



A yeast living ancestor reveals the origin of genomic introgressions

Melania d'Angiolo, Matteo de Chiara, Jia-Xing Yue, Agurtzane Irizar, Simon Stenberg, Karl Persson, Agnès Llored, Benjamin Barré, Joseph Schacherer, Roberto Marangoni, et al.

► To cite this version:

Melania d'Angiolo, Matteo de Chiara, Jia-Xing Yue, Agurtzane Irizar, Simon Stenberg, et al.. A yeast living ancestor reveals the origin of genomic introgressions. *Nature*, 2020, 587 (7834), pp.420-425. 10.1038/s41586-020-2889-1 . hal-03430153

HAL Id: hal-03430153

<https://hal.science/hal-03430153>

Submitted on 29 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A yeast living ancestor reveals the origin of genomic introgressions

Melania D'Angiolo¹, Matteo De Chiara¹, Jia-Xing Yue¹, Agurtzane Irizar¹, Simon Stenberg^{2,3}, Karl Persson³, Agnès Llored¹, Benjamin Barré¹, Joseph Schacherer⁴, Roberto Marangoni^{5,6}, Eric Gilson^{1,7}, Jonas Warringer³ and Gianni Liti^{1,*}

Affiliations

¹Université Côte d'Azur, CNRS, INSERM, IRCAN, Nice, France; ²Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences (UMB), Ås, Norway; ³Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden; ⁴Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France; ⁵Department of Biology, University of Pisa, Italy; ⁶CNR, Institute of Biophysics, Pisa, Italy; ⁷Department of Genetics, CHU, Nice, France.

*Corresponding Author (gianni.liti@unice.fr)

Summary

Genome introgressions drive evolution across the animal¹, plant² and fungal³ kingdoms. Introgressions initiate from archaic admixtures followed by repeated backcrossing to one parental species. However, how introgressions arise in reproductively isolated species, such as yeasts⁴, has remained unclear. Here, we discovered a clonal descendant of the ancestral yeast hybrid that founded the extant *S. cerevisiae* Alpechin lineage⁵, which carries abundant *S. paradoxus* introgressions. We show that this clonal descendant, hereafter defined as “living ancestor”, retained the ancestral genome structure of the first-generation hybrid with contiguous *S. cerevisiae* and *S. paradoxus* subgenomes. The ancestral first-generation hybrid underwent catastrophic genomic instability through more than a hundred mitotic recombination events, mainly manifesting as homozygous genome blocks generated by loss-of-heterozygosity. These homozygous sequence blocks rescue hybrid fertility by restoring meiotic recombination and are the direct origins of the introgressions present in the Alpechin lineage. We suggest a plausible route for introgression evolution through reconstruction of extinct stages and propose genome instability to allow hybrids to overcome reproductive isolation and enable introgressions to emerge.

Introgressions of genomic material between populations and species often reflect key historical and demographical events of the species evolution. For example, non-African and Oceanian human populations contain ~2% Neanderthal and 2-4% of Denisovan DNA respectively as a consequence of archaic admixture with extinct hominid populations⁶. Introgression begins with an ancestral hybridization followed by successive backcrossings with one parental species, resulting in bulk loss of haplotypes from the other subgenome⁷⁻⁹. However, many biological species pairs may not form hybrids, or show nearly complete hybrid sterility⁴. In the model genus *Saccharomyces*, sister species are known to hybridize^{10,11}, but the high sequence divergence between the hybrids' subgenomes inhibits meiotic recombination via the mismatch repair system^{12,13}. The resulting viable gametes are rare and are almost always non-recombined, unfit and reproductively isolated from their parental backgrounds¹⁴. Nevertheless, *Saccharomyces* introgressions are frequent and characterised by replacement of one parental haplotype by the other^{5,15-18}. These small introgression blocks are scattered throughout the genome and imply very extensive historical recombination, which is hard to reconcile with reproductive isolation. Hybrids do form between more distantly-related yeast species^{19,20}, but do not result in introgressions. This is likely to be explained by the hybrids' higher sequence divergence, which reinforces reproductive isolation by abolishing recombination, and also by incompatibility between subsets of genes²¹.

We recently sequenced the genomes of more than 1000 *S. cerevisiae* strains and described 26 well-defined lineages representing specific ecological niches or geographical areas⁵. Four of these lineages, Alpechin, Brazilian bioethanol, Mexican agave and French Guiana, present abundant introgressions of highly diverged (~12%) sequence from the sister species *S. paradoxus*, derived from at least two ancient admixture events. The Alpechin lineage, associated with olive oil production, carries the largest amount of introgressed material, ranging from 4% to 5% in individual strains and combined covering 8% of the genome⁵. Alpechin strains were mostly isolated from olive oil wastewater (Alpechin, in Spanish) in Spain. The Alpechin population has been recently resampled from similar olive-based environments, indicating that the clade is stable and characteristic of olive-domestication environments²². Here, we report the discovery of a *S. cerevisiae*-*S. paradoxus* hybrid that is the direct clonal descendant of an ancient hybridization event that founded the Alpechin lineage by sexual reproduction and backcrossing to *S. cerevisiae*. For shorthand, we refer to this strain as the "living ancestor", as it retains the genome structure of the ancestor, while noting that the present-day individual is separated from the actual ancestor by some evolutionary time. The ancestral hybrid genome was shaped by extensive genomic instability, resulting in mitotic recombination events that by loss-of-heterozygosity (LOH) generated over a hundred scattered blocks of homozygous DNA. We show that these LOH blocks restore recombination efficiency and gamete viability, explaining how the ancestral hybrid overcame the main barrier to introgression.

A clonal descendant of the ancestral hybrid

We identified a single natural *S. cerevisiae*-*S. paradoxus* living ancestor hybrid that co-exists with the *S. cerevisiae* Alpechin clade in olive oil related niches (**Supplementary Table 1**). We combined long- and short-read deep sequencing and haplotype structure information from meiotic gametes to infer its genome structure. The diploid hybrid contains the complete set of 16 chromosomes from each parent species. First, we inferred the phylogenetic relationship of both the subgenomes of the living ancestor against all major *S. cerevisiae* and *S. paradoxus* subpopulations^{5,23,24}. We identified the Alpechin lineage as the closest relative to the living ancestor's *S. cerevisiae* subgenome and the European *S. paradoxus* population as the closest relative to the *S. paradoxus* subgenome (**Fig. 1a**). Second, we compared the genomic locations of the *S. paradoxus* LOH blocks (with two copies of *S. paradoxus* DNA retained) in the living ancestor to the *S. paradoxus* introgressed regions in the Alpechin isolates (**Fig. 1b**, **Extended Data Fig. 1**, **Supplementary Tables 2-3**). Among the 114 Alpechin introgressions, 62 overlapped with the *S. paradoxus* LOH blocks in the living ancestor for 534736 bp (spanning 43872 markers). Given that the living ancestor has 9.7% of its genome in LOH with two *S. paradoxus* copies (as well as 5.5% with two *S. cerevisiae* copies) and the Alpechins have 8% of introgressed *S. paradoxus* genome, the expected number of markers randomly overlapping between LOH and introgressions would be only 8221 (two-tailed χ^2 test, $p < 0.00001$). We further compared the breakpoints of these 62 LOH blocks and the introgressed regions and found them to show remarkable genome-wide concurrence. Among 124 introgression boundaries, 90 were within 1 kb from the closest LOH boundary, with the vast majority sharing the exact same coordinates. These results imply that the *S. paradoxus* LOH blocks were the direct source of most of the Alpechins' introgressions.

We then explored whether newly arisen heterozygous mutations within *S. paradoxus* LOH blocks, accumulating over time within the two initially identical DNA sequences, were also present in the Alpechins' introgressions. Identical *de novo* mutations are highly unlikely to arise independently and are therefore compelling evidence of shared ancestry. We confidently called 207 heterozygous sites within *S. paradoxus* LOH blocks and for each site inferred the mutated and the ancestral state. Of the 156 *de novo* mutations on the *S. cerevisiae* haplotype of the LOH blocks, 125 (80%) were present also on the *S. paradoxus* introgression blocks of the Alpechins (**Fig. 1b**, **Extended Data Fig. 2**). In contrast, only 3 of the 51 *de novo* mutations present in the *S. paradoxus* haplotype were detected in the Alpechin population. This is uniquely compatible with a model in which most of the retained introgressions are direct descendants of the *S. paradoxus* DNA copied into the *S. cerevisiae* subgenome during the LOH block formation in the ancestral hybrid (**Supplementary Discussion 1**). Introgressions non-overlapping with LOH blocks likely emerged by integration of additional *S. paradoxus* subgenome during meiotic and mitotic recombination after the split with the living ancestor.

Taken together, these results unequivocally show that the discovered hybrid is closely related to the

Alpechin lineage, is a good proxy for the ancestral hybrid state and can serve as a “living ancestor” to understand the early evolution of the Alpechin lineage.

Genome instability in the ancestral hybrid

We found the living ancestor’s subgenomes to have remained phased, with contiguous *S. cerevisiae* and *S. paradoxus* haplotypes, meaning that this clonal descendant never underwent a single meiosis after the initial, ancient hybridization (**Fig. 1b, Extended Data Fig. 3**). In contrast, the exceptional number of LOH blocks suggested a massive mitotic genome instability event. We detected 102 LOH blocks, together representing 15.2% of its genome, dispersed across the 16 chromosome pairs (**Supplementary Table 2**). The LOH blocks are disproportionately common in some chromosomes (e.g. chr. IV) and have complex patterns consistent with multiple nested recombination events in others (e.g. chr. V). The LOH size varies over 4-orders of magnitude (0.085 to 245 kb), with events shorter than 10 kb being enriched (**Extended Data Fig. 4a**). We found more interstitial ($n=92$) than terminal LOH blocks ($n=10$) and the former were much smaller (11 vs. 67 kb, average), consistent with these two LOH types originating from distinct repair mechanisms²⁵. We detected a strong LOH bias for *S. paradoxus* rather than *S. cerevisiae* homozygosity blocks (78 vs. 24, covering 1092514 and 611623 bp respectively) and investigated this by evolving eight parallel mutation accumulation lines in which we propagated the living ancestor asexually for ~2220 generations through 120 single cell bottlenecks^{26,27}. Together, the mutation accumulation lines acquired 37 new LOH with no bias for either parental haplotype (**Extended Data Fig. 4b, Supplementary Table 2, Supplementary Discussion 2**). This is consistent with the parental bias towards *S. paradoxus* LOH blocks in the living ancestor being due to selection favouring retention of these blocks²⁸⁻³⁰, rather than to mutational mechanisms promoting their formation.

We also found signatures of genome instability in five gross chromosomal rearrangements and in chromosome I aneuploidy. The rearrangements are mostly restricted to the *S. paradoxus* subgenome and likely postdate the hybridization (**Extended Data Fig. 3, Supplementary Discussion 3**). We conclude that extensive genome instability followed the hybridization, resulting in exceptional levels of LOH and rearrangements.

LOH blocks rescue hybrid fertility

To validate that the living ancestor can overcome the reproductive isolation barrier, we tested whether it produces viable meiotic offspring. Indeed, the gamete viability, 3.93% (**Supplementary Table 4**), was 5.9-fold higher than the 0.67% observed in experimental *S. cerevisiae*-*S. paradoxus* crosses¹⁵. This level of gamete viability is remarkable considering the presence of multiple chromosomal rearrangements that strongly exacerbate the reproductive isolation^{15,31}. We estimated a theoretical gamete viability of 13.97%, if the living ancestor did not have the chromosomal rearrangements (**Supplementary Discussion 3**). One attractive explanation is that LOH blocks

mediate crossing-over by providing dispersed segments of sequence identity, and thereby promote correct chromosome segregation³². The genome sequences of 25 viable living ancestor's gametes revealed that 76% of the crossovers (125 of 165; two-tailed χ^2 test, $p=1.56 \times 10^{-104}$) occurred within 1 kb of a LOH block, validating this hypothesis (**Fig. 2a, b, Extended Data Fig. 5a**).

We detected 133 recombinant chromosomes in the 25 gametes, with 106 single, 23 double, 3 triple and 1 quadruple recombination events (**Extended Data Fig. 5a**). This corresponds to 6.6 crossovers per viable gamete (0.4 per chromosome), but it is an underestimation since crossovers on chromosome I, V and IX were undetected due to aneuploidy and large LOH. Nevertheless, it markedly exceeds the 2.7 crossovers per viable gamete previously measured in *S. cerevisiae*-*S. paradoxus* experimental hybrids¹³. The combination of novel crossovers and pre-existing LOH produces mosaic chromosomes with *S. paradoxus* haplotypes that are comparable in size to the Alpechins' introgression blocks (**Extended Data Fig. 5b-d**).

The crossovers mediated by LOH blocks promoted correct chromosome segregation and strongly limited the emergence of aneuploidies (mean of 1.12 per spore). LOH content and aneuploidy rate were negatively correlated ($r=-0.31$) across the gametes. Chromosomes involved in heterozygous inter-homolog translocations segregated to produce balanced genomes in viable gametes, underscoring their deleterious effect (**Extended Data Fig. 3, 5a, Supplementary Discussion 3**).

We probed whether the LOH blocks in the living ancestor also guide subsequent mitotic recombination events and new LOH events. We revisited the mutation accumulation lines and found at least 9/37 new LOH primed by pre-existing ones (**Extended Data Fig. 6a, Supplementary Table 2**). Next, we investigated *de novo* LOH formation in cell populations expanding clonally for a few generations and in cells returned to clonal growth after an aborted meiosis³³ (**Fig. 2c, d**). We observed higher LOH formation rate in the living ancestor compared to an experimental *S. cerevisiae*-*S. paradoxus* first-generation hybrid with subgenomes resembling those of the ancestral hybrid before the genome instability. This is consistent with the pre-existing LOH blocks promoting recombination. Sequencing the genomes of 18 clones derived from the living ancestor, we found that new LOH events were almost always mediated by pre-existing LOH blocks (19/21 events, **Fig. 2c, Extended Data Fig. 6a, Supplementary Table 2**).

Finally, we investigated if these new LOH blocks further increased the fertility of the living ancestor. An additional large LOH block, emerging in cells having expanded clonally for a few generations, increased the gamete viability by 47% (to a median of 5.77%), in line with that LOH blocks help restore fertility in interspecies hybrids (**Extended Data Fig. 6b, c, Supplementary Tables 4, 5, 10, Supplementary Discussion 4**). We conclude that the massive instability of the ancestral hybrid genome led to exceptional levels of LOH and that these blocks induced and guided both meiotic and mitotic recombination, thereby helping it overcome the sterility barrier.

Recreating the history of introgressions

The unique availability of a living ancestor and its modern descendants allowed us to reconstruct and probe the key evolutionary stages leading to genome introgressions (**Fig. 3, Extended Data Fig. 7a, Supplementary Tables 4, 5**). First, we investigated the reproductive capacity of an experimental *S. cerevisiae*-*S. paradoxus* first-generation hybrid. As expected^{12,14,15}, this hybrid produced gametes, i.e. it sporulated efficiently, but these were almost always inviable (median of 0.5% viability).

Second, we investigated a first generation of gametes derived from the living ancestor, which restored a diploid state by mating-type switching and haplo-selfing³⁴. This F1 generation had variable sporulation efficiency and gamete viability, with the detrimental effects of chromosomal aneuploidies explaining much of this variation (**Extended Data Fig. 7b-d**).

Third, we explored the reproductive capacity of the next evolutionary stage by backcrossing eight haploid-stable F1 living ancestor gametes to each parental species. The living ancestor gametes were more fertile with *S. cerevisiae* than with *S. paradoxus* (median 22.12% vs 13.89%, two-tailed Mann-Whitney test, $p=0.008$). The gamete viability of the backcrosses varied, with negative effects of the chromosomal aneuploidies (**Extended Data Fig. 7e**). The higher fertility with *S. cerevisiae* may help explain why this backcross was favoured, but randomness, *S. cerevisiae* having been more common and its backcross being fitter, may have also played a role.

Fourth, we induced meiosis in four extant Alpechin strains and found gamete viability to be generally high (median 93.17%; one outlier at 25%). One strain (AQA), containing a relatively high amount of heterozygous introgressions (**Extended Data Fig. 8**), produced gametes that were highly viable (89%), but half of these reproduced slowly asexually. By sequencing fit and unfit offspring, we mapped this monogenic growth defect to a locus located far from the closest introgression. Thus, we found no signs of introgressions impairing meiosis, sporulation or germination in the Alpechin strains.

Finally, we explored whether the Alpechin strains remained fertile with both parental species. We crossed a haploid Alpechin derivative with *S. cerevisiae* and *S. paradoxus*. Gametes of the *S. cerevisiae* (91.03%) but not the *S. paradoxus* backcross (0.74%) were viable, showing that the Alpechins are reproductively isolated from the *S. paradoxus* ancestor, despite retaining 5% of its genome as introgressions, while they are fertile with a *S. cerevisiae* lineage lacking *S. paradoxus* introgressions.

Overall, our reconstructions show how the Alpechin clade and its introgressions could evolve from the ancestral hybrid and its LOH regions. Molecular dating of the divergence indicates that over a million asexual generations have occurred since the LOH began, while a maximum of ~400000 asexual generations have occurred since the living ancestor and the Alpechin lineages split (**Fig. 3, Supplementary Table 6**). We estimated a maximum chronological time of ~4000 years since their last shared common ancestor, meaning that they likely diversified after the emergence of olive oil production³⁵.

The living ancestor's asexual fitness

We next sought to understand why the living ancestor and the Alpechin descendants coexist in nature, while intermediate stages have not been isolated. We therefore measured their clonal reproduction in many environments permitting mitosis, and their survival after nutrients depletion. The living ancestor was often fitter than both *S. cerevisiae* and *S. paradoxus* parent species and the experimental F1 hybrid (in 40, 83 and 28% of environments respectively, **Fig. 4a, Extended Data Fig. 9a, Supplementary Table 7**). Gametes derived from the living ancestor suffered widespread (77/82 environments) and severe (33% average increase in doubling time) defects in asexual reproduction. Extensive survival defects amplified, rather than compensated for, these defects (**Extended Data Fig. 9b**). We propose that the fitness defects arise from recessive incompatibilities between genes³⁰, which are unmasked in the gametes' mosaic *S. cerevisiae* and *S. paradoxus* genomes (**Extended Data Fig. 5b**).

Mating the haploid gametes of the living ancestor back to the haploid *S. cerevisiae* or *S. paradoxus* parents restored fitness. We found only rare (3/82 environments in *S. cerevisiae* backcross, 10/82 in *S. paradoxus* backcross) and weak (1% average increase of doubling time for both backcrosses) asexual reproduction defects relative to the living ancestor (**Fig. 4a, Extended Data Fig. 9a, Supplementary Table 7**). This is consistent with the recessive incompatibilities being masked by the first round of backcrossing.

Across many environments, the asexual reproduction of the living ancestor closely resembled that of a sub-group of the Alpechins and only rarely and weakly deviated from that of the Alpechin lineage as a whole (worse in 6%, better in 6% of environments; **Fig. 4b, Extended Data Fig. 10a**). The Alpechins better resembled their closest extant *S. cerevisiae* parent than their *S. paradoxus* parent, but they enjoyed some niche-specific advantages relative to the former (in 11% of environments, **Extended Data Fig. 10b**), reflecting adaptations potentially conferred by *S. paradoxus* introgressions. Supporting that introgressed genes were retained due to selection, we found them to overlap non-randomly (83/492, two-tailed χ^2 test, $p < 0.00001$) with genes independently introgressed and retained in other *S. cerevisiae* subpopulations⁵ (**Fig. 4c, d, Extended Data Fig. 10c**). These overlapping genes represent at least one additional and separate hybridization event, involving the American rather than the European *S. paradoxus* population. Further supporting selection, Alpechin introgressions were enriched in genes mediating interactions with the environment (**Supplementary Tables 8, 9**).

Discussion

We discovered a yeast clonal descendant of an ancient hybridization event that enabled us to understand the early evolution of the *S. cerevisiae* Alpechin lineage and its abundant *S. paradoxus* introgressions. We showed how genome instability in the ancestral hybrid allowed it to by-pass the

hybrid sterility barrier and initiated the formation of introgressions. Specifically, genome instability promoted recombination, which in turn reduced aneuploidies and substantially rescued the gametes' viability. Most of the asci contained single viable spores, preventing intra-tetrad mating, while the low fitness of self-mated progeny disfavoured haplo-selfing (**Fig. 3**). These conditions directed the gametes toward backcrossing, which restored both asexual fitness and much of the hybrid fertility in a single life cycle step. Additional mechanisms such as whole genome duplication^{36,37} and intercrosses among gametes might have contributed to the late phase of the introgressions' evolution.

How can an ancient hybridization state compete with its modern descendants in the same ecological niche? Wild yeasts enter meiosis once per 1000 generations³⁸ but the living ancestor has survived >1000x as long, without sexual re-shuffling of its genome and without entering into a protective spore state, even though it is capable of both sex and sporulation. This life cycle reprogramming likely reflects relaxation of selection for sex and gamete formation due to living in a human-made environment³⁹. Despite its long asexual evolution, the living ancestor remains competitive with the sexually active descendants, matching the fitness of the Alpechins in most niches and occasionally surpassing them. This is consistent with divergent adaptation, with selection maintaining two distinct stages of the introgression process in different sub-niches.

The *S. paradoxus* genes that introgressed into the *S. cerevisiae* Alpechins are not a random collection of functions, but what we expect given adaptive benefits. Introgressed genes from Denisovans and Neanderthals retained in modern human populations have driven adaptation to high altitude⁴⁰ and virus defense⁴¹, although widespread depletion of Neanderthal introgressions in functional genomic regions likely reflects a general selection against Neanderthal variants⁶.

Introgressions can be key to reconstructing species histories. In humans, the restriction of Neanderthal introgressions to out-of-Africa populations and of Denisovan introgressions to Oceanian and Asian populations inform on their historical range, co-existence and contacts. Analogously, the scarcity of *S. paradoxus* introgressions into *S. cerevisiae* before the out-of-China expansion and diversification of the latter, and their abundance after, inform on yeast species range, co-existence and contacts⁵. Introgressions from an unknown hominin into Denisovans⁴² and from an unknown yeast into Taiwanese *S. cerevisiae*⁵ inform on extinct or undiscovered species. The recent discovery of a first-generation Neanderthal-Denisovan hybrid individual provided direct evidence of admixture⁴³, like our yeast living ancestor hybrid. Unlike an ancient genome, access to the yeast living ancestor from which introgressions derive, and the capacity to reconstruct other evolutionary stages along the path, enables direct probing of the species history. Isolation of additional intermediate hybridization-to-introgression stages could further illuminate the emergence, evolution and function of interspecies introgressions.

References

1. Edelman, N. B. *et al.* Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599 (2019).
2. Arnold, B. J. *et al.* Borrowed alleles and convergence in serpentine adaptation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8320–8325 (2016).
3. Sun, Y. *et al.* Large-scale introgression shapes the evolution of the mating-type chromosomes of the filamentous ascomycete *Neurospora tetrasperma*. *PLoS Genet.* **8**, e1002820 (2012).
4. Greig, D. Reproductive isolation in *Saccharomyces*. *Heredity* **102**, 39–44 (2009).
5. Peter, J. *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* **556**, 339–344 (2018).
6. Wolf, A. B. & Akey, J. M. Outstanding questions in the study of archaic hominin admixture. *PLoS Genet.* **14**, e1007349–14 (2018).
7. Harrison, R. G. & Larson, E. L. Hybridization, introgression, and the nature of species boundaries. *J. Hered.* **105 Suppl 1**, 795–809 (2014).
8. Mallet, J. Hybridization as an invasion of the genome. *Trends Ecol. Evol. (Amst.)* **20**, 229–237 (2005).
9. Martin, S. H. & Jiggins, C. D. Interpreting the genomic landscape of introgression. *Curr. Opin. Genet. Dev.* **47**, 69–74 (2017).
10. Morales, L. & Dujon, B. Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiol. Mol. Biol. Rev.* **76**, 721–739 (2012).
11. Marsit, S. *et al.* Evolutionary biology through the lens of budding yeast comparative genomics. *Nat. Rev. Genet.* **18**, 581–598 (2017).
12. Hunter, N., Chambers, S. R., Louis, E. J. & Borts, R. H. The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *EMBO J.* **15**, 1726–1733 (1996).
13. Kao, K. C., Schwartz, K. & Sherlock, G. A genome-wide analysis reveals no nuclear dobzhansky-muller pairs of determinants of speciation between *S. cerevisiae* and *S. paradoxus*, but suggests more complex incompatibilities. *PLoS Genet.* **6**, e1001038 (2010).
14. Greig, D., Louis, E. J., Borts, R. H. & Travisano, M. Hybrid speciation in experimental populations of yeast. *Science* **298**, 1773–1775 (2002).
15. Liti, G., Barton, D. B. H. & Louis, E. J. Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics* **174**, 839–850 (2006).
16. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
17. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*. *Nat Comms* **5**, 4044 (2014).
18. Barbosa, R. *et al.* Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. *Genome Biol Evol* **8**, 317–329 (2016).
19. Gallone, B. *et al.* Interspecific hybridization facilitates niche adaptation in beer yeast. *Nature Ecology & Evolution* **3**, 1562–1575 (2019).
20. Langdon, Q. K. *et al.* Fermentation innovation through complex hybridization of wild and domesticated yeasts. *Nature Ecology & Evolution* **3**, 1576–1586 (2019).
21. Lee, H.-Y. *et al.* Incompatibility of nuclear and mitochondrial genomes causes hybrid sterility between two yeast species. *Cell* **135**, 1065–1073 (2008).
22. Pontes, A., Čadež, N., Gonçalves, P. & Sampaio, J. P. A Quasi-Domesticated Relic Hybrid Population of *Saccharomyces cerevisiae* × *S. paradoxus* Adapted to Olive Brine. *Front Genet* **10**, 449 (2019).
23. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31**, 872–888 (2014).
24. Yue, J.-X. *et al.* Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).
25. Symington, L. S., Rothstein, R. & Lisby, M. Mechanisms and regulation of mitotic recombination in *Saccharomyces cerevisiae*. *Genetics* **198**, 795–835 (2014).

26. Dutta, A. *et al.* Genome Dynamics of Hybrid *Saccharomyces cerevisiae* During Vegetative and Meiotic Divisions. *G3 (Bethesda)* **7**, 3669–3679 (2017).
27. Tattini, L. *et al.* Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes. *Mol. Biol. Evol.* **36**, 2861–2877 (2019).
28. Dunn, B. *et al.* Recurrent rearrangement during adaptive evolution in an interspecific yeast hybrid suggests a model for rapid introgression. *PLoS Genet.* **9**, e1003366 (2013).
29. Smukowski Heil, C. S. *et al.* Loss of Heterozygosity Drives Adaptation in Hybrid Yeast. *Mol. Biol. Evol.* **34**, 1596–1612 (2017).
30. Bozdag, G. O. *et al.* Engineering recombination between diverged yeast species reveals genetic incompatibilities. *bioRxiv* **5**, e15380–31 (2020).
31. Hou, J., Friedrich, A., de Montigny, J. & Schacherer, J. Chromosomal Rearrangements as a Major Mechanism in the Onset of Reproductive Isolation in *Saccharomyces cerevisiae*. *Current Biology* **24**, 1153–1159 (2014).
32. Rogers, D. W., McConnell, E., Ono, J. & Greig, D. Spore-autonomous fluorescent protein expression identifies meiotic chromosome mis-segregation as the principal cause of hybrid sterility in yeast. *PLoS Biol.* **16**, e2005066 (2018).
33. Laureau, R. *et al.* Extensive Recombination of a Yeast Diploid Hybrid through Meiotic Reversion. *PLoS Genet.* **12**, e1005781 (2016).
34. McClure, A. W., Jacobs, K. C., Zyla, T. R. & Lew, D. J. Mating in wild yeast: delayed interest in sex after spore germination. *Mol. Biol. Cell* **29**, mbcE18080528–3127 (2018).
35. Kaniewski, D. *et al.* Primary domestication and early uses of the emblematic olive tree: palaeobotanical, historical and molecular evidence from the Middle East. *Biol Rev Camb Philos Soc* **87**, 885–899 (2012).
36. Greig, D., Borts, R. H., Louis, E. J. & Travisano, M. Epistasis and hybrid sterility in *Saccharomyces*. *Proceedings of the Royal Society B: Biological Sciences* **269**, 1167–1171 (2002).
37. Charron, G., Marsit, S., Hénault, M., Martin, H. & Landry, C. R. Spontaneous whole-genome duplication restores fertility in interspecific hybrids. *Nat Comms* **10**, 4126–10 (2019).
38. 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. U.S.A.* **105**, 4957–4962 (2008).
39. Magwene, P. M. *et al.* Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1987–1992 (2011).
40. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
41. Enard, D. & Petrov, D. A. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell* **175**, 360–371.e13 (2018).
42. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 1–12 (2013).
43. Slon, V. *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* **531**, 504–116 (2018).

Figure Legends

Fig. 1 | The ancestor of the Alpechin lineage. **a**, Phylogenetic tree generated using a matrix of 2258 1-to-1 orthologous genes across 46 *Saccharomyces* genomes. We name and number the *S. cerevisiae* lineages as in the 1002 Yeast Genome Project, and the *S. paradoxus* lineages as in the Yeast Population Reference Panel (left boxes). Lineage codes are reported in the phylogenetic tree,

followed by strain name. The living ancestor *S. cerevisiae* subgenome (Living ancestor Sc) tightly clusters with the Alpechin lineage, while the *S. paradoxus* subgenome (Living ancestor Sp) has European ancestry. **b**, Circular plot of the living ancestor's (inner ring) and a pan-introgression Alpechin's (outer ring) genomes. Internal ribbons link regions involved in gross chromosomal rearrangements (grey intra-homolog and white inter-homolog) in the living ancestor, which were reverted to produce colinear maps. Genomic coordinates of *S. cerevisiae* and *S. paradoxus* chromosomes are based on the genome assemblies of the DBVPG6765 and CBS432 strains, respectively. The upper part represents a zoom-in of chromosome IV. Numbers refer to genome-wide overlap between *S. paradoxus* LOH and introgressions. The bottom part represents a zoom-in of the living ancestor's chromosomes XI with analysis of 5 heterozygous sites (circles) within a *S. paradoxus* LOH block. Yellow and grey circles respectively represent *de novo* mutations and ancestral alleles. Genotypes of the Alpechins (bottom box) indicate inheritance of *de novo* mutations from the *S. cerevisiae* haplotype. In contrast, *de novo* mutations are absent from European and Far East Asian *S. paradoxus*, consistent with their appearance post-dating hybridization and LOH formation.

Fig. 2 | LOH blocks guide recombination of the hybrid's subgenomes. **a**, Twenty-five haploid meiotic gametes were isolated from the living ancestor and whole-genome sequenced. Chromosome II meiotic recombination patterns in haploid gametes indicate that meiotic crossover sites (arrows) are highly enriched within LOH blocks (red lines). The number of spores that share a similar chromosome II recombination profile is reported. **b**, The bar plot represents the total number of recombination sites inside/outside LOH blocks genome-wide (two-tailed χ^2 test, $p=1.56 \times 10^{-104}$), ****: $p < 0.0001$. **c**, The living ancestor chromosome II was engineered with a heterozygous *URA3* gene at the native *LYS2* locus. This system enables us to select LOH events at the *LYS2* locus by plating diploid clones in 5-Fluoroorotic Acid (5-FOA) media, which selects for *URA3* loss. Six mitotic (T0) and twelve return-to-growth (RTG) clones were isolated and genome-sequenced. Chromosome II analysis of the diploid clones shows that most of the recombination events are mediated by pre-existing LOH homozygous blocks (red lines). The *URA3* gene is embedded in the living ancestor's *S. cerevisiae* chromosome II and therefore selects for events that generate a *S. paradoxus* homozygous region spanning the *LYS2* locus. **d**, LOH formation rate in an experimental *S. cerevisiae*-*S. paradoxus* F1 hybrid and the living ancestor using a *URA3* loss fluctuation assay, with heterozygous *URA3* engineered as illustrated in panel **c** for both strain backgrounds. The living ancestor has a higher rate both in mitotic growth (T0) and in returned-to-growth (RTG) experiments (two-tailed Mann-Whitney test, $p=0.008$ and $p=8.66 \times 10^{-05}$, respectively; $n=5$ biologically independent samples for each genetic background and condition). Boxes: horizontal line: median, upper/lower hinge: interquartile range (IQR), whiskers: max/min values. **: $p < 0.01$, ****: $p < 0.0001$

Fig. 3 | Reconstructing the origin of introgressions. Reconstruction of missing stages of the most likely route of the hybridization-to-introgression process. Plain lines point to naturally occurring stages (parental species, living ancestor and Alpechins), while dotted lines point to stages reproduced in the lab (experimental *S. cerevisiae*-*S. paradoxus* F1 hybrid, living ancestor gametes and backcrosses). Red and blue bars represent *S. paradoxus* and *S. cerevisiae* ancestry, respectively. The timespan inferred by molecular dating is represented by dotted grey lines and is expressed as number of asexual generations. White and grey circles respectively depict viable and dead meiotic gametes held as a tetrad within an ascus. Their low viability prevents intra-tetrad mating while their low fitness after germination disfavours haplo-selfing. The combination of these events promotes the backcrossing process leading to the extant Alpechin lineage.

Fig. 4 | The asexual fitness landscape. **a**, Relative doubling time for each hybridization-to-introgression stage resulting from the samples in each group (indicated in brackets) and their 8 replicates (40 for the living ancestor) across 82 environments. *S. cerevisiae* (Sc, $n=656$), *S. paradoxus* (Sp, $n=1312$), experimental first-generation hybrid (F1 hybrid, $n=1312$), living ancestor (LA, $n=3280$), living ancestor gametes (LA gametes, $n=16400$), backcross to *S. cerevisiae* (Bc Sc, $n=5248$), backcross to *S. paradoxus* (Bc Sp, $n=5248$), Alpechin ($n=10496$). Boxes: horizontal line: median, upper/lower hinge: IQR, whiskers: largest/smallest value within upper/lower hinge $\pm 1.5 \times$ IQR. **b**, Hierarchical clustering (Pearson's r , complete linkage) of the living ancestor (LA), the Alpechin isolates (three-letter names) and the closest living relatives of their *S. cerevisiae* (Sc) and *S. paradoxus* (Sp) parents. Strains were grouped based on doubling time similarity (**Extended Data Fig. 10a**) and tree branches colour indicates the four distinct phenotypic clusters. **c-d**, Overlap of introgressed genes (**c**) and regions (**d**) in four *S. cerevisiae* lineages with abundant *S. paradoxus* genetic material (red segments).

Methods

Natural and experimental yeast strains

The strains used in this work are listed in supplementary table 1. The living ancestor, CBS7002, was isolated from olive oil wastewater (Alpechin, in Spanish) by J. Santa Maria in the 1970s and obtained from the CBS yeast culture collection (<http://www.wi.knaw.nl/Collections/>). This strain was initially proposed as a type strain of the new species *S. hispalensis*, which is now regarded as synonymous to *S. cerevisiae*. The Alpechins were also mostly isolated from the same substrate and the same geographic location as the living ancestor. However, two Alpechins were isolated from Californian olive fermentations inoculated with European olives, one was isolated from the olive fruit fly and four were isolated from humans (as fungal infections or from faeces). Genome sequences of the Alpechins used in this study were previously reported⁵, whereas new samples associated with this work were sequenced with Illumina paired-end technology at the NGS platform of Institut Curie

according to the manufacturer's standard protocols.

We reconstructed the missing evolutionary stages linking the living ancestor and the Alpechins and leading to genome introgressions. First, we generated homozygous diploid *S. cerevisiae* (DBVPG6765, Wine/European subpopulation) and *S. paradoxus* (N17, European subpopulation) strains. Second, we mated a haploid *S. cerevisiae* DBVPG6765 and a haploid *S. paradoxus* N17 to generate an experimental *S. cerevisiae*-*S. paradoxus* F1 hybrid with subgenomes closely resembling those of the ancestral hybrid before the genome shock. Next, we induced meiosis in the living ancestor by a five-days incubation in potassium acetate (KAc) medium⁴⁴ and isolated 25 viable meiotic spores by tetrad dissection. The living ancestor is homothallic and therefore able to switch mating type during vegetative growth. 18 of these spores were auto-diploidized by mating-type switching and haplo-selfing to form homozygotic diploids, while 7 spores remained haploid due to the presence of chromosome III aneuploidies. The ploidy was further confirmed by flow cytometry analysis, as previously described⁵. Chromosome III contains the *MAT* locus, which controls the mating type. When both *MAT* alleles are present due to aneuploidies, spores are unable to switch mating type and haplo-selfing is inhibited⁴⁵.

We deleted the *HO* mating-type switch gene in the living ancestor and generated eight additional haploid-stable gametes (*ho::NAT*, either *ura3::KanMX* or *ura3::HygMX*). These spores were backcrossed to *S. paradoxus* N17 and to *S. cerevisiae* DBVPG6765 respectively to produce ($n=16$) diploid hybrids that were either ~75% *S. paradoxus* or ~75% *S. cerevisiae*.

We deleted the *HO* gene also in one diploid Alpechin (CPG) and generated a haploid-stable gamete (*MATa*, *ho::HygMX*, *ura3::KanMX*). This gamete was crossed with *S. paradoxus* N17, *S. cerevisiae* DBVPG6765 and with three haploid-stable living ancestor gametes.

Mutation accumulation lines

We evolved the living ancestor as 8 parallel mitotic mutation accumulation lines on YPD solid medium and passed them through a single-cell bottleneck every ~48 hours (~18.5 generations) at 30°C, for a total of 120 bottlenecks (~2220 generations). At each bottleneck, a random colony, having expanded from a single cell, was re-streaked to isolate single colony-forming units. To avoid selection bias, we picked the closest colony to the centre of each cultivation plate at each streak, regardless of its size. The ploidy of the evolved lines was measured by flow cytometry, as previously described⁵.

Rates of *de novo* LOH formation and return to growth (RTG) experiments

Rates of *de novo* LOH formation were measured for the living ancestor and the experimental *S. cerevisiae*-*S. paradoxus* first generation hybrid using a fluctuation assay, as previously described⁴⁶. In diploid strains, both *URA3* gene copies were deleted from their native locus on chromosome V and one *URA3* copy was inserted back into the *LYS2* locus on chromosome II, thereby replacing

one of the *LYS2* copies. Relative rates of *de novo* LOH formation were measured on 5-FOA plates as loss of this single *URA3* gene copy, reflecting the replacement of the chromosome region carrying the *URA3* allele with the homologous region from the other chromosome II copy, which carried the *LYS2* allele. These rates were measured in populations of both mitotically-growing cells (18 hours in YPD) as well as cells in which meiosis had been initiated (6 hours in KAc) and then aborted by returning the cells to mitotic growth (18 hours in YPD), as previously described²⁷. We performed 5 biological replicates (1 technical replicate for each) for each strain background and condition. We sequenced the genome of 6 mitotically growing clones and 12 return-to-growth clones derived from the living ancestor cells that had been selected to have lost the *URA3* allele. We confirmed the presence of a large *de novo* LOH on chromosome II in all analysed samples and checked the presence of additional *de novo* LOH events, as described in the paragraph “Mapping LOH and introgressions' boundaries”.

Estimation of sporulation efficiency and spore viability

We measured sporulation efficiency and spore viability as previously described⁴⁷, for all the stages of the hybridization-to-introgression process as well as for the genome-sequenced clones of the living ancestor derived by asexual growth (6 clones), return-to-growth (12 clones) and mutation accumulation protocol (8 clones) (**Supplementary Tables 4 and 5**). Sporulation was induced by incubation in liquid KAc medium and its efficiency was measured by counting ~200 cells at days 1 and 3 for the *S. cerevisiae* and *S. paradoxus* parents, the experimental F1 hybrid, the backcrosses and the crosses of the Alpechin strain with *S. cerevisiae* and *S. paradoxus*. All the other samples were measured at multiple time points ($t=0, 1, 2, 3, 4, 9$ days). Spores were then isolated by tetrad dissection and their viability was scored as the percentage of spores forming visible colonies after six days of growth on YPD at 30°C.

Bulk segregant analysis of Alpechin AQA growth defect

We induced sporulation of the Alpechin strain AQA (OS1357) in liquid KAc medium. Spores were isolated using a dissecting microscope and incubated in YPD medium for four days at 30°C. We pooled spores based on their colony size, forming ten pools. Six pools contained only big colonies while four contained only small colonies. Each pool was composed of 12 spores, except one (A504R10) that contained 11 spores. DNA was extracted from each pool and sequenced with Illumina technology (mean coverage of 55X). The markers used in the bulk segregant analysis were extracted from the Alpechin strain AQA by mapping onto the *S. cerevisiae* DBVPG6765 assembly, using BWA (v0.7.12). Post-processing steps, including sorting and duplicates removal were performed by SAMtools (v1.2) and picard tools (v2.8.0). Variant calling was performed by freebayes (v0.9.5). We extracted the markers assuming that the 2:2 segregation pattern in the spores was caused by a heterozygous SNV for which one of the two alleles was deleterious and recessive,

slowing the growth of haploid spores but not that of the Alpechin diploid parent. Therefore, we extracted only the SNVs whose allelic fraction (AF) was equal to 0.5 in the Alpechin AQA parent. Afterwards, we mapped Illumina reads of the pools on the *S. cerevisiae* DBVPG6765 assembly and performed the subsequent steps as described above. Then, we estimated the AF of the heterozygous markers in all the pools and excluded the ones which were not compatible with the following assumptions: i) All the pools must have a homozygous genotype for the marker. ii) Big and small pools must have a different allele for the marker. iii) The deleterious allele associated with the small phenotype must be the one with the lowest frequency in the *S. cerevisiae* population. Only one marker was compatible with these criteria and was therefore deemed to explain the variation in the spore's growth.

Estimating population size doubling time during clonal reproduction

We measured population size doubling time during clonal reproduction for all the hybridization-to-introgression stages. One Alpechin strain (BRK) was excluded from the analysis because of contamination. All yeast populations were stored at -80°C in 20% glycerol and were cultivated at 30°C in temperature and humidity-controlled cabinets. Yeast strains were revived from frozen 96-well stocks by robotic transfer (Singer RoToR; long pins) of a random sub-sample of each thawed population (~50.000 cells) to a Singer PlusPlate in 1536 array format on solid Synthetic Complete (SDC) medium composed of 0.14% Yeast Nitrogen Base (CYN2210, ForMedium), 0.50% $(\text{NH}_4)_2\text{SO}_4$, 0.077% Complete Supplement Mixture (CSM; DCS0019, ForMedium), 2.0% (w/w) glucose, pH buffered to 5.80 with 1.0% (w/v) succinic acid and 0.6% (w/v) NaOH. In every fourth position, fixed spatial controls (genotype: YPS128, *MATa/MATa*) were introduced to account for spatial variation across plates in the subsequent experimental stage. Controls were similarly subsampled from a separate 96-well plate and introduced to the pre-culture array (Singer RoTor; long pins). Populations were pre-cultivated for 72 hours at 30°C . For pre-cultures of nitrogen-limited environments, the background medium was modified to avoid nitrogen storing and later growth on stored nitrogen: CSM was replaced by 20 mg/L uracil (not converted into usable nitrogen metabolites) and $(\text{NH}_4)_2\text{SO}_4$ was reduced to growth-limiting concentrations (30 mg N/L). We cast all solid plates 24 hours prior to use, on a levelled surface, by adding 50 mL of medium in the same upper right corner of the plate. We removed excess liquid by drying plates in a laminar air-flow in a sterile environment. Pre-cultured populations were mitotically expanded until stationary phase (2 million cells; 72 h), were again subsampled (~50.000 cells; short pins) and transferred to experimental plates, containing the medium of interest (**Supplementary Table 7**). Synthetic grape must (SGM) was prepared as previously described⁴⁸. We tracked population size expansion using the Scan-o-matic system (v1.5.7, <https://github.com/Scan-o-Matic/scanomatic>). Plates were maintained undisturbed and without lids for the duration of the experiment (72 h) in high-quality desktop scanners (Epson Perfection V800 PHOTO scanners, Epson Corporation, UK) standing

inside dark, humid and thermostatic cabinets with intense air circulation. Images were analyzed and phenotypes were extracted and normalized against the fourth position controls using Scan-o-Matic⁴⁹. We extracted the normalized, relative population size doubling time for each experiment, D_r , by subtracting the control value for that position, $D_r = \log_2(D) - \log_2(D_{control,local})$. D_r is reported as the relative doubling time. We also report absolute doubling times, while maintaining the normalization, as $D_{normabs} = 2^{D_{norm}} \text{Mean}(D_{control,local})$. We performed 8 biological replicates per sample, except for the living ancestor for which we performed 40 biological replicates.

Estimation of chronological life span

Strains were grown overnight in liquid SDC medium, diluted 100x in 200 μ L of fresh SDC in a 96-well plate and incubated at 30 °C for 25 days. We started taking samples after the net growth in each cell population had ceased, 72 hours after incubation. At each time point, 5 μ L of cells were transferred in 100 μ L of staining solution (Phosphate-buffered saline + 3 μ M propidium iodide + 200 nM YO-PRO-1) in a 96-well plate and incubated 10 minutes in the dark at 30 °C. Cell viability was measured by high-throughput flow cytometry on a FACS-Calibur using the HTS module, as previously described⁵⁰. Cell debris and heavily damaged old necrotic cells were first removed based on forward scatter and side scatter at day 0. Then, cells were excited with the 488 nm laser and fluorescence was read with FL-1 and FL-3 filters, corresponding to YO-PRO-1 and propidium iodide fluorescence, respectively. Non-fluorescent cells were considered viable whereas fluorescent ones were considered dead. The final viability was calculated as the product of "Intact cells" and "Viable cells" (**Extended Data Fig. 9c**). We measured each sample once. Flow cytometry data were subsequently analysed using FlowJo (v10.3).

Sequencing and *de novo* assembly of the living ancestor's genome

The complex genome structure of the living ancestor makes full assembly challenging. Specifically, large blocks of LOH with near-identical sequence are interspersed in the highly diverged parental subgenomes and they are often merged by the assembler. We combined multiple sequencing and experimental approaches to overcome this challenge.

First, we deep-sequenced the living ancestor with both short (~200X, Illumina) and long (~146X, PacBio) read technology. DNA was extracted and sequenced for 2 x 100 (paired-end) cycles using the Illumina HiSeq2000. *De novo* assemblies from short reads were obtained using ABySS (v1.9.0)⁵¹. DNA was extracted and sequenced by PacBio technology at the Lausanne genomic technologies facility (Lausanne, Switzerland) and raw reads were processed by standard SMRT Analysis pipeline (v2.3.0) to obtain filtered subreads. The assembly was processed with the LRSDAY (v1.0.0) pipeline⁵². We performed *de novo* assembly using Canu (v1.5)⁵³ by setting the overall (diploid) genome size to 24 Mb. We then aligned the raw PacBio reads to our Canu assembly by pbalgn (distributed with GenomicConsensus v2.0.0) and polished the assembly by

Quiver (distributed with GenomicConsensus v2.0.0)⁵⁴, correcting sequencing errors. We further mapped the Illumina reads to the polished genome assembly for another round of error-correction, using Pilon (v1.22)⁵⁵. Scaffolds were manually visualised and inspected as dot-plots using the program MUMMERplot included in the suite MUMmer (v3.1)⁵⁶.

We additionally used the genome sequence of 25 living ancestor gametes to phase the living ancestor's subgenomes and to determine its karyotype structure. We analysed the genomes of the living ancestor's gamete and used the coverage to infer their segregation and recombination patterns. Most (65%, excluding aneuploidies, chromosome VII and chromosome XV) of the chromosomes had non-recombinant and uniform coverage profiles, that reflected the living ancestor's genome configuration. Several structural variants were evident using both the PacBio assembly and the segregation profiles and these were validated through manual inspection of the corresponding scaffolds.

Phylogenetic analysis of the living ancestor's subgenomes

We performed *de novo* genome assembly based on deep (~200X) Illumina sequencing data for the living ancestor and 32 representative strains sampled from the 1002 Yeast Genome Project⁵. The genome assembly was performed by ABySS (v1.9.0)⁵¹ with the option '-k 64'. All assembled scaffolds shorter than 1 kb were removed. The 33 assemblies were subsequently annotated using LRSDAY (v1.0.0)^{24,52}. For the living ancestor, we further partitioned its annotated genes into two subsets: the *S. cerevisiae*-like genes and the *S. paradoxus*-like genes, based on the tblastx (v2.2.31+) search against our curated yeast proteome database containing genes from 7 *S. cerevisiae* and 5 *S. paradoxus* representative strains²⁴. Each gene was classified as *S. cerevisiae*-like or *S. paradoxus*-like if its top hit (cutoff: sequence identity $\geq 90\%$ and match coverage $\geq 70\%$) matched with *S. cerevisiae* or *S. paradoxus*. We identified a total of 5489 *S. cerevisiae*-like and 5699 *S. paradoxus*-like genes and treated these two sets separately as two independent subgenomes. We refer to these as "Living ancestor Sc" and "Living ancestor Sp" respectively in our downstream gene orthology and phylogenetic analysis. We employed proteinortho (v5.15)⁵⁷ to identify orthologous gene groups shared among Living ancestor Sc and Living ancestor Sp, the other 32 representative *S. cerevisiae* strains and the 18 strains (7 *S. cerevisiae* + 5 *S. paradoxus* + 6 outgroup species from the *Saccharomyces* genus) that we analysed before²⁴. Based on this analysis, 2258 one-to-one orthologous groups shared among all these 52 strains (including Living ancestor Sc and Living ancestor Sp) were identified and used in the downstream phylogenetic analysis. The phylogenetic analysis was performed as previously described²⁴. The six *Saccharomyces* outgroup species are not displayed in the phylogenetic tree showed in Fig. 1a.

Mapping LOH and introgressions' boundaries

We annotated LOH and introgression regions based on the allelic fraction at a defined set of single

nucleotide markers derived from the alignment of DBVPG6765 (Wine/European *S. cerevisiae*) and CBS432 (European *S. paradoxus*) genome assemblies. We generated a list of 1059399 reliable markers between the two assemblies using the software nucmer included in the suite MUMmer (v3.1)⁵⁶, with the option “-C” to exclude markers with ambiguous position assignments. These markers reside within the core chromosome regions where near-perfect synteny is observed between the *S. cerevisiae* and *S. paradoxus* genomes. Because synteny is lost within subtelomeres²⁴, these regions were excluded from our analysis. The over 1M markers distributed over 11210867 bp of aligned genome gives a very dense map of polymorphic markers at an average distance of ~10 bp. We mapped Illumina reads on the *S. cerevisiae* DBVPG6765 assembly using BWA (v0.7.12). Read mapping runs were stored in the Sequence Alignment/Map (SAM) format and its binary form (BAM) and post-processing steps were performed by picard tools (v2.8.0). Sorting, indexing and per-position coverage calculation were performed using SAMtools (v1.2).

We used the software freebayes (v0.9.5) to call markers from the BAM files. Results were stored in Variant Call Format (VCF) files. We examined the fraction of the two possible alleles at every marker and used in-house Perl scripts to assign a genotype with the following distinguishing criteria: i) If the allelic fraction is equal to 0.5 the marker is in heterozygosity, ii) If the allelic fraction equals 0 or 1 the marker is in homozygosity for the *S. cerevisiae* or the *S. paradoxus* subgenome, respectively. Markers with a coverage <20 were considered unassigned and ignored in the subsequent analyses. For both the Alpechins (except YAE and YDM) and the original non-laboratory evolved living ancestor, we increased this threshold to <50 given their higher genome coverage.

In the case of the living ancestor and its derived clones, we annotated regions containing at least 10 consecutive markers in homozygosity (interspaced by less than 10 markers with a different state) as LOH. In the case of the Alpechins, we used the same parameters and annotated regions of homozygosity for the *S. paradoxus* allele as homozygous introgressions, and regions of heterozygosity as heterozygous introgressions. As additional evidence, we tested the reliability of LOH and introgression predictions by manually inspecting recombinant reads spanning their boundaries using IGV (v2.3.68). LOH and introgressions for which we could not find recombinant reads were excluded.

The same process was used to map introgression boundaries in strains belonging to three clades previously described⁵: Brazilian bioethanol, Mexican agave and French Guiana. For each of the introgressed clades (including Alpechin), we separately summed up the introgressions in their isolates in order to build a pan-introgression reference collection. In case multiple isolates within the same clade had different boundaries for an overlapping introgression, the largest event was kept.

Overall, we genotyped 104435 markers in LOH in the living ancestor (63375 in *S. paradoxus* LOH and 41060 in *S. cerevisiae* LOH). These markers were distributed along 1704137 bp (1092514 for

S. paradoxus LOH and 611623 for *S. cerevisiae* LOH), giving a total of 15.2% of genome in LOH (9.7% for *S. paradoxus* and 5.5% for *S. cerevisiae*). We genotyped 77955 introgressed markers in the Alpechin clade, distributed along 902805 bp (with a range of 467265 – 583627 bp in individual strains), giving a total of 8% of introgressed genome (ranging from 4% to 5% in individual strains). Subtelomeres were excluded from the genome size estimations.

We intersected the genomic coordinates of *S. paradoxus* LOH in the living ancestor with the genomic coordinates of introgressions in the Alpechins, using the program intersectBed included in the suite bedtools (v2.17). We retrieved the genomic positions in which *S. paradoxus* LOH and introgressions overlapped, for a total of 534736 overlapping bp (containing 43872 markers). We calculated the expected random number of overlapping markers (8221) as the total number of markers (1059399) multiplied by the percentage of genome in *S. paradoxus* LOH in the living ancestor (9.7%) and the percentage of introgressed genome in the Alpechin clade (8%).

Variant calling and SNVs analysis

Illumina reads were mapped to a concatenated reference assembly, which included both the *S. cerevisiae* DBVPG6765 and the *S. paradoxus* CBS432 assemblies²⁴. Competitive mapping against the two highly diverged species assemblies was performed by BWA (v0.7.12). Read mapping runs were stored in the Sequence Alignment/Map (SAM) format and its binary form (BAM). Sorting, indexing and per-position coverage calculation were performed using SAMtools (v1.2). Post-processing steps were performed by picard tools (v2.8.0). Single nucleotide variants (SNVs) were called from read mappings using freebayes (v0.9.5). We filtered SNVs to keep only the ones within LOH regions with two copies of the *S. paradoxus* subgenome. Subsequent filtering was performed based on the assumption that *de novo* mutations occurring after the LOH appearance should be in heterozygous state in the living ancestor's genome (AF=0.5). The allele present on the *S. paradoxus* reference genome was annotated as ancestral, whereas the alternative one was annotated as a *de novo* mutation. We confirmed that all the predicted *de novo* mutations were absent from the European and Far East Asian *S. paradoxus* populations (MAF=0)²³. We verified the presence of these mutations in 25 gametes derived from the living ancestor and retained only those segregating. We assigned the alleles to the two possible haplotypes (hosted on the *S. cerevisiae* or the *S. paradoxus* subgenome of the living ancestor) according to the species that contributed to the gametes' genomes for those specific flanking genomic positions. A total of 207 sites passed all these criteria. This set was used to investigate the relationship between the living ancestor and the Alpechin strains. We annotated the alleles present at these positions in the Alpechins' introgressions and determined the number of *de novo* mutations retained and the haplotype that was most likely inherited. We applied the same approach to the heterozygous SNVs within *S. cerevisiae* LOH regions and used the results to determine which model best explains the living ancestor-Alpechin origin.

Sequencing and genomic analysis of the living ancestor gametes

The living ancestor was sporulated in KAc medium⁴⁴. 25 spores were isolated using a dissecting microscope, incubated in YPD medium at 30°C and scored for viability after six days (**Supplementary Table 4**). We sequenced the living ancestor gametes with Illumina technology (mean coverage of 100X). Mapping and *de novo* assembly of their genomes were performed as described in the paragraph “Variant calling and SNVs analysis”. The sequencing depth of coverage was extracted using SAMtools (v1.2). Coverage analyses and statistical tests were performed using R (v3.5.3). Chromosomes with values close to 100X mean coverage for both species (200X total) across their entire length were considered as aneuploidies. Recombination sites were manually mapped as described in the paragraph “Mapping LOH and introgressions’ boundaries”.

Estimation of the living ancestor's base substitution rate and phenotypic impact

Read mapping and variant calling of the living ancestor before propagation (generation zero, G_0) and the mutation accumulation lines were performed as described in the paragraph “Variant calling and SNVs analysis”. We filtered the VCF files to keep only SNVs with quality higher than 100 and coverage higher than 20, and we excluded SNVs in subtelomeric regions²⁴.

We subtracted the SNVs present in the living ancestor G_0 from its derived mutation accumulation lines using the *vcf-isec* program included in the suite VCFtools (v0.1.16). Next, we searched for shared SNVs present in at least two mutation accumulation lines and removed them from the VCF files, based on the assumption that identical *de novo* mutations are highly unlikely to arise independently and derived from incorrect mapping or poor coverage. All the remaining SNVs were manually inspected using IGV (v2.3.68) and they were used to estimate the living ancestor's base substitution rate (μ) as $\mu=(n/bp)/g$, where n is the number of *de novo* mutations present in each line, bp is the (diploid) genome size of the living ancestor minus subtelomeric regions (22418023 bp) and g is the number of generations performed in the mutation accumulation experiment (2220). The median of base substitution rates among the eight lines was taken as the living ancestor's base substitution rate.

Finally, *de novo* mutations were annotated using the *S. cerevisiae* DBVPG6765 and the *S. paradoxus* CBS432 genome annotations²⁴ and their effect was predicted by Variant Effect Predictor (v87)⁵⁸ with the “--custom” parameter for custom reference genomes.

Molecular dating of genome instability and Alpechin emergence

The set of 207 SNVs described in the paragraph “Variant calling and SNVs analysis” was further refined for molecular dating analyses. We used the *S. paradoxus* CBS432 genome annotation²⁴ to extract SNVs sites that are either at the third codon position of coding regions or in intergenic regions. We retained 148 SNVs to estimate the timing since the LOH began in the living ancestor

and the split between the living ancestor and the Alpechin lineage. We first used a strict molecular clock model and estimated the number of generations (g) as $g=k_s/(2\mu)^{59}$. k_s is the substitution rate (SNVs in third position and intergenic regions/length of third positions plus intergenic regions in *S. paradoxus* LOH overlapping with introgressions) and μ is the per-nucleotide per-generation base substitution rate estimated through the analysis of the living ancestor's mutation accumulation lines (2.31×10^{-10}). Given that 148 substitutions were detected in the 312644 bp of the living ancestor's *S. paradoxus* LOH that overlapped with the Alpechins' introgressions, we estimated 1024316 generations since the genome shock has begun. This corresponds to a range of 2806 – 10243 years if we assume 100 – 365 clonal generations per year^{15,60}. Among the 148 SNVs, 90 have been inherited by at least one Alpechin isolate. We cannot confidently say whether the remaining 58 substitutions have been lost due to chromosome segregation or because they correspond to recent mutations that occurred after the split between the living ancestor and the Alpechin lineage. The 90 SNVs shared between the living ancestor and the Alpechin isolates resulted in a maximum age of 622895 generations occurring between the genome shock and the split. We therefore estimated that the Alpechins' birth occurred at the most 401421 generations ago, corresponding to a range of 1100 – 4014 years. This estimation assumes a similar generation time in nature in both branches after their split.

Next, we accounted for variation in the rate of the molecular clock using a random, local molecular clock model. First, we introduced mutations in the *S. paradoxus* CBS432 genome assembly in the positions corresponding to the 148 sites used in the previous molecular dating analysis, using the vcf-consensus program included in the suite VCFtools (v0.1.16). This allowed us to recreate living ancestor-like and Alpechin-like genome sequences. For the living ancestor-like sequence, we introduced mutations at all the 148 sites. For the Alpechin-like sequence, we introduced mutations at the 90 sites that retained mutations in the Alpechin lineage. We then extracted the sequences corresponding to the 312644 bp of the living ancestor's *S. paradoxus* LOH that overlapped with the Alpechins' introgressions from the *S. paradoxus* CBS432 genome assembly, from the living ancestor-like sequence and from the Alpechin-like sequence. We also extracted the sequences corresponding to the living ancestor's *S. paradoxus* LOH from the *S. cerevisiae* DBVPG6765 genome assembly.

We obtained four fasta sequences (DBVPG6765, CBS432, living ancestor and Alpechin) and aligned them using FSA (v1.15.9) with the “--refinement --exonerate --anchored” options⁶¹. The “--anchored” option of FSA further depends on MUMmer (v3.23)⁵⁶ and exonerate (v2.2.0)⁶² to assist the alignment. The aligned sequences were trimmed by trimAl (v1.4)⁶³ with the option “-gt 1.0” to remove gap positions. Finally, the filtered alignment was used to conduct a molecular dating analysis with BEAST2 (v2.6.2)⁶⁴, with the following parameter settings:

- Substitution model: HKY⁶⁵, with 4 Gamma category count and empirical base frequencies.
- Clock model: random local clock⁶⁶.

- Tree prior: calibrated Yule model⁶⁷.
- Molecular clock rate prior: Gamma (alpha=0.001, beta=1000).
- Speciation rate prior: Gamma (alpha=2.0, beta=2.0).
- Molecular clock rate prior: Gamma (alpha=2.0, beta=2.0).
- Calibration node: *S. cerevisiae* - *S. paradoxus* divergence time with a log normal distribution of $M=4.9$, $S=0.11$, which translates into a median divergence time of 4.87 million years ago (mya) with 95% central 95% probability range covering 4.06 mya - 5.84 mya (based on previous estimate⁶⁸).
- MCMC chain length: 100000000 with 1000 pre burn-in.
- MCMC chain logging frequency: every 1000 steps.

Two independent runs were performed and the convergence of these two MCMC runs was assessed by Tracer (v1.7.1) (<http://tree.bio.ed.ac.uk/software/tracer/>). TreeAnnotator (distributed with BEAST2) was used to generate the consensus tree with 10% burn-in. The resulting consensus tree was visualized with FigTree (v1.4.4) (<http://tree.bio.ed.ac.uk/software/figtree/>).

GO term analysis and annotation of LOH and introgressed genes

Introgressed genes were annotated using the CBS432 and DBVPG6765 genome annotations (**Supplementary Table 9**)²⁴. Standard GO term analysis was performed with the GO Term Finder tool available at *Saccharomyces* Genome Database (SGD). Significant GO terms were extracted by the algorithm implemented in the tool, with a False Discovery Rate (FDR) corrected α threshold of 0.05 (**Supplementary Table 8**).

References

- Liti, G., Warringer, J. & Blomberg, A. Isolation and Laboratory Domestication of Natural Yeast Strains. *Cold Spring Harb Protoc* **2017**, pdb.prot089052–6 (2017).
- Haber, J. E. Mating-Type Genes and MAT Switching in *Saccharomyces cerevisiae*. *Genetics* **191**, 33–64 (2012).
- Vázquez-García, I. *et al.* Clonal Heterogeneity Influences the Fate of New Adaptive Mutations. *Cell Reports* **21**, 732–744 (2017).
- De Chiara, M. *et al.* Domestication reprogrammed the budding yeast life cycle. *bioRxiv.org*. doi:10.1101/2020.02.08.939314
- Beltran, G., Novo, M., Rozès, N., Mas, A. & Guillamón, J. M. Nitrogen catabolite repression in *Saccharomyces cerevisiae* during wine fermentations. *FEMS Yeast Res.* **4**, 625–632 (2004).
- Zackrisson, M. *et al.* Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale. *G3 (Bethesda)* **6**, 3003–3014 (2016).
- Barré, B. *et al.* Intragenic repeat expansion in the cell wall protein gene HPF1 controls yeast chronological aging. *Genome Research* **30**, gr.253351.119–710 (2020).
- Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Research* **19**, 1117–1123 (2009).
- Yue, J.-X. & Liti, G. Long-read sequencing data analysis for yeasts. *Nat Protoc* **13**, 1213–1231 (2018).
- Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017).

54. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
55. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
56. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
57. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
58. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
59. Fay, J. C. & Benavides, J. A. Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet.* **1**, 66–71 (2005).
60. Rolland, T. & Dujon, B. Yeasty clocks: dating genomic changes in yeasts. *C. R. Biol.* **334**, 620–628 (2011).
61. Bradley, R. K. *et al.* Fast statistical alignment. *PLoS Comput. Biol.* **5**, e1000392 (2009).
62. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
63. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
64. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).
65. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
66. Drummond, A. J. & Suchard, M. A. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* **8**, 114 (2010).
67. Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**, 570–580 (2010).
68. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **175**, 1533–1545.e20 (2018).

Acknowledgements

We thank Anders Bergström, Gilles Fischer, Bertrand Llorente, Simone Mozzachiodi and Lorenzo Tattini for critical reading of the manuscript. This work was supported by Agence Nationale de la Recherche (ANR-11-LABX-0028-01, ANR-13-BSV6-0006-01, ANR-15-IDEX-01, ANR-16-CE12-0019 and ANR-18-CE12-0004) and the Swedish Research Council (2014-6547, 2014-4605 and 2018-03638). We thank the Institut Curie NGS platform for generating the sequence data. J.S. is a fellow of the University of Strasbourg Institute for Advanced Study and a member of the Institut Universitaire de France.

Author contributions

M.D., M.D.C., J.X.Y., J.W., G.L. designed the experiments; M.D., M.D.C., J.X.Y., performed the genomic analyses; M.D., A.I., A.L., performed and analysed sporulation efficiency, spore viability and LOH rates, S.S., K.P., performed and analysed the growth phenotype data, B.B. performed and analysed the chronological life span; E.G., R.M., J.S., J.W. and G.L. contributed with resources and reagents; J.W. and G.L. conceived and supervised the project; M.D., J.W. and G.L. wrote the paper.

Competing interest

The authors declare no competing interests.

Data availability

The genome sequences generated in this study are available at Sequence Read Archive (SRA), NCBI under the accession codes BioProject ID PRJNA594913, Biosample ID SAMN13540515 - SAMN13540586. Re-analysed sequencing data are available as part of the project ERP014555 in the Sequence Read Archive (SRA). The phenotype data are available within the supplementary information files. The source data underlying Figs. 1b, 2a-d, 4a-d and Extended Data Fig. 1a-b, 2, 4a-b, 5a-d, 6a-d, 7b-e, 8a, c, d, 9a-b, 10a-c are provided in the Source Data File. All the strains generated in this work are freely available upon request.

Code availability

Custom Bash and Perl scripts used in this study are available at:

<https://github.com/mdangiolo89/A-yeast-living-ancestor-reveals-the-origin-of-genomic-introgressions>

Extended data figure legends

Extended Data Fig. 1 | Genome-wide distribution of LOH and introgressions. **a**, Genotype maps of the genome of the living ancestor (top) and the 17 Alpechins. The living ancestor's DNA is coloured to denote *S. cerevisiae* (blue) and *S. paradoxus* (red) DNA, respectively. The Alpechins' DNA is coloured to denote the *S. cerevisiae* background (blue) and the interspersed introgressions that are homozygous (red) or heterozygous (grey) for *S. paradoxus* DNA. There is extensive overlap between the positions of *S. paradoxus* LOH in the living ancestor and introgressions in the Alpechins. **b**, Zoom-in of chromosome II. Colours as in panel **a**. White circles indicate centromeres. Genomic coordinates are based on the *S. cerevisiae* DBVPG6765 genome assembly.

Extended Data Fig. 2 | Alternative evolutionary models of the living ancestor's and Alpechins' origin. The genome-wide overlap between LOH, introgressions and mutated sites within LOH regions with two copies of *S. cerevisiae* or *S. paradoxus* DNA allowed us to test the alternative models 1 (left) and 2 (right). See Supplementary Discussion 1 for details. Top boxes report the number of introgressions and LOH, partitioned into overlapping and non-overlapping events. Bottom boxes report the number of *de novo* point mutations in LOH blocks of the living ancestor. Left and right boxes show mutations in *S. paradoxus* and *S. cerevisiae* LOH blocks, respectively. Numbers around the top chromosome pair report the total number of *de novo* mutations, partitioned into those occurring on the *S. cerevisiae* and *S. paradoxus* haplotype,

respectively. Numbers below the bottom chromosome pair report *de novo* mutations that are shared between the living ancestor and the Alpechins, again partitioned into those occurring on the *S. cerevisiae* and *S. paradoxus* haplotype, respectively. Haplotype phasing data is available only for the living ancestor. Patterns of overlap between *S. paradoxus* LOH/introgressions and shared *de novo* mutations between the living ancestor's *S. paradoxus* LOH and the Alpechin strains are compatible with model 1. In contrast, the sequence of events in model 2 makes it incompatible with the observed data: the emergence of *S. cerevisiae* LOH and their *de novo* mutations after the hybridization between an Alpechin and a *S. paradoxus* implies that none of these *de novo* mutations should be present in any Alpechin strain. The presence of 20 mutations which are shared between the Alpechins and the living ancestor rules out model 2 in favour of model 1.

Extended Data Fig. 3 | Genome structure of the living ancestor. Each coloured bar represents a chromosome, with dark tones representing *S. paradoxus* DNA and light tones representing *S. cerevisiae* DNA. Dark blocks within *S. cerevisiae* chromosomes represent *S. paradoxus* LOH, while light blocks within *S. paradoxus* chromosomes represent *S. cerevisiae* LOH. White circles indicate centromeres and numbers (from 1 to 5) annotate the recombination breakpoints of the gross chromosomal rearrangements listed in the top-right box. Subgenome assignment for rearranged chromosomes is based on centromere flanking sequences. Genomic coordinates of *S. cerevisiae* and *S. paradoxus* chromosomes are based on the genome assemblies of DBVPG6765 and CBS432 strains, respectively. We cannot confidently say whether intrinsic (genetic) or extrinsic (environmentally induced) genome instability explain the astounding genome instability signature, nor we can say whether it emerged from a single event or accumulated gradually.

Extended Data Fig. 4 | LOH size distribution. a, Histograms of LOH sizes in the living ancestor. The left panel reports the size histogram of *S. cerevisiae* LOH events (two copies of *S. cerevisiae* DNA retained and loss of the *S. paradoxus* DNA copy). Bin width is 4 kb. Conversely, the right panel reports the size histogram of the *S. paradoxus* LOHs (two copies of *S. paradoxus* DNA retained and loss of the *S. cerevisiae* DNA copy). Dark and light colours respectively indicate terminal and interstitial LOH events. Insets show x-axis zoom-ins up to 30 kb. Bin width is 2.5 kb. **b,** Size histograms of new LOH events in the mutation accumulation lines. Bin width is 15 kb. Insets show x-axis zoom-ins up to 30 kb. Bin width is 2.5 kb. Panel order and colour codes are as in panel **a**.

Extended Data Fig. 5 | Genome-wide genotype of the living ancestor gametes. a, We used the genome sequences of 25 living ancestor haploid gametes to infer their recombination landscape. Red and blue colours represent *S. paradoxus* and *S. cerevisiae* DNA, respectively. The upper part shows the living ancestor. Grey segments represent heterozygous regions for which both *S.*

paradoxus and *S. cerevisiae* DNA is present and correspond to duplicated regions derived from segregation of rearranged chromosomes or aneuploidies. Aneuploidies were all gain-of-copies, with chromosome III particularly prone to duplicate (7/25 gametes), consistent with the presence of a single LOH block encompassing the centromere, where crossovers are rare⁶⁹. Chromosomes with intra-homolog and inter-homolog gross chromosomal rearrangements are indicated with * and +, respectively. Genomic coordinates are based on the *S. cerevisiae* DBVPG6765 genome assembly. **b**, Percentage of the genome bearing *S. cerevisiae* or *S. paradoxus* DNA in the 25 living ancestor gametes. Colours as in panel **a**. **c**, The number of observed *S. cerevisiae* (blue) and *S. paradoxus* (red) haplotype blocks in three living ancestor gametes. These blocks derive from the combination of meiotic recombination events and pre-existing LOH. As comparison, we calculated the theoretical number of haplotype blocks expected from a single meiosis, considering only the number of crossovers, in an experimental *S. cerevisiae*-*S. paradoxus* F1 hybrid (data from Kao *et al.*¹³, brown bar), and in the living ancestor if LOH were not present (green bar). We estimated the number of expected haplotypes (y) as $y=16(x+1)$, where x represents the number of crossovers per chromosome. **d**, Size distribution of *S. paradoxus* haplotype blocks in three living ancestor gametes (upper panels) and of introgression blocks in three *Alpechins* (lower panels).

Extended Data Fig. 6 | Living ancestor clonal evolution. **a**, Genome-wide genotype maps of the living ancestor and its diploid clones derived by mutation accumulation line protocol (MAL) through 120 single-cell bottlenecks and return-to-growth (RTG) clones derived by inducing and then aborting meiosis. Grey represents heterozygous DNA, red represents LOH blocks with two *S. paradoxus* DNA copies, while blue represents LOH blocks with two *S. cerevisiae* DNA copies. Green arrows indicate new LOH recombination events, purple arrows indicate chromosome homogenisation by loss and re-synthesis events. Genomic coordinates are based on the *S. cerevisiae* DBVPG6765 genome assembly. **b**, Sporulation efficiency (at day 3) and **c**, spore viability of the living ancestor, its mutation accumulation lines (MAL, $n=8$), its mitotically growing (T0, $n=6$) and return-to-growth (RTG, $n=12$) clones. Boxes: horizontal line: median, upper/lower hinge: IQR, whiskers: largest/smallest value within upper/lower hinge $\pm 1.5 \times$ IQR. The bar plot (right) shows spore viability of each single living ancestor's mutation accumulation line. One line (A887R29) has a huge increase in spore viability due to tetraploidization. The pink dotted line indicates the spore viability of the living ancestor. **d**, Coverage plots of chromosome XI of a tetraploid (A887R29) and a diploid (A887R30) mutation accumulation lines. The plots show a balanced chromosome number for the A887R30 diploid clone (1 copy of *S. cerevisiae* and 1 copy of *S. paradoxus*) and an unbalanced one for the A887R29 tetraploid clone (3 copies of *S. cerevisiae* and 1 copy of *S. paradoxus*). Dots represent mean coverage values in 10 kb non-overlapping windows. Lines represent whole-chromosome coverage medians.

Extended Data Fig. 7 | Reconstruction of the hybridization-to-introgression model. **a**, The grey-dotted box contains a representation of the classic model for the origin of introgressions. The discovery of the living ancestor rewrote part of this mechanistic process and enabled us to test the downstream stages. For each stage of the model, we indicate the measured median gamete viability (%). Plain lines point to naturally occurring stages, while dotted lines point to stages derived in the laboratory. The initial *S. cerevisiae*-*S. paradoxus* F1 hybrid derived from an outbreeding event and this has been shown to be rare^{38,70}, thus one or very few hybridizations initiated the process. Next, the F1 hybrid expanded clonally to a possibly large population size, depending on nutrients availability in its ecological niche. This mitotic clonal expansion was followed by the genome instability and LOH formation phase. It is hard to predict the rate of genome instability, since it might be influenced by both strain background and the specific olive-oil related environment. However, genome instability of such magnitude might be rare and produce unviable combinations, likely reducing the ancestral hybrid population with LOH to one or few individuals. Next, the hybrid with LOH expanded mitotically, generating a population that kept evolving e.g. by accumulating additional LOH. Part of the population was maintained as asexual descendant, while part went through sexual reproduction and produced the Alpechin lineage. It is likely that the sporulation step was at the population level, therefore producing many spores. Next, one or more spores backcrossed to *S. cerevisiae*, which might occur rarely since it requires outbreeding. However, backcross of gametes from the ancestral hybrid was favoured by low gamete viability, which prevents intra-tetrad mating (that is the main way for spores to re-establish diploidy), while haplo-selfing was inhibited by fitness costs. We propose that this step bottlenecked the population to few individuals, consistent with our data showing that Alpechin strains share many of the LOH blocks. The living ancestor gametes were more fertile with *S. cerevisiae* than with *S. paradoxus*. We also crossed three living ancestor gametes to a haploid derivative of an Alpechin strain (CPG) and found fertility to be higher than in backcrosses with the most closely related *S. cerevisiae* lineage lacking introgressions. This is consistent with the Alpechin lineage being closer to the living ancestor. **b**, Sporulation efficiency after 3 days in sporulating conditions (central bar: median, upper/lower hinge: IQR, whiskers: largest/smallest value within upper/lower hinge $\pm 1.5 \times$ IQR) of hybridization-to-introgression stages. n =biologically independent samples. *S. cerevisiae* (Sc, $n=1$), *S. paradoxus* (Sp, $n=2$), experimental first-generation (F1) hybrid (F1 hybrid, $n=2$), living ancestor (LA, $n=1$), living ancestor gametes (LA gametes, $n=25$), backcross to *S. cerevisiae* (Bc Sc, $n=8$), backcross to *S. paradoxus* (Bc Sp, $n=8$), Alpechin (Alp, $n=16$), living ancestor gametes x Alpechin (LA gametes x Alp, $n=3$), *S. cerevisiae* x Alpechin (Sc x Alp, $n=1$), *S. paradoxus* x Alpechin (Sp x Alp, $n=1$). The *S. cerevisiae* x Alpechin cross (Sc x Alp) is a slow sporulator and requires additional time to sporulate, consistent with the presence of inefficient sporulation alleles in the Wine/European *S. cerevisiae* background⁷¹. **c**, Sporulation efficiency (at day 3) of living ancestor gametes grouped by number of aneuploidies (central bar: median, upper/lower hinge: IQR, whiskers: largest/smallest value within

upper/lower hinge $\pm 1.5 \times$ IQR). n =biologically independent samples. 0 aneuploidies ($n=5$), 1 aneuploidy ($n=6$), 2 aneuploidies ($n=12$), 3 aneuploidies ($n=1$), 4 aneuploidies ($n=1$). Pink dots indicate spores with an additional copy of chromosome III, which remained haploid and thus do not sporulate. **d**, Spore viability (central bar: median, upper/lower hinge: IQR, whiskers: largest/smallest value within upper/lower hinge $\pm 1.5 \times$ IQR) of hybridization-to-introgression stages. n =biologically independent samples. *S. cerevisiae* (Sc, $n=1$), *S. paradoxus* (Sp, $n=2$), experimental first-generation (F1) hybrid (F1 hybrid, $n=2$), living ancestor (LA, $n=1$), living ancestor gametes (LA gametes, $n=13$), backcross to *S. cerevisiae* (Bc Sc, $n=8$), backcross to *S. paradoxus* (Bc Sp, $n=8$), Alpechin (Alp, $n=4$), living ancestor gametes x Alpechin (LA gametes x Alp, $n=3$), *S. cerevisiae* x Alpechin (Sc x Alp, $n=1$), *S. paradoxus* x Alpechin (Sp x Alp, $n=1$). The living ancestor has higher spore viability than the experimental *S. cerevisiae*-*S. paradoxus* F1 hybrid. The spore viability of the backcrosses to *S. cerevisiae* compares favourably to that of the *S. paradoxus* backcrosses. The Alpechins are highly fertile with a Wine/European *S. cerevisiae* (Sc) background without introgressions and moderately fertile with the living ancestor gametes, whereas they have strong reproductive barriers with *S. paradoxus* (Sp). **e**, Percentage of viable spores for haploid-stable living ancestor gametes backcrossed to *S. cerevisiae* (blue line boxes) or *S. paradoxus* (red line boxes), grouped by their number of aneuploidies. Boxes: horizontal line: median, upper/lower hinge: interquartile range (IQR), whiskers: max/min values. n =biologically independent samples. 0 aneuploidies ($n=6$), 1 aneuploidy ($n=6$), 2 aneuploidies ($n=2$), 3 aneuploidies ($n=2$).

Extended Data Fig. 8 | Mapping the slow growth phenotype in the Alpechin AQA strain. **a**, Percentage of introgressed DNA in the Alpechin strains ranked for amount of heterozygous introgressions (grey, ordered first in the stack bar). Most introgressed material is in homozygous state (red, ordered second). Three strains, including AQA, have a considerable fraction of heterozygous introgressions. The rest of the genome up to 100% is *S. cerevisiae* background (blue, ordered third) and the y-axis is cut at 10%. **b**, Tetrad dissection of the Alpechin AQA strain showing a 2:2 colony size segregation pattern (two slow-growing and two normal-growing colonies in each meiosis), consistent with a single heterozygous variant underlying the growth defect. **c**, Each dot represents the allele frequency (AF) difference of reference and non-reference alleles across the genome between pools of big and small spores. Two regions show fixation of the opposite alleles in the big and small pools and chromosome XII is compatible with the assumption that the deleterious allele associated with the small phenotype has the lowest frequency in the *S. cerevisiae* population. **d**, Chromosome XII AF zoom and map of homozygous (red) and heterozygous (grey) introgressions on a *S. cerevisiae* genome backbone (blue) of the Alpechin strain AQA. The growth defect maps to the *SFI1* gene, which is far from the closest introgression, indicating that the slow growth phenotype is not driven by a deleterious effect of heterozygous introgressions masked by the *S. cerevisiae* allele. The white circle indicates the centromere. Genomic coordinates are based on the *S.*

cerevisiae DBVPG6765 genome assembly.

Extended Data Fig. 9 | The competitive fitness of the living ancestor. **a**, Mitotic growth (absolute doubling time, h) of the living ancestor ($n=40$) as compared to the closest extant relatives of its *S. cerevisiae* DBVPG6765 (upper left, $n=8$) and *S. paradoxus* N17 (upper middle, $n=16$) parents and a diploid experimental F1 hybrid between the two (upper right, $n=16$), the living ancestor's diploid gametes (lower left, $n=200$) and backcrosses of its haploid-stable gametes to *S. cerevisiae* (lower middle, $n=64$) and *S. paradoxus* parents (lower right, $n=64$). The number of observations (n) results from the combination of the strains in each group and the number of replicates performed in the experiment. Significant differences enumerated in the main text are indicated with asterisks or name (two-tailed Welsh test, FDR $q=0.05$) and presented in Supplementary Table 7. For the living ancestor gametes, we instead indicate (red text) the rare environments where no significant difference was found. **b**, The percentage of surviving cells (average \pm SD) for each step of the hybridization-to-introgression process during a 25-days chronological lifespan experiment. n =biologically independent samples. *S. cerevisiae* ($n=1$), *S. paradoxus* ($n=2$), experimental first-generation (F1) hybrid ($n=2$), living ancestor ($n=1$), living ancestor gametes ($n=25$), backcross to *S. cerevisiae* ($n=8$), backcross to *S. paradoxus* ($n=8$), Alpechin ($n=16$). **c**, Gating strategy used for the estimation of the percentage of surviving cells shown in panel **b**.

Extended Data Fig. 10 | The adaptive value of introgressions. **a**, Heatmap of the relative doubling time (mean of $n=40$ replicates for the living ancestor and $n=8$ replicates for all the other strains) of the living ancestor, its modern Alpechin descendants and the closest living relatives of its *S. cerevisiae* and *S. paradoxus* parents. Strains were grouped (left) based on doubling time similarity using hierarchical clustering (Pearson's r , complete linkage) and branch colours highlight four major clusters. Environments in which the living ancestor grows significantly (two-tailed Welsh test, FDR $q=0.05$) faster (green) or slower (red) than the Alpechins are coloured. **b**, Difference in relative doubling time (central bar: median, upper/lower hinge: IQR, whiskers: largest/smallest value within upper/lower hinge $\pm 1.5 \times$ IQR) between Alpechins ($n=128$) and *S. cerevisiae* ($n=8$) in the environments which favour Alpechins' growth (two-tailed Welsh test, FDR $q=0.05$). The number of observations (n) results from the combination of the strains in each group and the number of replicates performed in the experiment. Each dot represents the average among the 8 replicates of each of the 16 Alpechins. **c**, Genotype maps of the Alpechin, Mexican agave, French Guiana and Brazilian bioethanol clades in a shared introgressed region containing the *IRC7* (*YFR055W*) gene. This gene was previously reported to be introgressed in strains belonging to the Wine/European clade⁷². The *IRC7* gene present in the Alpechins has a European *S. paradoxus* ancestry, while Mexican Agave, Brazilian Bioethanol and French Guiana lineages inherited the *S. paradoxus*

American allele, underlying an independent introgression and maintenance of *IRC7* that supports adaptive introgression. Blue and red colours indicate *S. cerevisiae* and *S. paradoxus* regions, respectively. Subtelomeric regions are masked genome wide because they lack synteny between *S. cerevisiae* and *S. paradoxus* and cannot be aligned for markers genotyping, although inspection of this specific VI-R subtelomeric region in both *S. cerevisiae* and *S. paradoxus* genomes indicates that the telomeric repeats are very close to *IRC7*²⁴. The bottom panel shows genetic features annotation of this region from the *S. cerevisiae* Wine/European strain DBVPG6765.

References

69. Pan, J. *et al.* A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144**, 719–731 (2011).
70. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38**, 1077–1081 (2006).
71. Gerke, J., Lorenz, K. & Cohen, B. Genetic interactions between transcription factors cause natural variation in yeast. *Science* **323**, 498–501 (2009).
72. Roncoroni, M. *et al.* The yeast *IRC7* gene encodes a β -lyase responsible for production of the varietal thiol 4-mercapto-4-methylpentan-2-one in wine. *Food Microbiology* **28**, 926–935 (2011).