Ancient and recent origins of shared polymorphisms in yeast

Nicolò Tellini[1], Matteo De Chiara[1], Simone Mozzachiodi[1], Lorenzo Tattini[1], Chiara Vischioni[1], Elena S Naumova[2], Jonas Warringer[3], Anders Bergström[4,5] and Gianni Liti[1,#]

[1]CNRS, INSERM, IRCAN, Côte d'Azur University, Nice, France; [2]Kurchatov Complex for Genetic Research (GosNIIgenetika), National Research Center "Kurchatov Institute", Moscow, Russia; [3]Department of Chemistry and Molecular Biology, University of Gothenburg, 40530 Gothenburg, Sweden; [4]Ancient Genomics Laboratory, The Francis Crick Institute, London, UK; [5]School of Biological Sciences, University of East Anglia, Norwich, UK.

to whom correspondence should be addressed: G.L. gianni.liti@cnrs.fr

Summary

Shared genetic polymorphisms between populations and species can be ascribed to ancestral variation or to more recent gene flow. Here, we mapped shared polymorphisms in *Saccharomyces cerevisiae* and its sister species *Saccharomyces paradoxus*, which diverged 4-6 million years ago. We used a dense map of diagnostic markers (mean distance 15.6 bp) in 1,673 sequenced *S. cerevisiae* isolates to catalogue 3,852 introgressed blocks (≥ 5 consecutive markers) from *S. paradoxus*, with the majority being recent and clade-specific. The highly diverged wild Chinese *S. cerevisiae* lineages were depleted of introgressed blocks, but retained an excess of individual ancestral polymorphisms derived from incomplete lineage sorting, perhaps due to less dramatic population bottlenecks. In the non-Chinese *S. cerevisiae* lineages, we inferred major hybridisation events and detected cases of overlapping introgressed blocks across distinct clades due to either shared histories or convergent evolution. We experimentally engineered, in otherwise isogenic backgrounds, the introgressed *PAD1-FDC1* gene-pair that independently arose in two *S. cerevisiae* clades and revealed that it potentiates resistance against diverse antifungal drugs and ferulic acid. Overall, our study retraces histories of divergence and secondary contacts across *S. cerevisiae* and *S. paradoxus* populations and unveil a functional outcome.

Main

Genetic variation segregating in natural populations enables inference of species histories and past demographics. Extant populations diverge from ancestral progenitors by accumulating *de novo* variation that is exposed to both selection and neutral evolutionary forces. One source of genetic variation among extant populations is represented by ancestral polymorphisms that arose before populations diverged. The random assortment of ancestral polymorphisms can lead to incomplete allele sorting (incomplete lineage sorting, ILS), resulting in individuals from a population sharing alleles with a non-sister population[1–3]. An alternative source of shared alleles between non-sister populations is introgression, initiated by hybridisation[4], which is pervasive across the eukaryotic tree of life with potential adaptive or deleterious outcomes[5,6]. Although both ILS and introgression result in shared polymorphisms between non-sister populations, they emerge from processes unfolding at fundamentally different timescales, with introgression being more recent.

Genomic surveys of modern human populations[7], archaic humans[8,9] and great ape species[10] have elucidated the origin of genetic variation in our species. ILS has shaped the genetic relationships between our own genomes and those of our closest relatives, the bonobo and the common chimpanzee[10]. Introgression from archaic Neanderthal and Denisovan humans has introduced divergent haplotypes persisting in present-day human populations[11]. Some of this introgressed variation has contributed to local adaptation, including the adaptation to high altitude in Tibetans[12].

The model organism *Saccharomyces cerevisiae* and its closest relative *Saccharomyces paradoxus* diverged around 4.0-5.8 million years (My) ago[13]. Even though these species are reproductively isolated, introgressed DNA from *S. paradoxus* has been detected in few *S. cerevisiae* clades[14–16]. These events support at least two distinct pulses of introgression, between the American and between the European populations of the two species, respectively. However, these analyses have mostly relied on fragmented *de novo* assemblies derived from short reads or on mapping to a single reference, approaches which have not allowed the compilation of a high-resolution catalogue of introgressed DNA and inference of their origins. Furthermore, experimental demonstration of the functional implications of introgressed material in budding yeast are still lacking[17–19]. The occurrence of ILS in the *Saccharomyces* genus has been recently suggested[20] but not formally demonstrated. Given the ancient *S. cerevisiae* – *S. paradoxus* split (estimated time of divergence 3.27×10^8 generations), it seems unlikely that ILS polymorphisms would persist at such timescale given the short generation time and limited effective population size[17].

Recent studies described *S. cerevisiae* and *S. paradoxus* isolates collected worldwide and generated both short-read datasets[14,15,18,19,21–27] and high-quality whole-genome assemblies[28], paving the way for a deep investigation of shared alleles across these two species. Here, we exploit the thousands of *S. cerevisiae* genomes and the dense map of diagnostic markers between the two species to detect *S. paradoxus* alleles at high resolution. We identify and describe the major hybridisation events and report that most introgressed blocks derive from recent clade-specific hybridisations, while a handful of ancient introgressions are shared across distinct clades. We detect rare instances of recurrent introgression blocks and experimentally demonstrate the functional effect of *S. paradoxus PAD1-FDC1* alleles. We also formally test and quantify abundant ancestral ILS polymorphisms

in the *S. cerevisiae* Chinese lineages, providing novel insights into the evolutionary processes that shaped their genetic variation.

Results

Patterns of shared polymorphic markers across the *S. cerevisiae* species

We developed a genotyping framework based on a dense map of diagnostic markers to explore the landscape of *S. paradoxus* alleles segregating in the *S. cerevisiae* populations (Methods and Supplementary Fig. 1). We constructed a *S. cerevisiae consensus* (*S.c.c.*) genome by taking the most common allele among a set of strains representing the major phylogenetic clades within the species (Supplementary Table 1). We then ran pairwise whole-genome alignments of the *S.c.c.* genome against the five reference genome assemblies of the major *S. paradoxus* populations, to identify a set of 755,500 biallelic species-diagnostic single-nucleotide polymorphic markers (Fig. 1a). Finally, we mapped short-read sequencing data derived from 1,673 *S. cerevisiae* strains, onto the *S.c.c.* and European *S. paradoxus* (CBS432) genomes and genotyped the diagnostic markers (Fig. 1b, Supplementary Fig. 2 and Supplementary Table 2-3). We found that the strains of the Alpechin clade, characteristic of the olive oil-based environment, contain the highest number of *S. paradoxus* markers (mean 28,439 ± 3,455, single standard deviation; Fig. 1b), consistent with deriving from a recent hybridisation[14,19,29]. The Mexican Agave and the South American Mix II clades also showed high levels of *S. paradoxus* markers (18,598 ± 5,459 and 10,010 ± 3,075 respectively; Fig. 1b), consistent with secondary contacts between *S. paradoxus* and multiple North and South America *S. cerevisiae* populations[14,18].

Surprisingly, the highly diverged Chinese lineages (CHN-IX/Taiwanese, CHN-I and CHN-II) contain a large number of *S. paradoxus* markers (29,716 ± 148; 18,253 ± 501; 16,690 ± 593 respectively; Fig. 1b). Indeed, the number of *S. paradoxus* markers in the CHN-IX clade is comparable to the Alpechin clade, despite an overall lack of evidence for introgression[14,15]. The sole exception is a ~24 kb region on chromosome XI, which appear to have been introgressed from an unknown and possibly extinct *S. paradoxus* sister lineage (Supplementary Fig. 3).

We grouped consecutive *S. paradoxus* markers into blocks to define their sizes and boundaries (Supplementary Fig. 4, Supplementary Fig. 5 and Supplementary Table 4). We then counted the number of blocks, as well as the number of isolated *S. paradoxus* markers and measured the fraction of each *S. cerevisiae* genome that was covered by these two groups of *S. paradoxus* variants (Fig. 1c). Strains in the Alpechin clade had the highest fraction of their genome (3-5.9%) included in large introgression blocks but a relatively small number of isolated markers. In contrast, the Chinese lineages CHN-IX, CHN-I and CHN-II had very few introgression blocks, but a huge number of isolated *S. paradoxus* markers. These were scattered across the genomes, with no clear spatial clustering, and covered a much lower fraction of the genome (0.56-0.62%, 0.29-0.33% and 0.23-0.28% respectively). For example, the AHL Alpechin and the AMH CHN-IX strains have comparable numbers of *S. paradoxus* markers (28,566 and 29,826 respectively), but drastically different median inter-marker distances (11 bp vs. 194 bp) (Fig 1d). In other words, *S. paradoxus* markers in AHL clustered into fewer and larger blocks (e.g. 169 blocks ≥ 5 *S.p.* markers) that covered 4.8% of the genome while AMH showed many isolated markers

and fewer large blocks (76 blocks ≥ 5 *S.p.* markers), covering only 0.1% of the genome (Fig 1e and Supplementary Fig. 5). The difference is even more striking considering that 64 out of the 76 AMH blocks mapped to the single ~24 kb introgressed block on chromosome XI (Supplementary Fig. 3). These fundamentally different patterns in the genomic distribution of *S. paradoxus* markers across *S. cerevisiae* populations suggest different mechanistic origins and evolution.

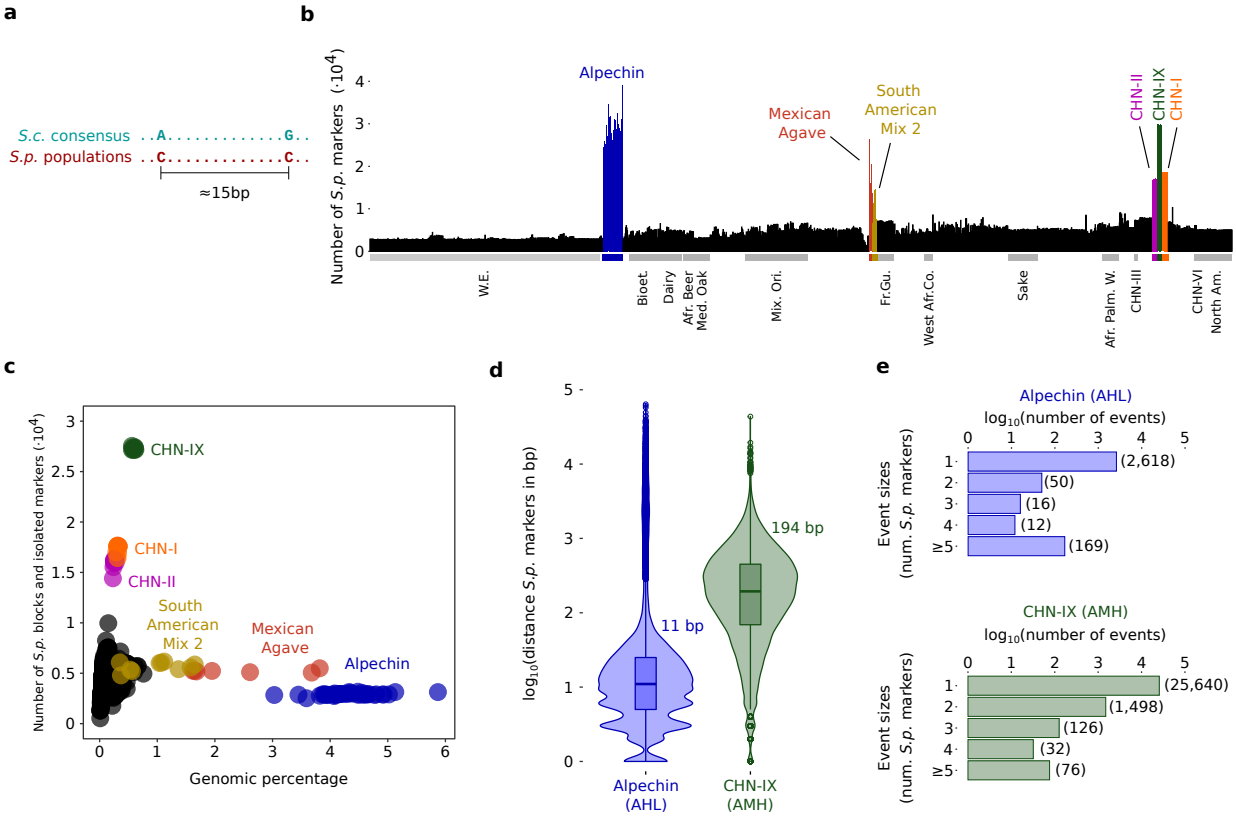


Figure 1 | The species wide landscape of *S. paradoxus* markers in *S. cerevisiae*. **a**, A diagnostic marker position is defined as a biallelic SNP between the *S. cerevisiae consensus* (*S.c.c.*) sequence and in all the *S. paradoxus* populations and occur genome-wide on average at 15 bp distance. **b**, Bar plot of number of *S. paradoxus* genotyped diagnostic markers (y-axis) across 1673 isolates of *S. cerevisiae* (x-axis) with selected clades highlighted (rectangles). Coloured clades have the highest number of *S. paradoxus* markers. **c**, The percentage of genome (x-axis) included within the diagnostic marker with *S. paradoxus* genotype. Consecutive *S. paradoxus* markers are joined into blocks and their size is defined by the first and last marker of the block. Isolated *S. paradoxus* markers lie between two markers with *S. cerevisiae* genotype and were counted as 1bp. The y-axis shows the total number of blocks and isolated markers. Relevant clades are coloured as in panel **b**. **d**, Distribution of the distance (in $\log_{10}$ base pairs) between consecutive *S. paradoxus* genotyped diagnostic markers in the AHL Alpechin and AMH CHN-IX strains. Box, interquartile range (IQR); whiskers, 1.5×IQR; thick horizontal line, median. Data points beyond the whiskers are outliers. **e**, Number of blocks of different sizes and isolated *S. paradoxus* markers detected in AHL and AMH strains partitioned by size. The size of each block is given by the number of consecutive diagnostic markers with *S. paradoxus* genotype. Size = 1 corresponds to isolated *S. paradoxus* markers. Blocks supported by at least 5 consecutive *S. paradoxus* diagnostic markers are combined into one category (≥5).

Deep coalescence in highly diverged *S. cerevisiae* lineages

We further explored the differences in patterns of *S. paradoxus* shared polymorphisms segregating among *S. cerevisiae* isolates by calculating D-statistics, which is based on counts of ABBA and BABA sites (A ancestral, B derived alleles) across a quartet of populations[8] (Methods). The D-statistic quantifies to what extent two sister

populations P1 and P2 differ in how often they share alleles with a donor population P3, at sites where P3 differs from an outgroup P4. No difference in the numbers of ABBA and BABA sites are expected under the null hypothesis of no introgression, and the statistical significance is assessed using block jackknife resampling. We used multiple whole genome alignments with the *S.c.c.* as P1, the European *S. paradoxus* CBS432 (*S.p.*) as donor species (P3), and *S. jurei* (*S.j.*) as outgroup (P4) to perform D-statistic tests across the entire *S. cerevisiae* collection (in turn P2). The quartet *S.c.c.*-AHL-*S.p.*-*S.j.* showed a negative D-value (-0.65), a net difference between ABBA and BABA sites (13,952 vs 2,964) and strong statistical support (Z-score: -16.87) formally supporting abundant *S. paradoxus* introgression in the Alpechin AHL strain (Fig. 2a). In contrast, the quartet *S.c.c.*-AMH-*S.p.*-*S.j.* showed a D-value close to 0 (0.0175), a comparable number of ABBA and BABA sites (12,695 vs 12,258) with no statistical support for *S. paradoxus* introgressions in the CHN-IX AMH strain (Z-score: 1.615). Thus, for AMH, we hypothesize that the presence of *S. paradoxus* alleles is the consequence of ILS, and not introgression (Fig. 2a, Supplementary Fig. 6).

To test the robustness of this result to the strains used in the ABBA-BABA test, we repeated the ABBA-BABA test but instead used a Wine European, CHN-IV and CHN-I *S. cerevisiae* strains in the P1 position. We found no evidence for introgression in AMH using any of these P1 variations (Supplementary Fig. 6a-b). We further investigated the ABBA and BABA positions in AMH that are shared across the ABBA-BABA tests performed with *S.c.c*, CHN-IV and CHN-I assemblies as P1 (Supplementary Table 5). We observed 18,152 shared ABBA+BABA sites which fall into intergenic (27.8%) and genic (72.2%) positions with an enrichment in 3rd codon positions (genic sites: 13,313; first: 2,052, second: 1,275; third: 9,787). This distribution of the ILS sites closely resembles the genome-wide distribution of diagnostic markers, suggesting no pervasive purifying selection acting on these ancestral polymorphisms.

Next, we inspected the presence of CHN-IX *S. paradoxus* alleles in the CHN-I and CHN-II *S. cerevisiae* populations. We observed that approximately 71.9% and 76.1% of the *S. paradoxus* alleles detected in the CHN-I (strain FJ7) and CHN-II (strain BAG) strains are shared with CHN-IX *S. paradoxus* alleles detected in AMH strain (Supplementary Figure 6c), consistent with their common origin as ancestral polymorphisms.

Finally, we extended the ABBA-BABA test varying the whole *S. cerevisiae* collection as P2 (Fig. 2b-c) and observed no evidence for *S. paradoxus* introgression in the CHN-II clade (e.g. BAG strain D: -0.0143, ABBA: 7782, BABA: 7562, and Z-score: -1.046), while CHN-I strains showed weakly but significantly positive D values (e.g. FJ7 strain D: 0.0528, ABBA: 8609, BABA: 9567, and Z-score: 4.834). However, the significance of the test for both CHN-II and CHN-I strains vary with the genome used as P1. Overall, we provided robust evidence that part of the genetic variants observed in the highly diverged CHN-IX *S. cerevisiae* population is driven by persisting ancestral alleles.

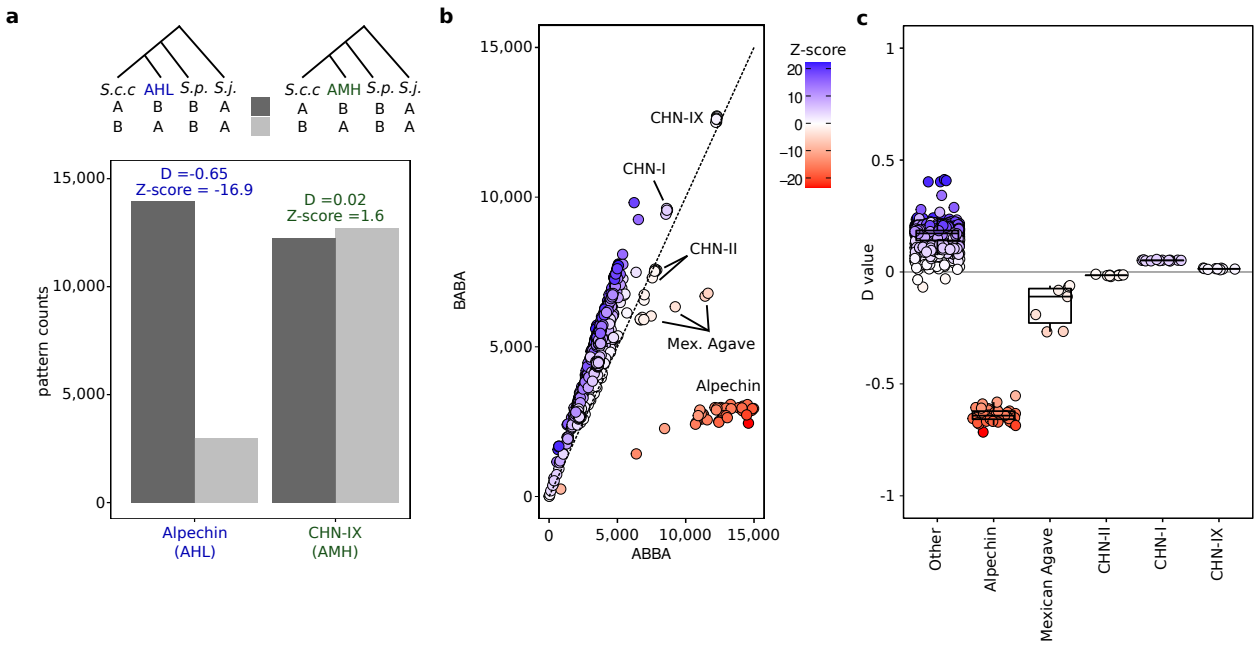


Figure 2 | Incomplete lineage sorting in Chinese lineages. a, Patterson's D statistics (ABBA-BABA test) with AHL and AMH in (P2) with the pattern *S.c.c*. (P1), CBS432 *S.p.* (P3) and *S. jurei* (P4). Bar plot indicates the number of ABBA and BABA patterns observed. "A" is the ancestral allele as imposed by P4, "B" is the derived allele. Only biallelic positions are included. **b** Scatter plot of ABBA and BABA patterns across the *S. cerevisiae* collection rotating in P2 with the pattern *S.c.c*. (P1), CBS432 *S.p.* (P3) and *S. jurei* (P4). The dotted line represents the diagonal. Relevant *S. cerevisiae* clades are labelled. **c**, Distributions of the Patterson's D value across the *S. cerevisiae* isolates. The dots colour gradient reflects the Z-score values. Alpechin and Mexican Agave strains show the strongest introgression signal, consistent with *S. paradoxus* diagnostic markers arranged in large introgressed blocks.

Major hybridisation events in *S. cerevisiae* history

To trace past admixture events between *S. cerevisiae* and *S. paradoxus* populations, we constructed a high-resolution strain-by-strain catalogue defining boundaries, positions, and frequencies of introgressed blocks (≥ 5 consecutive *S.p.* markers) across the *S. cerevisiae* clades (Supplementary Fig. 5, Supplementary Fig. 7, Supplementary Table 6). We observed an overall lack of association between introgression block breakpoints and meiotic recombination hallmarks (*p*-value = 0.91), consistent with their resection largely being shaped by mitotic recombination events[29]. Grouping the strains by shared introgressed blocks globally recapitulates their SNPs based phylogenetic relationships, reflecting that evolution from the original hybridisations to the current day introgressed blocks was shared within clades (Supplementary Fig. 8). We further investigated this evidence by extending our computational framework to identify the *S. paradoxus* population ancestry within the introgressed blocks. We also included four *S. cerevisiae* - *S. paradoxus* hybrids, with nearly complete subgenomes from both parental species, that may represent the ancestors of extant introgressed strains (Methods, Supplementary Fig. 1).

We unveiled two major hybridisation events that likely occurred on the European continent. The first can be traced to the AQF *S. cerevisiae* - *S. paradoxus* hybrid isolated in Spain (Fig. 3a) and recognized as a clonal descendant of the ancestor of the Alpechin lineage[29]. Despite the scattered geographical locations and broad sampling timeline (Supplementary Table 2), all Alpechin isolates extensively share common introgression

blocks (Supplementary Fig. 7 and Supplementary Fig. 8). The *S. paradoxus* introgressed sequence displays a uniformly high degree of similarity to the AQF hybrid sequence, thus supporting a single hybridisation origin of the entire clade (Fig. 3b).

The second European hybridisation event is represented by the OS162 *S. cerevisiae* - *S. paradoxus* hybrid, which was isolated from oak tree bark in England. Its genomic profile revealed a triploid hybrid, with one copy of *S. cerevisiae* and two copies of *S. paradoxus* subgenomes, devoid of loss of heterozygosity (LOH) regions (Fig. 3a), the precursors of introgressions[29]. Phylogenetic analysis placed the *S. paradoxus* subgenomes of OS162 in close relationship with the CBS432 European *S. paradoxus* reference strain (Fig. 3b), but not with AQF strain, supporting that distinct European *S. paradoxus* populations experienced interspecific hybridisation independently.

We also identified two hybrids on the American continent. The first includes the UFMG-CM-Y652 Brazilian strain (Fig. 3a and Supplementary Fig. 9) that coexists in a wild environment with multiple *S. cerevisiae* lineages with *S. paradoxus* introgressions[18]. We compared the *S. paradoxus* subgenome of the UFMG-CM-Y652 hybrid with the *S. paradoxus* introgression blocks found in the French Guiana, Mexican agave and the wild Brazilian *S. cerevisiae* strains. The phylogenies are not consistent with UFMG-CM-Y652 wild Brazilian hybrid being the direct ancestors of the three introgressed American *S. cerevisiae* clades, and that either the ancestor, has not been found, or that subsequent admixture events further moulded the genome of extant introgressed clades (Supplementary Table 7).

Finally, the second American hybrid, OS2389 strain, was isolated from a koa tree in Hawaii (Fig. 3a). The *S. paradoxus* subgenome of OS2389 hybrid showed different ancestry from the Hawaiian *S. paradoxus* lineage (UWOPS91-917.1), being instead related to the continental North (YPS138) and South (UFRJ50816) American *S. paradoxus* populations (Fig. 3c, Supplementary Fig. 9). Although OS2389 and the Brazilian hybrid UFMG-CM-Y652 do not show strong evidence of overlapping breakpoints at LOHs, both their *S. paradoxus* and *S. cerevisiae* subgenomes ancestries support a close parental origin making difficult to conclude if they originated from independent hybridisation events (Supplementary Fig. 9).

We further investigated the origin of the *S. paradoxus* subgenome on chromosome II (from 758,700 to 782,580 bp), which corresponds to a *S. paradoxus* introgressed block that is present in several wild Brazilian *S. cerevisiae* strains. Within this region, both the Brazilian hybrid UFMG-CM-Y652 and the Hawaiian hybrid OS2389 displayed the highest sequence similarity with the North American *S. paradoxus* (YPS138). In contrast, the wild *S. cerevisiae* Brazilian introgressed strains radiate from an independent branch, confirming that this block derives from a separate, and as of yet unknown, hybridization event (Fig. 3d). Overall, we identified at least two independent *S. paradoxus* - *S. cerevisiae* hybridizations that persisted as full hybrids in Europe and two hybrids in the American continent. One of the European hybridizations supports the single origin of the European Alpechin clade, whereas *S. paradoxus* introgressions in the American *S. cerevisiae* involves a more complex admixture scenario.

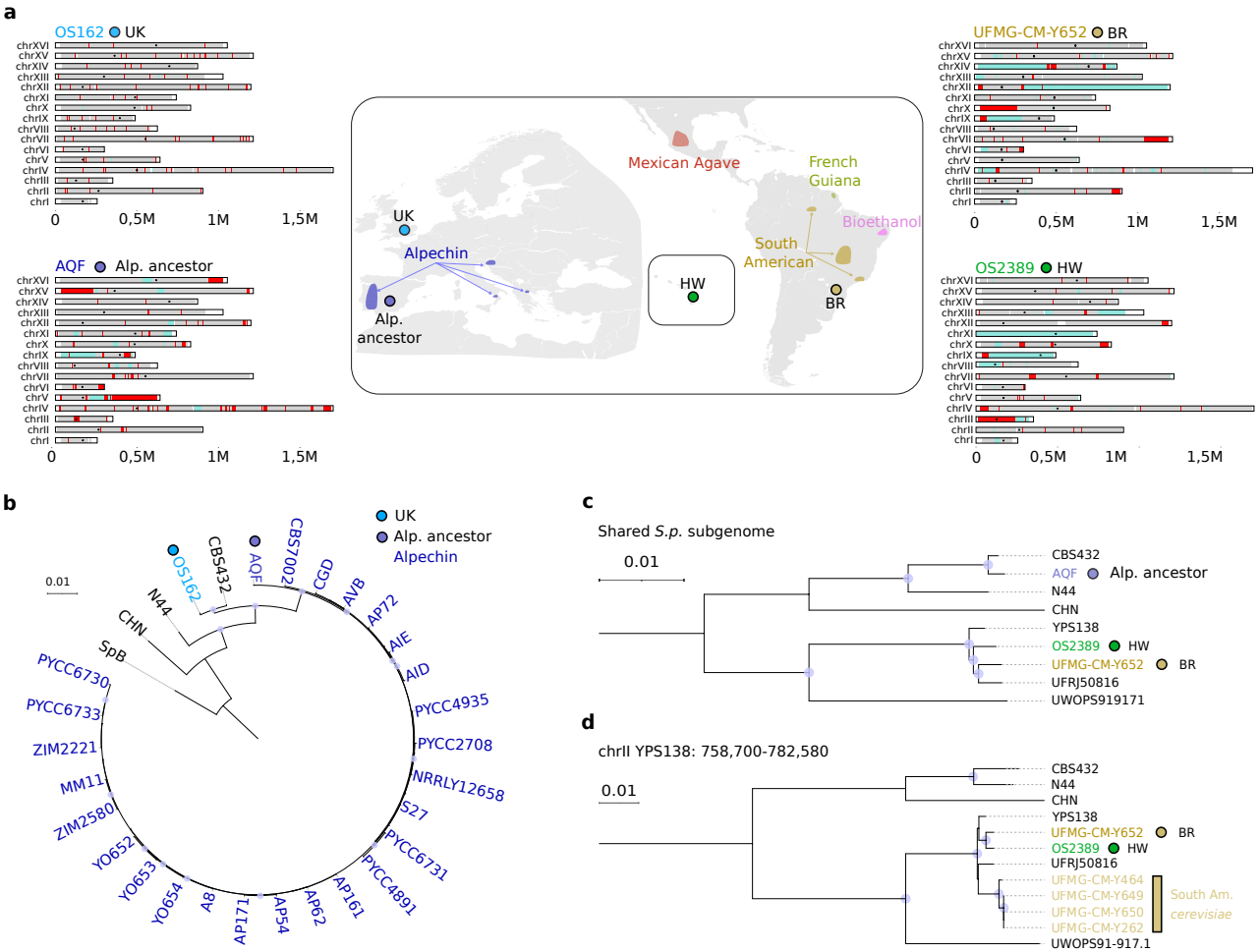


Fig. 3 | Independent hybridisation events. a, *S. cerevisiae* and *S. paradoxus* genomic composition of the Spanish (AQF), Brazilian (UFMG-CM-Y652, BR: Brazilian), and the newly discovered hybrids from UK (OS162) and Hawaii (OS2389, HW: Hawaiian). Colours represent homozygous *S. paradoxus* (red), homozygous *S. cerevisiae* (blue) and heterozygous (grey) regions. The map represents the geographic origin of the hybrids (coloured circles with black border) and introgressed clades (coloured areas). European and American continents are represented at different scales and Hawaii (small box) has been moved to fit the layout. **b**, Maximum likelihood phylogeny of the shared *S. paradoxus* components across hybrid strains with European *S. paradoxus* ancestry (AQF and OS162) and the Alpechin *S. cerevisiae* clade. **c**, Maximum likelihood phylogeny of the shared *S. paradoxus* subgenome of hybrid strains with the American *S. paradoxus* component. **d**, Maximum likelihood phylogeny of the *S. paradoxus* region on chrII:758,700-782,580 (YPS138 coordinates) introgressed in a subgroup of South American *S. cerevisiae* strains and both the Brazilian and Hawaiian hybrids (UFMG-CM-Y652 and OS2389). The blue circles represent the nodes with percentage of bootstrap SH-aLRT (approximate likelihood ratio test[30]) ≥ 95% and UFboot (ultrafast bootstrap approximation[31,32]) ≥ 95%. The size of the node reflects the value of UFboot. N44 Far East, CHN (BJ-DLS32-26) Chinese, SpB (14_101B) North American, UWOPS91-917.1 Hawaiian, UFRJ50816 South American *S. paradoxus*.

A shared introgression underlies an ancient admixture event

We calculated the density of *S. paradoxus* diagnostic markers with American or Eurasian ancestry across strains and observed a generally uniform ancestry within clades (Supplementary Fig. 10a). However, we detected a widespread high density of European-Far East Asian *S. paradoxus* markers across multiple lineages, including strains characterised by genome-wide introgressions *S. paradoxus* American ancestry (Fig. 4a). We found this signal to be driven by a single introgression that encompasses the chromosome III centromeric region (CEN3). This introgression was shared across eighty-three *S. cerevisiae* strains, 75 belonging to the non-Asian

domesticated clades while the other 8 belonging to the Asian domesticated clade (Fig. 4b). The introgression block is absent in the highly diverged Chinese wild lineages, implying that its origin postdates their divergence. The origin of the segment can be unambiguously assigned to the European *S. paradoxus,* with its phylogeny supporting a single origin, i.e. it derives from the same hybridization event (Fig 4c and Supplementary Fig. 10b). In all the 83 strains, the block encompassed the genes located between YCL002C and YCR005C, but in some strains it also spanned genes further away from CEN3 (up to YCL009C and YCR010C). A clustering based on the introgression boundaries showed no correlation with the *S. cerevisiae* populations phylogeny (Supplementary Fig. 10c-d). *S. cerevisiae* centromeres have low recombination rate and CEN3 is particularly refractory given its linkage with the *MAT* locus[33]. Thus, the position of this introgressed block likely increased its chance of persisting over the generations, and reduced the probability of it being lost or resected in the absence of strong deleterious effects. The low frequency in the population suggests that CEN3 introgression might not generally contribute to increasing fitness, nevertheless functional implications (e.g. related to centromere functions or flanking genes) cannot be ruled out.

The presence of an introgression with relatively conserved boundaries and shared ancestry across distinct domesticated *S. cerevisiae* strains points toward an ancestral event with a deep root in time. To test this scenario, we investigated the flanking regions outside the CEN3 introgression block in the industrial AHQ *S. cerevisiae* strain isolated in Taiwan from molasses (Supplementary Fig. 10e-f). Although the genomic background of AHQ is related to the Sake *S. cerevisiae* strains from the Asian domestication group, the 2-kb sequence upstream of the introgression block robustly branched with the Wine/European *S. cerevisiae* clade from the non-Asian domestication group (Supplementary Fig. 10f). This result, along with the common ancestry of the *S. paradoxus* CEN3 introgression, supports the evidence that the introgression was introduced into the Asian lineages through admixture with a Wine/European *S. cerevisiae* rather than from a distinct introgression event. In addition, the absence of the introgression in wild Chinese lineages is consistent with a *S. cerevisiae* / *S. paradoxus* hybridisation that occurred after the initial out-of-China event, while the spread of the introgression across the clades was likely favoured by domestication offering opportunities to move and interbreed[34]. Thus, we report an ancient secondary contact between *S. cerevisiae* and a European *S. paradoxus* that likely occurred on the European continent either before or in concomitance with the *S. cerevisiae* wine domestication, leaving behind CEN3 introgression block.

A convergent adaptive introgression

While most shared introgressions across *S. cerevisiae* strains are explained by common origin, events that arose independently and are maintained over time represent strong candidates of adaptive introgressions. We searched for introgressed genes that were independently acquired from different *S. paradoxus* populations and detected one striking case, corresponding to an introgression block encompassing the *PAD1/FDC1* gene pair (Fig. 4d). The two genes are functionally related and promote the transformation of the toxic compound cinnamic acid and related phenolic acids into potentially useful derivatives, like coumaric, caffeic, ferulic, and sinapic acids[35]. The *PAD1/FDC1* alleles are introgressed in 55 *S. cerevisiae* strains, belonging to the Alpechin, Bioethanol,

Mosaics R3, South American Mix 2 and Mix 3 clades (Supplementary Table 8), with all 40 Alpechin strains retaining such introgression blocks (Supplementary Fig. 7a). We found the *PAD1/FDC1* alleles were introgressed and retained in *S. cerevisiae* at least twice, from either the European or the American *S. paradoxus* subpopulations (Fig. 4d), strongly arguing for an adaptive role. The selection for retaining these *PAD1/FDC1* introgressions across the Alpechin and Bioethanol clades could be exerted by the abundance of cinnamic acid and other phenolic acids in the olive oil wastewater environment[36] and in intermediates of bioethanol production[37]. We experimentally tested this hypothesis by swapping the *PAD1-FDC1* alleles from a European *S. paradoxus* into the *S. cerevisiae* Wine/European background (strain DBVPG6765) that closely resembles the *S. cerevisiae* genomic background of the Alpechin strains but lacks introgressions. We then competed the DBVPG6765 strain with the introduced *S. paradoxus* and the native *S. cerevisiae PAD1-FDC1* alleles against each other, growing them in mixed populations in the presence of stresses representing selection pressures present in olive oil wastewater or brine (Ferulic acid, tyrosol, NaCl). We also included stresses (Phenolic azoles, Rapamycin) to which *S. cerevisiae* laboratory strain lacking *PAD1* or *FDC1* are known to be sensitive[38]. To measure the frequency of each strain in the mixed population (initial frequencies ~0.5), we tagged them with two distinct fluorescence proteins and used the relative amounts of these fluorescence as readouts. We found that the *S. paradoxus PAD1-FDC1* alleles conferred no fitness advantage in absence of stress, but that they improved growth in the presence of phenolic azoles, caffeine and ferulic acids (Fig. 4e). The *S. paradoxus PAD1-FDC1* alleles did not confer tolerance to NaCl, ethanol, rapamycin (which lacks phenolic structures) or tyrosol, which is abundant in olive wastewater, but not known to be degraded by yeast. To our knowledge, this is the first experimental validation of an adaptive *S. paradoxus* introgression in *S. cerevisiae*.

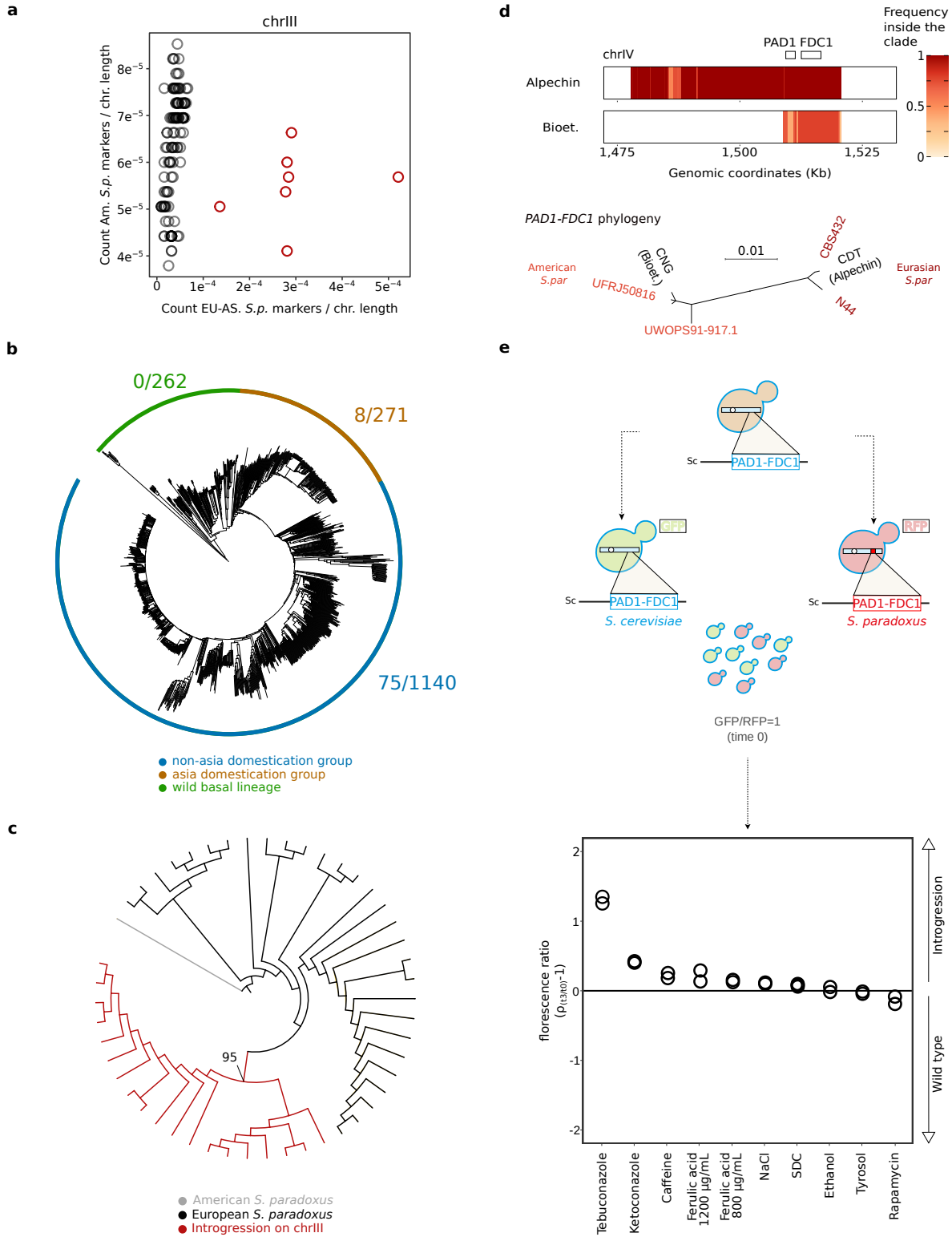


Fig. 4 | Recurrent introgressions in unrelated strains. a, Density of diagnostic markers with Eurasian (x-axis) and American (y-axis) *S. paradoxus* genotypes in *S. cerevisiae* strains with introgressed blocks. Red circles indicate a set of strains with the CEN3 introgressed block. The Alpechin strains were excluded to better visualise the other clades. **b**, Rooted neighbour joining phylogeny of the 1,673 *S. cerevisiae* strains. The coloured arcs annotate wild and domesticated origins, while numbers are counts of strains with the CEN3 introgression block. **c**, Ultrametric maximum likelihood phylogeny of the CEN3 introgression block in *S. cerevisiae* strains and the corresponding orthologous region in European *Saccharomyces paradoxus* strains. 95 indicates the percentage of phylogenies supporting the CEN3 introgression branch.

Branch length was removed to magnify the branch topology. **d**, *Upper panel:* heatmaps of the *S. paradoxus* introgression block at the end of chromosome IV, in the Alpechin and Bioethanol clades. Colour indicates the fraction of strains in each clade that carries the introgressed block. *Lower panel:* maximum likelihood unrooted phylogenetic tree underlies the distinct *S. paradoxus* ancestries of the *PAD-FDC1* introgressions **e,** Graphical representation of the genetic engineering of *PAD1-FDC1* introgression in the *S. cerevisiae* Wine/European DBVPG6765. Fitness in different environments (*x*-axis) of a *S. cerevisiae* DBVPG6765 strain carrying *S. paradoxus PAD1* and *FDC1* alleles compared to that of the WT DBVPG6765 strain. Fitness was measured by competing fluorescently tagged asexual versions of each strain over three growth cycles in each environment and measuring the intensity of each fluorescence, before and after the competition experiment. The y-axis reports the fluorescence ratio (DBVPG6765 with introgression/without introgression) after competition divided by the fluorescence ratio before competition, such that a positive number represents a frequency increase, i.e. higher fitness, of the strain carrying the *S. paradoxus PAD1* and *FDC1* alleles.

Discussion

Shared polymorphisms between populations can be ancient, predating their split, and persist by ILS. We reported abundant ILS polymorphisms in the highly diverged *S. cerevisiae* lineages that separated from the *S. cerevisiae* last common ancestor before the out-of China event. The abundance of ancestral polymorphisms in the highly diverged Asian lineages might reflect different evolutionary histories compared to lineages that post-date the out-of-China event. Asian *S. cerevisiae* lineages were isolated in a temperate forest environment[39,40], which can be perceived as the species ancestral ecological niche. Their persistence in a stable ecological niche without novel selective pressures and dispersal have resulted in fewer or less dramatic population bottlenecks maintaining a larger effective population size that favoured the persistence of ILS alleles. This finding somewhat mirrors African human populations, which share many variants with Neanderthals and Denisovans despite having little or no admixture from these archaic groups[7].

An alternative source of shared polymorphisms derives from the introgression process. We reported a detailed catalogue of introgression blocks in *S. cerevisiae* and unveiled multiple hybridisation events that shaped the species evolution after the out-of China dispersal. These hybridisations are not restricted to domestic environments, illustrating that they can occur in nature and perhaps can be selected for in human-made environments, where introgressed strains have been observed more frequently[14]. Strikingly, three out of the four described hybrids present a signature of genome instability in the form of LOH, which enables introgressions to emerge in reproductively isolated yeast species[29]. This finding supports that genome instability leading to LOH is a common mechanism and contributes to shaping the patterns of introgression across the genome. The remaining hybrid without LOH might represent a recent hybridisation event or did not yet experience the conditions that trigger the genome instability.

Hybrids containing full subgenomes from both species enables in-depth phylogenetic analysis. While the UK, Spanish and Brazilian hybrids have subgenome ancestries consistent with the modern *S. cerevisiae* and *S. paradoxus* populations in these regions, the Hawaiian hybrid subgenomes seem to have originated elsewhere. The *S. cerevisiae* subgenome is very close to the Ecuadorian isolates and the *S. paradoxus* subgenome does not relate to Hawaiian *S. paradoxus*, but instead to the American clade. Therefore, the Hawaiian hybrid or its founding parents likely migrated from the American mainland.

Overall, our study detected either complete hybrids with nearly full subgenomes from both parental species or isolates with up to 6% introgression. On the other hand, we have not detected isolates with more abundant

introgression that represent the F1 hybrid in their backcrossing stages. This is consistent with experimental data showing that gametes with chimeric genomes are highly unfit[29] as a consequence of widespread genetic incompatibilities[41]. Selection against introgressions has been observed in primates[42] and produced deserts depleted of introgression in the human genome[43], perhaps suggesting that the decay of introgressed material is rapid after the initial hybridisation.

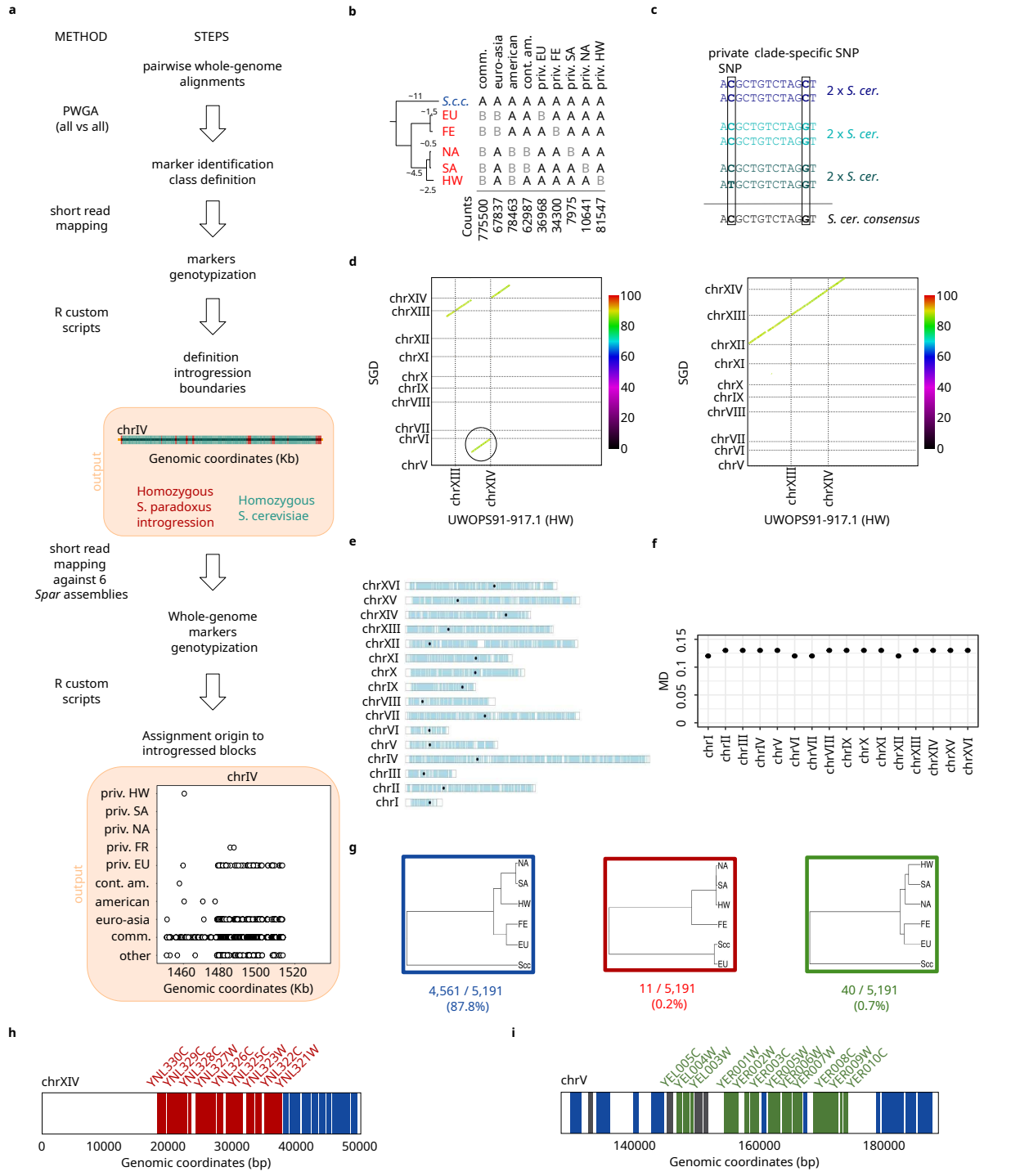
The yeast genetic toolbox enabled us to demonstrate the functional impact of the introgressed *PAD1-FDC1* gene pair on growth in the presence of antifungal drugs and ferulic acid. Such conditions might be relevant in the agricultural and industrial environments and underlie an adaptive potential for this introgression. These genes are known to metabolise cinnamic acid and to produce an off-flavour phenolic derivative (4-vinylguaiacol) in alcoholic beverages. Selection against the undesired 4-vinylguaiacol production has promoted the spread of loss-of-function alleles across the *S. cerevisiae* beer clades[22]. Both the Alpechin and Bioethanol isolates, that independently acquired and maintained the *PAD1-FDC1* introgression, are not used for fermented beverage production, but they share large swaths of the Wine/European and beer genetic backgrounds and potentially carried non-functional *PAD1-FDC1* alleles. An intriguing scenario is that introgressed material in yeast domesticated lineages helps to restore alleles that accumulated loss-of-function across many genes and pathways during domestication[44].

Data and code availability

The genomic data generated in this study and the related code are available at https://github.com/nicolo-tellini/S.cerevisiaeData. The developed computational pipeline, *intropipeline*, is available at https://github.com/nicolo-tellini/intropipeline.

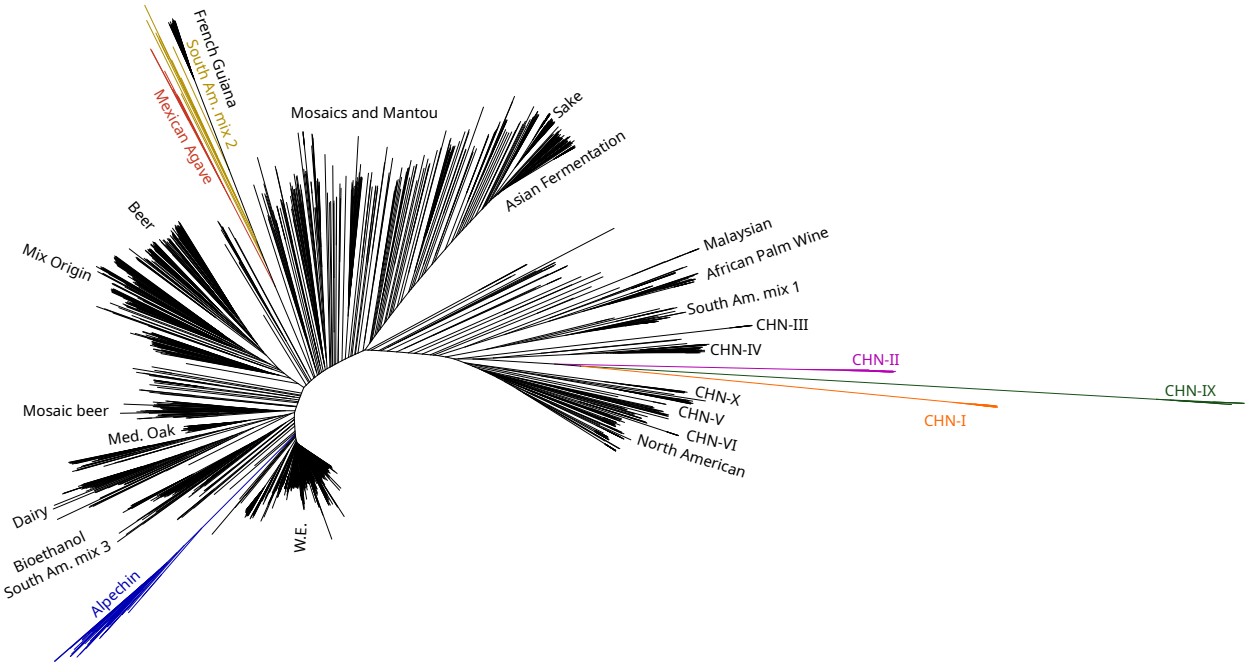
Acknowledgments

We thank Melania D'Angiolo, Eugenio Mancera, Nikolaos Vakirlis, Gilles Fischer, Etienne Danchin and Pedro Beltrao for discussions and critical reading of the manuscript. We also thank Vassiliki Koufopanou for sharing the strain OS162 and Ana Pontes and Jose Paulo Sampaio for providing information on the origin of the Alpechin and Brazilian sequenced strains. This work was supported by Agence Nationale de la Recherche (ANR-11-LABX-0028-01, ANR-15-IDEX-01, ANR-18-CE12-0004, ANR-20-CE12-0020, ANR-22-CE12-0015), Fondation pour la Recherche Médicale (EQU202003010413), UCA AAP Start-up Deep tech, CEFIPRA. NT was partially supported by the PhD fellowship program Region PACA.

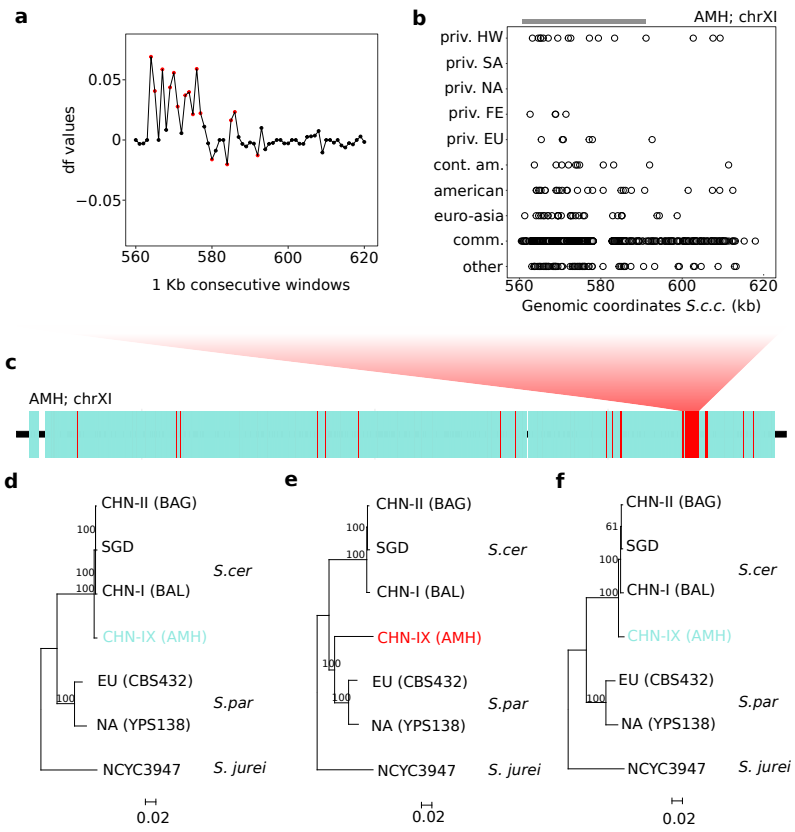


Supplementary Fig. 1 | Methods overview. a, Workflow of the pipeline. Abbreviations: PWGA: pairwise whole-genome alignment; CHR: chromosome. **b**, biallelic patterns across marker positions. A: *S.c.c.* allele; B: alternative allele. Numbers on the phylogeny represent whole-genome sequence divergence between *S.c.c.* and *S. paradoxus* and within the main *S. paradoxus* populations. In red the abbreviation of the main *S. paradoxus* populations. EU: European, FE: Far Eastern, NA: North American, SA: South American, HW: Hawaiian. The columns indicate the marker positions used to define the introgression boundaries (common abbr. comm.) and the origin (the remaining columns). The counts correspond to the number of marker positions available for each pattern. **c**, A cartoon of the strategy adopted to construct the *S.c.c.* sequence, which is explained in the methods. Briefly, for each clade of the 1011 collection, we picked 2 strains and extracted the SNPs against the SGD reference genome. We then used SGD genome as scaffold and changed the alleles in the positions in which the ALT allele was more frequent than the REF allele (freq. ≥ 0.75). **d**, Example of restoring the collinearity of

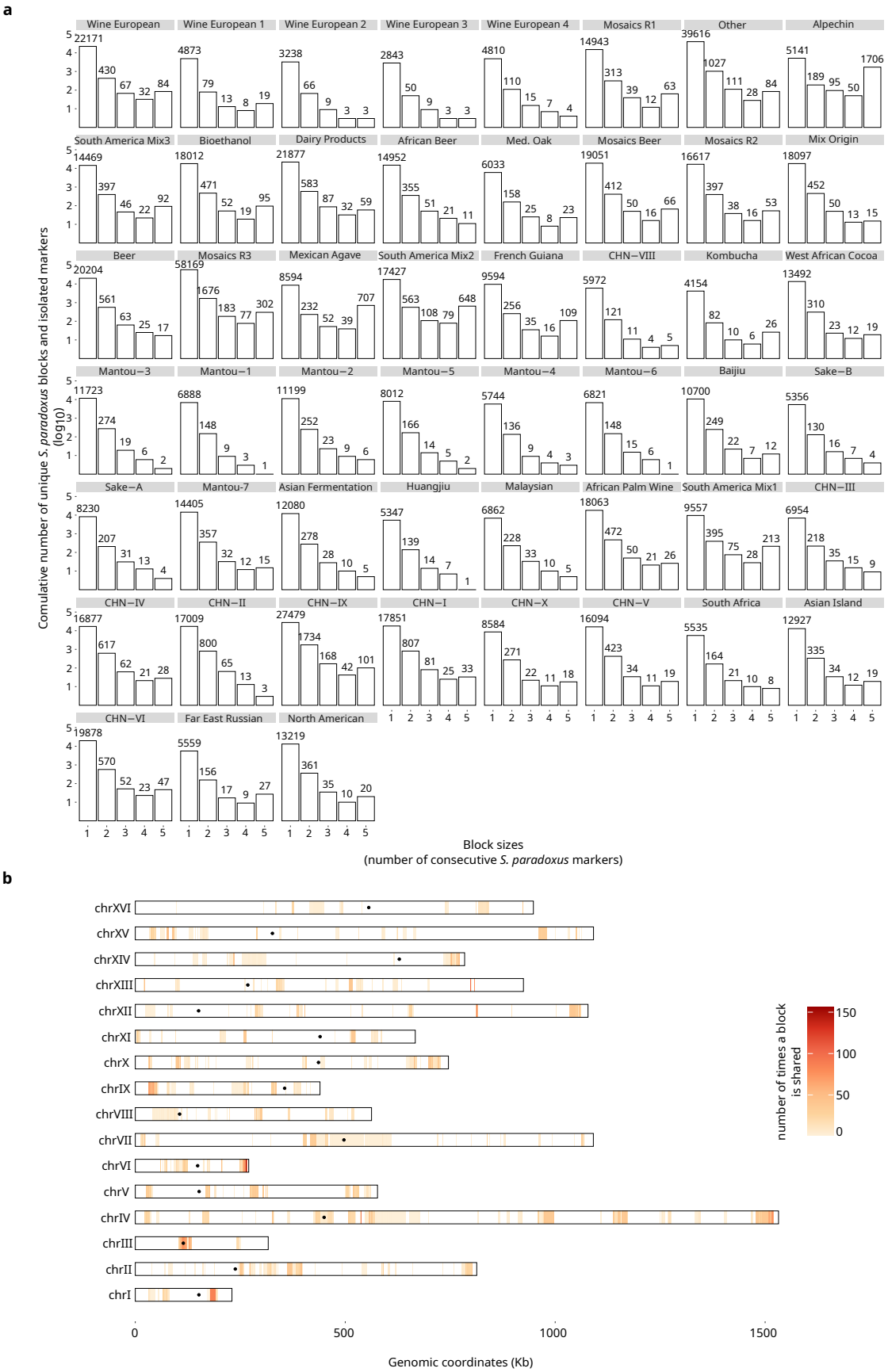
the translocated genomic region between *S.c.c.* and HW *S. paradoxus* on the translocation chromosome V/ chromosome XIII described in[28]. The colour of the line reflects the sequence divergence between *S.c.c.* and the HW *S. paradoxus*. **e**, The blue rectangles represent the genomic regions, in *S.c.c.* coordinates, which were effectively aligned across *S.c.c.* and all the *S. paradoxus* whole-genome assemblies. **f**. The per-chromosome marker density (MD) is measured as the number of diagnostic markers divided by the sum of the aligned regions depicted in the f. panel. **g**, UPGMA phylogenies across the genome assemblies of 5191 *S. cerevisiae* - *S. paradoxus* 1-to-1 orthologs. In blue, the most abundant individual gene topologies that follow the structure of the species tree. In red and green two contrasting gene topologies. **h**, zoom-in in the distribution of contrasting gene phylogenies across the assemblies. On the left, the *S. cerevisiae* introgression on chromosome XIV on the European population of *S. paradoxus*. In red, the genes with the phylogeny depicted in the central panel in **g**. On the right, the introgression / ancestral variation shared by the South American and the Hawaiian *S. paradoxus* on the region surrounding the centromeric position of chromosome V. In green, the genes with the phylogeny depicted in the right panel in **g**; in grey, gene trees with alternative topologies. In both the zoom-in, the blue rectangles represent genes whose phylogeny respects the species tree topology.



Supplementary Fig. 2 | Global *S. cerevisiae* phylogeny. Unrooted neighbour-joining tree of the 1673 *S. cerevisiae* strains analysed in this work. Coloured clades are discussed in the main text.

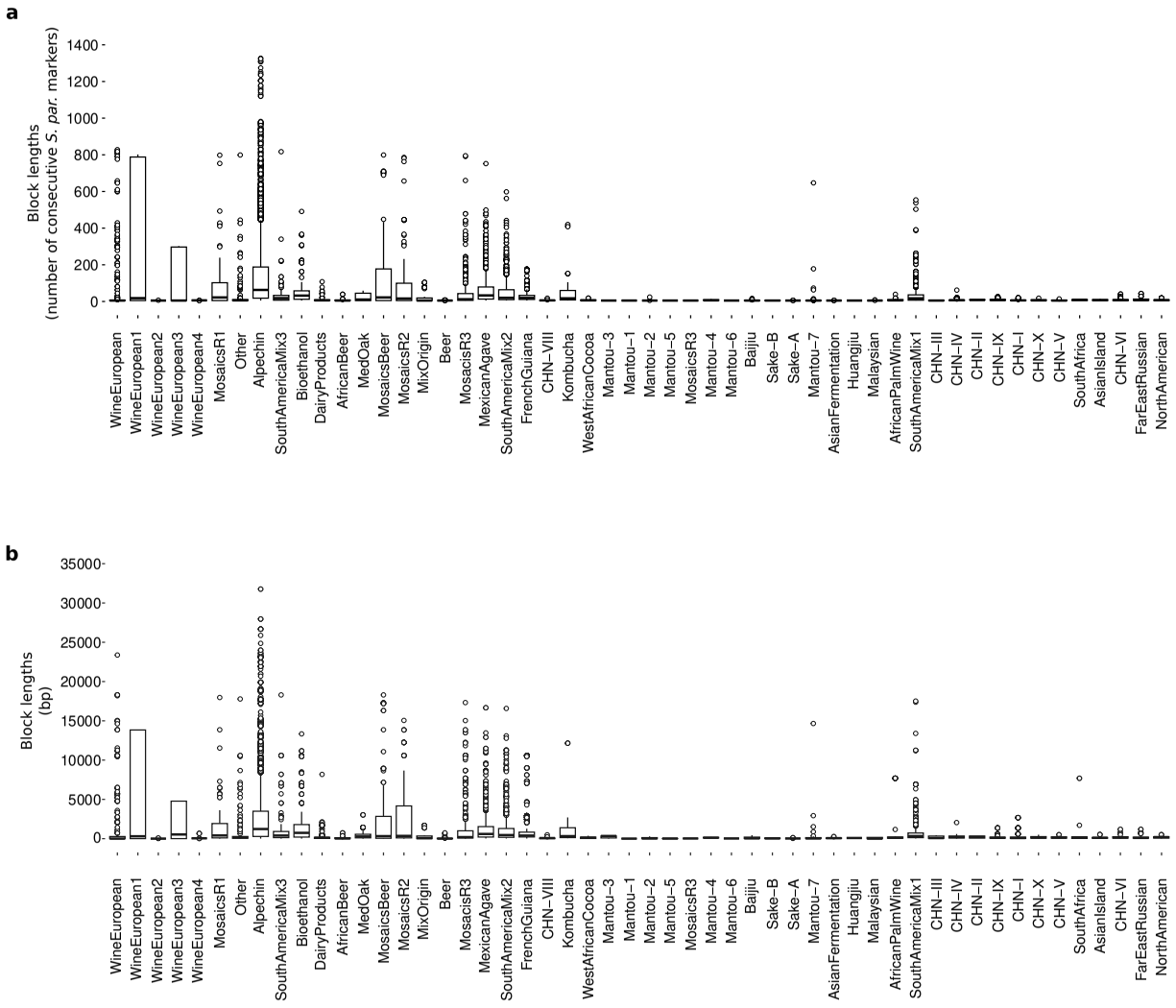


Supplementary Fig. 3 | Unknown origin introgression on chromosome XI in AMH. a, df statistics support the presence of the introgression on chromosome XI. Red dots represent genomics windows characterised by an absolute value of df 5 folds or higher compared to the average value across the chromosome. **b,** Polymorphism ancestry plot shows that the origin of the chromosome XI introgression cannot be retraced to any *S. paradoxus* population, consistent with its origin from an unknown sister species. The grey bar indicates the position of the introgression. **c**, Introgression boundaries in chromosome XI on AMH. The red blocks represent homozygous introgressions while the blue blocks are homozygous for *S. cerevisiae*. **d-f**, Neighbour-Join phylogenies (kimura 2-parameter substitution model) of the sequences spanning the genes *YKR064W-YKR078W* derived from *de novo* whole-genome assemblies[28,45,46]. The tree on panel **d** and **f** are derived from 25 kb flanking regions before and after the introgression, while the tree in panel **e** is the introgressed region.

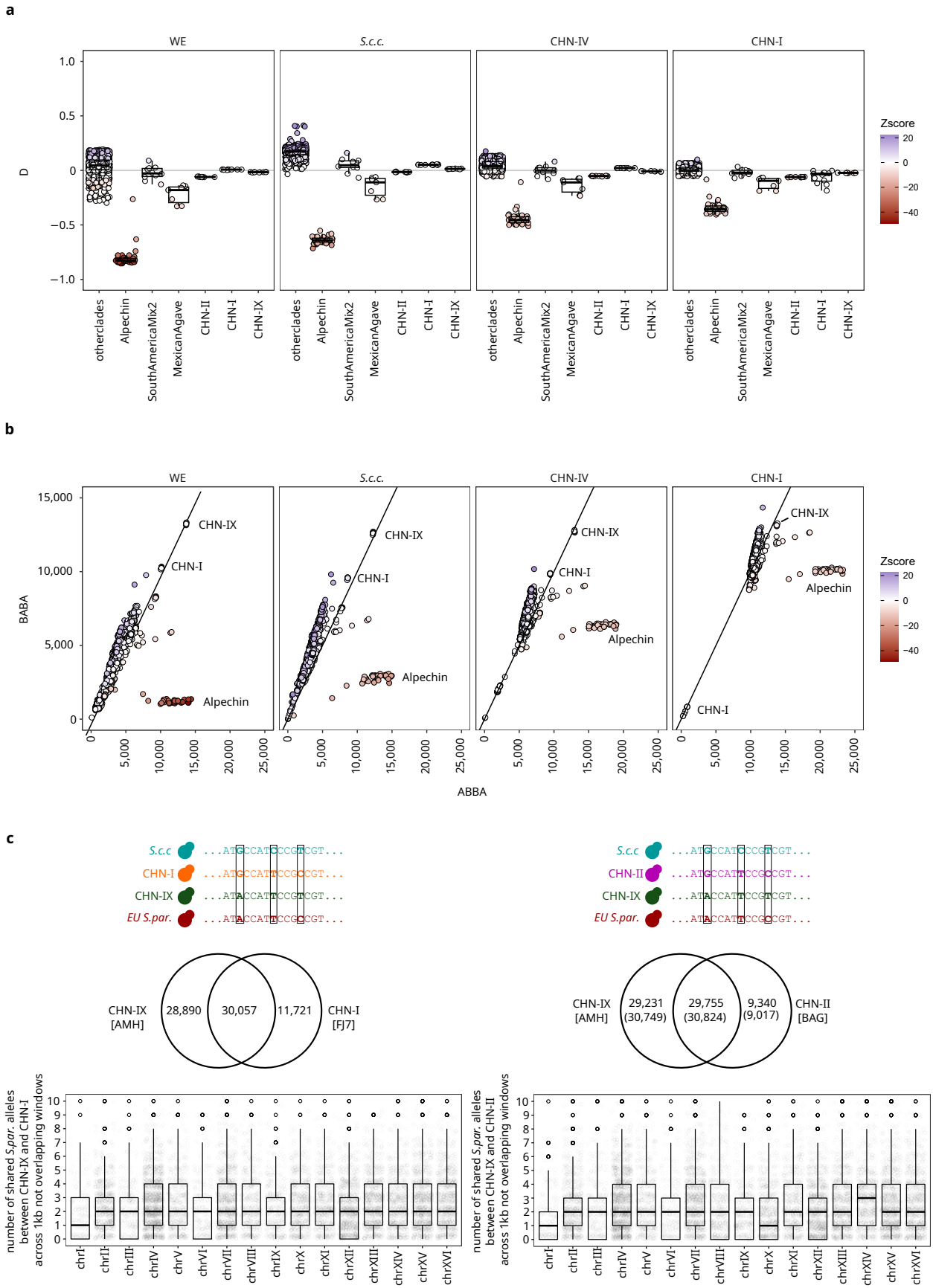


Supplementary Fig. 4 | Introgression blocks sizes and location. a, distribution of the diagnostic markers with *S. paradoxus* genotypes groped by size across different clades. For each clade, isolated markers and introgressed blocks with

the same boundaries are counted once. Overlapping blocks with different boundary coordinates are counted as separate events. The last column included all the blocks with at least 5 consecutive *S. paradoxus* markers. The Y-axis is on log10 scale. The absolute value of the counts is indicated on top of each column. The label "Other" indicates *S. cerevisiae* strains that could not be placed in a specific clade. **b**, physical positioning and frequency of the introgression blocks supported by ≥ 5 consecutive *S. paradoxus* markers across 1459 out of 1673 *S. cerevisiae* samples. The coloured scale reflects the number of times a specific block is shared across the 1459 *S. cerevisiae* strains. Two blocks shared by more than 156 *S. cerevisiae* strains were dropped to 156 to allow the visualisation of less common events. The genomic coordinates indicate positions in the *S.c.c. genome*.

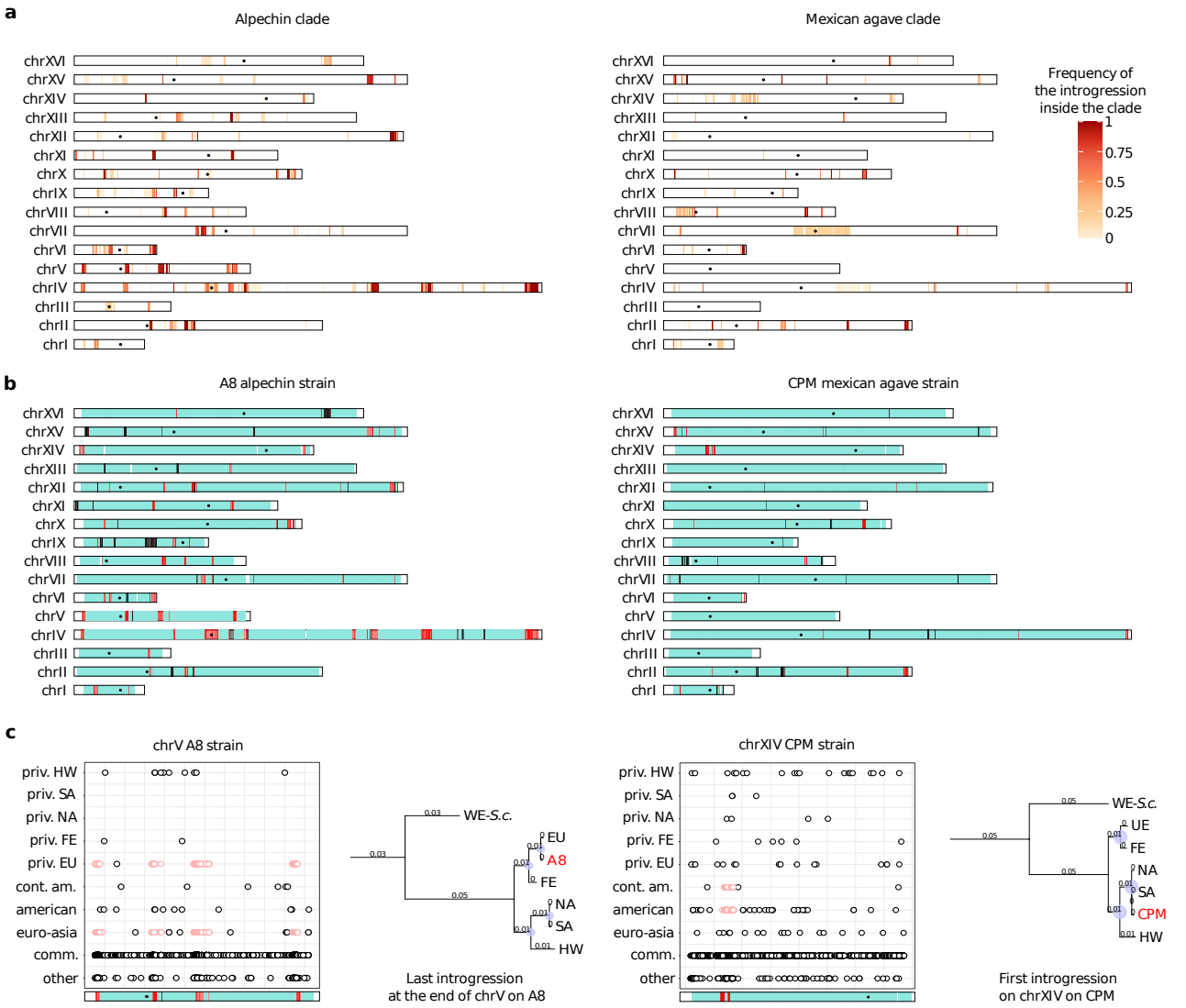


Supplementary Fig. 5 | Introgression block size. a, introgression lengths distributions across *S. cerevisiae* clades defined as the number of consecutive *S. paradoxus* markers. **b**, introgression lengths distributions across *S. cerevisiae* clades defined as the total length (in bp) of regions within *S. paradoxus* markers.

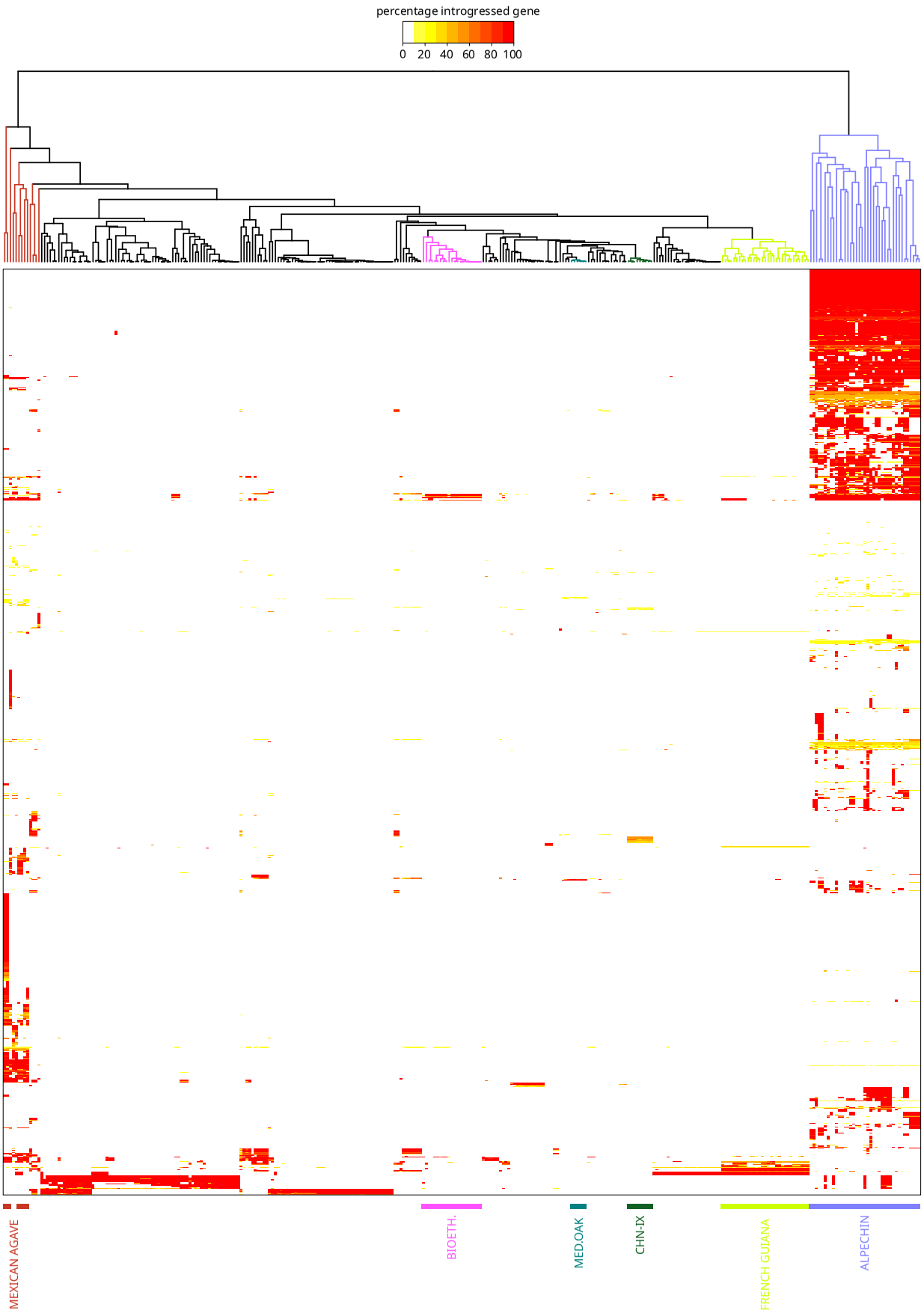


Supplementary Fig. 6 | Patterson's D statistics and whole-genome alignments. a, D values measured across the 1671 *S. cerevisiae* collection using different quartet input arrangements. The strains on the top of each plot represent the P1

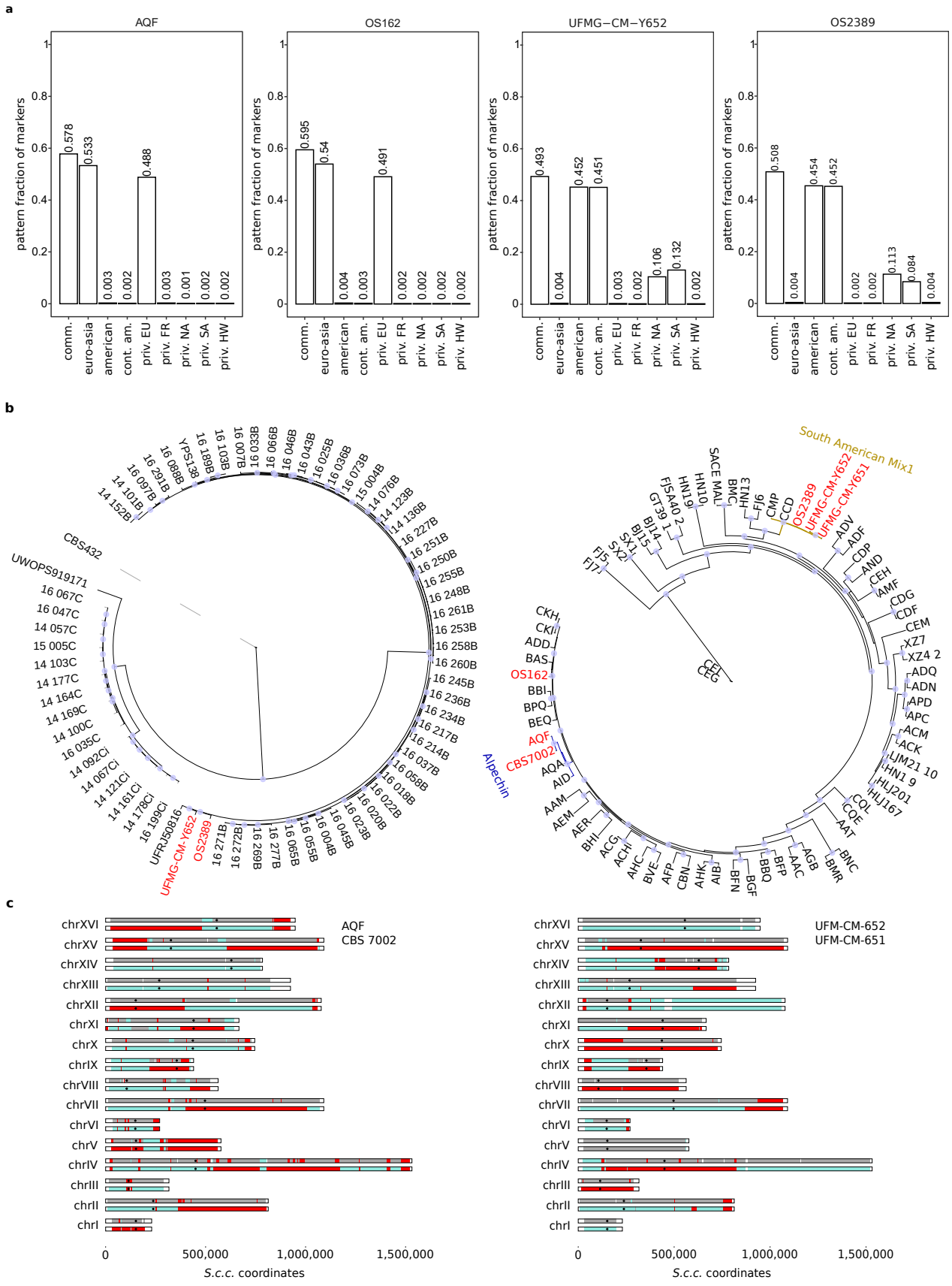
(WE: ADS, CHN-IV: BJ3 and CHN-I: BAL) population. P3 and P4 (Outgroup) are fixed and represented by the European *S. paradoxus* (CBS432) and *S. jurei,* respectively. Multiple D values are calculated, for each sample, by mean jack-knife resampling of genomic blocks. The gradient colour reflects the Z-score. D values associated with a Z-score equal or greater than the absolute value of 3 are considered statistically significant and the null hypothesis of absence of gene flow can be rejected. **b**, Absolute counts of both ABBA and BABA sites across the *S. cerevisiae* strains. **c**, *S. par.* alleles at polymorphic positions. On top, the cartoons represent examples of the polymorphic sites taken into account; in the middle, the Venn diagrams show the private and shared *S. paradoxus* alleles between the Chinese isolates; at the bottom, the boxplots show the distribution of the only shared *S. par.* alleles between the Chinese isolates. In the squared brackets the name of the strains; in the round brackets the number of sites obtained using the *de novo* genome assemblies[45] of BAG, AMH and CBS432 isolates depicted in the cartoon above.



Supplementary Fig. 7 | Highly introgressed clades. a, Frequency and genomic position of the introgressions detected across the Alpechin (N=40, on the left) and Mexican Agave (N=7, on the right) clades. **b**, Example of introgressions detected in two different strains (A8, Alpechin and CPM, Mexican agave). The red blocks represent homozygous *S. paradoxus* introgressions. The grey blocks are heterozygous regions while the blue blocks are homozygous *S. cerevisiae* regions. **c**, Plot that highlights the origin of diagnostic markers of the introgressions in chromosome V in the strain A8 and on chromosome XIV in the CPM strain. The origin of the two introgressions is supported by a maximum likelihood phylogeny. The circle represents nodes with SH-aLRT ≥ 80% and UFboot ≥ 95%. The size of the circle at the nodes reflects the value of SH-aLRT and UFboot. The Wine European *S. cerevisiae* is the outgroup. The numeric values represent the branch length.

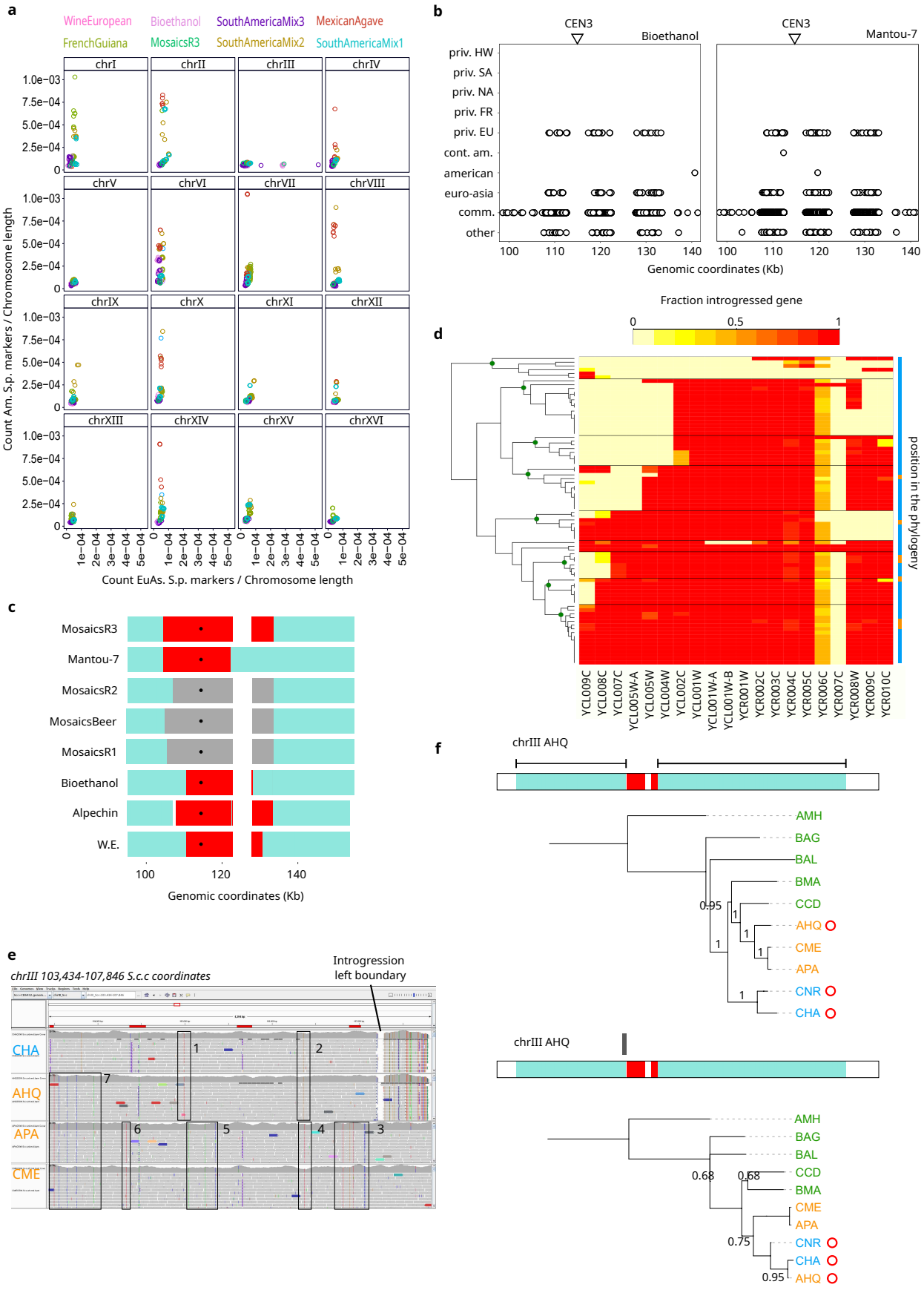

percentage introgressed gene
0 20 40 60 80 100
MEXICAN AGAVE
BIOETH.
MED.OAK
CHN-IX
FRENCH GUIANA
ALPECHIN

Supplementary Fig. 8 | Strains grouping by introgressed genes. The heatmap shows the *S. paradoxus* percentage of introgressed genes across *S. cerevisiae* strains with at least 5 introgressed genes. To reduce the complexity of the heatmap, we retained only the genes introgressed in at least one of the remaining *S. cerevisiae* strains (322 strains x 1305 genes). The hierarchical clustering was performed across the strains (columns). We coloured the branches of the strains for which the patterns strictly follow the clade division. Because of the dimension of the heatmap, to better visualise the patterns, the heatmap cells with a percentage value of gene introgressed between 0 and 10% were removed (white areas inside the heatmap).



Supplementary Fig. 9 | Hybrid subgenome ancestries. a, Fraction of allelic pattern, at marker positions, detected across the *S. paradoxus* subgenomes of the four hybrid isolates, **b**, Maximum likelihood phylogenies of the *S. paradoxus* (on the

left) and *S. cerevisiae* (on the right) subgenomes of hybrid isolates (red label). The *S. paradoxus* phylogeny was constructed including American *S. paradoxus* strains[47], while the *S. cerevisiae* phylogeny was constructed with a selection of *S. cerevisiae* strains from previous studies[14,15]. The circle represents nodes with SH-aLRT ≥ 95% and UFboot ≥ 95%. The size of the circle at the nodes reflects the value of SH-aLRT and UFboot (from 95 to 100). **c**, genomic profiles of hybrid-descendant pairs from Spain (on the left) and Brasil (on the right). Details of the strain and genomes origin are given in the methods (section *Saccharomyces* collections). Red and blue blocks represent homozygous *S. paradoxus* and *S. cerevisiae* regions respectively, while grey blocks are heterozygous regions.


a
WineEuropean
Bioethanol
SouthAmericaMix3
MexicanAgave
FrenchGuiana
MosaicsR3
SouthAmericaMix2
SouthAmericaMix1
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
chrXVI
Count Am. S.p. markers / Chromosome length
Count EuAs. S.p. markers / Chromosome length
b
CEN3
Bioethanol
Mantou-7
priv. HW
priv. SA
priv. NA
priv. FR
priv. EU
cont. am.
american
euro-asia
comm.
other
Genomic coordinates (Kb)
c
MosaicsR3
Mantou-7
MosaicsR2
MosaicsBeer
MosaicsR1
Bioethanol
Alpechin
W.E.
Genomic coordinates (Kb)
d
Fraction introgressed gene
position in the phylogeny
YCL009C
YCL008C
YCL007C
YCL005W-A
YCL005W
YCL004W
YCL002C
YCL001W
YCL001W-A
YCL001W-B
YCR001W
YCR002C
YCR003C
YCR004C
YCR005C
YCR006C
YCR007C
YCR008W
YCR009C
YCR010C
e
chrIII 103,434-107,846 S.c.c coordinates
Introgression left boundary
CHA
AHQ
APA
CME
f
chrIII AHQ
AMH
BAG
BAL
BMA
CCD
AHQ
CME
APA
CNR
CHA
chrIII AHQ
AMH
BAG
BAL
CCD
BMA
CME
APA
CNR
CHA
AHQ

Supplementary Fig. 10 | Ancient introgression on chromosome III. a, Density of *S. paradoxus* diagnostic markers across a subset of introgressed strains. The x-axis and y-axis respectively indicate the density of *S. paradoxus* diagnostic markers that support an Euroasiatic or American origin of the introgressions. Chromosome III has diagnostic markers with Euroasiatic ancestry also in clades with genome-wide introgressed blocks with American origins. **b**, Ancestry plot of diagnostic markers within the chromosome III introgression block in a Brazilian bioethanol and a Mantou 7 strains. The origin of diagnostic markers is attributed to different *S. paradoxus* populations (listed along the Y-axis). Genomic coordinates are from the *S.c.c.* genome. **c**, Map of introgressed blocks detected around chromosome III centromere (black dot) across different *S. cerevisiae* strains. Red and grey colours represent homozygous and heterozygous *S. paradoxus* introgression respectively, blue indicates the *S. cerevisiae* genome. **d**, heatmap with introgressed genes (x-axis) detected on the region encompassing the centromere of chromosome III across different *S. cerevisiae* strains. The strains are clustered by their introgression profile (left side), while the coloured bar (right side) indicates their phylogenetic assignment illustrated in Figure 4b. **e**, Alignment of the short reads against *S.c.c.* of a selection of four strains. CHA, Wine European, AHQ Mantou 7, APA and CME Sake strains. The rectangles 1 and 2 highlight shared SNPs between CHA and AHQ which are absent in the Sake strains. The rectangles 3, 4, 5, and 6 highlight private SNPs of the Sake clade. The rectangle 7 encloses the SNPs in AHQ shared with the Sake strains but absent in the Wine European strain (CHA). **f**, Maximum likelihood phylogenies of the arms of chromosome III, external to the introgressed block, and the 2 kb left-side flanking region of the introgression for a selection of strains. The introgression and the subtelomeric/telomeric regions are excluded. Red circles indicate strains with introgressed CEN3.

Methods

Saccharomyces collections

Genome sequences analysed in this study were previously published. The majority of the isolates belong to the 1011 *S. cerevisiae* collection[14]. We included additional datasets to widen the sample size of several clades and specifically added strains from the following sources: wild and industrials Chinese[15], wild Brazilian[18], olive oil[19], wild North American[27], beer[22,48], Brazilian Cachaça[21], flor[26], clinical[25] and industrial[23,24]. We checked the fastqs files quality with FastQC and assembled a collection of 1676 strains that covers all the currently known *S. cerevisiae* diversity in terms of both ecology and geography (Supplementary Table 2). The Alpechin strains from Greece and Italy have been recently sequenced by us and not included in the global analysis of this paper, but they present an introgression pattern consistent with the previously described ones.

We also included strains with a complete *S. cerevisiae* X *S. paradoxus* hybrid genome that had been either previously published or sequenced specifically for this study. We reanalysed the Alpechin living ancestor hybrid strain AQF[29]. This strain is the equivalent of the CBS7002, which was also sequenced by Pontes et al.[19] as monosporic isolate, thus explaining the recombined genome configuration in this study. Two strains, UFMG-CM-Y651 and UFMG-CM-Y652, isolated in Brazil[18] also harbour a hybrid genome structure. While the UFMG-CM-Y652 was sequenced in its natural state, the UFMG-CM-Y651 was sequenced as a monosporic isolate, consistent with its recombined genome profile. The genome composition of the UFMG-CM-Y651 is consistent with the original diploid strain being similar to the UFMG-CM-Y652 and therefore likely belonging to the same F1 hybrid population. Finally, we sequenced two hybrid strains: OS162 and OS2389 isolated in the UK and Hawaii respectively. These strains were sequenced at The Earlham genomics facilities (www.earlham.ac.uk) according to the LITE pipeline[49]. The resulting libraries were run as a 384-plex pool on two NovaSeq SP lanes with 150 bp paired-end reads. For each lane it generated 540M clusters passing filters (84%) with 92% bases >=Q30.

In addition to the *S. cerevisiae* strains and genomes, we used five reference-quality genome sequences representative of *S. paradoxus* populations[28] for detecting the *S. paradoxus* shared polymorphisms.

S. cerevisiae phylogeny

The short reads were mapped against the *Saccharomyces cerevisiae* consensus *S.c.c.* by *BWA (v.0.7.16a* options *-M -t*) and *samtools view (v. 1.7-12* options *-bS -F 1284*). Not primary aligned, unmapped and duplicated reads were filtered out. *samtools sort* and *index* (default options) were used to sort the *bam* files and produce the indexes (*bai*). We generated genome VCFs (gVCFs) by *bcftools mpileup* (options *-q 5 --annotate DP --skip-indels*) and *bcftools call* (*-m -Oz*). We streamlined the gVCFs by rewriting the structure with *bcftools query* (options -f '%CHROM \t %POS \t %REF \t %ALT \t %QUAL \t %INFO/DP \t %FORMAT\n') and performed custom filtering with *awk* (using a threshold of 20 for quality and 10 for depth). The phylogenetic tree was built in R with the *SNPRelate* package as follow: generation of *gds* file *snpgdsVCF2GDS* (option *method="biallelic-only"*), import gds file *snpgdsOpen*, construct the dissimilarity matrix comparing each pair of strains

(*snpgdsDiss*, options *autosome.only = F, missing.rate=0.01*), run *bionj* algorithm[50], ladderized the tree with *ladderize*. The phylogenetic tree was printed out using *iTol* with an unrooted layout (Supplementary Fig. 2). Based on the new extended phylogeny, we revised the clade nomenclature to accommodate geographical and ecological information. We remained faithful to previous naming with few exceptions. We grouped the clades "French Dairy"[14] and "Milk Chinese"[15] in a single "Dairy products" clade; we labelled two distinct Sake subclades (Sake-A and Sake-B). We kept the designation of "Asian fermentation" from the 1011 collection[14], despite these strains are not monophyletic in this extended phylogeny and are scattered in other Asian clades[15]. Wild chinese clades were maintained as in Duan et al., 2018. Finally, the wild Brazilian clades (B1, B2, B3 and B4, Barbosa 2016) and the Cachaça strains[21] were regrouped under the "South America Mix 1", "South America Mix 2" and "South America Mix 3".

***S. cerevisiae consensus* (*S.c.c.*) sequence construction**

To minimise the bias due to the use of a single reference sequence, we constructed a *S. cerevisiae* consensus genome. We selected 48 representative strains of the 1011 *S. cerevisiae* high coverage collection described in[14] (Supplementary Table 1) and we calculated the allele frequency across the selected strains for each single nucleotide polymorphism (SNPs). Alternative alleles with frequency higher or equal to 0.75 (15,559 positions) replaced the corresponding reference alleles in the SGD reference genome (strain: S288C; version: SGD R64-1-1).

Restoring collinearity in *S. paradoxus* genomes

The two genome assemblies from *S. paradoxus* isolates bearing large chromosomal rearrangements (UFRJ50816 and UWOPS91-917.1[28]) were modified to obtain a collinear structure with respect to the *S.c.c.* genome. We used the break points previously described and the genomic annotations of UFRJ50816 and UWOPS91-917.1 were modified accordingly. The restored collinearity was inspected by aligning the original and the modified *S. paradoxus* genomes to the *S.c.c.* genome with *MUMmer v3* (*nucmer* algorithm options --*mum*). We filtered the *delta* file with *delta-filter* (options *-q -r* and *-u)* to keep the one-to-one alignments and we inspected the result through *mummerplot* (options --postscript, --color).

***Saccharomyces cerevisiae* strains ploidy estimation**

We quantified ploidy of the isolates using the allele balance (AB) at heterozygous positions called with *freebayes* (*v.1.2.0* default options). The AB is defined as the number of reads supporting the alternative allele (ALT) divided by the total number of reads mapping to that position.

Definition of diagnostic markers dataset

The six assemblies of *S.c.c.*, CBS432, N44, YPS138, UFRJ50816 and UWOPS91-917.1 were masked in correspondence of both repetitive elements (LTR, Core X, Y', Ty elements) and rapidly evolving sequences such as telomeric regions, introns, pseudogenes and noncoding exons, replacing the sequences with '*N'* values.

We ran *MUMmer v3* (*nucmer* algorithm options *--mum*) to align the masked assemblies in pairwise combinations, chromosome by chromosome, taking advantage of their restored collinearity. The delta files were filtered with *delta-filter* (*-r -q -l 400*) keeping reference-query alignments longer than 400 bp and allowing for inversions. *show-snps* (options *-ClrT*) was used to extract the polymorphic positions. We identified a set of 1,257,196 markers positions; 1,205,445 (95.9%) were biallelic, 50,835 (4%) triallelic and 916 (0.07%) tetrallelic. We retained only the biallelic markers for the downstream analysis. We identified and flagged different allele patterns occurring at each biallelic markers position, across the assemblies, defining ten categories. The first set, which is the most abundant, consists of 775,500 biallelic markers in which *S. paradoxus* allele is conserved across the five assemblies, while the *S.c.c*. allele is different (species-specific markers). These alleles are considered to have arisen after the split between the species. The second set consists of 67,837 markers in which alleles are specific to the euroasiatic *S. paradoxus* populations (CBS432 and N44 assemblies), while *S.c.c.* allele is shared with the American *S. paradoxus* populations (YPS138, UFRJ50816 and UWOPS91-917.1 assemblies). The third set consists of 78,463 markers in which alleles are specific to the American *S. paradoxus* populations. *S.c.c.* allele is shared with the two Euroasiatic *S. paradoxus* populations. The fourth set consists of 62,987 markers in which the alleles are specific to the continental American *S. paradoxus* populations (YPS138, UFRJ50816 assemblies) while the *S.c.c.* allele is shared with the Euroasiatic *S. paradoxus* populations and the Hawaiian *S. paradoxus*. The fifth set consists of 36,968 markers in which one allele is private of the European *S. paradoxus* population (CBS432 assembly). The sixth set consists of 34,300 markers in which one allele is private to the Far Eastern *S. paradoxus* population (N44 assembly). The seventh set consists of 7,975 markers in which one allele is private to the North American *S. paradoxus* population (YPS138 assembly). The eighth set consists of 10,641 markers of which one allele is private to the South American *S. paradoxus* population (UFRJ50816 assembly). The ninth set consists of 81,547 markers of which one allele is private to the Hawaiian *S. paradoxus* population (UWOPS91-917.1 assembly). The tenth and last set consists of 49,227 markers in which the allelic patterns do not follow the phylogeny of the species. We collected the number of markers in 1-kb non-overlapping windows (Supplementary Fig. 1) and we measured the markers density across the chromosomes (density = number of markers in one chromosome / total size of the blocks aligned with *MUMmer*) showing that the markers are evenly distributed across the chromosomes (Supplementary Fig. 1).

Identification of *S. paradoxus* alleles and introgressions through marker patterns in *S. cerevisiae*

We used a modified version of MuLoYDH[51] to screen 1,673 *S. cerevisiae* strains to identify those enriched in *S. paradoxus* alleles using as input the dataset of species-specific markers stored as BED files with either *S.c.c.* or CBS432 coordinates.

The FASTA genomes were indexed with *samtools faidx* and *bwa index* (default options). The mapping of the short reads, against both *S.c.c.* and CBS432, was performed by piping *BWA (v.0.7.16a* options *-M -t*) and *samtools view (v.1.7-12* options *-bS -F 1284*). Not primary aligned, unmapped and duplicated reads were filtered out. *samtools sort* and *index* (default options) sorted the *bam* files and produced the indexes (*bai*). We collected

the statistics about the short read mapping (*samtools flagstat*) and the coverage (*samtools depth* option *-a* piped with *awk*). The markers positions were genotyped using the sorted *bam*, piping *samtools mpileup* (options *--positions -u -min-MQ5 --output-tags AD,ADF,ADR,DP,SP --skip-indels --redo-BAQ*) and *bcftools call* (*--ploidy -c -Oz*). The ploidy option was set up to either 1 or 2; whenever this information was missing (NA), or the ploidy was higher than 2, we performed the call with the default option (*--ploidy 2*). A preliminary step with *GEM-indexer* (v.1.423 options -c dna) followed by *GEM-mappability* (v.1.315 options -l) provided the read-length specific mappability information for both the FASTA genomes. *Control-FREEC* v10.7 was run to detect the copy number variants of the sample with known ploidy level. A per-sample configuration file was prepared with custom yeast parameters (*telomeric=4000, coefficentOfVariation=0.05, minExpectedGC=0.35, maxExpectedGC=0.40*); the significance of *Control-FREEC* predictions were assessed with *assess_significance.R* which returns two *p-values* (from a Wilcoxon and Kolmogorov-Smirnov test) for each CNVs. Only CNVs detected against both the references and supported by both p-values < 0.01 were retained. CNVs were only investigated on samples with known ploidy. markers genotyped against both the input assemblies were filtered following a strict strategy to keep the number of false positive calls low. We discarded marker positions from the mitochondrial genome and both telomeric and subtelomeric regions because of their known highly-content variability[28]. Furthermore, we extended the filtering on chrXIV from 1 to 38,500 because the European *S. paradoxus* population contains an introgression from *S. cerevisiae*[52] that would produce false-positive introgressions. Markers genotyped against *S.c.c.,* but not those against *S. paradoxus,* were quality filtered (QUAL > 20). In addition, we only retained markers whose genotypes were consistent across both mappings (example: REF A ALT G against *S.c.c.* must correspond to REF G ALT A against CBS432). For haploid and diploid strains, the positions embedded in CNVs were filtered out because the altered ploidy may compromise the step of genotyping. Introgressed segments, defined as regions with at least 5 consecutive markers presenting *S. paradoxus* alleles, were generated with custom R scripts (Supplementary Table 4).

To trace back the origin of the introgressions, we performed a second mapping that requires as input all six assemblies. We selected a few strains to maximise different patterns of introgressions. The FASTA genomes were indexed with *samtools faidx* and *bwa index* (default options). The mapping of the short reads, against *S.c.c.*, CBS432, YPS138, UFRJ50816 and UWOPS91-917.1 was performed piping *BWA* (*v.0.7.16a* options *-M -t*) and *samtools view* (*v. 1.7-12* options *-bS -F 1036*). *samtools sort* and *index* (default options) sorted the *bam* files and produced the indexes (*bai*). The markers positions were genotyped across the short-read sorted *bam*, piping *samtools mpileup* (options *--positions -u -min-MQ5 --output-tags AD,ADF,ADR,DP,SP --skip-indels --redo-BAQ*) and *bcftools call* (*--ploidy -c -Oz*). The markers were kept when the alternative allele was concordant in at least 4 out the 6 mappings. The presence of markers private to specific subpopulations was used to infer the origin of the event.

Comparing whole-genome phylogeny with gene-based trees

S. paradoxus has a well defined population structure. To identify genomic regions having a discordant evolutionary history, we extracted 5,191 1-to-1 orthologous genes across *S. paradoxus* and *S.c.c.* assemblies

based on genome annotations. We discarded ambiguously annotated, broken or double annotated sequences. For each ortholog we performed multisequence alignment with *MUSCLE* (*v3.8.31*) and we generated an UPGMA gene tree (MUSCLE with *-maketree default options*). We used the Robinson–Foulds metric (*RF.dist* function in R package phangorn v.2.10.0) to measure the distance of gene trees from the species tree. Most gene trees (4,569/5191; 88%) retraced the species tree structure and are informative for detecting potentially introgressed genes. Gene-based trees with a phylogeny discordant to the species phylogeny (622/5,191; 11%) were reported. Among these 569 were 2 operations distant from the species tree, 48 measured 4 operations and 6 trees 5 operations. At the end for each gene we measure the Kimura Two-Parameter distance matrix (function *distmat* of EMBOSS tools, *-nucmethod 2 -position 123,* Supplementary Table 9).

In order to assess the presence of *S. cerevisiae* introgression within *S. paradoxus,* we extracted and compared the phylogenies for 5,191 1-to-1 orthologs which were determined on the basis of the annotations of 5 *S. paradoxus* genomes, representing the available populations, and the *S. cerevisiae* reference genome.

Identification of introgressions from *S. paradoxus* using statistical tests

We developed a pipeline that, taking as input either short read datasets or *de novo* assemblies, performs Patterson's D[8], *f* [53,54] and *df* [55] statistics. These statistics rely on the comparison of alleles across 4 populations (P) whose phylogenetic relationship can be described as (((P1,P2),P3),O) where O stands for outgroup. In our study, P1 and P2 represent two different populations of *S. cerevisiae*, P3 a population of *S. paradoxus* while *S. jurei* is used as an outgroup in P4.

For each *S. cerevisiae* strain selected, we generated an artificial whole genome assembly by replacing the specific isolated alternative alleles on the scaffold of S. c. c. To do this, we mapped the selected *S. cerevisiae* short-reads against *S.c.c.* by piping *BWA* (*v.0.7.16a* options *-M -t*) and *samtools view* (*v.1.7-12* options *-bS*). Duplicated reads were filtered out. *samtools sort* and *index* (default options) sorted the *bam* files and produced the indexes (*bai*). The variant calling was performed piping *samtools mpileup* (options *--positions -u -min-MQ3 --output-tags AD,ADF,ADR,DP,SP --skip-indels --redo-BAQ*) and *bcftools call* (*-vm -Oz*). Variants with QUAL < 20 were discarded as well as INDELs and multiallelic positions. Finally, the alternative allele of the remaining variants were used to replace the allele on the *S.c.c* scaffold with a home-made script.

We then performed multiple whole-genome alignment with *progressiveMauve*[56] (released 2015-02-13) default option. All the alignments were run respecting the input order imposed by the phylogeny: (((P1,P2),P3),O) where P1 rotated between *S.c.c*., a Wine/European, a CHN-IV and a CHN-I isolate. Every *S. cerevisiae* isolate of the collection was used as P2, P3 is the *de novo* assembly of the European *S. paradoxus* (CBS432) and O is the *de novo* assembly of *S. jurei*. We extracted the SNPs from the *xmaf* file with *org.gel.mauve.analysis.SnpExporter*. The P2 alleles were considered reference (0 in the vcf) and only biallelic positions were kept. We took advantage of the availability of *S.c.c.* and CBS432 annotations for excluding the SNPs in telomeric and subtelomeric regions and SNPs located in different chromosomes when the chromosome of P1 P2 and P3 were compared as these genomes are collinear. The resulting VCF was equipped with a VCF header for downstream analysis.

PATTERSON' S D STATISTIC

We integrated the script written by Joana Meier (https://github.com/joanam/scripts/blob/master/convertVCFtoEigenstrat.sh), *convertVCFtoEigenstrat.sh* (rec=0.3cM/Kb) for converting the VCF file in the corresponding map, geno and ind files. Patterson's D statistics was run with ADMIXTOOLS (*v.7.0.2*) in R using the package admixr (*v.0.9.1*) function *d()*. Assuming around 2.2 million of SNPs per sample we tested different blgsize and we selected 0.12 which resulted in around 345 blocks for the jackknife resampling. A number of 345 blocks corresponds to around 6,300 SNPs per block; given 1 SNP every 5.5 bp (11,000,000 bp / 2,200,000 SNPs =5.5) each block cover a genome length of around 34,650 bps (6,300 snps * 5.5 bp) a distance at which the linkage disequilibrium approaches low r2 values close to the baseline[8,57,58]. The D values were reported together with the sd error, the Zscore and the number of sites for both ABBA and BABA patterns.

f STATISTICS

The *f* statistics estimates the fraction/percentage of genome introgressed. We ran the *f* statistic implemented in the R package PopGenome (*v.2.7.5*) with introgression.stats (*vcf,do.D = T, do.df=F,do.RNDmin = F, block.size = F, keep.site.info = TRUE*).

df STATISTICS

We subset the VCF chromosome-by-chromosome and ran the *df* statistic implemented in the R package *PopGenome* (*v.2.7.5*) in 200 bp non overlapping sliding windows; we then highlighted windows for which *df* values were 5 times higher than the whole-chromosome mean value.

PROOF OF CONCEPT

As a proof of concept we reconstructed the genome of an Alpechin strain known to be introgressed (AHL) and a CHNIX (AMH) strain for which one introgression block from an unknown *Saccharomyces spp* on chrXI was detected[14]. For both the strains we compared the results obtained with 1) the reconstructed false genome and 2) the *de novo* genome.

Introgression breakpoint and hotspot association

The association between the introgression breakpoints detected in the Alpechin strains and DSB/recombination hotspots was tested by means of the *regioneR* package (*v.1.30.0*)[59]. The association tests were performed by means of the function *overlapPermTest* setting the parameters *ntimes = 10000*, and *alternative = "greater"*. The fasta suite[60] (*fasta36 -b 3 -d 3 -m 8, version: 36.3.8d April, 2016*) was used to convert the coordinates of the hotspots regions detected in the original reference genomes (SGD_R62-1-1_20090218 and SGD_R58-1-1_20080305 for the hotspots reported by Pan et al. 2011[61] and Mancera et al. 2008[33], respectively) to the corresponding coordinates of the *S.c.c.* genome. Introgression breakpoint regions were defined, e.g., as the

genomic interval between the first marker of an introgression region and the closest flanking marker which does not belong to the same introgression region. Overall, we tested 3868 hotspot regions (median width 196 bp) and 13267 introgression breakpoints (median width 19 bp), detecting 482 overlaps. Our test did not reveal any association between the DSB/recombination hotspots and introgression breakpoints (p = 0.95). The robustness of the approach was previously assessed by applying the test to a collection of LOH breakpoints formed through the return-to-growth protocol and recombination hallmarks, revealing a strong association[62].

Phylogeny of the introgression on chromosome III

We collected 30 short read sequencing for European *S. paradoxus* populations sampled from Canada[47], USA[63], UK, Russia, Ukraine, Italy, Spain, Portugal, Germany and Latvia[64]; one North American *S. paradoxus* strain (YPS138)[28] that served as an outgroup and 18 representative strains with the introgression on chrIII from different clades: Wine European, Alpechin, Bioethanol, Mosaics beer, Mantou 7, Mosaics, the Alpechin hybrid described in D'Angiolo *et al* 2020[29] and its spore described by Pontes *et al* 2019[19]. For each strain, we ran *bwa* in a competitive mapping concatenating *S.c.c.* and CBS432 (European *S. paradoxus* genomes) and we removed PCR duplicated reads. As all the samples (except the hybrid) were homozygous and diploid we performed position genotypization (ploidy 2) restricted to the shortest shared introgressed area against chrIII of CBS432 from position 129,241 to 159,904 (*samtools mpileup -u -l CBS432.chrIII.bed --adjust-MQ 50 -min-MQ5 --output-tags AD,ADF,ADR,DP,SP --redo-BAQ -f CBS432.genome.fa sample | bcftools call -m -Oz > sample.chrIII.variants.gvcf.gz*). The resulting gvcf were filtered for the variant quality (vcftools --gzvcf --minQ 20 --recode --recode-INFO-all --out). For each strain we reconstructed its FASTA sequence by using the sequence of CBS432 as scaffold and replacing the positions corresponding to a variant site with the ALT allele. Since we wanted to compare intra-lineage short sequences we performed a stringent masking to keep only reliable positions. We masked the following segments: 129,241-132,000; 132,977-133,178; 153,001-153,799 and 159,530- 159,904; plus 1) the positions corresponding to INDELs; 2) multi allelic variant positions; 3) positions for which we were not able to identify the alleles in more than 4 samples and 4) positions missing against the outgroup (North American *S. paradoxus*) as we could not assume sequence identity with the European *S. paradoxus*. We then refined the masking. For each gene within the introgression, we counted the number of masked positions (Ns) and if more than 5% of the gene length was replaced by Ns the entire sequence was completely masked; 5/15 genes were completely masked (*YCL009C, YCL008C, YCL005W, YCL005W-A* and *YCR003W*). We joint the FASTA of the samples in a single multi FASTA file and we constructed the phylogeny with *MEGA X* (*v.10.1.8*) using Maximum Likelihood method and Kimura 2-parameter model (other options included complete deletion and initial tree generation by mean Neighbour-Join and BioNJ algorithms; non-coding nucleotide sequence were assumed). To conclude, we collected shared variants across the strains using the *UpsetR* package (version 1.4.0).

Phylogeny across shared *S. paradoxus* introgression

For each strain from South American mix1, South American mix 2, French Guiana and Mexican Agave clades we performed a competitive mapping using *bwa* and concatenating S.c.c. and European *S. paradoxus* CBS432 genomes and we removed PCR duplicated reads. We performed whole-genome position genotypization (*samtools mpileup -u -l CBS432.bed -min-MQ3 --output-tags AD,ADF,ADR,DP,SP --redo-BAQ -f CBS432.genome.fa sample | bcftools call -m -Oz > sample.gvcf.gz*). The resulting gvcfs were filtered for the variant quality and read depth; we retained only the biallelic positions and we discarded the INDELs (*vcftools --gzvcf --minQ 20 --minDP 10 --max-alleles 2 --remove-indels --recode --recode-INFO-all --out*). We reconstructed the FASTA sequence for each strain by using the CBS432 as scaffold the sequence and replacing the corresponding variant site positions with the ALT allele. We performed a masking to keep only reliable positions and replaced with N all the bases for which one sample selected for a specific introgression was missing the call. We joined the FASTA of the samples in a single multi-FASTA file. Only multi-FASTA with a reasonable number of masked positions were visually inspected for short read alignment and used to construct the phylogeny with *MEGA X* (*v.10.1.8*) using the Maximum Likelihood method and Kimura 2-parameter model (other options included complete deletion and initial tree generation by mean Neighbor-Join and BioNJ algorithms; non-coding nucleotide sequence were assumed). Shared variants across the strains were investigated using the *UpsetR* package (*v.1.4.0*).

Phylogeny of *S. cerevisiae* and *S. paradoxus* hybrid subgenomes

For each hybrid strain, we performed a competitive mapping using *bwa* and concatenating S288C reference genome (version: R64-1-1) and either the American *S. paradoxus* YPS138 for the American hybrids or the European *S. paradoxus* CBS432 genomes for the European hybrids. For the phylogeny of the *S. cerevisiae* subgenome, we included rappresentative strains from the main *S. cerevisiae* clades. From the bam files, we performed whole-genome position genotypization for each sample (*bcftools mpileup -E -Ou -q 5 -a DP -f genome.fa sample.bam | bcftools call -mO v | bcftools view --max-alleles 2 --exclude-types indels -e 'QUAL <= 20 || FORMAT/DP <= 10' | bcftools annotate -x ID,INFO,FILTER | bgzip > sample.gvcf.gz"*). We then merged the gvcfs in a single multi-sample gvcf file and kept only the positions against the S288C reference genome that were genotyped across all the samples. The multi-sample gvcf file was converted in the PHYLIP file format with vcf2phylip.py with options -m 1000 -o CEI. CEI is a CHN-IX *S. cerevisiae* strain and it was used as an outgroup. The PHYLIP file was the input for the construction of the phylogeny with the software IQ-tree (*iqtree -s file.phy -bb 1000 -alrt 1000 -bnni -nt AUTO -m TEST+ASC*). The phylogenetic tree was represented using *iTol* with a rooted circular layout (Supplementary Fig. 9b, panel on the right).

For the phylogeny of the *S. paradoxus* subgenome of the European hybrids we included: the Alpechin *S. cerevisiae* strains, the European *S. paradoxus* CBS432, and one representative for each Euroasiatic *S. paradoxus* population while the American *S. paradoxus* was used as outgroup. Insted, for the phylogeny of the *S. paradoxus* subgenome of the American hybrids we included the American *S. paradoxus* strains from Eberlein *et al* 2019[47] and Yue e*t al* 2017[28]. Using the gvcfs generated as described above, *w*e merged the sample gvcfs in a single multi-sample gvcf file but, we kept: the positions against the European *S. paradoxus* CBS432 reference genome

for the European strains and, the positions against the American *S. paradoxus* YPS138 reference genome for the American strains. In both the cases, we kept only the positions genotyped across all the samples. The two gvcf files, for the European and American S. paradoxus ancestry, were converted in PHYLIP files and the phylogenies obtained as described above. The phylogeny of the *S. paradoxus* ancestry of the European strains was depicted in Fig.3b while the phylogeny of the *S. paradoxus* ancestry of the American strains was depicted in Supplementary Fig. 9b (panel on the left). The same competitive mapping strategy was used, with different groups of strains, for the phylogenies depicted in Fig. 3c, 3d and Fig. 4d.

CRISPR/Cas9 plasmid assembly and introgression engineering

The introgression encompassing the genes *PAD1* and *FDC1* was engineered using CRISPR/Cas9 genome editing. The plasmid pUDP004 harbouring Cas9 was obtained from Addgene (https://www.addgene.org/101165/). The resistance to acetamide in pUDP004 was replaced with the resistance cassette to Nourseothricin generating the plasmid pL59. The plasmid pL59 was linearized using BsaI. Two gRNAs cutting on the border of the introgression were designed on UGENE and ordered as a synthetic oligo from Eurofins Genomics (™). The synthetic oligos were cloned into a single linearized plasmid by using the Gibson assembly kit (NEB, Gibson Assembly®) and the ligation reaction was carried out for 1 hour at 50 °C. The assembled plasmid was transformed into DH5-alpha competent bacteria by heat shock and the bacteria were incubated in 3 mL of LB broth for 1 hour to induce the synthesis of the antibiotic resistance molecules and then plated on LB plates containing 100 µg/µL of ampicillin. The following day, cells were screened by polymerase chain reaction (PCR) using primers to validate the correct golden gate assembly of the construct. Successfully transformed bacterial colonies were inoculated in LB broth containing 100 µg/uL of ampicillin and incubated overnight at 37 °C. Cells were harvested from the overnight incubation and the plasmid was extracted using the QIAprep Spin Miniprep Kit following the manufacturer's instructions. The introgressed region was amplified from the Alpechin strain OS872 using oligos with 60 nucleotides of homology flanking the *PAD1-FDC1* region in the Wine european strain DBVPG6765. Yeast samples were transformed using 50 µL of PCR reaction, and 200 ng of the constructed CRISPR/Cas9 plasmid using the lithium acetate protocol. Cells were then plated on selective media containing kanamycin (400 µg/mL) and incubated at 30 °C for 5 days. Candidate transformed clones were validated by PCR using a primer designed outside the artificially introgressed region and one inside the introgressed region. Positive clones were streaked on YPD (Yeast extract 1%, Peptone 2%, Dextrose 2%, Agar 2%) and grown for 2 days at 30 °C to allow plasmid loss. Plasmid loss was then confirmed by plating again the colonies in the selective medium and positive ones were patched on YPD and stored at -80 °C in 25% glycerol tubes.

Insertion of a cassette encoding fluorescent proteins at the HO locus

The fluorescent protein yGamillus was amplified by PCR from the plasmid pL42, while the fluorescent protein mRuby2 was amplified by PCR from the plasmid pL71. The sequence of yGamillus was obtained from the

Genescript construct pUC57-Kan-yGamillus TEF1ov. The sequence of mRuby2 was obtained from the plasmid pFA6a-link-yomRuby2-Kan generated by Kurt Thorn's lab[65].

The plasmid targeting the HO locus was generated using the pUDP004 backbone as described above, but in this case two gRNAs targeting the HO locus were used. The transformation of the cassettes bearing the fluorescent proteins to be inserted in the HO locus upon Cas9 editing was carried out as described above. Colonies were validated first by diagnostic PCR, using a primer binding inside the coding sequence of the fluorescent protein and one primer binding outside the HO locus and later by flow cytometry screening to validate the correct fluorescence of the proteins.

Competitive growth assay

The two competing strains ygl4147 (MATa, *ho*::yGamillus, *ura3::KanMX*) and ygl4151 (MATa, *ho*::mRuby, *ura3::KanMX*, *pad1-fdc1*^{Sc}::*PAD1-FDC1*^{Sp})were patched from glycerol stocks onto YPD plates and incubated overnight at 30°C. The following day part of the patch was transferred to two different falcon tubes per strain containing 5 mL liquid SDC media (0.67% YNB w/o amino acids, 2% dextrose) and grown overnight (16-18 hours) at 30°C with shaking at 220 rpm. The following day we measured the OD_{600} of the four independent cultures and transferred an equal amount of cells of the two competing strains (total OD_{600} = 1) from the overnight cultures to two new tubes containing SDC or an SDC based media to which we added one of the following reagents: ferulic acid (800 µg/mL or 1200 µg/mL), caffeine (0.002 mM), rapamycin (30 nM), ethanol (10% v/v), tebuconazole (0.0037 mg/mL), ketoconazole (0.0075 mg/mL), NaCl (0.6 mM) or tyrosol (0.6 mg/mL). We took 200 µL of cultures, which were sonicated and then analysed by flow cytometry using the filters B525-FITC and Y585-PE on the Cytoflex LX (Beckman Coulter) to evaluate the difference of the GFP/RFP ratio at T_0 in the population. The cultures were grown for 24 hours at 30°C with shaking at 220 rpm. The following day the OD_{600} of the cultures was measured and diluted to an OD_{600} of 1 in fresh media containing the same compounds that were added at the beginning of the experiment and repeated the same transfer the following day. Finally, we collected 200 µL from each culture and transferred them to a 1.5 mL Eppendorf tube. The samples were sonicated and then analysed following the same procedure used for the T_0 and using the same gating strategy as in the T_0. Events that were not included in the T_0 gates were discarded as they were often small (low FSC-H) compared to the median cell size of the population and were likely cell debris.

References

1. Nei, M., Suzuki, Y. & Nozawa, M. The Neutral Theory of Molecular Evolution in the Genomic Era. *Annu. Rev. Genomics Hum. Genet.* **11**, 265–289 (2010).
2. Schrempf, D. & Szöllösi, G. The Sources of Phylogenetic Conflicts. *No Commer. Publ. Authors Open Access Book* 24 (2020).
3. Sousa, V. & Hey, J. Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* **14**, 404–414 (2013).
4. Harrison, R. G. & Larson, E. L. Hybridization, Introgression, and the Nature of Species Boundaries. *J. Hered.* **105**, 795–809 (2014).
5. Taylor, S. A. & Larson, E. L. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nat. Ecol. Evol.* **3**, 170–177 (2019).
6. Suarez-Gonzalez, A., Lexer, C. & Cronk, Q. C. B. Adaptive introgression: a plant perspective. *Biol. Lett.* **14**, 20170688 (2018).
7. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
8. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722 (2010).
9. Slon, V. *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **561**, 113–116 (2018).
10. Mao, Y. *et al.* A high-quality bonobo genome refines the analysis of hominid evolution. *Nature* **594**, 77–81 (2021).
11. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol. CB* **26**, 1241–1247 (2016).
12. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
13. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533-1545.e20 (2018).
14. Peter, J. *et al.* Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
15. Duan, S.-F. *et al.* The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nat. Commun.* **9**, 2690 (2018).
16. Ono, J., Greig, D. & Boynton, P. J. Defining and Disrupting Species Boundaries in Saccharomyces. *Annu. Rev. Microbiol.* **74**, 477–495 (2020).
17. Clark, A., Dunham, M. J. & Akey, J. M. *The genomic landscape of* Saccharomyces paradoxus *introgression in geographically diverse* Saccharomyces cerevisiae *strains*. http://biorxiv.org/lookup/doi/10.1101/2022.08.01.502362 (2022) doi:10.1101/2022.08.01.502362.
18. Barbosa, R. *et al.* Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. *Genome Biol. Evol.* **8**, 317–329 (2016).
19. Pontes, A., Čadež, N., Gonçalves, P. & Sampaio, J. P. A Quasi-Domesticate Relic Hybrid Population of Saccharomyces cerevisiae × S. paradoxus Adapted to Olive Brine. *Front. Genet.* **10**, 449 (2019).
20. Peris, D. *et al. Macroevolutionary diversity of traits and genomes in the model yeast genus* Saccharomyces. http://biorxiv.org/lookup/doi/10.1101/2022.03.30.486421 (2022) doi:10.1101/2022.03.30.486421.
21. Barbosa, R. *et al.* Multiple Rounds of Artificial Selection Promote Microbe Secondary Domestication—The Case of Cachaça Yeasts. *Genome Biol. Evol.* **10**, 1939–1955 (2018).
22. Gallone, B. *et al.* Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell* **166**, 1397-1410.e16 (2016).
23. Gonçalves, M. *et al.* Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. *Curr. Biol.* **26**, 2750–2761 (2016).
24. Legras, J.-L. *et al.* Adaptation of S. cerevisiae to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Mol. Biol. Evol.* **35**, 1712–1727 (2018).
25. Ramazzotti, M. *et al.* Population genomics reveals evolution and variation of Saccharomyces cerevisiae in the human and insects gut. *Environ. Microbiol.* **21**, 50–71 (2019).
26. Coi, A. L. *et al.* Genomic signatures of adaptation to wine biological ageing conditions in biofilm-forming flor yeasts.

Mol. Ecol. **26**, 2150–2166 (2017).

27. Almeida, P. *et al.* A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol. Ecol.* **24**, 5412–5427 (2015).
28. Yue, J.-X. *et al.* Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).
29. D'Angiolo, M. *et al.* A yeast living ancestor reveals the origin of genomic introgressions. *Nature* **587**, 420–425 (2020).
30. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
31. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast Approximation for Phylogenetic Bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
32. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
33. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* **454**, 479–485 (2008).
34. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
35. Ramos-Cormenzana, A., Juárez-Jiménez, B. & Garcia-Pareja, M. P. Antimicrobial activity of olive mill wastewaters (alpechin) and biotransformed olive oil mill wastewater. *Int. Biodeterior. Biodegrad.* **38**, 283–290 (1996).
36. Jamoussi, B., Bedoui, A., Hassine, B. B. & Abderraba, A. Analyses of phenolic compounds occurring in olive oil mill wastewaters by GC–MS. *Toxicol. Environ. Chem.* **87**, 45–53 (2005).
37. Klinke, H. B., Thomsen, A. B. & Ahring, B. K. Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Appl. Microbiol. Biotechnol.* **66**, 10–26 (2004).
38. Hillenmeyer, M. E. *et al.* The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. *Science* **320**, 362–365 (2008).
39. Wang, Q.-M., Liu, W.-Q., Liti, G., Wang, S.-A. & Bai, F.-Y. Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Mol. Ecol.* **21**, 5404–5417 (2012).
40. Lee, T. J. *et al.* Extensive sampling of Saccharomyces cerevisiae in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Res.* **32**, 864–877 (2022).
41. Bozdag, G. O. *et al. Engineering recombination between diverged yeast species reveals genetic incompatibilities*. http://biorxiv.org/lookup/doi/10.1101/755165 (2019) doi:10.1101/755165.
42. Vilgalys, T. P. *et al.* Selection against admixture and gene regulatory divergence in a long-term primate field study. 8 (2022).
43. Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLOS Genet.* **14**, e1007349 (2018).
44. De Chiara, M. *et al.* Domestication reprogrammed the budding yeast life cycle. *Nat. Ecol. Evol.* **6**, 448–460 (2022).
45. O'Donnell, S. *et al. 142 telomere-to-telomere assemblies reveal the genome structural landscape in* Saccharomyces cerevisiae. http://biorxiv.org/lookup/doi/10.1101/2022.10.04.510633 (2022) doi:10.1101/2022.10.04.510633.
46. Naseeb, S. *et al.* Whole Genome Sequencing, *de Novo* Assembly and Phenotypic Profiling for the New Budding Yeast Species *Saccharomyces jurei*. *G3 GenesGenomesGenetics* **8**, 2967–2977 (2018).
47. Eberlein, C. *et al.* Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nat. Commun.* **10**, 923 (2019).
48. Gallone, B. *et al.* Interspecific hybridization facilitates niche adaptation in beer yeast. *Nat. Ecol. Evol.* **3**, 1562–1575 (2019).
49. Perez-Sepulveda, B. M. *et al.* An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biol.* **22**, 349 (2021).
50. Gascuel, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997).
51. Tattini, L. *et al.* Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes. *Mol. Biol. Evol.* **36**, 2861–2877 (2019).
52. Liti, G., Barton, D. B. H. & Louis, E. J. Sequence Diversity, Reproductive Isolation and Species Concepts in Saccharomyces. *Genetics* **174**, 839–850 (2006).
53. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**,

1817–1828 (2013).

54. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
55. Pfeifer, B. & Kapan, D. D. Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics* **20**, 207 (2019).
56. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* **5**, e11147 (2010).
57. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. *Proc. Natl. Acad. Sci.* **105**, 4957–4962 (2008).
58. Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature* **458**, 342–345 (2009).
59. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**, 289–291 (2016).
60. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 2444–2448 (1988).
61. Pan, J. *et al.* A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell* **144**, 719–731 (2011).
62. Mozzachiodi, S. *et al.* Aborting meiosis allows recombination in sterile diploid yeast hybrids. *Nat. Commun.* **12**, 6564 (2021).
63. Leducq, J.-B. *et al.* Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nat. Microbiol.* **1**, 15003 (2016).
64. Koufopanou, V. *et al.* Population Size, Sex and Purifying Selection: Comparative Genomics of Two Sister Taxa of the Wild Yeast Saccharomyces paradoxus. *Genome Biol. Evol.* **12**, 1636–1645 (2020).
65. Lee, S., Lim, W. A. & Thorn, K. S. Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PLoS ONE* **8**, e67902 (2013).