Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics

Calum J. Maclean,†,1 Brian P.H. Metzger,†,‡,1 Jian-Rong Yang,†,§,1 Wei-Chin Ho,1 Bryan Moyer,§,2 and Jianzhi Zhang*,1

1Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI
2Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI
†Present address: Department of Ecology and Evolution, University of Chicago, Chicago, IL
§Present address: Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong, China
‡Present address: Department of Ecology and Evolution, University of Chicago, Chicago, IL
*Present address: HudsonAlpha Institute for Biotechnology, Huntsville, AL
†These authors contributed equally to this work.
*Corresponding author: E-mail: jianzhi@umich.edu.
Associate editor: Ryan Hernandez

Abstract

The budding yeast Saccharomyces cerevisiae is the best studied eukaryote in molecular and cell biology, but its utility for understanding the genetic basis of phenotypic variation in natural populations is limited by inefficient association mapping due to strong and complex population structure. To overcome this challenge, we generated genome sequences for 85 strains and performed a comprehensive population genomic survey of a total of 190 diverse strains. We identified considerable variation in population structure among chromosomes and identified 181 genes that are absent from the reference genome. Many of these nonreference genes are expressed and we functionally confirmed that two of these genes confer increased resistance to antifungals. Next, we simultaneously measured the growth rates of over 4,500 laboratory strains, each of which lacks a nonessential gene, and 81 natural strains across multiple environments using unique DNA barcode present in each strain. By combining the genome-wide reverse genetic information gained from the gene deletion strains with a genome-wide association analysis from the natural strains, we identified genomic regions associated with fitness variation in natural populations. To experimentally validate a subset of these associations, we used reciprocal hemizygosity tests, finding that while the combined forward and reverse genetic approaches can identify a single causal gene, the phenotypic consequences of natural genetic variation often follow a complicated pattern. The resources and approach provided outline an efficient and reliable route to association mapping in yeast and significantly enhance its value as a model for understanding the genetic mechanisms underlying phenotypic variation and evolution in natural populations.

Key words: Saccharomyces cerevisiae, GWAS, growth rate, genome sequencing, population structure, drug resistance.

Introduction

Understanding the genetic basis of phenotypic variation is a major goal of modern biology. Model organisms play a prominent role in this endeavor because of the wealth of accumulated biological information and tools available for manipulating and examining these organisms. The budding yeast Saccharomyces cerevisiae has long been a favored eukaryotic model organism to molecular and cell biologists and was the first eukaryote to have its genome fully sequenced (Goffeau et al. 1996). Large-scale phenotyping of gene deletion (Giaever et al. 2002; Ohya et al. 2005; Hillenmeyer et al. 2008; Giaever and Nislow 2014) and overexpression (Sopko et al. 2006; Douglas et al. 2012) strains has provided extensive data on gene function, while the availability of genome sequences from closely related species (Kellis et al. 2003; Dujon et al. 2004; Scannell et al. 2011; Hittinger 2013; Liti et al. 2013) offers evolutionary insights into genotype–phenotype mapping.

Prior to 2009, the identification of genotype–phenotype relationships in yeast relied on only a few laboratory strains, which are now known to be phenotypic outliers (Liti et al. 2009; Warringer et al. 2011). Recent years have seen intensified research on the natural diversity and ecology of S. cerevisiae. For instance, S. cerevisiae has been isolated globally from diverse natural and man-made environments (Liti et al. 2009; Wang et al. 2012) and shown to harbor greater phenotypic diversity but much lower genetic diversity than its sister species S. paradoxus (Liti et al. 2009; Warringer et al. 2011). As such, S. cerevisiae has great potential for linking natural variation in phenotype to genetic variation in individual genes, a process that requires knowledge of both yeast genomic diversity and population structure. Such knowledge has accumulated primarily through low-coverage Sanger sequencing (Liti et al. 2009) and tiling array hybridization (Schacherer et al. 2009). These studies, as well as restriction-site associated DNA sequencing of a large strain set (Cromie et al. 2013), have revealed a complex population structure of S. cerevisiae. Higher-quality genomes produced by next-generation sequencing have further revealed the presence of both copy-number variants and genomic rearrangements

© The Author 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
All rights reserved. For permissions, please e-mail journals.permissions@oup.com

(Bergström et al. 2014; Hose et al. 2015; Strope et al. 2015), as well as the origins of domestic S. cerevisiae strains (Gallone et al. 2016; Gonçalves et al. 2016). Unfortunately, while strains representing pure lineages are often phenotypically distinct, many S. cerevisiae strains are mosaics with complex ancestry from multiple lineages due to human activity (Warringer et al. 2011). This strong and complex population structure has made genome-wide association study (GWAS), an important forward genetic method for detecting influential genetic variants in many species, difficult in yeast (Connelly and Akey 2012; Diao and Chen 2012). Consequently, the use of this otherwise powerful model species for systematic analysis of the genetic basis of natural phenotypic variation has been hindered.

To overcome this hurdle, we developed a resource for efficient GWAS in S. cerevisiae that simultaneously combines forward and reverse genetic analyses. While these approaches are individually commonly used in the yeast community (Smith et al. 2011; Liti and Louis 2012; Swinnen et al. 2012; Fay 2013; Giaever and Nislow 2014; Long et al. 2015), our approach allows both forward and reverse genetic information to be gained from a single experiment. To do so, we first generated genome sequences of 85 diverse S. cerevisiae strains that are genetically and phenotypically variable. Combining this data with available genome sequences from the literature, we assembled a dataset of 190 S. cerevisiae genomes and conducted a comprehensive population genomic analysis, identifying single nucleotide polymorphisms (SNPs) at ~3.5% of sites. From this information, we elucidated detailed phylogenetic relationships among strains and the broad population structure of the species. We detected genes from the newly sequenced genomes that are absent from the reference genome and demonstrated their expression and functions. We then barcoded the newly sequenced strains and simultaneously phenotyped them with over 4,500 single gene deletion strains by a high-throughput barcode-sequencing (bar-seq) method (Smith et al. 2009). With the assistance of this reverse genetic information, our GWAS identified potential causal genes responsible for growth rate variations in five of six environments examined. We experimentally verified a subset of these associations for high-temperature growth by a reciprocal hemizygosity test (Steinmetz et al. 2002), establishing the combination of simultaneous association mapping and reverse genetics as a powerful approach for unbiased identification of the genic basis underlying fitness variation among natural S. cerevisiae strains.

This resulted in an average coverage of 60× per genome (range 38–99×) (supplementary data S2, Supplementary Material online). On average, 6% of the reference genome was not covered by a read in each sequenced strain due to stochastic sampling of reads and/or strain differences in gene content as well as repeat elements. In total, we identified 311,287 single nucleotide polymorphisms (SNPs) and 15,884 insertions/deletions (indels).

To acquire a more comprehensive view of S. cerevisiae strain relations, we identified 105 additional strains that have publically available genome sequences (Hose et al. 2015; Strope et al. 2015) at the time of our analysis (August 2016), and applied our analysis pipeline to the Illumina sequencing reads of these strains. A neighbor-joining tree of all 190 strains was then constructed on the basis of a combined set of 421,773 SNPs (fig. 1b; supplementary fig. S1 and data S1–S2, Supplementary Material online). The tree was rooted using outgroup sequences from recently identified Chinese isolates (Wang et al. 2012). We recovered phylogenetic clustering based roughly on geographical and environmental origins of the strains, consistent with early observations made from fewer strains and SNPs (Liti et al. 2009; Schacherer et al. 2009). Clustering can be seen of strains into the West African, Malaysian/North American, Sake, Laboratory, and European/Wine groups previously identified. Additionally, we identified a “Bakery” clade that was previously suggested to exist by microsatellite-based analysis (Legras et al. 2007) and a new Natural clade of three wild strains (one from soil in Illinois and two on the gums of wild cherry trees from unknown locations) (fig. 1b). The remaining strains, originating from a wide variety of environments, form a group named “Mosaics” (fig. 1b; see below). While this work was under preparation for submission, two genomic studies focusing on closely related wine and beer strains were published (Gallone et al. 2016; Gonçalves et al. 2016). Because these domestic strains largely represent a single clade in our analysis, these two studies do not offer a species-wide view of yeast’s evolutionary history and the exclusion of these strains does not affect the overall phylogenetic patterns observed in our analysis.

To more closely examine the population structure of S. cerevisiae, we employed a model-based clustering algorithm implemented in fastSTRUCTURE (Raj et al. 2014) and identified seven distinct subpopulations (top row in fig. 1c) that are in agreement with the strain isolation sources and corroborate the clustering pattern seen in the phylogeny. In addition, many strains that do not fall within a specific clade are mosaics with ancestry from several lineages (top row in fig. 1c), supporting previous observations based on smaller data (Liti et al. 2009). We then conducted the fastSTRUCTURE analysis for each of the 16 chromosomes and observed prominent among-chromosome variations in population structure (fig. 1c). For example, the West African subpopulation is genetically distinct from other subpopulations for 10 of its 16 chromosomes, but is indistinct from the sake subpopulation for three chromosomes and indistinct from the Malaysian/North American subpopulation for another three chromosomes. This variation in chromosomal population structure is indicative of differences in the

Results

Genomic analysis of S. cerevisiae Reveals a Complex Population Structure

To identify genetic variation underlying phenotypic variation among S. cerevisiae strains, we generated genome sequences of 85 strains collected in six continents and from a variety of human-associated and wild environments (fig. 1a; supplementary data S1, Supplementary Material online). We obtained an average of 3.75 million 2×100-nucleotide paired-end reads per strain, approximately 97% of which were successfully mapped to the S288c reference genome.
Fig. 1. Geographical, environmental, and phylogenetic relationships of the 190 S. cerevisiae strains analyzed. (a) World map indicating the geographic locations where the analyzed strains were sampled. Colors represent the environment of isolation if known. (b) Maximum composite likelihood neighbor-joining tree of the 190 strains based on genome-wide SNP data. The environment type from which each strain was isolated is indicated as a colored circle. Branches are colored to denote clade. The scale bar represents 0.1% genome sequence divergence. Strain names in black are those sequenced in this work, while those in gray were sequenced previously. The same tree with bootstrap values is shown in supplementary figure S1, Supplementary Material online. Trees based on individual chromosomes are provided in supplementary figure S2, Supplementary Material online. (c) Population structures of the 190 strains. Strains are arrayed based on clade membership in panel b. Different
evolutionary histories of different chromosomes due to pervasive gene flow. Because *S. cerevisiae* reproduces largely asexually (Tsai et al. 2008), rare crosses between lineages can establish unique populations where distinct chromosome combinations persist in the absence of outbreeding. The observed differences in chromosomal population structure are not due to stochasticity in the population structure assessment because multiple runs on the same chromosome displayed only minor variations. Furthermore, phylogenies reconstructed using SNPs from individual chromosomes corroborate the fastSTRUCTURE results (supplementary fig. S2, Supplementary Material online).

The extent of linkage disequilibrium (LD) between SNPs is an important characteristic determining the highest possible resolution of association analysis; the lower the LD, the higher the resolution can be. We found that the mean LD measured by \( r^2 \) equals 0.0164 for SNPs within 100 nucleotides and it halves as the physical distance increases to 1,200 nucleotides (fig. 1d). This fast breakdown of linkage disequilibrium is similar to previous reports (Liti et al. 2009). Given that the average distance between the beginning of one gene and that of the next gene on the chromosome is \( \sim 2 \) kb in yeast, this result indicates that fine-scale mapping to the gene level should be theoretically possible by GWAS in this species. Nevertheless, due to the variation in recombination rate across the yeast genome, mapping resolution is expected to vary among genomic regions.

To further characterize the genetic variation present in *S. cerevisiae*, we conducted a comprehensive population genomic analysis of all 190 strains. The basic population genetic parameters are summarized in Table 1. Intronic and intergenic polymorphisms (\( \theta_W \)) and nucleotide diversities (\( \pi \)) are significantly lower than those at synonymous sites, suggesting pervasive purifying selection acting on noncoding regions. This is consistent with the fact that the compact nature of the yeast genome results in intergenic regions dense with promoter and other important regulatory elements and that yeast introns can regulate gene expression (Juneau et al. 2006; Parenteau et al. 2008). Consistent with previous analyses conducted using a smaller data set (Liti et al. 2009), we found no clear sign of positive selection using the majority of population genetic tests we applied, and the estimated fraction of adaptive amino acid substitutions was 0. However, removing the Wine/European strains from analysis resulted in the detection of positive selection in one test, suggesting that sampling biases and population structure may confound current approaches for detecting positive selection (supplementary figs. S3–S5 and tables S1–S2, Supplementary Material online; see “Materials and Methods” for details).

Finally, for the 85 newly sequenced genomes, we also analyzed and observed intron presence/absence polymorphisms (supplementary fig. S6, Supplementary Material online; see "Materials and Methods") and identified aneuploidy and large segmental duplications (supplementary fig. S7, Supplementary Material online; see "Materials and Methods"). Similar to the results of a previous in-depth investigation into *Saccharomyces* mitochondrial DNA from a largely different set of strains (Wolters et al. 2015), we found several differences in intron content between related isolates. We also detected many small- and large-scale duplication, many of which show no clear pattern of phylogenetic distribution (Hose et al. 2015).

**Some Non-Reference Genes Confer Drug Resistance**

Our knowledge of *S. cerevisiae* genome content and function is derived largely from a few laboratory strains, which are now known to be phenotypically atypical (Warringer et al. 2011). Furthermore, S288c, the strain from which the reference set of genes are defined, was constructed in the 1980s largely from a strain that was isolated in the late 1930s (Mortimer and Johnston 1986). Because the reference strain has been maintained in relatively benign and unvarying laboratory environments and has undergone repeated population bottlenecks, often to single individuals, the possibility arises that genes important for survival outside of the laboratory environment have been lost. As a consequence, identifying coding regions not present in the reference strain has the potential to explain phenotypic variation among strains. Indeed, recent reports have confirmed the presence of several nonreference genes in the genomes of natural and industrial *S. cerevisiae* strains (Novo et al. 2009; Borneman and Pretorius 2015; McIlwain et al. 2016). To further our understanding of the distribution and importance of nonreference genes, we additionally performed a *de novo* genome assembly of the Illumina sequencing reads obtained from our 85 strains (see supplementary data S3, Supplementary Material online for assembly statistics). We identified 181 distinct nonreference genes distributed across the phylogeny (fig. 2a; supplementary fig. S8 and data S4, Supplementary Material online). The majority were found to have a BLAST hit in a previously sequenced nonS288c *S. cerevisiae* strain, while others had best hits in other fungi or more distantly related organisms (fig. 2a). The phylogenetic distribution of nonreference genes did not follow a straightforward pattern; distantly related strains

### Table 1. Polymorphism (\( \theta_W \)) and Nucleotide Diversity (\( \pi \)) Per kb in Different Regions of the Yeast Genome.

<table>
<thead>
<tr>
<th>Region</th>
<th>( \theta_W ) Mean</th>
<th>( \theta_W ) SD</th>
<th>( \pi ) Mean</th>
<th>( \pi ) SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>5.63</td>
<td>0.018</td>
<td>2.67</td>
<td>0.015</td>
</tr>
<tr>
<td>Intronic</td>
<td>7.52</td>
<td>0.135</td>
<td>3.68</td>
<td>0.122</td>
</tr>
<tr>
<td>Synonymous</td>
<td>14.60</td>
<td>0.035</td>
<td>9.08</td>
<td>0.037</td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>3.63</td>
<td>0.010</td>
<td>1.37</td>
<td>0.007</td>
</tr>
<tr>
<td>Total</td>
<td>5.97</td>
<td>0.009</td>
<td>2.99</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Fig. 1 Continued colors show different inferred populations, which are indicated at the top of the panel. The Y-axis shows the fraction of SNPs coming from each inferred population. (d) Linkage disequilibrium (LD) decays as the physical distance between two linked sites becomes larger. LD is measured by \( r^2 \) between two linked sites minus the mean \( r^2 \) between two sites located on different chromosomes.
often share nonreference genes, suggesting independent gains via introgression or horizontal gene transfer, independent losses, or segregating polymorphism (fig. 2a; supplementary fig. S8, Supplementary Material online).

To begin characterization of these nonreference genes, we first estimated their expression levels in 23 S. cerevisiae strains with available transcriptome data generated by mRNA sequencing (RNA-seq) (Skelly et al. 2013). We found that on average, 54.9% of the nonreference genes examined had a higher expression level than each of the 5% most lowly expressed reference genes (fig. 2d; supplementary data S5, Supplementary Material online). Because the RNA-seq data were collected in a single benign environment, it is likely that more of the nonreference genes identified are expressed at
appreciable levels in the appropriate environments due to conditional expression.

To examine the functional importance of nonreference genes, we focused on two of them for which our initial BLAST search revealed their closest hit to be within the well annotated Lachancea thermotolerans genome. Following Liti et al. (2009) and Warringer et al. (2011), we used growth curves to determine the phenotypic consequences of deleting these nonreference genes on two aspects of strain growth, maximum growth rate and efficiency (see “Materials and Methods”). Non-Ref-129, identified from the Malaysian strain UWOPS87-2421, resembles the L. thermotolerans coding region KLTH0E00528g, which is annotated as a homolog of S. cerevisiae FLR1, a multi-drug transporter responsible for the efflux of drugs such as the widely used antifungal fluconazole (Gbska et al. 2006). Thus, Non-Ref-129, expressed even in a benign environment (5.25 RPKM; supplementary data S5, Supplementary Material online), may confer resistance to this important drug. We deleted Non-Ref-129 from haploid UWOPS87-2421 cells and exposed both the wild-type and deletion strains to various fluconazole concentrations to investigate the impact of gene deletion on strain growth (fig. 2c). Deleting Non-Ref-129 had a small but significant effect on the maximum growth rate (supplementary fig. S9a, Supplementary Material online) and a large effect on growth efficiency, especially when fluconazole concentration is comparable with a typical high-dose fluconazole treatment in clinical settings (>25 μg/ml) (Menichetti et al. 1996; Martin 1999) (fig. 2c). If Non-Ref-129 is indeed a drug transporter similar to FLR1, it may also be involved in resistance to diazaborine, benomyl, methotrexate, and other drugs (Brôco et al. 1999; Jungwirth et al. 2000).

To better understand the history of Non-Ref-129, we performed additional BLAST searches in other published S. cerevisiae genomes that were built de novo (Strope et al. 2015). We found in YJM653 an intact Non-Ref-129 and in YJM1250 an apparently pseudogenized Non-Ref-129 that is disrupted by an insertion; these two strains respectively reside at the edge of and within the Wine/European clade, both being highly diverged from UWOPS87-2421 (supplementary fig. S9a, Supplementary Material online). UWOPS87-2421 and YJM653 differ at only one nonsynonymous site and no synonymous sites in this 1644-nucleotide gene. The highly similar chromosomal locations of all three Non-Ref-129 genes in an unstable telomeric region (UWOPS87-2421, ChX:33185-34826; YJM653, ChX:30819-32462; YJM1250, ChX:22819-24457) suggest a single origin of Non-Ref-129 in S. cerevisiae.

The second nonreference gene experimentally studied, Non-Ref-67, was initially identified in strains CLIB272 and Y6 and found to be similar to the L. thermotolerans gene KLTH0H09460g, which is homologous to CRG1 in S. cerevisiae, a methyltransferase gene involved in lipid homeostasis and providing resistance to the phosphatase inhibitor cantharidin (Lissina et al. 2011). Unlike Non-Ref-129, Non-Ref-67 is not expressed in benign conditions (supplementary data S5, Supplementary Material online). However, because CRG1 expression increases by 40- to 50-fold upon exposure to cantharidin and other stresses (Lissina et al. 2011), it is possible that Non-Ref-67 expression is condition-specific. We deleted Non-Ref-67 from haploid CLIB272 cells and exposed wild-type and deletion strains to varying concentrations of cantharidin (fig. 2d; see “Materials and Methods”). The deletion of Non-Ref-67 does not alter the maximum growth rate consistently (supplementary fig. S9b, Supplementary Material online), but has a significant effect on growth efficiency at intermediate drug levels, with the deletion strain reaching only ~50% of the maximum OD of the wild-type strain at a cantharidin concentration of 12 μM (fig. 2d). Additional BLAST searches detected Non-Ref-67 in five S. cerevisiae genomes previously published (Strope et al. 2015). We identified two distinct versions of Non-Ref-67 distributed across the phylogeny (inset of supplementary fig. S9b, Supplementary Material online) that differ at 10 sites, including four nonsynonymous sites. However, the similar genomic location of these two variants, in a telomeric region of Ch. XV, suggests a single origin.

Two possible scenarios can explain the origin and phylogenetic distribution of each of the nonreference genes studied in depth here (Non-Ref-129 and Non-Ref-67). The first is that a gene arose via horizontal gene transfer after the separation of S. cerevisiae from S. paradoxus. The acquired gene may not have been fixed in S. cerevisiae if the transfer was recent. Alternatively, the gene may have been fixed, followed by multiple losses within S. cerevisiae. The two nonreference genes have relatively low levels of sequence identity to the L. thermotolerans genes mentioned, suggesting that the donor species have yet to be identified. The second scenario is that these nonreference genes arose from gene duplication. The relatively large sequence dissimilarities of these genes with their closest paralogs in S. cerevisiae suggest that the duplication events were ancient, implying multiple independent losses of these genes in several yeast species as well as within S. cerevisiae, which is possible given their subtelomeric locations. While the first scenario appears more parsimonious, both scenarios remain possible at this stage.

Bar-Seq Allows High-Throughput Simultaneous Phenotyping of Thousands of Strains
Accurate phenotyping is central to uncovering the genetic basis of phenotypic variation. Phenotyping different natural yeast strains has primarily relied on the production and analysis of growth curves (Warringer et al. 2011) or digital photography-based colony sizes (Bloom et al. 2013), which have limited throughput and resolution and can be time-consuming. We decided to adopt bar-seq (Smith et al. 2009) for phenotyping, which allows for the simultaneous measurement of the growth rates of all strains of interest in the same test tube through Illumina sequencing of strain-specific DNA barcodes. Bar-seq was originally designed to quantify the relative growth rates of S288c-derived gene deletion strains each carrying two unique pieces of 20-nucleotide DNA (barcodes) inserted at the time of strain construction (Winzeler et al. 1999; Giaever et al. 2002). We similarly constructed a panel of barcoded strains from a subset of the 85 strains sequenced in this work. Each carries two

doi:10.1093/molbev/msx151
unique barcodes which are not present in any deletion strain. This not only allows the bar-seq experiment of the natural strains but also that of natural strains and deletion strains all in one test tube (fig. 3a). Although our methodology could easily be used to expand the strain set in the future, at the time of our phenotypic analysis the additional 105 sequenced strains included in our phylogenetic analysis were unavailable.

We successfully inserted the unique barcodes, flanking a G418 sulfate resistance marker (KanMX4), at the HO (YDL227C) locus of 81 of the 85 diploid strains sequenced here (supplementary fig. S10 and data S1, Supplementary Material online; see "Materials and Methods"). Of these 81 heterozygous HO/hoΔ::KanMX4 strains, we obtained stable α and α haploids for 76 of them. We then created MATa hoΔ::HygMX4 (Hygromycin B resistance) and MATα hoΔ::NatMX4 (nourseothricin resistance) haploids through marker switching (supplementary data S1, Supplementary Material online; see "Materials and Methods").

To test the utility of these barcoded strains for studying the genetic basis of phenotypic variation, we combined the 81 barcoded diploid strains with the 4,653 diploid strains from the S. cerevisiae homozygous nonessential gene deletion collection to create a common starter pool to use across our experiments (fig. 3a). We grew this pool to a benign environment (YPD at 30 °C) as well as six stressful environments: high temperature (YPD at 40 °C), high salt/osmotic stress (YPD at 30 °C + 1.25 M NaCl), high ethanol (YPD at 30 °C + 7% ethanol), superoxide anions (YPD at 30 °C + 4 mM paraquat), oxidizing agents (YPD at 30 °C + 3 mM hydrogen peroxide), and a hypoxia mimetic (YPD at 30 °C + 1 mM cobalt.

**Fig. 3.** Simultaneous high-throughput phenotyping of 81 natural isolates and 4,521 gene deletion strains. (a) Flow chart showing the procedure of barcoding the natural isolates and simultaneously phenotyping natural isolates and gene deletion strains. (b) Heat map showing the fitness of 81 natural isolates relative to the reference strain (S288c-derived HO deletion strain) in each stressful environment, relative to that in the benign environment of YPD at 30 °C. The colored arrows show the color scheme for relative fitness, with the most extreme colors depicting the most extreme fitness values in each environment. Colored circles indicate the environmental origins of the natural isolates as in figure 1a. (c) The fitness of 81 natural isolates and 4,521 S288c derived gene deletion strains relative to the reference strain in the high temperature environment, relative to that in the benign environment. The natural isolates are shown by colored circles, based on the color scheme in figure 1a. The gene deletion strains are shown by black or gray circles depending on their locations on odd-numbered or even-numbered chromosomes, respectively, and are arranged by chromosomal position. Dashed vertical lines indicate the regions identified as significant by GWAS with those shown in red denoting those further investigated by hemizygosity tests in figure 4.
chlordan. We extracted genomic DNAs from the common starting pool and following each competition, produced bar-seq libraries, and quantified the sequencing read number of each barcode, a proxy for strain frequency, at each time point, by Illumina sequencing (fig. 3a). We estimated the fitness of each strain in a particular environment relative to the benign 30 °C YPD environment, relative to the BY4743-derived HO deletion strain (Giaever et al. 2002) to identify strains with particularly high or low relative fitness in the environment of interest (fig. 3b). While there is a clear phylogenetic component to the phenotypic similarity among some strains, being particularly apparent within the North American and Malaysian clades, other strains appear phenotypically diverged from the clades they are most closely related to. The bar-seq data also allowed us to determine the effect each gene deletion has across the tested environments, revealing that relative fitness varies greatly among the gene deletion strains, with 13.8% (623/4521) of gene deletion strains having significantly higher relative fitness and 18.3% (829/4521) having significantly lower relative fitness than the reference strain in high temperature (fig. 3c; supplementary data S6, Supplementary Material online). Similar patterns are observed in the other five environments examined (supplementary fig. S11 and data S6, Supplementary Material online). Broadly, the fitnesses of both the gene deletion strains and the natural strains are positively correlated across the environments tested (supplementary fig. S12, Supplementary Material online).

Combining Forward and Reverse Genetics Reveals SNPs Underlying Phenotypic Diversity

Using a multi-stage GWAS approach (supplementary fig. S13, Supplementary Material online; see “Materials and Methods”), we attempted to identify SNPs associated with relative fitness variation among the 81 natural strains. We discovered between 3 and 19 associated SNPs per environment after controlling for population structure for five of the six environments examined (supplementary data S7, Supplementary Material online). For example, we detected 13 SNPs associated with relative fitness at 40 °C. Five of the 13 SNPs map to a ∼16 kb region on Ch. XI that contains the ribosomal protein gene RPS21a, deletion of which is known to slow growth at high temperature (Sinha et al. 2008). Similarly, a second SNP on Ch. XI, located 51.5 kb from this cluster, is within 7 kb upstream and downstream of the genes DBP7, RPC37 and Gcn3, which again are known to reduce heat tolerance upon deletion, as well as SET3 and YKR023C, known to reduce stress tolerance when deleted.

While association studies rarely validate that the identified SNPs or linked regions are responsible for the observed phenotypic differences, validation can be performed in yeast. Such validations are an increasingly important step in understanding the signals identified by GWAS because many confounding factors—such as the strong population structure and linkage (Connelly and Akey 2012)—can lead to false signals that are hard to untangle without direct genetic manipulations. To this end, we used reciprocal hemizygosity tests to identify difference in fitness due to deletion of alternative alleles in hybrids of high- and low-fitness strains at 40 °C relative to 30 °C for genes surrounding each of several associated SNPs (fig. 4a; see “Materials and Methods”). In each competition experiment, one strain expressed yellow fluorescent protein (YFP), facilitating the quantification of relative fitness by flow cytometry (He et al. 2010). Reciprocal experiments with the YFP marker in opposing hybrid background were performed to remove any fitness effect of YFP expression. We chose two strains with high fitness (YPS128 and YJM320) and two with low fitness (W303, UWOPS05-227.2) at 40 °C relative to 30 °C and selected three significant SNPs for investigation, reciprocally deleting four to six genes surrounding each SNP (supplementary data S7, Supplementary Material online).

The first SNP of interest, the most significant identified, is located at site 490,822 of chromosome XI within the bidirectional promoter of GCN3 and BCH2, both annotated as having temperature related deletion phenotypes. In addition, two neighboring genes (DBP7 and RPC37) are similarly annotated. Our bar-seq data showed that deleting DBP7 from S288c drastically reduced fitness at 40 °C compared to that at 30 °C. To investigate if variations in these genes cause the fitness variation in natural populations, we individually deleted the alternative alleles in W303/YPS128 hybrids for each of the above four genes and compared their growth rates at 40 °C relative to that at 30 °C. As expected, the strain retaining the predicted high-fitness allele of DBP7 from YPS128 grew significantly faster than the strain retaining the predicted low-fitness allele from W303 at 40 °C relative to 30 °C (fig. 4b). No such significant difference was observed for the other three genes tested. Thus, variation in either function or expression of DBP7, which encodes a putative ATP-dependent RNA helicase of the DEAD-box family, likely contributes to fitness variation at 40 °C among these strains. Interestingly, the causal gene is two genes (3.5 kb) away from the significantly associated SNP identified in GWAS.

The second SNP of interest is located within the coding region of CNA1 (Ch XII, 1004315), a gene whose deletion is annotated as increasing stress susceptibility. We individually deleted the alternative alleles in W303/YPS128 hybrids for CNA1 and four neighboring genes and compared their growth rates. Unexpectedly, retention of the high-fitness allele resulted in a higher fitness than retention of the low-fitness allele at 40 °C relative to 30 °C for four of the five genes examined (fig. 4c). This suggests a complex genetic architecture and highlights that the associated SNPs are unlikely to be causal themselves, but instead simply markers for genomic regions harboring naturally occurring variation affecting the trait of interest. Interestingly, unlike the results for DBP7, deletion from S288c for two of these four genes did not have an appreciable impact on relative fitness, suggesting that genes underlying phenotypic variation in natural populations can differ from the genes affecting growth in the laboratory strain. As such, these results indicate the necessity of considering the genetic background when determining gene–phenotype relationships. Furthermore, the type of mutations may also matter, because only null mutations are present in the deletion collection, while some gain-of-function mutations are expected in natural strains.
The third region investigated contains a cluster of five significantly associated SNPs on Ch. XI. Three of these are located in the coding region of \textit{DYN1}, and do not affect the coding regions of the upstream genes \textit{RHO4} and \textit{TRM2}. We individually deleted alternative alleles of these three genes as well as three additional genes, including the thermally annotated gene \textit{RPS21A}, from the hybrid of UWOPS05-227.2 and YJM320. The results confirmed that the \textit{RHO4} and \textit{RPS21A} alleles have different effects on relative fitness at 40°C (fig. 4d). However, deleting \textit{RHO4} from S288c did not significantly alter the relative fitness at 40°C, again suggesting differences in genetic background and/or type of mutation underlying phenotypic variation in natural strains relative to a laboratory strain. Finally, for \textit{GLG1}, the strain carrying the predicted high-fitness allele was outcompeted by the strain carrying the predicted low-fitness allele, suggesting not only that an associated SNP may have multiple causal genetic variants but also that these variants may have opposite fitness effects.

**Discussion**

We have presented here a detailed overview of the genomic diversity within \textit{S. cerevisiae} by combining newly sequenced genomes with those previously published. The genome sequences, in conjunction with the genetically tractable haploids and diploids created, provide valuable resources to the community for understanding the genetic basis of phenotypic variation in yeast. This will not only be informative due to the wealth of biological information we have about yeast but will also be useful to society due to the wide use of diverse yeast in many industries.

We found strong population structure in \textit{S. cerevisiae} and significant variation in population structure and evolutionary history among different parts of the yeast genome. The incongruences in phylogeny and population structure among different genomic regions are likely due, in a large part, to mating between divergent strains. Meiotic products of such hybrids...
and their subsequent asexual competition can quickly lead to such patterns. Even without meiosis, an apparently rare event in yeast (Tsai et al. 2008), the production of beneficial aneuploids during clonal growth can occur, removing some of the chromosomes derived from the hybrid-forming strains.

We identified SNPs associated with variation in growth rate under several environmental stresses relative to a benign condition and validated a subset of these associations experimentally. These observations add much complexity to association mapping in yeast, especially if the goal is to identify causal genetic variants. For instance, only a minority of the associated SNPs were located within the causal genes validated by the hemizygosity test, suggesting that it is uncommon for a SNP identified by our association analysis to cause the observed fitness variation among the natural strains. Instead, the association analysis identified regions likely to harbor allelic variation affecting growth rate, and the reverse genetic data then more precisely located the causal genes. In addition to this problem, we found several cases where deleting a gene from a laboratory strain had no appreciable phenotypic effect, yet alternative alleles segregating in natural populations had different effects on growth rate. Because these genes are annotated as functional in the laboratory strain, our finding suggests that gain-of-function mutations relative to the laboratory strain may be involved in natural phenotypic variation and that forward genetics may sometimes identify a genetic basis that is invisible by reverse genetics of laboratory strains using gene deletion. Alternatively, these results may indicate that the effects of segregating variation depend on genetic background, i.e. epistasis. For high temperature growth, this view is consistent with previous work identifying variability in the gene-phenotype relationship across genetic backgrounds (Sinha et al. 2006; Cubillos et al. 2013). Finally, we also observed instances where retention of the assumed fitter allele resulted in lower fitness than retention of the assumed less fit allele. Because the genetic basis of high temperature growth is one of the most extensively studied traits in S. cerevisiae (Steinmetz et al. 2002; Sinha et al. 2006; Doniger et al. 2008; Parts et al. 2011; Edwards and Gifford 2012; Bloom et al. 2013; Cubillos et al. 2013), it remains to be determined the extent to which this variability in the genic basis of complex traits is observed for other phenotypes. That we observed all of the above phenomena in mapping only a single trait suggests that this key model system for understanding eukaryotic cell biology still has much to teach us about the genetic basis of phenotypic diversity. The combination of forward and reverse genetic approaches in this model system offers one way in which this complexity can begin to be unraveled.

Materials and Methods

Strains and Strain Construction

The strains sequenced in this work were obtained from the authors of two previous studies (Liti et al. 2009; Schacherer et al. 2009) and are listed in supplementary data S1, Supplementary Material online. The information of geographic location and environment of each strain is also provided when available. Most strains used are originally diploid and homothallic, and contain no tractable genetic marker, making tracking strains difficult and the maintenance of stable haploid strains, necessary for many studies, impossible. To produce a set of strains useful to the community, we adopted the approach used in the construction of the S. cerevisiae gene deletion collection to introduce drug resistance markers flanked by two unique, strain-identifying, 20-nucleotide DNA barcodes at the HO (YDL227C) locus of each strain (outlined in supplementary fig. S10, Supplementary Material online). This process simultaneously removed the strains’ ability to mating-type switch and introduced a reliable means for strain tracking. Diploid strains were transformed using the lithium acetate method (Cubillos et al. 2009) with minor alterations. The ~1 μg of transforming HO-targeting DNA contained a G418 sulfate resistance marker flanked by strain-specific barcodes and was produced by two successive polymerase chain reaction (PCR) amplifications. We first amplified the KanMX4 cassette from plasmid pFA6a-KanMX4 (Wach et al. 1994) using two 74-nucleotide primers each containing a unique 20-nucleotide barcode, the sequences necessary for its amplification (U1+U2 or D1+D2), and priming sites for the second PCR (supplementary data S8, Supplementary Material online). The second PCR used a dilution of the product of the first PCR as the template and added sequences homologous to regions upstream and downstream of HO for targeting and replacement of the locus (supplementary fig. S10, Supplementary Material online). The primers used in this PCR differed by strain to maintain lineage-specific SNPs in the region. A full list of the primers used can be found in supplementary data S8, Supplementary Material online. To ensure that the barcodes assigned to each strain are novel and maintain their compatibility with those in gene deletion strains (Giaever et al. 2002), molecular barcoded yeast (MoBY) ORF library (Ho et al. 2009), and existing technologies used to estimate barcode frequency, we employed unused barcode sequences already present on the widely used Tag4 array (Pierce et al. 2006). We confirmed successful insertion of the KanMX4 cassette by PCR and confirmed their sequence using Sanger sequencing. Although Cubillos et al. (2009) have previously produced a set of genetically tractable strains for a subset of genotypes studied, they were unsuitable for several reasons. First, the shorter barcode (6 bp vs. 20 bp) reduces the number of strains that can be confidently mixed due to potential misidentification of barcodes and therefore counts in the presence of sequencing errors. Second, the barcodes used are incompatible with the bar-seq methodology used for the gene deletion lines due to differences in flanking sequences. This precludes the simultaneous single tube determination of both wild and deletion line fitness values. Third, all strains were converted to uracil auxotrophs by deletion of URA3, possibly affecting strain growth. Together, these considerations led us to construct this new resource for the community.

From these heterozygous HO marked diploids (MATα/ MATα·HO/ho∆:Uptag-KanMX4-Downtag), stable haploid strains were obtained by sporulation on potassium acetate media followed by ascus digestion and tetrads dissection. G418 resistant colonies were identified by replication to YPD media containing 300μg/ml G418 sulfate (Gold Biotechnology, US).
Colony PCR was used to determine the mating type of individual colonies, and single MATa and MATα colonies were streaked to obtain a pure strain of each mating type. Samples were grown overnight and frozen at −80 °C in 20% glycerol for long-term storage. To allow for the easy formation of diploids between any two strains, we switched the drug resistance cassette carried by MATa and MATα strains to hygromycin B and nourseothricin, respectively (Gold Biotechnology). This was achieved by the standard LiAc method using a PCR product produced by the use of primers specific to the TEF promoter and terminator common to all three drug resistance cassettes (Wach et al. 1994; Goldstein and McCusker 1999).

Two sets of strains were treated slightly differently due to their genotypes. First, RM11 was previously made into a stable haploid strain by insertion of a KanMX4 cassette at the HO locus, resulting in the deletion of the targeting region we used in all other strains (Brem et al. 2002). To insert the appropriate barcodes into this background, unique homologous primers were used to target and replace the existing KanMX4 cassette with a HphMX4 marker amplified from plasmid pAG32 (Goldstein and McCusker 1999). Unique barcodes were then added and the cassette was switched back to KanMX4. Second, three strains (S288c, W303, and RM11) were already heterothallic haploids. After insertion of the barcoded cassette at the HO locus, these strains were transformed with plasmid pCM66, which contains a galactose inducible copy of HO and a nourseothricin drug resistance marker, to obtain strains of both mating types. After transformation, nourseothricin resistant cells were grown with galactose as the sole carbon source at 30 °C without shaking for 8 h to induce expression of HO. This allowed for mating-type switching and subsequent mother–daughter cell mating to produce diploids. Cells were then streaked for single colonies on YPD (1% yeast extract, 2% peptone, 2% glucose, and 2% agar) plates, and the ploidy of single colonies was checked by colony PCR using mating-type-specific primers. Diploid colonies were streaked for single colonies on fresh, nonselective, YPD plates and assayed for nourseothricin resistance. A single colony unable to grow in the presence of the drug, and therefore having lost the plasmid, was selected for each strain.

We attempted to produce genetically tractable strains for each of the 85 strains whose genomes we sequenced, but found some to be unamenable to our approach, either due to natural resistance to the drugs used or an inability to successfully sporulate and produce viable offspring of both mating types. The full details of all tractable strains created and the reason for missing strains are outlined in supplementary data S1, Supplementary Material online.

**Genome Sequencing**

Each of the 85 strains was streaked from frozen stocks onto YPD plates. Following 2 days of growth, a single colony was picked into 5 ml of liquid YPD media and grown to saturation (36 h at 30 °C with shaking). Cultures were centrifuged to collect cells, and DNA was extracted using standard methods. Dried DNA pellets were resuspended in 70 μl of Tris-EDTA (pH 8.0), the DNA was quantified, and the purity was assessed, before DNA storage at −80 °C. Illumina libraries were constructed using a protocol modified from a previous study (Rohland and Reich 2012). Briefly, 5 μg of genomic DNA was sheared using a Covaris S220 (duty cycle 10%, intensity 4, cycles/burst 200, time 55 s), of which 2 μg was used in library construction. To select DNA fragments of the desired size range (~400 nucleotides), we used DNA binding Magna beads to perform dual size selection. The fragments were blunt-end repaired, adapter ligated, and nick filled to repair the adapter overhangs. Finally, sequences necessary for multiplexing and cluster formation on an Illumina HiSeq2000 were added by PCR. Equal amounts of each library were combined and run across two paired-end 100-nucleotide lanes (43 strains in one lane and 42 in a second) of an Illumina HiSeq2000 at the University of Michigan DNA sequencing core.

**Read Mapping and SNP/Indel Calling**

Reads were first trimmed using Cutadapt (Martin 2011) to remove adapter sequences. Bowtie2 v2.1.0 (Langmead and Salzberg 2012) was used to map reads to the S288c reference (R64–1-1) genome under the sensitive local alignment mode, allowing up to 3 mismatches/indels per read. Pertinent statistics obtained during the mapping process are listed in supplementary data S2, Supplementary Material online. Paired reads were considered nonconcordant and discarded from further analysis if apparent mapping locations were more than 1,200 nucleotides apart or if the paired reads appeared to completely overlap one another. Paired reads were also removed from further analysis if either read was found to map ambiguously. Finally, we removed PCR duplicates by discarding all but one copy of any read pair found to map to exactly the same genomic position.

SAMtools v0.1.18 (Li et al. 2009) and VarScan v2.3.6 (Koboldt et al. 2012) were used to identify SNPs and indels within each genome. Only variants identified by both programs were used in downstream analysis. To further reduce false calls due to misalignment of reads to the reference genome, we removed variants that showed a significant strand bias (binomial P < 0.001) or invariant distance to the end of supporting reads (VDB < 0.0015) (Daneck et al. 2012). Only the most likely variant is listed for the indel and homozygous SNP lists. For the heterozygous SNP list, maximum likelihood genotype inferred by SAMtools is reported. To reduce errors in estimating allele frequencies, we used only segregating sites with reads covering the variant in each of the 85 strains except when identifying pseudogenizing variants.

**Phylogenetic Reconstruction**

We reconstructed a maximum composite likelihood neighbor-joining tree using MEGA 5.2 with all homozygous SNPs and all substitution types (Tamura et al. 2011). We allowed heterogeneous rates amongst lineages and heterogeneous rates amongst sites. Clades were identified in line with previous studies (Liti et al. 2009; Schacherer et al. 2009). To assess the strength of support for the phylogeny, we performed 1,000 bootstraps. Phylogenies of individual chromosomes were reconstructed using the same method.

**Population Structure**

To assess the population structure of the 190 strains, we used a model-based Markov Chain Monte Carlo (MCMC)
algorithm implemented in fastSTRUCTURE (Raj et al. 2014). For genome-wide population structure analysis, we randomly selected 10% of nonsingleton homozygous SNPs. One hundred runs of fastSTRUCTURE for each of $K = 2$–9 were performed, with other parameters set as the default. $K = 7$ was found to be the best. The population structure that exhibited the maximum mean likelihood was plotted using R. Finally, the population structure of each chromosome was determined using all homozygous SNPs on the specific chromosome at $K = 7$.

Linkage Disequilibrium
LD measured by $r^2$ between every pair of SNPs was calculated using custom code. We then computed the average $r^2$ for all SNP pairs with a distance in the range between $x = 99$ and $x$ nucleotides, where $x = 100, 200, 300, \ldots$, and 200,000. We computed the expected LD between unlinked SNPs by calculating the mean $r^2$ of 10,000 random pairs of SNPs located on different chromosomes. Following Schacherer et al. (2009), for each distance range, we presented the difference between an observed $r^2$ and the expected $r^2$ of unlinked SNPs in figure 1d.

Population Genomic Analysis of Natural Selection
We used only SNP sites that are dimorphic and for which the ancestral state could be unambiguously assigned in population genomic analysis. To infer SNP ancestral states, we took advantage of the published orthology information and multispecies genome sequences (Scannell et al. 2011). We used T-Coffee (Notredame et al. 2000) and the default settings in BioPerl to align the coding sequences of S. paradoxus, S. mikatae, and S. bayanus with the orthologous coding sequences of the S. cerevisiae reference sequence R64-1-1. Using these multiple-sequence alignments, we considered the states of S. paradoxus, S. mikatae, and S. bayanus for each SNP site and unambiguously assigned its ancestral state if at!least two of these outgroup species were in agreement.

We found polymorphisms that result in nonsense mutations to have much lower derived allele frequencies (DAFs) than nonsynonymous polymorphisms, which have lower DAFs than synonymous polymorphisms (supplementary fig. S3, Supplementary Material online). This pattern suggests that purifying selection against nonsense mutations is generally stronger than that against nonsynonymous mutations, which is in turn stronger than that against synonymous mutations. To examine whether purifying selection acts on synonymous mutations, especially in genes with strong codon usage bias (CUB), we measured CUB by codon-adaptation index (CAI) of yeast genes previously published (Qian et al. 2012b). We divided genes into two bins: those with CAI $> 0.6$ and those with CAI $\leq 0.6$. We found that synonymous polymorphisms in high-CAI genes tend to have lower DAFs than those in low-CAI genes (supplementary fig. S3, Supplementary Material online), supporting the hypothesis of purifying selection against synonymous mutations in genes with strong CUB. For each SNP category, we also calculated the population genetic statistics Tajima’s D (1989), Fu and Li’s $F$ (1993), and Fay and Wu’s $H$ (2000). Compared with synonymous polymorphisms, the more negative values of $D$ and $F$ for nonsynonymous and nonsense polymorphisms are consistent with the excess of rare alleles, and the less negative values of $H$ are consistent with the deficiency of common alleles (supplementary table S1, Supplementary Material online).

To assess potential positive selection at the protein level, we counted the number of synonymous polymorphisms ($P_S$), nonsynonymous polymorphisms ($P_n$), synonymous substitutions between S. cerevisiae and S. paradoxus ($D_S$) and nonsynonymous substitutions between S. cerevisiae and S. paradoxus ($D_n$) in each gene. The McDonald–Kreitman (1991) test was performed using a two-tailed Fisher’s exact test within R followed by a Bonferroni multiple-test correction. To calculate the proportion of amino acid substitutions driven by positive selection ($\alpha$) for each gene, we used S. paradoxus as an outgroup. We first determined if $D_N / P_N > D_S / P_S$. When $D_N / P_N > D_S / P_S$ we calculated $\alpha$ by $1 - D_N / P_N / (D_S / P_S)$; otherwise, we calculated $\alpha$’ by $1 - D_S / P_S / (D_N / P_N)$, which represents the fraction of nonsynonymous mutations under purifying selection. We found that the distribution of $\alpha$ is largely consistent with widespread purifying selection and relatively few instances of positive selection (supplementary fig. S4, Supplementary Material online). In addition, McDonald–Kreitman tests of individual genes failed to detect a significant signal of positive selection for any gene after Bonferroni correction. This result is consistent with a previous analysis of a smaller yeast dataset (Liti et al. 2009). McDonald–Kreitman tests suggested that 6.9% of genes are under significant purifying selection (Bonferroni corrected $P$-value $< 0.05$).

Because slightly deleterious alleles can bias the estimation of $\alpha$, we further estimated $\alpha$ using the approach proposed by Eyre-Walker and Keightley (2009). Briefly, their approach uses polymorphism data to estimate the distribution of fitness effects of new deleterious mutations (DFE) and then uses DFE to predict the number of neutral and adaptive substitutions. When using this approach with a two-epoch model, we found the estimated $\alpha$ to be $-0.11$ (supplementary table S2, Supplementary Material online), which is again consistent with the lack of signal of positive selection.

Recently, Messer and Petrov (2013) proposed a heuristic method to estimate $\alpha$ by considering $\alpha$ values for polymorphic sites with different levels of DAF. Following Messer and Petrov (2013), we applied the extended version of the McDonald–Kreitman test to all genes. Starting with all SNPs, we sequentially raised the threshold for DAF and recalculated $\alpha$ using only SNPs that pass the threshold. Using MATLAB, we applied a nonlinear least-square method to fit the data to the function of $x = a + be^{-c\alpha}$, where $x$ is the DAF threshold and $x_{\alpha}$ is the corresponding $\alpha$. We restricted $b < 0$ and $c > 0$. While the theoretical support of Messer and Petrov’s approach is lacking, this approach, surprisingly, converges on an estimated value of 0.55 for $\alpha$, suggesting that ~55% of between-species sequence divergence at nonsynonymous sites is due to positive selection (supplementary fig. S5A, Supplementary Material online). Interestingly, the plot of $\alpha$ vs. DAF showed a clear valley around intermediate DAFs.
that has not been observed previously. Suspecting the reason might be heterogeneous selection across strains, we partitioned strains based on phylogenetic clustering. We found that if we separated strains from the Wine/European cluster from all other strains and performed the same analysis on the two groups separately, the plots of $\alpha$ vs. DAF were dramatically different from each other. The extrapolated $\alpha$ values were $-0.36$ for the Wine/European cluster and 0.56 for all other strains. This difference in the estimate of $\alpha$ was not caused by sampling in general, but the specific partitions used (supplementary fig. S5b, c, Supplementary Material online). In addition, the $\alpha$ vs. DAF plot for strains not in the Wine/European cluster showed a better fit to an exponential curve than the combined analysis (adjusted $R^2 = 0.85$ without Wine/European strains vs. adjusted $R^2 = 0.69$ for all strains). Overall, these results suggest that the historical action of natural selection within these two groups has been different and that positive selection in yeast may be more common than initially expected, especially outside of domesticated wine strains. These results, however, are inconsistent with the results obtained using the method of Eyre-Walker and Keightley (2009), where no positive values are apparent when using either only strains in the Wine/European cluster or only strains outside the Wine/European cluster (supplementary table S2, Supplementary Material online).

De Novo Assembly and Identification of Non-Reference Genes

De novo genome assembly was performed using SOAPdenovo2 v2.04 (Luo et al. 2012) with $K = 51$ for all adapter-trimmed reads from each of the 85 strains we sequenced. Basic statistics of the genome assemblies are listed in supplementary data S3, Supplementary Material online. To examine the quality of the assemblies, we used BLASTn to search for the KanMX4 vector sequence present at the HO locus of 81 genomes. We then used it as an anchor to extract the strain-specific barcodes (UPTAG and DSTAG). We successfully recovered the strain-specific barcodes for all 81 genomes. To rule out false positive de novo gene calls, Exonerate v2.2.0 (Slater and Birney 2005) was used to align known genes in the reference genome of S288c to the assembled contigs. Known genes were localized onto contigs by prioritizing the Exonerate hits by (i) best hit with syntenic neighbor genes at either side, (ii) best hit with 100% query sequence coverage, (iii) hits longer than 200 nucleotides that do not overlap with better hits by more than 30 nucleotides—in case a gene is split among multiple contigs. Having identified the locations of known genes on de novo contigs, we used GeneMarkS v4.17 (Besemer et al. 2001) to perform gene predictions and compared these to the locations of known genes.

Predicted genes that showed no overlap with known gene locations were considered candidate nonreference genes. To avoid false positives caused by un-localized known genes, we used BLASTn to align the predicted genes with cDNA sequences of known genes. All predicted genes with hits covering 80% of the query, or with a >200-nucleotide region that is >90% identical with any reference gene, were removed. In order to classify the origin of remaining nonreference genes, we retrieved the best hit in the NCBI “nr” database reported by BLASTn and tBLASTx for each nonreference gene. If the best hit was in the reference genome, it was also removed unless there was a premature stop codon for the hit region in the reference genome. Finally, we removed nonreference genes with best hits to sequences derived from vectors, synthetic constructs, phages, bacterial genomes, or tandem elements (supplementary data S4, Supplementary Material online).

To access the expression levels of the nonreference genes, RNA-seq data of 23 S. cerevisiae strains generated by SOLiD were downloaded (Skelly et al. 2013). The color space RNA-seq reads from each strain were mapped to known and predicted nonreference genes using bowtie (Langmead et al. 2009), allowing up to two mismatches in color space. The best hits for each read were then used to calculate the Reads Per Kilobase of transcript per Million mapped reads (RPKM) for each gene.

Phenotypic Consequences of Deleting Non-Reference Genes

Two newly identified nonreference genes were deleted from the genetic backgrounds in which they were discovered using the G418 sulfate resistance cassette KanMX4 that was PCR-amplified from plasmid pFA6a–KanMX4 (Wach et al. 1994), followed by PCR confirmation (see supplementary data S8, Supplementary Material online for primer sequences). Non-Ref-129 was deleted from strain UWOPS87-2421, while Non-Ref-67 was deleted from strain CLIB272. Both backgrounds were MATa ho Δ::NatMX4 genotypes detailed below.

To test the phenotypic consequences of the deletion of these two genes, growth curves were obtained using a Bioscreen C (Growth Curves, USA). Non-Ref-129 deletion strain and wild-type strain were grown in Complete Supplemented Media (CSM) containing varying concentrations (0, 3, 6, 12.5, 25, 37.5, 50, 75, and 100 μg/ml) of flucanazole (Sigma Aldrich). Non-Ref-67 deletion strain and wild-type strain were also grown in CSM media but using varying concentrations (0, 1.5, 3, 6, 12, 25, 50, and 100 μM) of cantharidin. To initiate the growth, we grew 5 ml CSM cultures from frozen stocks for 24 h. Each strain was diluted 200 times into 350 μl of the appropriate drug-containing media in triplicate. Growth curves were collected at 30 °C for 60 h using the Wide-Band 420–580 nm filter every 20 min. Maximum growth rate (OD/h) was calculated from growth curve data as in a previous study (Warringer et al. 2011) and efficiency (Max OD) was calculated by taking the average of the 3rd–7th highest OD values recorded. Two-tailed t-tests were used to assess the significance of differences between genotypes. Three replications were performed for each growth curve.

Identification of Intron Losses

Based on S288c reference genome annotation, we built a sequence database containing all known exon–exon junctions up to 101 nucleotides from each side of the junction. To search for reads supporting an intron loss event, all reads were mapped to this database by Bowtie2. We required at least 95% coverage of the query read and that the read mapped
to at least 20 nucleotides in each of two adjacent exons. We further filtered ambiguous mappings on the reference genome by BLASTn with an E-value cutoff of 0.01. To search for reads supporting the presence of introns, the same procedure was conducted for all sequences annotated as exon–intron borders in the S288c reference genome. Finally, intron loss was declared in strains in which at least two reads span the exon–exon junction but no read spans the two corresponding exon–intron borders.

Identifying Potential Copy Number Variants (CNVs) and Aneuploidies
To assess gene duplication/deletion events, read pairs that were concordantly mapped to the reference genome following filtering of potential PCR duplicates were analyzed using Cufflinks v2.1.1 (Trapnell et al. 2010) to generate Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values for each coding sequence (CDS). Potential CNVs of individual genes, as well as aneuploidies and large scale duplications, were then identified by dividing each FPKM by that obtained from the same CDS in the reference strain S288c.

Simultaneous Phenotyping of Barcoded Yeast Strains
To phenotype all strains with unique barcodes, we used bar-seq (Smith et al. 2009) to estimate their relative growth rates in each of seven environments. The barcoded strains were mixed approximately equally and combined with the diploid homzygous gene deletion collection (Invitrogen 95401.H1Pool). Each nondeletion strain was present at approximately twice the initial population size of each gene deletion strain. The initial pool of strains was grown for approximately two generations in 25 ml of YPD media at 30 °C before the resulting culture, termed generation 0, was used to initiate competitions in each of the experiments. To reduce the effect of genetic drift, large populations were maintained throughout competitions with regular transfers to fresh media every 4–5 generations to maintain populations in exponential growth. Populations competed for approximately 30 generations (6 transfers) and samples were stored at −80 °C following each transfer. Following preliminary investigations, we chose to carry out in-depth analysis of the populations following the second transfer (~10 generations) in YPD at 30 °C, YPD at 40 °C, YPD + 1.25 M NaCl at 30 °C, YPD + 8% EtOH at 30 °C, YPD + 4 mM paraquat (superoxide) at 30 °C, YPD + 3 mM Hydrogen peroxide at 30 °C, and YPD + 1 mM cobalt chloride at 30 °C, respectively.

To determine the frequency of strains in the pooled population at a given time point, we extracted genomic DNA from samples using a Puregene Yeast/Bacteria DNA extraction kit (Qiagen). DNA barcodes were amplified by PCR using Accuprime pfX (Invitrogen). The primers used for barcode amplification also added sequences necessary for cluster formation and sequencing primer annealing on the Illumina platform. Because the downstream barcode is known to be missing in some deletion strains (Deutschbauer et al. 2005), only the upstream barcodes were used. Fifty base-pair single-end sequence reads were obtained using one lane of an Illumina Genome Analyzer Ix at the University of Michigan DNA Sequencing Core. The Illumina Pipeline software version 1.6 was used for base calling from the image data. Because all sequences started with the same 18 base pairs of the PCR primer region and this uniformity adversely affected base calling, we removed the first 18 sequencing cycles before base calling. We used the previously published “gene-barcode map” (Qian et al. 2012a) with the addition of our own strains’ barcode identities to assign each read to a particular strain allowing for only a single mismatch. We followed a previously published outline of bar-seq analysis (Robinson et al. 2014). We required that barcodes be represented by at least 40 total counts in populations analyzed before (generation 0) and after competition. Count numbers were then normalized using the TMM method implemented in the R package edger (Robinson et al. 2010). Because we are interested in phenotypes that are specific to a given condition, as opposed to the general fitness across environments, we computed fitness in a specific condition relative to that in YPD (the benign condition). Overall there were 79 natural S. cerevisiae strains and 4,498 deletion strains with usable data.

Let the number of reads for a genotype of interest at the beginning of the competitions be $N_0$ and the corresponding number for the reference genotype (i.e., the strain lacking HO) be $M_0$. Let the numbers of reads for the above two genotypes after competition in the benign environment of YPD at 30 °C be $N_g$ and $M_g$, respectively, and the corresponding numbers in a stressful environment of interest be $N_N$ and $M_N$, respectively. The competition lasted for $g = \sim 10$ generations in each environment. Note that the inaccuracy in $g$ does not affect comparison of relative fitness among strains because the same $g$ applies to all strains. The fitness of the genotype of interest relative to that of the reference genotype in the stressful environment, relative to that in the benign environment is $R = \left(\frac{N_g}{N_0}\right) \left(\frac{M_0}{M_g}\right)^{1/g} = \left(\frac{N_N}{N_0}\right) \left(\frac{M_0}{M_N}\right)^{1/g}$.

Two biological replicates of the competition and bar-seq were conducted in YPD at 30 °C. Based on the two estimates of $N_i/M_1$, we tested whether $N_2/M_2$ is significantly different from $N_1/M_1$ (i.e., whether $R$ is significantly different from 1), followed by FDR correction for multiple testing.

GWAS
We followed a multistep GWAS approach (Listgarten et al. 2012, 2013) outlined in supplementary fig. S13, Supplementary Material online. For each environment, we first scaled and centered the relative growth rates. We removed all sites with a minor allele frequency below 5% across the set of strains for which we had phenotype data. We then removed sites for which data were missing in >5% of strains and all mitochondrial sites. The remaining 123,121 sites were converted into a 0, 0.5 and 1 format to represent homozygous nonreference, heterozygous, and homozygous reference states in each strain, respectively.

Besides true positives, individual SNPs can be significantly associated with a phenotype because of chance or the confounding effect of population structure. Because over- or under-correction of population structure can lead to a loss of statistical power or false positives, respectively, we performed
the GWAS in a number of steps to attempt to find an appropriate balance. First, we performed an association between the normalized phenotype values and each SNP without controlling for population structure by a simple linear regression method without covariates. Second, we used this unstructured association to rank all SNPs based on their statistical significance of association. From this list, we performed a series of associations by maximum likelihood using EMMA (Kang et al. 2008). To control for population structure, we estimated a kinship matrix based on a specific set of SNPs. To define this set, we started with the 1,000 SNPs most significantly associated with the phenotype in the unstructured analysis and then successively added the next 1,000 most significant SNPs from the unstructured analysis until 120,000 SNPs were included. For each association, the genomic control factor, lambda, was calculated using gcpl2 within R (Devlin and Roeder 1999). We identified the minimal kinship set that controlled for population structure based on where lambda first hit 1, or if it failed to do so, was minimized (supplementary fig. S12, Supplementary Material online). Third, we ran an additional series of associations centered on the 3,000 SNP region identified by lambda using kinship sets in 50 SNP windows. Again, we found the smallest kinship set where lambda hit 1 or was at its minimum. Finally, we performed an association for the 500 most significant SNPs from the unstructured association. In each case, we used the estimated kinship set that was optimal for lambda, minus any SNPs within 10 kb of the focal SNP, to estimate population structure. Any variant in this final association with a P-value below 0.0001 (i.e., 5%/500) was classified as significant. For these SNPs, we identified the coding region nearest to the variant as well as its immediate neighbors. These candidate SNPs generally cover a 4–10 kb region which is approximately 4–8 times the range over which LD is seen to break down.

Experimental Validation of GWAS Results
To confirm the validity of our GWAS approach and to narrow down the genes responsible for the observed phenotypic variations within the regions surrounding the significant SNPs, we performed reciprocal hemizigosity tests (Steinmetz et al. 2002). We chose to concentrate on a subset of SNPs significantly associated with relative fitness at 40 °C. The same candidate genes were deleted from haploid MATa hoΔ::HygMX4 backgrounds that showed a high relative fitness and a low relative fitness using a KanMX4 marker. Appropriate strain hybrids were formed by mating of these deletion strains to MATz cells. Two types of hybrid were formed to allow for the relative fitness of reciprocal hemizygotes to be ascertained. The first was formed by mating of deletion-carrying haploids to MATz hoΔ::NatMX4 of the other genetic background and the second was formed by mating to MATz hoΔ::TDH3p-YFP-NatMX4 cells. This resulted in four hybrid strains for each gene to be tested, fluorescent and nonfluorescent hybrids lacking the high-fitness allele (ΔH-F and ΔH-NF) and fluorescent and nonfluorescent hybrids lacking the low-fitness allele (ΔL-F and ΔL-NF). All strains were frozen at −80 °C until needed. To perform competitive fitness assays, freezer samples were inoculated into 1 ml of YPD and grown with shaking for 24 h at 30 °C. Diploid hybrids were paired so that each competition included two hybrids carrying reciprocal deletions of the alternative alleles and identifying markers. That is, ΔH-F was paired with ΔL-NF in the first competition, and ΔL-F was paired with ΔH-NF in the second competition. Competitions were performed by mixing equal volumes of the two strains to form the t₀ population which was diluted 1,000 fold into 1.5 ml of fresh YPD media pre-warmed at 40 °C. Strains competed for 36 h, or ~10 generations, at 40 °C with shaking till saturation (t₁). All competitions were performed in 96 deep-well plates with four replicates of each competition. The frequencies of fluorescent and nonfluorescent cells within t₀ and t₁ populations were determined by flow cytometry using a BD Accuri C6 attached to a hypercyt sampling robot. After data cleaning to remove artifacts, the relative fitness of the competing strains was calculated using F = (# of ΔH cells at t₁/ # of ΔL cells at t₁) / G, where G = 10 denotes the number of generations of competition. The same experiment was repeated in YPD at 30 °C. Relative fitness was computed by dividing fitness at 40 °C by fitness at 30 °C. Five replicate competitions were performed for each set of strains and the relative fitness was calculated by averaging over the reciprocal YFP-marked strain pairs for each gene. Statistical significance was obtained by t tests.

Data Access
The yeast genome sequences determined in this work have been deposited to NCBI with the BioProject ID of PRJNA320792.

Supplementary Material
Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments
We thank Gianni Liti and Leonid Kruglyak for sharing the yeast strains that are sequenced in this study, Audrey Gasch for help accessing read data from previously sequenced strains, and two anonymous reviewers for constructive comments. This work was supported by a research grant from the U.S. National Science Foundation (MCB-1329578) to J.Z.

References
Bróco N, Tenreiro S, Viegas CA, Sá-Correia I. 1999. FLR1 gene (ORF YBR008c) is required for benomyl and methotrexate resistance in

Supplementary Material online.


