
YJM1444
YJM1450
YJM1478
YJM1615
YJM428
YJM450
YJM451
YJM456
YJM789
69 Strains (includes S288c)
S. cerevisiae group 1
S288c-type
(n=93)
YJM1304
YJM248
YJM320
YJM541
YJM554
YJM555
YJM689
S. cerevisiae group 2
Introgression from
S. paradoxus
(n=7)
S. paradoxus
S. mikatae
S. arboricolus
S. bayanus
S. kudriavzevii
100
100
99
85
99
0.02
(C) YHL008C
YJM1399
YJM1248
YJM1439
YJM627
SK1
YJM195
YJM1342
YJM1190
YJM1199
YJM1202
YJM1326
YJM450
YJM681
YJM1418
YJM1273
YJM1402
YJM1434
YJM1573
YPS163
YJM1447
YJM1208
YJM1450
YJM1615
YJM470
YJM456
YJM1388
YJM1389
YJM1460
YJM1592
Sigma1278b
YJM689
62 Strains
S. cerevisiae group 1
(n=93)
S288C
YJM1307
YJM1401
YJM1419
YJM451
YJM1400
YJM1479
S. cerevisiae group 2
S288c-type
(n=7)
S. paradoxus
S. kudriavzevii
100
100
93
100
100
100
0.02
(D) ZRT1
S288c
YJM1129
YJM1389
YJM1526
YJM1592
YJM1388
YJM1419
YJM1460
7 Strains
YJM1190
YJM689
YJM789
YJM1338
YJM1443
S. cerevisiae group 1
S288c-type
(n=20)
YJM1248
YJM1439
YJM627
SK1
YJM195
5 Strains
6 Strains
YJM541
YJM554
YJM555
YJM681
YJM1399
YJM1342
YJM693
YJM1400
YJM1479
S. cerevisiae group 2
(n=25)
YJM1418
YJM470
YJM428
Sigma1278b
YJM1447
YJM1083
YJM1250
YJM1386
YJM1381
YJM1307
YJM1549
YJM451
YJM1078
YJM1252
YJM248
40 Strains
S. cerevisiae group 3
(n=50)
S. paradoxus
S. mikatae
S. kudriavzevii
S. bayanus
S. arboricolus
100
100
94
100
100
97
94
98
0.02

Figure S2: Neighbor-joining trees of different ORFs from the 100-genomes *S. cerevisiae* strains and sibling species. (A) Phylogeny of *ADE13* ORF. All *S. cerevisiae* sequences are highly similar and cluster together. (B) Phylogeny of *CDC10* ORF. Sequences cluster into two groups. Group 1 is S288c type and has a majority of strains. Group 2 has 7 strains and clusters closely with *S. paradoxus* sequence indicating an introgression of this sequence from *S. paradoxus* to strains of this group. (C) Phylogeny of *YHL008C* ORF. Sequences cluster into two groups. Group 1 has a majority of strains. Group 2 has 7 strains including S288c. In this case there is no known source of introgression. (D) Phylogeny of *ZRT1* ORF. Sequences cluster into three groups with 100% bootstrap support, except 5 strains (YJM1418, YJM470, YJM428, YJM1447, and Sigma1278b) that do not cluster with any of the three groups with strong support. From this phylogeny, it seems like there has been putative introgression in 80 of the 100 strains. The five strains that do not cluster are recombinant types of the three groups.

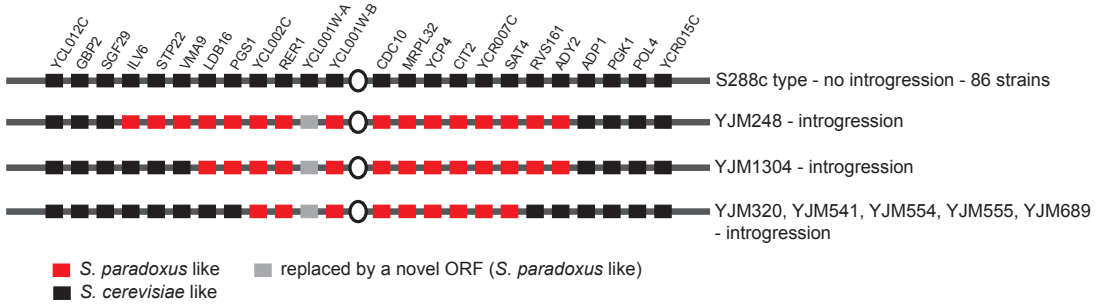


Figure S3: Structure of an introgressed cluster of *S. paradoxus* genes in seven *S. cerevisiae* strains.
The structure of the chromosome 3 introgression in these seven strains is consistent with a single introgression event the size of which was reduced by recombination.

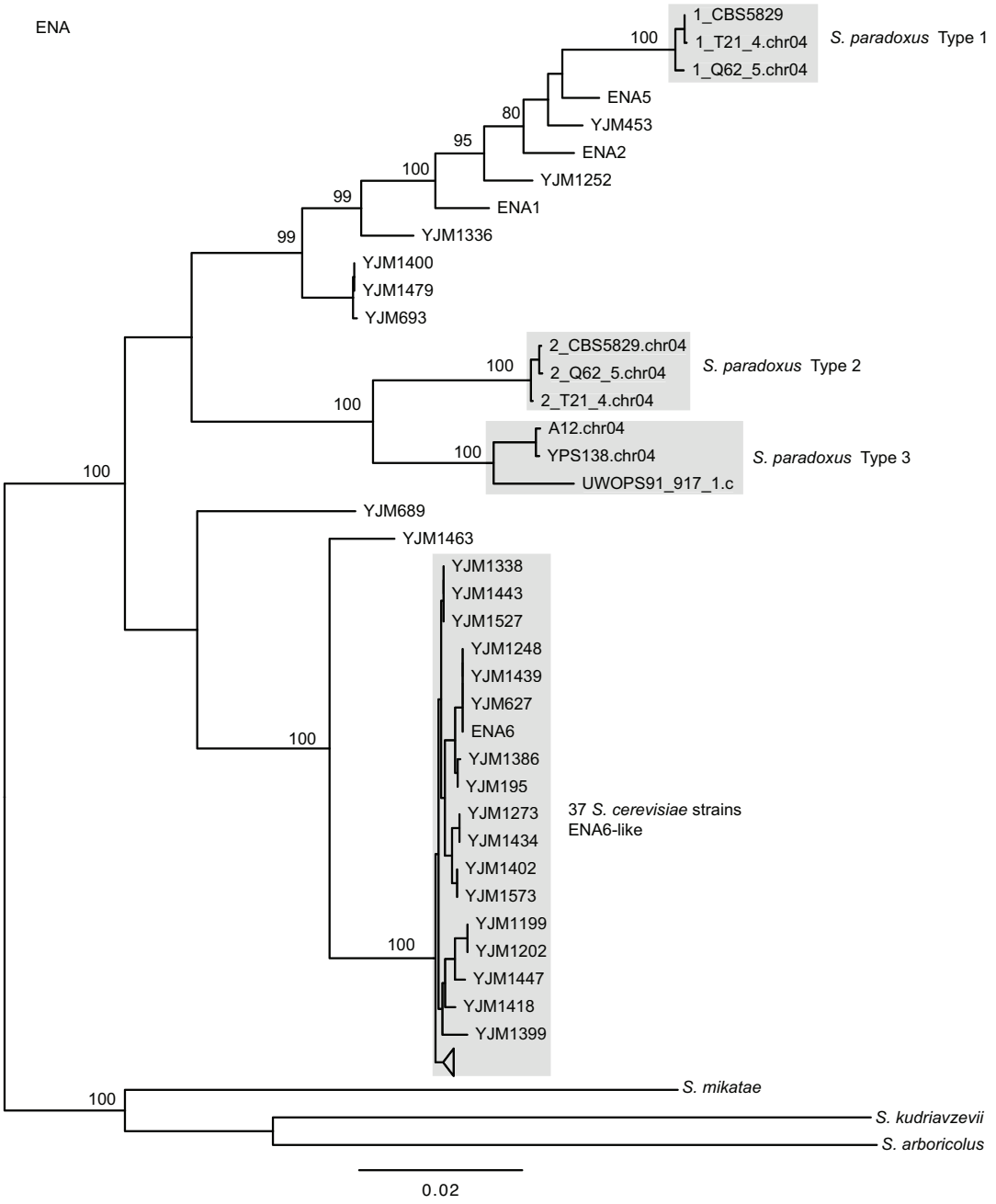


Figure S4: Neighbor-joining tree of *ENA* ORF from *S. cerevisiae* strains, 6 *S. paradoxus* strains, and sibling species with bootstrap values. Only single copy *ENA* sequences from *S. cerevisiae* and one or two copies (where found) of *ENA* sequences from *S. paradoxus* were used in tree construction. The *S. paradoxus* sequences form three clusters. The three clusters of *S. paradoxus ENA* in the tree are also seen when the *ENA* of 36 *S. paradoxus* (Liti et al. 2009) strains are used (data not shown). One *S. cerevisiae* ENA cluster includes *ENA6*. The other *S. cerevisiae ENA* sequences do not form any tight cluster but are consistent with recombination between *S. paradoxus* and *S. cerevisiae ENA*.

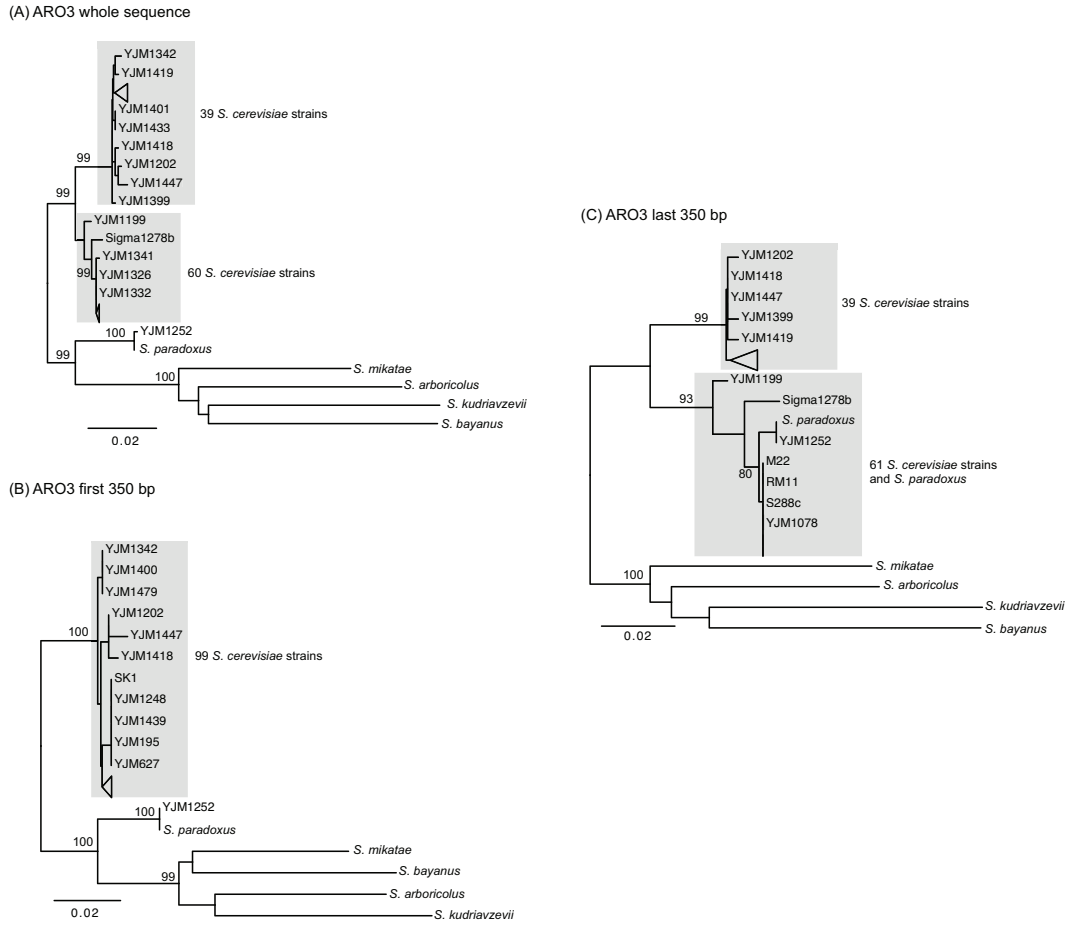
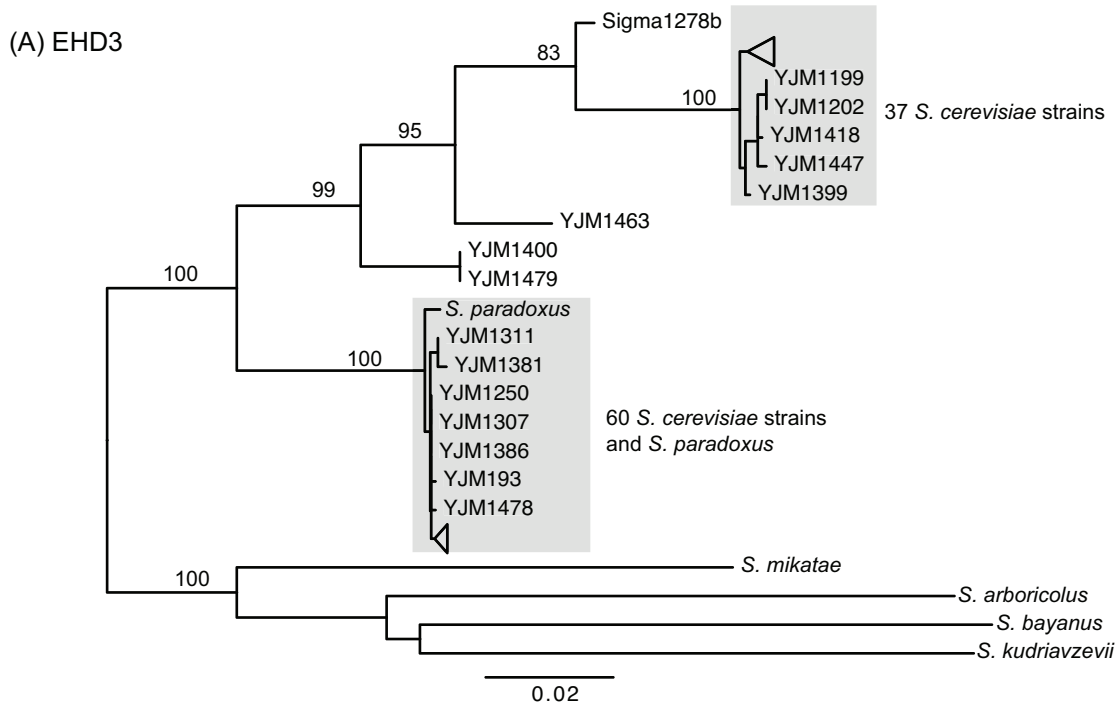


Figure S5: Neighbor-joining tree of *ARO3* ORF from the 100-genomes *S. cerevisiae* strains and sibling species with bootstrap values. Two clusters of *S. cerevisiae ARO3* are seen when the entire open reading frame is used to construct the tree. One strain (YJM1252), which clusters outside of the *S. cerevisiae* clades, clusters instead with *S. paradoxus*. When only the first 350 bp of the ORF is used to build the tree, the two *S. cerevisiae* clusters disappear. When the last 350 bp of the ORF is used to build the tree, the two *S. cerevisiae* clusters are farther apart, with one cluster also containing *S. paradoxus*. This suggests the presence of recombinant (*S. cerevisiae*×*S. paradoxus*) *ARO3* in 60 *S. cerevisiae* strains, with YJM1252 having a complete introgression of the *ARO3* ORF.


(A) EHD3
Sigma1278b
83
100
YJM1199
YJM1202
YJM1418
YJM1447
YJM1399
37 S. cerevisiae strains
95
99
YJM1463
YJM1400
YJM1479
100
100
S. paradoxus
YJM1311
YJM1381
YJM1250
YJM1307
YJM1386
YJM193
YJM1478
60 S. cerevisiae strains
and S. paradoxus
S. mikatae
100
S. arboricolus
S. bayanus
S. kudriavzevii
0.02

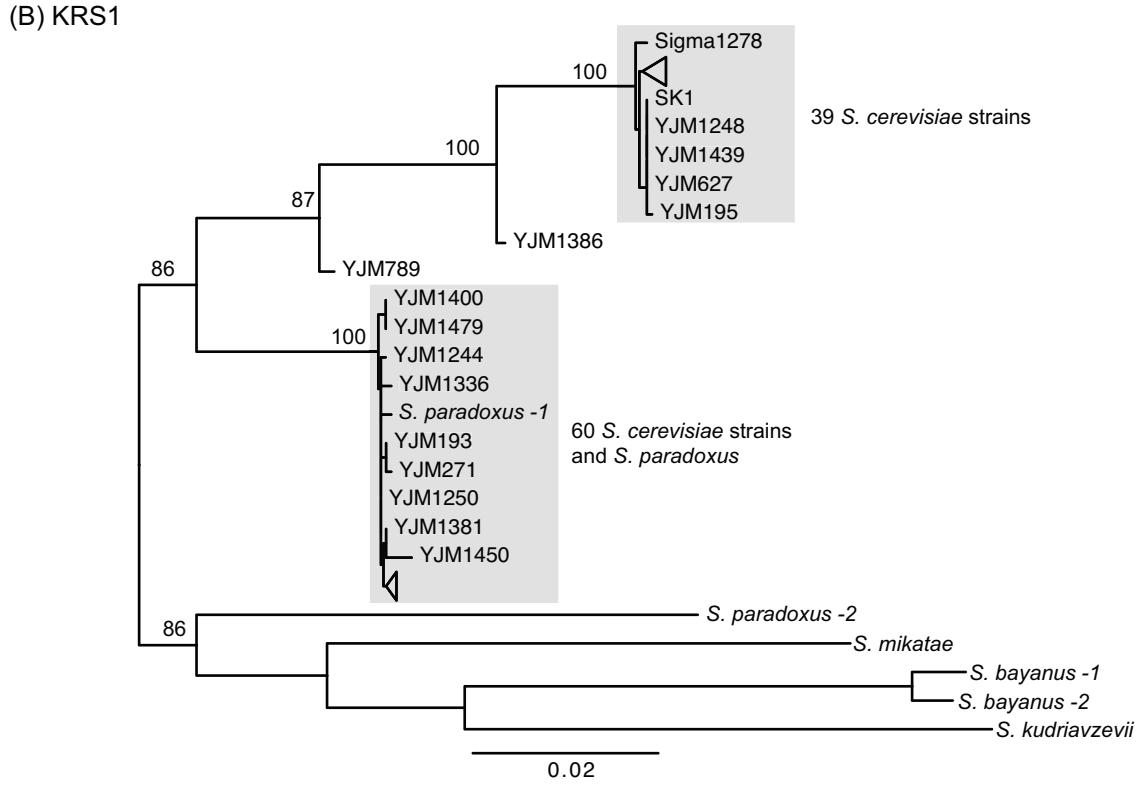

(B) KRS1
Sigma1278
100
SK1
YJM1248
YJM1439
YJM627
YJM195
39 S. cerevisiae strains
100
87
YJM1386
86
YJM789
YJM1400
YJM1479
100
YJM1244
YJM1336
S. paradoxus -1
YJM193
YJM271
YJM1250
YJM1381
YJM1450
60 S. cerevisiae strains
and S. paradoxus
S. paradoxus -2
86
S. mikatae
S. bayanus -1
S. bayanus -2
S. kudriavzevii
0.02

Figure S6: Neighbor-joining tree of *EHD3* (A) and *KRS1* (B) ORFs from the 100-genomes *S. cerevisiae* strains and sibling species with bootstrap values. In both trees, the *S. cerevisiae* sequences fall into two clusters, with a few strains not falling into either one (recombinant types). Sixty *S. cerevisiae* strains cluster with *S. paradoxus* in each of the trees consistent with introgression.

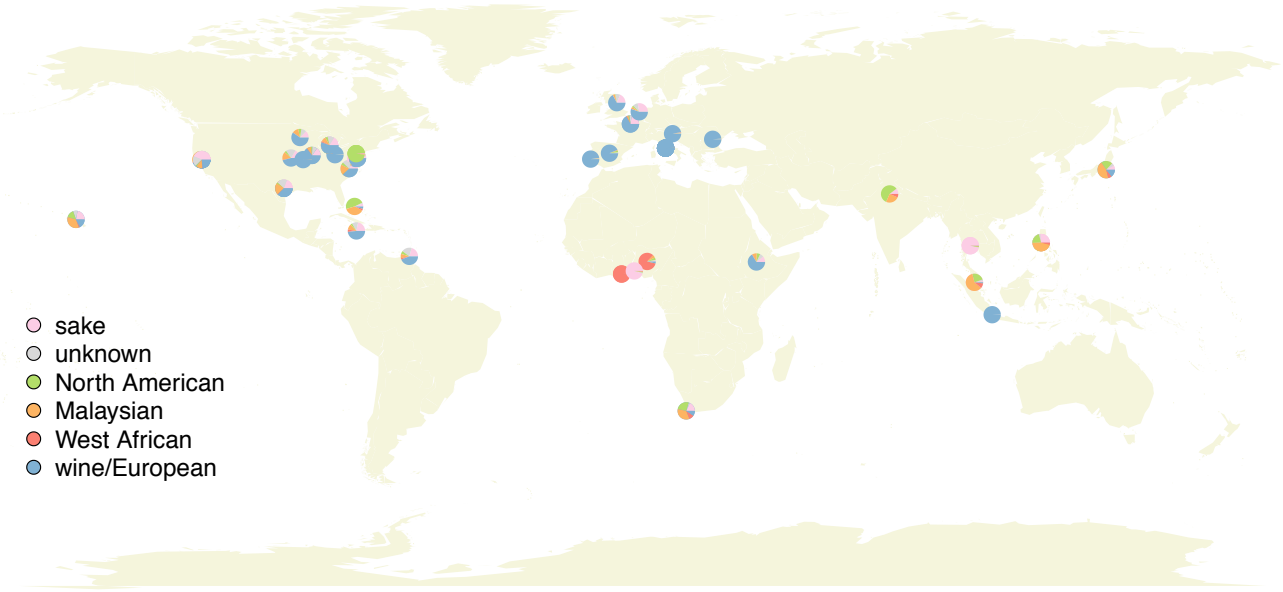


Figure S7: Geographic distribution of the different populations of *S. cerevisiae* strains. World map showing locations, where known, of the *S. cerevisiae* strains sequenced in this study. Each strain is represented by a single glyph divided and colored according to fractional ancestry of the strain as shown in Figure 3. Placement of strains on map is approximate; for some strains only broad-scale geographic information on collection location is available, and in instances where multiple strains were collected near the same location, glyphs have been spread out to facilitate viewing of ancestry.

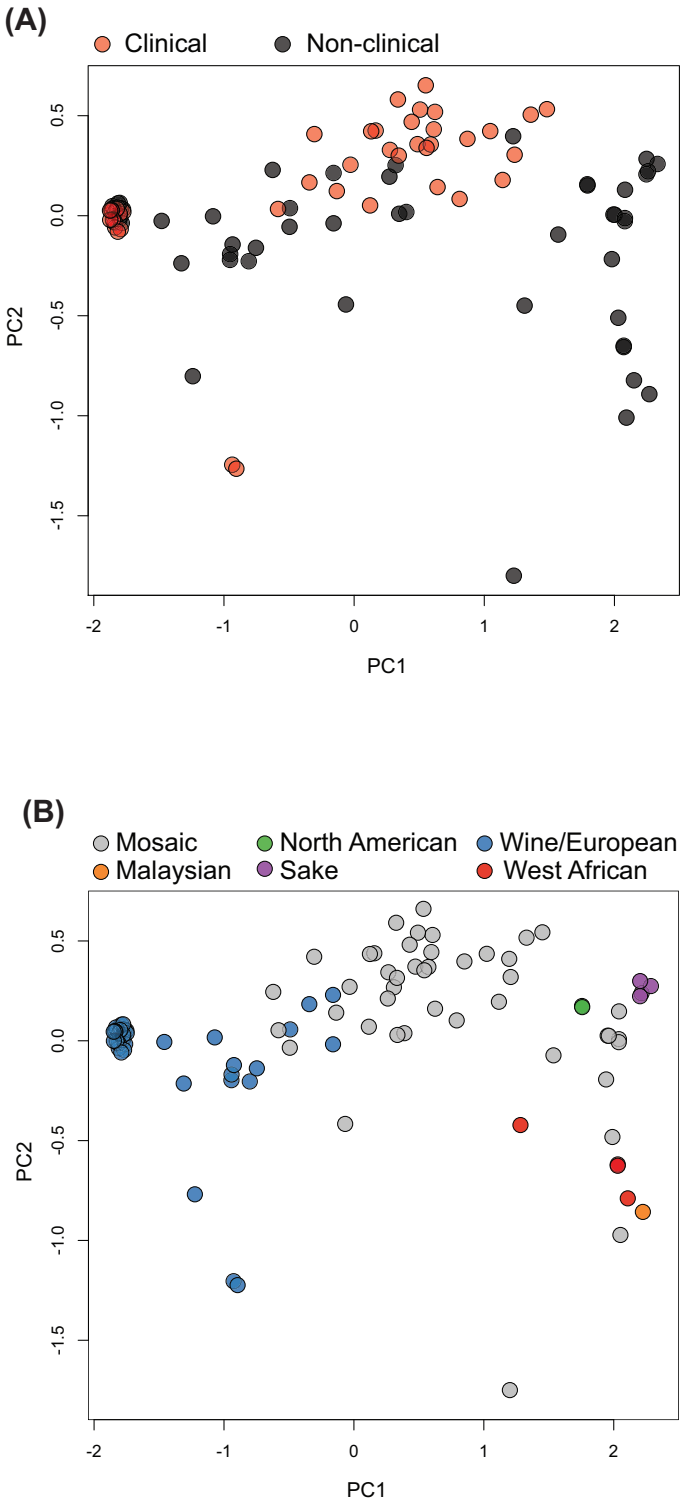


Figure S8: Principal component analysis of the 93 sequenced strains plus S288c. PCA plots of sequence identity between 94 strains (S288c and 93 strains) using 218 kb (total) of sequence (excluding introgressions and large indels) gathered from all 16 chromosomes: (A) Clinical vs. non-clinical strains; (B) The six populations of the 94 strains.

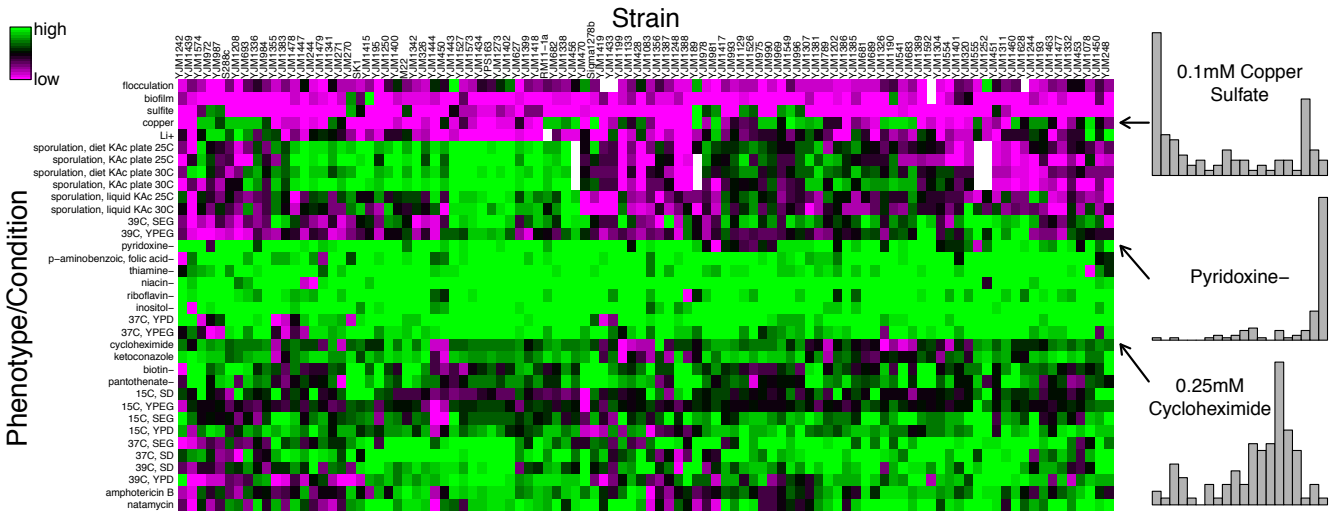
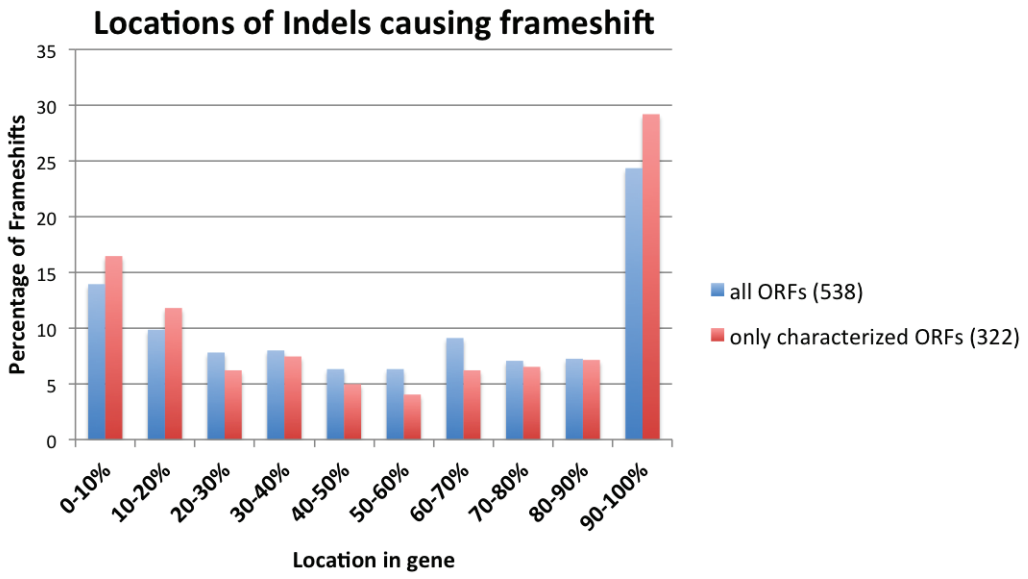


Figure S9: Heatmap illustrating phenotypic variability in the 100-genomes strains. Each row represents a phenotype measured in all or nearly all of the 100 strains. Phenotype measurements are comprised of ratios of growth under treatment to control conditions (for growth phenotypes) or are shown in appropriate measurement units for the phenotype (e.g. percent sporulated for sporulation phenotypes). Phenotypes are scaled to have mean 0 and variance 1 to facilitate comparison. The color bar (upper left) illustrates the color scheme distinguishing low from high phenotypic values; white boxes indicate missing data. On right, histograms for individual phenotypes illustrate examples of phenotypic distributions for three phenotypes.

(A)



(B)

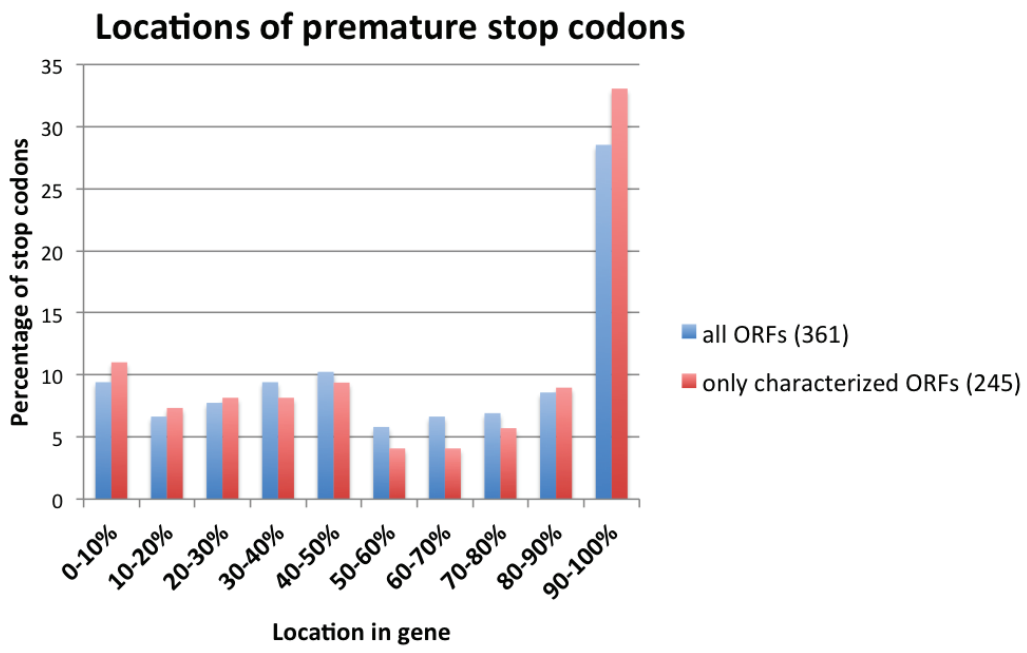


Figure S10: Locations of genes with potentially inactivating frameshift and premature stop codons. Genes analyzed (total 4522) do not include duplicated genes and genes with introns. (A) Locations of frameshifts. (B) Locations of premature stop codons.

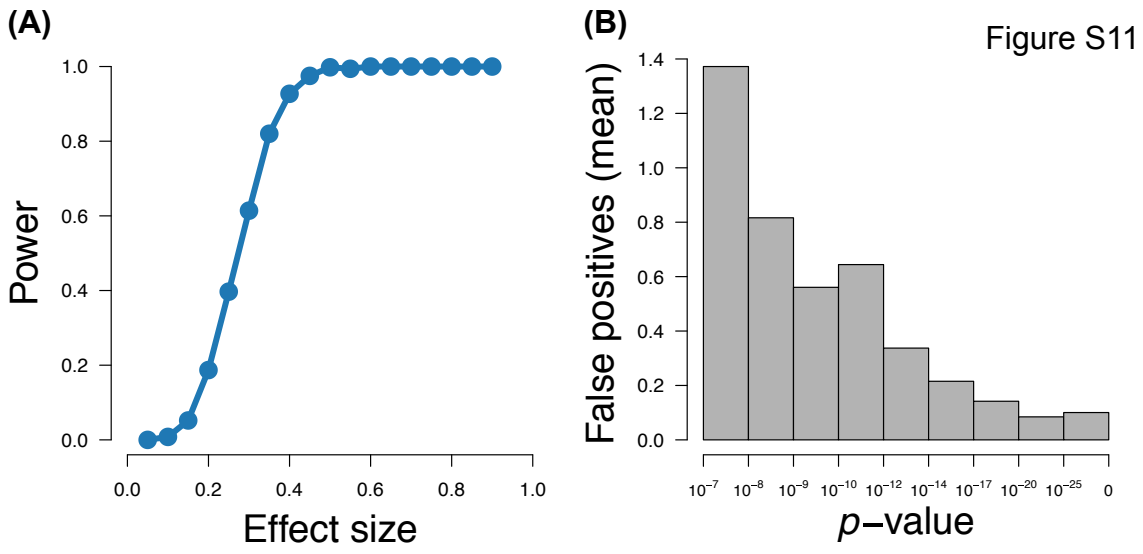


Figure S11: Power and false positive rate for simulated tests of genotype-phenotype association.
Both plots use a threshold of $p < 1\times10^{-7}$ as significance level. (A) Power as a function of causal SNP effect size (proportion of phenotypic variation explained by allele at causal SNP). Power is defined as the fraction of 1,000 simulations in which the causal SNP was called as significant ($p < 1\times10^{-7}$). (B) *P*-values of false positive SNPs. A false positive is defined as a SNP with $p < 1\times10^{-7}$ that is not within 25 kb of the causal SNP. Each bar shows the mean number of false positives across simulations with *p*-value between the values indicated by the axis ticks on each side of the bar. Mean number of false positives is calculated across 18,000 simulations (1,000 simulations for each effect size from 5% to 90%).

Supplemental Table Legends (Table S1 – S19)

Table S1. The 100-genomes strains with their parental isolate name, geographic and environmental origin, as well as population; other relevant strains are also included.

All strains listed in column 1 have been deposited in and should be requested from the Fungal Genetics Stock Center (FGSC; http://www.fgsc.net)

SUH = Stanford University Hospital

CBS, NCYC, NRRL, and Phaff strains and isolates should be requested from the CBS (Centraalbureau voor Schimmelcultures; http://www.cbs.knaw.nl/), NCYC (National Collection of Yeast Cultures; http://www.ncyc.co.uk/), NRRL (Northern Regional Research Laboratory, now the National Center For Agricultural Utilization Research; http://nrrl.ncaur.usda.gov/), and Phaff Yeast Culture (http://phaffcollection.ucdavis.edu) collections, respectively.

Most of the YJM isolates listed in column 2 have been deposited into and should be requested from the Phaff Yeast Culture (http://phaffcollection.ucdavis.edu) collection.

Table S2: Error summary of the 93 assembled *S. cerevisiae* genomes and N50 of AbySS assembled scaffolds.

Under sequence, the version number reflects the in house version number of the assembled sequence.

N's denote either a single base uncertainty in the sequence, or a longer string of N's indicating a gap in the assembly or an ambiguity in the assembly.

Ambiguity codons refers to the use of the standard ambiguity codons K (G or T) M (A or C) R (A or G) Y (C or T) S (C or G) W (A or T) B (C or G or T) V (A or C or G) H (A or C or T) D (A or G or T). In most cases, the ambiguity codons are found in duplicate, near identical sequences in the assembly. No N's or ambiguity codons are present in the mitochondrial genome sequences.

Incomplete telomeres: The number of chromosomes ends, out of 32 possible per strain, where the assembly does not extend into the telomerase-generated terminal telomeric repeat.

Fraction incomplete telomeres: The fraction of chromosomes where assembly did not extend to telomerase-generated sequence.

Errors found by pilon: The number of errors the pilon program identified, excluding those at the incomplete telomere ends, and at N's and ambiguity bases, to prevent double counting of these assembly problems. Multiple errors reported by pilon within a 100 base window are counted as only a single error.

Fraction protein associated errors: The fraction of pilon detected errors within or within ~50 bases of a protein coding gene, and ranges from .16 to .58. These errors are found in a total of 159 different protein-coding genes. No errors were detected by pilon in the remaining 5600 proteins across all 93 genomes. These nuclear genomes are approximately 13,000,000 bases, and 67% protein coding. No errors were detected in the mitochondrial sequence. In this analysis, the *COS*, *PAU, YRF*, telomeric Y' helicases, and Ty encoded proteins were excluded.

Total identified errors: The sum of the N's, ambiguity codons, missing telomeres, and pilon detected errors.

Sequence coverage is based on counting the number of Illumina sequence reads that align to the 5189 bp (total) of the unique nuclear protein-coding regions of *URA3* (803 bp), *LEU2* (1094 bp), *HIS4* (2399 bp), and *HIS1* (893 bp) using BWA.

Table S3. PCR primers for RFLP strain confirmations. Nineteen primer pairs, in combination with restriction digestion, were identified to distinguish the strains in this study; (0) = PCR product is not cut and (1) = PCR product is cut by the designated enzyme.

Table S4. Spore viability of the 100 strains when crossed with the S288c background strain YJM1617, and the PCR test result for the *ECM34-SSU1* translocation in the 100 strains.

Table S5. List of genes that are introgressed, or putatively introgressed, deleted, and novel in one

or more of the 93 strains.

(A) Introgressed genes relative to S288c. This list does not include subtelomeric and duplicated genes. Gene notations: (1) = S288c-like, not introgressed; (0) = deleted; (P) = likely introgression from *S. paradoxus* (sequence identity ≥ 96%). (s) = closest sequence identity (< 96%) is to *S. paradoxus, cerevisiae, kudriavzevii, mikatae, bayanus,* or *arboricolus*; possible introgressions from unknown or unsequenced species of *Saccharomyces* most closely related to these species.

(B) Deleted genes relative to S288c. This list does not include subtelomeric and duplicated genes. (1) denotes the presence of the gene, (0) denotes the absence. In addition, those genes that are detected as introgressed are denoted in the same way as (A).

(C) Novel putative genes relative to S288c. These sequences are not present or not annotated in S288c. Also included are some genes previously known to be absent in S288c that are found in one or more of the 93 strains.

Table S6. GO results for introgressed genes, deleted genes and genes with frameshifts and premature stop codons.

Table S7. Results of tests for non-independence between genetic variants (introgressed loci, frameshift/stop polymorphisms, novel genes, deleted genes) and population membership or clinical status.

Table S8. Approximate numbers of SNPs and Indels in the 93 genomes relative to s288c.

Table S9. Copy numbers, based on coverage, of Ty1-Ty5 in the 93 strains and S288c.

Table S10. Locations of LTRs in the 93 strains. (A) LTRs that are present in S288c and one or more of the 93 strains. (B) LTRs that are not present in S288c but present in one or more of the

93 strains. (C) Total numbers of positions of 4 types of LTRs that are present in the 93 strains and S288c.

Table S11. List of genes with LTR insertions, the strain names in which they appear, and the location of insertion.

Table S12. Genes with ORF length polymorphisms, relative to the reference S288c. Genes (excluding sub-telomeric, duplicated, and intron-containing genes) with ORF length polymorphisms, relative to the reference S288c, due to frameshift indels (FS) and premature stop codons (Stop); (0) denotes those genes with no FS or premature stop.

Table S13. The presence/absence in the 93 strains of polymorphisms previously described to be phenotypically relevant.

Polymorphisms previously described to be phenotypically relevant and their presence/absence in the 93 strains. Also shown are other likely inactivating polymorphisms in these genes. The strain name of the reference sequence used for the analysis is given in parentheses next to the gene name on the first row, followed by the type of polymorphism surveyed. Previously described polymorphisms: *MKT1* (Deutschbauer and Davis 2005; Sinha et al. 2006; Demogines et al. 2008a; Dimitrov et al. 2009); *END3* (Sinha et al. 2006); *NCS2* (Sinha et al. 2008); *MIP1* (Baruffini et al. 2007; Dimitrov et al. 2009); *RAD5* (Fan et al. 1996; Demogines et al. 2008a); *AMN1* (Ronald et al. 2005); *FLO8* (Liu et al. 1996); *CYS4* (Kim and Fay 2007); *BUL2* (Kwan et al. 2011); *RDS2* (Diezmann and Dietrich 2011); *IME1* (Gerke et al. 2009); *RME1* (Deutschbauer and Davis 2005; Gerke et al. 2009); *RSF1* (Gerke et al. 2009); *TAO3* (Deutschbauer and Davis 2005); *RAS2* (Ben-Ari et al. 2006); *TRM10* (Torabi and Kruglyak 2011); *CDC28* (Lee et al. 2011); *SSY1* (Brown et al. 2008); *GPA1* (Lang et al. 2009); *SAL1* (Dimitrov et al. 2009); *CAT5*

(Dimitrov et al. 2009); *GAL3* (Warringer et al. 2011); SSD1 (Jorgensen et al. 2002); *HO* (Meiron et al. 1995; Ekino et al. 1999); *PMS1* (Demogines et al. 2008b; Argueso et al. 2009); *MLH1* (Demogines et al. 2008b; Argueso et al. 2009); *HAP1* (Gaisne et al. 1999); *CTR3* (Knight et al. 1996); *PDR5* (Wei et al. 2007); *VMA1/VDE* (Gimble and Thorner 1992); *AQY1* (Will et al. 2010); *AQY2* (Will et al. 2010).

Table S14. *ENA* copy numbers and flanking gene types in the 93 strains. *ENA* copy numbers were determined by *ENA* sequence coverage analysis.

Table S15. *CUP1* copy numbers and the five types of junction sequences of the *CUP1* repeats in the 93 strains. The *CUP1* copy numbers for the 93 sequenced strains from this work were determined by the *CUP1* sequence coverage analysis. For YJM1077 (SK1) and YJM1281 (YPS163), *CUP1* copy number was determined by PCR. For YJM789 (isogenic with YJM145) and S288C, *CUP1* copy number was from (Fogel et al. 1988; Zhao et al. 2014).

Table S16. The numbers of SNPs and indels within the 218 kb region of the 93 strains as compared to S288C.

Table S17. 49 phenotypes measured in the 100 genomes strains (see Supplemental Material for phenotyping methods).

Table S18. Summary and complete results of genotype-phenotype association mapping.

Table S19. List of the 93 strains sequenced in this study and their accession numbers for each chromosome in GenBank and for each strain in SRA.