Article

Current Biology

Phased polyploid genomes provide deeper insight into the multiple origins of domesticated *Saccharomyces cerevisiae* beer yeasts

Highlights

- We sequenced and phased 35 *S. cerevisiae* polyploid beer-related genomes
- The three main beer-yeast subpopulations were derived from different admixtures
- Highly divergent genes are related to functions relevant to the brewing environment
- Phased genomes revealed differential evolutionary trajectories of individual genes

Authors

Omar Abou Saada, Andreas Tsouris, Chris Large, Anne Friedrich, Maitreya J. Dunham, Joseph Schacherer

Correspondence

schacherer@unistra.fr

In brief

Abou Saada et al. sequenced and phased the genomes of 35 beer-related *S. cerevisiae* genomes, providing insight into the genetic makeup of this polyphyletic population. They show that highly divergent genes reflect adaptations to the brewing environment and highlight independent, parallel domestication events in *ADH2*, *MAL11*, and *PAD1*.

Saada et al., 2022, Current Biology *32*, 1350–1361
March 28, 2022
https://doi.org/10.1016/j.cub.2022.01.068

CellPress

CellPress

Current Biology

Article

Phased polyploid genomes provide deeper insight into the multiple origins of domesticated *Saccharomyces cerevisiae* beer yeasts

Omar Abou Saada,[1] Andreas Tsouris,[1] Chris Large,[2] Anne Friedrich,[1] Maitreya J. Dunham,[2] and Joseph Schacherer[1,3,4,5,*]
[1]Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France
[2]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[3]Institut Universitaire de France (IUF), Paris, France
[4]Twitter: @jo_schacherer
[5]Lead contact
*Correspondence: schacherer@unistra.fr
https://doi.org/10.1016/j.cub.2022.01.068

SUMMARY

Yeasts, and in particular *Saccharomyces cerevisiae*, have been used for brewing beer for thousands of years. Population genomic surveys highlighted that beer yeasts are polyphyletic, with the emergence of different domesticated subpopulations characterized by high genetic diversity and ploidy level. However, the different origins of these subpopulations are still unclear as reconstruction of polyploid genomes is required. To gain better insight into the differential evolutionary trajectories, we sequenced the genomes of 35 *Saccharomyces cerevisiae* isolates coming from different beer-brewing clades, using a long-read sequencing strategy. By phasing the genomes and using a windowed approach, we identified three main beer subpopulations based on allelic content (European dominant, Asian dominant, and African beer). They were derived from different admixtures between populations and are characterized by distinctive genomic patterns. By comparing the fully phased genes, the most diverse in our dataset are enriched for functions relevant to the brewing environment such as carbon metabolism, oxidoreduction, and cell wall organization activity. Finally, independent domestication, evolution, and adaptation events across subpopulations were also highlighted by investigating specific genes previously linked to the brewing process. Altogether, our analysis based on phased polyploid genomes has led to new insight into the contrasting evolutionary history of beer isolates.

INTRODUCTION

The yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) is a well-studied model organism with a long history of human domestication due to its fermentation ability. It has unknowingly been leveraged by early humans to ferment foods and has been domesticated in various ecological niches. Notably, there is evidence of the domestication of *S. cerevisiae* in the cheese, wine, bread, sake, cacao, coffee bean, and beer industries.[1–8] The domestication process began long before Louis Pasteur's identification of the brewer's yeast *S. cerevisiae* and Emil Hansen's isolation of pure cultures for use in the Carlsberg brewery in 1883,[9] likely accelerated through backslopping: the practice of collecting a part of the fermentation product that still contains living cells and using it to inoculate the next fermentation, thereby improving its efficiency. Backslopping is a driver of domestication, which can accelerate the adaptation of yeasts to human preferences.[10] It is particularly illustrated through the widespread inactivation of two genes, *PAD1* and *FDC1*, whose product 4-vinyl guaiacol (4VG) produces an undesirable off-flavor in beer,[4] though it contributes to a desirable flavor profile in some Belgian beers. This adaptation is a striking example of domestication given that the yeasts used in beer brewing are a polyphyletic group, with some strains more closely related to European wine or sake isolates than to other beer strains.[5] Two main industrial beer subpopulations,[4] named Beer 1 and Beer 2, were identified.[4,5,8] The Beer 1 group, mostly composed of polyploid ale strains, has been shown to derive from admixture between close relatives of European and Asian wine strains.[8]

Industrialized beer-brewing strains have had to adapt only to the brewing environment, which typically has high alcohol concentrations, high osmotic pressure, and low pH. Adaptations can mean remodeling the cell wall,[11,12] degrading protein aggregates[13] caused by ethanol denaturation, controlling pH by vacuolar acidification,[14] or controlling osmotic pressure via the inactivation of aquaporins.[5] However, the life cycle of these industrialized strains also shields them from the wild, in which they have reduced fitness.[4] African beers, on the other hand, do not undergo these same industrialization processes. Similar to wine yeasts, which cannot grow in grape must year round and have to maintain their ability to survive in vineyard environments, African beer yeasts must remain adapted to their local environments as traditional African fermentation methods offer less stable environments than industrial methods. Traditional African beer-making methods rely on the presence of native *S. cerevisiae* (and other yeasts) on the brewing ingredients.

The fermentation processes for African beers typically start with an initial spontaneous fermentation usually driven by lactic acid bacteria (LAB).[15,16] An alcoholic fermentation follows, either spontaneously[16,17] by explicit backslopping methods[18] or indirectly by backslopping methods such as the reuse of tools or containers that allow well-adapted microorganisms from successful previous fermentations to drive the fermentation process.[15,16] The life cycle of African beer yeasts contrasts with the industrialized, highly specialized beer-brewing yeasts grown as pure cultures.[19,20] Backslopping, a known driver of domestication,[4] has certainly shaped the genomes of industrialized beer-brewing yeasts, and very plausibly that of African beer yeasts as well, though less extensively. Comparing the two groups may reveal genes relevant to adaptations to brewing environments and uncover convergent evolution processes.

2In this study, we further characterize the origins of modern industrial ale-brewing strains, finding that the previously described Beer 1 and Beer 2 groups differ in the proportion of European/Asian alleles, renaming the groups to Asian dominant and European dominant, and show that the alleles of the African beer strains are closest to European wine and French dairy. We also phased the genomes of all 35 strains and determined the genetic distances between strains and between groups, finding that the mean divergence between African beer strains and modern ale-brewing strains is under 0.35%. Using phased genome data, we calculated the intra-strain divergence and found that the Asian dominant strains have the highest mean intra-strain divergence at 0.21%, followed closely by European dominant and African beer strains at 0.20% and 0.16% divergence, respectively. By comparing the fully phased genes in our dataset, we determined the level of divergence between gene haplotypes and identified those that reach the highest level of pairwise diversity. We detected genes of interest such as *ROQ1* that are required for denatured protein degradation, *YPS* genes that are involved in cell wall remodeling, and several *IMA* genes that are involved in isomaltose utilization.[12,13,21] Finally, we also investigated genes that present evidence of domestication (*MAL11*, *PAD1*, *FDC1*, *GAL2*, *ADH2*, and *SFA1*) and provided evidence of convergent evolution in the loss of function of the *FDC1* gene in African beer and Asian dominant groups. We also found that the *ADH2* and *SFA1* genes appear to be undergoing the same human selection as the *PAD1* and *FDC1* genes to suit human preferences for beer flavor by reducing fusel alcohol formation,[22] a source of off-flavors in beer when present in high concentrations.[23]

RESULTS

Selection of beer isolates, sequencing and genome phasing

To dissect the genetic diversity and genomic architecture of beer-brewing yeasts, we selected 35 strains from diverse clades based either on their known use in fermenting beers or on their high genetic similarity to beer-brewing strains (Table S1). Beer-brewing strains of *S. cerevisiae* are polyphyletic, forming at least three distinct clades: one clade of African beer strains and two clades of modern ale strains named Beer 1 and Beer 2, which are believed to have different origins.[4] It has been shown that Beer 1 strains are a polyploid admixture of European and Asian wine strains.[8] The African beer and Beer 1 groups typically have higher ploidies, between 3n and 5n for the African beer group and typically 4n for the Beer 1 group. Strains from the Beer 2 group are not typically polyploid. Adding to the diversity and complexity of the population of beer yeasts, some isolates used in breweries have genomes consistent with European wine strains, and others have genomes that cluster with other strains of mixed origins. For our study, we selected 8 African beer strains, 16 Beer 1 strains, and 5 Beer 2 strains. We also selected 6 beer-brewing strains from outside of these three clades, including 2 European wine strains isolated from breweries, 3 strains from the mixed-origin clade and 1 from the mosaic region 1 clade.[6]

Given the polyploid nature of a majority of the strains selected, we sequenced all 35 strains with Oxford Nanopore long reads in order to phase their genomes. We used publicly available short-read data for nearly all of the strains selected[6,8] (Table S2). Only strain YMD4285 was sequenced by Illumina for this study. We aimed to obtain at least 80X theoretical coverage with our long reads for most strains in order to obtain accurate and contiguous phasing results. We reached the target of 80X coverage in 26 out of 35 strains, with the remaining 9 ranging from 14.4X theoretical coverage to 77.6X (Table S2). Whenever possible, we downsampled our long-read data to 80X of the best reads, obtaining mean read lengths up to 37.7 kb (mean 21.2 kb) and mean read quality scores up to 15.9 (mean 14.4). In cases where we could not downsample to 80X, we used all of the sequencing reads for our analyses. The Illumina short-read data we used ranges from 144X to 368X (mean 281X), with mean quality scores ranging from 28.8 to 34.8 (mean 32.9).

We recently developed a phasing algorithm and pipeline, nPhase,[24] which phases a genome using short reads, long reads, and a reference sequence. The short reads are mapped to the reference sequence and variant called, which serves as a list of high-confidence SNP positions. The long reads are also mapped to the same reference sequence and iteratively clustered together, according to the similarity between reads at these previously defined SNP positions. The iterative clustering ends when only distinct clusters remain, which are different from each other. nPhase is a ploidy agnostic phasing method, and it makes no attempt to coerce the results to a given or estimated ploidy; it only detects when the existing clusters should not be merged together. nPhase also provides a cleaning algorithm, which removes small clusters, and attempts to improve the contiguity of phasing results and reduce noise at little cost to accuracy by applying simple heuristics.[24] We used our dataset of accurate short reads and phase-informative long reads to phase all 35 strains using nPhase,[24] applying the nPhase cleaning algorithm to improve the contiguity of our results (Figure 1).

Without a ground truth, we could not assess the accuracy of our phasing results; however, we could assess their contiguity. We used the L90 metric, which we define here as the minimal number of haplotigs to cover at least 90% of all reads, and the L90 per chromosome, which is simply the L90 divided by the number of chromosomes times the ploidy. The L90 per chromosome for the phasing of a triploid strain of *S. cerevisiae* is therefore the L90 divided by 3*16. If the value is close to 1, we have close to a contiguous phasing; if it is much higher, the phasing

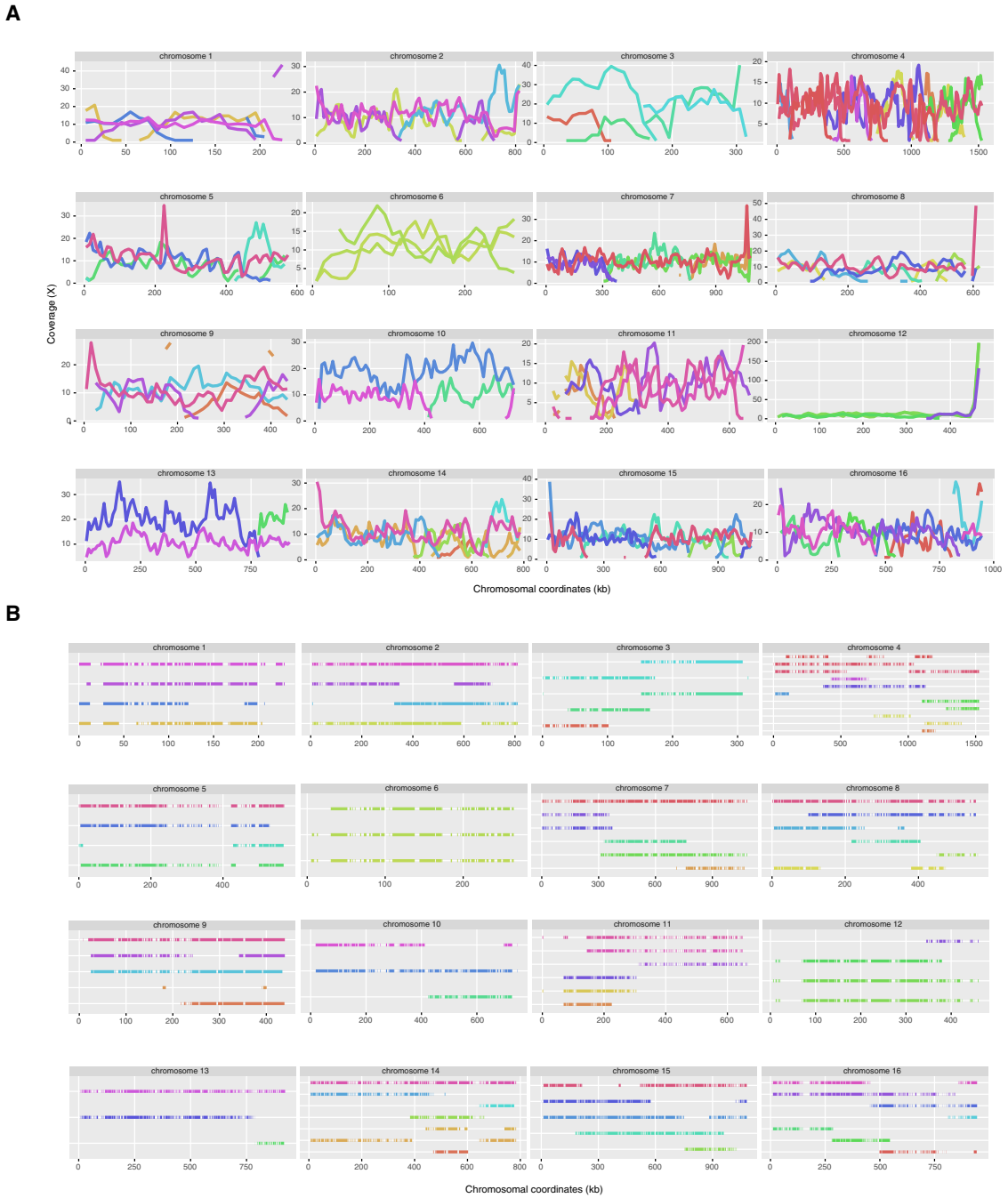

Figure 1. Cleaned phasing results for all 35 strains

(A and B) For each strain we present two plots drawn for its cleaned nPhase phasing predictions, generated from the raw predictions. In (A) we show the coverage of each predicted haplotig, in (B) we provide an overview of where the haplotigs are along the genome. We present here the results for strain AQH; the results for the remaining 34 strains can be found on this Zenodo repository: https://zenodo.org/record/5718147. See also Figure S2.

is increasingly fragmented, and if much lower, increasingly likely not to have correctly distinguished between haplotypes. We report in our raw results an L90 per chromosome that ranges from 1 to 2.6 (with an outlier at 4.3 due to low long-read coverage), with a mean L90 per chromosome of 1.6 (Table S1). After applying the cleaning pipeline available for nPhase, we improved the contiguity, reducing the range of L90 per chromosome to between 0.8 and 2.3 (with the same outlier at 3.6). The cleaning step also substantially reduced the average total number of haplotigs from 198 to 100.

The phasing correctly predicted a number of suspected and known aneuploidies such as the 6 different chromosome losses of the tetraploid strain BBG and the extra copy of chromosome 3 in the diploid strain CPB, though not all, with aneuploidies in

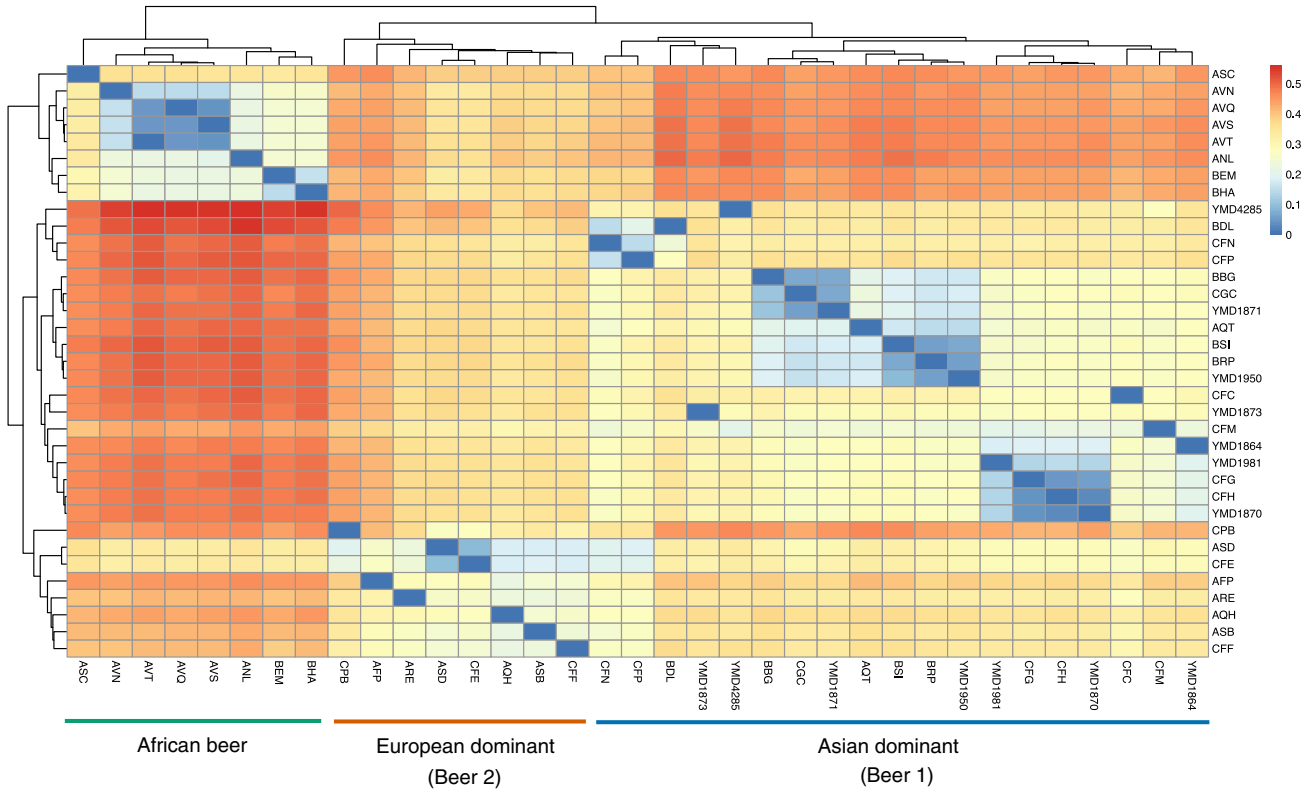

Figure 2. Inter-strain divergence levels using 10 kb haplotype windows

This heatmap represents the mean inter-strain divergence between each pair of the 35 strains used in this study. The values were calculated by comparing all 10 kb haplotype windows between strains and range from 0% divergence for strains compared to themselves to 0.56% between the most different strains. Hierarchical clustering was performed and suggests that the strains can be divided into three main groups; strains CFP and CFN are attributed to the right-most group despite an ambiguous profile suggesting close similarity to the group at the center of the heatmap. See also Tables S5, S6, and S7.

strains such as CFM remaining unclear after phasing, potentially due to lower read lengths (Table S4).

Inter-strain divergence reveals three groups of strains

The standard way of estimating the divergence between two strains uses unphased genomes to calculate their distance based on allelic differences. Using this method, we obtained a mean inter-strain divergence of 0.58% across all strains, with a maximum divergence of 1% when comparing AVS and YMD4285 (Table S5). This method is not well adapted to polyploidy, as it does not take into account the complexity of these genomes, leading to inaccurate representations of the differences in genetic content between strains. It does not, for example, reveal whether two strains may have a subgenome or haplotype in common. There are two main barriers to obtaining this type of information: it requires access to phased haplotypes, and the question is complicated by recombination events and loss of heterozygosity (LOH) events. Using our dataset of phased haplotypes, we propose a more accurate metric of inter-strain divergence for polyploids that takes their haplotypes into account. For each pair of strains, A and B for example, we calculate the distance between 10-kb regions of all haplotypes, keeping only the match with the lowest divergence. We allow a 10-kb region of a haplotype in strain B to match with several 10-kb regions in strain A's haplotypes. Under this definition, the similarity between A and B can be different from the similarity between B and A. This calculation of inter-strain divergence allows us to estimate divergence based on the allelic content of each strain (Figure 2). Using this method, we updated our mean inter-strain genetic divergence numbers from 0.58% to 0.36%, and the highest level of inter-strain divergence dropped to 0.56%, obtained when comparing strains ANL and YMD4285 (Table S6). The previously most divergent strains AVS and YMD4285 are 0.55% divergent using this calculation method.

This inter-strain divergence based on haplotypes reveals three main groups of strains in our dataset, defined by a higher similarity to one another than to other strains: the African beer group (8 strains), the Beer 2 group to which we can add two European wine strains and the mosaic region 1 strains (8 strains), and finally the Beer 1 group to which we can add the 3 mixed-origin strains (19 strains). Two of the three mixed-origin strains, CFP and CFN, could arguably be assigned to either group though the third, BDL, resembles the Beer 1 group more closely. Despite the polyphyletic nature of the population, we can reorganize the strains in our dataset into three major groups.

Three main groups differ by proportions and origin of allele content

In order to elucidate the difference between the Beer 1 and Beer 2 groups, and to start characterizing the allele content of African

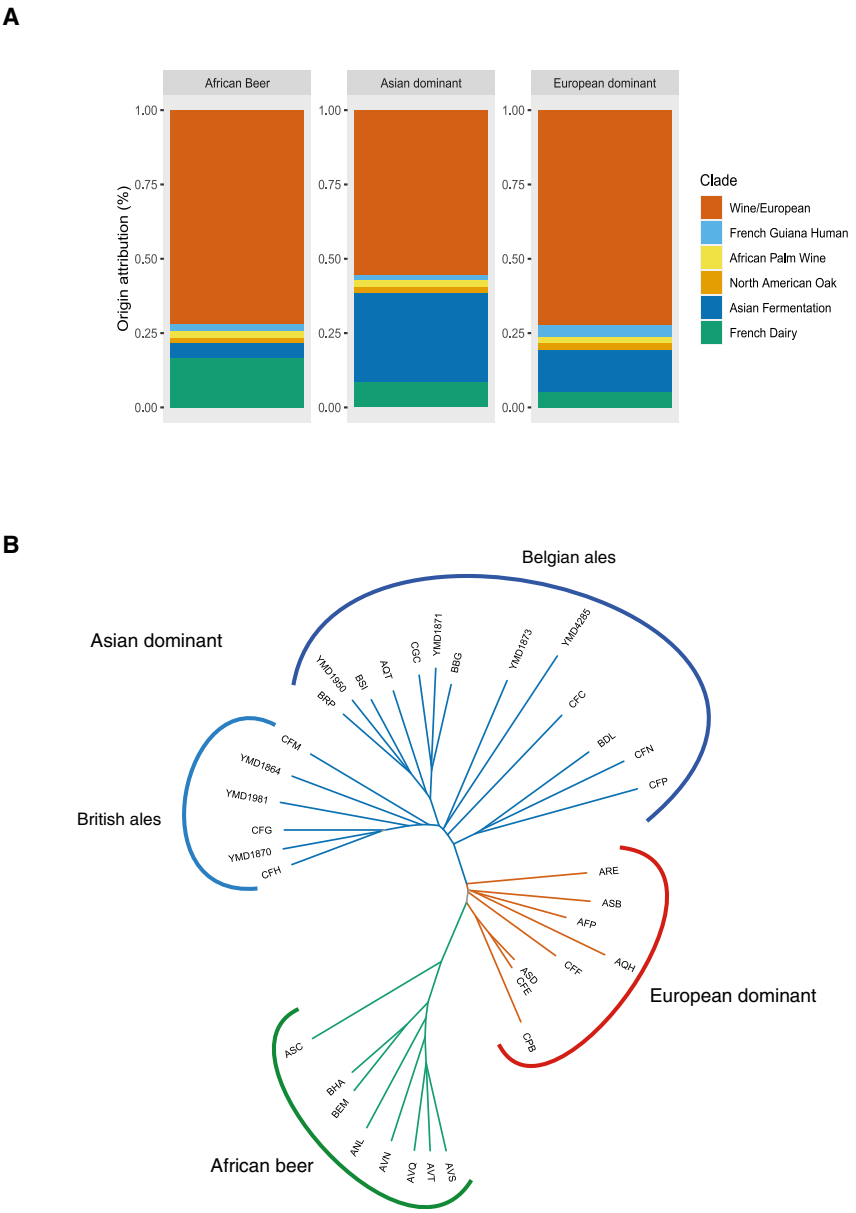

Figure 3. Three groups of beer strains differ by allelic origin

We phased the genomes of all 35 strains, and for each haplotype we identified SNPs that are markers of known clades such as French dairy or European wine. We then attributed each haplotype to the clade with the highest signal, in blocks of 20 kb, finding three different profiles: African beer, European dominant, and Asian dominant.

(A) In this figure, we show that all three groups have a high European wine signal. Strains attributed to the African beer also have a high French dairy signal, whereas the difference between the Asian dominant and European dominant strains is their level of Asian fermentation alleles. The Asian dominant group has a higher signal for Asian fermentation than the European dominant group.

(B) This dendrogram, generated from a SNP matrix using Illumina data, has been colored to represent the origin group attributed to each strain, and it shows that the European dominant group is between the Asian dominant and African beer groups.

beers, we used a windowed approach to compare each strain's haplotypes with their closest match in the clades described in the 1,011 *S. cerevisiae* genomes survey.[6] For each of these previously described clades, we identified the polymorphisms that are specific to it according to the sequencing data of the population.[6] Then, for each strain, we divided each of its haplotypes into 20-kb windows and identified all of the clade-specific polymorphisms in the window. We assigned each 20-kb region of each haplotype to the clade that had the highest signal. We did not use all of the clades described in the 1,011 yeast genomes survey.[6] We excluded clades known to derive from older populations such as Brazilian bioethanol, which shares a close relationship with European wine,[25] and West African cocoa that is an admixture of European wine, Asian fermentation, and North American oak.[3] We also excluded clades for which our dataset had too few strains to contribute sufficient data; this excluded clades such as Ecuadorean and Far East Russian for which we had fewer than 10 strains each. Finally, we obviously excluded the clades we are trying to study. We did not include the beer clades we are investigating as well as the mosaic or mixed-origin clades. We therefore limited our allele content comparison to the following six clades: European wine (for which we merged all wine clades), North American oak, Asian fermentation (we merged sake and Asian fermentation), French dairy, African palm wine, and the French Guiana subpopulations.

Through this windowed approach we can confirm the reorganization of our strains into three groups based on their similar origin profiles (Figure 3A). British and Belgian/German ales and mixed-origin strains (i.e., the Beer 1 group) have the same origin profile (Figure 4), mainly composed of European wine and Asian fermentation alleles, with the largest signal of Asian fermentation markers out of all three groups, forming the Asian dominant group. This method of estimating the origin of the allele content of these strains corroborates the admixed origin previously described[8] (Figure S1). African beers have a large signal of European wine alleles and differ from the other two groups by their higher signal of French dairy alleles. The final group, containing all of the mosaic beers and two European wine strains (i.e., the Beer 2 group), is characterized by its high level of European wine and low but still significant Asian fermentation signal, and it resembles the profile of Asian dominant strains where the balance between Asian fermentation and European wine alleles has been inverted.

We then generated a dendrogram based on all genomic SNPs to place the strains in relation to one another (Figure 3B). We found that the European dominant group is in between the

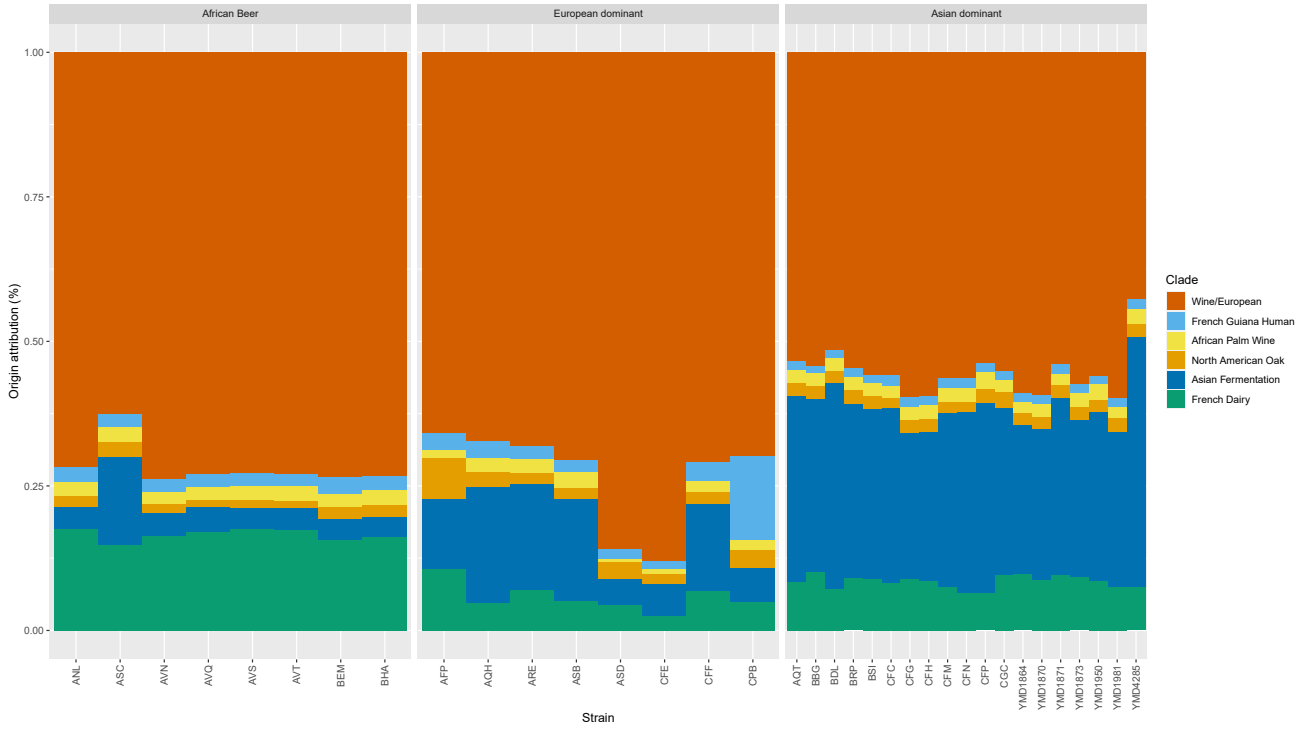

Figure 4. Origin profiles per strain per group

For each strain, we attributed 20-kb regions of its haplotypes to the clade with the highest similarity. This figure shows, for each strain, the proportion of regions attributed to each clade. Each page corresponds to one of the three groups we identified, based on these profiles: European dominant strains, Asian dominant strains, and African beer strains. Strains within the same group have very similar profiles. See also Figure S1.

African beer group and the Asian dominant group. Consistent with previous reports, we also distinguished two ale groups that correspond to geographical origin, i.e., British ales (CFG, CFH, YMD1864, YMD1870, and YMD1981) that cluster together on one branch of the dendrogram, along with USA strain CFM, and Belgian/German ales that cluster on the adjacent branch (BRP, YMD1950, BSI, AQT, and YMD1871), alongside CGC, a USA strain isolated from an olive fly and BBG and a strain isolated from the water of the Morava river in Slovakia. The Belgian/German strains YMD1873, CFC, and CFF are also found along the main branch of the Asian dominant group on this dendrogram, alongside YMD4285, BDL, and CFN. We will hereafter refer to the 5 British strains, and the USA strain that clusters with them, as the British ales and all other Asian dominant strains as the Belgian/German Ales.

We can modify the previously described method of calculating inter-strain divergence, comparing haplotypes within a strain to one another. Using this method, we found that overall African beers are the least self-diverse, with 0.16% mean self-divergence, likely owing to their highly polyploid nature. The most self-diverse are Asian dominant strains with 0.21% self-divergence, and the European dominant strains are not far behind with 0.20% mean self-divergence. The Asian dominant and European dominant strains reach slightly higher self-divergence levels than African beers. The lower extremes of African beer strains are likely due to its higher ploidy, and therefore a higher likelihood exists for each 10-kb region not to be too distant from one of the several other haplotypes.

Intra-strain divergence of African beers varies from 0.12% in the least self-divergent strain to 0.17% in the most self-divergent strain. In European dominant, it varies from 0.05% to 0.27%. The two least self-divergent European dominant strains are at 0.05% and 0.09%, with the third least at a much higher 0.21%. In Asian dominant strains the self-divergence levels vary from 0.09% to 0.35%, though the minimum and maximum are slightly extreme outliers, with the next least self-divergent and next most self-divergent strains at 0.16% and 0.28%, respectively (Table S7).

Genes with highest divergence enriched in functions relevant to brewing environment

We phased the genomes of 35 strains of *S. cerevisiae* associated with beer brewing, selected from a diverse set of isolates comprising three main clades and three associated clades. All of these isolates have had to adapt to the brewing environment or were very closely genetically related to beer-brewing strains. The main groups have adapted independently to the brewing environment, and we expect that a survey of the genes with the most diverse haplotypes in our dataset will reveal genes that have undergone rapid deterioration due to being redundant, pseudogenes that were under no selective pressure, and genes of interest for adaptation to the brewing environment, which were put under selective pressure.

To investigate this, we extracted all of the fully phased genes in our dataset and calculated the pairwise divergence between all phased copies (Figure 5). Phased African beer genes are on average 0.23% divergent from one another, slightly lower than

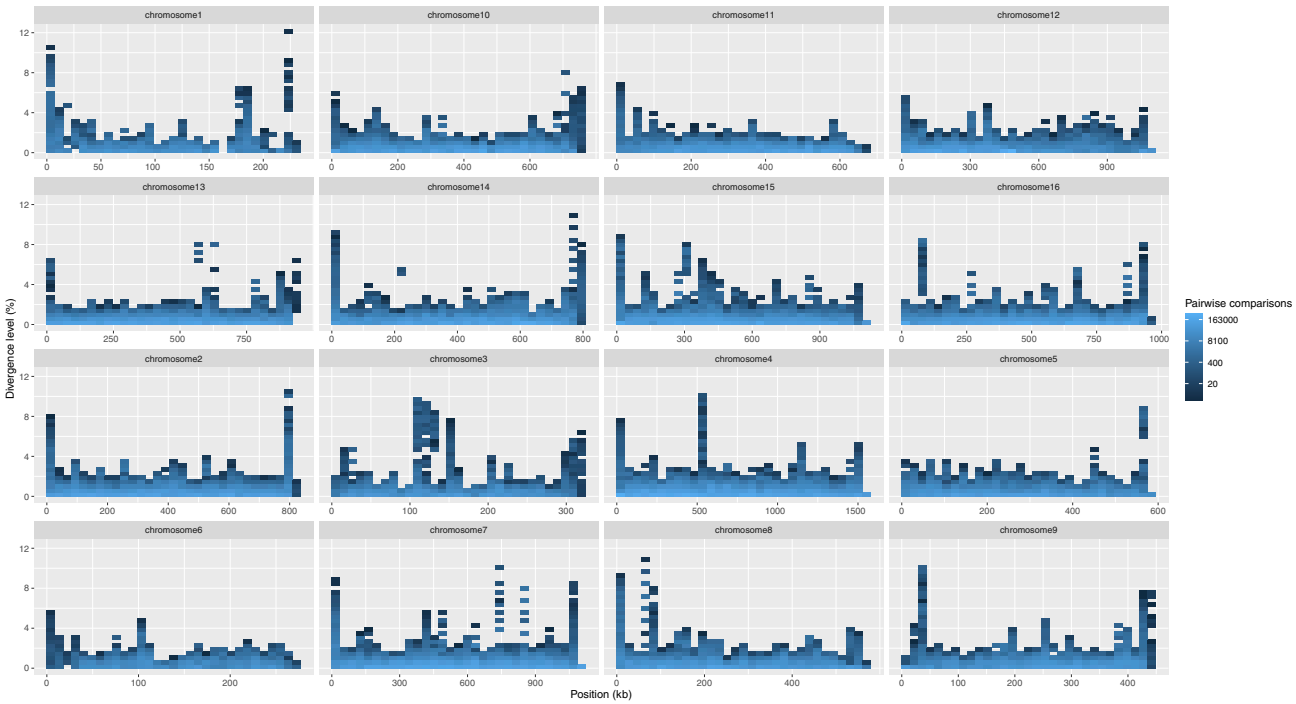

Figure 5. Distribution of divergence levels for fully phased genes using all strains

We identified all of the fully phased genes in our dataset and compared them with one another in a pairwise manner, then plotted the divergence levels and their distribution along the genome as a heatmap. The y axis represents the divergence level as a percentage, and the x axis is the position along the chromosome. The color represents the number of supporting pairwise comparisons, which are shown here by using a log scale for visual clarity. We observe that more highly divergent genes tend to be less frequent and rarely reach above 4% divergence. The subtelomeric regions appear to have more highly divergent genes; however, we also observe many other regions within the chromosomes with high levels of divergence, so they are not limited to the telomeres. See also Tables S8 and S9.

the average 0.24% of European dominant strains and 0.32% of Asian dominant strains. In all groups, a minority of genes show significant divergence levels within their group, reaching over 4% divergence levels. When all phased copies are compared with one another, the average divergence level rises to 0.36%, and a few more gene alleles are found with a pairwise divergence level over 4%, pointing to genes that have very divergent haplotypes across different groups but not necessarily within them.

We then identified the 144 genes that have a pairwise divergence level of 4% or higher (Table S8). Of these 144 genes, 57 had a verified annotation, 54 were uncharacterized, and the remaining 33 were dubious genes. We subjected our list of 57 verified genes to a Gene Ontology (GO) term finder analysis to identify enrichment in processes, function, and cellular component localization. We found that our list of 57 highly divergent genes is enriched for carbon metabolic processes for various carbon sources (e.g., maltose, galactose, sucrose), galactose transport, and cell wall organization. These genes are also enriched in cell wall structural constituents, dehydrogenase activity, and transmembrane sugar transporters, and enriched in genes whose products localize to the cell periphery, cell wall, and vacuoles (Table S9).

Notable genes of interest include *ROQ1*, which directs the SHRED pathway to degrade proteins denatured by high alcohol concentrations,[10] *YPS* genes involved in cell wall remodeling to resist oxidative and osmotic stress,[12,26] and *CTT1*, a catalase expressed in response to oxidative stress.[27] Deeper investigation into the genes highlighted and the diverse haplotypes observed and their potential functional consequences would be of significant interest for further understanding the changes required for wild yeast to adapt to the brewing environment, examples of convergent and/or divergent evolutionary trajectories.

Industrial domestication markers: The *MAL11*, *PAD1*, and *FDC1* genes

Our dataset corroborates and expands on previously reported findings for the *MAL11*, *PAD1*, and *FDC1* genes.[4] These genes, highlighted in Gallone et al.,[4] are evidence of the domestication of beer yeasts to suit industrial needs and human flavor preferences. We describe here our observations for these genes in our dataset (Figures 6A and 6B; Table S10).

Maltose utilization is an industrially relevant phenotype in beer brewing, due to the high maltose content obtained after malting grain. Maltose is typically the main fermentable carbon source in wort, the brewing solution to undergo fermentation. The *MAL11* gene codes for an effective maltose transporter, shown to be present in the Asian dominant strains but inactivated in the European dominant group by frameshift-inducing indels.[4] There are two reported frameshift-inducing indels, 1772CA➥C and 1175A➥AT. We report that *MAL11* is present and intact in half of the African beer strains and absent in the others. In our dataset, *MAL11* suffered inactivation by homozygous frameshift-inducing indels in all European dominant

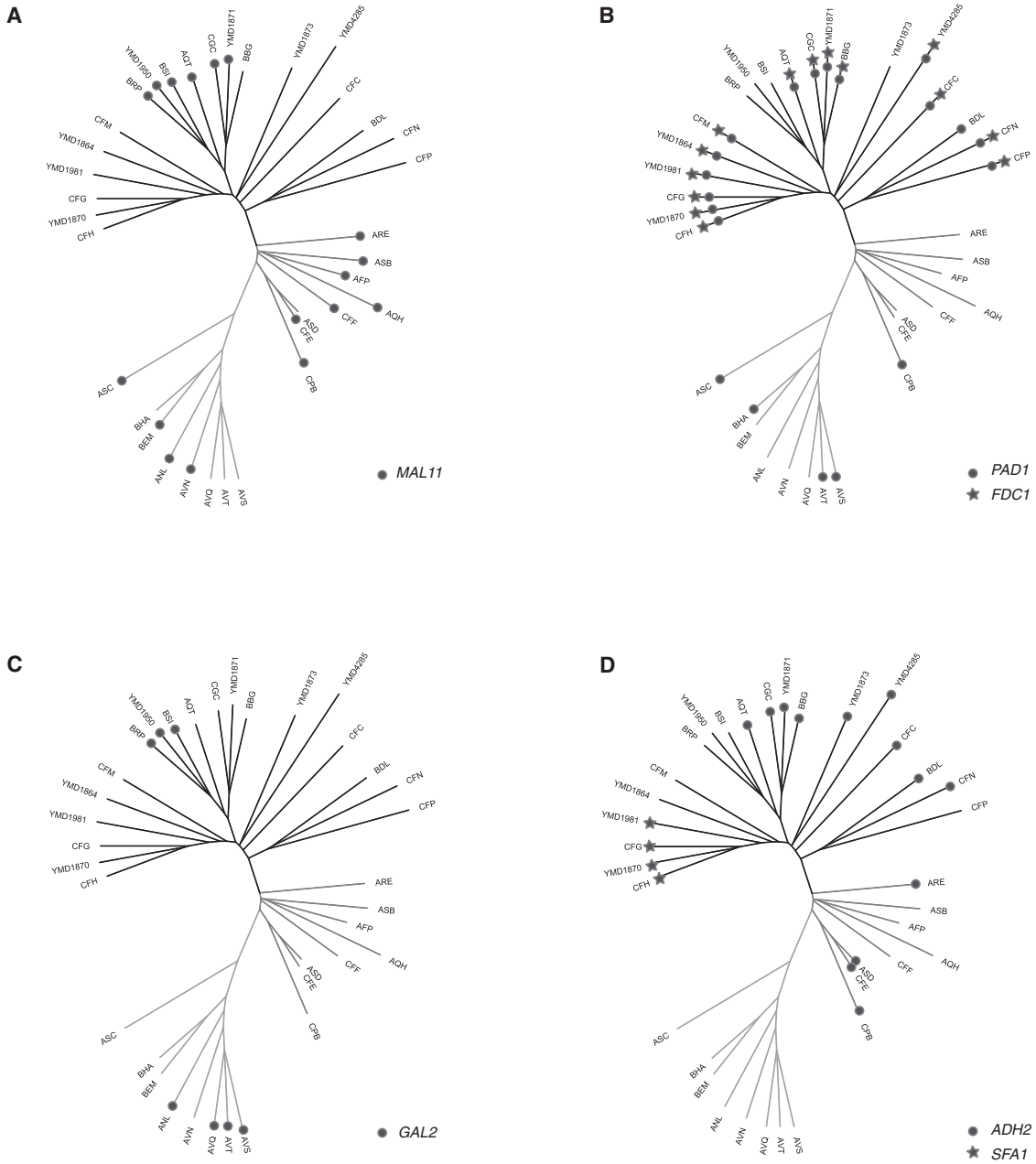

Figure 6. Status of 6 genes of interest in the 35 strains

These four panels are based on the dendrogram in Figure 3B and the detailed haplotype information given in Table S10 for the 6 genes *MAL11*, *PAD1*, *FDC1*, *GAL2*, *ADH2*, and *SFA1* for all 35 strains. For each gene and strain, we display a black or red cross on the strains that harbor at least one inactivated copy. The color of the cross is only to distinguish between two genes on (B and D).

(A) In (A) we can observe that the *MAL11* gene is present in some strains of the African beer group and absent from our sample of European dominant strains.

(B) Reiterates previously known results on the inactivation of the *PAD1*/*FDC1* genes and shows this domestication event also occurs in the African beer strains.

(C) Shows the loss of *GAL2* copies in almost half of our sample of African beer strains.

(D) Finally, in (D) we observe the loss of *ADH2* in Belgian and European dominant strains, and the loss of *SFA1* in most British ales. See also Tables S10 and S11.

strains except for ARE, which displays both known frameshift-inducing indels heterozygously, and ASD which is intact. Similarly, all of the Belgian ale strains have at least a heterozygous indel except for BBG which is intact. Finally, we found that none of the British ale strains in our dataset displayed any frameshift-inducing indels.

The *PAD1* and *FDC1* genes code for proteins that participate in the formation of 4VG, a compound that yields a potent off-flavor in beer,[4] and their function therefore leads to an inferior product by human standards. The inactivation of these genes has previously been identified as evidence of human domestication of beer yeasts due to their effects on beer flavor.[4] In our

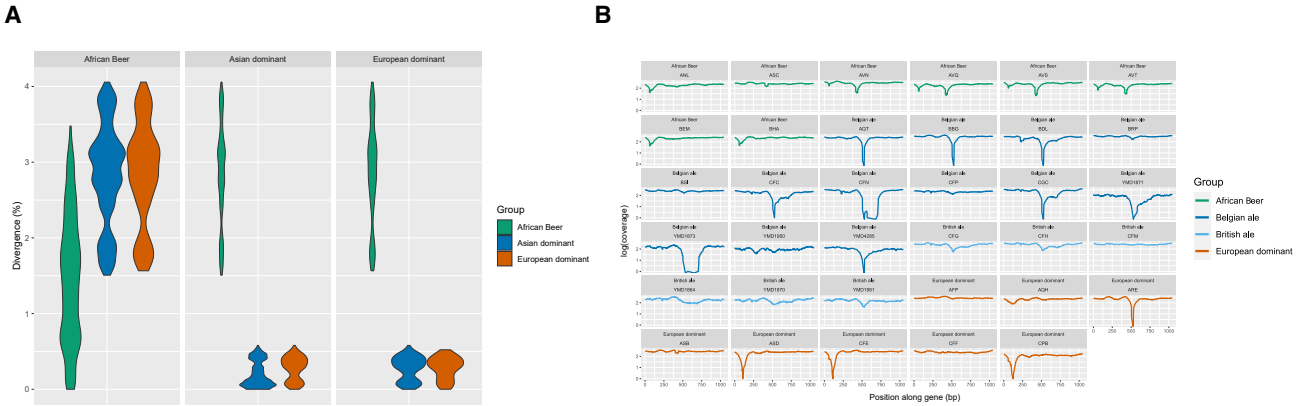

Figure 7. Divergence levels of *GAL2* and coverage levels of *ADH2*

(A) This graph represents the divergence levels of *GAL2* obtained via pairwise comparisons of all fully phased haplotypes for this gene in our dataset. We note that the divergence level (y axis, given as a percentage) is high when comparing African beer haplotypes of *GAL2* to one another, and even higher when compared with European dominant or Asian dominant copies of *GAL2*. The European dominant and Asian dominant haplotypes of *GAL2* are not very divergent from each other, with a maximum of 0.5% divergence. However, the African beer haplotypes of *GAL2* are at minimum 1.5% divergent from European dominant and Asian dominant strains, and at most 4% divergent from them.

(B) We extracted the coverage levels of Illumina reads in the region corresponding to the gene *ADH2* (chromosome XIII: 873291–874337). This graph shows the coverage level of *ADH2* for each strain, using a log scale on the y axis to represent coverage for ease of interpretation. The x axis represents the position along the gene, starting at 0 for the first position of the CDS. The strains are colored according to the group, except for the Asian dominant group that is subdivided into British and Belgian ales. We observe a shared homozygous deletion in the middle of *ADH2* among 9 of the 13 Belgian ale strains and 1 of the European dominant strains. We also note the presence of a shared homozygous deletion in the beginning of *ADH2* in 3 of the 8 European dominant strains.

dataset, we make corroborating observations. In fact, the *PAD1* gene presents no frameshift-inducing indels and appears intact in all European dominant and African beer strains; however, it is fully inactivated by nonsense mutations in all haplotypes of British ale strains and in over half of the Belgian ale strains. In addition, the *FDC1* gene appears to be intact in European dominant strains. In Asian dominant strains, it is inactivated through the frameshift-inducing indel 495T➡TA. This indel is present homozygously in all haplotypes of British ale strains and in the majority of Belgian ale strains. Only three Belgian ale strains appear to have intact copies of *FDC1*. African beers also present a frameshift-inducing indel, which inactivates their copy of *FDC1*; however, it's a different indel than the one observed in Asian dominant strains. In African beers we have the indel 35AC➡A, which is present at least heterozygously in half of the strains in our dataset. The other African beer strains have an intact copy of *FDC1*.

Overall, we found that the African beer strains bear previously reported markers of domestication through the presence of *MAL11* and the independent inactivating indel observed in *FDC1* for some of the strains. By contrast, the industrialized Beer 2 strains do not present the industrially favorable genotypes, consistent with the previously reported observation that they exhibit fewer signs of domestication than Beer 1 strains.[4] These results further support the notion that traditional beer-brewing methods, such as those used in African beer brewing, are drivers of domestication.

Phasing diverse populations reveals distinct evolutionary trajectories

To leverage the diversity of our dataset and explore some of the highly divergent genes described above, we calculated, for each full gene haplotype, the mean distance to all of the haplotypes of each group. This gave us insight into the conservation and divergence of genes among all strains. We describe our observations for *GAL2*, *ADH2*, and an associated gene, *SFA1* (Figures 6C and 6D; Table S10).

The haplotypes of the *GAL2* gene are highly diverse in African beer strains

The fermentation environment of African beer strains is known to typically harbor a variety of LAB strains that proliferate during the initial spontaneous fermentation. French dairy strains, which also share their environment with LAB strains, compete with them by consuming all of the available sugars faster. However, the typical *GAL* pathway in *S. cerevisiae* is repressed by the presence of glucose, a more efficient sugar that the yeast will metabolize first. Once the environment is depleted of glucose, growth stalls as the yeast cells switch to galactose utilization. Adaptations to the *GAL* pathway that address this competitive disadvantage have been shown in French dairy strains. The high affinity glucose/galactose transporter *GAL2* has been shown not to undergo glucose repression and allows for the simultaneous assimilation of both glucose and galactose. These modifications permit them to avoid the shift that occurs when switching from glucose to galactose, thereby improving their competitive fitness.[28,29]

In our dataset, copies of *GAL2* are very similar to one another and present no frameshifts in European dominant and Asian dominant strains; however, we observe a spectrum of copies of *GAL2* in African strains ranging from 1.56% genetic divergence to the closest non-African versions of *GAL2* to 4.05% genetic divergence with the most distant versions (Figure 7A; Table S11). Four African beer strains harbor haplotypes at divergence levels with non-African strains between 1.5% and 4.0%. The remaining 4 African beer strains all have a narrower range

of haplotype divergence, around 2.5% for 1 and over 3.0% for the other 3. The 4 strains with the higher range of diversity among their own haplotypes are the same ones harboring frameshift-inducing indels, homozygously for 1 strain and heterozygously for the other 3 (Figure 6C; Table S10).

This diversity in copies and the inactivation of certain alleles of *GAL2* exclusively found in African beer strains may represent adaptations to sharing their environment with LAB strains, which would parallel but remain independent with the adaptations observed in French dairy strains.

The *ADH2* and *SFA1* genes present further evidence of domestication in Asian dominant and European dominant strains

At high concentrations, fusel alcohols are considered a potent off-flavor in beer. *S. cerevisiae* has six genes involved in the final step of the Ehrlich pathway for fusel alcohol formation,[22] i.e., the *ADH* alcohol dehydrogenase family *ADH1* to *ADH5*, and *SFA1*, an alcohol dehydrogenase and glutathione-dependent formaldehyde dehydrogenase. *ADH1*, *ADH3*, *ADH4*, and *ADH5* convert acetaldehyde to ethanol; however, *ADH2* performs the inverse reaction and oxidizes ethanol into acetaldehyde.

In our dataset, all African beer strains and British ale strains have at least one intact copy of *ADH2*. Half of the European dominant strains and the majority of Belgian ale strains suffer from homozygous deletions. Also, 9 Belgian ale strains and 1 European dominant strain all present a homozygous deletion of approximately 26 bp in the middle of *ADH2*, which is much larger in strains CFN and YMD1873 (Figure 7B). A different deletion of about 25 bp at the beginning of the gene is observed in 3 European dominant strains. The *ADH2* gene has previously been a target for inactivation for industrial beer-brewing purposes owing to its role in forming the off-flavor acetaldehyde and reducing alcohol content.[30] These observed deletions and high genetic diversity may reflect evidence of domestication.

This potential domestication event does not affect British ale strains. However, we make complementary observations in our dataset, as all strains appear to have intact copies of the *SFA1* gene except for 4 of 6 British ale strains. These strains present either a deletion leading to a frameshift and a subsequent premature stop or have at least one haplotype with a nonsense mutation. This inactivation of several alleles of *SFA1* only in British ale strains may be evidence of a domestication event that runs parallels and complements the disruption of *ADH2* in Belgian ale and European dominant strains, likely in connection to their role in fusel alcohol production.[23]

DISCUSSION

We phased 35 strains of *S. cerevisiae* that are either used in beer brewing or are in clades that have a large proportion of beer-brewing strains, according to the 1,011 *Saccharomyces cerevisiae* genomes survey.[6] A little under a quarter of strains are diploids (n = 8), with all others being polyploids that range from 3n to 5n. We phased all 35 strains using nPhase, obtaining contiguous results with an average of 1.2 haplotigs per chromosome to phase 90% of reads.

We used a windowed approach on these phased haplotypes to estimate their pairwise divergence levels based on phased genetic content, revealing that our dataset seems to comprise three large groups of strains that are more similar to one another than to other strains. Using a different windowed approach, we then estimated the allelic origins of these beer strains by assigning their haplotypes to different clades.[6] We found that they all contain an important proportion of European wine alleles, and that we can categorize them into the same three distinct groups, this time based on their allelic origin profiles: Asian dominant strains, European dominant strains, and African beer strains. The Asian dominant strains correspond to the previously defined Beer 1 group[4] whose origin as a polyploid admixture of Asian and European wine alleles has previously been described.[8] The European dominant group corresponds to the previously defined Beer 2 group,[4] again an admixture of Asian and European wine that however differs from the Asian dominant group by its lower proportion of Asian fermentation alleles. Finally, we characterized the allele content of the African Beers as having a strong European wine signal and a higher French dairy signal than the other groups.

African beer-brewing methods are significantly less industrialized and typically follow traditional means,[15] which for *S. cerevisiae* translates to a mode of life that must remain adapted to the wild and to environments with other microorganisms, notably the LAB that proliferate during the initial spontaneous fermentation step that typically precedes *S. cerevisiae*'s alcoholic fermentation.[15,16] French dairy strains of *S. cerevisiae* that share an environment with LAB have been shown to adapt their *GAL* pathway to disable its glucose repression and more rapidly drain the environment of sugar to outcompete other organisms.[28,29] We found possible evidence of a similar adaptation to sharing an environment with LAB in the extensive changes to *GAL2* we observed in African beer strains, and the presence of multiple different haplotypes of *GAL2* within each strain. We propose that these modifications may disable or attenuate glucose repression, or confer some other advantage to *S. cerevisiae* strains sharing an environment with LAB.

We also found that despite less obvious domestication pressures, some African beer strains show known signs of domestication. It has been shown that the *FDC1* gene is inactivated in a large number of industrialized beer strains, and not in wild strains, due to its role in forming the undesirable off-flavor compound 4VG.[4] In half of the African beer strains in our dataset, we observed that *FDC1* was inactivated by a frameshift mutation different from the one that affects Asian dominant strains, suggesting an independent domestication event for this gene.

Finally, we propose that two complementary domestication events occurred in European dominant strains and British and Belgian ale strains. The alcohol dehydrogenases *SFA1* and *ADH2* can both contribute to the last step of the formation of fusel alcohol,[22] which in high concentrations are undesirable[23] (in fact, fusel is a German word for bad or cheap liquor). A deletion in the middle of *ADH2* is widely present in Belgian ale strains and at the beginning of *ADH2* in half of the European dominant strains in our dataset, whereas premature stops in *SFA1* are observed in British ale strains, suggesting independent and complementary domestication events that should have a similar effect of lowering the overall concentration of fusel alcohol in the final brew.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Saccharomyces cerevisiae
- METHOD DETAILS
  - Strain selection, DNA extraction & sequencing
  - Phasing and cleaning using nPhase
  - Aneuploidy identification
  - Pairwise haplotype divergence
  - Dendrogram creation
  - Allele content origin attribution
  - Divergence between gene haplotypes
  - Gene Ontology Term Finder calculations on SGD
  - Identifying frameshifts & premature stops
  - Coverage plots for the *MAL11*, and *ADH2* genes
  - Data visualization tools
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2022.01.068.

ACKNOWLEDGMENTS

We thank Justin Fay and Viktor Boer for helpful discussions. We also thank the members of the PhenoVar ANR consortium. This work was supported by a National Institutes of Health (NIH) grant R01 (GM101091-01), an Agence Nationale de la Recherche grant (ANR-16-CE12-0019), and a European Research Council (ERC) Consolidator grant (772505).

AUTHOR CONTRIBUTIONS

O.A.S. and J.S. designed the study. A.T. and C.L. conducted the experiments. O.A.S. and A.F. performed the analysis. M.J.D. and J.S. contributed with resources and reagents. O.A.S and J.S. wrote the paper. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 14, 2021
Revised: December 2, 2021
Accepted: January 21, 2022
Published: March 28, 2022

REFERENCES

1. Legras, J.-L., Merdinoglu, D., Cornuet, J.-M., and Karst, F. (2007). Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. Mol. Ecol. *16*, 2091–2102.
2. 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.
3. Ludlow, C.L., Cromie, G.A., Garmendia-Torres, C., Sirr, A., Hays, M., Field, C., Jeffery, E.W., Fay, J.C., and Dudley, A.M. (2016). Independent origins of yeast associated with coffee and cacao fermentation. Curr. Biol. *26*, 965–971.
4. Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell *166*, 1397–1410.e16.
5. Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., Hutzler, M., Gonçalves, P., and Sampaio, J.P. (2016). Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. Curr. Biol. *26*, 2750–2761.
6. Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature *556*, 339–344.
7. Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., Sanchez, I., Couloux, A., Guy, J., Franco-Duarte, R., et al. (2018). Adaptation of *S. cerevisiae* to fermented food environments reveals remarkable genome plasticity and the footprints of domestication. Mol. Biol. Evol. *35*, 1712–1727.
8. Fay, J.C., Liu, P., Ong, G.T., Dunham, M.J., Cromie, G.A., Jeffery, E.W., Ludlow, C.L., and Dudley, A.M. (2019). A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. PLoS Biol. *17*, e3000147.
9. Barnett, J.A., and Lichtenthaler, F.W. (2001). A history of research on yeasts 3: Emil Fischer, Eduard Buchner and their contemporaries, 1880–1900. Yeast *18*, 363–388.
10. Steensels, J., Gallone, B., Voordeckers, K., and Verstrepen, K.J. (2019). Domestication of industrial microbes. Curr. Biol. *29*, R381–R393.
11. Udom, N., Chansongkrow, P., Charoensawan, V., and Auesukaree, C. (2019). Coordination of the cell wall integrity and high-osmolarity glycerol pathways in response to ethanol stress in *Saccharomyces cerevisiae*. Appl. Environ. Microbiol. *85*, e00551–e00519.
12. Gagnon-Arsenault, I., Tremblay, J., and Bourbonnais, Y. (2006). Fungal yapsins and cell wall: a unique family of aspartic peptidases for a distinctive cellular function. FEMS Yeast Res. *6*, 966–978.
13. Szoradi, T., Schaeff, K., Garcia-Rivera, E.M., Itzhak, D.N., Schmidt, R.M., Bircham, P.W., Leiss, K., Diaz-Miyar, J., Chen, V.K., Muzzey, D., et al. (2018). SHRED is a regulatory cascade that reprograms Ubr1 substrate specificity for enhanced protein quality control during stress. Mol. Cell *70*, 1025–1037.e5.
14. Charoenbhakdi, S., Dokpikul, T., Burphan, T., Techo, T., and Auesukaree, C. (2016). Vacuolar H+-ATPase protects *Saccharomyces cerevisiae* Cells against ethanol-induced oxidative and cell wall stresses. Appl. Environ. Microbiol. *82*, 3121–3130.
15. Holzapfel, W.H. (2002). Appropriate starter culture technologies for small-scale fermentation in developing countries. Int. J. Food Microbiol. *75*, 197–212.
16. Johansen, P.G., Owusu-Kwarteng, J., Parkouda, C., Padonou, S.W., and Jespersen, L. (2019). Occurrence and importance of yeasts in indigenous fermented food and beverages produced in sub-Saharan Africa. Front. Microbiol. *10*, 1789.
17. Adebo, O.A. (2020). African sorghum-based fermented foods: past, current and future prospects. Nutrients *12*, 1111.
18. Bokulich, N.A., and Bamforth, C.W. (2013). The microbiology of malting and brewing. Microbiol. Mol. Biol. Rev. *77*, 157–172.
19. Lengeler, K.B., Stovicek, V., Fennessy, R.T., Katz, M., and Förster, J. (2020). Never change a brewing yeast? Why not, there are plenty to choose from. Front. Genet. *11*, 582789.
20. Whittington, H.D., Dagher, S.F., and Bruno-Bárcena, J.M. (2019). In Production and conservation of starter cultures: from "backslopping" to controlled fermentations. How Fermented Foods Feed a Healthy Gut Microbiota: A Nutrition Continuum, M.A. Azcarate-Peril, R.R. Arnold, and

J.M. Bruno-Bárcena, eds. (Springer International Publishing), pp. 125–138, 10.1007/978-3-030-28737-5_5.

21. Teste, M.-A., François, J.M., and Parrou, J.-L. (2010). Characterization of a new multigene family encoding isomaltases in the yeast *Saccharomyces cerevisiae*, the IMA family. J. Biol. Chem. *285*, 26815–26824.
22. Dickinson, J.R., Salgado, L.E.J., and Hewlins, M.J.E. (2003). The catabolism of amino acids to long chain and complex alcohols in *Saccharomyces cerevisiae*. J. Biol. Chem. *278*, 8028–8034.
23. Hazelwood, L.A., Daran, J.-M., van Maris, A.J.A., Pronk, J.T., and Dickinson, J.R. (2008). The Ehrlich pathway for fusel alcohol production: a century of research on *Saccharomyces cerevisiae* metabolism. Appl. Environ. Microbiol. *74*, 2259–2266.
24. Saada, O.A., Tsouris, A., Eberlein, C., Friedrich, A., and Schacherer, J. (2021). nPhase: an accurate and contiguous phasing method for polyploids. Genome Biol. *22*, 1–27.
25. Jacobus, A.P., Stephens, T.G., Youssef, P., González-Pech, R., Ciccotosto-Camp, M.M., Dougan, K.E., Chen, Y., Basso, L.C., Frazzon, J., Chan, C.X., et al. (2021). Comparative genomics supports that Brazilian bioethanol *Saccharomyces cerevisiae* comprise a unified group of domesticated strains related to cachaça spirit yeasts. Front. Microbiol. *12*, 644089.
26. Krysan, D.J., Ting, E.L., Abeijon, C., Kroos, L., and Fuller, R.S. (2005). Yapsins are a family of aspartyl proteases required for cell wall integrity in *Saccharomyces cerevisiae*. Eukaryot. Cell *4*, 1364–1374.
27. Verbelen, P.J., Saerens, S.M.G., Van Mulders, S.E., Delvaux, F., and Delvaux, F.R. (2009). The role of oxygen in yeast metabolism during high cell density brewery fermentations. Appl. Microbiol. Biotechnol. *82*, 1143–1156.
28. Duan, S.F., Shi, J.Y., Yin, Q., Zhang, R.P., Han, P.J., Wang, Q.M., and Bai, F.Y. (2019). Reverse evolution of a classic gene network in yeast offers a competitive advantage. Curr. Biol. *29*, 1126–1136.e5.
29. Boocock, J., Sadhu, M.J., Durvasula, A., Bloom, J.S., and Kruglyak, L. (2021). Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. Science *371*, 415–419.
30. Wang, Z.-Y., Wang, J.-J., Liu, X.-F., He, X.-P., and Zhang, B.-R. (2009). Recombinant industrial brewing yeast strains with ADH2 interruption using self-cloning GSH1+CUP1 cassette. FEMS Yeast Res. *9*, 574–581.
31. Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., Caradec, C., Davidas, S., Cruaud, C., Liti, G., et al. (2017). *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience *6*, 1–13.
32. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
33. Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics *20*, 289–290.
34. 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.
35. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008.
36. Engel, S.R., Dietrich, F.S., Fisk, D.G., Binkley, G., Balakrishnan, R., Costanzo, M.C., Dwight, S.S., Hitz, B.C., Karra, K., Nash, R.S., et al. (2014). The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3 (Bethesda) *4*, 389–398.
37. Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics *20*, 3710–3715.
38. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997.
39. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165.
40. Wong, B. (2011). Points of view: color blindness. Nat. Methods *8*, 441.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
Ligation Sequencing Kit	Oxford Nanopore Technologies	Catalog number: SQK-LSK109
Native Barcoding Expansion 1-12	Oxford Nanopore Technologies	Catalog number: EXP-NBD104
QIAGEN Genomic-tip 100/G kit	QIAGEN	Catalog number: 10243
Broad-range DNA quantification kit	Qubit	Catalog number: Q32850
Deposited Data		
long-read sequencing of 35 strains of *Saccharomyces cerevisiae*	European Nucleotide Archive; listed in Table S2	BioProject accession: PRJEB46384
short-read sequencing of 35 strains of *Saccharomyces cerevisiae*	European Nucleotide Archive; listed in Table S2	BioProject accession: PRJEB13017
Experimental Models: Organisms/Strains		
35 polyploid beer-related strains of *Saccharomyces cerevisiae*	1,011 strains collection; listed in Table S1	N/A
Software and Algorithms		
GATK v4.1.5.0	https://github.com/broadinstitute/gatk	N/A
R package ape v5.5	https://cran.r-project.org/web/packages/ape	N/A
R package SNPRelate v3.14	https://www.bioconductor.org/packages/release/bioc/html/SNPRelate.html	N/A
filtlong v0.2.0	https://github.com/rrwick/Filtlong	N/A
bcftools v1.11	https://github.com/samtools/bcftools	N/A
Gene Ontology Term Finder	https://www.yeastgenome.org/goTermFinder	N/A
bwa-mem v0.7.17	https://github.com/lh3/bwa	N/A
bamCoverage v3.5.0	https://github.com/deeptools/deepTools	N/A
pheatmap v1.0.12	https://cran.r-project.org/web/packages/pheatmap/index.html	N/A
FigTree v1.4.4	http://tree.bio.ed.ac.uk/software/figtree/	N/A
ggplot2 v3.3.3	https://ggplot2.tidyverse.org	N/A
R programming language v4.0.2	https://www.r-project.org/about.html	N/A
nPhase v1.1.3	https://github.com/OmarOakheart/nPhase	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Joseph Schacherer (schacherer@unistra.fr).

Materials availability

This study did not generate new unique materials.

Data and code availability

Oxford Nanopore and Illumina sequencing data have been deposited to the EBI (European Bioinformatics Institute) and are publicly available as of the date of publication. The accession number of the study is listed in the key resources table. Illumina short read data for the *Saccharomyces cerevisiae* strains is taken from the 1,011 yeast genomes project and their SRA accession numbers are listed in the key resources table.

This paper does not report original code.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Saccharomyces cerevisiae

For this study, we focused on a population of 35 of *Saccharomyces cerevisiae* isolates from diverse clades based either on their known use in beer fermentation or on their high genetic similarity to beer-brewing strains (Table S1). These strains were subset from the 1,011 yeast strain collection[6] and were isolated to single colonies on solid YPD-agar plates supplemented with ampicilin. A single colony from each strain was then grown in 30ml of YPD media until saturation (48 hours) at 30°C with agitation prior to DNA extraction.

METHOD DETAILS

Strain selection, DNA extraction & sequencing

The DNA of 35 strains was extracted from 30 mL cultures (single colony, 48h growth at 30°C) using the QIAGEN Genomic-tip 100/G kit with the recommended manufacturer's genomic DNA buffer set. The manufacturer's protocol was followed as recommended and final DNA was eluted in 100-200 μl water. DNA was quantified with the broad-range DNA quantification kit from Qubit. Genomic DNA was migrated on a 1.5% agarose gel to check for degradation.

For the long-read sequencing we used the Oxford Nanopore Technology (Oxford, UK). Libraries for sequencing using the MinION and were prepared as described in Istace et al.[31] using the Ligation Sequencing Kit SQK-LSK109. We barcoded strains with the Native Barcoding Expansion 1-12 (EXP-NBD104) to multiplex up to 12 samples per sequencing reaction. Sequencing statistics are given in Table S2.

Phasing and cleaning using nPhase

We used filtlong v0.2.0 (https://github.com/rrwick/Filtlong) to subset our nanopore long reads to 80X (estimated as 12 500 000 * 80 bases), then used the nPhase pipeline[24] v1.1.3 with default parameters to phase each strain using its long and short reads and the R64 reference sequence of *S. cerevisiae*. Once we obtained raw results using the nPhase pipeline command, we ran the nPhase cleaning command using default parameters to improve contiguity and eliminate short, uninformative haplotigs. Phasing results are given in Table S3.

Aneuploidy identification

Aneuploidy information for most strains was obtained from Peter *et al.*[6] For the 7 remaining strains aneuploidy was determined based on allele frequency plots (Figure S2). Aneuploidy information is given in Table S4.

Pairwise haplotype divergence

nPhase outputs a file with the suffix ".variants.tsv" which indicates, for each predicted haplotig, the SNPs that were phased. We use this file along with the reference sequence of *S. cerevisiae* to infer the full sequences of our haplotypes and split them into 10kb windows. Then, for each pair of strains, we compared every full 10 kb haplotype window to all of the haplotypes fully covering the same window in the opposite strain and only kept the lowest divergence value.

This method extends to the calculation of internal divergence levels, with the difference that instead of comparing the haplotypes of one strain to the haplotypes of another, we compared the haplotypes of one strain to each other. We again keep the lowest value, but we do not allow a 10 kb haplotype block to compare to itself.

Being ploidy agnostic, nPhase tends to group homozygous regions together so there may be an over-estimation of divergence, however nPhase also doesn't take indels into account so there may be an under-estimation of divergence. It's unclear which bias has the stronger effect, or the extent of the effect of these limitations.

Dendrogram creation

A genotyping matrix was constructed with the GenotypeGVCFs function of GATK[32] that was run on individual gvcf files generated by GATK's HaplotypeCaller method. This matrix was used to build a neighbor-joining tree with the R packages ape[33] and SNPrelate.[34] To that end, the gvcf matrix was converted into a gds file and individual dissimilarities were estimated for each pair of isolates with the snpgdsDiss function. The bionj algorithm was then run on the obtained distance matrix.

Pairwise differences between the studied strains was estimated from the non-shared SNPs positions obtained with bcftool[35] isec with -n -1 -c all options run on individual gvcf files.

Allele content origin attribution

In order to investigate the origins of these beer strains we used a windowed approach to split the haplotypes predicted by nPhase into 20 kb windows and compared them to 6 of the clades described in Peter *et al.*[6]: European wine (we merged all of the European wine subclades), the clinical French Guiana human, African palm wine, North American Oak, Asian fermentation (we merged the Sake and Asian fermentation clades) and French Dairy.

For each clade, we consider that a position is a marker of this clade if it has a Minor Allele Frequency (MAF) $\geq$ 25% within the clade and is not present in more than one of the other 5 clades at a MAF $\geq$ 25%. Then for each 20kb window of each haplotype we attributed the clade with the highest number of markers.

Divergence between gene haplotypes

To calculate the pairwise divergence between genes we used the latest annotation of *S. cerevisiae* available on SGD (Release 64-2-1 of the S288C reference genome[36]) and extracted the positions of genes. We then extracted each gene's sequence in the reference genome and for each strain we used the strainName.variant.tsv file generated by nPhase to extract all predicted haplotypes, only keeping the variants that fall within each gene's sequence and inferring them into the reference sequence. We only kept gene haplotypes which had full predictions, we did not keep any incompletely inferred genes. We then proceeded to a pairwise comparison of every gene haplotype in our dataset, calculating the divergence as the number of mismatching positions divided by the length of the gene.

Gene Ontology Term Finder calculations on SGD

Based on the divergence calculations described above, we then identified all genes for which at least one pairwise comparison led to a divergence level $\geq$ 4%. We only keep genes whose ORF classification is listed as "Verified" in the annotation, not "Uncharacterized" or "Dubious". Then we input that list with default parameters into the Gene Ontology Term Finder[37] available on the *Saccharomyces* Genome Database (SGD) website at the following url: https://www.yeastgenome.org/goTermFinder

Identifying frameshifts & premature stops

We identified indels that cause frameshift mutations by manual inspection of the Illumina VCF files generated by the nPhase pipeline using bwa-mem[38] for mapping and GATK 4.0[32] for variant calling. Premature stops were identified by identifying stop codons in the previously generated inferred gene haplotypes.

Coverage plots for the *MAL11*, and *ADH2* genes

We generated the data for our gene coverage plots of *MAL11* and *ADH2* using bamCoverage[39] v3.5.0 with a window size of 1.

Data visualization tools

The heatmap with clustering (Figure 2) was generated using pheatmap v1.0.12 (https://cran.r-project.org/web/packages/pheatmap/index.html), the dendrogram is viewed in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and other figures were generated using ggplot2 v3.3.3 (https://ggplot2.tidyverse.org) on the R programming language v4.0.2 (https://www.r-project.org/about.html). We used a color palette intended for interpretability by people with colorblindness[40].

QUANTIFICATION AND STATISTICAL ANALYSIS

Data were visualized using software cited in the data visualization section in the method details. The only statistical tests were those calculated by the Gene Ontology Term Finder[37] on the Saccharomyces Genome Database (SGD) using default parameters (p-value is significant if <0.01), n is given in Table S9 (total_num_annotations).