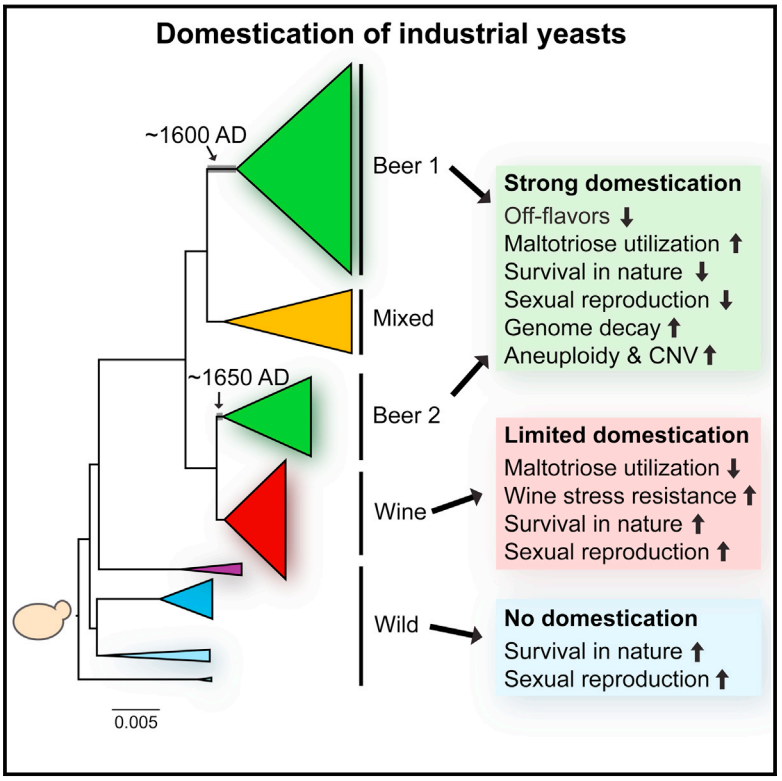
Article

Cell

Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts

Graphical Abstract



Highlights

- We sequenced and phenotyped 157 *S. cerevisiae* yeasts
- Present-day industrial yeasts originate from only a few domesticated ancestors
- Beer yeasts show strong genetic and phenotypic hallmarks of domestication
- Domestication of industrial yeasts predates microbe discovery

Authors

Brigida Gallone, Jan Steensels, Troels Prahl, ..., Guy Baele, Steven Maere, Kevin J. Verstrepen

Correspondence

steven.maere@psb.vib-ugent.be (S.M.), kevin.verstrepen@biw.vib-kuleuven.be (K.J.V.)

In Brief

The history and domestication of yeast used for making beer and other types of alcohol are revealed through genomic and phenotypic analyses.

Gallone et al., 2016, Cell *166*, 1397–1410
September 8, 2016
http://dx.doi.org/10.1016/j.cell.2016.08.020

CellPress

Article

Cell

Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts

Brigida Gallone,[1,2,3,4,11] Jan Steensels,[1,2,11] Troels Prahl,[5] Leah Soriaga,[6] Veerle Saels,[1,2] Beatriz Herrera-Malaver,[1,2] Adriaan Merlevede,[1,2] Miguel Roncoroni,[1,2] Karin Voordeckers,[1,2] Loren Miraglia,[8] Clotilde Teiling,[9] Brian Steffy,[9] Maryann Taylor,[10] Ariel Schwartz,[6] Toby Richardson,[6] Christopher White,[5] Guy Baele,[7] Steven Maere,[3,4,*] and Kevin J. Verstrepen[1,2,12,*]

[1]Laboratory for Genetics and Genomics, Centre of Microbial and Plant Genetics (CMPG), KU Leuven, Kasteelpark Arenberg 22, 3001 Leuven, Belgium
[2]Laboratory for Systems Biology, VIB, Bio-Incubator, Gaston Geenslaan 1, 3001 Leuven, Belgium
[3]Department of Plant Systems Biology, VIB, 9052 Gent, Belgium
[4]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Gent, Belgium
[5]White Labs, 9495 Candida Street, San Diego, CA 92126, USA
[6]Synthetic Genomics, 11149 North Torrey Pines Road, La Jolla, CA 92037, USA
[7]Department of Microbiology and Immunology, Rega Institute, KU Leuven, 3000 Leuven, Belgium
[8]Encinitas Brewing Science, 141 Rodney Avenue, Encinitas, CA 92024, USA
[9]Illumina, 5200 Illumina Way, San Diego, CA 92122, USA
[10]Biological & Popular Culture (BioPop), 2205 Faraday Avenue, Suite E, Carlsbad, CA 92008, USA
[11]Co-first author
[12]Lead Contact
*Correspondence: steven.maere@psb.vib-ugent.be (S.M.), kevin.verstrepen@biw.vib-kuleuven.be (K.J.V.)
http://dx.doi.org/10.1016/j.cell.2016.08.020

SUMMARY

Whereas domestication of livestock, pets, and crops is well documented, it is still unclear to what extent microbes associated with the production of food have also undergone human selection and where the plethora of industrial strains originates from. Here, we present the genomes and phenomes of 157 industrial *Saccharomyces cerevisiae* yeasts. Our analyses reveal that today's industrial yeasts can be divided into five sublineages that are genetically and phenotypically separated from wild strains and originate from only a few ancestors through complex patterns of domestication and local divergence. Large-scale phenotyping and genome analysis further show strong industry-specific selection for stress tolerance, sugar utilization, and flavor production, while the sexual cycle and other phenotypes related to survival in nature show decay, particularly in beer yeasts. Together, these results shed light on the origins, evolutionary history, and phenotypic diversity of industrial yeasts and provide a resource for further selection of superior strains.

INTRODUCTION

Since prehistoric times, humans have exploited the capacity of the common baker's yeast *Saccharomyces cerevisiae* to convert sugars into ethanol and desirable flavor compounds to obtain foods and beverages with a prolonged shelf-life, enriched sensorial palate, improved digestibility, and an euphoriant effect due to the presence of ethanol (Michel et al., 1992; Steensels and Verstrepen, 2014). Whereas the use of pure cultures started well after the pioneering work of Pasteur and Hansen in the 19th century, early brewers, winemakers, and bakers had already learned that inoculating unfermented foods with a small portion of fermented product resulted in fast and more predictable fermentations. This so-called "backslopping" might have resulted in yeast lineages that grew continuously in these man-made environments and lost contact with their natural niches, providing a perfect setting for domestication. However, strong evidence for this hypothesis is still missing and it remains unclear whether industrial yeast diversity is shaped by selection and niche adaptation (domestication) or neutral divergence caused by geographic isolation and limited dispersal (Goddard and Greig, 2015; Warringer et al., 2011).

Domestication is defined as human selection and breeding of wild species to obtain cultivated variants that thrive in man-made environments, but behave suboptimally in nature. Typical signs of domestication, including genome decay, polyploidy, chromosomal rearrangements, gene duplications, and phenotypes resulting from human-driven selection, have been reported in crops, livestock, and pets (Driscoll et al., 2009; Purugganan and Fuller, 2009). Several studies have recently investigated the *S. cerevisiae* population by sequencing the genomes of hundreds of different strains, providing a first glimpse of the complex evolution of this species (Almeida et al., 2015; Borneman et al., 2011, 2016; Liti et al., 2009; Magwene et al., 2011; Schacherer et al., 2009; Strope et al., 2015). However, most of these studies focused primarily on yeasts from wild and clinical habitats and often include only a limited set of industrial strains, mainly originating from wine. Moreover, most studies use haploid derivatives instead of natural strains and can therefore not explore typical patterns of domestication like polyploidy, aneuploidy,

and heterozygosity. The use of haploids also excludes a large fraction of industrial *S. cerevisiae* strains that have lost the ability to sporulate, such as the vast majority of beer yeasts. Nevertheless, some studies already revealed signs of domestication in wine strains, such as an increased resistance to copper (present in grapevine pesticides) and sulfite (used as a preservative in wine) (Pérez-Ortín et al., 2002; Warringer et al., 2011). An in-depth investigation of strains originating from other industrial niches is still lacking.

Here, we describe the high-quality sequencing, de novo assembly, annotation, and extensive phenotyping of 157 *S. cerevisiae* strains used for the industrial production of beer, wine, bread, spirits, saké, and bioethanol, in their natural ploidy. Our data reveal that industrial yeasts are genetically and phenotypically distinct from wild strains and stem from only a limited set of ancestral strains that have been adapting to man-made environments. They further diversified into five clades: one including Asian strains such as saké yeasts, one mostly containing wine yeasts, a mixed clade that contains bread and other yeasts, and two separate families of beer yeasts. While most clades lack strong geographical substructure, one of the beer clades contains geographically isolated subgroups of strains used in continental Europe (Belgium/Germany), the United Kingdom, and a recent sublineage of United States beer yeasts that diverged from the British subclade during colonization. Interestingly, these beer yeast lineages exhibit clear and profound hallmarks of domestication, more so than the other lineages. The shift from variable, complex, and often harsh environments encountered in nature to more stable and nutrient-rich beer medium favored specialized adaptations in beer yeasts, but also led to genome decay, aneuploidy, and loss of a functional sexual cycle. Specifically, we find evidence for active human selection, demonstrated by convergent evolution for efficient fermentation of beer-specific carbon sources, mainly through mutations and duplications of the *MAL* (maltose) genes, as well as nonsense mutations in *PAD1* and *FDC1*, which are involved in the production of 4-vinyl guaiacol (4-VG), an undesirable off-flavor in beer. Our results further suggest that beer yeast domestication was initiated hundreds of years ago, well after the first reported beer production, but before the discovery of microbes. Together, our results reveal how today's industrial yeasts are the outcome of centuries of human domestication and provide a new resource for further selection and breeding of superior variants.

RESULTS

Niche and Geography Drive Diversification

To examine the evolutionary history of industrial yeasts, we sequenced the genomes of 157 *S. cerevisiae* isolates originating from various sources in their natural ploidy to a median coverage of 135× (min = 26×, max = 403×) (for details on data analysis, see STAR Methods). This collection includes 102 industrial beer strains, 19 wine strains, 11 spirit strains, 7 saké strains, 7 strains isolated from spontaneous fermentations, 5 bioethanol strains, 4 bread strains, and 2 laboratory strains (Table S1). Interestingly, ten of these *S. cerevisiae* beer strains are used for commercial production of lager beers, which were believed to be exclusively produced by strains of the genetically related *Saccharomyces pastorianus*. After de novo assembly of each of the genomes, we inferred a maximum-likelihood phylogenetic tree based on codon alignments for 2,020 concatenated single-copy nuclear genes shared by each of the 157 isolates and the outgroup species *Saccharomyces paradoxus* (Figure S1A). Additionally, we included a representative set of 24 previously sequenced strains belonging to the main established lineages of the *S. cerevisiae* phylogeny (Liti et al., 2009; Strope et al., 2015), extending the number of strains to 181 (Figure 1A). Trees constructed from the original and extended datasets are congruent and show five main lineages that contain the majority of industrial yeasts: Wine (bootstrap support 100%), Beer 1 (86%), Beer 2 (56%), Asia (100%), and a Mixed lineage (99%) containing yeasts used in different industries. Three of these lineages (Beer 1, Beer 2, and Mixed) were not previously described.

Next, we studied the population structure in a filtered set of 53,929 polymorphic sites accounting for 2,454,052 SNPs across all strains, using the Bayesian model-based clustering approach implemented in fastStructure (Raj et al., 2014) (Figures 1B and S1B). This analysis yields a population structure that is highly consistent with the major lineages defined in the phylogeny and identifies mosaicism in 17% of the strains (in which the estimated ancestry $Q < 0.8$ for $K = 8$ ancestral populations). The population structure is further supported by a principal component analysis (PCA) on the same SNP data (Figure 1C).

Further analysis of the phylogeny and population structure reveals that the evolutionary divergence of industrial yeasts is shaped by both their industrial application and geographical origin. First, most yeasts cluster together according to the industry in which they are used and are clearly separated from the wild or clinical yeasts that have previously been sequenced. This was further confirmed by constructing a larger phylogeny, based on nine genomic regions, that includes the vast majority of all sequenced *S. cerevisiae* strains, 450 isolates in total (Figure S1C; Table S2). Wine and saké yeasts cluster in the previously identified Wine and Asia lineages (Liti et al., 2009). The majority of beer yeasts (85.3%) are found in two main lineages (Beer 1 and Beer 2) that are only distantly related. The Mixed clade harbors 7.8% of all beer strains (most of which are atypical beer yeasts that are used for bottle refermentation of strong Belgian ales) and contains all bread strains. Interestingly, spirit strains lack this clear phylogenetic relationship, as they are highly mosaic and scattered throughout the tree, suggesting that these strains might be the result of breeding by modern-day yeast companies that sell yeasts for spirits production. Moreover, because spirit yeasts are typically not re-used after fermentation, they likely had less opportunity to diverge into a separate clade.

Within and between the lineages, we also observed geographical patterns. For example, most saké yeasts form a monophyletic group and cluster together with wild isolates and bioethanol strains from China, while South American bioethanol strains are closely related to strains used to produce cachaça, a Brazilian sugarcane spirit. Moreover, the Beer 1 clade consists of three separate subpopulations, each reflecting geographically distinct groups: Belgium/Germany, Britain, and the United States. The absence of genetic admixture among these subpopulations indicates that these strains diverged allopatrically after the initial split

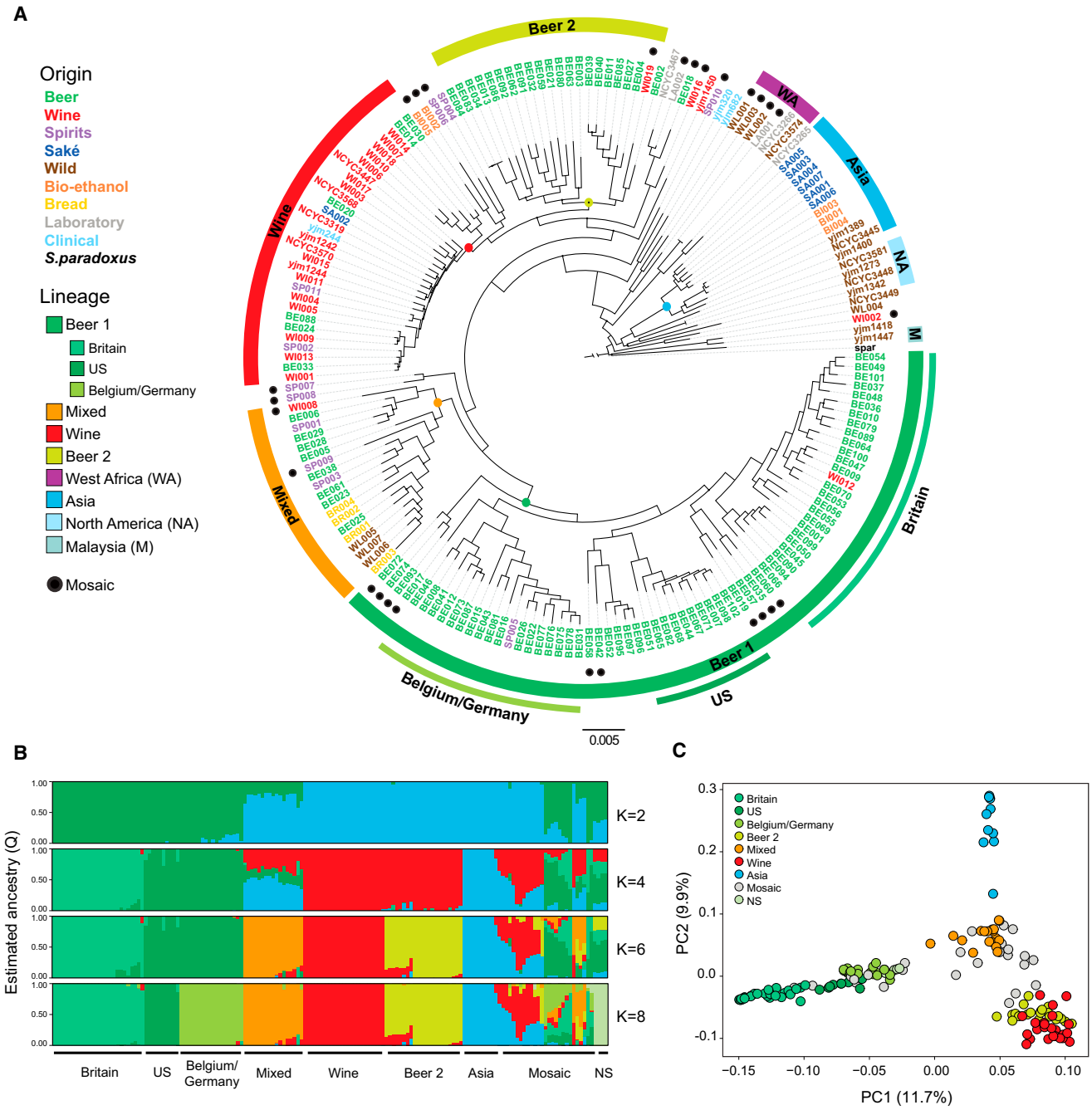


Figure 1. Phylogeny and Population Structure of Industrial *S. cerevisiae* Strains

(A) Maximum likelihood phylogenetic tree of all *S. cerevisiae* strains sequenced in this project supplemented with a representative set of 24 previously sequenced strains (Liti et al., 2009; Strope et al., 2015) and using *Saccharomyces paradoxus* as an outgroup. Black dots on nodes indicate bootstrap support values <70%. Color codes indicate origin (names) and lineage (circular bands). The basal splits of the five industrial lineages are indicated with colored dots. Mosaic strains identified in this study are indicated with black dots next to the strain codes. Branch lengths reflect the average number of substitutions per site. Scale bar, 0.005 substitutions per site.

(B) Population structure identified in the 157 surveyed strains. The vertical axis depicts the fractional representation of resolved populations (colors) within each strain (horizontal axis, strains listed in Figure S1C) for K = 2, 4, 6, and 8 assumed ancestral populations (where K = 8 maximizes the marginal likelihood and best explains the data structure). Mosaic strains (i.e., strains that possess <80% ancestry from a single population) are visualized as a separate group.

(C) Principal component projection, using the same set of SNPs as in Figure 1B. Colors represent different populations. WA, West Africa; NA, North America; M, Malaysia; NS, not specified.

See also Figure S1 and Tables S1, S2, and S8.

Table 1. Genetic Diversity within Each Subpopulation of Industrial *S. cerevisiae* Strains

Subpopulation	Number of Strains	Analyzed Sites	Segregating Sites	π	Θ_w
Britain	26	12,018,937	101,881	3.13E-03	1.88E-03
United States	10	11,973,239	72,559	2.31E-03	1.72E-03
Belgium/Germany	18	12,017,007	108,560	3.12E-03	2.19E-03
Mixed	17	12,043,532	132,188	4.35E-03	2.69E-03
Wine	24	12,052,956	114,133	1.59E-03	2.15E-03
Beer 2	21	12,063,361	142,745	2.95E-03	2.77E-03
Asia	10	12,035,745	99,879	2.39E-03	2.36E-03

The number of strains per subpopulation, the amount of analyzed and segregating sites, as well as nucleotide diversity (π) and population mutation rate (Watterson's θ, θ_w) are indicated.

(Figure 1B). Moreover, the high nucleotide diversity within each of the Beer 1 sublineages exceeds that within the Wine population, suggesting that the split did not happen recently (Table 1). Compared to Beer 1, Beer 2 is more closely related to the Wine lineage and includes 20.6% of all brewing strains. However, in contrast to the Beer 1 group, the Beer 2 lineage lacks geographic structure and contains yeasts originating from Belgium, the United Kingdom, the United States, Germany, and Eastern Europe. The presence of two major genetically distinct sources of beer yeasts hints toward two independent European domestication events, one of which is at the origin of both the Wine and Beer 2 clade.

Remarkable Structural Variation in Beer Yeasts

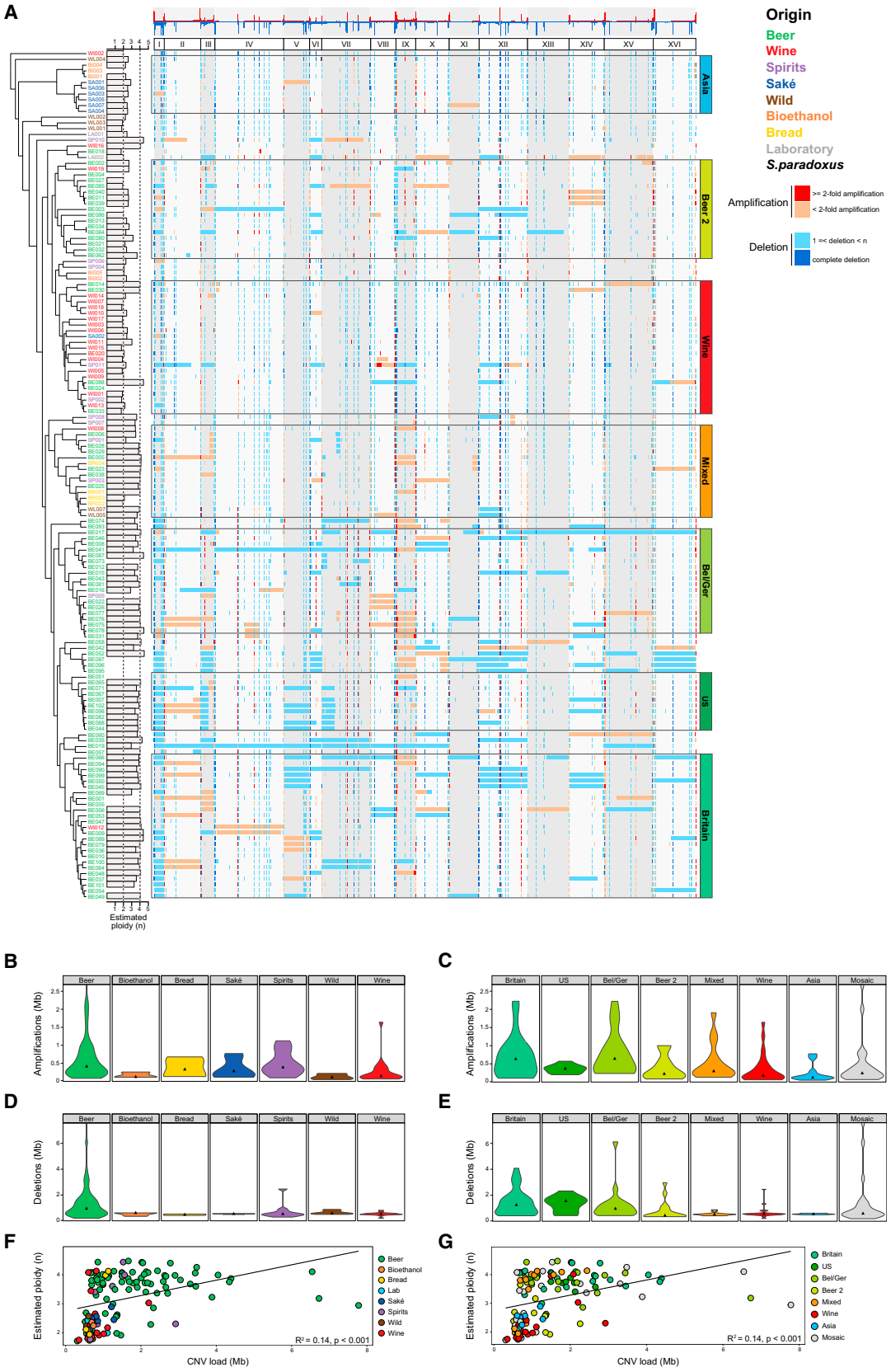
Variation in genome structure, such as polyploidy, aneuploidy, large segmental duplications, and copy-number variations (CNVs), have repeatedly been found in association with domestication and adaptation to specific niches in experimentally evolved microbes (Bergström et al., 2014; Borneman et al., 2011; Dunham et al., 2002; Dunn et al., 2012; Pavelka et al., 2010; Rancati et al., 2008; Selmecki et al., 2009; Voordeckers et al., 2015) and in association with domestication of higher organisms (Purugganan and Fuller, 2009).

Sequencing the yeast strains in their natural ploidy allowed analysis of gross chromosomal rearrangements and aneuploidies (Figure 2A). We detected a staggering 15,288 deletion and amplification events across all strains, covering on average 1.57 Mb per strain. The size of the regions ranges from complete chromosomes (resulting in aneuploidies) to small local variations of a few kilobases (kb), all of which we will refer to as "CNVs." The extent of deletions significantly exceeds that of amplifications, respectively 1.07 Mb and 0.50 Mb on average per strain (2.15-fold difference, Wilcoxon signed rank test, $p < 0.001$). We observed significant variation among strains originating from different industries in the total frequency of CNV events (ANOVA F test, $p < 0.001$) and the fraction of the genome affected (ANOVA F-test, $p < 0.001$) (Figure S2). Pairwise comparisons of subpopulations and industries show no significant differences in the load of amplifications between strains from different industries or subpopulations, but we detected significant differences in the load of deletions between strains from the wine (median = 0.51 Mb) and beer (median = 0.94 Mb) industry (Tukey honest significant difference [HSD], $p < 0.05$) (Figures 2B–2E). This high incidence of CNV in beer strains goes together with a high incidence of polyploidy and aneuploidy ($R^2 \sim 0.14$, $p < 0.001$; average genome content of 3.52, SD = 0.67, Figures 2A, 2F, and 2G), which is linked to extensive chromosomal loss and general genome instability (Sheltzer et al., 2011).

CNVs are not uniformly spread across the genome. Considering subtelomere lengths of 33 kb (Brown et al., 2010), on average 39.7% of subtelomeric nucleotide positions are affected by CNV events compared to 9.54% of non-subtelomeric nucleotide positions (4.1-fold difference, Wilcoxon signed-rank test, $p < 0.001$). However, not all subtelomeres are equally prone to CNV: most variability is detected in ChrI, ChrVII, ChrVIII, ChrIX, ChrX, ChrXII, ChrXV, and ChrXVI (Figure 2A). Gene ontology (GO) enrichment analysis reveals that genes involved in nitrogen and carbon metabolism, ion transport, and flocculation are most heavily influenced by CNVs (Table S3), which is in line with previous results (Bergström et al., 2014; Dunn et al., 2012). Interestingly, some CNVs seem linked to specific environments (Table S4), suggesting that CNVs may underlie niche adaptation. For example, many genes involved in uptake and breakdown of maltose (present in saké medium, main carbon source in beer, but absent from grape must) are amplified in beer and saké-related subpopulations, while they are often lost in strains from the Wine subpopulation (false discovery rate [FDR] q value < 0.001).

Relaxed Selection on Sex and Survival in Nature

Apart from selection for industrial traits, domestication is also characterized by relaxed selection and potential loss of costly traits that are not beneficial in the man-made environment (Driscoll et al., 2009). In order to chart the phenome of our collection and investigate signs of selection for some traits and loss of others, 82 phenotypes, such as aroma production, sporulation characteristics, and tolerance to osmolytes, acids, ethanol, and low and high temperatures, were measured in all strains (Figures 3A and S3; Table S5). Hierarchical clustering of the phenotypes resolves the main phylogenetic lineages and reveals a moderate correlation between genotype and phenotype distances between strains (Spearman correlation ~0.33), which is further increased (Spearman correlation ~0.36) when mosaic strains, for which genetic distance has no straightforward evolutionary interpretation, are omitted (Figure 3A). Moreover, the clustering splits the collection into two main phenotypic subgroups: one largely overlapping with the Beer 1 clade that contains the majority of the Belgium/Germany, United States, and



(legend on next page)

Britain beer yeasts as well as mosaic strains containing major genome fractions of these subpopulations and a second one where the remaining genetic subpopulations are strongly over-represented (Fisher's exact test, Bonferroni corrected $p < 0.001$). Overall, strains from the Beer 1 clade perform poorly in general stress conditions that are not usually encountered in the brewing environment (Figure S3; Table S6). In contrast, strains from the Wine subpopulation show superior performance in general stress conditions, which likely reflects the high-sugar and high-alcohol environments encountered in wine-making, as well as survival in potentially nutrient-poor and harsh natural environments in between the grape harvest seasons.

Saccharomyces cerevisiae is a facultative sexual organism. While its main mode of reproduction is clonal, sporadic sporulation can help to survive periods of stress (Briza et al., 1990). It has been shown that in yeast, sexual reproduction is beneficial when adapting to new, harsh niches, but plays a lesser role in more favorable environments (Goddard et al., 2005; McDonald et al., 2016). Our data show that there are large systematic differences in the reproductive lifestyle of yeasts inhabiting different industrial niches: 44.4% of the Beer 1 population is obligate asexual, while this trait ranges between 0% and 21% in the other populations (Figure 4A) and is absent in wild strains. Furthermore, over 80% of the non-mosaic Beer 1 strains that are able to sporulate show little or no spore viability (Figure 4B). Additionally, beer yeast lineages generally show a high level of heterozygosity, especially Beer 1. Compared to the Wine clade for example, strains from the Beer 1 and Beer 2 clade have on average 5.10-fold (Tukey HSD, $p < 0.001$) and 2.04-fold (Tukey HSD, $p = 0.06$) more heterozygous sites, respectively (Figures 4C, S4A, and S4B). The lack of genetic admixture suggests that this heterozygosity was acquired during long periods of asexual reproduction, rather than through outbreeding. Further analysis of the correlation between sexual lifestyle and genome structure shows that spore viability is weakly anticorrelated with the heterozygosity level ($R^2 \sim 0.17$; $p < 0.001$) and the fraction of the genome associated with large (>20 kb) amplifications and deletions ($R^2 \sim 0.16$; $p < 0.001$), while sporulation efficiency is only significantly anticorrelated with the latter ($R^2 \sim 0.19$; $p < 0.001$) (Figures 4D–4G).

Together, this indicates that the genome of beer yeasts, but not wine yeasts, show signs of decay and loss of survival skills outside a specific man-made environment, probably caused by their long (estimated >75,000 generations) and uninterrupted growth in rich medium.

Selection for Industrial Phenotypes

A key hallmark of domestication is phenotypic adaptation to artificial, man-made niches and accentuation of traits desirable for humans. Phenotypic evaluation of the strains for industrially relevant traits (including aroma production, ethanol production, and fermentation performance) shows that many strains harbor phenotypic signatures linked to their industrial application. The ability to accumulate high concentrations of ethanol, for example, seems tightly linked to industrial niche. Beer 1 strains typically generate only 7.5%–10% v/v of ethanol, while strains used for the production of high-alcohol products like saké, spirits, wine, and especially bioethanol, can produce up to 14.5% v/v (Figure 3B; Table S6).

With the exception of a few wine yeast characteristics (see earlier) (Figures 3C and 3D), it remains unclear whether genetic and phenotypic variation between *S. cerevisiae* lineages is primarily caused by human-driven selection and domestication, or if neutral genetic drift or non-human selection are involved. To assess this further, we compared the phenotypic behavior of different subpopulations for two industrially relevant traits for which the genetic underpinnings are largely known, namely maltotriose fermentation and the production of 4-vinyl guaiacol (4-VG), the main compound responsible for phenolic off-flavors (POF). Beer yeasts show a significantly higher capacity to metabolize maltotriose, a carbon source specifically found in beer medium (Figure 3E; Table S6). Efficient utilization of maltotriose correlates with the presence of a specific allele (*AGT1*) of the sugar transporter *MAL11*, known to show high affinity for maltotriose (phenotypic variability explained by SNPs in *MAL11* ~77.40%, SE 0.5%). This allele is only present in Beer 1 subpopulations and some mosaic strains, while the complete *MAL1* locus (including the *MAL11* gene) is absent in the Wine subpopulation (Table S7). Interestingly, strains of the Beer 2 subpopulation are generally able to ferment maltotriose but contain various frameshift mutations in *MAL11* and show a reduced CNV for the complete *MAL1* locus, suggesting that other, yet unknown mechanisms facilitate maltotriose uptake in this lineage, and maltotriose metabolism evolved convergently in the Beer 1 and Beer 2 lineages.

Yeasts used for the production of alcoholic beverages ideally should not produce undesirable aromas. Although tolerated in some specialty beers, the presence of 4-VG, a compound with a spicy, clove-like aroma, is generally undesired in saké, wine, and most beer styles. Two genes, phenylacrylic acid decarboxylase (*PAD1*) and ferulic acid decarboxylase (*FDC1*), both

Figure 2. Ploidy and Copy-Number Variation in Industrial *S. cerevisiae* Strains

(A) Genome-wide visualization of copy-number variation (CNV) profiles, with the aggregate profile across all strains depicted on the top. Estimates for the nominal ploidy (n) values of the strains are represented by a bar chart next to the strain codes. Heat map colors reflect amplification (red shades) or deletion (blue shades) of genomic fragments. A distinction is made between completely deleted fragments (dark blue) and fragments of which at least one copy is still present (light blue). Similarly, highly amplified fragments (copy number ≥2-fold the basal ploidy) are depicted in dark red, while low and moderately amplified fragments (copy number <2-fold the basal ploidy) are depicted in orange. For strains with no estimated ploidy available, colors are only indicative of the presence of amplifications (orange) or deletions (light blue). Roman numbers indicate chromosome number. Strains are clustered according to their genetic relatedness as determined in Figure S1A. Origin (name colors) and population (colored rectangles) are indicated on the figure.

(B–E) Violin plots describing the density of amplifications and deletions across different industries and subpopulations. Triangles indicate the median within each group.

(F and G) Correlations between levels of CNV load (Mb) and estimated ploidy (n), by industry and subpopulations.

See also Figure S2 and Tables S3 and S4.

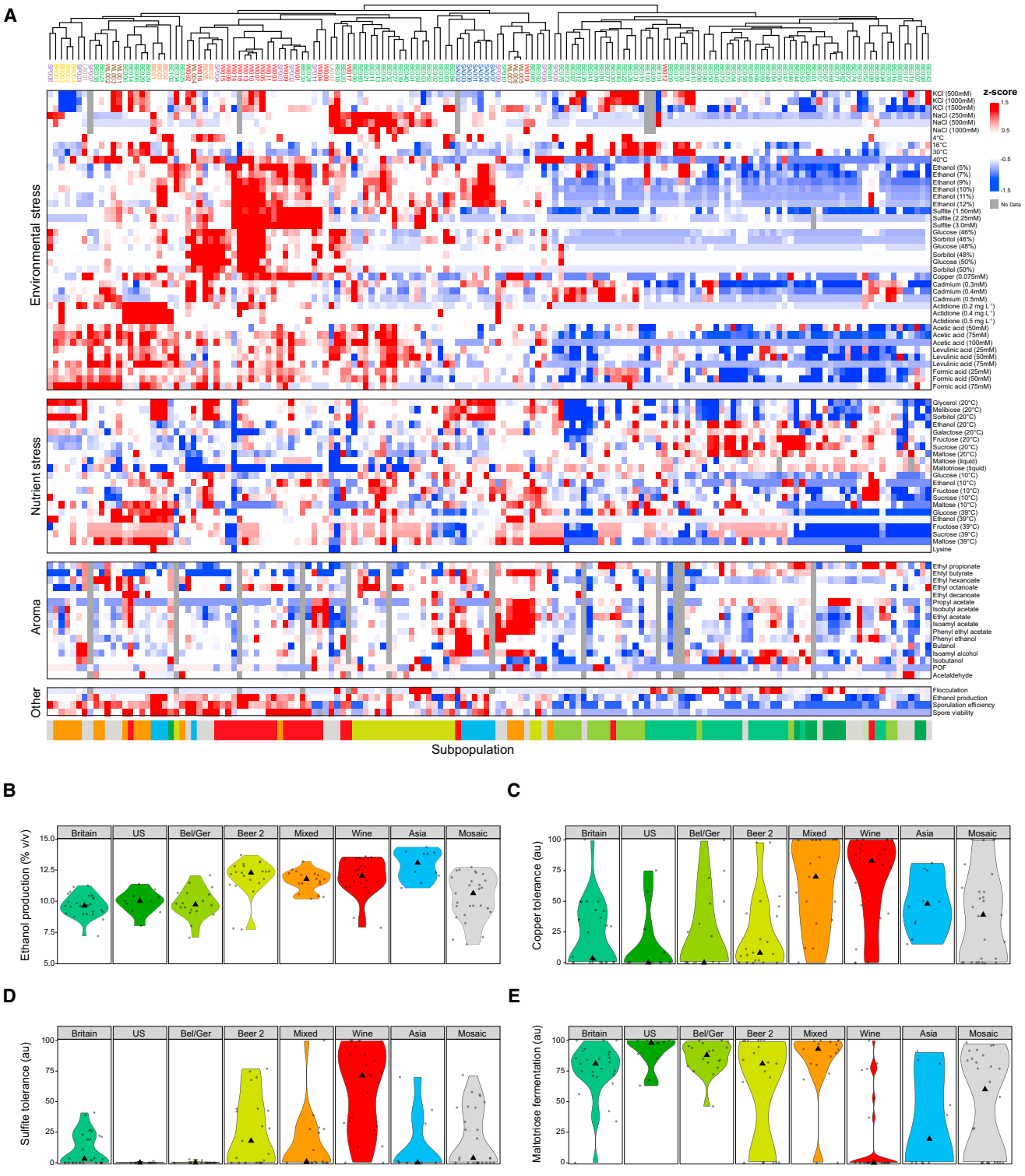


Figure 3. Trait Variation of Industrial *S. cerevisiae* Strains

(A) Heat map representation of phenotypic diversity within industrial *S. cerevisiae* strains. Phenotypic values are calculated as *Z* scores (normalized values) and colored according to the scale on the right. Missing values are represented by gray shadings. Strains are hierarchically clustered based on phenotypic behavior. Strain names are colored according to geographical origin, as in Figure 1A. The corresponding subpopulation of each strain is indicated by the colored bar below the figure, according to the color code of Figure 1B.

(B) Ethanol production (depicted as % v/v^{-1}) of all strains from different subpopulations in fermentation medium containing 35% glucose.

(legend continued on next page)

located as a cluster in the subtelomeric region of ChrIV, control 4-VG production. *PAD1* encodes a flavin prenyltransferase that catalyzes the formation of a flavin-derived cofactor, which is required by Fdc1 for decarboxylation of the precursor ferulic acid (White et al., 2015). Pad1 and Fdc1 help to detoxify phenylacrylic acids found in plant cell walls (Mukai et al., 2010). Therefore, it would be expected that, unless there is counterselection, activity of these genes is retained. Interestingly, phenotypic profiling reveals that many industrial strains have lost the ability to produce 4-VG, while it is generally retained in wild strains, as well as in bakery and bioethanol strains (Figure 5B). In these cases, 4-VG production is likely less detrimental, either because the flavor disappears during baking, or the product is not destined for consumption. Sequence analysis shows that many industrial strains, especially beer and saké strains, acquired loss-of-function mutations (SNPs and/or frameshift InDels) in *PAD1* and/or *FDC1*, while this was never observed in strains from natural environments or bioethanol production (Figure 5A). Moreover, different sublineages acquired different disruptive mutations, hinting to the presence of diverse convergent adaptive strategies in response to human selection against 4-VG production.

To investigate the origin and the maintenance of the phenotypic diversity in 4-VG production, we used Bayesian inference to reconstruct the ancestral phenotypic state in the two key genes *PAD1* and *FDC1* using BEAST (Drummond et al., 2012) (Figure 5C). Shifts from 4-VG$^+$ to 4-VG$^-$ and vice versa occurred frequently after the initial split from *S. paradoxus*. In both the *PAD1* and the *FDC1* trees, an early subclade containing most Beer 1 strains acquired loss-of-function mutations at the base of the clade, suggesting that already very early during domestication of the Beer 1 lineage, a 4-VG$^-$ variant was derived from the 4-VG$^+$ ancestor. Several other loss- and gain-of-function mutations occurred across both trees, most notably the loss-of-function mutation in *FDC1* of the Asian saké (but not bioethanol) strains.

Interestingly, a strong incongruence between single gene trees and the strain phylogeny is present for three beer strains used in the production of German Hefeweizen beers (BE072, BE074, and BE093). Hefeweizen (wheat) beer is a traditional German beer style and one of the few styles where a high 4-VG level is desirable because it contributes to the typical smoky, spicy aroma of these beers. Phylogenetically, Hefeweizen yeasts cluster within the Beer 1 lineage, but they are shown to be highly mosaic, containing genomic fragments of all three Beer 1 subclades (mainly from Belgium/Germany). Only a small fraction (~8%–13%) of the genome originates from the Wine subpopulation, but this fraction includes the subtelomeric region of ChrIV, containing a functional *PAD1* and *FDC1* allele. This suggests that hybridization between different domesticated subpopulations yielded variants combining the typical traits of beer yeasts, including maltotriose fermentation, with a particular trait from a wine strain (4-VG production) that is only desirable in special beer styles.

Creating Superior Hybrid Yeasts through Marker-Assisted Breeding

Apart from yielding insight into the origins of today's industrial yeasts, our results also open new routes for the creation of new superior strains. The availability of genomic data and the increasing number of polymorphisms that are known to contribute to industrially relevant phenotypes enables rapid DNA-based selection of superior segregants and hybrids in large-scale breeding schemes. Such marker-assisted breeding is already intensively used for crop and livestock breeding, because it circumvents labor-intensive and time-consuming phenotyping. As proof-of-concept, we combined our genomic and phenotypic data to obtain new hybrids with altered aromatic properties using marker-assisted breeding. Specifically, a 4-VG producing beer strain harboring a heterozygous loss-of-function mutation in *FDC1* (strain BE027) was selected and sporulated to obtain segregants. Next, the *FDC1* allele of the segregants was genotyped using mismatch PCR. Two segregants, one harboring the loss-of-function allele and one harboring the functional allele, were crossed with segregants of SA005, an Asian saké strain with a homozygous non-functional *FDC1* allele, resulting in hybrids with good beer fermentation characteristics but drastically different aroma profiles (4-VG$^+$ versus 4-VG$^-$) that suit specific beer styles (Figure 5D).

Domestication Predates Microbe Discovery

Despite its wide use in industry and as a model organism, little is known about the ecology and evolutionary history of *S. cerevisiae*. Moreover, because early brewers, winemakers, and bakers were unaware of the existence of yeast, there is no record of how yeasts made their way into these processes, nor how yeasts were propagated and shared. As a result, it has proven difficult to estimate when specific industrial lineages originated. Moreover, current demographic and molecular clock models of *S. cerevisiae* employ the experimentally determined mutation rate of the haploid lab strain S288c in rich growth medium (Lynch et al., 2008), while it is known that the mutation rate is heavily influenced by the genetic background (Filteau et al., 2015), ploidy (Sheltzer et al., 2011), growth speed (van Dijk et al., 2015), and environmental stress (Voordeckers et al., 2015), factors that are likely very different for industrial, wild, and lab yeasts. However, our dataset, and specifically the Beer 1 clade, provides a strong tool for dating beer yeast divergence. First, given the absence of a functional sexual cycle and lack of admixture, exclusively clonal reproduction can be assumed. Second, our data show that United States beer yeasts are related closest to European beer yeasts, suggesting that they were imported from Europe during colonization, rather than stemming from indigenous wild United States yeasts. More specifically, United States beer yeasts seem phylogenetically most closely related to British beer yeasts (Figure 1A), which is

(C) Growth of all strains from different subpopulations on medium supplemented with 0.075 mM copper, relative to growth on medium without copper.
(D) Growth of all strains from different subpopulations on medium supplemented with 2.25 mM sulfite, relative to growth on medium without sulfite.
(E) Growth of all strains from different subpopulations in medium containing 1% w v^{-1} maltotriose as the sole carbon source, relative to growth on medium with 1% w/v^{-1} glucose. au, arbitrary units; Bel/Ger, Belgium/Germany.
See also Figure S3 and Tables S5, S6, and S7.

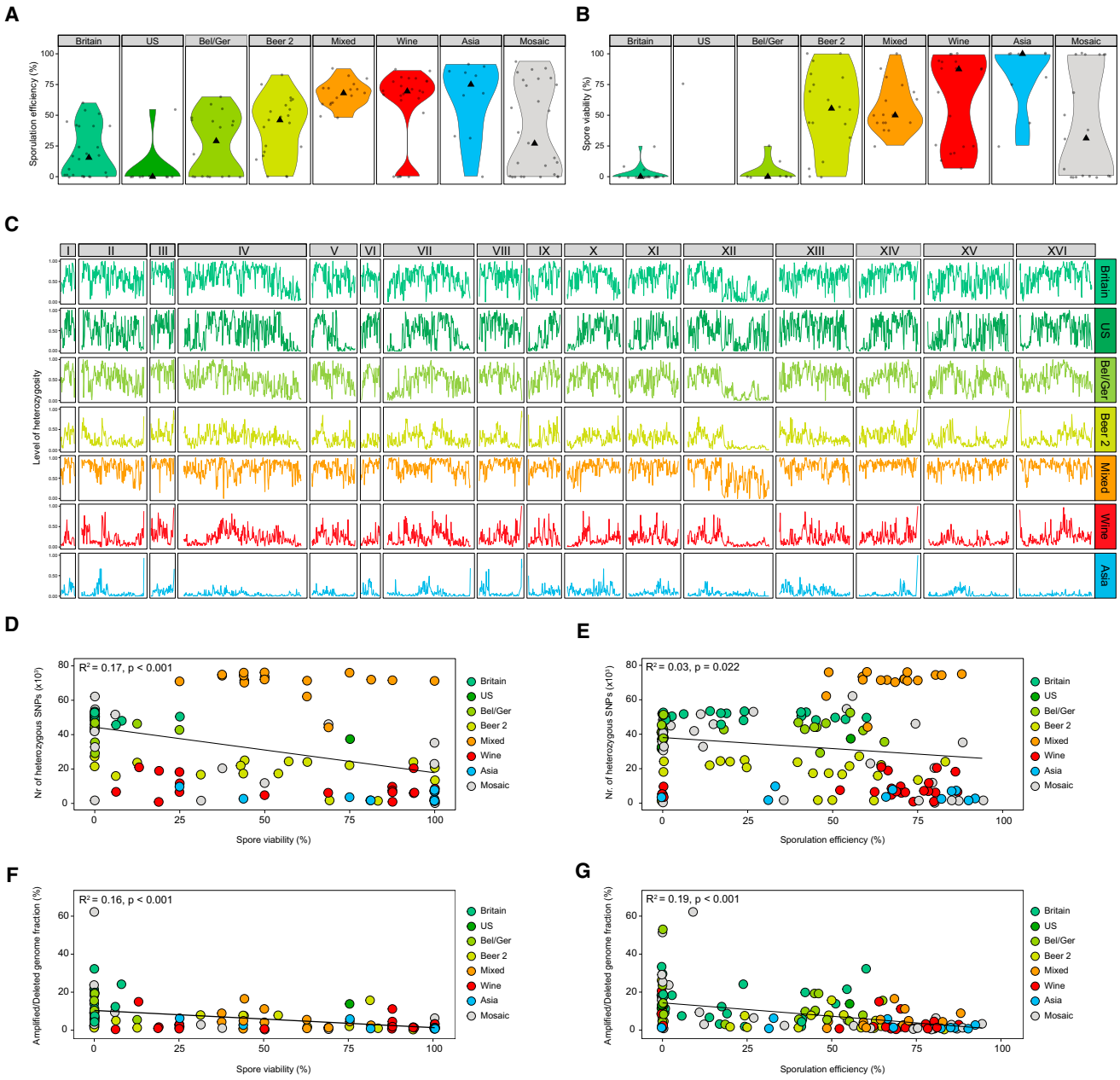


Figure 4. The Reproductive Lifestyle of Industrial *S. cerevisiae* Strains

(A) Violin plots depicting sporulation efficiency of all strains from different subpopulations.

(B) Violin plots depicting spore viability of all sporulating strains from different subpopulations.

(C) Visualization of the level of heterozygosity across the genome of the different subpopulations, calculated as the ratio of heterozygous/homozygous SNPs in 10 kb windows.

(D–G) Scatter plots depicting the correlation between the number of heterozygous loci and spore viability (D) or sporulation efficiency (E), and the correlation between the fraction of the genome subjected to large (>20 kb) structural variation and spore viability (F) or sporulation efficiency (G). Dot colors indicate subpopulations similar to the color code of Figure 1B.

See also Figure S4.

confirmed by the average per-site nucleotide divergence (d_{xy}), which is significantly lower between Britain and the United States (average $d_{xy} = 1.97 \times 10^{-3}$) than between Belgium/Germany and the United States strains (average $d_{xy} = 2.26 \times 10^{-3}$) (Wilcoxon signed-rank test, $p < 0.001$) (Table S8). This suggests that the origin of the United States brewing strains can be traced back to the introduction of beer culture in the United States by early 17th century British settlers (Van Wieren, 1995). Third, in contrast

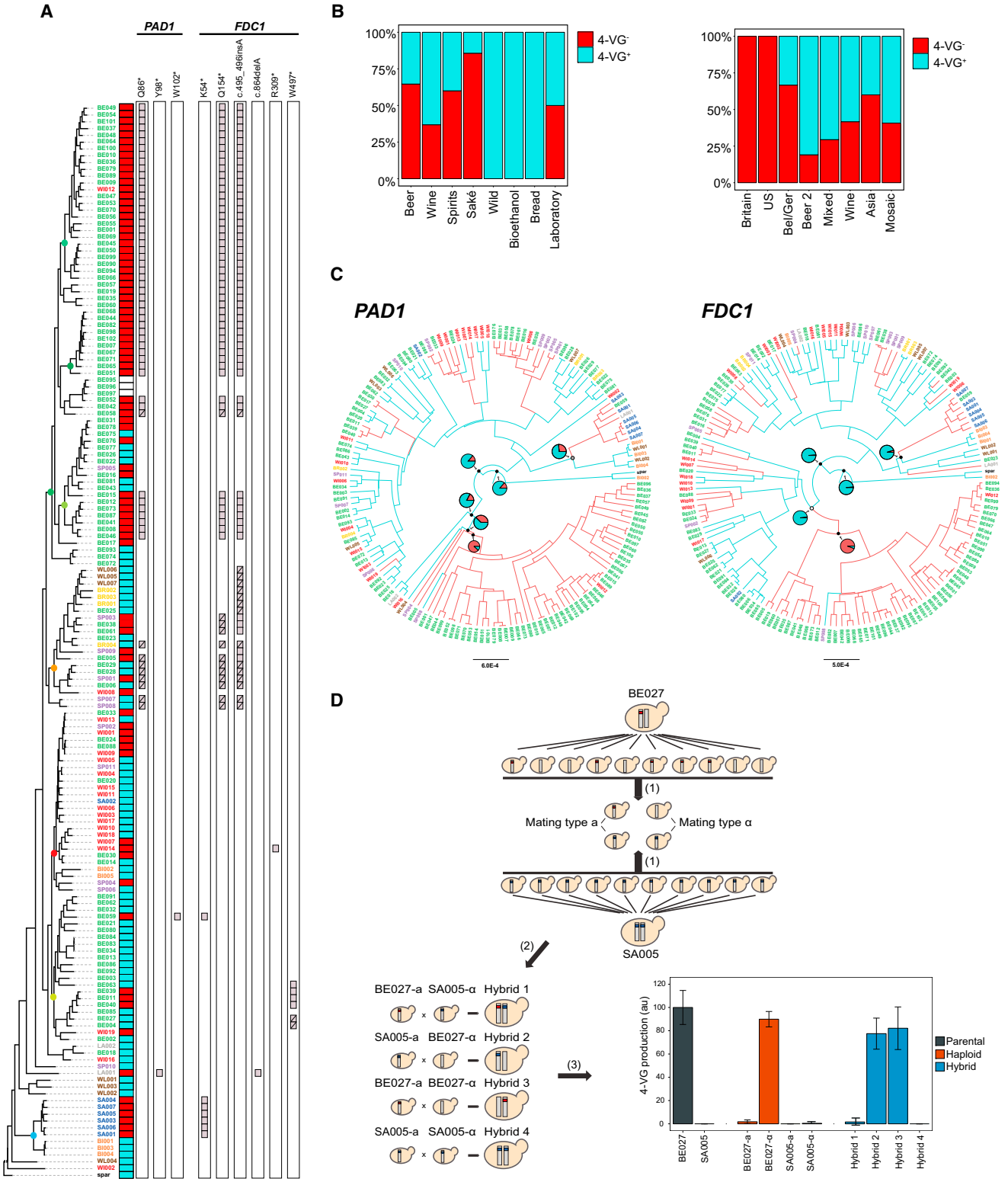


Figure 5. Production of 4-Vinyl Guaiacol by Industrial *S. cerevisiae* Strains

(A) Distribution of loss-of-function SNPs and frame-shift mutations in *FDC1* and *PAD1* of industrial *S. cerevisiae* strains. Gray boxes indicate the presence of the loss-of-function mutation, diagonal bars indicate heterozygosity at this site. Strains are clustered according to the strain phylogeny and strain names are colored according to their origin. Basal splits of the five industrial lineages are indicated with colored dots (see Figure 1A).

(legend continued on next page)

to wine, beer is not produced seasonally but throughout the whole year, which provides fermenting yeast with a predictable and stable growth environment. Yeast cells undergo about three doublings during one batch of beer fermentation, which takes ~1 week. Moreover, brewers typically recycle yeasts from a finished fermentation to inoculate a new batch, which implies that beer yeasts are continuously growing in their industrial niche. Together, these facts make it possible to estimate the number of generations to be ~150/year.

Based on estimates of the number of generations per year and the divergence time between United Kingdom and United States beer strains, we calculated the average mutation rate in a brewing environment to be 1.61–1.73E-08/bp/generation. While this value differs from previous assumptions, it is similar to the measured mutation rate in a diploid yeast strain that was subjected to 2 years of artificial evolution in a high-ethanol environment (Voordeckers et al., 2015). Moreover, mutations likely also occur in the second phase of beer fermentations, when cells are no longer dividing, which implies that the mutation rate per generation in industrial conditions should be higher than what is measured under conditions where the cells are dividing frequently, as is usually the case in laboratory experiments (Loewe et al., 2003). Using these data, the last common ancestor of the three major Beer 1 subclades (Belgium/Germany, United Kingdom, and the United States) is estimated to date from AD 1573–1604, suggesting that domestication started around this time. Interestingly, this coincides with the gradual switch from home-centered beer brewing where every family produced their own beer, to more professional large-scale brewing, first in pubs and monasteries and later also in breweries (Hornsey, 2003). The last common ancestor of Beer 2 is estimated to be more recent, between AD 1645–1671. This suggests that beer yeast domestication started before the discovery of microbes and the isolation of the first pure yeast cultures by Emil Hansen in the Carlsberg brewery in 1883, but well after the invention of beer production, estimated to have occurred as early as 3000 BC (Michel et al., 1992). Although it is difficult to assess how many different yeast strains were domesticated and in which industrial context these domestications occurred, the limited number of clades of industrial yeasts and the clear segregation of wild and industrial yeasts suggests that today's industrial yeasts originated from a limited set of ancestral strains, or closely related groups of ancestral strains.

DISCUSSION

Together, our results show that today's industrial *S. cerevisiae* yeasts are genetically and phenotypically separated from wild stocks due to human selection and trafficking. Specifically, the thousands of industrial yeasts that are available today seem to stem from only a few ancestral strains that made their way into food fermentations and subsequently evolved into separate lineages, each used for specific industrial applications. Within each cluster, strains are sometimes further subdivided along geographical boundaries, as is the case for the Beer 1 clade, which is divided into three main subgroups. However, further subclustering of beer yeasts according to beer style was generally not observed, which may not be surprising as it is common practice for brewers to use only one yeast strain within their brewery for the production of a wide array of different beers. Notable exceptions are yeasts associated with the few beers that largely depend on very specific yeasts characteristics, such as Hefeweizen beers. Another exception may include those beers for which production is restricted to a specific geographic area, such as Belgian Saisons or British Stouts.

We further show that industrial yeasts were clearly subjected to domestication, which is reflected in their genomes and phenomes. Interestingly, domestication seems strongest in beer yeasts, which demonstrate domestication hallmarks such as decay of sexual reproduction and general stress resistance, as well as convergent evolution of desirable traits like maltotriose utilization. Yeasts from the Beer 1 clade show the clearest signs of domestication, possibly because Beer 2 only diverged more recently from other sublineages. Many of these domestication features may have simply been the result of the yeasts' adaptation to their new industrial niches. However, for some traits, it is likely that humans actively intervened, e.g., by selecting strains that do not produce undesirable off-flavors, which our analysis identifies as *PAD1* or *FDC1* nonsense mutants.

The presence of a strong domestication signature in beer yeast genomes agrees well with the common practices in the brewing industry. Beer yeasts are typically recycled after each fermentation batch, and because beer is produced throughout the year, this implies that beer yeasts are continuously growing in their industrial niche. By contrast, wine yeasts can only grow in wine must for a short period every year, spending the rest of their lives in and around the vineyards or in the guts of insects (Bokulich et al., 2014; Christiaens et al., 2014; Stefanini et al., 2016). During these nutrient-poor periods, wine yeasts likely undergo few mitotic doublings, yet they may undergo sexual cycles and even hybridize with wild yeasts (Stefanini et al., 2016). Moreover, only a very small portion of the yeasts may find their way back into the grape must when the next harvest season arrives, while trillions of cells are being transferred to the next batch during backslopping in beer production. This results in large

(B) Percentage of strains within each origin (left) and population (right) capable of producing 4-vinyl guaiacol (4-VG). Red, 4-VG$^-$; turquoise, 4-VG$^+$.
(C) Phylogenetic trees and ancestral trait reconstruction of *PAD1* and *FDC1* genes. Branches are colored according to the most probable state of their ancestral nodes, turquoise (4-VG$^+$) or red (4-VG$^-$). Pie charts indicate probabilities of each state at specific nodes, turquoise (4-VG$^+$) or red (4-VG$^-$); posterior probability for the same nodes is indicated by a dot: black dot, 90%–100%; gray, 70%; white, 42%. Branch lengths reflect the average numbers of substitutions per site (compare scale bars).
(D) Development of new yeast variants with specific phenotypic features by marker-assisted breeding. Two parent strains (BE027 and SA005) were sporulated and, using genetic markers, segregants with the desired genotype were selected (1). Next, breeding between segregants from different parents (outbreeding) or the same parent (inbreeding) were performed (2). This breeding scheme yields hybrids with altered aromatic properties that can directly be applied in industrial fermentations (3). 4-VG production is shown relative to the production of BE027. Yeast genomes are represented by gray bars, loss-of-function mutations in *FDC1* as red (W497*) and blue (K54*) boxes within the gray bars. Error bars represent one SD from the mean.
See also Table S5.

effective population sizes for beer, but not wine, yeasts. The differences in industrial practices between beer brewing and wine-making likely had three important consequences. First, beer yeasts evolved faster than wine strains (Figure 1A). This resulted in a large genetic diversity within beer yeasts, while wine yeasts are genetically more homogeneous (Table 1). Second, after the initial domestication event, some beer yeasts were contained in the brewery and diverged allopatrically, leading to geographically defined subpopulations mirroring human traffic and colonization. Third, beer strains generally lost their ability to reproduce sexually. This, combined with continuous cultivation in a mild growth environment, made them susceptible to genetic drift and fixation of deleterious alleles that would otherwise be purged by evolutionary competition in harsh conditions. Hence, these asexual populations continuously accumulated deleterious mutations in an irreversible manner, a process known as Muller's ratchet (Muller, 1964). We propose that continuous clonal reproduction and relaxed selection for general stress resistance and famine likely allowed genome decay in beer yeasts and resulted in yeasts specialized in thriving in a man-made niche like beer fermentations, but not in natural environments. Both these characteristics (genome decay and niche specialization) are considered to be key characteristics of domestication.

Our study does not only provide insight into the domestication origin of industrial yeasts, it may also help to select and breed new superior strains. The genome sequences, phylogenetic tree, and phenome data can be used to set up marker-assisted breeding schemes similar to those routinely used for the breeding of superior crops and livestock (Takeda and Matsuoka, 2008).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Strain Collection
- METHOD DETAILS
  - DNA Extraction
  - Library Prep and Whole Genome Sequencing
  - De Novo Assembly
  - Annotation
  - Core Genome Analysis and Identification of Single Copy Genes
  - Reference-Based Alignments and Variant Calling
  - Phylogenetic Analyses
  - Population Structure and Diversity Analysis
  - Time Divergence Estimate
  - Copy-Number Variation Analysis
  - Character Evolution Analysis
  - Determination of Cell Ploidy
  - Phenotypic Analysis
  - Development of Artificial Hybrids
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY
  - Data Resources

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2016.08.020.

An audio PaperClip is available at http://dx.doi.org/10.1016/j.cell.2016.08.020#mmc9.

AUTHOR CONTRIBUTIONS

Conceptualization, B.G., J.S., T.P., L.M., S.M., and K.J.V.; Formal Analysis, B.G., J.S., L.S., M.R., A.M., and K.V.; Investigation, B.G., J.S., T.P., A.M., L.S., V.S., B.H.-M., and M.T.; Resources, L.M., B.S., C.T., C.W., and K.J.V.; Writing, B.G., J.S., S.M., and K.J.V.; Supervision, T.P., T.R., A.S., C.W., G.B., S.M., and K.J.V.

ACKNOWLEDGMENTS

We thank all K.J.V. and S.M. laboratory members for their help and suggestions. We thank K. Wolfe, S. Oliver, and P. Malcorps for their valuable feedback on the manuscript. We acknowledge D. Brami, A. Bass, J. Kurowski, K. Fortmann, and N. Parker for data collection and data review. Additionally we acknowledge S. Rombauts, Y. Lin, R. de Jonge, and V. Storme for fruitful discussions on data analysis. B.G. acknowledges funding from the VIB International PhD Program in Life Sciences. J.S. acknowledges funding from IWT and KU Leuven. K.J.V. acknowledges funding from an ERC Consolidator grant CoG682009, HFSP program grant RGP0050/2013, KU Leuven NATAR Program Financing, VIB, EMBO YIP program, FWO, and IWT. S.M. acknowledges funding from VIB and Ghent University. The funders had no role in study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Received: March 24, 2016
Revised: June 8, 2016
Accepted: August 8, 2016
Published: September 8, 2016

REFERENCES

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J.L., Serra, M., Dequin, S., Couloux, A., et al. (2015). A population genomics insight into the Mediterranean origins of wine yeast domestication. Mol. Ecol. *24*, 5412–5427.

Baele, G., and Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. Bioinformatics *29*, 1970–1979.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., and Alekseyenko, A.V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. *29*, 2157–2167.

Bergström, A., Simpson, J.T., Salinas, F., Barré, B., Parts, L., Zia, A., Nguyen Ba, A.N., Moses, A.M., Louis, E.J., Mustonen, V., et al. (2014). A high-definition view of functional genetic variation from natural yeast genomes. Mol. Biol. Evol. *31*, 872–888.

Bokulich, N.A., Thorngate, J.H., Richardson, P.M., and Mills, D.A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. Proc. Natl. Acad. Sci. USA *111*, E139–E148.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Borneman, A.R., Desany, B.A., Riches, D., Affourtit, J.P., Forgan, A.H., Pretorius, I.S., Egholm, M., and Chambers, P.J. (2011). Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of *Saccharomyces cerevisiae*. PLoS Genet. *7*, e1001287.

Borneman, A.R., Forgan, A.H., Kolouchova, R., Fraser, J.A., and Schmidt, S.A. (2016). Whole genome comparison reveals high levels of inbreeding and strain

redundancy across the spectrum of commercial wine strains of *Saccharomyces cerevisiae*. G3 (Bethesda) *6*, 957–971.

Briza, P., Breitenbach, M., Ellinger, A., and Segall, J. (1990). Isolation of two developmentally regulated genes involved in spore wall maturation in *Saccharomyces cerevisiae*. Genes Dev. *4*, 1775–1789.

Brown, C.A., Murray, A.W., and Verstrepen, K.J. (2010). Rapid expansion and functional divergence of subtelomeric gene families in yeasts. Curr. Biol. *20*, 895–903.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Christiaens, J.F., Franco, L.M., Cools, T.L., De Meester, L., Michiels, J., Wenseleers, T., Hassan, B.A., Yaksi, E., and Verstrepen, K.J. (2014). The fungal aroma gene *ATF1* promotes dispersal of yeast cells through insect vectors. Cell Rep. *9*, 425–432.

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

Dittmar, J.C., Reid, R.J., and Rothstein, R. (2010). ScreenMill: a freely available software suite for growth measurement, analysis and visualization of high-throughput screen data. BMC Bioinformatics *11*, 353.

Driscoll, C.A., Macdonald, D.W., and O'Brien, S.J. (2009). From wild animals to domestic pets, an evolutionary view of domestication. Proc. Natl. Acad. Sci. USA *106* (*Suppl 1*), 9971–9978.

Drummond, A.J., and Suchard, M.A. (2010). Bayesian random local clocks, or one rate to rule them all. BMC Biol. *8*, 114.

Drummond, A.J., Suchard, M.A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. *29*, 1969–1973.

Dunham, M.J., Badrane, H., Ferea, T., Adams, J., Brown, P.O., Rosenzweig, F., and Botstein, D. (2002). Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA *99*, 16144–16149.

Dunn, B., Richter, C., Kvitek, D.J., Pugh, T., and Sherlock, G. (2012). Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. Genome Res. *22*, 908–924.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Filteau, M., Hamel, V., Pouliot, M.C., Gagnon-Arsenault, I., Dubé, A.K., and Landry, C.R. (2015). Evolutionary rescue by compensatory mutations is constrained by genomic and environmental backgrounds. Mol. Syst. Biol. *11*, 832.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Goddard, M.R., and Greig, D. (2015). *Saccharomyces cerevisiae*: a nomadic yeast with no niche? FEMS Yeast Res. *15*, fov009.

Goddard, M.R., Godfray, H.C.J., and Burt, A. (2005). Sex increases the efficacy of natural selection in experimental yeast populations. Nature *434*, 636–640.

Hornsey, I.S. (2003). A History of Beer and Brewing (Cambridge: The Royal Society of Chemistry).

Hutter, S., Vilella, A.J., and Rozas, J. (2006). Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics *7*, 409.

Huxley, C., Green, E.D., and Dunham, I. (1990). Rapid assessment of S. cerevisiae mating type by PCR. Trends Genet. *6*, 236.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

Kozlov, A.M., Aberer, A.J., and Stamatakis, A. (2015). ExaML version 3: a tool for phylogenomic analyses on supercomputers. Bioinformatics *31*, 2577–2579.

Kück, P., and Meusemann, K. (2010). FASconCAT: convenient handling of data matrices. Mol. Phylogenet. Evol. *56*, 1115–1118.

Kuhn, R.M., Karolchik, D., Zweig, A.S., Trumbower, H., Thomas, D.J., Thakkapallayil, A., Sugnet, C.W., Stanke, M., Smith, K.E., Siepel, A., et al. (2007). The UCSC genome browser database: update 2007. Nucleic Acids Res. *35*, D668–D673.

Lanfear, R., Calcott, B., Ho, S.Y.W., and Guindon, S. (2012). Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. *29*, 1695–1701.

Legras, J.L., and Karst, F. (2003). Optimisation of interdelta analysis for *Saccharomyces cerevisiae* strain characterisation. FEMS Microbiol. Lett. *221*, 249–255.

Lemey, P., Rambaut, A., Drummond, A.J., and Suchard, M.A. (2009). Bayesian phylogeography finds its roots. PLoS Comput. Biol. *5*, e1000520.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341.

Liu, Y., Schröder, J., and Schmidt, B. (2013). Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinformatics *29*, 308–315.

Loewe, L., Textor, V., and Scherer, S. (2003). High deleterious genomic mutation rate in stationary phase of *Escherichia coli*. Science *302*, 1558–1560.

Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., Hartl, D.L., and Thomas, W.K. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc. Natl. Acad. Sci. USA *105*, 9272–9277.

Magwene, P.M., Kayıkçı, Ö., Granek, J.A., Reininga, J.M., Scholl, Z., and Murray, D. (2011). Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA *108*, 1987–1992.

McDonald, M.J., Rice, D.P., and Desai, M.M. (2016). Sex speeds adaptation by altering the dynamics of molecular evolution. Nature *531*, 233–236.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Michel, R.H., McGovern, P.E., and Badler, V.R. (1992). Chemical evidence for ancient beer. Nature *360*, 24.

Mukai, N., Masaki, K., Fujii, T., Kawamukai, M., and Iefuji, H. (2010). *PAD1* and *FDC1* are essential for the decarboxylation of phenylacrylic acids in *Saccharomyces cerevisiae*. J. Biosci. Bioeng. *109*, 564–569.

Muller, H.J. (1964). The relation of recombination to mutational advance. Mutat. Res. *106*, 2–9.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M., and Stamatakis, A. (2010). How many bootstrap replicates are necessary? J. Comput. Biol. *17*, 337–354.

Pavelka, N., Rancati, G., Zhu, J., Bradford, W.D., Saraf, A., Florens, L., Sanderson, B.W., Hattem, G.L., and Li, R. (2010). Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. Nature *468*, 321–325.

Peng, Y., Leung, H., Yiu, S., and Chin, F. (2010). IDBA – a practical iterative de Bruijn Graph de novo assembler. In Research in Computational Molecular Biology, B. Berger, ed. (Springer), pp. 426–440.

Pérez-Ortín, J.E., Querol, A., Puig, S., and Barrio, E. (2002). Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. Genome Res. *12*, 1533–1539.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., and Lercher, M.J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. *31*, 1929–1936.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Purugganan, M.D., and Fuller, D.Q. (2009). The nature of selection during plant domestication. Nature *457*, 843–848.

R Development Core Team (2011). R: A language and environment for statistical computing (R Foundation for Statistical Computing).

Raj, A., Stephens, M., and Pritchard, J.K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573–589.

Rambaut, A., Suchard, M.A., Xie, D., and Drummond, A.J. (2014) Tracer v1.6. http://beast.bio.ed.ac.uk/Tracer.

Rancati, G., Pavelka, N., Fleharty, B., Noll, A., Trimble, R., Walton, K., Perera, A., Staehling-Hampton, K., Seidel, C.W., and Li, R. (2008). Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. Cell *135*, 879–893.

Ranwez, V., Harispe, S., Delsuc, F., and Douzery, E.J.P. (2011). MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. PLoS ONE *6*, e22594.

Schacherer, J., Shapiro, J.A., Ruderfer, D.M., and Kruglyak, L. (2009). Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature *458*, 342–345.

Selmecki, A.M., Dulmage, K., Cowen, L.E., Anderson, J.B., and Berman, J. (2009). Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. PLoS Genet. *5*, e1000705.

Sheltzer, J.M., Blank, H.M., Pfau, S.J., Tange, Y., George, B.M., Humpton, T.J., Brito, I.L., Hiraoka, Y., Niwa, O., and Amon, A. (2011). Aneuploidy drives genomic instability in yeast. Science *333*, 1026–1030.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. *33* (Web Server issue, suppl 2), W465–W467.

Steensels, J., and Verstrepen, K.J. (2014). Taming wild yeast: potential of conventional and nonconventional yeasts in industrial fermentations. Annu. Rev. Microbiol. *68*, 61–80.

Steensels, J., Meersman, E., Snoek, T., Saels, V., and Verstrepen, K.J. (2014). Large-scale selection and breeding to generate industrial yeasts with superior aroma production. Appl. Environ. Microbiol. *80*, 6965–6975.

Stefanini, I., Dapporto, L., Berná, L., Polsinelli, M., Turillazzi, S., and Cavalieri, D. (2016). Social wasps are a *Saccharomyces* mating nest. Proc. Natl. Acad. Sci. USA *113*, 2247–2251.

Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., Dietrich, F.S., and McCusker, J.H. (2015). The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. *25*, 762–774.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. *34* (Web Server issue, suppl 2), W609–W612.

Takeda, S., and Matsuoka, M. (2008). Genetic approaches to crop improvement: responding to environmental and population changes. Nat. Rev. Genet. *9*, 444–457.

van Dijk, D., Dhar, R., Missarova, A.M., Espinar, L., Blevins, W.R., Lehner, B., and Carey, L.B. (2015). Slow-growing cells within isogenic populations have increased RNA polymerase error rates and DNA damage. Nat. Commun. *6*, 7972.

Van Wieren, D.P. (1995). American Breweries II (Eastern Coast Breweriana Association).

Voordeckers, K., Kominek, J., Das, A., Espinosa-Cantú, A., De Maeyer, D., Arslan, A., Van Pee, M., van der Zande, E., Meert, W., Yang, Y., et al. (2015). Adaptation to high ethanol reveals complex evolutionary pathways. PLoS Genet. *11*, e1005635.

Wang, Q.-M., Liu, W.-Q., Liti, G., Wang, S.A., and Bai, F.Y. (2012). Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. Mol. Ecol. *21*, 5404–5417.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. Nature *449*, 54–61.

Warringer, J., Zörgö, E., Cubillos, F.A., Zia, A., Gjuvsland, A., Simpson, J.T., Forsmark, A., Durbin, R., Omholt, S.W., Louis, E.J., et al. (2011). Trait variation in yeast is defined by population history. PLoS Genet. *7*, e1002111.

White, M.D., Payne, K.A.P., Fisher, K., Marshall, S.A., Parker, D., Rattray, N.J.W., Trivedi, D.K., Goodacre, R., Rigby, S.E.J., Scrutton, N.S., et al. (2015). UbiX is a flavin prenyltransferase required for bacterial ubiquinone biosynthesis. Nature *522*, 502–506.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C., and Weir, B.S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics *28*, 3326–3328.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
2-methoxy-4-vinylphenol (4-VG)	Sigma-Aldrich	Cat#W267511; CAS: 7786-61-0
3-Methylbutanol	Sigma-Aldrich	Cat#I9392; CAS: 123-51-3
Acetic acid	VWR	Cat#VELC1005.2500; CAS: 64-19-7
Agar powder bacteriological	VWR	Cat#89132-792; CAS: 9002-18-0
Antimycin A	Sigma-Aldrich	Cat#A8674; CAS: 1397-94-0
Bacto Peptone	BD Biosciences	Cat#211820
Cadmium sulfate 8/3-hydrate	Sigma-Aldrich	Cat#20920; CAS: 7790-84-3
Chloroform	VWR	Cat#PROL22711.324; CAS: 67-66-3
Copper(II) sulfate	Sigma-Aldrich	Cat#451657; CAS: 7758-98-7
Cycloheximide	Sigma-Aldrich	Cat#C7698; CAS: 66-81-9
D-(−)-Fructose	Sigma-Aldrich	Cat#47740; CAS: 57-48-7
D-(+)-Glucose monohydrate	Sigma-Aldrich	Cat#49159; CAS: 14431-43-7
D-(+)-Maltose monohydrate	Sigma-Aldrich	Cat#M5885; CAS: 6363-53-7
D-(+)-Melibiose	Sigma-Aldrich	Cat#63630; CAS: 585-99-9
Diethyl ether	VWR	Cat#PROL23811.292; CAS: 60-29-7
D-Sorbitol	Sigma-Aldrich	Cat#S1876; CAS: 50-70-4
Ethanol absolute	VWR	Cat#PROL20802.321; CAS: 64-17-5
Ferulic acid	Sigma-Aldrich	Cat#128708; CAS: 537-98-4
Formic acid	Sigma-Aldrich	Cat#6440; CAS: 64-18-6
Galactose	Fisher	Cat#15061-0010; CAS: 59-23-4
Glycerol	Sigma-Aldrich	Cat#G5516; CAS: 56-81-5
Levulinic acid	Sigma-Aldrich	Cat#L2009; CAS: 123-76-2
L-Lysine	Sigma-Aldrich	Cat#L5501; CAS: 56-87-1
Maltotriose hydrate	Sigma-Aldrich	Cat#851493; CAS: 207511-08-8
Potassium acetate	VWR	Cat#PROL26667.293; CAS: 27-08-2
Potassium chloride	VWR	Cat#26764-298; CAS: 7447-40-7
Propidium iodide solution	Sigma-Aldrich	Cat#P4864; CAS: 25535-16-4
Roti phenol	Carl Roth	Cat#38.3; CAS: 108-95-2
SC Amino acid mixture	MP biomedicals	Cat#114400022
Sodium chloride	Fisher scientific	Cat#S/3160/60; CAS: 7647-14-5
Sodium sulfite	Sigma-Aldrich	Cat#S0505; CAS: 7757-83-7
Sucrose	Sigma-Aldrich	Cat#84100; CAS: 57-50-1
TRIS acetate – EDTA buffer solution	Sigma-Aldrich	Cat#93296
Yeast Extract powder	LabM	Cat#MC001; CAS: 013-01-2
Yeast Nitrogen Base Without Amino Acids	Sigma-Aldrich	Cat#Y0626
Yeast Nitrogen Base Without Amino Acids and Ammonium Sulfate	Sigma-Aldrich	Cat#Y1251
Zymolyase 100T	AMSBIO	Cat#120493-1; CAS: 37340-57-1
Critical Commercial Assays		
Genomic-tip 100/G	QIAGEN	Cat#10243
Genomic DNA buffer set	QIAGEN	Cat#19060
MasterPure Yeast DNA Purification Kit	Epicenter	Cat#MPY80200
Nextera XT DNA Library Preparation Kit	Illumina	Cat#FC-131-1024

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
SPRI works HT	Beckman Coulter	Cat#B06938
Deposited Data		
De novo assembly	DDBJ/ENA/GenBank	BioProject PRJNA323691
Experimental Models: Organisms/Strains		
See Table S1 for list of strains sequenced	This paper	N/A
Sequence-Based Reagents		
FDC1-W497-FW: 5′-TGCAGATCAGATGGC TTTTG-3′	This study	N/A
FDC1-W497-RV-STOP: 5′-GCAATTATTTATA TCCGTACCTTTTT-3′	This study	N/A
FDC1-W497-RV-ALT: 5′-GCAATTATTTATATC CGTACCTTTTC-3′	This study	N/A
Delta12: 5′-TCAACAATGGAATCCCAAC-3′	Legras and Karst, 2003	N/A
Delta21: 5′-CATCTTAACACCGTATATGA-3′	Legras and Karst, 2003	N/A
P86: 5′-ACTCCACTTCAAGTAAGAGTT-3′	Huxley et al., 1990	N/A
P87: 5′-GCACGGAATATGGGACATACTT-3′	Huxley et al., 1990	N/A
P88: 5′-AGTCACATCAAGATGGTTTAT-3′	Huxley et al., 1990	N/A
Software and Algorithms		
Agilent Chemstation software	Agilent Technologies, USA	https://www.agilent.com/en-us/products/software-informatics/massspec-workstations/gc-msd-chemstation-software
AUGUSTUS (v2.5)	Stanke and Morgenstern 2005	http://bioinf.uni-greifswald.de/augustus/; RRID: SCR_008417/
BEAST (v1.8.2)	Drummond et al., 2012	http://beast.bio.ed.ac.uk/; RRID: SCR_010228
Burrows-Wheeler Aligner (BWA) (v0.6.1)	Li and Durbin, 2009	http://bio-bwa.sourceforge.net/; RRID: SCR_010910
CD-HIT (v4.6)	Fu et al., 2012	http://weizhongli-lab.org/cd-hit/; RRID: SCR_007105
ExaML (v3)	Kozlov et al., 2015	http://sco.h-its.org/exelixis/web/software/examl/index.html
FASconCAT (v1.0)	Kück and Meusemann, 2010	https://www.zfmk.de/en/research/research-centres-and-groups/fasconcat
fastStructure (v1.0)	Raj et al., 2014	https://rajanil.github.io/fastStructure/
FigTree (v1.4.2)		http://tree.bio.ed.ac.uk/software/figtree/; RRID: SCR_008515
Gene-E	Broad Institute	http://www.broadinstitute.org/cancer/software/GENE-E/
Genome Analysis Toolkit (GATK) (v2.7.2)	Broad Institute, McKenna et al., 2010	https://www.broadinstitute.org/gatk/; RRID: SCR_001876
Idba_ud (v1.1.1)	Peng et al., 2010	http://i.cs.hku.hk/∼alse/hkubrg/projects/idba_ud/
liftOver	Kuhn et al., 2007	http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver
LogCombiner	Drummond et al., 2012	http://beast.bio.ed.ac.uk/logcombiner
MACSE (v1.01b)	Ranwez et al., 2011	http://bioweb.supagro.inra.fr/macse/index.php?menu=intro&option=intro
MAFFT (v7.187)	Katoh and Standley, 2013	http://mafft.cbrc.jp/alignment/software/; RRID: SCR_011811
Musket	Liu et al., 2013	http://musket.sourceforge.net/homepage.htm#latest
PAL2NAL (v14)	Suyama et al., 2006	http://www.bork.embl.de/pal2nal/
PartitionFinder (v1.1.1)	Lanfear et al., 2012	http://www.robertlanfear.com/partitionfinder/
Picard Tools (v1.56)	Broad Institute	http://broadinstitute.github.io/picard/; RRID: SCR_006525

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PLINK (v1.07)	Purcell et al., 2007	http://pngu.mgh.harvard.edu/~purcell/plink/; RRID: SCR_001757
PopGenome (v2.1.6)	Pfeifer et al., 2014	https://cran.r-project.org/web/packages/PopGenome/index.html
RaxML (v8.1.3)	Stamatakis, 2014	http://sco.h-its.org/exelixis/web/software/raxml/index.html; RRID: SCR_006086
SAS (v9.4)		http://www.sas.com/en_us/software/sas9.html; RRID: SCR_008567
ScreenMill macro	Dittmar et al., 2010	http://www.rothsteinlab.com/tools/
SnpEff (v3.3)	Cingolani et al., 2012	http://snpeff.sourceforge.net/; RRID: SCR_005191
SNPRelate (v1.6.4)	Zheng et al., 2012	http://bioconductor.org/packages/release/bioc/html/SNPRelate.html
Splint	This paper	Available upon request
Tracer (v1.6)	Rambaut et al., 2014	http://tree.bio.ed.ac.uk/software/tracer/
trimAl (v1.2)	Capella-Gutiérrez et al., 2009	http://trimal.cgenomics.org/introduction
Trimmomatic (v0.30)	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic; RRID: SCR_011848
Variscan (v2.0.3)	Hutter et al., 2006	http://www.ub.edu/softevol/variscan/
VcfTools (v0.1.10;0.1.14)	Danecek et al., 2011	http://vcftools.sourceforge.net/; RRID: SCR_001235

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests may be directed to, and will be fulfilled by the corresponding author Kevin J. Verstrepen (kevin.verstrepen@biw.vib-kuleuven.be).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strain Collection

For this study, a set of 157 *Saccharomyces cerevisiae* strains was sequenced, phenotyped and analyzed. Strains were obtained from historical yeast collections of the VIB Laboratory for Systems Biology (KU Leuven, Belgium) and White Labs (USA). While detailed background information on many of these strains is limited, their geographical origin was in most cases documented and is listed in Table S1. Beer strains mainly originate from the main fermentation (82/102) or bottle conditioning (10/102) of ale beers. While lager beer fermentations are usually carried out by *S. pastorianus*, an interspecific hybrid of *S. cerevisiae* and *S. eubayanus*, we identified 10 *S. cerevisiae* strains typically employed in lager fermentations and included them in the selection.

All strains are long-term stored in −80°C using a glycerol-based standard storage medium (peptone 1% w v^{-1}, yeast extract 0.5% w v^{-1}, glucose 1% w v^{-1}, glycerol 25% v v^{-1}).

METHOD DETAILS

DNA Extraction

For strains BE001-043, BI001-005, BR001-004, LA001, SA001-007, SP001-007, NA001-004 and WI001-018, genomic DNA was prepared using the QIAGEN genomic tip kit (QIAGEN, Germany) according to recommended protocols. Final DNA concentrations were measured using Qubit (Thermo Fisher Scientific, USA). For the other strains, genomic DNA was extracted with the MasterPureTM Yeast DNA Purification Kit (Epicenter, USA), but with some modifications to the recommended protocols. Three mL of each of the liquid yeast cultures were pelleted by centrifugation at 20,800 x g for 5min in 1.7mL micro centrifuge tubes. 300 μl of yeast cell lysis solution was added to each micro centrifuge tube along with 1 μl of 5g L^{-1} RNase A. The cells were resuspended by vortex mixing each micro centrifuge tube for 10 s. Each tube was incubated at 65°C for 15 min and was then chilled on ice for 5 min. Next, 150 μl of protein precipitation reagent was added to each tube and the tubes were vortexed for 10 s. The suspensions were then centrifuged for 10 min at 20,800 x g to pellet the cellular debris. The supernatants were transferred to clean 1.7mL micro centrifuge tubes. Next, 500 μl of isopropanol was added to each tube. Each tube was then inverted several times to precipitate the DNA, which was then pelleted by centrifugation at 20,800 x g for 10 min. The supernatants were discarded and the pellets were washed with 500 μl of ice-cold 70% ethanol and briefly centrifuged. The ethanol was removed by pipetting. The DNA pellets were centrifuged again for 10 min at 20,800 x g to remove any remaining ethanol. Each DNA pellet was then suspended in 35 μl of TE buffer and stored at

4°C. After isolation, the purified DNA was quantified using fluorimetric methods and diluted to the optimal concentration for library construction.

Library Prep and Whole Genome Sequencing

For strains BE001-043, BI001-005, BR001-004, LA001, SA001-007, SP001-007, NA001-004 and WI001-018, paired-end sequencing libraries (100bp) with a mean insert size of 300bp were prepared and run according to the manufacturer's instructions on an Illumina HiSeq2000 at the EMBL GeneCore facility, Heidelberg (http://genecore3.genecore.embl.de/genecore3/). For the other strains, libraries were prepared using the Nextera XT sample preparation kit. A total of 50ng of yeast DNA was fragmented and tagged with DNA adapters by the Nextera transposome resulting in adaptor-ligated DNA fragments. The DNA was purified and PCR-amplified to add the dual indexes as well as the common adapters required for cluster generation and sequencing. All samples were pooled together and clustered on board a HiSeq 2500 instrument at Illumina (San Diego, USA). Samples were sequenced in both Rapid Run Mode and High Output Mode using 2 × 100 bp paired-end reads.

De Novo Assembly

For each library, low-quality and ambiguous reads were trimmed using Trimmomatic (v0.30) (Bolger et al., 2014). After k-mer based read correction with musket (Liu et al., 2013), reads were assembled using idba_ud (Peng et al., 2010). De novo assemblies were evaluated by mapping back the reads and also by checking BLAST matches of assembled contigs to the SILVA database for rDNA classification. Each de novo assembly was scaffolded against the *S. cerevisiae* S288c reference genome assembly (http://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/S288C_reference_genome_R64-1-1_20110203.tgz). The liftOver workflow (Kuhn et al., 2007) was used to determine the coordinates of contigs from each newly assembled strain relative to the reference strain (http://genomewiki.ucsc.edu/index.php/Minimal_Steps_For_LiftOver). Scaffolded contigs mapped to each reference strain chromosome were combined into a ''pseudo-molecule,'' with the placed contigs stitched together with gaps indicated by ''N.'' Unplaced contigs (including alternative lower scoring matches) were kept. Unplaced contigs less than 300 nucleotides were not included in the final assembly (Table S1).

Annotation

The genome annotation of the *S. cerevisiae* S288c reference genome (nr. of genes = 6,692) was downloaded from the UCSC (version Apr2011/sacCer3) Table Browser in GPR format. FASTA records were renamed to match the chromosome naming convention in the GPR file. The liftOver workflow was used to create a coordinate conversion file (chain file) between the *S. cerevisiae* S288c genome and each newly scaffolded assembly. Using the chain file, the coordinates of the *S. cerevisiae* S288c genes were ''lifted'' to each new genome assembly. Lifted genes were considered valid if they did not contain internal stop codons. Independently, the gene prediction tool AUGUSTUS v2.5 (Stanke and Morgenstern, 2005) was used to predict genes for each new strain using the provided training set/model for *S. cerevisiae* S288c with the following parameters (–noInFrameStop = true–maxDNAPieceSize = 1000000–progress = false–uniqueGeneId = true–keep_viterbi = false). The annotated and predicted genes, using liftOver and AUGUSTUS respectively, were combined with priority given to the liftOver annotation when the predictions overlapped (Table S1).

Core Genome Analysis and Identification of Single Copy Genes

Across the 157 annotated *S. cerevisiae* genomes, 986,179 genes were annotated and predicted in total, with an average of 6,281 genes per genome (min = 6,099 genes, max = 6,655 genes). CD-HIT (v4.6) was used to approximate a non-redundant set of putative translations across the 157 genomes (parameters, -c 0.7 -M 3200 -T 0 -d 60) (Fu et al., 2012). Using a 70% amino acid identity threshold, the collection of 986,179 translated genes was reduced to 8,410 clusters. A total of 3,519 clusters contained exactly one gene from each of the 157 assembled genomes. The 3,519 clusters represent an approximation of the *S. cerevisiae* core genome across the 157 genomes evaluated. A conservative set of single copy genes was identified across the 157 genome assemblies, the *Saccharomyces paradoxus* genome and an additional set of 24 *S. cerevisiae* strains assembled in recent studies (Table S1) (Bergström et al., 2014; Liti et al., 2009; Strope et al., 2015). First, one-to-one ortholog pairs were extracted from previously identified orthologs between *S. paradoxus* (NRRL-Y17217) and *S. cerevisiae* S288c (https://portals.broadinstitute.org/regev/orthogroups/orthologs/Scer-Spar-orthologs.txt) (Wapinski et al., 2007). A total of 5,096 one-to-one ortholog pairs were identified, with a subset of 5,084 annotations mapped to the *S. cerevisiae* S288c reference gene set used. The set of 5,084 genes was filtered to a smaller subset of 2,417 genes based on i) inclusion in the set of 3,333 *S. cerevisiae* S288c ORFS that could be mapped by liftOver across all 157 genomes, and ii) inclusion in the set of 3,519 clusters uniquely represented in each of the 157 strains based on CD-HIT results. Lastly, the presence and the single-copy status of the 2,417 genes were investigated in the additional 24 previously sequenced *S. cerevisiae* strains (http://www.moseslab.csb.utoronto.ca/sgrp/download.html and http://www.ncbi.nlm.nih.gov/genbank with accession numbers as reported in (Strope et al., 2015) - last access June 2015), further reducing the selection to a conservative set of 2,026 genes. The final set included 2,020 single-copy genes after removing six highly fragmented sequences.

Reference-Based Alignments and Variant Calling

To identify single nucleotide polymorphisms (SNPs) and short insertions and deletions (InDels), reads were pre-processed by filtering low quality and ambiguous reads, adapters and PhiX contaminations, using Trimmomatic (v0.30) (Bolger et al., 2014). Clean reads

were mapped to the *S. cerevisiae* reference genome S288c (R64-1-1, EF4-Ensemble Release 74) with the Burrows-Wheeler Aligner (BWA, v0.6.1) using default parameters except for –q 10 (Li and Durbin, 2009). Non-primary alignments and non-properly paired reads were filtered out and duplicate reads were marked using Picard Tools (v1.56) (http://picard.sourceforge.net). Before variant calling, reads were locally realigned in order to eliminate false positives due to misalignment of reads, which was followed by a base quality score recalibration step, using the Broad Institute Genome Analysis Toolkit (GATK v2.7.2) (McKenna et al., 2010). SNP and InDel discovery and genotyping was performed across all 157 strains simultaneously, to minimize false positive calls, with a minimum base quality score of 20, a standard minimum confidence threshold for calling of 50 and a standard emitted confidence of 20. Sites with total quality by depth < 2.00 and mapping quality < 40, genotype quality < 30 and genotype depth < 5 were filtered out using GATK Variant Filtration. For SNP calling, sites overlapping InDels, sites with more than 50% missing genotypes and multiallelic sites were filtered out using VcfTools (v0.1.10;v0.1.14) (Danecek et al., 2011). The final set of SNPs included a total of 421,361 biallelic segregating sites accounting for a total of 10,576,934 SNPs across all strains. SnpEff (v3.3) (Cingolani et al., 2012) was used to annotate and predict the effect of the variants.

Phylogenetic Analyses

Phylogenetic Tree for the Sequenced Collection - Figure S1A

Multiple sequence alignments (MSAs) for the 2,020 amino acid sequences identified above were generated using MAFFT (v7.187), with default settings and 1,000 refinement iterations (Katoh and Standley, 2013). Codon alignments were obtained from MSAs of predicted amino acid sequences and the corresponding DNA sequences by the PAL2NAL program (v14) (Suyama et al., 2006). Quality checks and format conversions were performed using trimAl (v1.2) (Capella-Gutiérrez et al., 2009). The full set of codon alignments were concatenated into a supermatrix using FASconCAT (v1.0) (Kück and Meusemann, 2010). The resulting supermatrix included 158 taxa and 2,782,494 positions, 99.174% nucleotides, 0.826% gaps and 0% ambiguities. The matrix was partitioned based on all 2,020 gene blocks and all three codon positions within each block, resulting in 6,060 distinct data partitions accounting for 144,171 distinct alignment patterns. Twenty completely random starting trees and 20 randomized stepwise-addition parsimony starting trees were obtained using RAxML (v8.1.3) (Stamatakis, 2014). Robinson-Foulds (RF) distances were computed between all trees in both the fully random and parsimony tree sets, to avoid systematic bias due to low diversity in starting trees. Because the stepwise addition algorithm generated a set of starting trees with low diversity, all the subsequent analyses were conducted with fully random starting trees. Twenty maximum-likelihood (ML) tree searches were performed on each of the 20 fully random starting trees under the GTRGAMMA model (4 discrete rate categories) using ExaML (v3) and the rapid hill climbing algorithm (-f d) (Kozlov et al., 2015). During the ML search, the alpha parameter of the model of rate heterogeneity and the rates of the GTR model of nucleotide substitutions were optimized independently for each partition. The branch lengths were optimized jointly across all partitions. For each starting tree, the best tree was selected based on the highest log-likelihood score. Parameters and branch lengths were re-optimized on the best 20 topologies with ExaML (-f E) using the median of the four categories for the discrete approximation of the GAMMA model of rate heterogeneity (-a). The tree with the best overall log-likelihood score of all 20 tree inferences was considered the final ML tree. Non-parametric bootstrap analysis was performed on the concatenated matrix using RaxML (v8.1.3). The a posteriori boot-stopping criterion (Pattengale et al., 2010) (MRE bootstrapping convergence criterion) was applied to define the number of replicates. After every 50 replicates, the set of bootstrapped trees generated so far is repeatedly (100 permutations) split in two equal subsets, and the Weighted Robinson-Foulds (WRF) distance is calculated between the majority-rule consensus trees of both subsets (for each permutation). Low WRF distances (< 3%) for > = 99% of permutations were used to indicate bootstrapping convergence. Convergence was reached after 250 replicates: average weighted Robinson-Foulds distance (WRF) = 1.86%, percentage of permutations in which the WRF was $\leq$ 3.00 = 100%. The tree was visualized and rooted in FigTree (v1.4.2) using *S. paradoxus* as the outgroup (http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenetic Tree for the Extended Collection - Figure 1A

In order to compare our strain collection with previously sequenced strains, we included 24 additional isolates, previously described in Liti et al. (2009) (re-sequenced in Bergström et al. (2014)) and Strope et al. (2015) (Table S1). Generation of MSAs and construction of the concatenation matrix was performed as described earlier. The resulting supermatrix included 2,785,239 positions, 99.077% nucleotides, 0.922% gaps and 0.001% ambiguities. The matrix was partitioned based on all 2,020 gene blocks and all three codon positions within each block, resulting in 6,060 distinct data partitions, accounting for 163,920 distinct alignments patterns. The ML searches and re-optimization were run on 30 fully random starting trees as described above, using RAxML (v8.1.3) and ExaML (v3). Non-parametric bootstrap analysis was performed as described above. Convergence was reached after 250 replicates: average weighted Robinson-Foulds distance (WRF) = 2.10%, percentage of permutations in which the WRF was $\leq$ 3.00 = 99%. The tree was visualized and rooted in FigTree using *S. paradoxus* as the outgroup.

Multi-locus Phylogeny - Figure S1C

Nine partial genes previously used to genetically characterize 99 Chinese isolates of *S. cerevisiae* (Wang et al., 2012) were recovered from 194 previously sequenced genomes (Table S2) and from the 157 isolates sequenced in this study, for a total of 450 strains. Each gene was aligned with MAFFT (v7.187) and the final MSAs were concatenated with FASconCAT (v1.0). The concatenated alignment was trimmed using trimAl (v1.2) with the automated1 option, optimized for ML tree reconstruction. The resulting supermatrix included 19,254 positions, 83.348% nucleotides and 16.652% gaps. The matrix was partitioned based on gene blocks in nine distinct partitions with joint branch length optimization. ML search on 30 fully random starting trees and non-parametric bootstrap analysis were

performed as described above. Convergence was reached after 550 replicates: average weighted Robinson-Foulds distance (WRF) = 2.10%, percentage of permutations in which the WRF was $\leq$ 3.00 = 99%. The tree was visualized and rooted in FigTree using *S. paradoxus* as the outgroup.

Population Structure and Diversity Analysis

The model-based Bayesian algorithm fastSTRUCTURE (v1.0) was used to detect and quantify the number of populations and the degree of admixture in the 157 sequenced genomes (Raj et al., 2014). The set of 421,361 biallelic segregating sites identified above was filtered further by removing SNPs with a minor allele frequency (MAF) < 0.05 and SNPs in linkage-disequilibrium, using PLINK (v1.07) (Purcell et al., 2007) with a window of size 50 SNPs advanced by 5 SNPs at a time and an r^2 threshold of 0.5. fastSTRUCTURE was run on a filtered set of 53,929 segregating sites, varying the number of ancestral populations (*K*) between 1 and 10 using the simple prior implemented in fastSTRUCTURE. The number of iterations varied from 10 at $K = 1$ up to 80 at $K = 10$. $K = 8$ was found to be optimal, i.e., scoring the highest marginal likelihood (log-marginal likelihood: −0.7298459788, total iterations: 60). Analysis of estimated ancestry (Q) matrices and plotting were performed in R (v3.1) (R Development Core Team, 2011). The same set of 53,929 SNPs was used to perform a principal component analysis (PCA) as implemented in the SNPrelate package (v1.6.4) (Zheng et al., 2012). Whole-genome levels of polymorphism were estimated using Variscan (v2.0.3) (Hutter et al., 2006), considering only sites with valid high-quality alleles (> Q40) in at least 80% of strains within the group (defined with the NumNuc parameter adjusted for each group together with CompleteDeletion = 0 and FixNum = 0).

Time Divergence Estimate

In order to estimate the timing of the split leading to the Beer 1 and Beer 2 clade, the mutation rate of yeasts in a beer fermentation environment was calculated. This calculation was based on four assumptions:

US Beer Yeasts Originate from UK Beer Yeasts

Phylogenetic analysis of strains sequenced in this study and previously sequenced wild isolates reveals that US beer yeasts are genetically closely related to European beer yeasts, but not to strains isolated from natural sources in the US. This strongly suggests migration of strains from Europe to the US after colonization. Moreover, the average per-site nucleotide divergence (d_{XY}) further indicates that these US strains likely originate from the UK (Table S8). The per-site nucleotide diversity between subpopulations was calculated on the set of 2,020 genes used for the inference of the strain phylogeny using the R package PopGenome (Pfeifer et al., 2014).

The Split between US and UK Beer Yeasts Happened between 1607 and 1637

The first permanent English settlement in North America was established in 1607 in Jamestown, Virginia. In 1609, American ''help wanted'' advertisements appear in London seeking brewers for this colony, indicating the importance of beer brewing in early colonial America. In 1637, the well-known colonist Captain Sedgwick founded the first authoritatively recorded brewery in the Massachusetts Bay Colony. Therefore, it is likely that even in early colonial America, beer was produced using yeasts that were brought in by English settlers.

Beer Yeasts Reproduce Clonally during Beer Brewing

During beer brewing, yeasts do not face long periods of nutrient starvation. Nutrient starvation is generally required to induce sporulation and thus initiate sexual reproduction. Indeed, our analyses of the sexual lifestyle of beer yeasts revealed that most beer strains, especially from Beer 1, lost the ability to produce viable spores (only observed in ~6% of the strains, and only in response to severe nutrient-poor conditions), further indicating that sexual reproduction is not favorable, and definitely not common, for beer yeasts during brewing.

Beer Yeasts Undergo around 150 Doublings per Year

As beer fermentations take around one week, and yeasts on average undergo a bit under three doublings per fermentation, it can be estimated that they undergo about 150 generations per year.

Using these parameters, an estimated mutation rate per site per generation was calculated from the formula $k = 2\mu t$, where k is the average per-site nucleotide divergence between US and UK strains (1.97E-03, see Table S8), μ is the mutation rate per site per generation and t is the time in generations, assuming 150 generations per year and divergence of US and UK strains between 1607 and 1637. This calculation yields a mutation rate of 1.61-1.73E-08/bp/generation, a value approximately 50x greater than the value typically calculated in haploid laboratory strains in non-stressful conditions (Lynch et al., 2008). While this value might seem high, it is not unreasonable for several reasons.

First, the mutation rate estimates obtained here are comparable to those measured in a directed evolution experiment performed in 6%–12% ethanol (Voordeckers et al., 2015). Although the conditions that were used in this directed evolution experiment differ from real beer fermentations, they do show that ethanol has a drastic effect on the mutation rate. Moreover, a 6%–12% ethanol concentration should be fairly comparable to the ethanol concentrations encountered in beer brewing over the past 400 years. It is a common misunderstanding that the alcohol percentage of ale beer in the past centuries used to be much lower than it is now. Indeed, in the time span that we considered in our calculations (the past 400 years), high-alcohol beers were being produced (Stan Hieronymus, Anders Kissmeyer, Martyn Cornell and Ron Pattinson, personal communication). The generally low ethanol tolerance of beer yeasts (as compared to wine and spirits yeasts, for example) and the presence of other stress factors during industrial fermentation (nutrient starvation for example) might further contribute to an increased mutation rate.

Second, yeast only undergoes about three cell divisions during beer fermentations, which generally take place in the first 48 hr of the fermentation. After this, the yeast cells are further exposed to high ethanol concentrations for several days, and it has been shown for several microbes that in this state of quiescence, mutations can still accumulate (Loewe et al., 2003). Hence, mutations can also occur in the second phase of fermentation, when the cells are not dividing, which implies that the mutation rate (per generation) in industrial growing conditions should be higher than what is measured under conditions where the cells are dividing frequently, as is usually the case in laboratory experiments.

Given the mutation rate estimate μ = 1.61-1.73E-08/bp/generation, an average of 150 generations/year and an average divergence *dxy* = 2.14E-03 substitutions/site between the UK/US and Belgium/Germany subclades in the Beer1 lineage, the last common ancestor of the major Beer 1 subclades is calculated to have existed until *dxy* / (2 μ * 150) = ~443-412 years ago. A similar calculation for Beer 2 (*dxy* = 1.79E-03 substitutions/site between the earliest diverging Beer 2 subclades) suggests that the last common ancestor of Beer 2 existed until ~371-345 years ago. Given the limited amount of information that could be used for dating, both ages should be considered only rough approximations.

Copy-Number Variation Analysis

Copy-number variations (CNVs) were identified on the reference-based alignments. Initial read depth profiles were obtained for each isolate based on the average read depth calculated in non-overlapping windows of 1000bp. In 68 samples (BE044-BE102, LA002, SP008-SP011, NA005-NA007, WI019), a deviation in read depth was detected: instead of fluctuating around a constant line, the read depth profile showed a convex trend with high depth at the terminal regions of the chromosomes that gradually decreased toward the center. These samples also showed high local variance. This bias in coverage is further referred to as a ''smiley pattern.'' Since conventional methods for CNV detection rely on read depth as a proxy for copy number, these methods were not applicable on the ''smiley pattern'' strains. To tackle this problem, a custom-built algorithm was developed, dubbed Splint (available upon request), which instead measures the size of discontinuities in read depth by using a discontinuous spline regression technique. In Splint, the data were modeled as the product of the bias and the copy number of each region, plus error. Here, the bias was assumed to be a continuous curve (expected depth as a function of chromosomal location), modeled as a smoothing spline. The copy number on the other hand is a piecewise constant function, with discontinuities at breakpoints in between regions of constant copy number. This was modeled as a sum of indicator functions, one for each region. After regression, the fitted value of the coefficient of each indicator function is proportional to the copy number in the corresponding region. The regression method requires the locations of the discontinuities as input values. Initially, these are located in a rough manner by comparing the 50kb regions to the left and right of each 1000 bp window. If the difference between the median depth in the left and right regions is small, the frame is not likely to contain a copy-number breakpoint; if the difference is large, it may contain a breakpoint. This measure is smoothed (by moving average) and corrected for linear bias by subtracting a linear trendline. Peaks that exceed 2.5 times the sample-wide median, in absolute value, are annotated as breakpoints. However, this method only gives rough coordinates of discontinuities, delimiting large regions of constant copy number. After this rough estimation, an initial regression was run, and a hidden Markov model (HMM) was used to find regions where the regressed values are significantly different from the data. The HMM accepts deviance of the estimated curve from the data as input signals (15% greater than, 15% lesser than, or approximately equal), and aggregates high densities of deviant signals into output states (under-estimation, over-estimation or correct estimation of copy number; better results were obtained when a special state was reserved for total deletions. The windows where the state changes are seen as likely breakpoints. The regression and HMM were re-evaluated until no more deviating regions could be found. The regression coefficients of the piecewise constant function in the final regression are proportional to the copy number in the corresponding regions, but the proportionality constant depends on the shape and scale of the continuous (spline) factor in the regression, which is different for each chromosome. The form of the spline is such that its value is always 1 in the left telomere for each chromosome. Using the regression coefficients of the piecewise function as a proportional proxy for the copy numbers implicitly assumes that the bias is the same for each chromosome at the left telomere. We observe that the smiley pattern is generally similar on both sides of the chromosomes, so we repeat the regression setting the spline value at the right telomere at 1, and instead use the means of the two sets of regression coefficients to estimate the copy number. Splint was run using frames of 1000bp and 500bp. Shorter frames will result in higher resolution of the CNV calls at the cost of an increased rate of false positive calls. Because results that depend on the window size were not deemed robust, only CNVs found in both 1000bp and 500bp window analyses were used in the final results. The functional enrichment analysis of CNV-driven genes was carried out using the Gorilla database (Eden et al., 2009) using the complete set of *S. cerevisiae* genes as the reference. False discovery rate (FDR) Benjamini & Hochberg adjusted $q \leq 0.05$ were considered significant.

Character Evolution Analysis

Ancestral character states for the production of the phenolic off-flavor 4-vinyl guaiacol (4-VG) were estimated based on the sequences of the two key genes *PAD1* and *FDC1*. The protein-coding nucleotide sequences of *PAD1* and *FDC1* were retrieved from the 157 de novo assemblies obtained in this study and from the outgroup species *S. paradoxus.* Because the annotation procedure described above excluded all genes including internal stop codons, a local BLAST database was set up for all the genomes and BLASTN searches were performed (1E-04 E-value cut-off) using the *PAD1* and *FDC1* coding sequences from the *S. cerevisiae* strain S288c reference genome (R64-2-1). We found that in one bioethanol strain (BI002), both genes resulted from an introgression event from *S. paradoxus*. *S. paradoxus* introgression events involving *PAD1* and *FDC* have been previously reported for the Brazilian

bioethanol strain BG1 (Dunn et al., 2012). In another bioethanol strain (BI006), *PAD1* was shown to be a chimeric gene between the *S. cerevisiae* and *S. paradoxus* allele. This latter strain was therefore excluded from the analysis. Considering the presence of stop codons and frameshift mutations in some of the sequences, the sets of protein-coding sequences for either *PAD1* or *FDC1* were aligned with MACSE (v1.0b), a tool that prevents the disruption of the underlying codon structure when aligning non-functional sequences (Ranwez et al., 2011). Each MSA was partitioned based on codon positions according to three schemes: i) CP_{111} – all codon positions are combined; ii) CP_{112} – first and second codon positions are combined and third position is independent; iii) CP_{123} all codon positions are independent. The optimal partitioning scheme and the best-fit nucleotide substitution model for each partition of the two MSAs were estimated using the software PartitionFinder (v1.1.1) (Lanfear et al., 2012). For all analyses, branch lengths were linked between partitions and 24 substitutions models were considered for each partition. Model and partitioning scheme were selected based on the Bayesian information criterion (BIC). For *PAD1*, the best scheme was obtained with CP_{112}, and K80 and HKY+G (gamma-distributed rate heterogeneity across sites using four rate categories) were the best-fit nucleotide substitution models assigned for the two partitions respectively. For *FDC1* the best scheme was obtained with CP_{123}, and HKY was selected as the best-fit nucleotide substitution model for each of the three partitions. All phylogenetic analyses and ancestral state reconstructions were performed in BEAST (v1.8.2) (Drummond et al., 2012). The trait was treated as discrete and one of two character states were assigned to each isolate based on GC analysis (see further): production of 4-VG, state = 1; no production of 4-VG, state = 0. Monophyly was imposed on all the *S.cerevisiae* strains with the exception of strain BI002. The phylogenetic tree and ancestral state for all internal nodes were simultaneously inferred for each gene, to account for phylogenetic uncertainty. The partitioned MSAs were examined under three different clock models: i) a global molecular clock with fixed evolutionary rates; ii) an uncorrelated relaxed molecular clock with an underlying lognormal distribution (UCLD) on the evolutionary rates; iii) a random local molecular clock (RLC) that allows different evolutionary rates in sub-regions of the phylogenetic tree (Drummond and Suchard, 2010). Additionally, each clock model was tested in combination with an asymmetric versus a symmetric model of trait evolution (Lemey et al., 2009). A pure-birth Yule speciation prior and a random starting tree were used for all the analyses. The suitable number of iterations to allow convergence and proper mixing of the Markov chain Monte Carlo (MCMC) runs was determined for each MSA-molecular clock-trait model combination using Tracer (v1.6) (Rambaut et al., 2014). Effective Sample Size (ESS) > 100 was reached for all parameters in each run. Model selection was performed in BEAST by comparing marginal likelihoods estimated using path sampling and stepping-stone sampling with a chain length of 2 million generations sampling every 200 steps (Baele and Lemey, 2013; Baele et al., 2012). Model comparison showed strong support for the RLC model and asymmetry in the evolution of the trait for both *PAD1* and *FDC1*. A second independent run for both *PAD1* and *FDC1* was performed under the most supported scheme to ensure convergence to the same topology. LogCombiner (Drummond et al., 2012) was used to remove burn-in trees (10%) and to resample to a frequency of 10,000. The final Maximum Clade Credibility (MCC) trees were obtained in TreeAnnotator (Drummond et al., 2012) on 19,998 trees in total for each gene. MCC trees were visualized in FigTree (v1.4.2).

Determination of Cell Ploidy

DNA content of the sequenced strains (a measure for the ploidy level) was determined by staining cells with propidium iodide (PI) and analysis of 50,000 stained cells by flow cytometry on a BD Influx (BD Biosciences, USA). The fluorescent signal of previously established haploid (BY4742), diploid (BY4743) and tetraploid (BR001) strains was used to generate a calibration curve. Highly flocculent strains were excluded from the analysis (missing bar charts Figure 2A).

Phenotypic Analysis

Flavor Production and Flocculation in Fermentation Conditions

To assess the production of aroma-active compounds, lab-scale fermentation experiments were performed. These fermentations were performed in rich growth medium (YPGlu 10%; peptone 2% w v^{-1}, yeast extract 1% w v^{-1}, glucose 10% w v^{-1}). Yeast precultures were shaken overnight at 30°C in test tubes containing 5mL of yeast extract (1% w v^{-1}), peptone (2% w v^{-1}) and glucose (4% w v^{-1}) medium (YPGlu 4%). After 16 hr of growth, 0.5mL of the preculture was used to inoculate 50mL of YPGlu 4% medium in 250mL Erlenmeyer flasks, and this second preculture was shaken at 30°C for 16 hr. This second preculture was used for inoculation of the fermentation medium (YPGlu 10%) at an initial optical density (at 600nm; OD_{600}) of 0.5, roughly equivalent to 10^7 cells mL^{-1}. The fermentations, performed in 250mL Schott bottles with a water lock placed on each bottle, were incubated statically for 7 days at 20°C. Weight loss was measured daily to estimate fermentation progress. After 7 days, the fermentations were stopped, filtered (0.15 mm paper filter) and samples for chromatographic analysis, density and ethanol measurements were taken.

Headspace gas chromatography coupled with flame ionization detection (HS-GC-FID) (Agilent Technologies, USA) was used for the quantification of yeast aroma production. The GC was calibrated for 16 important aroma compounds, including esters (ethyl acetate, isobutyl acetate, propyl acetate, isoamyl acetate, phenyl ethyl acetate, ethyl propionate, ethyl butyrate, ethyl hexanoate, ethyl octanoate, ethyl decanoate), higher alcohols (isoamyl alcohol, isobutanol, butanol, phenyl ethanol), acetaldehyde and 4-VG. The GC was equipped with a headspace autosampler (PAL system, CTC analytics, Switzerland) and contained a DB-WAXETER column (length, 30 m; internal diameter, 0.25 mm; layer thickness, 0.5 μm, Agilent Technologies, USA) and N_2 was used as the carrier gas. Samples were heated for 25 min at 70°C in the autosampler. The injector block and FID temperatures were both kept constant at 250°C. Samples of 5mL filtered fermentation medium were collected in 15 ml glass tubes containing 1.75 g of sodium chloride each. These tubes were immediately closed and cooled to minimize evaporation of volatile compounds. The oven temperature

was held at 50°C for 5 min, after which it increased to 80°C at 4°C min^{-1}. Next, it increased to 200°C at 5°C min^{-1} and was held at 200°C for 3 min. Results were analyzed with the Agilent Chemstation software (Agilent Technologies, USA).

Additionally, after fermentation, the flocculation character of each strain was scored visually from 1 (not flocculent) to 6 (extremely flocculent, big flocs).

Ethanol Accumulation Capacity

To assess the maximal ethanol accumulation capacity of all strains, fermentation tests were performed in rich medium containing 35% w v^{-1} glucose. Yeast precultures were shaken overnight at 30°C in test tubes containing 5mL YPGlu 4%. After 16 hr of growth, 0.5mL of the preculture was used to inoculate 50mL of YPGlu 4% medium in 250mL Erlenmeyer flasks, and this second preculture was shaken at 30°C for 16 hr. This preculture was used for inoculation of the fermentation medium (peptone 2% w v^{-1}, yeast extract 1% w v^{-1}, glucose 35% w v^{-1}; YPGlu 35%) at an initial OD_{600} of 0.5, roughly equivalent to 10^7 cells mL^{-1}. The fermentations, performed in 250 ml Schott bottles with a water lock placed on each bottle, were incubated statically for 14 days at 30°C. Weight loss was measured daily to estimate fermentation progress. After 14 days, the fermentations were stopped, filtered (0.15 mm paper filter) and samples for ethanol measurements [performed with the Alcolyzer Beer DMA 4500M (Anton Paar, Austria)] were taken.

Screening for Environmental and Nutrient Stress Tolerance

All strains were tested in robot-assisted, high-throughput spotting assays in several conditions. All isolates were evaluated on YPGlu 2% agar (Yeast Extract 1% w v^{-1}, Peptone 2% w v^{-1}, Glucose 2% w v^{-1}, agar 2% w v^{-1}) for (i) temperature tolerance (4°C - 16°C - 30°C - 40°C), (ii) sugar- and/or osmotolerance using increasing concentrations of glucose and sorbitol (final osmolyte concentration of 46 - 48 - 50% w v^{-1}), (iii) acid tolerance using increasing concentrations of acetic (50 - 75 - 100mM), levulinic (25 - 50 - 75mM) and formic acid (50 - 75mM), (iv) sulphite tolerance using increasing concentrations of SO_2 (1.50 - 2.25 - 3.00mM), (v) ethanol tolerance using increasing concentrations of ethanol (5 - 7 - 9 - 10 - 11 - 12 - 13% v v^{-1}), (vi) actidione (= cycloheximide) tolerance using increasing concentrations of actidione (0.2 - 0.4 mgL^{-1}), (vii) halotolerance using increasing concentrations of NaCl (250 - 500 - 100mM) and KCl (500 - 1000 – 1500mM) and (viii) metal tolerance using 0.075mM copper and increasing concentrations of cadmium (0.3 - 0.4 - 0.5mM). Stressor concentrations were selected based on previous pilot experiments with a broader concentration range (data not shown). Additionally, temperature tolerance (10°C - 39°C) on three different carbon sources (glucose, fructose, sucrose, ethanol and maltose) was assessed on YP agar supplemented with 2% w v^{-1} of one of the carbon sources (or, in case of ethanol, 2% v v^{-1}). For each of these experiments, growth on YPGlu 2% agar on 20°C was used as a control condition.

For utilization of different carbon sources, experiments were performed on SC (Synthetic Complete) agar containing 2% w v^{-1} galactose, glycerol, melibiose, sorbitol, ethanol, fructose, sucrose or maltose as sole carbon source. Additionally, maltose and maltotriose were assessed in liquid medium (see further). Growth on SC 2% glucose at 20°C was used as a control.

Prior to the experiment, the 96-well microtiter plates containing the isolates (stored at −80°C) were thawed and spotted on YPGlu 2% agar and incubated at 30°C for 48 hr. Next, 96-well plates containing 150μl of YPGlu 2% in each well were inoculated with the isolates and incubated overnight at 30°C on a microtiter plate shaking platform (Heidolph Instruments, Germany) at 600rpm, allowing the cells to reach stationary phase. Then, the OD_{600} of all wells was measured using a microtiter plate reader (Molecular Devices, Sunnyvale, USA). Subsequently, the cell density was manually adjusted to $OD_{600} \approx 0.1$ in a second 96-well microtiter plate using sterile deionized water in order to standardize the starting cell density for all isolates. This plate was used as the source plate for spotting the test media. In order to maximize throughput and reproducibility, a high-density array robot (Singer Instruments, UK) was used for all spotting or replication steps. After spotting, all plates were sealed using plastic paraffin film and all plates (except for plates used in the thermotolerance assays) were incubated at 20°C. After 4-14 days of incubation (depending on the experiment), all plates were scanned using a high-definition scanner (Seiko Epson, Japan). Scanned images were processed using ImageJ combined with the ScreenMill macro (Dittmar et al., 2010). Data were processed by calculating relative growth compared to the control condition, and subsequent normalization by conversion to z-scores (Table S5). Heat maps were obtained using the Gene-E software (http://www.broadinstitute.org/cancer/software/GENE-E/). Strains were hierarchically clustered based on phenotypic behavior using a centered Pearson correlation metric and average linkage mapping.

Maltose and Maltotriose Fermentation Capacity in Liquid Medium

For maltose and maltotriose fermentation capacity, experiments were performed in 96 well plates with 150 μl SC liquid medium containing 1% (w v^{-1}) of maltose or maltotriose, supplemented with 3 mg L^{-1} antimycin to block respiration. Pregrowth was performed as described above, and cells were inoculated at $OD_{600} \approx 0.1$. OD_{600} was assessed after 4 days of growth at 20°C (shaken, 900rpm). Growth in SC liquid medium containing 1% glucose at 20°C was used as a control.

4-VG Production

Screening for 4-VG production was assessed by measuring 4-VG production of each strain in medium enriched in the 4-VG precursor, ferulic acid. Strains were pregrown on YPGlu 2% agar, a single colony was picked and directly inoculated in a GC vial filled with 5mL of test medium (YPGlu 2% supplemented with 100mg L^{-1} ferulic acid). Vials were capped (but not completely closed) and wrapped with plastic paraffin film and statically incubated at 30°C for 3 days. Next, 4-VG concentration was measured using GC-FID as described earlier. Strains were scored as 4-VG^+ if the concentration produced was significantly higher compared to the non-inoculated fermentation medium.

Investigation of the Yeast's Sexual Life Cycle

Sporulation was induced on minimal sporulation medium [1% (w v^{-1}) KAc, 0.05% (w v^{-1}) amino acids, 2% (w v^{-1}) agar] at 23°C after pre-growth in YPGlu 2%. Tetrad dissection of 4 tetrads of each strains was carried out using a Singer micromanipulator (Singer Instruments, UK), and mating-type determination of all germinated spores was carried out by mating-type PCR.

Development of Artificial Hybrids

Sporulation, tetrad dissection, and mating type characterizations were performed as described earlier. For SNP genotyping, PCR primers were developed for the W497* stop-gained mutation in *FDC1*. To detect segregants carrying the stop-gained mutation, following primers were used: W497-FW (5′-TGCAGATCAGATGGCTTTTG-3′), W497-RV-STOP (5′-GCAATTATTTATATCCGTACCT TTTT-3′). To detect the alternative allele (without the stop-gained mutation), following primers were used: W497-FW (5′-TGCAGATC AGATGGCTTTTG-3′), W497-RV-ALT (5′-GCAATTATTTATATCCGTACCTTTTC-3′).

To hybridize haploid segregants, a direct mating approach as described in Steensels et al., (2014), was performed. The segregants were first streaked to single colonies on a YPGlu 2% agar plate. One colony of each segregant was picked, and both were mixed on a second YPGlu 2% agar plate. Ten microliters of distilled water were added to the mixed cell cultures to increase mixing efficiency. The plates were dried and incubated at room temperature for 10-12 hr. A small fraction of the spot was picked with a toothpick and streaked to single colonies on a fresh YPGlu 2% agar plate. After 48 hr of incubation, the diploid status of the resulting colonies was verified by mating type PCR, the presence of both genomes in the hybrid by Interdelta Analysis (Legras and Karst, 2003).

QUANTIFICATION AND STATISTICAL ANALYSIS

Standard statistical analyses were conducted in RStudio (v.0.98.994) (https://www.rstudio.com/) with custom scripts.

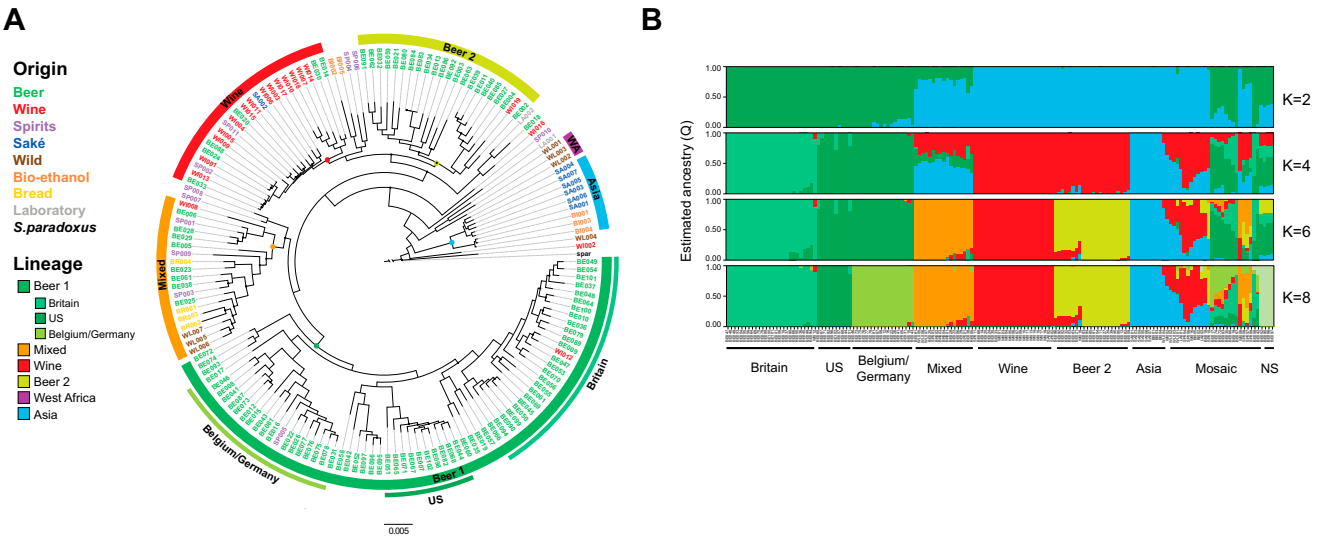
Preprocessing of the phenotypic data to produce Figure 3A and Table S5 consisted of a conversion to z-scores, calculated as follows: z-score = $(X_i - \mu)/\sigma$, where X_i is the data value for strain i, μ the mean of all strains and σ the standard deviation across all strains. To facilitate direct comparison between different strains for a specific trait, a reference condition for each environmental stressor-related trait was determined. This reference condition is defined as the most stringent condition (i.e., the condition with the highest stressor concentration or most extreme temperature) where around 50% of the investigated strains still managed to reach a colony area greater than 10% of their colony area in the control condition (YPGlu 2% agar, 20°C).

The phenotypic variability explained by the set of mutations identified in *MAL11* gene was computed as following: first, all SNPs and InDels were searched in pairwise comparison for high correlations (> 0.90). A Ward clustering was additionally performed to retain or exclude mutations according to the clusters identified. Second, a linear regression analysis was performed on individual mutations; the mutation was retained if significantly associated with the phenotype, with $p < 0.001$ considered as significant. Last, REML (restricted maximum likelihood) analysis with completely random effects was carried out on the final set of SNPs in SAS (v9.4).

DATA AND SOFTWARE AVAILABILITY

Data Resources

The accession numbers for the de novo assembly data reported in this paper are DDBJ/ENA/GenBank: Bioproject ID, PRJNA323691, Biosample ID SAMN05190362-SAMN05190518, and MBUB00000000-MCAB00000000 (Table S1).



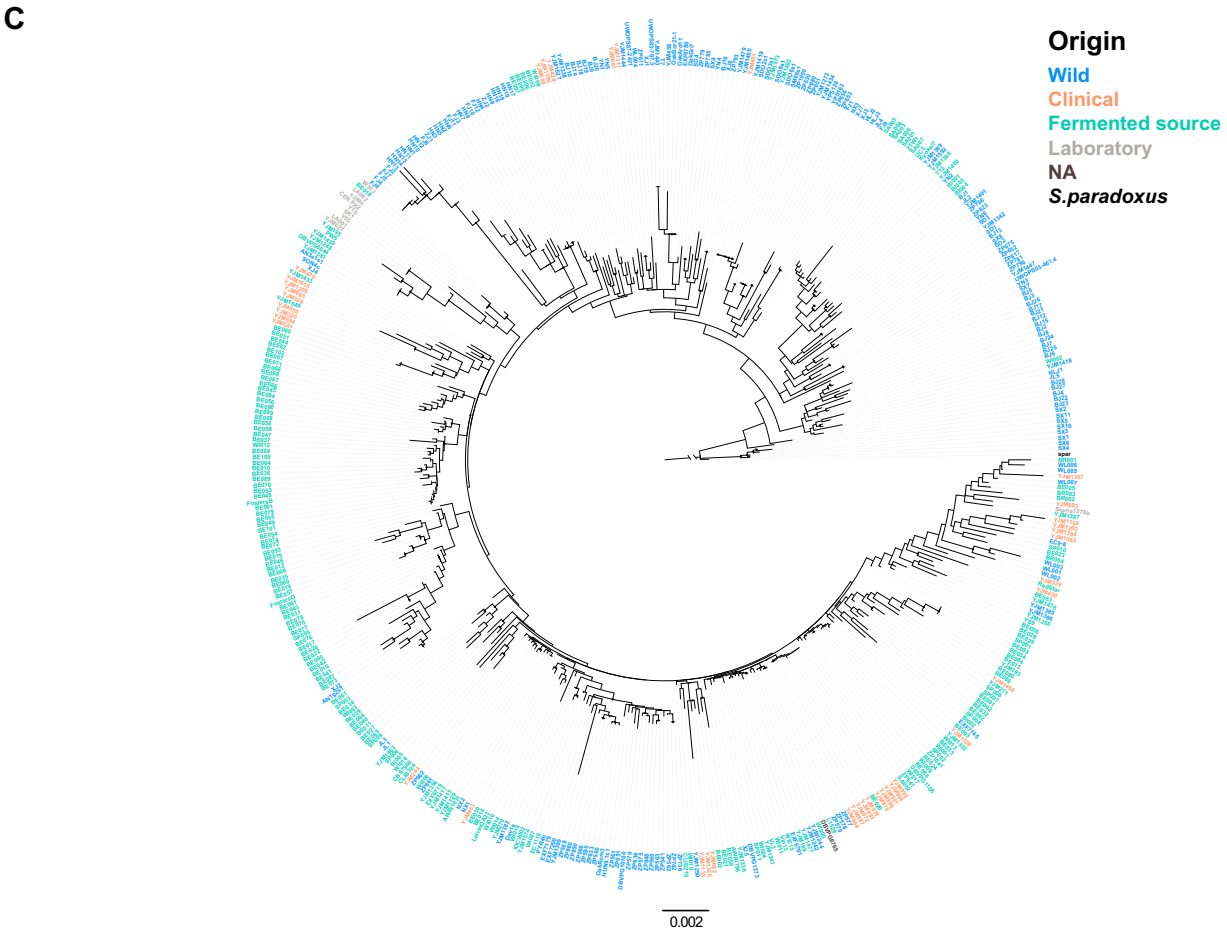


Figure S1. Phylogeny and Population Structure of the Industrial *S. cerevisiae* Strains, Related to Figures 1A–1C

(A) Phylogenetic tree of strains sequenced in this study, using *Saccharomyces paradoxus* as an outgroup. The tree was inferred from the concatenation matrix of 2,020 single copy orthologs. Black dots on nodes indicate bootstrap support values < 70%. Color codes indicate origin (names) and lineage (circular bands). The basal splits of the five industrial lineages are indicated with a colored dot. Branch lengths reflect the average number of substitutions per site (scale bar = 0.005 substitutions per site).

(B) Population structure plot, with strain codes indicated.

(legend continued on next page)

(C) Maximum likelihood phylogenetic tree inferred from a concatenated alignment of nine partial genes of 450 *S. cerevisiae* isolates, using *S. paradoxus* as an outgroup. Strains are colored according to origin. For a list of the included strains and corresponding references, see Table S2. Branch lengths reflect the average number of substitutions per site (scale bar = 0.002 substitutions per site). Dots indicate nodes with bootstrap support values > 50%. Font color codes indicate origin: wild (blue), clinical (orange), fermented source (green), laboratory (gray), not available (NA) (dark gray), *S. paradoxus* (black).

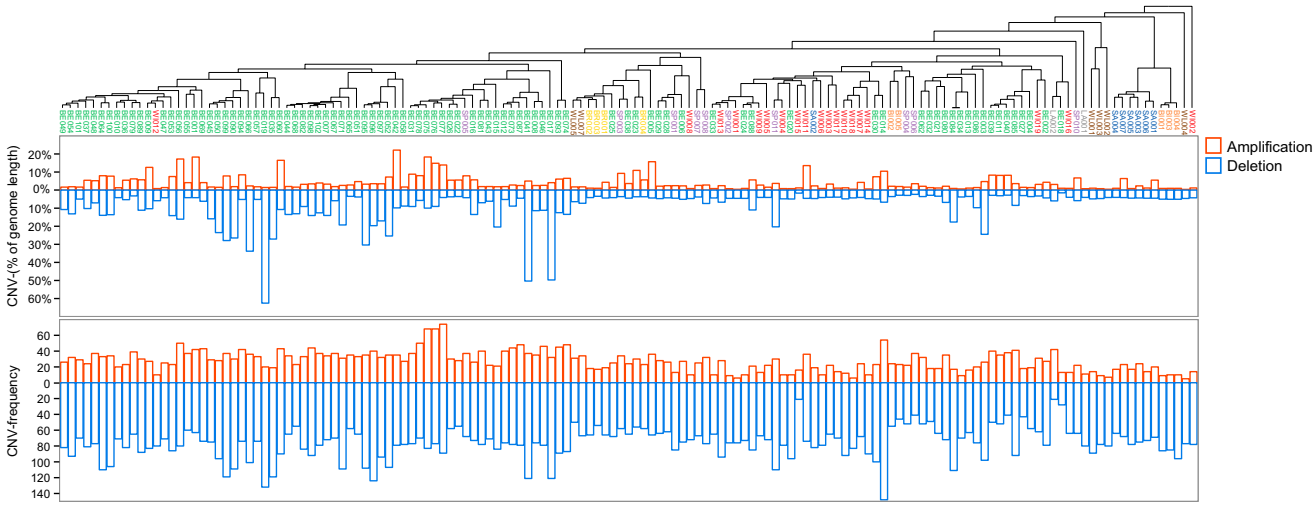


Figure S2. Copy-Number Variability across the Phylogenetic Tree, Related to Figure 2

Distribution of amplifications (red) and deletions (blue) in each strain, expressed in percentage of the genome affected (top) and in number of CNV events (bottom). The phylogenetic tree is described in Figure S1A.

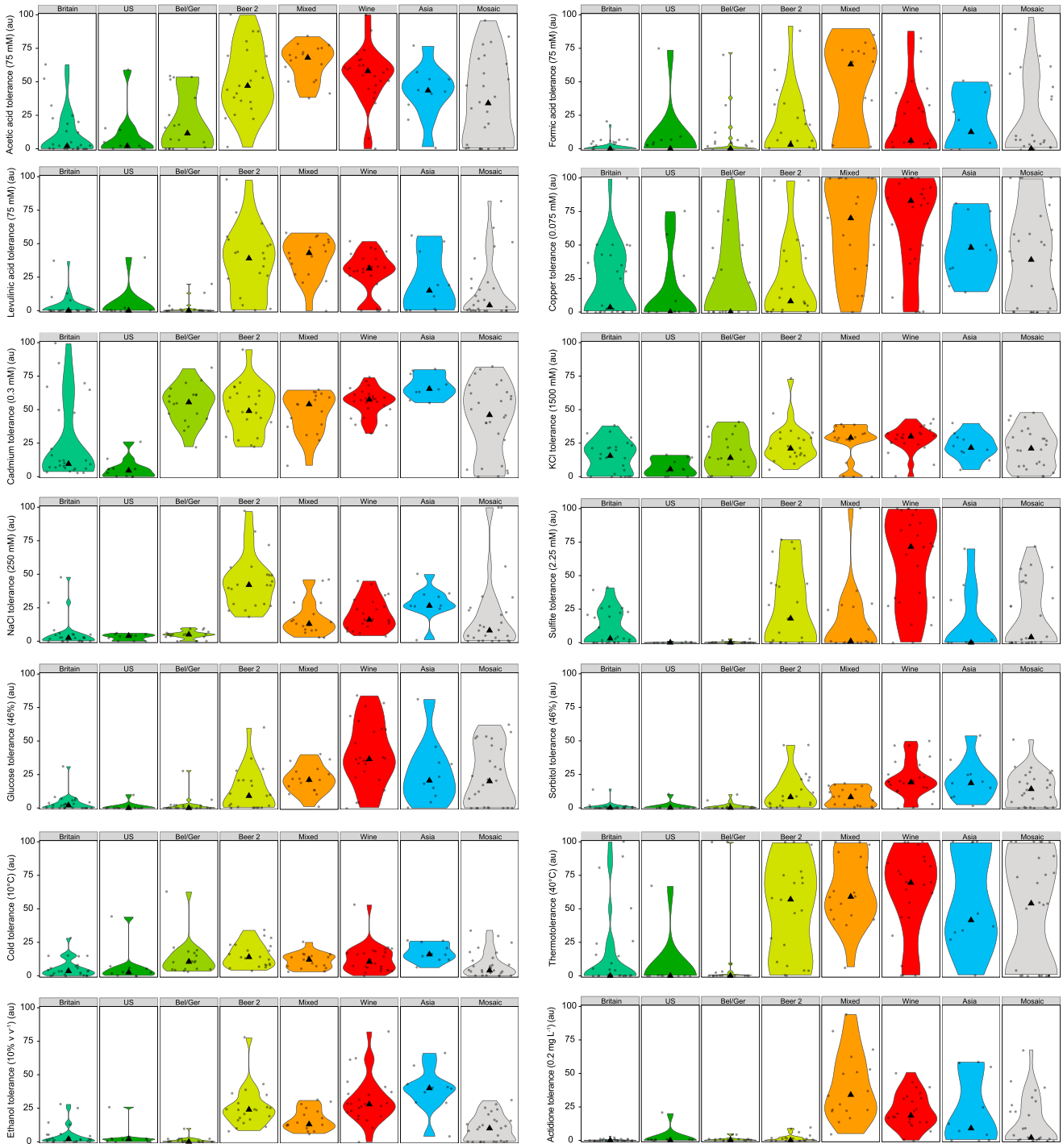
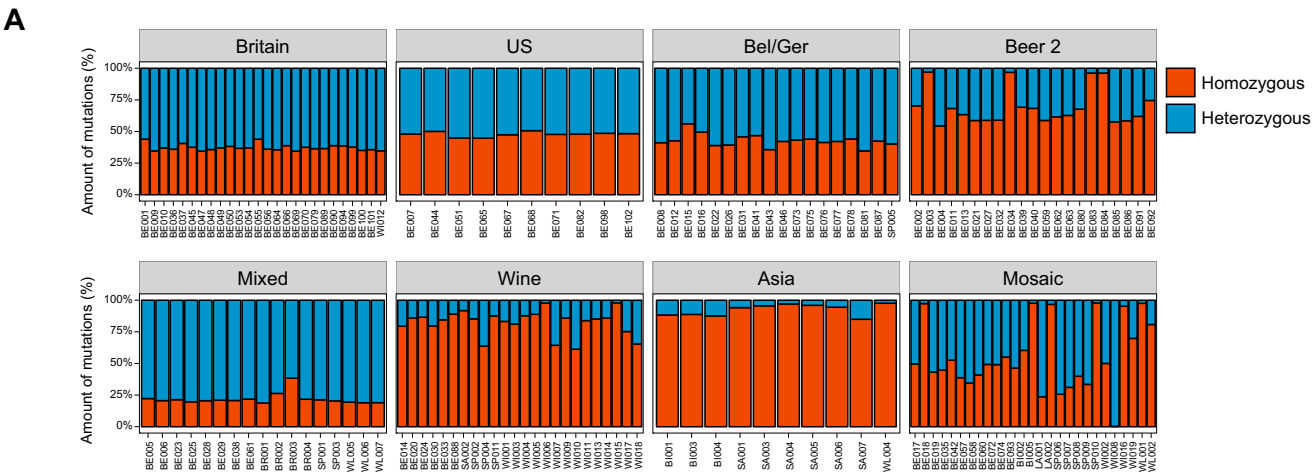
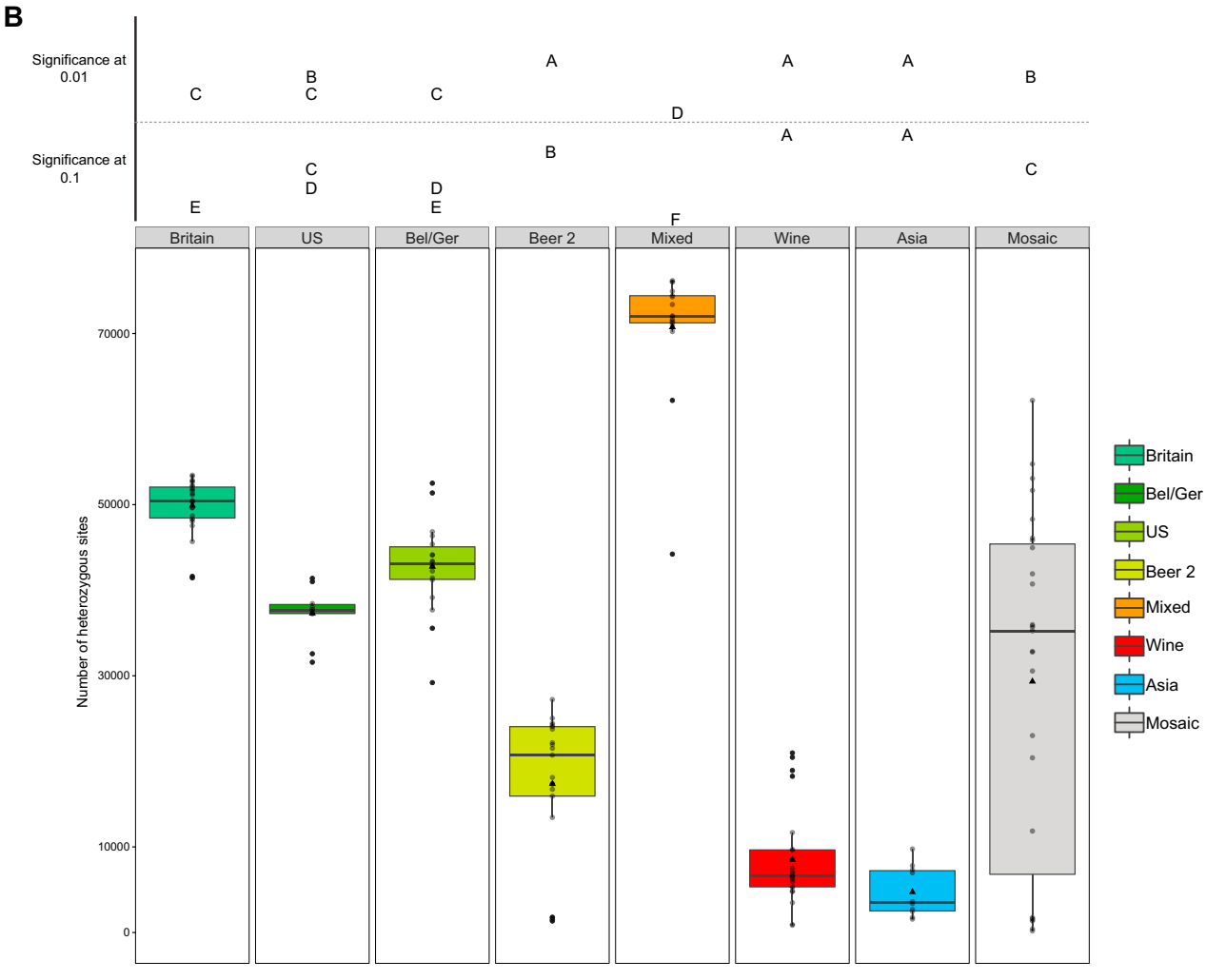


Figure S3. Trait Variation in Industrial *S. cerevisiae* Strains, Related to Figure 3

Graphic representation (violin plots) of trait variation within and between the subpopulations for different environmental stressors. Triangles represent median values for each subpopulation. All values are depicted as relative growth compared to growth on medium without the stressor. Statistical analysis for each trait is given in Table S6. au = arbitrary units.





(legend on next page)

Figure S4. Heterozygosity of Industrial Yeasts, Related to Figure 4

(A) Percentage of total SNPs identified as heterozygous or homozygous in each strain. Boxes depict subpopulations and bar colors indicate the percentage of homozygous (red) and heterozygous (blue) SNPs.

(B) Box plots depicting the total number of heterozygous sites per subpopulation. The mean number of heterozygous sites for each comparison group is indicated by a triangle and the median by a horizontal line. Groups sharing the same letter (top) are not significantly different at the 10% or 1% level.