We sequenced the genome of Saccharomyces cerevisiae strain YJM789, which was derived from a yeast isolated from the lung of an AIDS patient with pneumonia. The strain is used for studies of fungal infections and quantitative genetics because of its extensive phenotypic differences to the laboratory reference strain, including growth at high temperature and deadly virulence in mouse models. Here we show that the ~12-Mb genome of YJM789 contains ~60,000 SNPs and ~6,000 indels with respect to the reference S288c genome, leading to protein polymorphisms with a few known cases of phenotypic changes. Several ORFs are found to be unique to YJM789, some of which might have been acquired through horizontal transfer. Localized regions of high polymorphism density are scattered over the genome, in some cases spanning multiple ORFs and in others concentrated within single genes. The sequence of YJM789 contains clues to pathogenicity and spurs the development of more powerful approaches to dissecting the genetic basis of complex hereditary traits.

Results and Discussion

Genome and Comparison. We sequenced the genome of strain YJM789 by using a shotgun approach, generating >170,000 sequence reads, followed by finishing to close gaps of the nonrepetitive portions of the genome, which yielded an additional ~4,000 reads. After assembly, 11.8 Mb of high-quality genome sequence were obtained. The coverage corresponds to 98% of the S288c genome as determined from chromosome-by-chromosome alignments of the two genome sequences [see Fig. 1 for chromosome XIV and supporting information (SI) Fig. 5 for the entire genome]. The 16 YJM789 nuclear chromosomes are covered by 31 contigs (see SI Tables 1 and 2) and the mitochondrial genome (mtDNA) by a single contig.

ORFs and Horizontal Transfer. Employing three methods, we predicted 5,904 ORFs in the nuclear genome of YJM789, of which 5,509 have a reciprocal-best-hit ortholog in S288c (see

© 2007 by The National Academy of Sciences of the USA

PNAS | July 31, 2007 | vol. 104 | no. 31 | 12825–12830
repeats may be responsible for the absence of *faecium* N-acetyltransferases (GNAT) from different bacteria, with the presence of these two genes in YJM789 was confirmed by PCR analysis of genomic DNA from YJM789 and S288c corroborated the inversion. Analogsous analysis of the vineyard isolate RM11-1a (35) and a sequence comparison to *Saccharomyces paradoxus*, the closest species to *S. cerevisiae* that has its genome sequenced (36), shows that, in both YJM789 and RM11-1a, this region is inverted relative to S288c and *S. paradoxus*. In addition, a translocation was detected between chromosomes VI and X, wherein an element of 18 kb on chromosome VI in S288c (base pairs 11,626–30,088) is found on chromosome X in YJM789 (base pairs 1–18,478 on contig 100). This translocation was confirmed independently by PCR amplification across the breakpoints and by sequencing the ends of the amplicons.

Highly Polymorphic Chromosomal Regions. Within the aligned regions of the YJM789 and S288c genomes, we identified 59,361 high-confidence SNPs that are scattered throughout the genome (SI Table 3 and SI Fig. 7 present the SNP distribution for each individual chromosome). SNP density is 6.1 per kilobase on average but is far from constant across the genome and across individual chromosomess, with chromosome I having the highest average SNP density of 19.7 per kilobase. There is a discrete region of 12 kb on chromosome I that is highly polymorphic (Fig. 3). The abrupt transitions from low-to-high and high-to-low SNP density at its boundaries prompted us to analyze this chromosome I region in more detail.

Close examination of the highly polymorphic region on chromosome 1 (Fig. 3A) showed that the 12-kb sequence contains 2,356 SNPs and 187 indels, accounting for >50% of the total chromosome I polymorphisms. The region in S288c encompasses five members of the nonessential DUP240 gene family encoding putative integral membrane proteins (37); *UIP3, YAR028W, YAR029W, PRM9*, and *MST28*. The corresponding region of YJM789 contains the orthologs for *UIP3, YAR028W, PRM9, MST28*, as well as an ORF (yorf4.01.161.113) unique to YJM789. Recurrent deletion and ectopic recombination has been suggested to underlie the diversity of the DUP240 gene family regions among *S. cerevisiae* strains (38).

We compared the sequence of this region from S288c, RM11-1a, YJM789, the sibling species *S. paradoxus*, and six of the *S. cerevisiae* strains previously examined (Fig. 3 B–F) (38). Phylogenetic analysis of the DUP240 region on chromosome I shows that YJM789 is markedly distinct from all other *S. cerevisiae* strains but is similar to *S. paradoxus* (Fig. 3 C and E). As determined from a sequence outside this region on chromosome I, a different phylogenetic relationship exists among the strains (Fig. 3D). Indeed, the nucleotide similarity between YJM789
and *S. paradoxus* within the region appears higher (93%) than along the rest of chromosome I (Fig. 3B). Although several large indels exist between YJM789 and *S. paradoxus*, *S. paradoxus* is the organism with the highest similarity to YJM789 currently in the GenBank database. The average sequence identity is 85% between these two genomes (Fig. 3F).

Although rearrangements characterize variation in DUP240 ORFs among several *S. cerevisiae* strains (38), the degree of separation of YJM789 from other *S. cerevisiae* strains, the similarity between YJM789 and *S. paradoxus*, and the divergence in phylogeny to genes outside this region were unexpected. Introgression between YJM789 and *S. paradoxus* or a closely related species is a possibility that can account for these observations. Indeed, *S. paradoxus* and *S. cerevisiae* share similar habitats (39), and hybrids between the two are found in nature (40). Although hybrids between *S. cerevisiae* and other members of the *Saccharomyces* sensu stricto are predominantly sterile, rare viable offspring containing DNA from both species have been produced (41–43), providing a putative way for introgression to occur.

### Highly Polymorphic ORFs

Regions of high polymorphism density also are found localized to individual ORFs. The absence of YJM789 DNA hybridization to oligonucleotides representing several S288c ORFs has been reported and, in some cases, interpreted as missing sequences rather than highly polymorphic regions (26). The genealogy sequence enables an investigation of this issue. We have obtained high-quality sequences covering the vicinities of the proposed ORF regions (SI Table 4). Six of these ORFs indeed appear to be absent in YJM789 (YHR054C, YIL080W, YIL082W, YIL113W, and YIL114W). However, 22 ORFs are confirmed to be present but highly polymorphic in YJM789, including six genes in the highly polymorphic chromosome I region.

One notable example of a highly polymorphic ORF is *PDR5*, which encodes a multidrug transporter. *PDR5* is among the genes predicted absent from the YJM789 genome (26). Sequencing shows that it is present but that it contains >250 SNPs (no indels), resulting in 5.3% amino acid differences between YJM789 and S288c. Because the regions flanking the *PDR5* ORF are similar between both strains, the diverged region is highly localized (Fig. 4). The origin of the divergence seen in *PDR5* are unclear. The closest matching sequence to YJM789 *PDR5* in GenBank is S288c PDR5. *PDR5* is the closest paralog to *PDR5* in both strains, yet there is >25% divergence at the protein level between *PDR5* and *PDR5* in each genome. Because this divergence is higher than the divergence observed for the two *PDR5* orthologs (5.3%) (SI Fig. 8), ectopic recombination between *PDR5* and *PDR5* may not be the cause of high divergence between YJM789 and S288c. Interestingly, the corresponding gene products in *S. paradoxus* and, potentially, in RM11-1a are both truncated. The *PDR5* gene product in YJM789 appears to be inactive for at least one substrate, resulting in cycloheximide hypersensitivity in this strain (25). No obvious loss-of-function mutations (frameshift or nonsense) were detected in the coding sequence, although such mutations might have been anticipated if there had been selection for loss of Pdr5p function or if there had been random genetic drift after inactivation.

### Indels

Indels between the genome sequences of YJM789 and S288c were identified by using chromosome-by-chromosome examinations of the sequence alignments to reveal the physical gaps (see Fig. I for indel analysis results for chromosome XIV and SI Fig. 5 for the other 13 chromosomes). Within the high-quality YJM789 sequence, 275,836 bp were identified in the S288c genome that are absent in YJM789, and 48,764 bp in high-quality YJM789 sequence, 275,836 bp were identified in the S288c genome that are absent in YJM789, and 48,764 bp in

Wei et al.

PNAS | July 31, 2007 | vol. 104 | no. 31 | 12827
Fig. 3. Highly polymorphic region on chromosome I. (A) SNP distribution between YJM789 and S288c determined from a 1-kb sliding window over the nonrepeat sequence of S288c chromosome I. (B) Clustergram of the sequence similarity of chromosomes I of YJM789 compared with S288c, RM11-1a, and S. paradoxus (S. para) using a 1-kb sliding window. (C) Clustergram of the sequence similarity of the high polymorphism region of YJM789 chromosome I compared with S288c, RM11-1a, and S. paradoxus using a 100-bp sliding window. (D) Phylogeny of chromosome I sequences excluding the interval containing the high polymorphism region. The phylogenetic tree was constructed from nucleotide sequence alignments generated by using the program VISTA and the neighbor-joining method of the PHYLIP package. (E) Phylogeny of chromosome I sequences using a 1-kb sliding window. (F) Alignments of YJM789, S288c, and S. paradoxus over the high polymorphism region using S288c (Upper) or YJM789 (Lower) as the reference sequence. The y axis represents the sequence similarity between two genomes along the reference sequence (graphs generated in VISTA). Sequence identity is shown for each pairwise comparison in a 100-bp sliding window. Note that differences in sequence lengths arise because of indels between YJM789 and S288c. Genes, as encoded in S288c, are represented by colored boxes: red, verified ORFs; pink, uncharacterized ORFs; gray, dubious ORFs; black, tRNAs and long terminal repeats.

compared with S0 in S288c. Ty3 and Ty4 were not found and are suspected to be absent from the YJM789 genome, a result supported by the absence of hybridization to probes covering these genes during array analysis (26).

Gene Product Polymorphisms and Their Phenotypic Consequences.

Many orthologs between YJM789 and S288c contain nucleotide polymorphisms that affect either the sequence or length of the corresponding gene products. The 5% most variable genes with nonsynonymous polymorphisms (and no indels) are found to be significantly enriched in unknown functions (SI Fig. 10 for gene ontology category comparison). Gene product length polymorphisms resulting from in-frame or out-of-frame indels, ORF fusions, nonsense mutations (SNPs), and Ty polymorphisms (SI Data Sets 2 and 3 list selected ORFs of each category) are less abundant and likely to impact gene product functions.

There are cases where two or more ORFs annotated as separate in S288c appear to be a single ORF in YJM789 (SI Data Set 3). One case is NFT1, annotated in S288c as two genes (YKR103W and YKR104W). The stop codon becomes a tyrosine-encoding TAT codon in YJM789, as well as in several other Saccharomyces species, resulting in a longer ORF (44). Another case involves the S288c ORFs, YJL107C and PRM10, which appear to be a single ORF in YJM789 and other fungi (45, 46).

Although sequence information alone is inadequate for predicting the phenotypic consequence of polymorphisms, there are a few cases for which such consequences can be proposed. One example is the inactivating missense polymorphism found in the
S288c AMN1 gene, which is responsible for yielding nonclumpy cells. RM11-1a cells do not separate efficiently, form clumps, and lack this polymorphism (46). Likewise, the YJM789 AMN1 gene contains the same SNP as that of RM11-1a. Although YJM789 is less clumpy than RM11-1a, it is much more clumpy than S288c (Gael Yvert, personal communication).

In addition, several genes with product length polymorphisms appear to be functional in YJM789 but apparently not functional in S288c (SI Data Set 5). For example, the S288c HAP1 gene contains a partially inactivating Ty1 insertion, resulting in a hap1 hypomorphic mutation (47), whereas the YJM789 HAP1 gene has no insertion. The S288c FLO8 gene (a transcription factor required for pseudohyphal formation) contains an inactivating amber mutation (48), yet YJM145 (a diploid isogenic with YJM789) forms abundant pseudohyphae (20), and its FLO8 ORF has no amber mutation.

Many genes bearing intragenic tandem repeats have different frame repeat numbers between S288c and YJM789. Several of these genes encode cell surface proteins (such as TIR1, HIS150, FIT1, AGA1, MNN4, and FLO10), and their variation in tandem repeats may generate functional cell surface variability that has been reported to contribute to a rapid adaptation to the environment and possibly host immune evasion (49).

Mitochondrial Genome. The mitochondrial sequence of YJM789 is collinear with that of strain S288c and approximately the same size (86,214 vs. 85,779 bp). For much of the mitochondrial genome, sequence identity is >98%. Nevertheless, these genomes differ in several ways. Strain YJM789 is missing ~54 GC-rich transposable elements (50) but contains 17 that are not present in S288c, none of which disrupt verified genes.

One particularly interesting gene present in the YJM789 mitochondrial genome is a matruse, an ortholog of Candida stellata cox-2. In YJM789, this ortholog is encoded partly in an additional intron of COXI (intron 4). In addition to the cox-2 ortholog, three regions of high sequence divergence exist. Intron 6 of COXI, which is 1,487 bp, is approximately the same length as intron 5 of COXI of S288c (1,365 bp) but essentially shows no sequence conservation. High variation is found between the ATP6 gene and tRNA-Glu and includes the region encoding the putative RF3 matruse protein. Furthermore, between COX2 and RNA-Phe is a 1,417-bp region, which encodes the hypothetical RF1 gene that has only 72% identity to the corresponding region of S288c.

Implications. The YJM789 genome sequence is marked by extensive polymorphisms relative to the laboratory strain S288c throughout the nuclear and mitochondrial genomes. The ~60,000 SNPs scattered over the genome alignment represent a SNP frequency of 1 in 164 bp (0.6%), which is higher than the divergence between human beings (0.1–0.01%). High SNP frequencies together with indels could account for the reduced spore viability seen in crosses of these two strains: 87.4% for YJM789/S288c background hybrids, compared with 97.6% for S288c/S288c (19). In comparison with SNPs, the number of indels measuring >100 bp (269) is moderate. Although the indels involve ~324 kb (within the aligned 98% of the S288c genome), much of them represent repeat sequences, which could suggest that SNPs might be a primary cause of heritable phenotypic variation between these two strains.

Although the idea of horizontal gene transfer has been accepted in bacteria (51), eukaryotic genomes were initially thought to be units that do not exchange genetic information (52). The YJM789 genome provides preliminary evidence to suggest a putative horizontal transfer of YJM-GNAT from bacteria and a potential introgression of an ~12-kb chromosome I sequence from closely related yeast. Although further analysis with sequences of more yeast strains will be informative for proof, the possibilities of such horizontal genetic exchanges are in line with an increasing number of reports describing introgression or horizontal genetic exchange in Saccharomyces sensu stricto species (36, 40, 53–55).

Finally, we made the YJM789 genome a free-to-access resource that marks an initial step toward a more complete set of reference sequences for the S. cerevisiae species. The benefits of complete genome information of several individuals can soon be explored. One key application will be the development of new technologies to interrogate the genome content of several S. cerevisiae strains by including, for example, polymorphism-specific probes on tiling microarrays. These technologies have promise to further advance applications in yeast to define the contribution of sequence variants to heritable traits. Importantly, applied to YJM789, these technologies will help us to understand how sequence polymorphisms change the information encoded in the genome to confer pathogenicity.

Materials and Methods

Gene Prediction and Comparison. We used three different gene-prediction methods to identify potential ORFs: (i) directly

---

Fig. 4. Polymorphism density across PDR5 between YJM789 and S288c. (A) Polymorphism distribution on chromosome XV from kilobases 600 to 640. Dashed lines indicate the start and stop positions of the PDR5 ORF. (B) The distribution of nonsynonymous and synonymous substitutions within the PDR5 ORF as determined from a 900-bp sliding window (each slide is 90 bp). The possibilities for nonsynonymous and synonymous substitutions were calculated as described previously (61). Red, nonsynonymous substitutions; green, synonymous substitutions; blue horizontal bars, transmembrane domains; vertical bars at the bottom, substitution sites.
mapping genes by using the S288c-verified ORFs from the *Saccharomyces* database (56). (ii) ORF calling based on the positions of start and stop codons, and (iii) GLIMMER5 (57). Ortholog assignments were required to meet all of the following criteria: reciprocal best match with an E value of the high-scoring segment pair (HSP) less than 10^-6, an identity of ≥40%, and a match length of at least 75% of both protein lengths. Homologs were identified in cases for which no reciprocal best hit was obtained as the nearest S288c best hit homolog, with threshold requirements as described above. YJM789 genes without S288c homologs were defined as YJM789-specific.

**ORF Annotation.** The gene names, functional descriptions, and gene ontology categories for the YJM789 genes with S288c orthologs or homologs were annotated according to their S288c counterparts. The annotation of specific genes of YJM789 was based on comparison to the nonredundant database. For the YJM789 genes with S288c homologs that contained frameshift, indel, or missense mutations, the nature of the potentially inactivating polymorphism was identified. Complete annotations are provided in SI Data Set 1.

**Genome Alignment.** The public software BLASTN (58) and MUMessage (59) were used to align the sequences of the high-quality contigs of YJM789 to the individual S288c chromosome sequences. The results of these two analyses were checked manually and combined. In regions of disagreement between the alignment programs, the alignment with the highest sequence similarity was chosen. The polymorphism sets were derived from the final S188c-

**Additional Materials and Methods.** Further details are found in SI Text.

We thank Gael Yvert for information regarding YJM789 clumsiness and Michael Knop, Rui Wang-Sattler, Guohui Ding, Kang Tu, Lin Tao, Hao Xu, Hong Yu, and Ziliang Qian for helpful discussion, suggestions, and assistance with calculations. The RM11-1a sequence was obtained from the *S. cerevisiae* RM11-1a Sequencing Project, Broad Institute of Harvard and Massachusetts Institute of Technology. This work was supported by National Institutes of Health Grants HG02052 (to R.W.D.), GM008717 (to R.W.D. and L.M.S.), and HG000205 (to R.W.D. and L.M.S.); China National Basic Key Research Program Grants 2003CB715901 (to Y.L.), 2006CB501606 (to Y.L.), 2006CB910703 (to Y.L.), 2004CB720103 (to Z.C.), and 2006AA02Z317 (to Z.C.); China National Natural Science Foundation Grants 30500107 and 30670953 (to Z.C.); and Science and Technology Commission of Shanghai Municipality Grant 06PJ14072 (to Z.C.).

SI Figure 6
PDR15_S288c
PDR15_YJM789

PDR5_S288c
PDR5_YJM789

SI Figure 6

0.05
ORFs with high variation

All ORFs

Different GO categories

Fractions of different GO categories

SI Figure 10