Quantifying incomplete lineage sorting and introgression throughout Saccharomyces cerevisiae evolutionary history

Nicolò Tellini

UNIVERSITÉ CÔTE D'AZUR | ÉCOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTÉ

THÈSE DE DOCTORAT

Quantification du tri des lignées incomplètes et de l'introgression au cours de l'histoire évolutive de *Saccharomyces cerevisiae*

Nicolò TELLINI

Genomique des populations et traits complexes
IRCAN-CNRS UMR7284-INSERM U1081

Présentée en vue de l'obtention du grade de docteur en
Sciences de la vie et de la santé
d'Université Côte d'Azur

Dirigée par : Gianni LITI, DR, CNRS, Université Côte d'Azur

Soutenue le : 05/12/2023

Devant le jury, composé de :
Etienne DANCHIN, DR, INRA, Université Côte d'Azur, President du Jury

Tatiana GIRAUD, DR, CNRS, Université Paris-Saclay, Rapporteur

Delphine SICARD, DR2, INRAE, Université de Montpellier, Rapporteur

Anders BERGSTRÖM, UEA, University of East Anglia, Examinateur

IRCAN | cnrs | RÉGION SUD PROVENCE ALPES CÔTE D'AZUR | Agence Nationale de la Recherche ANR | Fondation pour la Recherche Médicale

QUANTIFICATION DU TRI DES LIGNÉES INCOMPLÈTES ET DE L'INTROGRESSION AU COURS DE L'HISTOIRE ÉVOLUTIVE DE *Saccharomyces cerevisiae*

Quantifying incomplete lineage sorting and introgression throughout Saccharomyces cerevisiae evolutionary history

Nicolò TELLINI

Jury :

Président du jury

Etienne DANCHIN, DR, INRA, Université Côte d'Azur
President du Jury

Rapporteurs

Tatiana GIRAUD, DR, CNRS, Université Paris-Saclay, Rapporteur
Delphine SICARD, DR2, INRAE, Université de Montpellier, Rapporteur

Examinateurs

Anders BERGSTRÖM, UEA, University of East Anglia, Examinateur

Directeur de thèse

Gianni LITI, DR, CNRS, Université Côte d'Azur

Université Côte d'Azur

Nicolò TELLINI

Quantification du tri des lignées incomplètes et de l'introgression au cours de l'histoire évolutive de Saccharomyces cerevisiae

viii+126 p.

Ce document a été préparé avec LaTeX2e et la classe `these-ISSS` version v. 2.10.

Impression : output.tex – 29/2/2024 – 8:48

Révision pour la classe : these-ISSS.cls,v 2.10 2020/06/24 14:16:37 mpelleau

Quantification du tri des lignées incomplètes et de l'introgression au cours de l'histoire évolutive de *Saccharomyces cerevisiae*

Résumé

L'étude de la distribution de la variation génétique au sein des populations et entre elles permet de mieux comprendre l'histoire évolutive d'une espèce. De plus, l'accès à de grands ensembles de données génomiques permet d'étudier les processus évolutifs qui façonnent la variation ségrégative de l'espèce. Dans ce travail, nous retraçons l'histoire évolutive de l'espèce *Saccharomyces cerevisiae* par la détection et la classification de polymorphismes partagés avec son espèce sœur *Saccharomyces paradoxus*. Nous identifions des polymorphismes partagés acquis par hybridation suivie d'introgression et des polymorphismes qui persistent à travers le processus de spéciation et de diversification résultant en des cas de triage incomplet des lignées (ILS).Nous définissons un ensemble de données de polymorphismes nucléotidiques simples diagnostiques bialléliques entre *Saccharomyces cerevisiae* et *Saccharomyces paradoxus* que nous utilisons comme marqueur diagnostique pour décrire la composition génomique de 1,673 *S. cerevisiae*, pour lesquels un séquençage à lecture courte du génome entier était publiquement disponible. Nous développons une méthode basée sur les marqueurs pour la détection et la classification des marqueurs de diagnostic organisés soit en 1) blocs de marqueurs *S. paradoxus* consécutifs, soit en 2) marqueurs *S. paradoxus* isolés à l'échelle du génome. Pour les blocs, nous décrivons les limites, et la distribution dans la collection *S. cerevisiae* et nous retraçons l'origine de *S. paradoxus* par comparaison de séquences avec des assemblages de génomes entiers télomère à télomère des principales populations de *S. paradoxus*. Pour un événement récurrent, nous avons effectué un test pour évaluer l'effet sur la condition physique de porter un haplotype *S. paradoxus* à un locus unique englobant une paire de gènes impliqués dans la dégradation de composés toxiques pour la levure. Nous avons démontré que l'haplotype *S. paradoxus* confère un avantage par rapport à l'haplotype *S. cerevisiae* dans des conditions environnementales caractéristiques de la niche habitée par la population *S. cerevisiae*. Pour les marqueurs isolés, nous appliquons une méthode classique de détection des signatures d'un tri lignager incomplet, qui peut expliquer l'excès de marqueurs S. paradoxus distribués à l'échelle du génome dans certaines populations de *S. cerevisiae*. Nous montrons des preuves convaincantes de la rétention d'allèles ancestraux dans une seule population sauvage de *S. cerevisiae* qui se trouve à la racine de l'espèce. Nous émettons l'hypothèse que la persistance d'une telle variation ancestrale est due à la possibilité réduite de croisement avec d'autres populations de *S. cerevisiae* dans la nature, en raison du nombre réduit de générations et des goulets d'étranglement moins spectaculaires qu'ont connu les autres lignées au cours de la dispersion et de la domestication.Dans l'ensemble, nous avons retracé l'histoire de la divergence et des contacts secondaires entre les populations de *S. cerevisiae* et de *S. paradoxus* et dévoilé un cas convaincant d'introgression interespèces avec un résultat fonctionnel.

Mots-clés : Saccharomyces, triage incomplet des lignées, hybridation, introgression, introgression adaptative, introgression ancestrale.

Quantifying incomplete lineage sorting and introgression throughout *Saccharomyces cerevisiae* evolutionary history

Abstract

The study of the distribution of the genetic variation within and across populations provides insights into the evolutionary history of a species. Moreover, the access to large genomic datasets allows the investigation of the evolutionary processes that shape the species' segregating variation. In this work, we retrace the evolutionary history of the species *Saccharomyces cerevisiae* through the detection and classification of shared polymorphisms with its sister species *Saccharomyces paradoxus*. We identify shared polymorphisms acquired because of hybridization followed by introgression and polymorphisms that persist across the process of speciation and diversification resulting in instances of incomplete lineage sorting (ILS). We define a dataset of biallelic diagnostic single nucleotide polymorphisms between *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* that we use as diagnostic markers to describe the genomic composition of 1,673 *S. cerevisiae* genomes, for which short-read whole-genome sequencing were publicly available. We develop a marker-based method for the detection and classification of diagnostic markers organized either in 1) blocks of consecutive *S. paradoxus* markers or in 2) genome-wide isolated *S. paradoxus* markers. For the blocks, we describe the boundaries, the distribution across the *S. cerevisiae* collection and we retrace the *S. paradoxus* origin by sequence comparison with telomere-to-telomere whole genome assemblies from the main *S. paradoxus* populations. For a recurrent introgression event on chromosome IV, we performed an assay to evaluate the fitness effect of carrying a *S. paradoxus* haplotype at a single locus encompassing a gene-pair involved in the degradation of toxic compounds for the yeast. We demonstrate that the *S. paradoxus* haplotype confers an advantage over the *S. cerevisiae* haplotype in environmental conditions that are characteristic of the niche that the *S. cerevisiae* population inhabits. For the isolated markers, we apply a classical method for detecting signatures of incomplete lineage sorting, which can explain the excess of genome-wide distributed *S. paradoxus* markers in certain populations of *S. cerevisiae*. We show convincing evidence of retention of ancestral alleles in a single wild *S. cerevisiae* population that diverged early from other lineages. We speculate about the persistence of such ancestral variation due to reduced possibility of outcrossing with other *S. cerevisiae* populations into the wild, or due to fewer generations and less dramatic bottlenecks that were instead experienced by the other lineages during dispersal and domestication. Overall, we retraced histories of divergence and secondary contacts across *S. cerevisiae* and *S. paradoxus* populations and unveiled a compelling case of interspecies introgression with a functional outcome.

Keywords: Saccharomyces, incomplete lineage sorting, hybridization, introgression, adaptive introgression, ancestral introgression.

Table of Contents

CHAPTER 1

Incomplete lineage sorting

In this first chapter I describe the process of incomplete lineage sorting

1.1 Evolution and the study of genetic variation

Evolution is largely driven by heritable alterations that living organisms pass on from one generation to the next. At the DNA level, genetic changes are inherited by descendants. The most common DNA change is the single nucleotide polymorphism (SNP) in which one nucleotide is replaced by another one. SNPs play a remarkable role in the evolutionary history of populations as SNPs contribute greatly to genetic diversity. Genetic diversity makes a population better equipped for facing diverse environmental conditions in the struggle for life ensuring or facilitating the persistence of the population itself. SNPs accumulate over time and can be used to infer relatedness at the level of individuals and populations. The study of the relatedness and evolutionary history of populations is the main goal of phylogenetics. A common approach consists of investigating evolutionary patterns among contemporary individuals taking advantage of genomic data. The evolutionary affiliations that emerge are hypothesised relationships supported by a degree of dissimilarity and are depicted with a phylogenetic tree where each lineage is a line of biological entities that change over time, connected by ancestry-descent relationships (Hull 1980; Neto 2019). Untangling and describing the evolutionary history of different populations is a challenging task due to the complexity of the evolutionary processes. For the sake of clarity and simplicity, I am going to use the terms species/population, and speciation/diversification interchangeably throughout this chapter.

1.2 The fate of genetic variants with regard to the size of the population

One of the most accepted evolutionary theories is the nearly neutral theory of evolution proposed in 1975 by Ohta (Ohta 1973). Ohta proposed that most SNPs have slightly deleterious effects and segregate as nearly neutral rather than under a completely neutral regime as previously proposed with the neutral theory of evolution (Kimura 1991). The frequency of a slightly deleterious mutation inside a population, compared to a completely neutral variant, depends on the effective population size. When a population is small the fate of the variant is determined by the genetic drift i.e. by random sampling alleles, generation after generation. Under genetic drift alone one allele at a variant site will eventually become fixed in the population (Fig. 1.1).

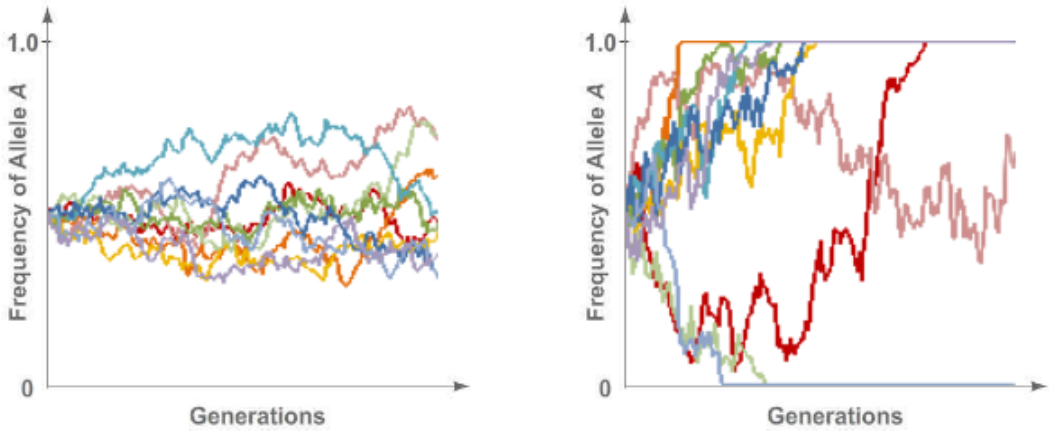

Figure 1.1 – Simulation of genetic drift. On the left starting frequency of A = 0.5, population size of 1000 individuals. On the right starting frequency of A = 0.5, the population size is 100 individuals. In both cases, I simulated 10 runs of 200 generations using the simulator online tool 'Genetic Drift Simulation - W. H. Freeman'

In contrast, in a large population, the stochastic fluctuation of the allele frequency of the variant, due to drift, becomes a negligible aspect and natural selection can instead eliminate the deleterious variants. Moreover, as empirically shown by Chao & Carr 1993 the effective population size is affected by the generation time of a species, so species with longer generation time have smaller population size, and species with a shorter generation time a larger population size. Nevertheless, generation time and population size compete against each other. A long generation time reduces the mutation rate per year, but a small population give the possibility for variants to drift and fix thereby increasing the fixation rate. Conversely, a short generation time increases the mutation rate per year but a large population influences the fate of the variants by selection rather than drift and in the long term slightly deleterious alleles are lost thereby decreasing the fixation rate (Ridley 2004). In addition, population histories impact the fate of variants through expansion and shrinking of population sizes. For example, both founder effects and bottlenecks affect the population size. The migration of a small group of individuals that settles in a new niche or a different place drastically impacts the genetic diversity of the new population as the immigrants expand from a small subgroup of the initial population, leading to a loss of genetic variation in the newly forming population. Scholefield & Greenberg 2007 showed the founder effect on a haplotype responsible for Huntington's disease in two South African populations. On the other hand, environmental changes, diseases, hunting, predation or habitat destruction can be responsible for reducing the population size of a species and result in higher levels of drift and decreased genetic

variation as, for example, has been observed in the American bison (Hedrick 2009). Overall, the size of the population plays a central role in determining the fate of the variation of a species.

1.3 The intimate relation between species and gene trees

A species tree depicts a summary of the relationships among species. Back in the years, the availability of few gene sequences/loci from a restricted group of individuals offered a limited view of the evolutionary history of species. For example, the Neoaves clade of birds includes up to 95% of bird species today known. The Neoaves clade experienced a rapid radiation that impairs the reconstruction of a robust phylogeny. In 2008, a first phylogeny based on 19 loci for 169 species highlighted previously unrecognized relationships and changed others (Hackett et al. 2008, Fig.1.2 A). In 2014, along with the G10K project, which aims to sequence one genome from each vertebrate genus, Jarvis et al. 2014 took advantage of whole genome data of 48 species representing all orders of Neoaves and found out that deeper branches exhibit higher gene tree incongruence because of incomplete lineage sorting (Fig.1.2 B).

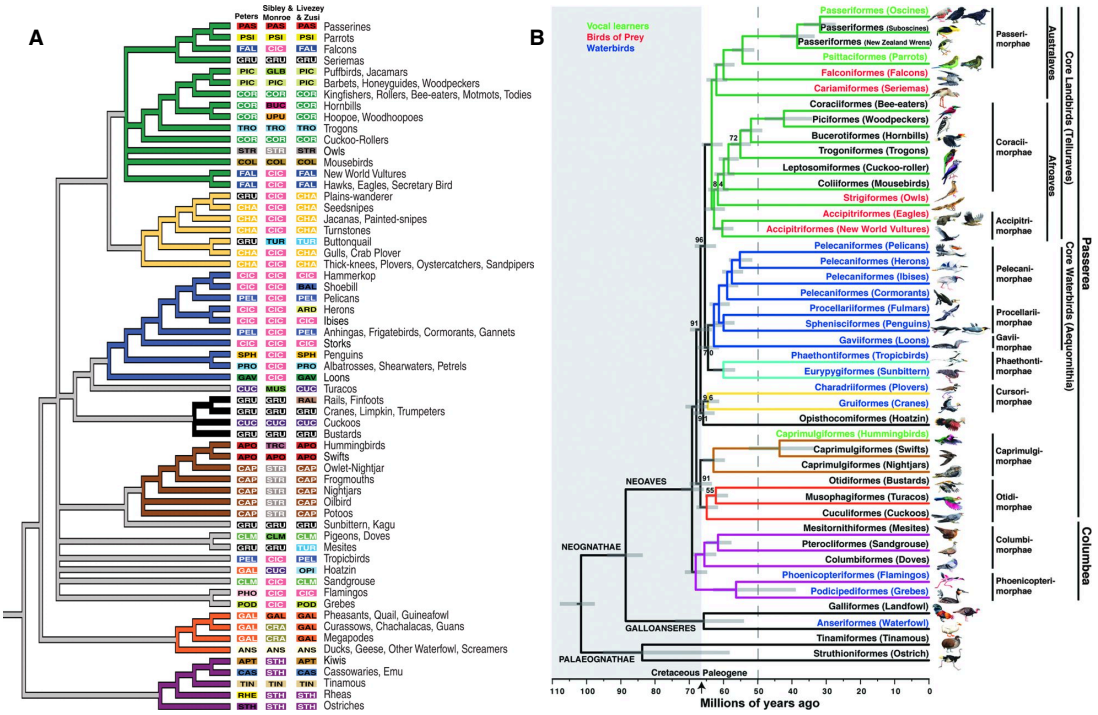

Figure 1.2 – In panel **A**, the phylogenetic tree of 19 loci across 169 species of birds, strongly supported nodes (100% bootstrap) were collapsed. Grey polytomic branches represent unresolved nodes. In panel **B**, time-calibrated phylogeny of 48 bird species from whole-genome sequencing data. The grey rectangle highlights deep unresolved nodes. Adapted from Hackett et al. 2008 and Jarvis et al. 2014.

The advent of whole-genome sequencing approaches allowed the inclusion of hundreds or thousands of genes from large groups of individuals providing a better inference of the species tree. However, a phylogeny based on whole-genome data which introduces several genes relies on the assumption that the topology of the species tree reflects the most common gene tree typologies. We can consider three species A, B and C and their species tree along with the overall topological gene structures which occur with the frequency of 82% and the other two at 9% each (Fig. 1.3). Discordant gene trees are part of the evolutionary history of the species (Maddison 1997).

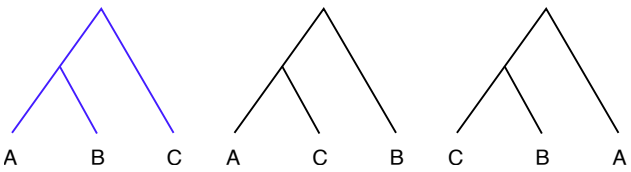

Figure 1.3 – The blue phylogeny is the most common phylogeny (species phylogeny) among three generic species A, B, and C. Discordant phylogenies are depicted in black.

For example, the diversification of the butterfly genus *Heliconious* was boosted by rapid speciation events driven by adaptive species radiation, a process triggered by ecological opportunities such as the availability of new resources or the loss of predators and opened the route to ancestral populations to adapt to new environmental niches. The *H. erato* clade, in particular, was shaped by rapid radiations, widespread hybridization and introgression interrupting the bifurcation history of the species, resulting in conflicting gene topologies (Fig.1.4).

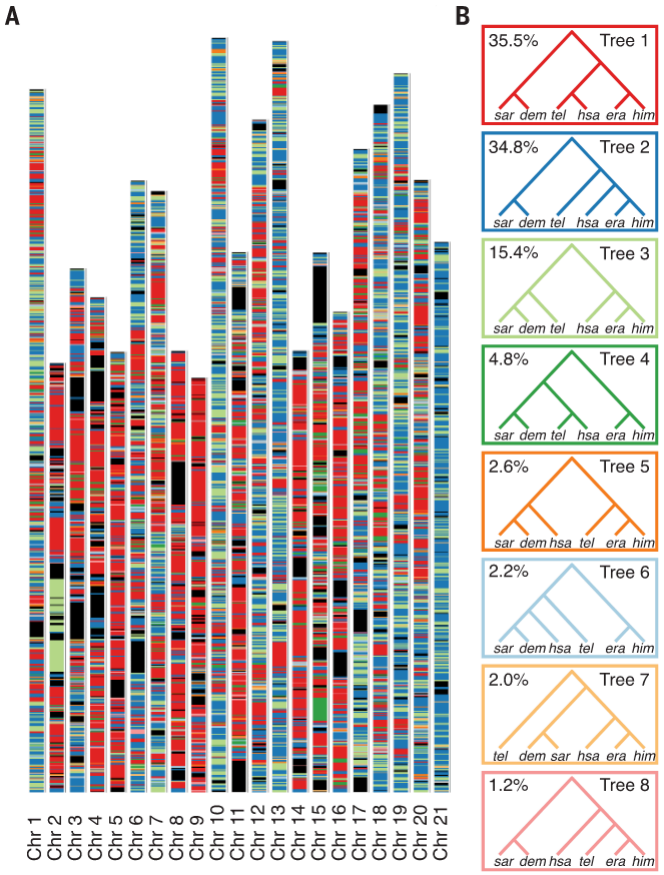

Figure 1.4 – The genus *Heliconious* is a species-rich group of butterflies. The figure describes the complexity and heterogeneity of the genome ancestry inside the clade *H. erato* and *H. sara*. In **A**, coloured rectangles represent tree topologies (consecutive 50-kb windows) and colours match the topologies depicted in **B**. In **B** are shown the eight most common topologies. Up to 70.3% of the windows are described by two different topologies (Tree1 and Tree2). Adapted from Edelman et al. 2019.

Thus, the evolutionary relationships of species are complex and the building blocks of this complexity, for example, the evolutionary history of each single gene, potentially experienced and followed different evolutionary pathways. Often, neighbouring genes on a chromosome evolve following a similar path because of linkage disequilibrium but this might not be the case when

we consider genes at unlinked loci. As Maddison 1997 proposed we can think about species tree as a *"conglomerate of gene trees where the containing species tree descends and branches, while within its branches a number of contained gene trees descend and branch"* (Fig.1.5 A). A species tree depicted as a cloud of overlapping gene trees allows us to visualize the multitude of topologies that might be in conflict with the species tree (Fig.1.5 B).

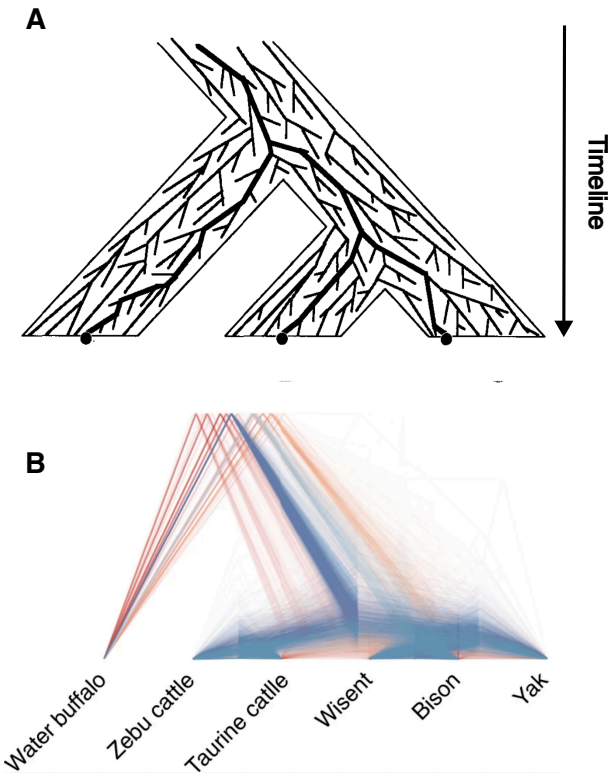

Figure 1.5 – In panel **A** a graphical summary of gene trees extending over a generation. The bold line highlights a single gene tree that descended from an ancestral species to today's living species. The dots are different alleles. In panel **B**, K. Wang et al. 2018 showed the overlap of 3306 gene trees with bootstrap support of at least 75% across all nodes, in the *Bovini* tribe. The blue lines indicate the most frequent nuclear phylogeny, which is shared by only 53.5% of the 3306 gene trees. Panel A adapted from Maddison 1997 and panel B adapted from K. Wang et al. 2018.

These topological inconsistencies are the result of different evolutionary processes. For the rest of this chapter, I will focus on describing the incomplete lineage sorting while a more comprehensive review of the causes and effects of topological tree discordance can be found in (Steenwyk et al. 2023).

1.4 Incomplete lineage sorting

A clear qualitative understanding of incomplete lineage sorting can be facilitated by introducing a directional timescale to a three-way species phylogenetic tree (Fig.1.6).

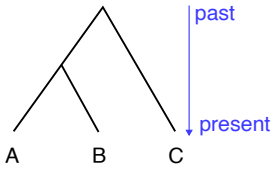

Figure 1.6 – A generic three species phylogeny with a timeline.

If we follow the species tree in a time-forward direction, we can define lineage sorting as the removal of ancestral polymorphisms generating monophyletic species (Paul Moran & Irv Kornfield 1993). When the ancestral species fail to remove ancestral polymorphisms, the retention of ancestral alleles manifests, in phylogenetics, as *hemiplasy* i.e the topological discordance between the gene tree and species tree (Avise & Robinson 2008). The unsuccessful removal of ancestral alleles results in *incomplete lineage sorting* (ILS). For example, given three generic populations (A, B and C) we can depict the phylogenetic discordance between the gene and the species tree by overlapping their topologies (Fig.1.7). At the root of the phylogeny, the A-B-C ancestral species contains two generic ancestral alleles (α and β). At the first speciation event, the allele β becomes characteristic of the C species, which survives to our time, while, on the other branch, α and β persist in the A-B ancestral species and ultimately sort at the second speciation event leaving the species A with the α allele and the species B with the β allele. Attempting to reconstruct the phylogenetic relationship starting from this allelic position results in a phylogenetic incongruence between the gene and the species tree (Fig.1.7). The ancestral α and β alleles also sort : β in A and α in B (not shown in Fig.1.7). Both the phylogenies in Fig.1.7 are characterized by a time span between the two speciation events, in which the α and the β alleles persist in the A/B ancestral population (A-B branch). Intuitively, we can think that if the time between the two consecutive speciation events is short and/or the effective population size of the ancestral population is large, the ILS will be more likely. Extending the reasoning to a large number of genes and assuming that their most common evolutionary trend follows a nearly neutral evolutionary scenario, a short speciation time does not allow the ancestral population to fix one of the ancestral alleles, by means of genetic drift, resulting in lineages that inherit such variation at some of the loci making them

similar despite not being closely related. Similarly, a large effective population size contributes to the persistence of variation in the ancestral population. Citing Maddison 1997 *"the different genes are competing for the same locus in the genome, and the probability that two copies will find themselves sitting on the same chair* (persist at the same locus) *each time the music stops* (after the speciation event) *will depend in a fairly simple way on the number of chairs*" i.e. the population size. So, although the ILS is a source of phylogenetic incongruence that obscures the evolutionary relationships among species, this evolutionary process results in the retention of ancestral alleles in the descending species, contributing towards the genetic diversity (Rivas-González et al. 2023) and phenotypic variation (Feng et al. 2022) of extant populations.

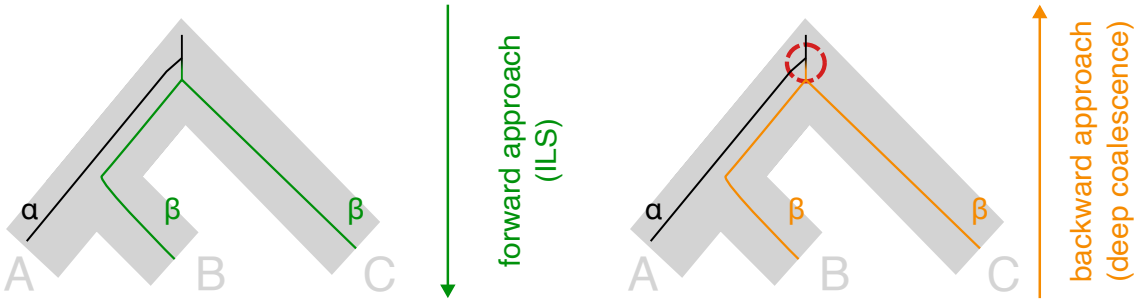

Figure 1.7 – On the left, overlap between the species tree (in the background in light grey) of three generic species A, B, C ; overlaid with a phylogenetic tree of a biallelic locus with α and β alleles (solid lines). The green lines depict the sorting of the β allele. The green arrow depicts the direction to follow to describe the repartition of the alleles as incomplete lineage sorting (from the past to the present). The sorting of α and β alleles is in conflict with the species tree : the β allele of the **B** population branches with the β allele of the **C** population. **B** and **C** populations seem monophyletic, but the species tree tells us that they are paraphyletic. On the right, overlap between the species tree (in the background in light grey) of three generic species A, B, and C. On top of it, is a phylogenetic tree of a biallelic locus with α and β alleles. The orange lines depict the sorting of the β allele. The orange arrow depicts the direction to follow to describe the coalescence of the alleles (from the present to the past). The α allele of population A and β allele of population B coalesce further back in time (at the node indicated with a red circle) compared to the population split (A-B node on the species tree). So, the coalescence between α and β occurs deep in the phylogeny (deep coalescence). For simplicity, only one discordant phylogeny is included in the figure.

1.4.1 Deep coalescence

For completeness, the same process can be also described in a time-backwards direction (Fig.1.7). Instead of lineage sorting, the term *coalescence* refers to the case where different lineages find their common ancestor at the most recent branching point (Avise et al. 1983). Consequently, *hemiplasy* is due to the failure of present-day lineages to coalesce at the more recent common ancestor. Instead of ILS, we refer to *deep coalescence* because the coalescence is successfully

reached further back in time, predating the speciation event (Maddison 1997). A detailed and rigorous mathematical treatment of the coalescent theory and its implementation, along with its extended version that accounts for ancestral polymorphisms, the multi-species coalescent (MSC) theory, can be found in Kingman 1982 and Mirarab et al. 2021. The most powerful applications of the MSC allow the estimation of parameters of the ancestral population such as species divergence times, and effective population sizes as well as phylogenetic tree reconstruction (Liu et al. 2019). For example, a recent paper provided an accurate fossil-free estimation of speciation times, which is consistent with previous estimates from fossil records, and ancestral population sizes along the primate phylogeny (Rivas-González et al. 2023).

1.5 Detection of ILS

The methods for the detection and the quantification of the ILS can be divided in two main groups :

— Coalescent-based methods

— Symmetry-based methods

1.5.1 Coalescent-based methods

I previously introduced *the coalescent* from a qualitative point of view. However, *the coalescent* is more than a qualitative model. Published in Kingman 1982, *the coalescent* is a mathematical model used for describing stochastic processes in a time-backward direction, such as the random coalescence of alleles to the most recent common ancestor. Compared to classical evolutionary models, the time-backwards approach offers an advantageous way to perform simulations. The computationally easier and more efficient implementation of the time-backwards approach makes *the coalescent* one of the most attractive approaches for several applications (Rosenberg & Nordborg 2002). For example, the coalescent hidden Markov model (CoalHMM) is a computational technique that allows the estimation of the ancestral population size, the speciation time and the recombination rate which are all parameters that affect the probability of observing ILS (Dutheil et al. 2009). Recently, Rivas-González et al. 2023 applied the CoalHMM to a multi-species alignment of primates to estimate the frequencies of ILS across 29 ancestral nodes of the primate phylogeny and found out that up to 64% of the genome, at individual nodes, is affected by ILS.

1.5.2 Symmetry-based methods

The symmetry-based methods are summary statistics based on the frequency of occurrence of discordant trees. Given the stochastic nature of the inheritance of the ancestral alleles, for all the discordant phylogenies, equal occurrences across the descending species are expected.

1.5.2.1 ABBA-BABA test

Given four populations indicated with the labels P1, P2, P3 and a P4 that serves as an outgroup (or 'O'), we select biallelic variants. We then define the allele in O ancestral (or 'A') and the other, the derived allele (or 'B'). In case of incomplete lineage sorting between P1 and P2 the segregation of the alleles A and B, in the P1-P2 ancestral population can be resolved in two different ways : A sorts toward the population P1 and B sorts toward the population P2 and, vice versa, A sorts toward the population P2 and and B sorts toward the population P1. The intrinsic stochasticity of the process of segregation gives rise to incomplete lineage sorting (Fig. 1.8).

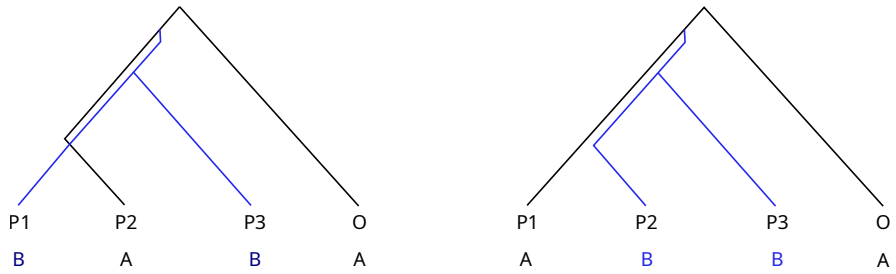

Figure 1.8 – The phylogenies depicted show the possible segregation of ancestral alleles and the occurrence of the ABBA and BABA patterns.

An equal or comparable occurrence of 'ABBA' and 'BABA' patterns is indicative of the absence of gene flow and we can assume that the frequencies we observe are ascribed to random segregation of the alleles. The frequencies are compared by calculating a value named *D*, also named either *D statistics* or *Patterson's D statistics* because it was first proposed by Nick Patterson in Green et al. 2010. The *D* value is computed as follows :

$$D = \frac{nBABA - nABBA}{nBABA + nABBA}$$

where *n* represents the number of 'BABA' or 'ABBA' sites across the alignment. The equation returns a *D* value between -1 and 1. When D is equal to 0, *nBABA* - *nABBA* are equal to 0, otherwise

one of the two patterns has higher frequency than the other one. A standard error for the *D* statistic can then be computed using a block jackknife approach. Briefly, the genome is divided into a number of non-overlapping blocks whose size is larger than the extent of linkage disequilibrium because nearby sites in the genome are expected to have similar histories. A distribution of *D* pseudo values is constructed by measuring the *D* value, each time excluding a single block of the genome in turn. (Green et al. 2010). From the estimate of the average and standard error of *D*, we can compute the *Z* score to test whether *D* significantly deviates from 0. A *Z* score for which $|Z| > 3$ is considered to be significant while a score $|Z| < 3$ is considered to be not significant. It is important to highlight that this test was developed to detect gene flow between P3 and one of P1 and P2, a condition that is reflected by a large disproportion between the number of 'BABA' and 'ABBA' patterns. The null hypothesis tested by the approach is the absence of gene flow between P3 and one of P1 and P2. A $|Z| > 3$ allows the rejection of the null hypotheses of the absence of gene flow, while a $|Z| < 3$ does not allow the rejection of the null hypotheses and the small difference between 'BABA' and 'ABBA' is ascribed to the random segregation of the alleles rather than gene flow.

1.6 Biological implications of the ILS

In recent years, ILS has been detected across various organisms and recognised as the natural outcome of rapid speciation in large populations. In this last paragraph, I summarize evidence of the role of ILS in biology and its relation with genomic elements. Scally et al. 2012 retraced the evolutionary history of humans and the great ape's chimpanzee and gorilla. They reported up to 30% of the loci exhibiting ILS with an equal amount of the discordant genealogies (i.e. *((H, G), C)* and *((C, G), H)*) that contrast the species genealogy *((H, C), G)*. One of the most outstanding results concerns the distribution of the ILS for genic and intergenic regions. Scally et al. 2012 reported a marked drop in the average number of ILS sites within coding exons which extends for hundreds of kbp from the coding gene (Fig.1.9).

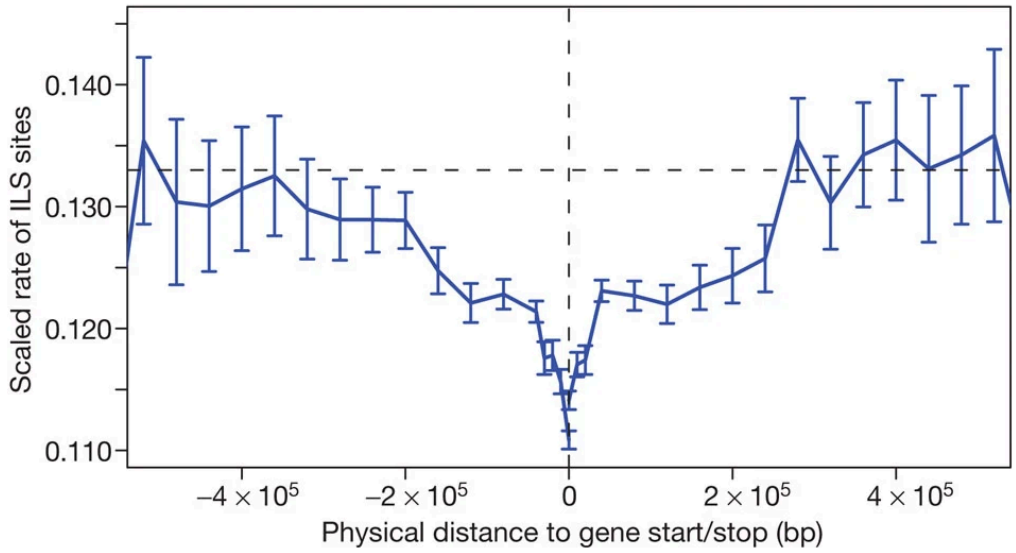

Figure 1.9 – The blue line represents the variation in ILS upstream and downstream of to the nearest gene. The horizontal dashed line is the average value outside 300 kbp from the nearest gene. Adapted from Scally et al. 2012.

The drop observed in ILS sites has been attributed to the effect of linkage that the loci experienced in relation to selected polymorphisms. This effect goes hand in hand with a low *dN/dS* ratio reflecting purifying selection acting on genes which, on average, are expected to evolve under evolutionary constraints. Similarly, Rivas-González et al. 2023, taking advantage of the high-quality assemblies released by Kuderna et al. 2023, extended the study of the impact of the ILS across 50 primate species confirming an overall decrease in the level of ILS in exons compared to intergenic regions. Scally et al. 2012 highlighted a difference in the levels of gene expression between closely related primates, such as chimpanzees and humans, in genes with a relatively high level of ILS compared to those with a low level of ILS. Kuderna et al. 2023 performed a gene ontology enrichment analyses on both genes with low and high ILS revealing that the former are enriched in housekeeping genes with basic cellular functions while genes associated with the immune system showed the highest level of ILS among genes, especially in the major histocompatibility complex, a large locus containing a set of polymorphic genes, maintained by balancing selection, essential for binding fragments of peptides derived from pathogens. Noteworthy is also the lower level of ILS detected by both Scally et al. 2012 and Kuderna et al. 2023, between human and chimpanzee in the former and across several nodes of the primate phylogeny in the latter, on the sex chromosome X compared to autosomes, suggesting a reduced effective population size of chromosome X

in a primate common ancestors. In a recent paper Feng et al. 2022 detected high percentages of ILS among the genomes of the South American monito del monte and Australian marsupials, i.e. *Diprotodontia* (wallaby and koala) and *Dasyuromorphia* (Tasmania devil and Antechinus), and demonstrated by transgenic techniques the direct functional link between an incompletely sorted site and a skeletal morphological characteristic, thereby moving the study of the functional impact of ILS one step forward. Briefly, Feng et al. 2022 first quantified the ILS and highlighted incompletely sorted morphological traits (Fig.1.10 A) and subsequently engineered by CRISPR-Cas9 in a mouse embryo, a non-synonymous point mutation in a gene, showing significant ILS signal, involved in the development of the thoracic vertebrae. The differences that emerged, comparing the T1 and T2 spinous process between wild-type and mutated mice (Fig.1.10 B), provided the proof-of-concept and compelling evidence for the role that the ancestral allele played in the morphological structure of the thoracic vertebrae in marsupials.

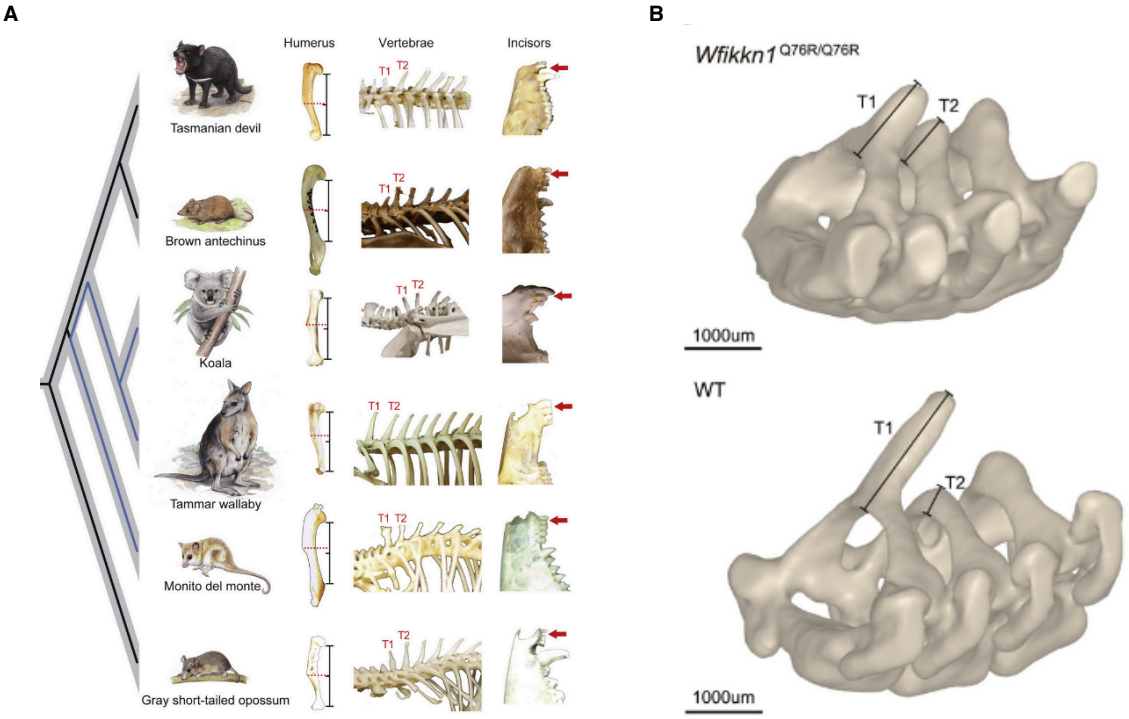

Figure 1.10 – On the left side of panel **A** the overlap between the species phylogeny and the discordant phylogeny that brings monito del monte and the Australian *Diprotodontia* (wallaby and koala) phylogenetically closer. On the right side of panel **A** some morphological characteristics of marsupials reflect incompletely sorted features. In panel **B** the T1 and T2 spinous process of the thoracic vertebrae of mutated and wild type mouse. Note that the mutation results in T1 and T2 having comparable length, the same morphological feature shared by monito del monte and the *Diprotodontia*. Adapted from Feng et al. 2022.

CHAPTER 2

Hybridization and introgression

In this second chapter I describe the process of hybridization and introgression

2.1 Species concept and reproductive isolation

Species are *"groups of actually or potentially interbreeding natural populations which are reproductively isolated from other such groups"* (Ernst Mayr 1942). The term species designates populations reproductively isolated although not a single ubiquitous property can be provided to define a universal concept of species so that several definitions exist (see De Queiroz 2007). The origin of a reproductive barrier is a key event in the speciation process. A reproductive barrier rises as a by-product of divergence. In the simplest scenario, two geographically separated populations evolve apart and fix different alleles, because of drift or adaptation to different niches and thus evolve a degree of reproductive isolation (Ridley 2004), for example, because of incompatible alleles at different loci (Dobzhansky-Muller incompatibilities). Although the reproductive barriers preserve the genetic integrity of a species, over the last decades, several studies on different organisms questioned the idea of rigid reproductive barriers, shifting to a permeable concept of reproductive barriers so that hybridization and gene flow have been emerging as a common aspect of evolution (Taylor & Larson 2019). In this chapter, I will describe the processes of hybridization and introgression.

2.2 Hybridization

Hybridization is the process of interbreeding between individuals from distinct populations or species, an event known as secondary contact. The species interbreed at a geographical location defined as a hybrid zone, producing F1 hybrids, i.e. individuals that inherit half of the genetic material from species and half from the other species, that eventually can generate viable offspring with complex recombined genomes. The secondary contact can be also human-mediated as a consequence of volunteer or accidental introductions of wild populations in a specific environment. For example, since the early 1900s, in the Foreste Casentinesi, Monte Falterona, Campigna National Park, the Camaldolese monks, carried out intensive restocking efforts with selectively bred specimens of Atlantic trout (*Salmo trutta*). These activities have significantly altered the distribution of native trout and caused widespread local extinction of the indigenous populations of Mediterranean trout (*Salmo macrostigma* or *Salmo cettii*) through hybridization. From 2013 to 2018, the LIFE STREAMS project promoted the recovery of native species through the reintro-

duction of Mediterranean trout and the restocking of Italian Apennines rivers (*LIFE STREAMS project* 2023). Although hybridization may accidentally cause extinction (Todesco et al. 2016), in the last years evidence of its role in shaping the evolutionary history of species has been documented, leading to a reevaluation of its role in the evolutionary process (Taylor & Larson 2019). Hybridization may lead to the retention of introgressed adaptive material (adaptive introgression) from an extant or extinct species (ancient hybridization). One example, that closely concerns our species, is the presence of *Neanderthal* (Reilly et al. 2022) and *Denisovan* (Huerta-Sánchez et al. 2014) introgressed DNA in modern *Homo sapiens* populations as well as the recent recovery of "Denny", a first-generation daughter of a Denisovan father and a Neanderthal mother (Warren 2018). The outcome of these secondary contacts is not limited to the examples mentioned. Although uncommon, hybridization can contribute to the speciation event (Hybrid speciation) in which hybrid individuals evolve reproductive barriers against their parental lineages and adapt to a new niche. For example, the homoploid hybridization that involved the sunflowers *Helianthus annuus* and *Helianthus petiolaris* led to the birth of hybrid lineages that are reproductively isolated from their parent and adapted to saline habitats (Christian Lexer 2003).

2.3 Introgression

The introgression is the process of backcrossing of an interspecific hybrid (i.e. a cross of two species from the same genus Khan & Croser 2004) with one of its parental species. Introgression leads to the transfer of genetic material from one species (donor species) into the gene pool of the other (recipient species). The process results in new individuals with a mosaic genome in which the major component is represented by the parental species the hybrid backcrossed with (i.e. the recipient species), while most of the genetic material of the donor species is lost during the backcrossing or negatively selected against, except for loci that eventually contribute to the improvement of the fitness of individuals in the ecological niche they occupy. Thus, introgression represents an evolutionary process of opportunities, in which the donor species provides a wealth of diverged and ready-to-be-used functional haplotypes to the recipient species on which selection can operate. For example, Jones et al. 2018 showed that the seasonal camouflage of *Lepus americanus* is influenced by the introgressed regulatory region upstream the *Agouti* locus (Fig.2.1). *Lepus americanus* moults to white fur during the autumnal period, while some maintain

brown fur. Brown-fur *Lepus americanus* inherited the allelic regulatory region from a black-tailed jackrabbit (*Lepus californicus*) and the allele reaches high frequency in populations of *Lepus americanus* living in a region where snowfall is not heavy, due to a more favourable camouflage that helps to elude predators. Moran et al. 2021 summarizes three emerging principles in the context of purged and maintained DNA of the donor species :

1. Initial rapid and massive loss of the foreign DNA followed by a period of either a slow removal or trim at the boundaries of adaptive alleles. The initial contribution to the donor and the recipient species is expected to be 50% and decreases over time due to the recombination that breaks down the haplotypes. As the hybrid backcrosses with one of the parental species, most of the foreign DNA of the donor species is progressively lost. In addition, genetic incompatibilities between the species are removed over the generations. Veller et al. 2023 predicted the rapid purging to last a few tens of generations in a population of segregating hybrids. Under a backcross scenario, this effect is expected to be more rapid. At late stages, a slow period of purging of the remaining incompatible foreign DNA or trimming of the boundaries of the positively selected haplotypes is excepted primarily as a function of the recombination rate of the species.

2. Genomic regions with functionally relevant roles experience less introgression. Gene-rich genomic regions as well as conserved genomic elements are refractory to introgression. For example, sex chromosomes in humans, *Drosophila* (Presgraves 2008), *Heliconious* (Martin et al. 2013) and mouse (Payseur et al. 2004) are depleted in introgression. In modern humans, there is less Neanderthal introgression closer to coding regions. On the other hand, in yeast, introgressed regions in the Alpechin/Olive *Saccharomyces cerevisiae* clade were enriched in genes that mediate the interaction with the external environment (D'Angiolo et al. 2020) challenging the observations from other organisms.

3. Recombination plays a pivotal role in genome stabilization although evidence that correlates low and high recombination rates with the occurrence of introgression is contrasting and still debated (Moran et al. 2021). Mechanistically, the process is mediated by homologous recombination in which a certain level of sequence homology allows the replacement of some of the loci in the recipient species with their counterpart from the donor species largely resulting in the replacement of orthologous sequences.

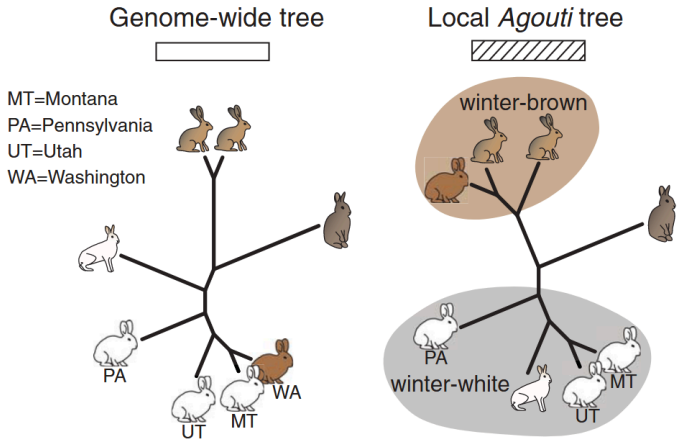

Figure 2.1 – In the left panel, the genome-wide phylogeny of hares : at the top, the silhouette of two black-tailed jackrabbits, at the bottom the silhouettes of four *Lepus americanus* (three white one brown); on the left a mountain hare and on the right a European rabbit. In the right panel, the local topology of the *Agouti* locus. The brown-fur *Lepus americanus* from Washington (WA) branches inside the black-tailed jackrabbit group. The shadows in the background of the clades highlight the hares living in places where, during the winter, the landscape is white because of heavy snowfall (winter white) and landscape where the landscape retains typical autumn colours (winter brown). Adapted from Jones et al. 2018.

2.4 Timing, frequency and duration of introgression

Inferring the timing of introgression allows an understanding of the consequences of the process itself and the evolutionary history of the species involved. For example, Leducq et al. 2016 proposed that the North America *C** lineage (*SpC**) of the yeast *Saccharomyces paradoxus* is the result of the homoploid hybrid speciation between the North America *C* (*SpC*) and the North America *B* lineages (*SpB*) (Fig.2.2).

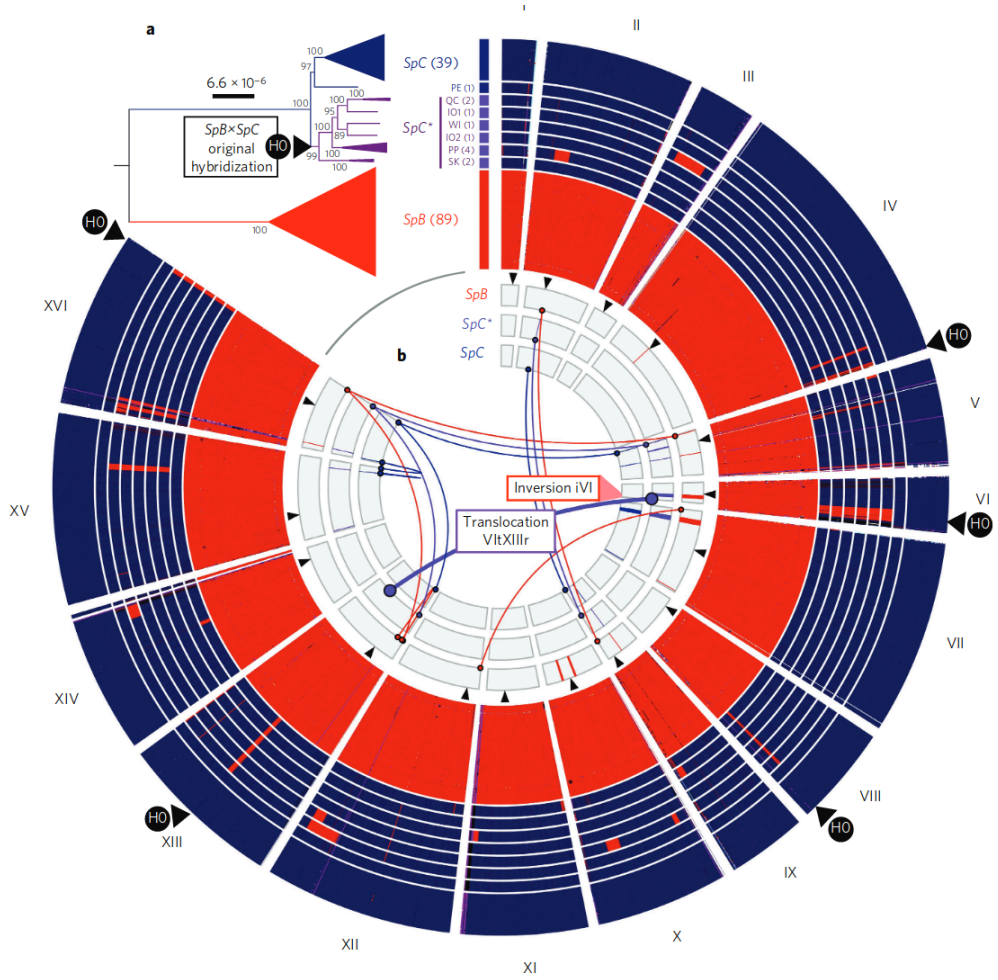

Figure 2.2 – In panel **A** the phylogenetic relationships among North American *S. paradoxus* sublineages (*SpC*, *SpB* and *SpC**). The numbers at the node are bootstraps. H0 the initial hybridization step (*SpC* x *SpB*). On the right of the phylogeny, two-letter codes are geographical locations in North America. The numbers inside the round brackets count the number of samples. The red and blue doughnut plots depict the ancestries of the lineages across the chromosomes. In panel **B** the chord diagram depicts the main structural variants across lineages : translocations, inversions and telomere exchange. Adapted from Leducq et al. 2016.

Hibbins & Hahn 2019 developed a model based on the multispecies network coalescent to determine the timing of introgression and simulated several scenarios that account for homoploid hybrid speciation providing a set of population parameters estimated by the data provided in Leducq et al. 2016 and previous studies in the sister species *Saccharomyces cerevisiae*. Hibbins & Hahn 2019 observed that introgression between *SpC** and *SpB* occurred after the speciation between *SpC* and *SpC**, rejecting the hypothesis that *SpC** is, instead, the result of homoploid hybrid speciation as proposed in Leducq et al. 2016. Thus the timing of the introgression is an interesting

and challenging aspect to evaluate that can contribute to elucidating the evolutionary history of the species.

Another aspect of the introgression process is the frequency and the duration of the admixture, i.e. the number of times that two species admixed along their evolutionary history (referred to as "pulses") and how long such events lasted. For example, reviewing recent discoveries and open questions in the study of archaic hominin admixture Wolf & Akey 2018 highlights that, initially, admixture with Neanderthals occurred only once because non-Africans carry a comparable level of Neanderthal ancestry. However, further studies reported that Neanderthal ancestry varies among Asians, Native Americans and Europeans, with the latter having a lower level of ancestry (0.1-0.5%). This opened the way to further speculation about the number of pulses as well as the possibility that the Neanderthal ancestry was diluted by admixture within modern human populations. The time window within which the introgression may occur is strictly linked to the timing of when two populations established strong reproductive barriers that either fully impair the viability of hybrids' descendants (i.e. making the hybrids sterile) or make the occurrence of viable descendants so infrequent that introgressed individuals become a rare exception. Other reproductive barriers that do not involve genetic factors, for example, behavioural, temporal, geographical or morphological exist too but they will not be discussed further.

2.5 Adaptive introgression

Adaptive introgression represents cases of positively selected ancestry from the donor species that confer a fitness advantage. In populations with large effective population sizes, adaptive introgressions are expected to increase in frequency because natural selection promotes their maintenance in the environment, potentially reaching fixation. At the local scale, a consequence of this sweep is the increase in the allele frequencies at nearby polymorphisms that are in linkage disequilibrium with the adaptive locus promoting hitchhiking. At the same time, a longer persistence of the adaptive introgression may result in differential shortening of the haplotype because of recombination, increasing the heterozygosity at the edges of the adaptive haplotype in the population (Moran et al. 2021). Suarez-Gonzalez et al. 2018 proposed a series of steps for claiming the adaptive value of introgression stating that multiple lines of evidence have to be provided. Although the focus of the paper was on plants, these principles can be applied to virtually all organisms :

1. Identification of the introgressed regions ruling out the persistence of ancestral genetic variation i.e. incomplete lineage sorting ;
2. Detection of a signature of positive selection, i.e. the selective sweep characteristic of a region under selection highlighting its persistence as well as the higher frequency in the population of the introgressed haplotype than expected by chance ;
3. Demonstrating that the inherited haplotype has an adaptive relevant phenotypic variation that provides an advantage in the environment ;
4. Direct measurement of a fitness effect of the introgressed region in the recipient species.

Undoubtedly, establishing the adaptive significance of introgressed DNA is challenging and, in some cases, impossible. For example, species at risk of extinction, described as threatened or otherwise protected, cannot be incorporated into experiments. A full characterization of the genomic elements included in the introgressed haplotype is a prerequisite for experimental tests. In addition, depending on the specific trait, the feasibility of phenotypic tests can dramatically vary. Traits controlled by a single gene or for which a single gene contributes significantly to the phenotypic expression are easier to test than polygenic traits. Some phenotypes, such as the resistance to a drug or the ability to grow in the presence of a specific nitrogen or carbon source, can be easily tested in the laboratory, while other phenotypes may require complex experimental setups. Song et al. 2011 detected a polymorphic region on chromosome 7 of the house mice (*Mus musculus*) embedding the gene *vkorc1* inherited by introgression from the Algerian mouse (*Mus spretus*). The Algerian allele confers resistance to warfarin, an anticoagulant used as a medication but also as rat poison since the 1950s when it was introduced in rodenticides. Over time, the use of such poison selected for house mice with the Algerian allele of *vkorc1* contributed to a selective sweep of the allele.

2.6 Detection of introgression

2.6.1 The fixation index (F_{ST})

F_{ST} is a measure of population differentiation estimated by the expected allele frequencies between two or more subpopulations :

$$\mathrm{F_{ST}} = \frac{H_T - H_S}{H_T}$$

H_T : expected heterozygosity of the whole population ;

H_S : is a weighted average of the expected heterozygosity of all the subpopulations.

The H_T is the heterozygosity expected under Hardy–Weinberg equilibrium. It is calculated based on the allele frequencies of the population after one generation of random mating. Assuming biallelic sites, the genotype frequencies will be p^2, $2pq$, and q^2, where p^2 and q^2 are the frequencies of homozygous genotypes and $2pq$ the frequency of heterozygous genotypes. For example, if we assume a biallelic position, with $p = 0.3$, $q = 0.7$, we can either calculate the heterozygosity as

$$\mathrm{H_T} = 2pq = 2 \cdot 0.3 \cdot 0.7 = 0.42$$

or

$$\mathrm{H_T} = 1 - \sum_{i=1}^{nvariants} p_i^2 = 1 - [(0.3)^2 + (0.7)^2] = 1 - (0.09 + 0.49) = 0.42$$

The H_S is a weighted average of the expected heterozygosity of all the subpopulations. Assuming a simple case of two subpopulations :

$$\mathrm{H_S} = \frac{n_1}{n} \cdot HT_{\mathrm{pop1}} + \frac{n_2}{n} \cdot HT_{\mathrm{pop2}}$$

Let's see the simple case in which we have 50 individuals divided exactly into 25 individuals in each subpopulation and for both the subpopulations $p = 0.3$, $q = 0.7$. We will have :

$$\mathrm{H_T} = 0.42$$

$$\mathrm{H_S} = \frac{n_1}{n} \cdot HT_{\mathrm{pop1}} + \frac{n_2}{n} \cdot HT_{\mathrm{pop2}} = \frac{25}{50} \cdot 0.42 + \frac{25}{50} \cdot 0.42 = \frac{1}{2} \cdot 0.42 + \frac{1}{2} \cdot 0.42 = 0.21 + 0.21 = 0.42$$

So that,

$$\mathrm{F_{ST}} = \frac{H_T - H_S}{H_T} = \frac{0.42 - 0.42}{0.42} = 0$$

This is an extreme case in which $\mathrm{F_{ST}}$ = 0 means complete sharing of genetic material. The two populations share the same allele frequencies. To detect introgression, this value is measured in consecutive windows along the genome and a value close to 0 is expected under a scenario of introgression as the sequences are expected to be almost identical between donor and recipient species. This measure has a clear limitation, in that we are assuming that locally two populations share the same alleles because of gene flow but natural selection can be a confounding factor (Rosenzweig et al. 2016, see Cruickshank & Hahn 2014). In addition, if the introgressed DNA is fixed inside a subpopulation and the same DNA is identical in the donor population, the calculation of $\mathrm{F_{ST}}$ has an obvious mathematical limitation : there is no heterozygosity so that $\mathrm{H_T}$ = 0 and $\mathrm{F_{ST}}$ is not defined. Additionally, if the introgression is recent and did not have enough time to sweep and increase in frequency, the $\mathrm{F_{ST}}$ may fail to detect its presence.

2.6.2 d_{xy}

Given two different species X and Y, d_{xy} is the average number of differences between one sequence randomly chosen from population X and another sequence randomly chosen from population Y. Low values of d_{xy} may indicate a recent introgression. This measure meets the same limitations listed for the $\mathrm{F_{ST}}$ in terms of sequence identity. d_{xy} is given by the formula :

$$\mathrm{d_{xy}} = \frac{1}{n_x \cdot n_y} \sum_{i=1}^{n_x} \sum_{j=1}^{n_j} k_{\mathrm{ij}}$$

where k represents the number of differences between a pair of sequences. The subscripts i and j denote sequences from populations X and Y, respectively, while n_x and n_y the number of sequences sampled from the two populations (Wakeley 1996).

2.6.3 ABBA-BABA test and related statistics

The ABBA-BABA test was previously described in subsection 1.5.2.1. The test was developed to detect gene flow between P3 and one of P1 and P2. Gene flow is reflected by a large disproportion between the number of 'BABA' and 'ABBA' patterns. However the ABBA-BABA test is a genome-wide summary statistic ; it does not inform on the number of introgressed loci or the position, the physical extent or the origin of the introgression. A modified version of this test, named D_3 and D_{FOIL}, have been suggested : the former for accommodating cases in which an outgroup is not available but compared to the ABBA-BABA evaluates both phylogenetic discordance and the branch lengths (under introgression phylogenetic discordance and shorter branch lengths are expected, see Hahn & Hibbins 2019) ; the latter, instead, accommodates a five-taxon phylogeny (Pease & Hahn 2015). Recently Martin et al. 2015, tested the use of the ABBA-BABA test across the genome in consecutive sliding windows highlighting the fact that the test is unreliable for this purpose. The ABBA-BABA test has a large variance when applied to small genomic windows. As an example, we can think about the extreme case in which a single informative site is present and it happens to be an 'ABBA' site, even in the absence of introgression. With nABBA = 1 and nBABA = 0, the D measure is -1.

2.6.4 Statistics for sliding-window approaches

Examples of statistics based on sliding-window approach include f_d (Martin et al. 2015), f_{dM} (Malinsky et al. 2015), and the distance fraction d_f (Pfeifer & Kapan 2019).

2.6.4.1 f_d and f_{dM}

The f_d statistic originated from the ABBA-BABA test and it uses the allele frequencies of ancestral and derived alleles (A and B) instead of the ABBA-BABA counts to identify introgressed loci across the genome, specifically detecting introgression between P2 and P3. In terms of allele frequencies, we can rewrite the ABBA-BABA formula replacing B with p_{ij} and A with 1-p_{ij}, where

p_{ij} is the allele frequency of B (derived allele) at position *i* in the population *j* and A is expressed as complementary frequency that sum to 1. In this case, ABBA for a given site *i* can be written as :

$$\text{ABBA}_\text{i} = (1 - p_{\text{i}1}) \cdot p_{\text{i}2} \cdot p_{\text{i}3} \cdot (1 - p_{\text{i}4})$$

and BABA :

$$\text{BABA}_\text{i} = p_{\text{i}1} \cdot (1 - p_{\text{i}2}) \cdot p_{\text{i}3} \cdot (1 - p_{\text{i}4})$$

The sum of all the ABBA_i and BABA_i is used for the calculation of the D value. Green et al. 2010 proposed to make use of the numerator of the D formula, expressed as :

$$S(P1, P2, P3, O) = \sum_{i=1}^{nalleles} [(1 - p_{\text{i}1}) \cdot p_{\text{i}2} \cdot p_{\text{i}3} \cdot (1 - p_{\text{i}4})] - \sum_{i=1}^{nalleles} p_{\text{i}1} \cdot (1 - p_{\text{i}2}) \cdot p_{\text{i}3} \cdot (1 - p_{\text{i}4})$$

Assuming the following relation among populations : ((P1, P2), P3), O), the fraction of the genome shared between P2 and P3, after the split between P1 and P2, can be estimated by comparing the observed value of *S* to a value of *S* estimated under a scenario of 100% introgression from P3 to P2. In this case, then P2 resembles a lineage of P3 so we can write the formula :

$$f = \frac{S(P1, P2, P3, O)}{S(P1, P3a, P3b, O)}$$

Martin et al. 2015 addressed a few limitations of this implementation by (i) considering P3a, and P3b as identical instead of similar (f_{hom}) and fixing subsequent issues (discussed in Material and Methods in Martin et al. 2015) and by (ii) calculating f defining a new denominator which includes a generic donor population (PD) for each site i.e. :

$$fd = \frac{S(P1, P2, P3, O)}{S(P1, PD, PD, O)}$$

PD is either P2 or P3 and it is chosen alternatively based on the population with the higher frequency of the derived allele (B) so that f_d remains below 1. Although, f_d performs better than D as statistics across consecutive small windows it does only detect gene flow from P2 to P3 or vice versa but not from P1 to P3. To address this limitation Malinsky et al. 2015 proposed a **M**odified version of f_d, $f_{d\mathbf{M}}$. f_{dM}, like D, is symmetrically distributed around zero (if there is no introgression) and quantifies shared variation between P3 and P2 (positive values) or between P3 and P1 (negative values). The calculation of f_{dM} depends on the frequency of the derived allele in P1 and P2 (Malinsky et al. 2021). If the frequency of the derived allele in P2 is higher or equal to P1 then $f_{dM} = f_d$ otherwise, if the frequency of the derived allele in P1 is higher than in P2 then :

$$f_{\mathrm{dM}} = \frac{S(P1, P2, P3, O)}{-S(PD, P2, PD, O)}$$

2.6.4.2 The distance fraction (d_f)

The distance fraction (d_f) combines the pairwise nucleotide diversity d_{xy} and the ABBA-BABA test. Pfeifer & Kapan 2019, considering bi-allelic SNPs only, rewrite d_{xy} as function of the allele frequencies instead of the counts, where B has frequency q and A 1-q, and the same for D and f_d. In addition, Pfeifer & Kapan 2019 incorporates the 'BBAA' pattern into the denominator. Overall, d_f provides an estimate of the amount of introgression and it is less sensitive to the timing of gene flow compared to f_{dM} (Pfeifer & Kapan 2019 and Malinsky et al. 2021).

CHAPTER 3

The Saccharomyces genus

In this third chapter I describe the Saccharomyces genus

3.1 *Saccharomyces* species

The *Saccharomyces* are yeasts, single-cell eukaryotic microorganism members of the fungi kingdom and the *Ascomycota* division. The etymology of the word *Saccharomyces* derives from the Greek roots of *sacchar-* i.e. sugar + *-myces* i.e. fungus (Merriam-Webster.com s. d.) proposed by *J. Meyen* in 1838. The first species classification was based on phenetic data which followed the first molecular methods on DNA leading to the definition of a group of *Saccharomyces sensu lato* and a group of *Saccharomyces sensu stricto*. This terminology is no longer in use, and species boundaries are defined based on evidence of reproductive isolation and genomic divergence, in accordance with the biological species concept and phylogenetic analysis. The genus *Saccharomyces* consist of eight species : *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. jurei*, *S. kudriavzevii*, *S. arboricola*, *S. eubayanus* and *S. uvarum* (Fig. 3.1).

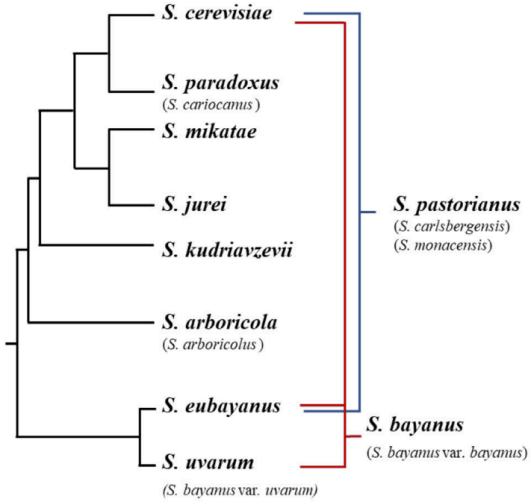

Figure 3.1 – Phylogenetic tree of the main *Saccharomyces* species (in bold). The names in parentheses are the names of the previous classification. The blue and red lines highlight the *Saccharomyces* component of two hybrid strains commonly used in industrial fermentation. *Saccharomyces bayanus* is used in wine-making and cider fermentation while *Saccharomyces pastorianus* for the production of lager beer. Adapted from Alsammar & Delneri 2020.

3.2 *Saccharomyces cerevisiae* genomics and evolution

Saccharomyces cerevisiae is the most known and studied species in the genus. It is vastly exploited in the production of fermented products across the world (Parapouli et al. 2020, Onyema et al. 2023, Molinet & Cubillos 2020) and it is one of the most versatile model organisms across

several fields of research (Duina et al. 2014). Because of this broad interest, it was the first eukaryotic organism to have its genome sequenced in 1996 (type strain S288C). The haploid genome of *Saccharomyces cerevisiae* is $\sim$ 12 million base pairs in length organised in 16 linear chromosomes and dense in intron-free genes ($\sim$ 6,000 genes). The main ploidy of *Saccharomyces cerevisiae* is the diploid (2n) and euploid state. In some industrial environments, polyploids (3n and 4n) and aneuploid isolates are common (Gallone et al. 2016, Al Safadi et al. 2010). In 2009, the initial high-quality whole-genome sequencing project using short-read technology for 36 isolates identified five distinct populations (Liti et al. 2009). In 2015, a larger whole-genome sequencing project for 100 isolates reported a collection of isolates from clinical samples with mosaic genomic compositions, underscoring the opportunistic role of *S. cerevisiae* in clinical environments (Strope et al. 2015). In 2018, a second large-scale whole-genome sequencing project (Peter et al. 2018) scaled up to 1,011 *S. cerevisiae* isolates divided in 26 main populations (Fig.3.2 B). The previous discovery of wild Chinese isolates (Q.-M. Wang et al. 2012) and the large sequencing effort of isolates from a broad ecological and geographical source support a single out-of-China origin of the species (Peter et al. 2018). The same year a third large-scale whole-genome sequencing effort introduced several *S. cerevisiae* from typical Asiatic fermented food and beverages and from primaeval and secondary forests in China and far East Asia (Duan et al. 2018). These latter wild isolates gave a strong contribution to the known genetic diversity of the species (Peter et al. 2018, Duan et al. 2018) and support China and in general far East Asia as the centre of origin for the species. Moreover, Duan et al. 2018 provided evidence for the onset of domestication of the species in China and proposed a nearly neutral model of evolution for wild populations of *S. cerevisiae*. This hypothesis is supported by 1) the absence of genomic domestication signatures in wild clades (De Chiara et al. 2022), 2) their strong population structure along with high sequence divergence (Duan et al. 2018, Q.-M. Wang et al. 2012) and 3) the absence of evidence of admixture and scarce positive selection (Duan et al. 2018) even though wild populations living in sympatry are excellent sporulators i.e. to reproduce meiotically (Bai et al. 2022, De Chiara et al. 2022). Also, numerous studies suggest that structural variations (SVs) play a role in the onset of reproductive isolation in *S. cerevisiae* (Q.-M. Wang et al. 2012, Liti et al. 2006, Hou et al. 2014). The recent release of 142 telomere-to-telomere *de novo* assemblies (O'Donnell et al. 2022), a renewed interest in the investigation of the ecology and genomics in the wild (Mozzachiodi et al. 2022, Peris et al. 2023) along with the shift of the analytical paradigm from the use of a single reference genome to genome

graphs (Garrison et al. 2018, Eizenga et al. 2020) will improve our understanding of the processes that shaped the evolution of the species in the genus.

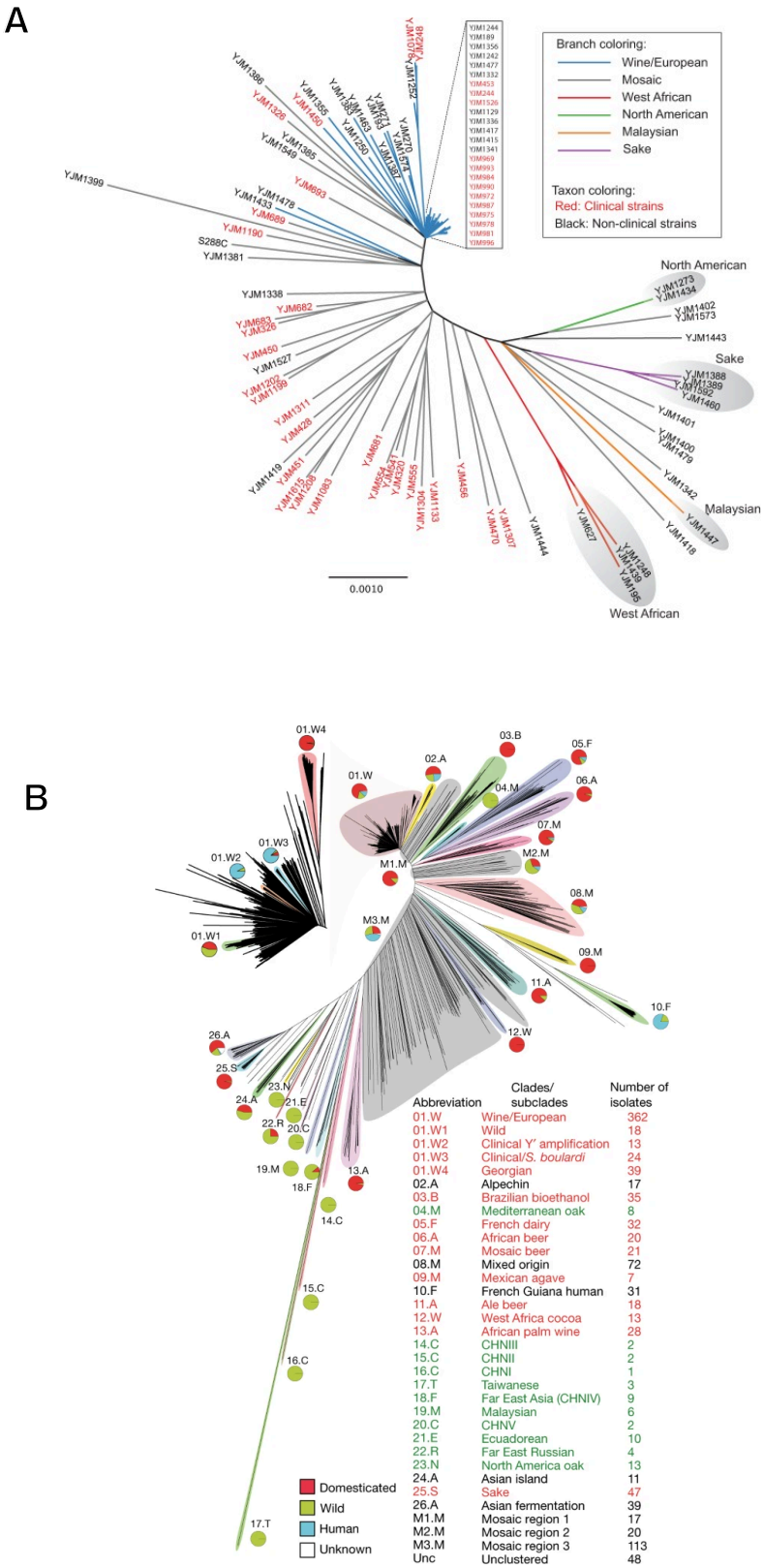

Figure 3.2 – Neighbor-joining trees of the 100 (**A**) and 1011 (**B**) *S. cerevisiae* collections. Adapted from Strope et al. 2015 and from Peter et al. 2018.

3.3 *Saccharomyces paradoxus*

Saccharomyces paradoxus, the sister species of *Saccharomyces cerevisiae*, inhabits temperate forests of the northern hemisphere and lives associated with plants of the order *Fagales* such as beech, oak and birch. Its distribution is confined to North America (G. I. Naumov 1998, Leducq et al. 2016, Eberlein et al. 2019), Hawaii (G. I. Naumov 1999), Europe (Naumov 1986) and Asia/Far East Asia (G. Naumov et al. 1993). The preference for lower temperatures might explain its absence in subtropical and tropical forests (Robinson et al. 2016). *S. paradoxus* presents an overall strong population structure with no evidence of admixture across continents (Liti et al. 2009), no positive selection (Mousseau et al. 2020) and no signature of domestication (He et al. 2022, Fig. 3.3). Migration of isolates across the ocean is likely to have been human-mediated and, in any case, restricted to European strains collected in North America (named *SpA* Leducq et al. 2016). The genomes of *S. paradoxus* and *S. cerevisiae* are highly collinear and diverged at a level of around 12%. The only exceptions are represented by a few lineages, in both species, where structural events drastically rearranged the genome (Yue et al. 2017). In *S. paradoxus*, this includes UFRJ 50816, a strain sampled in Brazil previously known as *S. cariocanus* (G. I. Naumov et al. 1995), and UWOPS.91-917-1, a strain from Hawaii (Yue et al. 2017), while in *S. cerevisiae* massive rearrangements are found in isolates of the Malaysian population (Yue et al. 2017).

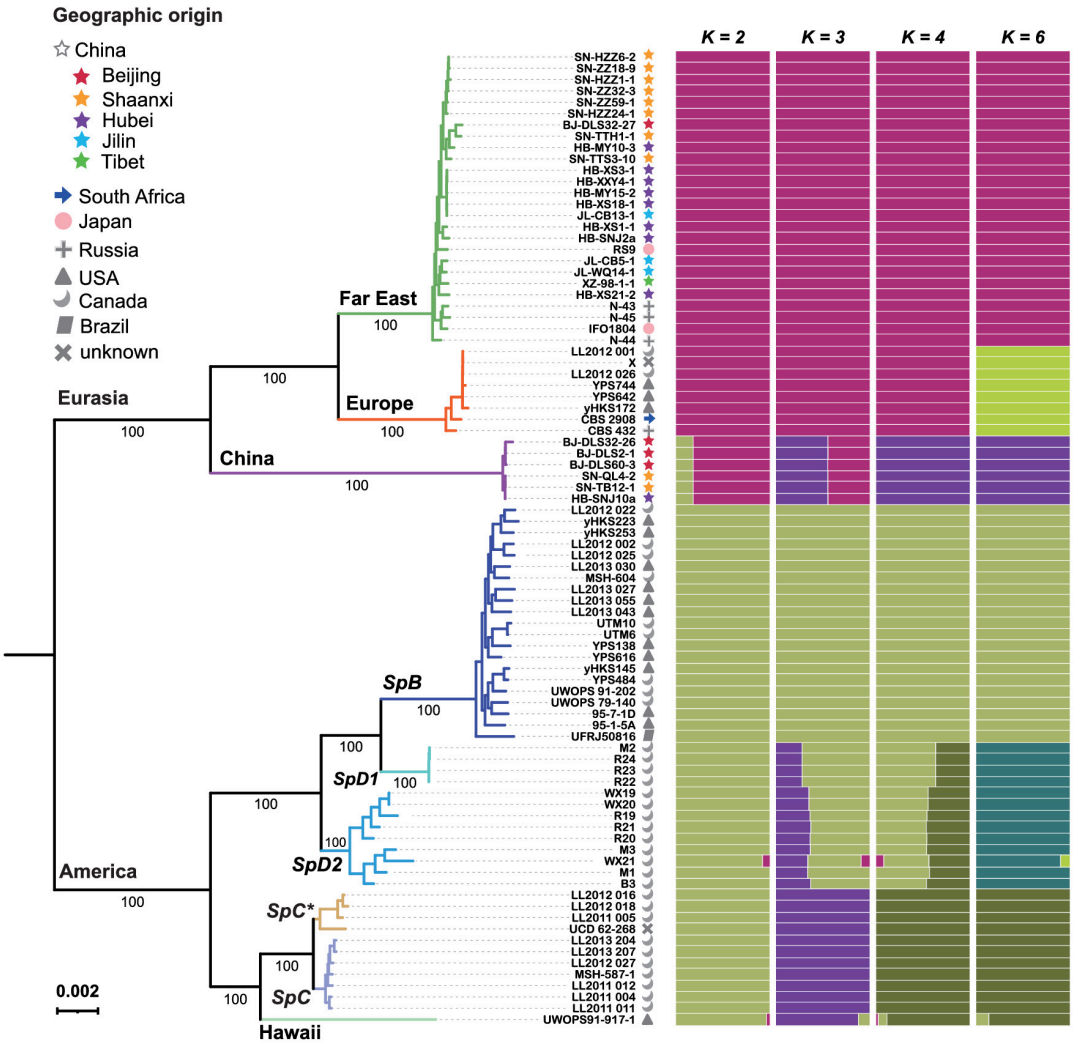

Figure 3.3 – Maximum likelihood tree and ADMIXTURE plot of main *Saccharomyces paradoxus* populations. Adapted from He et al. 2022.

3.4 *Saccharomyces* life cycle

Saccharomyces yeasts exist mainly as haploid (n) and diploid (2n) cells, with the diploid state being the more common. Sex determination is established by a single locus on chromosome III, known as the mating type locus. Haploid cells harbour either the **a** or the α mating locus and can reproduce asexually through the process of mitosis (Fig. 3.4). Alternatively, haploid cells with opposite mating type can mate to form a diploid cell (Fig. 3.4). The cell that commits to asexual reproduction, known as the mother cell, produces a daughter cell through a process known as

budding, where the daughter cell forms and then buds off from the mother cell (Herskowitz 1988). The asexual reproduction is common under favourable nutritional conditions and it is the main way *Saccharomyces*, both haploids or diploids, form colonies. Depending on environmental conditions and genetic factors, a single mother cell can produce more than one daughter cell. In response to adverse conditions such as starvation, a diploid cell undergoes sporulation. The sporulation reduces the chromosome content to a haploid state through the meiotic process. A single diploid cell undergoes meiosis to form four haploid spores (gametes), i.e. a tetrad, each with opposite mating types in pairs : two **a** and two α gametes (Fig. 3.4). The spores are enclosed inside a structure called *ascus*. When conditions become favourable, the spores germinate and may follow different directions (Ono et al. 2020).

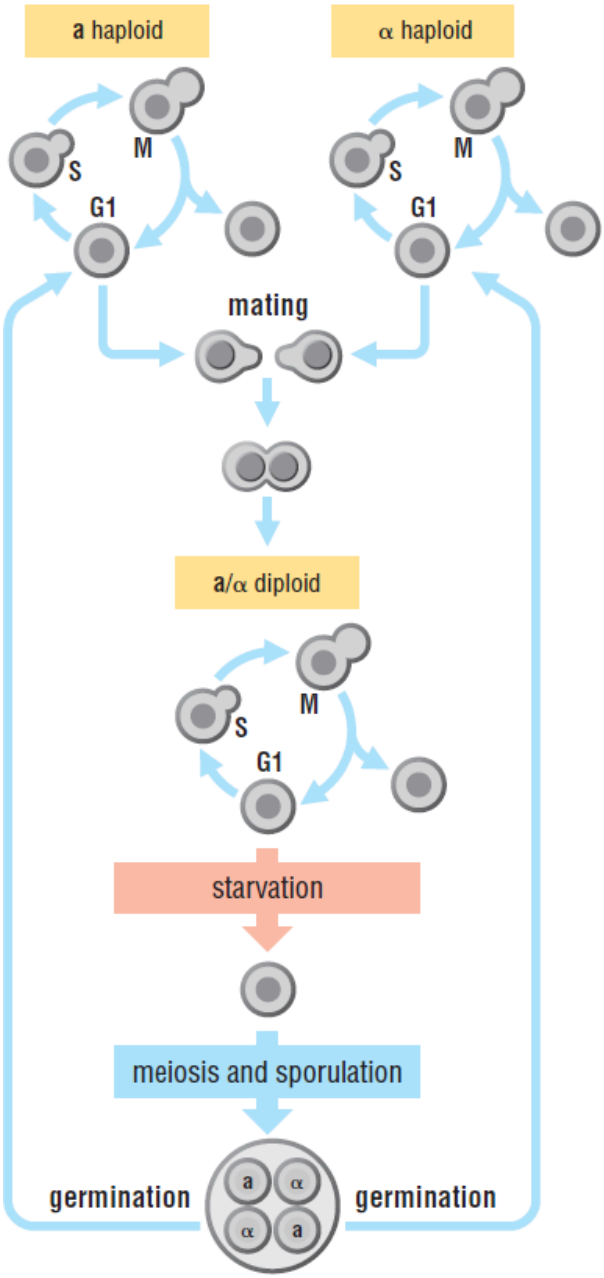

Figure 3.4 – Schematic of *Saccharomyces* life cycle. Adapted from Morgan 2007.

The germinated spores often mate inside the *ascus*, reestablishing the diploid state, a process known as intra-tetra mating (Fig. 3.5). Alternatively, the germinated spores are released outside the *ascus* and return to haploid vegetative growth. After a mitotic cell division, the haploid mother cell can switch its mating type (homothallism ability) and mates with its daughter cell (autodiploidization or haploselfing) to form a diploid cell (Fig. 3.5). Rarely, the germinated spores mate

with germinated spores from a different *ascus* (inter-tetrad mating, Fig. 3.5). Even more rarely, the inter-tetrad mating involves germinated spores from *asci* originating from different *Saccharomyces* species (Fig. 3.5) (Tsai et al. 2008). This latter process is known as outcrossing and is fundamental to the interspecies hybridization and introgression processes.

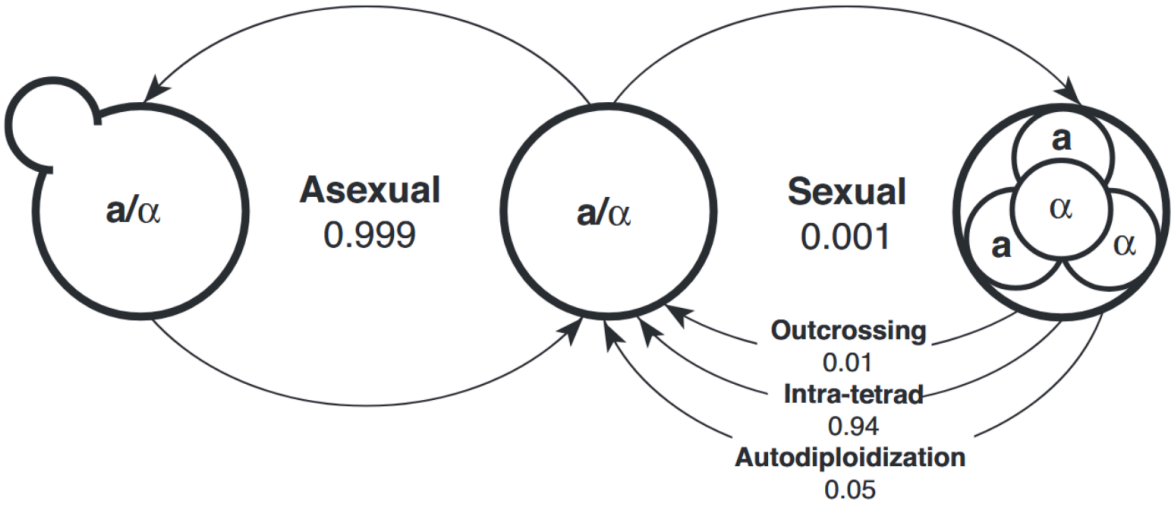

Figure 3.5 – Life cycle quantification for the European *S. paradoxus*. Adapted from Tsai et al. 2008.

3.5 *Saccharomyces* reproductive isolation

The process of introgression requires the overcoming of several reproductive barriers between species. Both pre-and post-zygotic barriers either hamper spore mating or reduce/abort spore viability. One example of a pre-zygotic barrier is the geographical isolation between different species. The occupation of distinct (ecologically different) or separated (physically distant) niches prevents physical contact between species, leading to reproductive isolation. Although long-term geographical separation contributes to allopatric diversification, this does not preclude the possibility that if the geographical barrier is removed or if species are forced to mate, they can produce viable offspring. As it is commonly observed in the *Saccharomyces* genus, *Saccharomyces* species can be forced to mate resulting in viable individuals.

A high variation in spore germination timing, leading to asynchronous spore germination in genetically identical natural isolates of *Saccharomyces paradoxus*, has been shown in the laboratory under controlled conditions (Stelkens et al. 2016). Stelkens et al. 2016 demonstrated that

asynchronous spore germination might serve as a bet-hedging mechanism, allowing yeast populations to survive in the wild where environmental changes can occur abruptly. The variation in spore germination timing (Maclean & Greig 2008) and how it is affected under different environmental conditions (Plante & Landry 2021), represents both an example of temporal and ecological pre-zygotic barrier, phenomenons not extensively explored in *Saccharomyces* to date. In addition, factors such as mating pheromones, mating-type-specific peptide signals, and adhesin proteins are not considered to significantly influence species recognition or discrimination in *Saccharomyces* mating partner choice (Ono et al. 2020).

On the other hand, post-zygotic barriers are well characterised and recognized across the whole *Saccharomyces* genus. Although *Saccharomyces* species form viable hybrid individuals, the hybrids are completely sterile or poorly fertile. The low spore viability is the result of different, sometimes concurrent, mechanisms that determine post-zygotic barriers : errors of failure in chromosome segregation, structural chromosomal rearrangements, and genetic incompatibilities (nuclear-nuclear or nuclear-mitochondrial incompatibilities) (Ono et al. 2020).

3.5.1 Chromosomal missegregation, structural rearrangements and genetic incompatibilities

Hybrid spore inviability is often due to improper chromosome segregation. Following meiosis, inviable spores contain aneuploid chromosomes and the lack of chromosomes in haploid spores leads to their inability to germinate. These missing chromosomes are inherited by other spores from the same *ascus*, during meiosis I (Ono et al. 2020). At the molecular level, the missegregation is the result of incorrect crossover. The mismatch repair system of the cell prevents recombination between divergent or erroneously aligned sequences maintaining the genome integrity (Hunter et al. 1996). Thus, the mismatch repair system contributes to preventing recombination and increasing chromosome non-disjunction between divergent sequences. In addition to missegregation, chromosomal structural rearrangements reduce spore viability. A single translocation can compromise the viability of at least one spore out of four since it will be missing a copy of the genes that have been translocated, rendering the spore inviable. An example can be observed in the *Saccharomyces paradoxus* species. The *Saccharomyces paradoxus* population sampled from Brazil is reproductively isolated from the North American *Saccharomyces paradoxus* po-

pulation. The sequence of the DNA showed that the two populations are only 0.1% divergent and that chromosomal rearrangements, in the Brazilian *Saccharomyces paradoxus* population, are at the source of the sterility. Other genetic incompatibilities not extensively studied yet are the Bateson-Dobzhansky-Muller Incompatibilities (BDMIs). BDMIs include allelic incompatibilities between loci in the nuclear and mitochondrial genome.

3.6 *Saccharomyces* hybrids

Saccharomyces hybrids are common in industrial environments. The most studied hybrids were historically named *S. bayanus* and *S. pastorianus*. These names were maintained as they are still commonly used in the industry (Fig. 3.1). In 2011, Nguyen et al. 2011 disclosed the mosaic composition of the type strain *S. bayanus* CBS380^{T} by sequencing a set of genes. From the comparison with publicly available *Saccharomyces spp.* sequences, they found out that CBS380^{T} presented : a component of *S. uvarum*, an unknown species that was provisionally named *S. lagerae* (later assigned to *S. eubayanus* Libkind et al. 2011), and *S. cerevisiae*. The industrial hybrid *S. pastorianus*, previously know as *Saccharomyces carlsbergensis*, is used for brewing lager. The ability to ferment at low temperatures is the result of the hybrid vigour arising from its parental species i.e. *Saccharomyces cerevisiae* and *Saccharomyces eubayanus*. Specifically, *Saccharomyces cerevisiae* contributes to *S. pastorianus*' ability to ferment maltotriose while *Saccharomyces eubayanus* contributes to *S. pastorianus*' ability to grow at low temperatures. The origin of the hybrid is still debated. The scarce evidence of *Saccharomyces eubayanus* in the European continent contrasted the wide availability of *S. pastorianus*, since its first description in the 14th century in breweries, opening a debate on the first hybridization event. In 2022, *S. eubayanus* was sampled in Ireland (Bergin et al. 2022), nevertheless not a single *S. eubayanus* component matches the *S. eubayanus* subgenome of *S. pastorianus*, which instead resembles a mosaic of different *S. eubayanus* ancestries (Fig. 3.6). This evidence contrasts with the hybridization between *S. cerevisiae* and a single wild lineage of *S. eubayanus*, leaving open several questions about the origin of *S. pastorianus*.

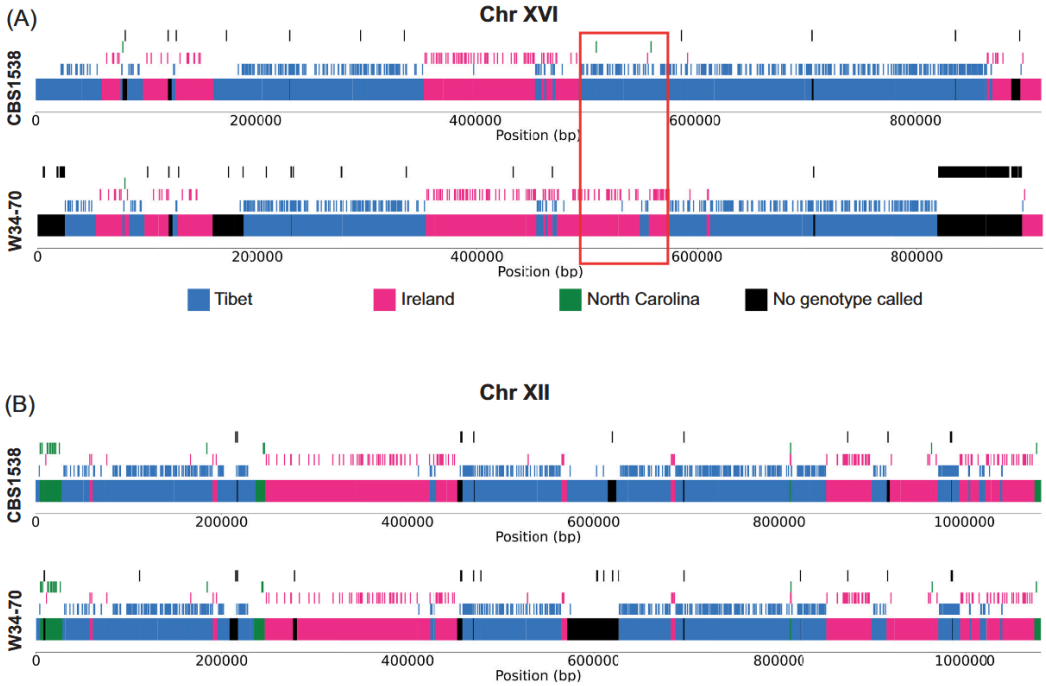

Figure 3.6 – Different *S. eubayanus* ancestry in *S. pastorianus* across chromosome XVI (panel A) and chromosome XII (panel B). Different colours correspond to different ancestries. The red rectangle shows a marked difference between the strain CBS 1538 and W34-70. The two strains are members of two groups of *S. pastorianus*. CBS 1530 is a Saaz isolate of *S. pastorianus*; W34-70 a Frohberg isolate of *S. pastorianus*, two different strains that greatly differ in ploidy, chromosome content and structure (Alsammar & Delneri 2020). Adapted from Bergin et al. 2022.

The research on new flavours and properties promoted the production of new commercial hybrid strains. For example, VIN7 is a broadly used hybrid between *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii* used in wine fermentation. *S. cerevisiae* x *S. kudriavzevii* hybrids show drastic differences in ploidy and the amounts of *S. kudriavzevii* ancestry (Erny et al. 2012). Borneman et al. 2016 analysed the whole genomes of a group of industrial strains from the Australian Wine Research Institute (AWRI), bringing to the light a few *Saccharomyces* complex hybrids including rare crosses between *S. cerevisiae* and *S. paradoxus* and *S. cerevisiae* and *S. mikateae*, a couple of strains obtained in the laboratory and used during the *prise de mousse* i.e. during the creation of bubbles in sparkling wine. Gallone et al. 2019 reanalyzed the whole genome of industrial hybrid isolates exploited in brewing. It turned out that their subgenomes adapted to different industrial niches and originated from specific subclades. For example, the *S. cerevisiae* subgenome of *S. pastorianus* originates from the Beer 1 subclade of *S. cerevisiae*, the same *S. cerevisiae* isolates involved in ale beer production, while in the hybrid *S.cerevisiae* x *S. kudriavzevii* VIN7 the

S. cerevisiae subgenome originates from a wine European isolate. In most of the *S. cerevisiae* x *S. kudriavzevii* hybrids involved in the production of Belgian traditional beer, the *S. cerevisiae* subgenome creates a monophyletic group (within the Beer 2 subclade). Members of the Beer 2 clade are found in Lambic and Trappist beers : the former are known for their spontaneous fermentation and complex flavours, while the latter is brewed by monks, using different methods and coming in a large variety of styles. In a parallel project Langdon et al. 2019 arrived at similar conclusions. Early in 2023 a complex hybrid zone between *S. cerevisiae* and *S. paradoxus* has been identified in the fermentation environment of the Mexican agave, used for the production of mescal and tequila in Mexico : a process that is carried out as spontaneous fermentation i.e. without the use of a starter (data unpublished). This is the first evidence of such massive *S. cerevisiae* and *S. paradoxus* hybridization in fermenting anthropic environments. Previously, the only available evidence of the presence of a "hybrid zone" in an anthropic environment between *S. cerevisiae* and *S. paradoxus* was the characterization of a hybrid sampled from olive oil wastewater (D'Angiolo et al. 2020) along with *S. cerevisiae* isolates with evidence of introgression collected from the same environment, even though the manufacturing process of olive oil does not require a fermentation process (Pontes et al. 2019). More recently, the addition of the species *S. jureii* to the genus opened the way to the investigation of new aromas in craft beer production by crossing *S. cerevisiae* and *S. jureii* (Giannakou et al. 2021). In wild environments, hybrids are rare. The most convincing evidence of a wild interspecies hybrid was described in Barbosa et al. 2016. Two *S. cerevisiae* x *S. paradoxus* hybrid samples were collected from a non-recognised plant in the National Park Foz do Iguaçu in Brazil. It may be worth noticing that no wild *S. paradoxus* isolates were reported, suggesting either a migration of these strains or the presence of a yet unsampled *S. paradoxus* population.

3.7 Hybridization and introgression between *S. cerevisiae* and *S. paradoxus*

S. cerevisiae x *S. paradoxus* hybrids are viable and successfully reproduce mitotically. On the other hand, these hybrids are largely sterile with less than 1% spore viability (Hunter et al. 1996). The sequence divergence between these two species (≈12%) is the main factor that impairs the spore viability. The anti-recombination barrier, mediated by the mismatch repair system, prevents the proper pairing and crossover between homologous chromosomes. The resulting spores show

frequent aneuploidies and low levels of recombination (Hunter et al. 1996). However, despite this reproductive isolation, an introgression event was reported, in 2007, on chromosome I of *Saccharomyces cerevisiae* strain YJM789. The fragment originates from *Saccharomyces paradoxus* and encompasses five genes encoding for putative integral membrane proteins (Wei et al. 2007). The year before, Liti et al. 2006 described an introgressed region at the beginning of chromosome XIV in the European *Saccharomyces paradoxus* population from *Saccharomyces cerevisiae*, suggesting that the introgression process occurred bidirectionally between the two species. The 100-cerevisiae collection revealed introgressed DNA of variable size with 318 ORFs from *S. paradoxus*. In particular, the strains YJM1078, YJM1252, and YJM248 counted more than two hundred introgressed genes each, the majority of which were shared, indicating a recent common origin (Strope et al. 2015). More recently, introgression has been reported to be a common process that impacted the evolutionary history of different populations of *Saccharomyces cerevisiae* (Duan et al. 2018, Barbosa et al. 2016, Pontes et al. 2019, Peter et al. 2018). The *Saccharomyces* pangenome of the 1011 collection includes 913 introgressed ORFs, unambiguously assigned to *S. paradoxus*, out of 6,081 non-redundant non-reference ORFs. Among the 26 *S. cerevisiae* clades, four showed high levels of introgression and these were all associated with human activity. Notably, Peter et al. 2018 reported a strong match between the geographic origins of the *S. cerevisiae* and the corresponding regional population of *S. paradoxus*. The introgression in the Brazilian Bioethanol, Mexican agave and French Guiana *S. cerevisiae* populations were ascribed to the American *S. paradoxus* population while introgression in the European Alpechin clade was ascribed to the European population of *S. paradoxus*. From a mechanistic perspective, two possible routes have been proposed to overcome the hybrid sterility and initiate the introgression process, in yeast. In 2020, D'Angiolo et al. 2020 proposed genome instability as a trigger of structural recombination and recovery of spore viability while. In 2021, Mozzachiodi et al. 2021 showed that a switch from the meiotic cell cycle to the mitotic cell cycle before the cells overcome the commitment point to complete meiosis, results in recombined diploid hybrids. In the last few years, the massive effort carried out by the scientific community in detecting introgression unveiled how commonly the process impacted the evolutionary history of species although the evidence for the adaptive value of this process is still scarce(Taylor & Larson 2019). Also in *S. cerevisiae*, one of the model organisms with the largest genomic toolbox, the adaptive value of the introgression remains unexplored. Future studies of introgression require efforts in understanding the functional implications of introgression (Suarez-

Gonzalez et al. 2018), the ecological context in which hybridization and introgression occur and persist (Porretta & Canestrelli 2023) and the role played in driving the evolutionary trajectory of a population (Pontes et al. 2019, Kleindorfer et al. 2019).

CHAPTER 4

Ancient and recent origins of shared polymorphisms in yeast

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

1. Institute for Research on Cancer and Aging, Nice, 2. Department of Animal Medicine, Production and Health, University of Padova, 3. Kurchatov Complex for Genetic Research (GosNIIgenetika), National Research Center "Kurchatov Institute", 4. University of Gothenburg, 5. School of Biological Sciences, University of East Anglia

4.1 Preamble

This chapter is adapted from :

Tellini Nicolò, Matteo De Chiara, Simone Mozzachiodi, Lorenzo Tattini, Chiara Vischioni, Elena Naumova, Jonas Warringer, Anders Bergström, and Gianni Liti. *Ancient and recent origins of shared polymorphisms in yeast. (2023), Research Square*.

A peer-reviewed copy of the following manuscript has been accepted for publication under the DOI : https://doi.org/10.1038/s41559-024-02352-5.

4.2 Abstract

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

4.3 Main

Genetic variation segregating in natural populations enables inference of species histories and past demographics. Extant populations diverge from ancestral progenitors by accumulating *de novo* variation that is exposed to both selection and neutral evolutionary forces. One source of ge-

netic variation among extant populations is represented by ancestral polymorphisms that arose before populations diverged. Concisely, the random segregation of ancestral neutral polymorphisms can lead to incomplete allele sorting (ILS), resulting in individuals from a population sharing alleles with a non-sister population (Nei et al. 2010, Schrempf & Szöllösi 2020, Sousa & Hey 2013). The probability of observing ILS can be quantified by the coalescence theory and depends on both the effective population size (N_e) and the speciation time ($T_{speciation}$) (Rannala et al. s. d., Kingman 1982). In simple terms, the larger the N_e and the shorter the $T_{speciation}$, the greater are the chances of ILS (Maddison 1997). An alternative source of shared alleles between non-sister populations is introgression, initiated by hybridisation (Harrison & Larson 2014), which is pervasive across the eukaryotic tree of life with potential adaptive or deleterious outcomes (Taylor & Larson 2019, Suarez-Gonzalez et al. 2018). The hybridization event is often followed by successive rounds of backcrossing to one parental population, leading to the depletion of the other population subgenome. Introgression blocks patterns are shaped by a complex relationship of selection and recombination properties (Moran et al. 2021, Martin & Jiggins 2017). Although both ILS and introgression result in shared polymorphisms between non-sister populations (Steenwyk et al. 2023), they emerge from processes unfolding at fundamentally different timescales, with introgression being more recent. Genomic surveys of modern human populations (Bergström et al. 2020), archaic humans (Green et al. 2010, Slon et al. 2018) and great ape species16 have elucidated the origin of genetic variation in our species. ILS has shaped the genetic relationships between our genomes and those of our closest relatives, the bonobo and the common chimpanzee (Mao et al. 2021). Introgression from archaic *Neanderthal* and *Denisovan* humans has introduced divergent haplotypes persisting in present-day human populations (Sankararaman et al. 2016). Some of this introgressed variation has contributed to local adaptation, including the adaptation to high altitude in Tibetans (Huerta-Sánchez et al. 2014). The model organism *Saccharomyces cerevisiae* and its closest relative *Saccharomyces paradoxus* diverged around 4.0−5.8 million years (My) ago (Shen et al. 2018). Even though these species are reproductively isolated, introgressed DNA from *S. paradoxus* has been detected in few *S. cerevisiae* clades (Peter et al. 2018, Duan et al. 2018, Ono et al. 2020). These events support at least two distinct pulses of introgression, between the American and between the European populations of the two species, respectively. However, these analyses have mostly relied on fragmented *de novo* assemblies derived from short reads or on mapping to a single reference, approaches which have not allowed the compilation of

a high-resolution catalogue of introgressed DNA and inference of their origins. Furthermore, experimental demonstration of the functional implications of introgressed material in budding yeast is still lacking (Clark et al. 2022, Barbosa et al. 2016, Pontes et al. 2019). The occurrence of ILS in the *Saccharomyces* genus has been recently suggested (Peris et al. 2023) but not formally demonstrated. Given the ancient *S. cerevisiae* ˘ *S. paradoxus* split (estimated time of divergence $3.27x10_8$ generations), it seems unlikely that ILS polymorphisms would persist at such a timescale given the short generation time and limited effective population size (Clark et al. 2022). Recent studies described *S. cerevisiae* and *S. paradoxus* isolates collected worldwide and generated both short-read datasets (Peter et al. 2018, Duan et al. 2018, Pontes et al. 2019, Gallone et al. 2016, Barbosa et al. 2018, Gonçalves et al. 2016, Legras et al. 2018, Ramazzotti et al. 2019, Coi et al. 2017, Almeida et al. 2015) and high-quality whole-genome assemblies (Yue et al. 2017), paving the way for a deep investigation of shared alleles across these two species. Here, we exploit the thousands of *S. cerevisiae* genomes and the dense map of diagnostic markers between the two species to detect *S. paradoxus* alleles at high resolution. We identify and describe the major hybridisation events and report that most introgressed blocks derive from recent clade-specific hybridisations, while a handful of ancient introgressions are shared across distinct clades. We detect rare instances of recurrent introgression blocks and experimentally demonstrate the functional effect of *S. paradoxus PAD1-FDC1* alleles. We also formally test and quantify abundant ancestral ILS polymorphisms in the *S. cerevisiae* Chinese lineages, providing novel insights into the evolutionary processes that shaped their genetic variation.

4.4 Results

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

We developed a genotyping framework based on a dense map of single nucleotide diagnostic markers to explore the landscape of *S. paradoxus* alleles segregating in the *S. cerevisiae* populations (Methods and Supplementary Fig. 4.6). We constructed a *S. cerevisiae* consensus (S.c.c.) genome by taking the most common allele among a set of strains representing the major phylogenetic clades within the species (Supplementary Table 1). We then ran pairwise whole-genome alignments of the *S.c.c.* genome against the five reference genome assemblies of the major *S. paradoxus* populations, to identify a set of 755,500 biallelic species-diagnostic single-nucleotide

polymorphic markers (Fig. 4.1a). Finally, we mapped short-read sequencing data derived from 1,673 *S. cerevisiae* strains, onto the *S.c.c.* and European *S. paradoxus* (CBS432) genomes and genotyped the diagnostic markers (Fig. 4.1b, Supplementary Fig. 4.7 and Supplementary Table 2-3). We found that the strains of the Alpechin clade, characteristic of the olive oil-based environment, contain the highest number of *S. paradoxus* markers (mean $28,439 \pm 3,455$, single standard deviation; Fig. 4.1b), consistent with deriving from a recent hybridisation (Peter et al. 2018, Pontes et al. 2019, D'Angiolo et al. 2020). The Mexican Agave and the South American Mix II clades also showed high levels of *S. paradoxus* markers ($18,598 \pm 5,459$ and $10,010 \pm 3,075$ respectively; Fig. 4.1b), consistent with secondary contacts between *S. paradoxus* and multiple North and South America *S. cerevisiae* populations (Peter et al. 2018, Barbosa et al. 2016). Surprisingly, the highly diverged Chinese lineages (CHN-IX/Taiwanese, CHN-I and CHN-II) contain a large number of *S. paradoxus* markers ($29,716 \pm 148$; $18,253 \pm 501$; $16,690 \pm 593$ respectively; Fig. 4.1b). Indeed, the number of *S. paradoxus* markers in the CHN-IX clade are comparable to the Alpechin clade, despite an overall lack of evidence for introgression (Peter et al. 2018, Duan et al. 2018). The sole exception is a 24 kb region on chromosome XI, which appears to have been introgressed from an unknown and possibly extinct *S. paradoxus* sister lineage (Supplementary Fig. 4.8). We grouped consecutive *S. paradoxus* markers into blocks to define their sizes and boundaries (Supplementary Fig. 4.9, Supplementary Fig. 4.10 and Supplementary Table 4). We then counted the number of blocks, as well as the number of isolated *S. paradoxus* markers and measured the fraction of each *S. cerevisiae* genome that was covered by these two groups of *S. paradoxus* variants (Fig. 4.1c). Strains in the Alpechin clade had the highest fraction of their genome ($3-5.9\%$) included in large introgression blocks but a relatively small number of isolated markers. In contrast, the Chinese lineages CHN-IX, CHN-I and CHN-II had very few introgression blocks, but a huge number of isolated *S. paradoxus* markers. These were scattered across the genomes, with no clear spatial clustering, and covered a much lower fraction of the genome ($0.56-0.62\%$, $0.29-0.33\%$ and $0.23-0.28\%$ respectively). For example, the AHL Alpechin and the AMH CHN-IX strains have comparable numbers of *S. paradoxus* markers (28,566 and 29,826 respectively), but drastically different median inter-marker distances (11 bp vs. 194 bp) (Fig. 4.1d). In other words, *S. paradoxus* markers in AHL clustered into fewer and larger blocks (e.g. 169 blocks ≥ 5 *S.p.* markers) that covered 4.8% of the genome while AMH showed many isolated markers and fewer large blocks (76 blocks ≥ 5 *S.p.* markers), covering only 0.1% of the genome (Fig. 4.1e and Supple-

mentary Fig. 4.10). The difference is even more striking considering that 64 out of the 76 AMH blocks mapped to the single ≈ 24 kb introgressed block on chromosome XI (Supplementary Fig. 4.8). These fundamentally different patterns in the genomic distribution of *S. paradoxus* markers across *S. cerevisiae* populations suggest different mechanistic origins and evolution.

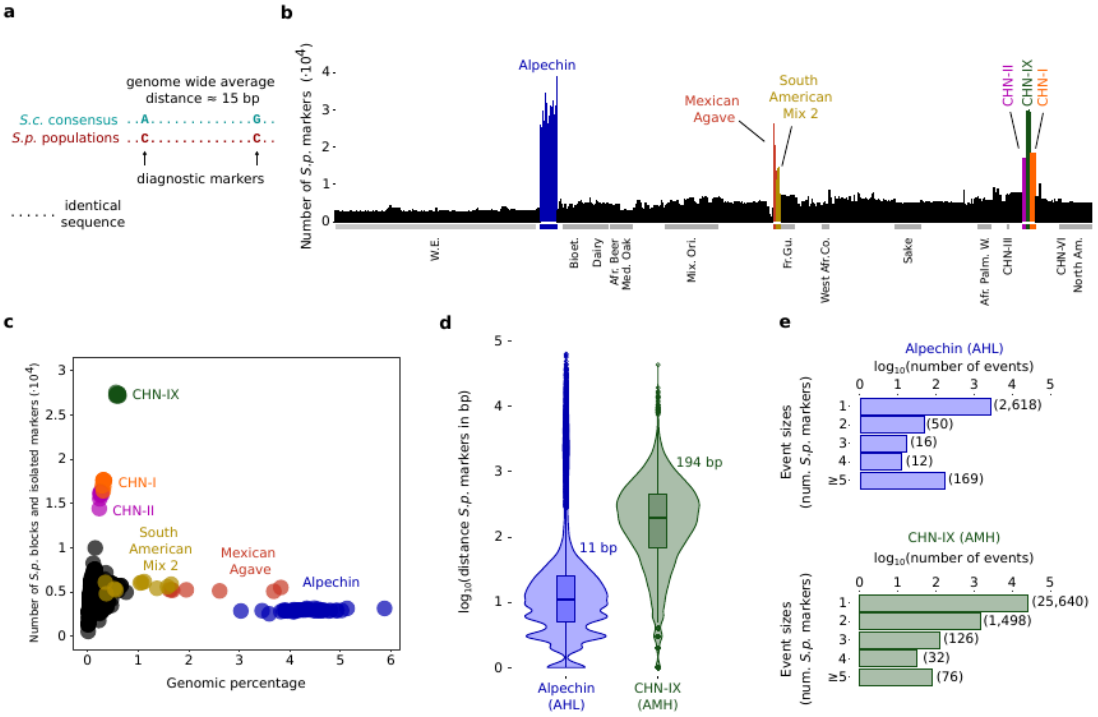

Figure 4.1 – **The species wide landscape of *S. paradoxus* markers in *S. cerevisiae***. **a**, A diagnostic marker position is defined as a biallelic SNP between the *S. cerevisiae* consensus (*S.c.c.*) sequence and in all the *S. paradoxus* populations and occur genome-wide on average at a 15 bp distance. **b**, Bar plot of number of *S. paradoxus* genotyped diagnostic markers (y-axis) across 1,673 isolates of *S. cerevisiae* (x-axis) with selected clades highlighted (rectangles). Coloured clades have the highest number of *S. paradoxus* markers. **c**, The percentage of genome (x-axis) included within the diagnostic marker with *S. paradoxus* genotype. Consecutive *S. paradoxus* markers are joined into blocks and their size is defined by the first and last marker of the block. Isolated *S. paradoxus* markers lie between two markers with *S. cerevisiae* genotype and were counted as 1bp. The y-axis shows the total number of blocks and isolated markers. Relevant clades are coloured as in panel **b**. **d**, Distribution of the distance (in $\log_{10}$ base pairs) between consecutive *S. paradoxus* genotyped diagnostic markers in the AHL Alpechin and AMH CHN-IX strains. Box, interquartile range (IQR); whiskers, 1.5×IQR; thick horizontal line, median. Data points beyond the whiskers are outliers. **e**, Number of blocks of different sizes and isolated *S. paradoxus* markers detected in AHL and AMH strains partitioned by size. The size of each block is given by the number of consecutive diagnostic markers with *S. paradoxus* genotype. Size equal 1 corresponds to isolated *S. paradoxus* markers. Blocks supported by at least 5 consecutive *S. paradoxus* diagnostic markers are combined into one category (≥5).

4.4.2 Deep coalescence in highly diverged S. cerevisiae lineages

We further explored the differences in patterns of *S. paradoxus* shared polymorphisms segregating among *S. cerevisiae* isolates by calculating D-statistics, which is based on counts of ABBA and BABA sites (A ancestral, B derived alleles) across a quartet of populations (Green et al. 2010) (Methods). The D-statistic quantifies to what extent two sister populations P1 and P2 differ in how often they share alleles with a donor population P3, at sites where P3 differs from an outgroup P4. No difference in the numbers of ABBA and BABA sites is expected under the null hypothesis of no introgression, and the statistical significance is assessed using block jackknife resampling. We used multiple whole genome alignments with the *S.c.c.* as P1, the European *S. paradoxus* CBS432 (*S.p.*) as donor species (P3), and *S. jurei* (*S.j.*) as outgroup (P4) to perform D-statistic tests across the entire *S. cerevisiae* collection (in turn P2). The quartet *S.c.c.*-AHL-*S.p.*-*S.j.* showed a negative D-value (-0.65), a net difference between ABBA and BABA sites (13,952 vs 2,964) and strong statistical support (Z-score : -16.87) formally supporting abundant *S. paradoxus* introgression in the Alpechin AHL strain (Fig. 4.2a). In contrast, the quartet *S.c.c.*-AMH-*S.p.*-*S.j.* showed a D-value close to 0 (0.0175), a comparable number of ABBA and BABA sites (12,695 vs 12,258) with no statistical support for *S. paradoxus* introgressions in the CHN-IX AMH strain (Z-score : 1.615). Thus, for AMH, we hypothesize that the presence of *S. paradoxus* alleles are the consequence of ILS, and not introgression (Fig. 4.2a, Supplementary Fig. 4.11). To test the robustness of this result to the strains used in the ABBA-BABA test, we repeated the ABBA-BABA test but instead used a Wine European, CHN-IV and CHN-I *S. cerevisiae* strains in the P1 position. We found no evidence for introgression in AMH using any of these P1 variations (Supplementary Fig. 4.11 a-b). We further investigated the ABBA and BABA positions in AMH that are shared across the ABBA-BABA tests performed with *S.c.c*, CHN-IV and CHN-I assemblies as P1 (Supplementary Table 5). We observed 18,152 shared ABBA+BABA sites which fall into intergenic (27.8%) and genic (72.2%) positions with an enrichment in 3rd codon positions (genic sites : 13,313 ; first : 2,052, second : 1,275 ; third : 9,787). This distribution of the ILS sites closely resembles the genome-wide distribution of diagnostic markers, suggesting no pervasive purifying selection acting on these ancestral polymorphisms. Next, we inspected the presence of CHN-IX *S. paradoxus* alleles in the CHN-I and CHN-II *S. cerevisiae* populations. We observed that approximately 71.9% and 76.1% of the *S. paradoxus* alleles detected in the CHN-I (strain FJ7) and CHN-II (strain BAG)

strains are shared with CHN-IX *S. paradoxus* alleles detected in AMH strain (Supplementary Fig. 4.11c), consistent with their common origin as ancestral polymorphisms. Finally, we extended the ABBA-BABA test varying the whole *S. cerevisiae* collection as P2 (Fig. 4.2b-c) and observed no evidence for *S. paradoxus* introgression in the CHN-II clade (e.g. BAG strain D : -0.0143, ABBA : 7782, BABA : 7562, and Z-score : -1.046), while CHN-I strains showed weakly but significantly positive D values (e.g. FJ7 strain D : 0.0528, ABBA : 8609, BABA : 9567, and Z-score : 4.834). However, the significance of the test for both CHN-II and CHN-I strains varies with the genome used as P1. Overall, we provided robust evidence that part of the genetic variants observed in the highly diverged CHN-IX *S. cerevisiae* population is driven by persisting ancestral alleles.

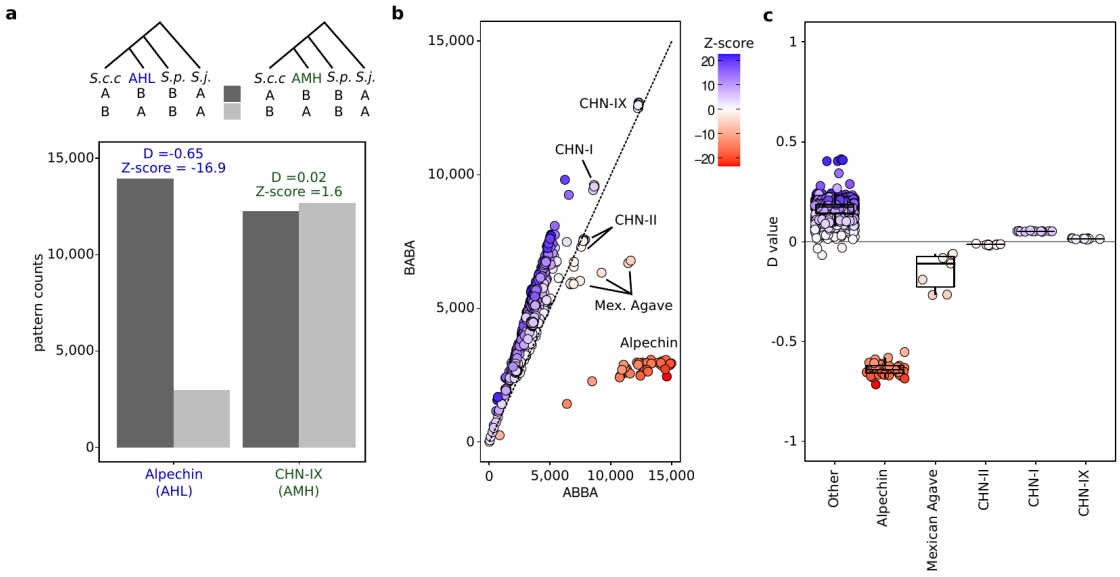

Figure 4.2 – **Incomplete lineage sorting in Chinese lineages. a**, Patterson's D statistics (ABBA-BABA test) with AHL and AMH in (P2) with the pattern *S.c.c.* (P1), CBS432 S.p. (P3) and *S. jurei* (P4). The bar plot indicates the number of ABBA and BABA patterns observed. "A" is the ancestral allele as imposed by P4, "B" is the derived allele. Only biallelic positions are included. **b**, Scatter plot of ABBA and BABA patterns across the *S. cerevisiae* collection rotating in P2 with the pattern *S.c.c.* (P1), CBS432 *S.p.* (P3) and *S. jurei* (P4). Nine strains resulted in empty vcf files and were excluded from the plot. The dotted line represents the diagonal. Relevant *S. cerevisiae* clades are labelled. **c**, Distributions of the Patterson's D value across the *S. cerevisiae* isolates. The dot's colour gradient reflects the Z-score values. Alpechin and Mexican Agave strains show the strongest introgression signal, consistent with *S. paradoxus* diagnostic markers arranged in large introgressed blocks. Box, interquartile range (IQR); whiskers, 1.5×IQR; thick horizontal line, median. Outliers are not shown.

4.4.3 Major hybridisation events in S. cerevisiae history

To trace past admixture events between *S. cerevisiae* and *S. paradoxus* populations, we constructed a high-resolution strain-by-strain catalogue defining boundaries, positions, and frequencies of introgressed blocks (supported by ≥ 5 consecutive diagnostic markers with *S. paradoxus* genotype, methods) across the *S. cerevisiae* clades (Supplementary Table 6). We observed an overall lack of association between introgression block breakpoints and meiotic recombination hallmarks (p-value = 0.91), consistent with their resection largely being shaped by mitotic recombination events (D'Angiolo et al. 2020). Grouping the strains by shared introgressed blocks globally recapitulates their SNPs-based phylogenetic relationships, reflecting that evolution from the original hybridisations to the current day introgressed blocks were shared within clades (Supplementary Fig. 4.12). We further investigated this evidence by extending our computational framework to identify the *S. paradoxus* population ancestry within the introgressed blocks. We also included four *S. cerevisiae* - *S. paradoxus* hybrids, with nearly complete subgenomes from both parental species, that may represent the ancestors of extant introgressed strains (Methods, Supplementary Fig. 4.6). We unveiled two major hybridisation events that likely occurred on the European continent. The first can be traced to the AQF *S. cerevisiae* - *S. paradoxus* hybrid isolated in Spain (Fig. 4.3a) and recognized as a clonal descendant of the ancestor of the Alpechin lineage (D'Angiolo et al. 2020). Despite the scattered geographical locations and broad sampling timeline (Supplementary Table 2), all Alpechin isolates extensively share common introgression blocks (Supplementary Fig. 4.12). The *S. paradoxus* introgressed sequence displays a uniformly high degree of similarity to the AQF hybrid sequence, thus supporting a single hybridisation origin of the entire clade (Fig. 4.3b). The second European hybridisation event is represented by the OS162 *S. cerevisiae* - *S. paradoxus* hybrid, which was isolated from oak tree bark in England. Its genomic profile revealed a triploid hybrid, with one copy of *S. cerevisiae* and two copies of *S. paradoxus* subgenomes, devoid of loss of heterozygosity (LOH) regions (Fig. 4.3a), the precursors of introgressions (D'Angiolo et al. 2020). Phylogenetic analysis placed the *S. paradoxus* subgenomes of OS162 in close relationship with the CBS432 European *S. paradoxus* reference strain (Fig. 4.3b), but not with AQF strain, supporting that distinct European *S. paradoxus* populations experienced interspecific hybridisation independently. We also identified two hybrids on the American continent. The first includes the UFMG-CM-Y652 Brazilian strain (Fig. 4.3a and Supplementary

Fig. 4.13) that coexists in a wild environment with multiple *S. cerevisiae* lineages with *S. paradoxus* introgressions (Barbosa et al. 2016). We compared the *S. paradoxus* subgenome of the UFMG-CM-Y652 hybrid with the *S. paradoxus* introgression blocks found in the French Guiana, Mexican agave and the wild Brazilian *S. cerevisiae* strains. The phylogenies are not consistent with UFMG-CM-Y652 wild Brazilian hybrid being the direct ancestors of the three introgressed American *S. cerevisiae* clades, and that either the ancestor, has not been found, or that subsequent admixture events further moulded the genome of extant introgressed clades (Supplementary Table 7). Finally, the second American hybrid, OS2389 strain, was isolated from a koa tree in Hawaii (Fig. 4.3a). The *S. paradoxus* subgenome of OS2389 hybrid showed different ancestry from the Hawaiian *S. paradoxus* lineage (UWOPS91-917.1), being instead related to the continental North (YPS138) and South (UFRJ50816) American *S. paradoxus* populations (Fig. 4.3c, Supplementary Fig. 4.13). Although OS2389 and the Brazilian hybrid UFMG-CM-Y652 do not show strong evidence of overlapping breakpoints at LOHs, both their *S. paradoxus* and *S. cerevisiae* subgenomes ancestries support a close parental origin making it difficult to conclude if they originated from independent hybridisation events (Supplementary Fig. 4.13). We further investigated the origin of the *S. paradoxus* subgenome on chromosome II (from 758,700 to 782,580 bp), which corresponds to a *S. paradoxus* introgressed block that is present in several wild Brazilian *S. cerevisiae* strains. Within this region, both the Brazilian hybrid UFMG-CM-Y652 and the Hawaiian hybrid OS2389 displayed the highest sequence similarity with the North American *S. paradoxus* (YPS138). In contrast, the wild *S. cerevisiae* Brazilian introgressed strains radiate from an independent branch, confirming that this block derives from a separate, and as of yet unknown, hybridization event (Fig. 4.3d). Overall, we identified at least two independent *S. paradoxus* - *S. cerevisiae* hybridizations that persisted as full hybrids in Europe and two hybrids in the American continent. One of the European hybridizations supports the single origin of the European Alpechin clade, whereas *S. paradoxus* introgressions in the American *S. cerevisiae* involves a more complex admixture scenario.

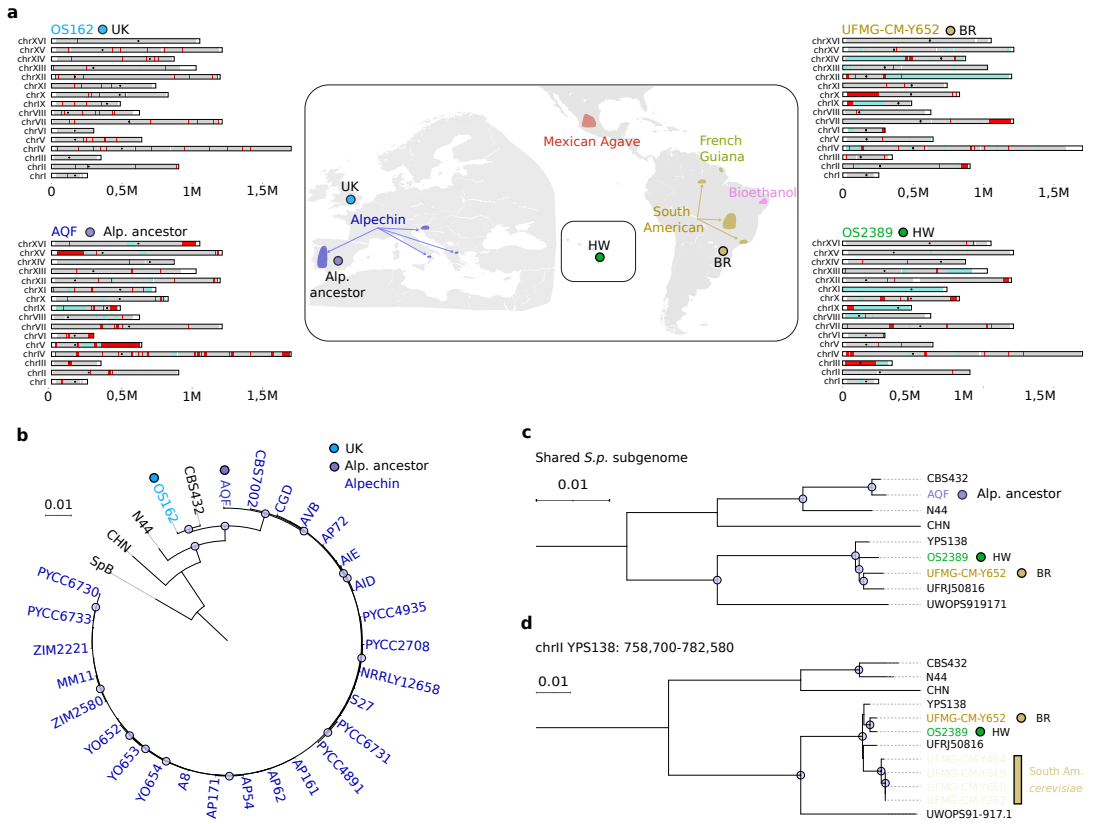

Figure 4.3 – **Independent hybridisation events. a**, *S. cerevisiae* and *S. paradoxus* genomic composition of the Spanish (AQF), Brazilian (UFMG-CM-Y652, BR : Brazilian), and the newly discovered hybrids from the UK (OS162) and Hawaii (OS2389, HW : Hawaiian). Colours represent homozygous *S. paradoxus* (red), homozygous *S. cerevisiae* (blue) and heterozygous (grey) regions. The map represents the geographic origin of the hybrids (coloured circles with black borders) and introgressed clades (coloured areas). European and American continents are represented at different scales and Hawaii (small box) has been moved to fit the layout. **b**, Maximum likelihood phylogeny of the shared *S. paradoxus* components across hybrid strains with European *S. paradoxus* ancestry (AQF and OS162) and the Alpechin *S. cerevisiae* clade. **c**, Maximum likelihood phylogeny of the shared *S. paradoxus* subgenome of hybrid strains with the American *S. paradoxus* component. **d**, Maximum likelihood phylogeny of the *S. paradoxus* region on $chrII : 758,700 - 782,580$ (YPS138 coordinates) introgressed in a subgroup of South American *S. cerevisiae* strains and both the Brazilian and Hawaiian hybrids (UFMG-CM-Y652 and OS2389). Across all the phylogenies, the circles at the nodes represent the percentage of bootstrap SH-aLRT (approximate likelihood ratio test (Guindon et al. 2010) $\geq 95\%$ and UFboot (ultrafast bootstrap approximation Minh et al. 2013, Hoang et al. 2018) $\geq 95\%$. The size of the node reflects the value of UFboot. N44 Far East, CHN (BJ-DLS32 − 26) Chinese, SpB ($14_{1}01$B) North American, UWOPS91 − 917.1 Hawaiian, UFRJ50816 South American *S. paradoxus*.

4.4.4 A shared introgression underlies an ancient admixture event

We calculated the density of *S. paradoxus* diagnostic markers with American or Eurasian ancestry across strains and observed a generally uniform ancestry within clades (Supplementary Fig. 4.14a). However, we detected a widespread high density of European-Far East Asian *S. paradoxus*

markers across multiple lineages, including strains characterised by genome-wide introgressions *S. paradoxus* American ancestry (Fig. 4.4a). We found this signal to be driven by a single introgression that encompasses the chromosome III centromeric region (CEN3). This introgression was shared across eighty-three *S. cerevisiae* strains, 75 belonging to the non-Asian domesticated clades while the other 8 belonging to the Asian domesticated clade (Fig. 4.4b). The introgression block is absent in the highly diverged Chinese wild lineages, implying that its origin postdates their divergence. The origin of the segment can be unambiguously assigned to the European *S. paradoxus*, with its phylogeny supporting a single origin, i.e. it derives from the same hybridization event (Fig. 4.4c and Supplementary Fig. 4.14b). In all the 83 strains, the block encompassed the genes located between *YCL002C* and *YCR005C*, but in some strains, it also spanned genes further away from CEN3 (up to *YCL009C* and *YCR010C*). A clustering based on the introgression boundaries showed no correlation with the *S. cerevisiae* populations phylogeny (Supplementary Fig. 4.14c-d). *S. cerevisiae* centromeres have a low recombination rate and CEN3 is particularly refractory given its linkage with the MAT locus (Mancera et al. 2008). Thus, the position of this introgressed block likely increased its chance of persisting over the generations and reduced the probability of it being lost or resected in the absence of strong deleterious effects. The low frequency in the population suggests that CEN3 introgression might not generally contribute to increasing fitness, nevertheless, functional implications (e.g. related to centromere functions or flanking genes) cannot be ruled out. The presence of an introgression with relatively conserved boundaries and shared ancestry across distinct domesticated *S. cerevisiae* strains point toward an ancestral event with a deep root in time. To test this scenario, we investigated the flanking regions outside the CEN3 introgression block in the industrial AHQ *S. cerevisiae* strain isolated in Taiwan from molasses (Supplementary Fig. 4.14e-f). Although the genomic background of AHQ is related to the Sake *S. cerevisiae* strains from the Asian domestication group, the 2-kb sequence upstream of the introgression block robustly branched with the Wine/European *S. cerevisiae* clade from the non-Asian domestication group (Supplementary Fig. 4.14f). This result, along with the common ancestry of the *S. paradoxus* CEN3 introgression, supports the evidence that the introgression was introduced into the Asian lineages through admixture with a Wine/European *S. cerevisiae* rather than from a distinct introgression event. In addition, the absence of introgression in wild Chinese lineages is consistent with a *S. cerevisiae* / *S. paradoxus* hybridisation that occurred after the initial out-of-China event, while the spread of the introgression across the clades was likely favoured by

domestication offering opportunities to move and interbreed (Liti et al. 2009). Thus, we report an ancient secondary contact between *S. cerevisiae* and a European *S. paradoxus* that likely occurred on the European continent either before or in concomitant with the *S. cerevisiae* wine domestication, leaving behind CEN3 introgression block.

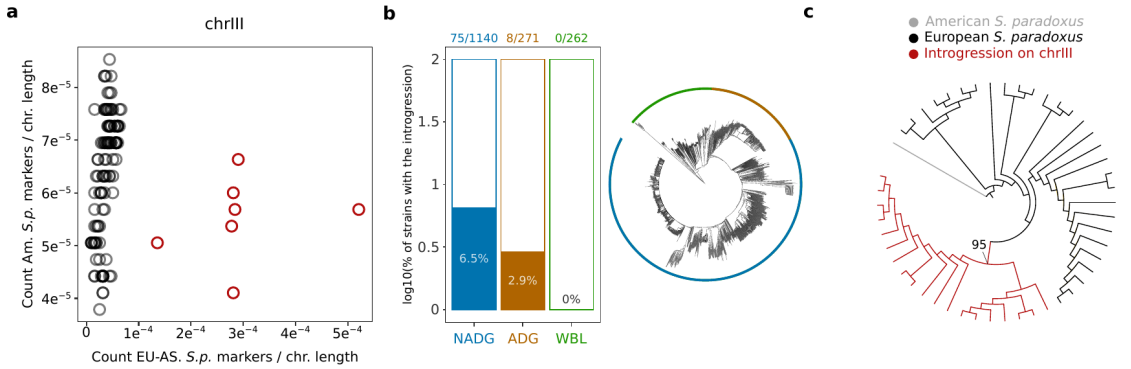

Figure 4.4 – **Ancient introgression on chromosome III. a**, Density of diagnostic markers with Eurasian (x-axis) and American (y-axis) *S. paradoxus* genotypes in *S. cerevisiae* strains with introgressed blocks. Red circles indicate a set of strains with the CEN3 introgressed block. The Alpechin strains were excluded to better visualise the other clades. **b**, On the left, barplot with the percentage of isolates with the introgression on CHR III ($\log_{10}$ scale). On the right, rooted neighbour-joining phylogeny of the 1,673 *S. cerevisiae* strains. The colours annotate wild basal lineage (WBL) and both non-asian and asian domesticated groups (NADG and ADG), while numbers are counts of strains with the CEN3 introgression block. **c**, Ultrametric maximum likelihood phylogeny of the CEN3 introgression block in *S. cerevisiae* strains and the corresponding orthologous region in European *Saccharomyces paradoxus* strains. 95 indicates the percentage of phylogenies supporting the CEN3 introgression branch. The branch length was removed to magnify the branch topology.

4.4.5 A convergent adaptive introgression

While most shared introgressions across *S. cerevisiae* strains are explained by common origin, events that arose independently and are maintained over time represent strong candidates of adaptive introgressions. We searched for introgressed genes that were independently acquired from different *S. paradoxus* populations and detected one striking case, corresponding to an introgression block encompassing the *PAD1-FDC1* gene pair (Fig. 4.5a-b). The two genes are functionally related and promote the transformation of the toxic compound cinnamic acid and related phenolic acids into potentially useful derivatives, like coumaric, caffeic, ferulic, and sinapic acids (Ramos-Cormenzana et al. 1996). The *PAD1-FDC1* alleles are introgressed in 55 *S. cerevisiae* strains, belonging to the Alpechin, Bioethanol, Mosaics R3, South American Mix 2 and Mix 3 clades (Supplementary Table 8), with all 40 Alpechin strains retaining such introgression blocks (Sup-

plementary Fig. 4.12a). We found the *PAD1-FDC1* alleles were introgressed and retained in *S. cerevisiae* at least twice, from either the European or the American *S. paradoxus* subpopulations (Fig. 4.5b), strongly arguing for an adaptive role. The selection for retaining this *PAD1-FDC1* introgressions across the Alpechin and Bioethanol clades could be exerted by the abundance of cinnamic acid and other phenolic acids in the olive oil wastewater environment (Jamoussi et al. 2005) and in intermediates of bioethanol production (Klinke et al. 2004). We experimentally tested this hypothesis by swapping the functional copy *PAD1-FDC1* alleles (Fig. 4.5c) from an alpechin *S. cerevisiae* (CPG) into the Wine/European *S. cerevisiae* genome (strain DBVPG6765) that closely resembles the *S. cerevisiae* genomic background of the Alpechin strains but lacks introgressions. Thus, we performed a competitive growth assay with DBVPG6765 wild type (DBVPG6765^{WT}) and DBVPG6765 with the Alpechin introgressed *PAD1-FDC1* copy (DBVPG6765^{ALP}). We tagged DBVPG6765^{WT} and DBVPG6765^{ALP} with two distinct fluorescent proteins (yGamillus, GFP and mRuby2, RFP) ; we then mixed DBVPG6765^{WT} with DBVPG6765^{ALP} in equal proportion ($\approx$ 50% DBVPG6765WT $\approx$ 50% DBVPG6765ALP) and let the strains grow and compete in different stressful environments (Fig. 4.5c). We tracked the growth in the presence of selective stresses characteristic of the olive oil wastewater or brine environments (Ferulic acid, tyrosol, NaCl) as well as phenolic azoles and rapamycin, to which *S. cerevisiae* laboratory strain lacking *PAD1* or *FDC1* activity are known to be sensitive (Hillenmeyer et al. 2008). At the end, we quantified the RFP and GFP fluorescence with the cytometer at two different time points, T_0 (starting point) and T_3 (end competitive experiment) and calculated the ratio $(RFP/GFP)_{T_3}/(RFP/GFP)_{T_0}$ (Methods, Fig. 4.5d). We found that the Alpechin *PAD1-FDC1* alleles conferred no fitness advantage in the absence of stress, but improved the growth in the presence of phenolic azoles, caffeine and ferulic acids (Fig. 4.5d and Supplementary Fig. 4.15). On the other hand, the Alpechin alleles did not confer tolerance to NaCl, ethanol, rapamycin or tyrosol, which is abundant in olive wastewater, but not known to be degraded by yeast. To our knowledge, this is the first experimental validation of an adaptive *S. paradoxus* introgression in *S. cerevisiae*.

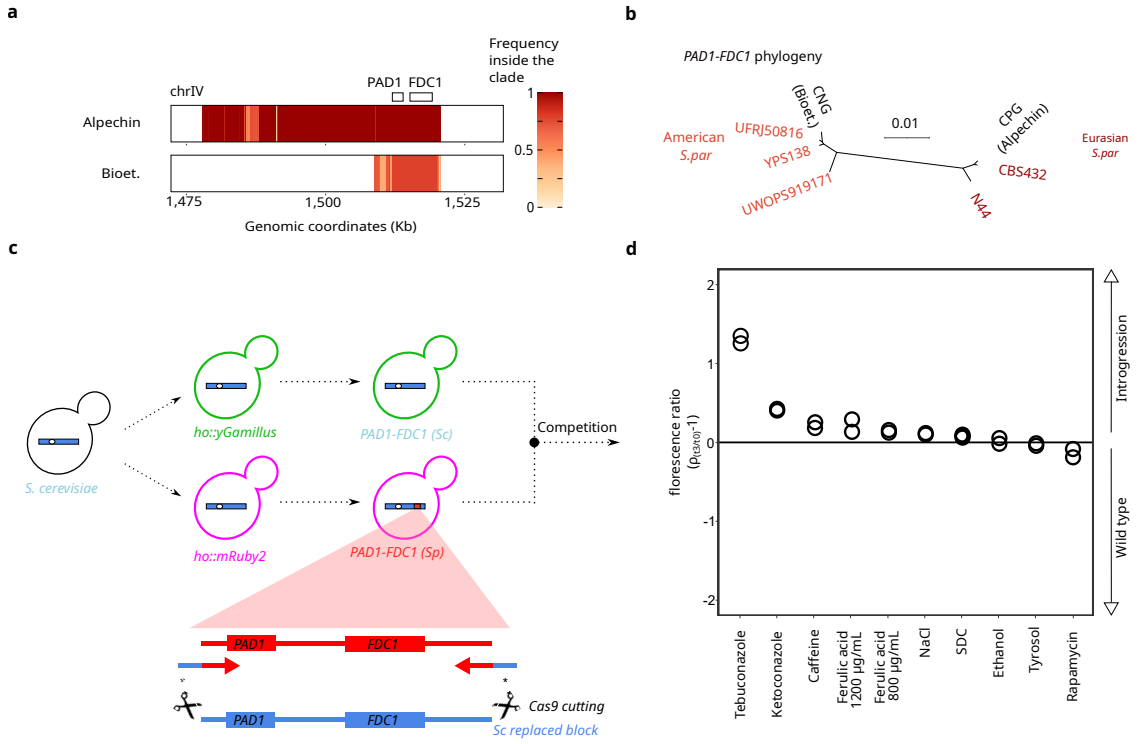

Figure 4.5 – **Adaptive introgression of the genes *PAD1/FDC1***. **a**, Heatmaps of the *S. paradoxus* introgression block at the end of chromosome IV, in the Alpechin and Bioethanol clades. The colour indicates the fraction of strains in each clade that carries the introgressed block. *PAD1-FDC1* positioning is indicative. **b**, Maximum likelihood unrooted phylogenetic tree underlies the distinct *S. paradoxus* ancestries of the *PAD-FDC1* introgressions. **c**, Graphical representation of the genetic engineering of *PAD1-FDC1* introgression in the *S. cerevisiae* Wine/European DBVPG6765. **d**, Fitness in different environments (x-axis) of a *S. cerevisiae* DBVPG6765 strain carrying the Alpechin *PAD1-FDC1* alleles (DBVPG6765^{ALP}) compared to DBVPG6765^{WT} strain. Fitness was measured by competing fluorescently tagged asexual versions of each strain over three growth cycles in each environment and measuring the intensity of each fluorescence, before and after the competition experiment. The y-axis reports the fluorescence ratio as $[(RFP/GFP)_{T_3}/(RFP/GFP)_{T_0}] - 1$ (named $\rho(t3/t0) - 1$) i.e the ratio between DBVPG6765^{ALP} and DBVPG6765^{WT} fluorescence after the competition (T_3) and before competition (T_0), such that a positive number represents higher fitness for the strain carrying the *S. paradoxus PAD1-FDC1* alleles (DBVPG6765^{ALP}). Circles represent biological replicas, four replicas were run for the non-stress condition (SDC) and two replicas for each of the stresses.

4.5 Discussion

Shared polymorphisms between populations can be ancient, predating their split, and persist by ILS. We reported abundant ILS polymorphisms in the highly diverged *S. cerevisiae* lineages that separated from the *S. cerevisiae* last common ancestor before the out-of-China event. The abundance of ancestral polymorphisms in the highly diverged Asian lineages might reflect different evolutionary histories compared to lineages that post-date the out-of-China event. Asian *S. cere-*

visiae lineages were isolated in a tropical forest environment (Q.-M. Wang et al. 2012, T. J. Lee et al. 2022), which can be perceived as the species ancestral ecological niche. Their persistence in a stable ecological niche without novel selective pressures and dispersal has resulted in fewer or less dramatic population bottlenecks maintaining a larger effective population size that favoured the persistence of ILS alleles. This finding somewhat mirrors African human populations, which share many variants with Neanderthals and Denisovans despite having little or no admixture from these archaic groups (Bergström et al. 2020). An alternative source of shared polymorphisms derives from the introgression process. We reported a detailed catalogue of introgression blocks in *S. cerevisiae* and unveiled multiple hybridisation events that shaped the species' evolution after the out-of-China dispersal. These hybridisations are not restricted to domestic environments, illustrating that they can occur in nature and perhaps can be selected for in human-made environments, where introgressed strains have been observed more frequently (Peter et al. 2018). Strikingly, three out of the four described hybrids present a signature of genome instability in the form of LOH, which enables introgressions to emerge in reproductively isolated yeast species (D'Angiolo et al. 2020). This finding supports that genome instability leading to LOH is a common mechanism and contributes to shaping the patterns of introgression across the genome. The remaining hybrid without LOH might represent a recent hybridisation event or did not yet experience the conditions that trigger the genome instability. Hybrids containing full subgenomes from both species enable in-depth phylogenetic analysis. While the UK, Spanish and Brazilian hybrids have subgenome ancestries consistent with the modern *S. cerevisiae* and *S. paradoxus* populations in these regions, the Hawaiian hybrid subgenomes seem to have originated elsewhere. The *S. cerevisiae* subgenome is very close to the Ecuadorian isolates and the *S. paradoxus* subgenome does not relate to Hawaiian *S. paradoxus*, but instead to the American clade. Therefore, the Hawaiian hybrid or its founding parents likely migrated from the American mainland. Overall, our study detected either complete hybrids with nearly full subgenomes from both parental species or isolates with up to 6% introgression. On the other hand, we have not detected isolates with more abundant introgression that represent the F1 hybrid in their backcrossing stages. This is consistent with experimental data showing that gametes with chimeric genomes are highly unfit (D'Angiolo et al., 2020) as a consequence of widespread genetic incompatibilities (Bozdag et al. 2019). Selection against introgressions has been observed in primates (Vilgalys et al. 2022) and produced deserts depleted of introgression in the human genome (Wolf & Akey 2018), perhaps suggesting that the decay of

introgressed material is rapid after the initial hybridisation. The yeast genetic toolbox enabled us to demonstrate the functional impact of the introgressed *PAD1-FDC1* gene pair on growth in the presence of antifungal drugs and ferulic acid. Such conditions might be relevant in the agricultural and industrial environments and underlie an adaptive potential for this introgression. These genes are known to metabolise cinnamic acid and to produce an off-flavour phenolic derivative (4-vinyl guaiacol) in alcoholic beverages. Selection against the undesired 4-vinyl guaiacol production has promoted the spread of loss-of-function alleles across the *S. cerevisiae* beer clades (Gallone et al. 2016). Both the Alpechin and Bioethanol isolates, which independently acquired and maintained the *PAD1-FDC1* introgression, are not used for fermented beverage production, but they share large swaths of the Wine/European and beer genetic backgrounds and potentially carried non-functional *PAD1-FDC1* alleles. An intriguing scenario is that introgressed material in yeast domesticated lineages helps to restore alleles that accumulated loss of function across many genes and pathways during domestication (De Chiara et al. 2022).

4.6 Data availability

The gVCF of the *S. cerevisiae* collection and raw data of the diagnostic markers and introgressed blocks coordinates are available at https://bitbucket.org/yeastgenomics. The genome sequences generated in this study are available at the European Nucleotide Archive (ENA) under the accession code PRJEB71987.

4.7 Code availability

The developed computational pipelines and scripts are available at https://bitbucket.org/yeastgenomics.

4.8 Acknowledgments

We thank Melania D'Angiolo, Eugenio Mancera, Nikolaos Vakirlis, Gilles Fischer, Etienne Danchin and Pedro Beltrao for discussions and critical reading of the manuscript. We also thank Vassiliki Koufopanou for sharing the strain OS162 and Ana Pontes and Jose Paulo Sampaio for providing information on the origin of the Alpechin and Brazilian sequenced strains. This

work was supported by Agence Nationale de la Recherche (ANR-11-LABX-0028-01, ANR-15-IDEX-01, ANR-18-CE12-0004, ANR-20-CE12-0020, ANR-22-CE12-0015), Fondation pour la Recherche Médicale (EQU202003010413), UCA AAP Start-up Deep tech, CEFIPRA. NT was partially supported by the PhD fellowship program Region PACA.

4.9 Author contributions

N.T., M.D.C. and G.L. conceived the project and designed the experiments ; N.T. developed the bioinformatic pipeline ; N.T., M.D.C., L.T. designed and performed the genomic analyses ; S.M. and C.V. designed and performed the PAD1-FDC1 functional characterisation ; E.S.N., A.B., J.W. and G.L. contributed with resources and reagents ; G.L. supervised the project ; N.T. and G.L. wrote the paper with input from all the other authors.

4.10 Competing interests

The authors declare no competing interests.

4.11 Supplementary

4.11.1 Methods

Saccharomyces collections

Genome sequences analysed in this study were previously published. The majority of the isolates belong to the 1011 *S. cerevisiae* collection (Peter et al. 2018). We included additional datasets to widen the sample size of several clades and specifically added strains from the following sources : wild and industrial Chinese (Duan et al. 2018), wild Brazilian (Barbosa et al. 2016), olive oil (Pontes et al. 2019), wild North American (Almeida et al. 2015), beer (Gallone et al. 2016,Gallone et al. 2019), Brazilian Cachaça (Barbosa et al. 2018) , flor (Coi et al. 2017), clinical (Ramazzotti et al. 2019) and industrial (Gonçalves et al. 2016,Legras et al. 2018). We checked the fastqs files' quality with FastQC and assembled a collection of 1,676 strains that covers all the currently known *S. cerevisiae* diversity in terms of both ecology and geography (Supplementary Table 2). The Alpechin strains from Greece and Italy have been recently sequenced by us and not

included in the global analysis of this paper, but they present an introgression pattern consistent with the previously described ones. We also included strains with a complete *S. cerevisiae* X *S. paradoxus* hybrid genome that had been either previously published or sequenced specifically for this study. We reanalysed the Alpechin living ancestor hybrid strain AQF (D'Angiolo et al. 2020). This strain is the equivalent of the CBS7002, which was also sequenced by Pontes et al. 2019 as a monosporic isolate, thus explaining the recombined genome configuration in this study. Two strains, UFMG-CM-Y651 and UFMG-CM-Y652, isolated in Brazil (Barbosa et al. 2016) also harbour a hybrid genome structure. While the UFMG-CM-Y652 was sequenced in its natural state, the UFMG-CM-Y651 was sequenced as a monosporic isolate, consistent with its recombined genome profile. The genome composition of the UFMG-CM-Y651 is consistent with the original diploid strain being similar to the UFMG-CM-Y652 and therefore likely belonging to the same F1 hybrid population. Finally, we sequenced two hybrid strains : OS162 and OS2389 isolated in the UK and Hawaii respectively. These strains were sequenced at The Earlham genomics facilities (www.earlham.ac.uk) according to the LITE pipeline (Perez-Sepulveda et al. 2021). The resulting libraries were run as a 384-plex pool on two NovaSeq SP lanes with 150 bp paired-end reads. For each lane, it generated 540M clusters passing filters (84%) with 92% bases $\geq$ Q30. In addition to the *S. cerevisiae* strains and genomes, we used five reference-quality genome sequences representative of *S. paradoxus* populations (Yue et al. 2017) for detecting the *S. paradoxus* shared polymorphisms.

***S. cerevisiae* phylogeny**

The short reads were mapped against the *Saccharomyces cerevisiae* consensus *S.c.c.* by BWA (v.0.7.16a options -/M -/t) and samtools view (v. 1.7 − 12 options -bS -F 1284). Not primary aligned, unmapped and duplicated reads were filtered out. samtools sort and index (default options) were used to sort the bam files and produce the indexes (bai). We generated genome VCFs (gVCFs) by bcftools mpileup (options -q 5 –annotate DP –skip-indels) and bcftools call (-m -Oz). We streamlined the gVCFs by rewriting the structure with bcftools query (options -f '%CHROM t %POS t %REF t %ALT t %QUAL t %INFO/DP t %FORMATn') and performed custom filtering with awk (using a threshold of 20 for quality and 10 for depth). The phylogenetic tree was built in R with the SNPRelate package as follows : generation of gds file snpgdsVCF2GDS (option method="biallelic-only"), import gds file snpgdsOpen, construct the dissimilarity matrix comparing each pair of strains (snpgdsDiss, options autosome.only = F, missing.rate=0.01), run bionj

algorithm (Gascuel 1997), laddered the tree with ladderize function. The phylogenetic tree was printed out using iTol with an unrooted layout (Supplementary Fig. 2). Based on the new extended phylogeny, we revised the clade nomenclature to accommodate geographical and ecological information. We remained faithful to the previous naming with few exceptions. We grouped the clades "French Dairy" (Peter et al. 2018) and "Milk Chinese" (Duan et al. 2018) in a single "Dairy products" clade ; we labelled two distinct Sake subclades (Sake-A and Sake-B). We kept the designation of "Asian fermentation" from the 1011 collection (Peter et al. 2018), despite these strains are not monophyletic in this extended phylogeny and are scattered in other Asian clades (Duan et al. 2018). Wild Chinese clades were maintained as in Duan et al. 2018. Finally, the wild Brazilian clades (B1, B2, B3 and B4, from Barbosa et al. 2016) and the Cachaça strains (Barbosa et al. 2018) were regrouped under the "South America Mix 1", "South America Mix 2" and "South America Mix 3".

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

To minimise the bias due to the use of a single reference sequence, we constructed a *S. cerevisiae* consensus genome. We selected 48 representative strains of the 1011 *S. cerevisiae* high coverage collection described in Peter et al. 2018 (Supplementary Table 1) and we calculated the allele frequency across the selected strains for each single nucleotide polymorphism (SNPs). Alternative alleles with frequencies higher or equal to 0.75 (15,559 positions) replaced the corresponding reference alleles in the SGD reference genome (strain : S288C ; version : SGD R64-1-1).

Restoring collinearity in *S. paradoxus* genomes

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

***Saccharomyces cerevisiae* strains ploidy estimation**

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

Definition of diagnostic markers dataset

The six assemblies of *S.c.c.*, CBS432, N44, YPS138, UFRJ50816 and UWOPS91 − 917.1 were masked in correspondence of both repetitive elements (LTR, Core X, Y', Ty elements) and rapidly evolving sequences such as telomeric regions, introns, pseudogenes and noncoding exons, replacing the sequences with 'N' values. We ran MUMmer v3 (nucmer algorithm options –mum) to align the masked assemblies in pairwise combinations, chromosome by chromosome, taking advantage of their restored collinearity. The delta files were filtered with delta-filter (-r -q -l 400) keeping reference-query alignments longer than 400 bp and allowing for inversions. show-snps (options -ClrT) was used to extract the polymorphic positions. We identified a set of 1,257,196 marker positions ; 1,205,445 (95.9%) were biallelic, 50,835 (4%) triallelic and 916 (0.07%) tetrallelic. We retained only the biallelic markers for the downstream analysis. We identified and flagged different allele patterns occurring at each biallelic marker position, across the assemblies, defining ten categories. The first set, which is the most abundant, consists of 775,500 biallelic markers in which *S. paradoxus* allele is conserved across the five assemblies, while the *S.c.c.* allele is different (species-specific markers). These alleles are considered to have arisen after the split between the species. The second set consists of 67,837 markers in which alleles are specific to the euroasiatic *S. paradoxus* populations (CBS432 and N44 assemblies), while *S.c.c.* allele is shared with the American *S. paradoxus* populations (YPS138, UFRJ50816 and UWOPS91 − 917.1 assemblies). The third set consists of 78,463 markers in which alleles are specific to the American *S. paradoxus* populations. *S.c.c.* allele is shared with the two Euroasiatic *S. paradoxus* populations. The fourth set consists of 62,987 markers in which the alleles are specific to the continental American *S. paradoxus* populations (YPS138, UFRJ50816 assemblies) while the *S.c.c.* allele is shared with the Euroasiatic *S. paradoxus* populations and the Hawaiian *S. paradoxus*. The fifth set consists of 36,968 markers in which one allele is private of the European *S. paradoxus* population (CBS432 assembly). The sixth set consists of 34,300 markers in which one allele is private to the Far Eastern *S. paradoxus* population (N44 assembly). The seventh set consists of 7,975 markers in which one allele is private to the North American *S. paradoxus* population (YPS138 assembly). The eighth set consists of 10,641 markers of which one allele is private to the South American *S. paradoxus*

population (UFRJ50816 assembly). The ninth set consists of 81,547 markers of which one allele is private to the Hawaiian *S. paradoxus* population (UWOPS$91-917.1$ assembly). The tenth and last set consists of 49,227 markers in which the allelic patterns do not follow the phylogeny of the species. We collected the number of markers in 1-kb non-overlapping windows (Supplementary Fig. 1) and we measured the markers' density across the chromosomes (density = number of markers in one chromosome / total size of the blocks aligned with MUMmer) showing that the markers are evenly distributed across the chromosomes (Supplementary Fig. 1).

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

We used a modified version of MuLoYDH57 to screen 1,673 *S. cerevisiae* strains to identify those enriched in *S. paradoxus* alleles using as input the dataset of species-specific markers stored as BED files with either *S.c.c.* or CBS432 coordinates. The FASTA genomes were indexed with samtools faidx and bwa index (default options). The mapping of the short reads, against both *S.c.c.* and CBS432, was performed by piping BWA (v.0.7.16a options -M -t) and samtools view (v.$1.7-12$ options -bS -F 1284). Not primary aligned, unmapped and duplicated reads were filtered out. samtools sort and index (default options) sorted the bam files and produced the indexes (bai). We collected the statistics about the short read mapping (samtools flagstat) and the coverage (samtools depth option -a piped with awk). The markers positions were genotyped using the sorted bam, piping samtools mpileup (options –positions -u -min-MQ5 –output-tags AD,ADF,ADR,DP,SP –skip-indels –redo-BAQ) and bcftools call (–ploidy -c -Oz). The ploidy option was set up to either 1 or 2; whenever this information was missing (NA), or the ploidy was higher than 2, we performed the call with the default option (–ploidy 2). A preliminary step with GEM-indexer (v.1.423 options -c dna) followed by GEM-mappability (v.1.315 options -l) provided the read-length specific mappability information for both the FASTA genomes. Control-FREEC v10.7 was run to detect the copy number variants of the sample with known ploidy levels. A per-sample configuration file was prepared with custom yeast parameters (telomeric=4000, coefficentOfVariation=0.05, minExpectedGC=0.35, maxExpectedGC=0.40); the significance of Control-FREEC predictions was assessed with assess_significance.R which returns two p-values (from a Wilcoxon and Kolmogorov-Smirnov test) for each CNVs. Only CNVs detected against both the references and supported by both p-values < 0.01 were retained. CNVs were only investigated on samples with known ploidy. Markers genotyped against both the input assemblies were filtered following a

strict strategy to keep the number of false positive calls low. We discarded marker positions from the mitochondrial genome and both telomeric and subtelomeric regions because of their known highly-content variability (Yue et al. 2017). Furthermore, we extended the filtering on chrXIV from 1 to 38,500 because the European *S. paradoxus* population contains an introgression from *S. cerevisiae* (Liti et al. 2006) that would produce false-positive introgressions. Markers genotyped against *S.c.c.*, but not those against *S. paradoxus*, were quality filtered (QUAL $>$ 20). In addition, we only retained markers whose genotypes were consistent across both mappings (example : REF A ALT G against *S.c.c.* must correspond to REF G ALT A against CBS432). For haploid and diploid strains, the positions embedded in CNVs were filtered out because the altered ploidy may compromise the step of genotyping. We generated a catalogue of introgression blocks supported by at least 5 consecutive diagnostic markers with *S. paradoxus* genotype by a custom-made R scripts (Supplementary Table 4). The selected value of $\geq$5 consecutive diagnostic markers was based on the introgression blocks abundance and distribution across the genome and represented a conservative threshold to retain highly confident events for a subset of follow-up analysis. To trace back the origin of the introgressions, we performed a second mapping that requires as input all six assemblies. We selected a few strains to maximise different patterns of introgressions. The FASTA genomes were indexed with samtools faidx and bwa index (default options). The mapping of the short reads, against *S.c.c.*, CBS432, YPS138, UFRJ50816 and UWOPS91 $-$ 917.1 were performed piping BWA (v.0.7.16a options -M -t) and samtools view (v. 1.7 $-$ 12 options -bS -F 1036). samtools sort and index (default options) sorted the bam files and produced the indexes (bai). The markers positions were genotyped across the short-read sorted bam, piping samtools mpileup (options –positions -u -min-MQ5 –output-tags AD,ADF,ADR,DP,SP –skip-indels –redo-BAQ) and bcftools call (–ploidy -c -Oz). The markers were kept when the alternative allele was concordant in at least 4 out of the 6 mappings. The presence of markers private to specific subpopulations was used to infer the origin of the event.

Comparing whole-genome phylogeny with gene-based trees

S. paradoxus has a well-defined population structure. To identify genomic regions having a discordant evolutionary history, we extracted 5,191 1-to-1 orthologous genes across *S. paradoxus* and *S.c.c.* assemblies based on genome annotations. We discarded ambiguously annotated, broken or double-annotated sequences. For each ortholog, we performed multisequence alignment with MUSCLE (v3.8.31) and we generated a UPGMA gene tree (MUSCLE with -maketree default

options). We used the Robinson–Foulds metric (RF.dist function in R package phangorn v.2.10.0) to measure the distance of gene trees from the species tree. Most gene trees ($4,569/5191$; 88%) retraced the species tree structure and are informative for detecting potentially introgressed genes. Gene-based trees with a phylogeny discordant to the species phylogeny ($622/5,191$; 11%) were reported. Among these 569 were 2 operations distant from the species tree, 48 measured 4 operations and 6 trees 5 operations. In order to assess the presence of *S. cerevisiae* introgression within *S. paradoxus*, we extracted and compared the phylogenies for 5,191 1-to-1 orthologs which were determined on the basis of the annotations of 5 *S. paradoxus* genomes, representing the available populations, and the *S. cerevisiae* reference genome.

Identification of introgressions from S. paradoxus using statistical tests

We developed a pipeline that, taking as input either short read datasets or de novo assemblies, performs Patterson's D (Green et al. 2010), f (Martin et al. 2013,Durand et al. 2011) and df (Pfeifer & Kapan 2019) statistics. These statistics rely on the comparison of alleles across 4 populations (P) whose phylogenetic relationship can be described as (((P1,P2),P3),O) where O stands for outgroup. In our study, P1 and P2 represent two different populations of S. cerevisiae, P3 a population of *S. paradoxus* while *S. jurei* is used as an outgroup in O. For each *S. cerevisiae* strain selected, we generated an artificial whole genome assembly by replacing the specific isolated alternative alleles on the scaffold of *S.c.c.* To do this, we mapped the selected *S. cerevisiae* short-reads against *S.c.c.* by piping BWA (v.0.7.16a options -M -t) and samtools view (v.$1.7-12$ options -bS). Duplicated reads were filtered out. samtools sort and index (default options) sorted the bam files and produced the indexes (bai). The variant calling was performed piping samtools mpileup (options –positions -u -min-MQ3 –output-tags AD,ADF,ADR,DP,SP –skip-indels –redo-BAQ) and bcftools call (-vm -Oz). Variants with QUAL $<$ 20 were discarded as well as INDELs and multiallelic positions. Finally, the alternative allele of the remaining variants was used to replace the allele on the *S.c.c* scaffold with a homemade script. We then performed multiple whole-genome alignments with progressiveMauve (Darling et al. 2004) (released $2015-02-13$) default option. All the alignments were run respecting the input order imposed by the phylogeny : (((P1,P2),P3),O) where P1 rotated between *S.c.c.*, a Wine/European, a CHN-IV and a CHN-I isolate. Every *S. cerevisiae* isolate of the collection was used as P2, P3 is the de novo assembly of the European *S. paradoxus* (CBS432) and O is the de novo assembly of *S. jurei*. We extracted the SNPs from the xmaf file with org.gel.mauve.analysis.SnpExporter. The P2 alleles were considered reference (0 in the vcf)

and only biallelic positions were kept. We took advantage of the availability of *S.c.c.* and CBS432 annotations for excluding the SNPs in telomeric and subtelomeric regions and SNPs located in different chromosomes when the chromosomes of P1 P2 and P3 were compared as these genomes are collinear. The resulting VCF was equipped with a VCF header for downstream analysis.

Patterson's D statistics

We integrated the script written by Joana Meier (convertVCFtoEigenstrat.sh option rec=0.3 cM/Kb) for converting the VCF file in the corresponding map, geno and ind files. Patterson's D statistics were run with ADMIXTOOLS (v.7.0.2) in R using the package admixr (v.0.9.1) function d(). Assuming around 2.2 million SNPs per sample we tested different blgsize and we selected 0.12 which resulted in around 345 blocks for the jackknife resampling. A number of 345 blocks corresponds to around 6,300 SNPs per block ; given 1 SNP every 5.5 bp ($11,000,000bp/2,200,000$ SNPs =5.5) each block covers a genome length of around 34,650 bps ($6,300snps * 5.5bp$) a distance at which the linkage disequilibrium approaches low r^2 values close to the baseline (Green et al. 2010, Tsai et al. 2008, Schacherer et al. 2009). The D values were reported together with the sd error, the Zscore and the number of sites for both ABBA and BABA patterns.

f statistics

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

df statistics

We subset the VCF chromosome-by-chromosome and ran the df statistics implemented in the R package PopGenome (v.2.7.5)65 in 200 bp non-overlapping sliding windows ; we then highlighted windows for which df values were 5 times higher than the whole-chromosome mean value.

Proof of concept

As a proof of concept, we reconstructed the genome of an Alpechin strain known to be introgressed (AHL) and a CHNIX (AMH) strain for which one introgression block from an unknown Saccharomyces spp on chrXI was detected (Peter et al. 2018). For both the strains we compared the results obtained with 1) the reconstructed false genome and 2) the *de novo* genome.

Introgression breakpoint and hotspot association

The association between the introgression breakpoints detected in the Alpechin strains and DSB/recombination hotspots was tested by means of the regioneR package (v.1.30.0) (Gel et al.

2016). The association tests were performed by means of the function overlapPermTest setting the parameters ntimes = 10000, and alternative = "greater". The fasta suite (Pearson & Lipman 1988) (fasta36 -b 3 -d 3 -m 8, version : 36.3.8d April 2016) was used to convert the coordinates of the hotspots regions detected in the original reference genomes ($SGD_R62-1-1_20090218$ and $SGD_R58-1-1_20080305$ for the hotspots reported by Pan et al. 2011 and Mancera et al. 2008, respectively) to the corresponding coordinates of the *S.c.c.* genome. Introgression breakpoint regions were defined, e.g., as the genomic interval between the first marker of an introgression region and the closest flanking marker which does not belong to the same introgression region. Overall, we tested 3868 hotspot regions (median width 196 bp) and 13,267 introgression breakpoints (median width 19 bp), detecting 482 overlaps. Our test did not reveal any association between the DSB/recombination hotspots and introgression breakpoints (p = 0.95). The robustness of the approach was previously assessed by applying the test to a collection of LOH breakpoints formed through the return-to-growth protocol and recombination hallmarks, revealing a strong association (Mozzachiodi et al. 2021).

Phylogeny of the introgression on chromosome III

We collected 30 short read sequencing for European *S. paradoxus* populations sampled from Canada (Eberlein et al. 2019), USA (Leducq et al. 2016), UK, Russia, Ukraine, Italy, Spain, Portugal, Germany and Latvia (Koufopanou et al. 2020); one North American *S. paradoxus* strain (YPS138) (Yue et al. 2017) that served as an outgroup and 18 representative strains with the introgression on chrIII from different clades : Wine European, Alpechin, Bioethanol, Mosaics beer, Mantou 7, Mosaics, the Alpechin hybrid described in D'Angiolo et al. 2020 and its spore described by Pontes et al. 2019. For each strain, we ran bwa in a competitive mapping concatenating *S.c.c.* and CBS432 (European *S. paradoxus* genomes) and we removed PCR duplicated reads. As all the samples (except the hybrid) were homozygous and diploid we performed position genotypization (ploidy 2) restricted to the shortest shared introgressed area against chrIII of CBS432 from position 129,241 to 159,904 (samtools mpileup -u -l CBS432.chrIII.bed –adjust-MQ 50 -min-MQ5 –output-tags AD,ADF,ADR,DP,SP –redo-BAQ -f CBS432.genome.fa sample | bcftools call -m -Oz > sample.chrIII.variants.gvcf.gz). The resulting gvcf were filtered for the variant quality (vcftools –gzvcf –minQ 20 –recode –recode-INFO-all –out). For each strain, we reconstructed its FASTA sequence by using the sequence of CBS432 as a scaffold and replacing the positions corresponding to a variant site with the ALT allele. Since we wanted to compare intra-lineage

short sequences we performed a stringent masking to keep only reliable positions. We masked the following segments : $129,241 - 132,000$; $132,977 - 133,178$; $153,001 - 153,799$ and $159,530 - 159,904$; plus 1) the positions corresponding to INDELs; 2) multi allelic variant positions; 3) positions for which we were not able to identify the alleles in more than 4 samples and 4) positions missing against the outgroup (North American *S. paradoxus*) as we could not assume sequence identity with the European *S. paradoxus.* We then refined the masking. For each gene within the introgression, we counted the number of masked positions (Ns) and if more than 5% of the gene length was replaced by Ns the entire sequence was completely masked; 5/15 genes were completely masked (*YCL009C*, *YCL008C*, *YCL005W*, *YCL005W-A* and *YCR003W*). We joined the FASTA of the samples in a single multi-FASTA file and we constructed the phylogeny with MEGA X (v.10.1.8) using the Maximum Likelihood method and Kimura 2-parameter model (other options included complete deletion and initial tree generation by mean Neighbour-Join and BioNJ algorithms; non-coding nucleotide sequence were assumed).

Phylogeny across shared S. paradoxus introgression

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

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

For each hybrid strain, we performed a competitive mapping using bwa and concatenating S288C reference genome (version : R64 − 1 − 1) and either the American*S. paradoxus* YPS138 for the American hybrids or the European *S. paradoxus* CBS432 genomes for the European hybrids. For the phylogeny of the *S. cerevisiae* subgenome, we included representative strains from the main *S. cerevisiae* clades. From the bam files, we performed whole-genome position genotypization for each sample (bcftools mpileup -E -Ou -q 5 -a DP -f genome.fa sample.bam | bcftools call -mO v | bcftools view –max-alleles 2 –exclude-types indels -e 'QUAL $\leq$ 20 || FORMAT/DP $\leq$ 10' | bcftools annotate -x ID,INFO,FILTER | bgzip $>$ sample.gvcf.gz"). We then merged the gvcfs in a single multi-sample gvcf file and kept only the positions against the S288C reference genome that was genotyped across all the samples. The multi-sample gvcf file was converted in the PHYLIP file format with vcf2phylip.py with options -m 1000 -o CEI. CEI is a CHN-IX *S. cerevisiae* strain and it was used as an outgroup. The PHYLIP file was the input for the construction of the phylogeny with the software IQ-tree (iqtree -s file.phy -bb 1000 -alrt 1000 -bnni -nt AUTO -m TEST+ASC). The phylogenetic tree was represented using iTol with a rooted circular layout (Supplementary Fig. 8b, panel on the right). For the phylogeny of the *S. paradoxus* subgenome of the European hybrids we included : the Alpechin *S. cerevisiae* strains, the European *S. paradoxus* CBS432, and one representative for each Euroasiatic *S. paradoxus* population while the American *S. paradoxus* was used as an outgroup. Instead, for the phylogeny of the *S. paradoxus* subgenome of the American hybrids we included the American *S. paradoxus* strains from Eberlein et al. 2019 and Yue et al. 2017. Using the gvcfs generated as described above, we merged the sample gvcfs in a single multi-sample gvcf file but, we kept : the positions against the European *S. paradoxus* CBS432 reference genome for the European strains and, the positions against the American *S. paradoxus* YPS138 reference genome for the American strains. In both cases, we kept only the positions genotyped across all the samples. The two gvcf files, for the European and American *S. paradoxus* ancestry, were converted in PHYLIP files and the phylogenies were obtained as described above. The phylogeny of the *S. paradoxus* ancestry of the European strains was depicted in Fig. 4.3b while the phylogeny of the *S. paradoxus* ancestry of the American strains was depicted in Supplementary Fig. 8b (panel on the left). The same competitive mapping strategy was used, with different groups of strains, for the phylogenies depicted in Fig.4.3c, Fig. 4.3d and Fig. 4.4d.

CRISPR/Cas9 plasmid assembly and introgression engineering

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

Insertion of a cassette encoding fluorescent proteins at the HO locus

The fluorescent protein yGamillus was amplified by PCR from the plasmid pL42, while the fluorescent protein mRuby2 was amplified by PCR from the plasmid pL71. The sequence of yGamillus was obtained from the Genescript construct pUC57-Kan-yGamillus TEF1ov. The sequence of mRuby2 was obtained from the plasmid pFA6a-link-yomRuby2-Kan generated by Kurt Thorn's lab (S. Lee et al. 2013). The plasmid targeting the HO locus was generated using the pUDP004

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

Competitive growth assay

The two competing strains ygl4147 (MATa, ho : :yGamillus, ura3 : :KanMX) and ygl4151 (MATa, ho : :mRuby, ura3 : :KanMX, pad1-fdc1Sc : :PAD1-FDC1Sp) were patched from glycerol stocks onto YPD plates and incubated overnight at 30°C. The following day part of the patch was transferred to two different falcon tubes per strain containing 5 mL liquid SDC media (0.67% YNB w/o amino acids, 2% dextrose) and grown overnight (16-18 hours) at 30°C with shaking at 220 rpm. The following day we measured the OD600 of the four independent cultures and transferred an equal amount of cells of the two competing strains (total OD600 = 1) from the overnight cultures to two new tubes containing SDC or an SDC-based media to which we added one of the following reagents : ferulic acid (800 µg/mL or 1200 µg/mL), caffeine (0.002 mM), rapamycin (30 nM), ethanol (10% v/v), tebuconazole (0.0037 mg/mL), ketoconazole (0.0075 mg/mL), NaCl (0.6 mM) or tyrosol (0.6 mg/mL). We took 200 µL of cultures, which were sonicated and then analysed by flow cytometry using the filters B525-FITC and Y585-PE on the Cytoflex LX (Beckman Coulter) to evaluate the difference of the GFP/RFP ratio at T_0 in the population. The cultures were grown for 24 hours at 30°C with shaking at 220 rpm. The following day the OD600 of the cultures was measured and diluted to an OD600 of 1 in fresh media containing the same compounds that were added at the beginning of the experiment and repeated the same transfer the following day. Finally, we collected 200 µL from each culture and transferred them to a 1.5 mL Eppendorf tube. The samples were sonicated and then analysed following the same procedure used for the T_0 and using the same gating strategy as in the T_0. Events that were not included in the T_0 gates were discarded as they were often small (low FSC-H) compared to the median cell size of the population and were likely cell debris.

4.11.2 Figures

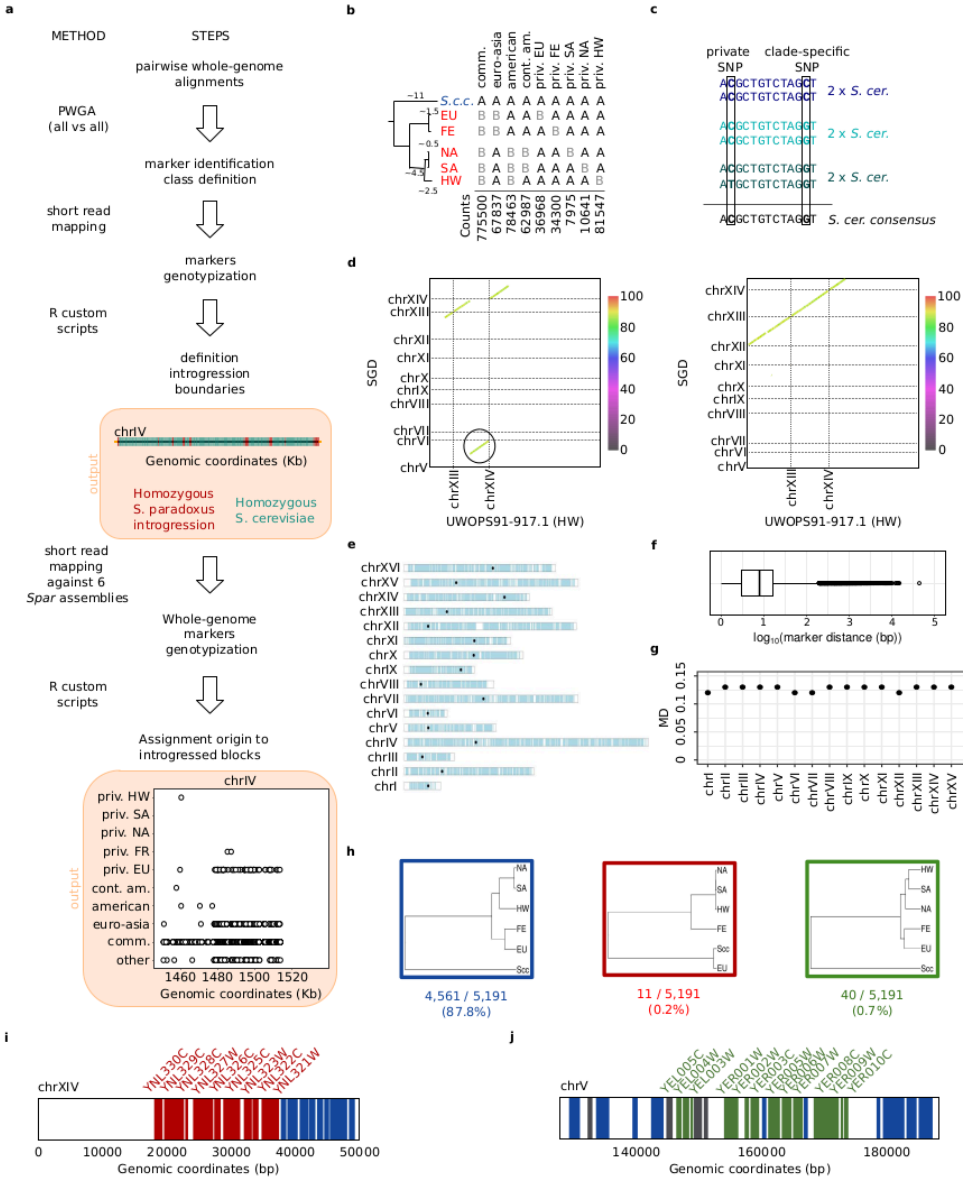

Figure 4.6 – **Methods overview**. **a**, Workflow of the pipeline. Abbreviations : PWGA : pairwise whole-genome alignment ; CHR : chromosome. **b**, biallelic patterns across marker positions. A : *S.c.c.* allele ; B : alternative allele. Numbers on the phylogeny represent whole-genome sequence divergence between *S.c.c.* and *S. paradoxus* and within the main *S. paradoxus* populations. In red the abbreviation of the main *S. paradoxus* populations. EU : European, FE : Far Eastern, NA : North American, SA : South American, HW : Hawaiian. The columns indicate the marker positions used to define the introgression boundaries (common abbr. comm.) and the origin (the remaining columns). The counts correspond to the number of marker positions available for each pattern. **c**, A cartoon of the strategy adopted to construct the *S.c.c.* sequence, which is explained in the methods. Briefly, for each clade of the 1,011 collection, we picked 2 strains and extracted the SNPs against the SGD reference genome. We then used SGD genome as a scaffold and changed the alleles in the positions in which the ALT allele was more frequent than the REF allele (freq. ≥ 0.75). **d**, Example of restoring the collinearity of the translocated genomic region between *S.c.c.* and HW *S. paradoxus* on the translocation chromosome V/ chromosome XIII described in Yue et al. 2017. The colour of the line reflects the sequence divergence between *S.c.c.* and the HW *S. paradoxus*. **e**, The blue rectangles represent the genomic regions, in *S.c.c.* coordinates, which were effectively aligned across *S.c.c.* and all the *S. paradoxus* whole-genome assemblies. **f**, Distribution of the diagnostic markers distance along the genome (1st Qu. : 3, mean : 14.59, median : 8, 3rd Qu. :16). Box, interquartile range (IQR) ; whiskers, 1.5×IQR ; thick horizontal line, median. Circles represent outliers. **g**, The per-chromosome marker density (MD) is measured as the number of diagnostic markers divided by the sum of the aligned regions depicted in the g. panel. **h**, UPGMA phylogenies across the genome assemblies of 5,191 *S. cerevisiae* - *S. paradoxus* 1-to-1 orthologs. In blue, the most abundant individual gene topologies follow the structure of the species tree. In red and green two contrasting gene topologies. **i**, zoom in on the distribution of contrasting gene phylogenies across the assemblies. On the left, the *S. cerevisiae* introgression on chromosome XIV on the European population of *S. paradoxus*. In red, the genes with the phylogeny are depicted in the central panel in i. On the right, the introgression / ancestral variation shared by the South American and the Hawaiian *S. paradoxus* on the region surrounding the centromeric position of chromosome V. In green, the genes with the phylogeny are depicted in the right panel in j ; in grey, gene trees with alternative topologies. In both the zoom-in, the blue rectangles represent genes whose phylogeny respects the species tree topology.

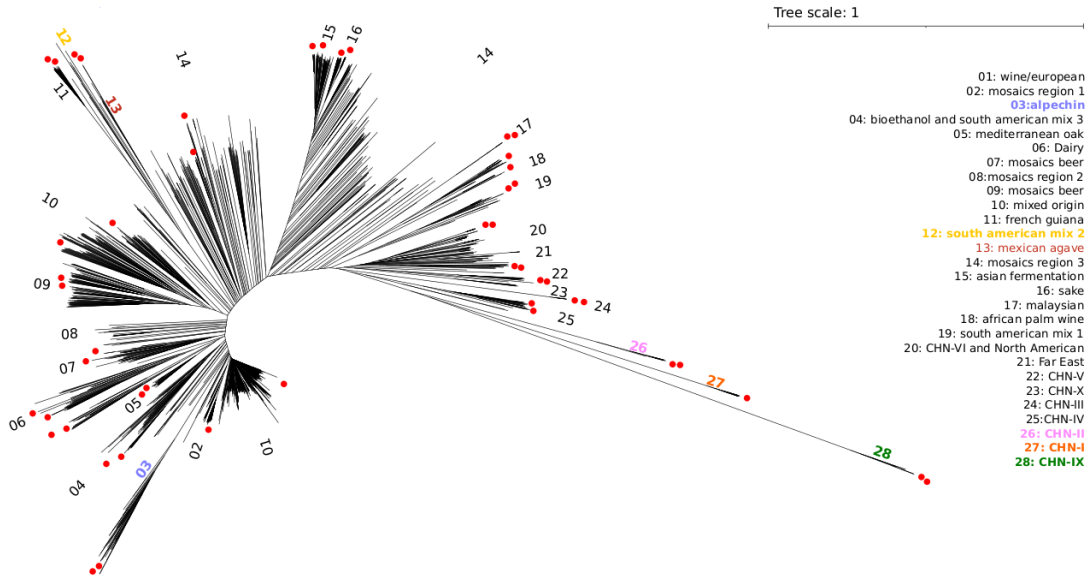

Figure 4.7 – **Global *S. cerevisiae* phylogeny**. Unrooted neighbour-joining tree of the 1,673 *S. cerevisiae* strains analysed in this work. Coloured clades are discussed in the main text.

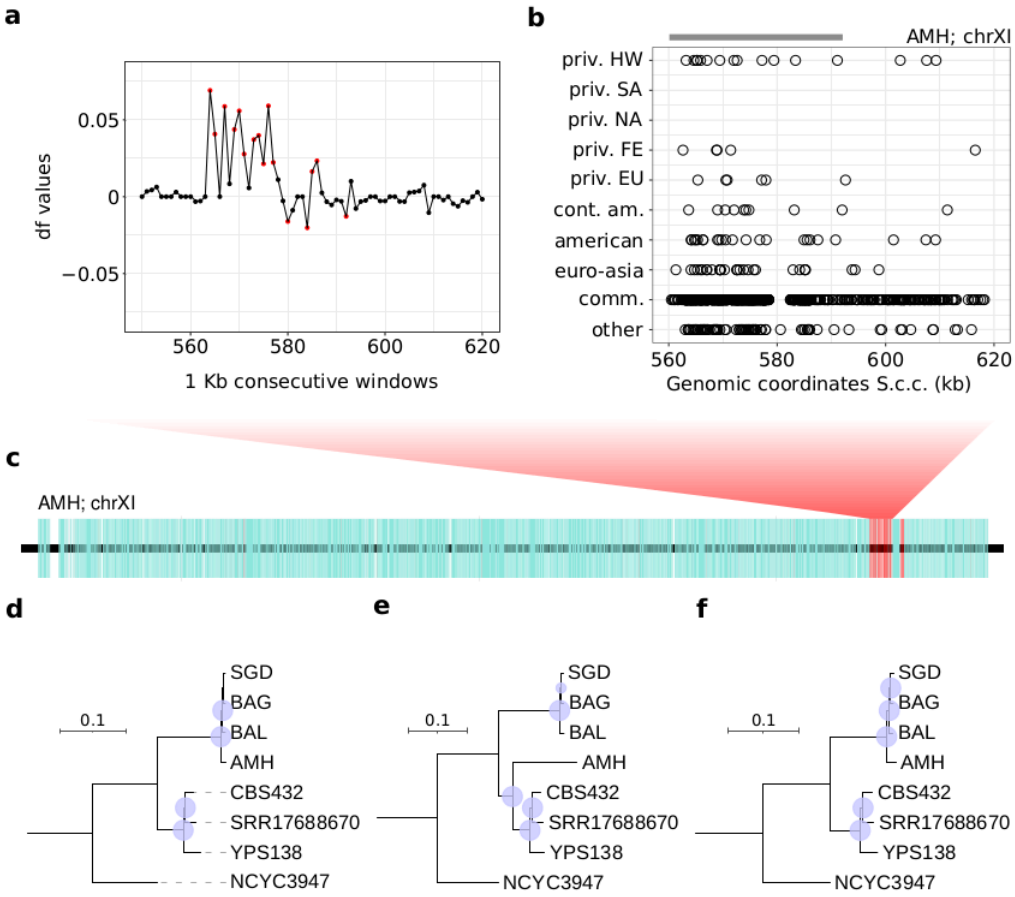

Figure 4.8 – **Unknown origin introgression on chromosome XI in AMH. a**, df statistics support the presence of the introgression on chromosome XI. Red dots represent genomics windows characterised by an absolute value of df 5 folds or higher compared to the average value across the chromosome. **b**, Polymorphism ancestry plot shows that the origin of the chromosome XI introgression cannot be retraced to any *S. paradoxus* population, consistent with its origin from an unknown sister species. The grey bar indicates the position of the introgression. **c**, Introgression boundaries in chromosome XI on AMH. The red blocks represent homozygous introgressions while the blue blocks are homozygous for S. cerevisiae. **d-f**, Maximum likelihood phylogenies of the sequences spanning the genes *YKR064W-YKR078W* derived from de novo whole-genome assemblies (Yue et al. 2017, O'Donnell et al. 2022, Naseeb et al. 2018). SGD : *S. cerevisiae* reference genome, BAG : CHN-II *S. cerevisiae*, BAL : CHN-I *S. cerevisiae*, AMH : CHN-IX *S. cerevisiae*, CBS432 : European *S. paradoxus*, SRR17688670 : Chinese *S. paradoxus*, YPS138 : American *S. paradoxus* and NCYC3947 : *Saccharomyces jurei* (outgroup). The tree on panel **d** and **f** are derived from 25 kb flanking regions before and after the introgression, while the tree in panel **e** is the introgressed region. The blue circles at the nodes represent the percentage of SH-aLRT $\geq$ 95% and UFboot $\geq$ 95%.

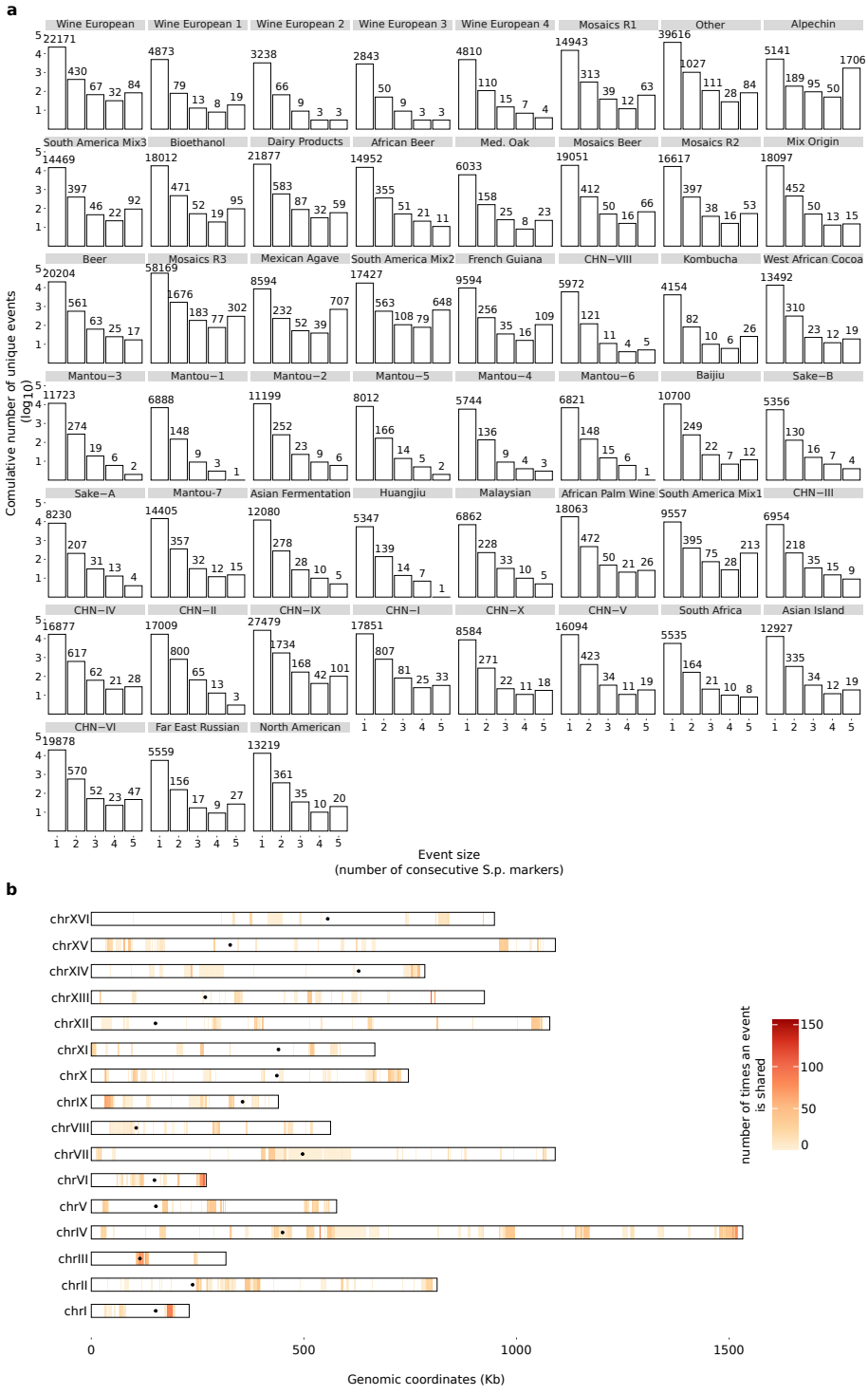

Figure 4.9 – **Introgression blocks sizes and location. a**, distribution of the diagnostic markers with *S. paradoxus* genotypes groped by size across different clades. For each clade, isolated markers and introgressed blocks with the same boundaries are counted once. Overlapping blocks with different boundary coordinates are counted as separate events. The last column included all the blocks with at least 5 consecutive *S. paradoxus* markers. The Y-axis is on $\log_{10}$ scale. The absolute value of the counts is indicated at the top of each column. The label "Other" indicates *S. cerevisiae* strains that could not be placed in a specific clade. **b**, physical positioning and frequency of the introgression blocks supported by ≥ 5 consecutive *S. paradoxus* markers across 1,459 out of 1,673 *S. cerevisiae* samples. The coloured scale reflects the number of times a specific block is shared across the 1,459 *S. cerevisiae* strains. Two blocks shared by more than 156 *S. cerevisiae* strains were dropped to 156 to allow the visualisation of less common events. The genomic coordinates indicate positions in the *S.c.c.* genome.

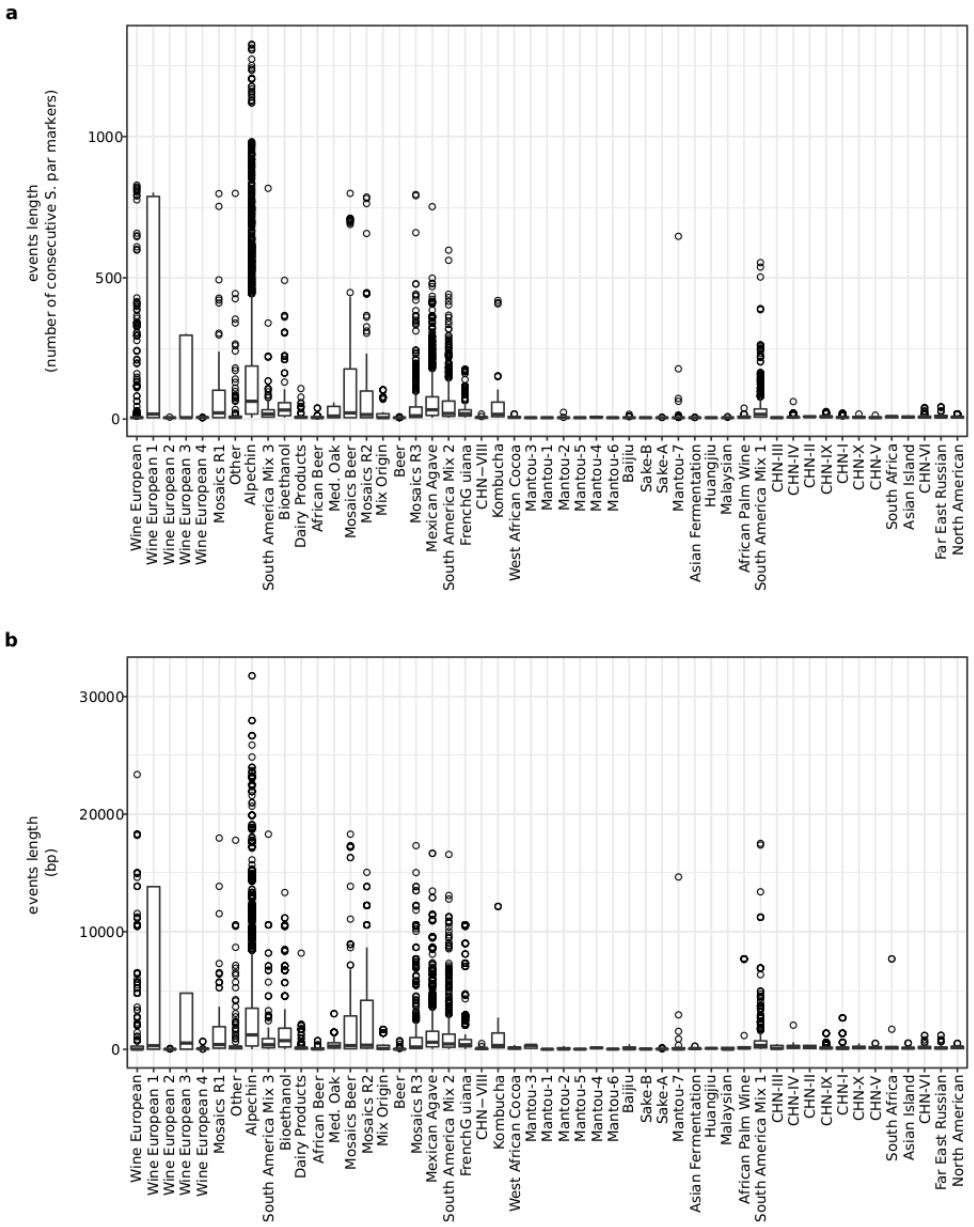

Figure 4.10 – **Introgression block size. a**, introgression lengths distributions across *S. cerevisiae* clades defined as the number of consecutive *S. paradoxus* markers. **b**, introgression lengths distributions across *S. cerevisiae* clades defined as the total length (in bp) of regions within *S. paradoxus* markers. Box, interquartile range (IQR) ; whiskers, 1.5×IQR ; thick horizontal line, median. Fully coloured data points beyond the whiskers are outliers. Empty dots represent values of clades with a number of values ≤ 20.

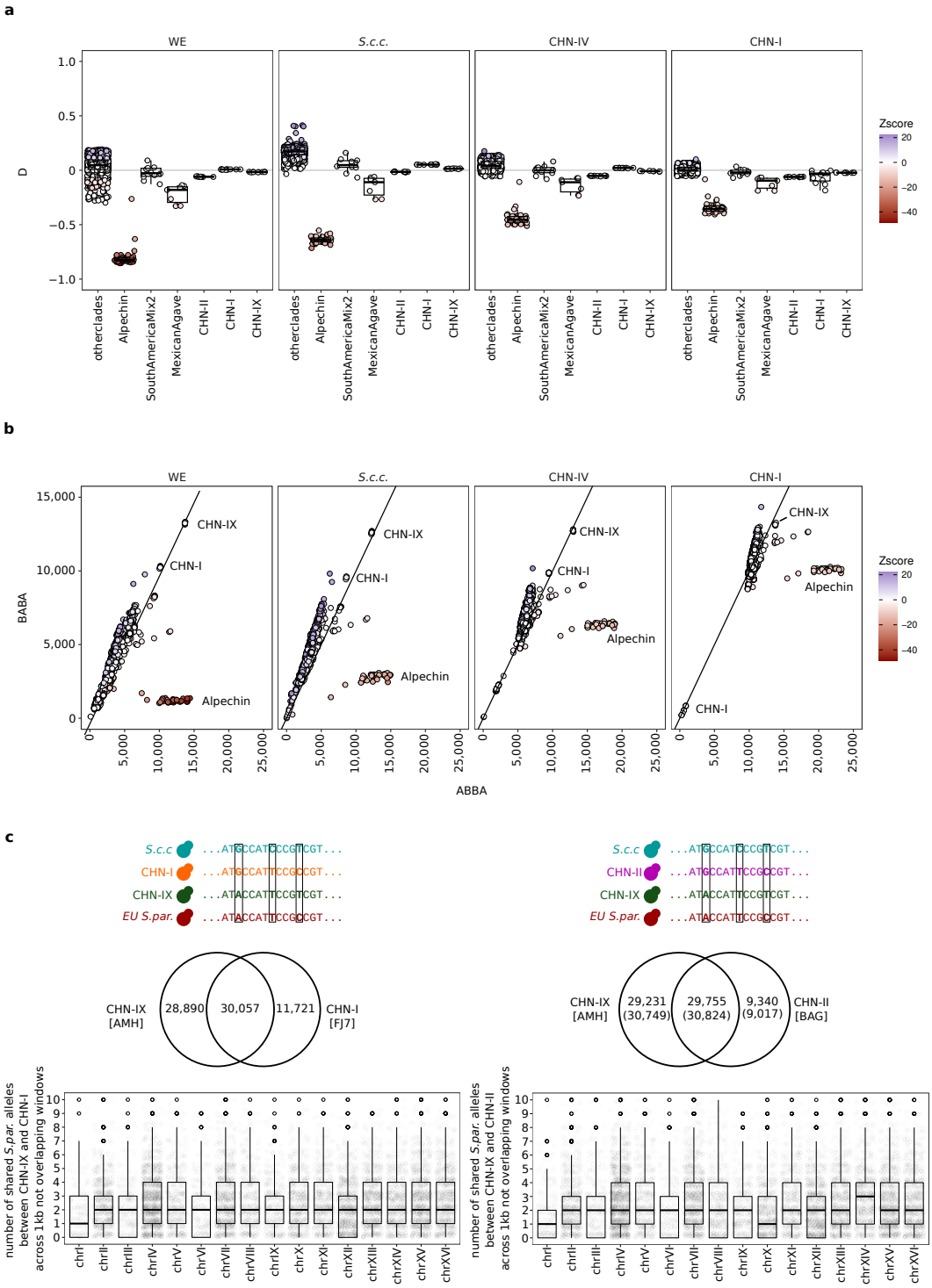

Figure 4.11 – **Patterson's D statistics and whole-genome alignments. a**, D values measured across the 1,671 *S. cerevisiae* collection using different quartet input arrangements. The strains on the top of each plot represent the P1 (WE : ADS, CHN-IV : BJ3 and CHN-I : BAL) population. P3 and P4 (Outgroup) are fixed and represented by the European *S. paradoxus* (CBS432) and *S. jurei*, respectively. Multiple D values are calculated, for each sample, by mean jack-knife resampling of genomic blocks. The gradient colour reflects the Z-score. D values associated with a Z-score equal to or greater than the absolute value of 3 are considered statistically significant and the null hypothesis of the absence of gene flow can be rejected. Box, interquartile range (IQR) ; whiskers, 1.5×IQR ; thick horizontal line, median. Outliers are not shown. **b**, Absolute counts of both ABBA and BABA sites across the *S. cerevisiae* strains. **c**, *S. par.* alleles at polymorphic positions. On top, the cartoons represent examples of the polymorphic sites taken into account ; in the middle, the Venn diagrams show the private and shared *S. paradoxus* alleles between the Chinese isolates ; at the bottom, the boxplots show the distribution of the only shared *S. par.* alleles between the Chinese isolates. In the squared brackets the name of the strains ; in the round brackets the number of sites obtained using the *de novo* genome assemblies51 of BAG, AMH and CBS432 isolates depicted in the cartoon above. Box, interquartile range (IQR) ; whiskers, 1.5×IQR ; thick horizontal line, median. Outliers are not shown.

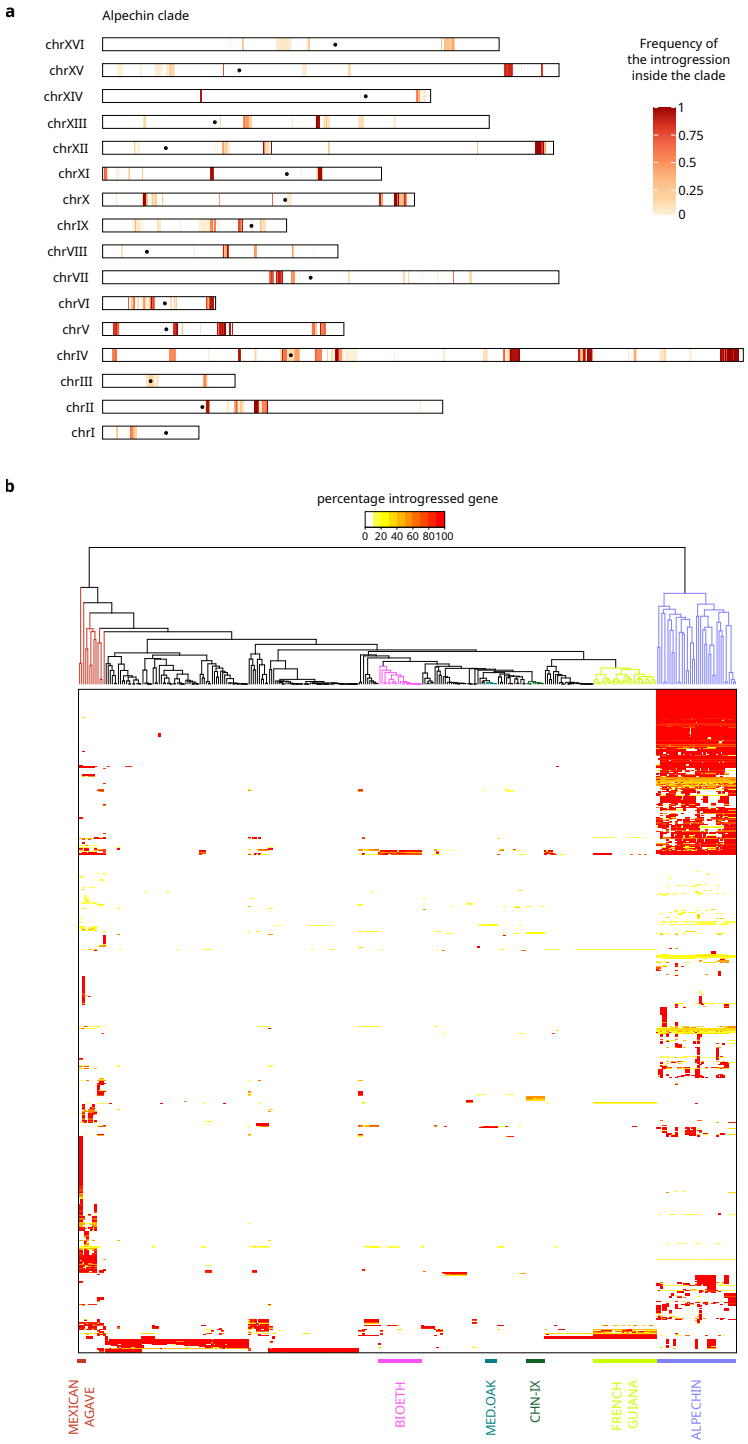

Figure 4.12 – **Highly introgressed clades. a**, Frequency and genomic position of the introgressions detected across the Alpechin (N=40, on the left) and Mexican Agave (N=7, on the right) clades. **b**, The heatmap shows the *S. paradoxus* percentage of introgressed genes across *S. cerevisiae* strains with at least 5 introgressed genes. To reduce the complexity of the heatmap, we retained only the genes introgressed in at least one of the remaining *S. cerevisiae* strains (322 strains x 1305 genes). The hierarchical clustering was performed across the strains (columns). We coloured the branches of the strains for which the patterns strictly follow the clade division. Because of the dimension of the heatmap, to better visualise the patterns, the heatmap cells with a percentage value of gene introgressed between 0 and 10% were removed (white areas inside the heatmap).

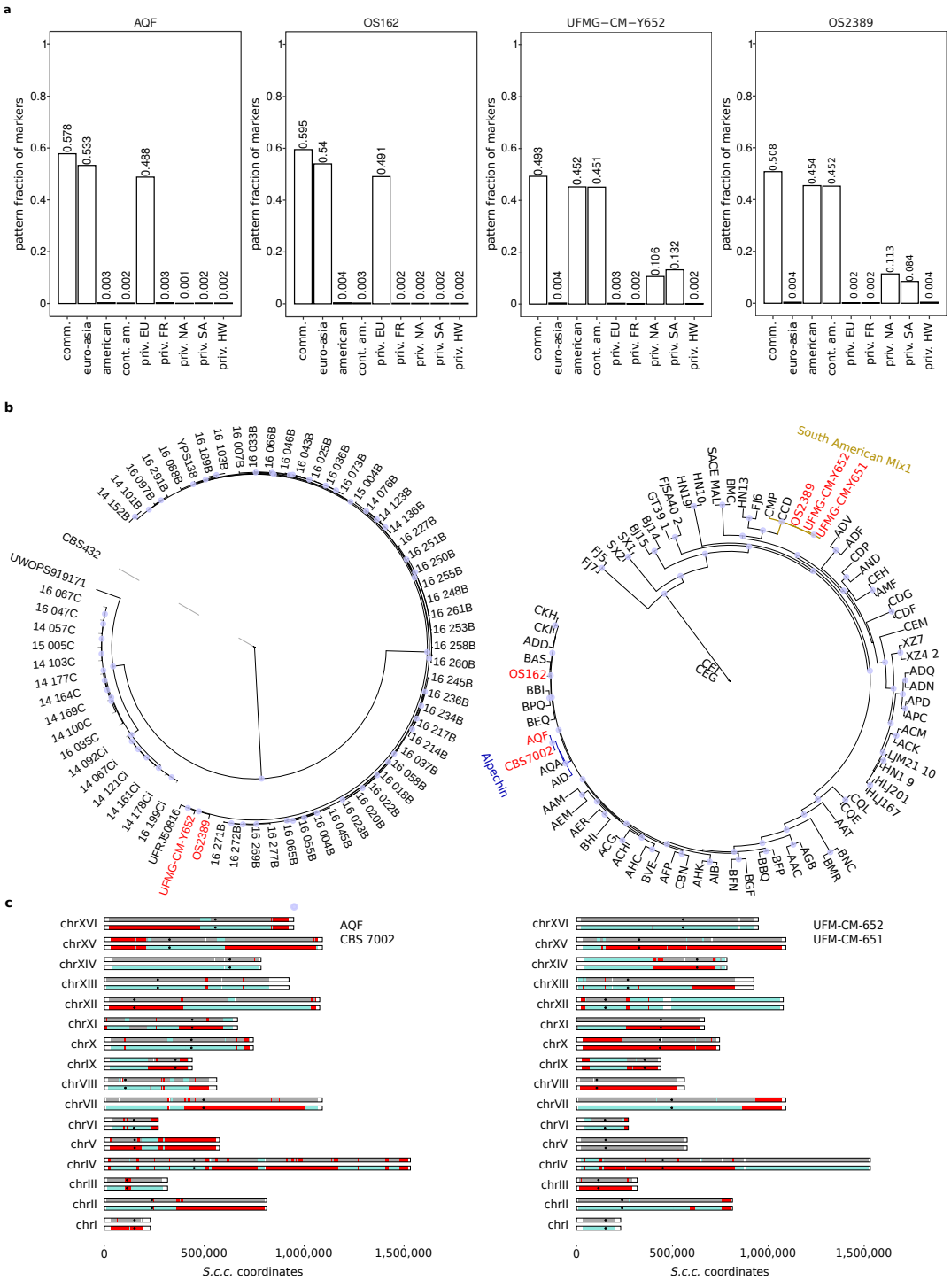

Figure 4.13 – **Hybrid subgenome ancestries. a**, Fraction of allelic pattern, at marker positions, detected across the *S. paradoxus* subgenomes of the four hybrid isolates, **b**, Maximum likelihood phylogenies of the *S. paradoxus* (on the left) and *S. cerevisiae* (on the right) subgenomes of hybrid isolates (red label). The *S. paradoxus* phylogeny was constructed including American *S. paradoxus* strains (Eberlein et al. 2019), while the *S. cerevisiae* phylogeny was constructed with a selection of *S. cerevisiae* strains from previous studies (Peter et al. 2018,Duan et al. 2018). The circle represents nodes with SH-aLRT $\geq$ 95% and UFboot $\geq$ 95%. The size of the circle at the nodes reflects the value of SH-aLRT and UFboot (from 95 to 100). **c**, genomic profiles of hybrid-descendant pairs from Spain (on the left) and Brazil (on the right). Details of the strain and genome origin are given in the methods (section *Saccharomyces* collections). Red and blue blocks represent homozygous *S. paradoxus* and *S. cerevisiae* regions respectively, while grey blocks are heterozygous regions.

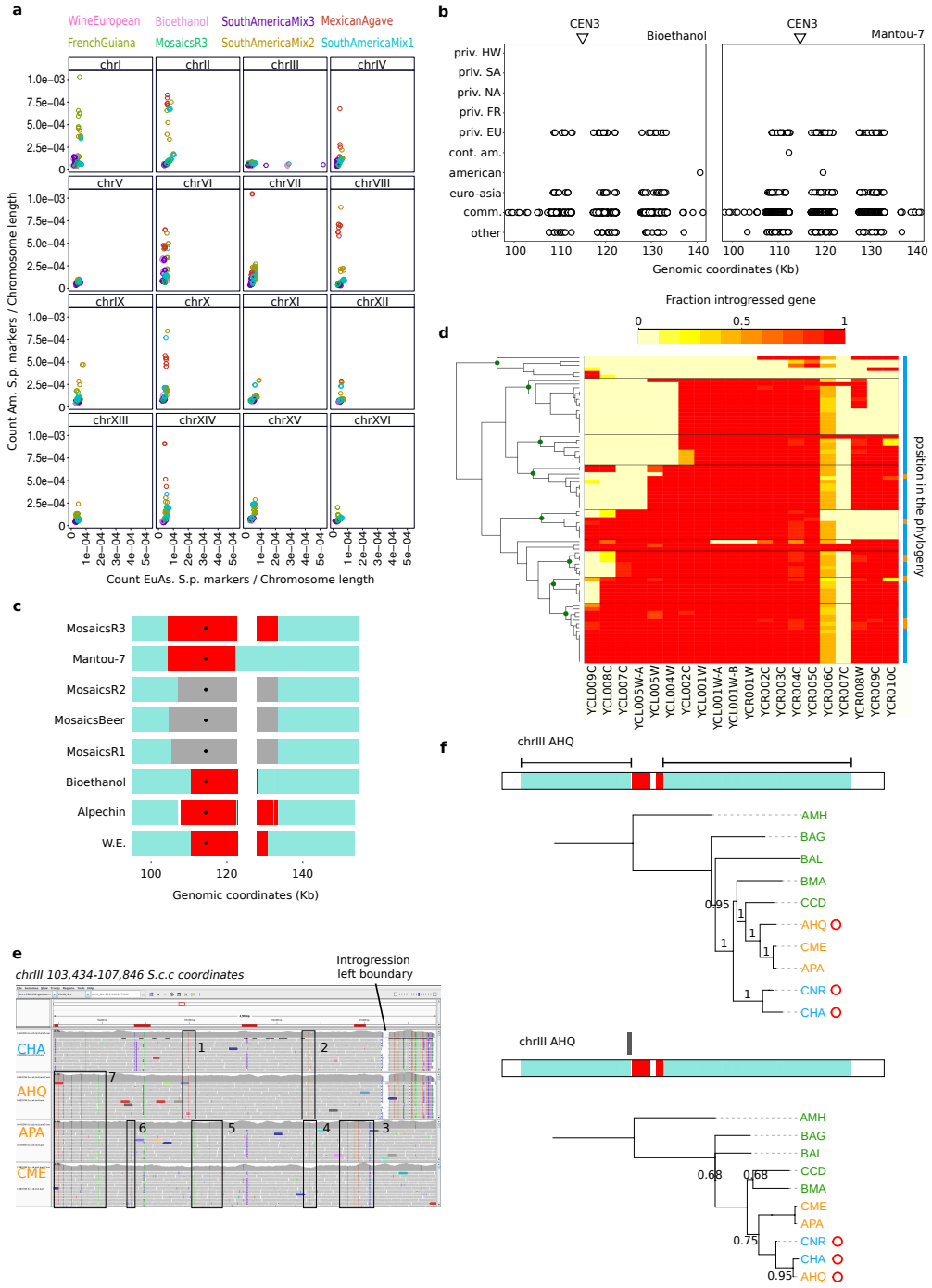

Figure 4.14 – **Ancient introgression on chromosome III. a**, Density of *S. paradoxus* diagnostic markers across a subset of introgressed strains. The x-axis and y-axis respectively indicate the density of *S. paradoxus* diagnostic markers that support a Euroasiatic or American origin of the introgressions. Chromosome III has diagnostic markers with Euroasiatic ancestry and also in clades with genome-wide introgressed blocks with American origins. **b**, Ancestry plot of diagnostic markers within the chromosome III introgression block in Brazilian bioethanol and a Mantou 7 strains. The origin of diagnostic markers is attributed to different *S. paradoxus* populations (listed along the Y-axis). Genomic coordinates are from the *S.c.c.* genome. **c**, Map of introgressed blocks detected around chromosome III centromere (black dot) across different *S. cerevisiae* strains. Red and grey colours represent homozygous and heterozygous *S. paradoxus* introgression respectively, blue indicates the *S. cerevisiae* genome. **d**, heatmap with introgressed genes (x-axis) detected on the region encompassing the centromere of chromosome III across different *S. cerevisiae* strains. The strains are clustered by their introgression profile (left side), while the coloured bar (right side) indicates their phylogenetic assignment illustrated in 4.4b. **e**, Alignment of the short reads against *S.c.c.* of a selection of four strains. CHA, Wine European, AHQ Mantou 7, APA and CME Sake strains. The rectangles 1 and 2 highlight shared SNPs between CHA and AHQ which are absent in the Sake strains. The rectangles 3, 4, 5, and 6 highlight private SNPs of the Sake clade. The rectangle 7 encloses the SNPs in AHQ shared with the Sake strains but absent in the Wine European strain (CHA). **f**, Maximum likelihood phylogenies of the arms of chromosome III, external to the introgressed block, and the 2 kb left-side flanking region of the introgression for a selection of strains. The introgression and the subtelomeric/telomeric regions are excluded. Red circles indicate strains with introgressed CEN3.

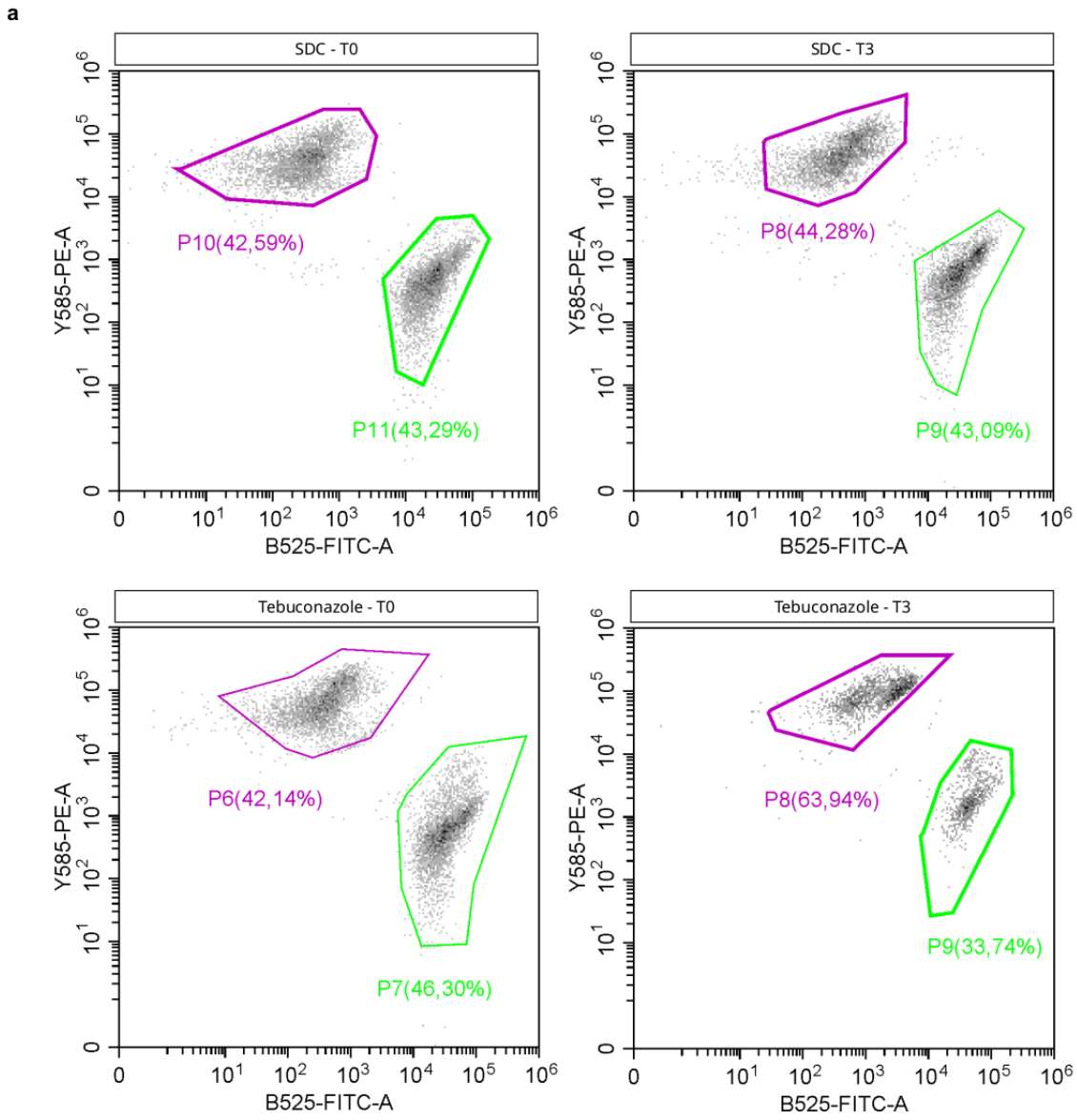

Figure 4.15 – **FACS experiment results**. **a**, Flow-cytometry cell distribution at time T_0 (left panels) and T_3 (right panels) in SDC (panels at the top) and Tebuconazole (0.0037 mg/mL, panels at the bottom). The gating of the T_0 population was used to monitor the ratio of RFP/GFP. The violet shapes surround the population of cells with the *PAD1-1FDC1* alpechin introgressed copy labelled with mRuby2 ($DBVPG6765^{ALP}$) while the green shapes surround the population of cells with the *PAD1-FDC1 S. cerevisiae* copy labelled with yGamillus ($DBVPG6765^{WT}$).

CHAPTER 5

Conclusions and Perspectives

5.1 Evidence of ancestral alleles in highly divergent wild Chinese population and evolutionary modes

By studying the distribution of the alleles segregating at specific diagnostic genomic positions, which discriminate most of the today's known *S. cerevisiae* and *S. paradoxus* populations, we reported evidence of ancestral variation persisting in contemporary wild *S. cerevisiae* populations sampled in primaeval forests in China, especially in the CHN-IX/Taiwanese population. Peris et al. 2023 reported cases of loss of monophyly in the relation ((*S.cer*, CHN-IX), *S.par*) detecting discrepancies in which CHN-IX occurred either 1) as an outgroup of both the species [((*S.cer*, *S.par*), CHN-IX)] or 2) an early split of *S. paradoxus* species [((CHN-IX, *S.par*), *S.cer*)]. Thus, we proposed that the retention of ancestral alleles might result from a low number of generations, moderate population bottleneck and persistence in the same environment since the split from the last common ancestor *S.cer-S.par*. This evidence raises questions about the evolutionary modes of wild populations because the persistence of ancestral variation is expected as a result of neutrally segregating sites when populations remain large. Estimates on the population size of wild populations are absent in *S. cerevisiae* but, for *S. paradoxus*, Tsai et al. 2008 estimated the effective population size of two wild populations to be in the order of $\approx 10^7$. In addition, wild *S. cerevisiae* populations are highly monophyletic, homozygous, show small fractions of genes under selection (Duan et al. 2018) and when found living in sympatry do not show signs of intercrossing among them. As reported in Bai et al. 2022, CHN-III and CHN-V *S. cerevisiae* populations collected from Wuzhi Mountain in the tropical island of Hainan do not show evidence of admixture and this is

true also for the CHN-I population sampled in the same island. This might be explained by both 1) the presence of an undetected reproductive barrier and 2) rare opportunities for outcrossing which nevertheless contrast with their excellent ability to sporulate, i.e. to reproduce meiotically. Q.-M. Wang et al. 2012 was able to explain part of the reproductive isolation across some isolates due to structural variations (SVs). The recent release of 142 *S. cerevisiae* telomere-to-telomere assemblies and future larger sequencing efforts will offer the opportunity to investigate in detail the role of SVs in wild populations. *Saccharomyces* species reproduce both mitotically and meiotically, but we have a limited understanding of the frequencies at which these processes occur and the impact they have on the evolution of the species in the wild. Future studies should aim to further quantify the frequency and the possible outcomes of the meiotic process in the wild, which is expected to be prevalent as wild isolates face recurrent environmental changes. Another challenging aspect concerns the investigation of the phenotypic impact of *S. paradoxus* alleles segregating in the *S. cerevisiae* background. An interesting opportunity would be the comparison across wild populations isolated across different continents such as the Ecuadorean, North American and Mediterranean *S. cerevisiae* which represent the completely wild remnants postdating the single out-of-China event. The approaches we exploited take advantage of the datasets available at the time this project started, but identification of the ILS sites should be reevaluated as *de novo* assemblies are released. Software such as ASTRAL can be used to estimate the population tree, given a set of gene trees, under the multi-species coalescent model and can provide the opportunity to evaluate and resolve population relationships inside the *S. cerevisiae* phylogeny, accounting for ILS. Moreover, the coalescent Hidden Markov Model approach can be used across multi-sequence whole-genome alignments to identify ILS regions but also to estimate population parameters and diversification times that, as shown in Rivas-González et al. 2023, are as accurate as fossil estimates in primates. These approaches are potentially promising for the study of *Saccharomyces* evolution, in the absence of fossil records.

5.2 *S. cerevisiae* x *S. paradoxus* hybrids are rare but occur occasionally

Because of the high sequence divergence between *S. cerevisiae* and *S. paradoxus* ($\approx$ 12%), the assessment of polymorphic positions across the genomes easily allows the identification of hy-

brid strains. In this project, we characterized 4 *S. cerevisiae* x *S. paradoxus* hybrids. The isolates collected in Brazil by Barbosa et al. 2016, UFMG-CM-Y652 and UFMG-CM-Y651, represent strong evidence of the presence of a complex hybrid zone in the wild between *S. cerevisiae* and *S. paradoxus* as the distribution of their parental ancestries varies notably along the genome. The *S. paradoxus* ancestry of the Hawaiian OS2389 hybrid is not traceable to the only known Hawaiian *S. paradoxus* (UWOPS91−917.1), but to closely related North/South American *S. paradoxus* (i.e. SpB). However, the *S. paradoxus* ancestry of these hybrids do not match the ancestry of the *S. paradoxus* introgressions detected in *S. cerevisiae* isolates collected from close areas, suggesting the possibility of co-occurring unsampled *S. paradoxus* populations in the area. Similarly, the *S. cerevisiae* component groups with the South American mix 1 clade suggesting that the hybrid either migrated to Hawaii or that both the parental isolates are also present in the Hawaiian archipelagos. The scarce evidence of both of *S. cerevisiae* and *S. paradoxus* from Hawaii and South America do not allow us to provide a conclusive explanation of its origin. Interestingly, OS2389 was sampled in Hawaii from a koa tree, one of the most abundant endemic plants in the archipelagos which represents a unique niche. The UK hybrid OS162 collected from an oak tree shows Wine/European *S. cerevisiae* and European *S. paradoxus* ancestries in line with its geographical location. It is the only hybrid with full ancestry components of both the species and harbours one copy of *S. cerevisiae* and two copies of *S. paradoxus* subgenome. Taken together, these isolates demonstrate that the occurrence of hybridization in the wild is evident but rare. Although the strains that we described span a large geographical range, they have been sampled during a long time window of over 50 years, from the 1970s when the Hawaiian isolate was sampled. Future efforts in sampling from wild environments could help us to understand the ecology, geographical boundaries and prevalence of hybridization of *Saccharomyces* yeasts.

5.3 Introgressed regions provide insights into common evolutionary histories and offer opportunities for adaptation to diverse ecological niches.

The detection of the introgression on chromosome III across isolates from domestic environments reinforces the evidence of secondary contact between the European *S. paradoxus* population and, probably, the European *S. cerevisiae* population, from which most of the non-Chinese domestic isolates derived. These interactions resulted in a parallel exchange of DNA in which *S. cerevisiae* retained the introgression on chromosome III and the European *S. paradoxus* the introgression on chromosome XIV (Liti et al. 2006). This latter sweep reached fixation in the European *S. paradoxus* population, suggesting either a functional role or a recent strong bottleneck. In the case of the introgression on chromosome III of *S. cerevisiae*, although it occurs at low frequencies across isolates of different clades, we can not rule out a functional implication of the event. On the other hand, the introgression encompassing the *PAD1-FDC1* gene pair happened independently in two distinct events of hybridization, in Europe and America, representing a clear case of convergent evolution in the Alpechin and Bioethanol populations, respectively. The proof of its functional effect improving the fitness in isolates subjected to stressful environmental conditions characteristic of olive oil manufacturing is the first evidence of the adaptive role of *S. paradoxus* introgression on *S. cerevisiae* and likely contributed to the adaptation of the population to this niche. Other introgressed genes, especially in the Alpechin clade, are good candidates for functional investigation and future studies should aim to unveil their impact. However, the quantitative contributions of these events to phenotypic traits are nontrivial to disentangle, given the complexity of the phenotypes and the possible epistatic interactions among different introgression blocks, that may need to be tested in combination.

Acknowledgement

I express my deepest gratitude to Gianni Liti for his guidance and teachings. The patience, dedication, and attention to detail that Gianni dedicates to his work are reasons for inspiration. The positive and healthy professional environment he has created in the lab is one of his most remarkable accomplishments, and it has been a great help to me during this long journey. I want to thank the jury members Etienne Danchin, Tatiana Giraud, Delphine Sicard, and Anders Bergström for their work in reviewing this work and having constructive discussions during the defence. Completing my dissertation would not have been possible without the financial support from the Agence Nationale de la Recherche (ANR), Fondation pour la Recherche Médicale (FRM), and the fellowship program Region PACA. Thanks to Sakshi for all the support, inside and outside the lab, and for being a partner in crime along this adventure ; thank you for all the moments that ended well and for those that ended worse but that in the end have brought only the best. Thanks to Matteo and Lorenzo for the mentoring and the long, crazy discussions about variables, methods and unlikely explanations to even less likely questions. Thanks for providing apartments and medicines too. Endless gratitude to Agnes for her unlimited assistance in everything since 2018, the year I first set foot in the lab. Thanks to Melania, Ben, Chiara, and Simone for the chats in the canteen, the anxiety, the days at the beach, the anxiety, the dinners out, the anxiety, the failed experiments, and the successful ones, from which I could only indirectly feel the frustration and the joy, but, above all, the perseverance and dedication. Thanks to Danielle for all the time she carved out to help me prepare the thesis defence presentation ; it was a very instructive experience. Thanks to Xanita, the best benchmate I ever had. Thanks to Federica and Simon, two great brains, enjoying your PhD journey. Thank you to everyone who has passed through or been a part of the lab for a short or longer time, thank you to all the colleagues at IRCAN ! Un ringraziamento immenso a Sara, Fabio, Lorenzo e Matteo, amici di una vita sempre presenti, a tutta la famiglia e i parenti per il sotegno continuo ovunque vada, qualunque cosa faccia. Thanks !

Références

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., … Sampaio, J. P. (2015, novembre). A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular Ecology*, *24*(21), 5412–5427. Consulté le 2021-05-15, sur http://doi.wiley.com/10.1111/mec.13341 doi: 10.1111/mec.13341

Al Safadi, R., Weiss-Gayet, M., Briolay, J., & Aigle, M. (2010, septembre). A polyploid population of Saccharomyces cerevisiae with separate sexes (dioecy). *FEMS Yeast Research*, *10*(6), 757–768. Consulté le 2023-09-23, sur https://doi.org/10.1111/j.1567-1364.2010.00660.x doi: 10.1111/j.1567-1364.2010.00660.x

Alsammar, H., & Delneri, D. (2020, mars). An update on the diversity, ecology and biogeography of the Saccharomyces genus. *FEMS Yeast Research*, *20*(3), foaa013. Consulté le 2023-09-12, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7150579/ doi: 10.1093/femsyr/foaa013

Avise, J. C., & Robinson, T. J. (2008, juin). Hemiplasy : A New Term in the Lexicon of Phylogenetics. *Systematic Biology*, *57*(3), 503–507. Consulté le 2023-06-14, sur https://academic.oup.com/sysbio/article/57/3/503/1666092 doi: 10.1080/10635150802164587

Avise, J. C., Shapiva, J. F., Daniel, W., & Lansman, A. (1983, décembre). Mitochondrial DNA Differentiation during the Speciation Process in Peromyscusl. *Molecular Biology and Evolution*, *1*(1), 38–56. Consulté sur https://academic.oup.com/mbe/article/1/1/38/1508292?login=true doi: 10.1093/oxfordjournals.molbev.a040301

Bai, F.-Y., Han, D.-Y., Duan, S.-F., & Wang, Q.-M. (2022, janvier). The Ecology and Evolution of the Baker's Yeast Saccharomyces cerevisiae. *Genes*, *13*(2), 230. Consulté le 2023-09-21, sur https://www.mdpi.com/2073-4425/13/2/230 doi: 10.3390/genes13020230

Barbosa, R., Almeida, P., Safar, S. V., Santos, R. O., Morais, P. B., Nielly-Thibault, L., … Sampaio, J. P. (2016, février). Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. *Genome Biology and Evolution*, *8*(2), 317–329. Consulté le 2021-05-15, sur https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evv263 doi: 10.1093/gbe/evv263

Barbosa, R., Pontes, A., Santos, R. O., Montandon, G. G., de Ponzzes-Gomes, C. M., Morais, P. B., … Sampaio, J. P. (2018, août). Multiple Rounds of Artificial Selection Promote Microbe Secondary Domestication—The Case of Cachaça Yeasts. *Genome Biology and Evolution*, *10*(8), 1939–1955. Consulté le 2021-05-15, sur https://academic.oup.com/gbe/article/10/8/1939/5047776 doi: 10.1093/gbe/evy132

Bergin, S. A., Allen, S., Hession, C., Ó Cinnéide, E., Ryan, A., Byrne, K. P., … Butler, G. (2022, décembre). Identification of European isolates of the lager yeast parent *Saccharomyces eubayanus*. *FEMS Yeast Research*, *22*(1), foac053. Consulté le 2023-09-13, sur https://academic.oup.com/femsyr/article/doi/10.1093/femsyr/foac053/6874782 doi: 10.1093/femsyr/foac053

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., … Tyler-Smith, C. (2020, mars). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, *367*(6484), eaay5012. Consulté le 2021-08-25, sur https://www.sciencemag.org/lookup/doi/10.1126/science.aay5012 doi: 10.1126/science.aay5012

Borneman, A. R., Forgan, A. H., Kolouchova, R., Fraser, J. A., & Schmidt, S. A. (2016, avril). Whole Genome Comparison Reveals High Levels of Inbreeding and Strain Redundancy Across the Spectrum of Commercial Wine Strains of *Saccharomyces cerevisiae*. *G3 Genes|Genomes|Genetics*, *6*(4), 957–971. Consulté le 2023-09-13, sur https://academic.oup.com/g3journal/article/6/4/957/6055620 doi: 10.1534/g3.115.025692

Bozdag, G. O., Ono, J., Denton, J. A., Karakoc, E., Hunter, N., Leu, J.-Y., & Greig, D. (2019, septembre). *Engineering recombination between diverged yeast species reveals genetic incompatibilities* (preprint). Evolutionary Biology. Consulté le 2022-11-14, sur http://biorxiv.org/lookup/doi/10.1101/755165 doi: 10.1101/755165

Chao, L., & Carr, D. E. (1993, avril). THE MOLECULAR CLOCK AND THE RELATIONSHIP BETWEEN POPULATION SIZE AND GENERATION TIME. *Evolution*, *47*(2), 688–690. Consulté le 2023-09-26, sur https://doi.org/10.1111/j.1558-5646.1993.tb02124.x doi: 10.1111/j.1558-5646.1993.tb02124.x

Christian Lexer, L. H. R. (2003). *Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization.* Consulté le 2023-09-08, sur https://www.science.org/doi/10.1126/science.1086949 doi: 10.1126/science.1086949

Clark, A., Dunham, M. J., & Akey, J. M. (2022, août). *The genomic landscape of Saccharomyces paradoxus introgression in geographically diverse Saccharomyces cerevisiae strains* (preprint). Evolutionary Biology. Consulté le 2022-10-14, sur http://biorxiv.org/lookup/doi/10.1101/2022.08.01.502362 doi: 10.1101/2022.08.01.502362

Coi, A. L., Bigey, F., Mallet, S., Marsit, S., Zara, G., Gladieux, P., … Legras, J. L. (2017). Genomic signatures of adaptation to wine biological ageing conditions in biofilm-forming flor yeasts. *Molecular Ecology*, *26*(7), 2150–2166. Consulté le 2021-05-15, sur https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14053 (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14053) doi: https://doi.org/10.1111/mec.14053

Cruickshank, T. E., & Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, *23*(13), 3133–3157. Consulté le 2023-09-11, sur https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12796 (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12796) doi: 10.1111/mec.12796

Darling, A. C., Mau, B., Blattner, F. R., & Perna, N. T. (2004, juillet). Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394–1403. Consulté le 2021-07-09, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC442156/ doi: 10.1101/gr.2289704

De Chiara, M., Barré, B. P., Persson, K., Irizar, A., Vischioni, C., Khaiwal, S., … Liti, G. (2022, février). Domestication reprogrammed the budding yeast life cycle. *Nature Ecology & Evolution*, *6*(4), 448–460. Consulté le 2022-10-25, sur https://www.nature.com/articles/s41559-022-01671-9 doi: 10.1038/s41559-022-01671-9

De Queiroz, K. (2007, décembre). Species Concepts and Species Delimitation. *Systematic Biology*, *56*(6), 879–886. Consulté le 2023-09-08, sur https://academic.oup.com/sysbio/article/56/6/879/1653163 doi: 10.1080/10635150701701083

Duan, S.-F., Han, P.-J., Wang, Q.-M., Liu, W.-Q., Shi, J.-Y., Li, K., ... Bai, F.-Y. (2018, juillet). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature Communications*, *9*(1), 2690. Consulté le 2021-05-15, sur https://www.nature.com/articles/s41467-018-05106-7 (Number : 1 Publisher : Nature Publishing Group) doi: 10.1038/s41467-018-05106-7

Duina, A. A., Miller, M. E., & Keeney, J. B. (2014, mai). Budding Yeast for Budding Geneticists : A Primer on the Saccharomyces cerevisiae Model System. *Genetics*, *197*(1), 33–48. Consulté le 2023-09-23, sur https://doi.org/10.1534/genetics.114.163188 doi: 10.1534/genetics.114.163188

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011, août). Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. Consulté le 2021-07-09, sur https://academic.oup.com/mbe/article/28/8/2239/1052492 (Publisher : Oxford Academic) doi: 10.1093/molbev/msr048

Dutheil, J. Y., Ganapathy, G., Hobolth, A., Mailund, T., Uyenoyama, M. K., & Schierup, M. H. (2009, septembre). Ancestral Population Genomics : The Coalescent Hidden Markov Model Approach. *Genetics*, *183*(1), 259–274. Consulté le 2023-08-14, sur https://academic.oup.com/genetics/article/183/1/259/6063098 doi: 10.1534/genetics.109.103010

D'Angiolo, M., De Chiara, M., Yue, J.-X., Irizar, A., Stenberg, S., Persson, K., ... Liti, G. (2020, novembre). A yeast living ancestor reveals the origin of genomic introgressions. *Nature*, *587*(7834), 420–425. Consulté le 2021-05-15, sur https://www.nature.com/articles/s41586-020-2889-1 (Number : 7834 Publisher : Nature Publishing Group) doi: 10.1038/s41586-020-2889-1

Eberlein, C., Hénault, M., Fijarczyk, A., Charron, G., Bouvier, M., Kohn, L. M., ... Landry, C. R. (2019, février). Hybridization is a recurrent evolutionary stimulus in wild yeast speciation. *Nature Communications*, *10*(1), 923. doi: 10.1038/s41467-019-08809-7

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., ... Mallet, J. (2019). Genomic architecture and introgression shape a butterfly radiation.

Eizenga, J. M., Novak, A. M., Sibbesen, J. A., Heumos, S., Ghaffaari, A., Hickey, G., ... Garrison, E. (2020, août). Pangenome graphs. *Annual review of genomics and human genetics*, *21*, 139–162. Consulté le 2023-09-23, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8006571/ doi: 10.1146/annurev-genom-120219-080406

Ernst Mayr. (1942). *Systematics And The Origin Of Species*. Consulté le 2023-09-08, sur http://archive.org/details/in.ernet.dli.2015.20284

Erny, C., Raoult, P., Alais, A., Butterlin, G., Delobel, P., Matei-Radoi, F., ... Legras, J. L. (2012, mai). Ecological Success of a Group of Saccharomyces cerevisiae/Saccharomyces kudriavzevii Hybrids in the Northern European Wine-Making Environment. *Applied and Environmental Microbiology*, *78*(9), 3256–3265. Consulté le 2023-09-13, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346444/ doi: 10.1128/AEM.06752-11

Feng, S., Bai, M., Rivas-González, I., Li, C., Liu, S., Tong, Y., ... Zhang, G. (2022, mai). Incomplete lineage sorting and phenotypic evolution in marsupials. *Cell*, *185*(10), 1646–1660.e18.

Consulté le 2023-08-07, sur https://linkinghub.elsevier.com/retrieve/pii/S0092867422003440 doi: 10.1016/j.cell.2022.03.034

Gallone, B., Steensels, J., Mertens, S., Dzialo, M. C., Gordon, J. L., Wauters, R., ... Verstrepen, K. J. (2019, novembre). Interspecific hybridization facilitates niche adaptation in beer yeast. *Nature Ecology & Evolution*, *3*(11), 1562–1575. Consulté le 2021-03-06, sur http://www.nature.com/articles/s41559-019-0997-9 doi: 10.1038/s41559-019-0997-9

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., ... Verstrepen, K. J. (2016, septembre). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell*, *166*(6), 1397–1410.e16. doi: 10.1016/j.cell.2016.08.020

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., ... Durbin, R. (2018, octobre). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, *36*(9), 875–879. Consulté le 2023-09-23, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6126949/ doi: 10.1038/nbt.4227

Gascuel, O. (1997, juillet). BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695. Consulté le 2022-11-15, sur https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025808 doi: 10.1093/oxfordjournals.molbev.a025808

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., & Malinverni, R. (2016, janvier). regioneR : an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, *32*(2), 289–291. Consulté le 2023-01-05, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708104/ doi: 10.1093/bioinformatics/btv562

Giannakou, K., Visinoni, F., Zhang, P., Nathoo, N., Jones, P., Cotterrell, M., ... Delneri, D. (2021, décembre). Biotechnological exploitation of Saccharomyces jurei and its hybrids in craft beer fermentation uncovers new aroma combinations. *Food Microbiology*, *100*, 103838. Consulté le 2023-09-13, sur https://www.sciencedirect.com/science/article/pii/S0740002021001039 doi: 10.1016/j.fm.2021.103838

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., ... Sampaio, J. (2016, octobre). Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts. *Current Biology*, *26*(20), 2750–2761. Consulté le 2021-05-15, sur https://linkinghub.elsevier.com/retrieve/pii/S0960982216309848 doi: 10.1016/j.cub.2016.08.040

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... Paabo, S. (2010, mai). A Draft Sequence of the Neandertal Genome. *Science*, *328*(5979), 710–722. Consulté le 2021-07-13, sur https://www.sciencemag.org/lookup/doi/10.1126/science.1188021 doi: 10.1126/science.1188021

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010, mai). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. Consulté le 2023-01-05, sur https://doi.org/10.1093/sysbio/syq010 doi: 10.1093/sysbio/syq010

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C. K., Braun, E. L., Braun, M. J., ... Yuri, T. (2008, juin). A Phylogenomic Study of Birds Reveals Their Evolutionary History. *Science*, *320*(5884), 1763–1768. Consulté le 2021-08-30, sur http://science.sciencemag

.org/content/320/5884/1763 (Publisher : American Association for the Advancement of Science Section : Report) doi: 10.1126/science.1157704

Hahn, M. W., & Hibbins, M. S. (2019, décembre). A Three-Sample Test for Introgression. *Molecular Biology and Evolution*, *36*(12), 2878–2882. Consulté le 2023-09-11, sur https://doi.org/10.1093/molbev/msz178 doi: 10.1093/molbev/msz178

Harrison, R. G., & Larson, E. L. (2014, décembre). Hybridization, Introgression, and the Nature of Species Boundaries. *Journal of Heredity*, *105*(S1), 795–809. Consulté le 2021-10-24, sur https://academic.oup.com/jhered/jhered/article/2961884/Hybridization, doi: 10.1093/jhered/esu033

He, P.-Y., Shao, X.-Q., Duan, S.-F., Han, D.-Y., Li, K., Shi, J.-Y., ... Bai, F.-Y. (2022). Highly diverged lineages of Saccharomyces paradoxus in temperate to subtropical climate zones in China. *Yeast*, *39*(1-2), 69–82. Consulté le 2023-09-23, sur https://onlinelibrary.wiley.com/doi/abs/10.1002/yea.3688 (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/yea.3688) doi: 10.1002/yea.3688

Hedrick, P. W. (2009). Conservation genetics and North American bison (Bison bison). *The Journal of Heredity*, *100*(4), 411–420. doi: 10.1093/jhered/esp024

Herskowitz, I. (1988). Life Cycle of the Budding Yeast Saccharomyces cerevisiae. *MICROBIOL. REV.*, *52*.

Hibbins, M. S., & Hahn, M. W. (2019, mars). The Timing and Direction of Introgression Under the Multispecies Network Coalescent. *Genetics*, *211*(3), 1059–1073. Consulté le 2023-08-19, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6404246/ doi: 10.1534/genetics.118.301831

Hillenmeyer, M. E., Fung, E., Wildenhain, J., Pierce, S. E., Hoon, S., Lee, W., ... Giaever, G. (2008, avril). The Chemical Genomic Portrait of Yeast : Uncovering a Phenotype for All Genes. *Science*, *320*(5874), 362–365. Consulté le 2023-02-09, sur https://www.science.org/doi/10.1126/science.1150021 doi: 10.1126/science.1150021

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018, février). UFBoot2 : Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. Consulté le 2023-01-05, sur https://doi.org/10.1093/molbev/msx281 doi: 10.1093/molbev/msx281

Hou, J., Friedrich, A., de Montigny, J., & Schacherer, J. (2014, mai). Chromosomal rearrangements as a major mechanism in the onset of reproductive isolation in Saccharomyces cerevisiae. *Current biology : CB*, *24*(10), 1153–1159. Consulté le 2023-09-23, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4067053/ doi: 10.1016/j.cub.2014.03.063

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., ... Nielsen, R. (2014, août). Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, *512*(7513), 194–197. Consulté le 2021-10-05, sur https://www.nature.com/articles/nature13408 (Bandiera_abtest : a Cg_type : Nature Research Journals Number : 7513 Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Genetic variation Subject_term_id : genetic-variation) doi: 10.1038/nature13408

Hull, D. L. (1980). Individuality and selection. *Annual review of ecology and systematics*, *11*(1), 311–332. (Publisher : Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA)

Hunter, N., Chambers, S. R., Louis, E. J., & Borts, R. H. (1996, avril). The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *The EMBO journal*, *15*(7), 1726–1733.

Jamoussi, B., Bedoui, A., Hassine, B. B., & Abderraba, A. (2005, janvier). Analyses of phenolic compounds occurring in olive oil mill wastewaters by GC–MS. *Toxicological & Environmental Chemistry*, *87*(1), 45–53. Consulté le 2022-10-18, sur http://www.tandfonline.com/doi/abs/10.1080/02772240400026757 doi: 10.1080/02772240400026757

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., ... Zhang, G. (2014, décembre). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*(6215), 1320–1331. Consulté le 2023-09-27, sur https://www.science.org/doi/10.1126/science.1253451 (Publisher : American Association for the Advancement of Science) doi: 10.1126/science.1253451

Jones, M. R., Mills, L. S., Alves, P. C., Callahan, C. M., Alves, J. M., Lafferty, D. J. R., ... Good, J. M. (2018, juin). Adaptive introgression underlies polymorphic seasonal camouflage in snowshoe hares. *Science*, *360*(6395), 1355–1358. Consulté le 2023-09-09, sur https://www.science.org/doi/10.1126/science.aar5273 doi: 10.1126/science.aar5273

Khan, T. N., & Croser, J. S. (2004, janvier). PEA | Overview. In C. Wrigley (Ed.), *Encyclopedia of Grain Science* (pp. 418–427). Oxford : Elsevier. Consulté le 2023-08-19, sur https://www.sciencedirect.com/science/article/pii/B0127654909001269 doi: 10.1016/B0-12-765490-9/00126-9

Kimura, M. (1991). The neutral theory of molecular evolution : A review of recent evidence. *The Japanese Journal of Genetics*, *66*(4), 367–386. doi: 10.1266/jjg.66.367

Kingman, J. F. C. (1982, septembre). The coalescent. *Stochastic Processes and their Applications*, *13*(3), 235–248. Consulté le 2023-06-21, sur https://www.sciencedirect.com/science/article/pii/0304414982900114 doi: 10.1016/0304-4149(82)90011-4

Kleindorfer, S., Custance, G., Peters, K. J., & Sulloway, F. J. (2019, juin). Introduced parasite changes host phenotype, mating signal and hybridization risk : Philornis downsi effects on Darwin's finch song. *Proceedings of the Royal Society B : Biological Sciences*, *286*(1904), 20190461. Consulté le 2023-09-25, sur https://royalsocietypublishing.org/doi/10.1098/rspb.2019.0461 (Publisher : Royal Society) doi: 10.1098/rspb.2019.0461

Klinke, H. B., Thomsen, A. B., & Ahring, B. K. (2004, novembre). Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. *Applied Microbiology and Biotechnology*, *66*(1), 10–26. Consulté le 2022-10-18, sur http://link.springer.com/10.1007/s00253-004-1642-2 doi: 10.1007/s00253-004-1642-2

Koufopanou, V., Lomas, S., Pronina, O., Almeida, P., Sampaio, J. P., Mousseau, T., ... Burt, A. (2020, septembre). Population Size, Sex and Purifying Selection : Comparative Genomics of Two Sister Taxa of the Wild Yeast Saccharomyces paradoxus. *Genome Biology and Evolution*, *12*(9), 1636–1645. Consulté le 2021-11-26, sur https://academic.oup.com/gbe/article/12/9/1636/5872529 doi: 10.1093/gbe/evaa141

Kuderna, L. F. K., Gao, H., Janiak, M. C., Kuhlwilm, M., Orkin, J. D., Bataillon, T., ... Bonet, T. M. (2023). A global catalog of whole-genome diversity from 233 primate species.

Langdon, Q. K., Peris, D., Baker, E. P., Opulente, D. A., Nguyen, H.-V., Bond, U., ... Hittinger, C. T. (2019, novembre). Fermentation innovation through complex hybridization of wild and domesticated yeasts. *Nature Ecology & Evolution*, *3*(11), 1576–1586. Consulté le 2021-03-06, sur http://www.nature.com/articles/s41559-019-0998-8 doi: 10.1038/s41559-019-0998-8

Leducq, J.-B., Nielly-Thibault, L., Charron, G., Eberlein, C., Verta, J.-P., Samani, P., ... Landry, C. R. (2016, janvier). Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature microbiology*, *1*, 15003. Consulté le 2021-07-05, sur https://doi.org/10.1038/nmicrobiol.2015.3 doi: 10.1038/nmicrobiol.2015.3

Lee, S., Lim, W. A., & Thorn, K. S. (2013, juillet). Improved Blue, Green, and Red Fluorescent Protein Tagging Vectors for S. cerevisiae. *PLoS ONE*, *8*(7), e67902. Consulté le 2022-09-29, sur https://dx.plos.org/10.1371/journal.pone.0067902 doi: 10.1371/journal.pone.0067902

Lee, T. J., Liu, Y.-C., Liu, W.-A., Lin, Y.-F., Lee, H.-H., Ke, H.-M., ... Tsai, I. J. (2022, mars). Extensive sampling of *Saccharomyces cerevisiae* in Taiwan reveals ecology and evolution of predomesticated lineages. *Genome Research*, genome ;gr.276286.121v2. Consulté le 2023-09-21, sur http://genome.cshlp.org/lookup/doi/10.1101/gr.276286.121 doi: 10.1101/gr.276286.121

Legras, J.-L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., ... Dequin, S. (2018, juillet). Adaptation of S. cerevisiae to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Molecular Biology and Evolution*, *35*(7), 1712–1727. Consulté le 2021-05-15, sur https://doi.org/10.1093/molbev/msy066 doi: 10.1093/molbev/msy066

Libkind, D., Hittinger, C. T., Valério, E., Gonçalves, C., Dover, J., Johnston, M., ... Sampaio, J. P. (2011, août). Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences*, *108*(35), 14539–14544. Consulté le 2023-09-12, sur https://www.pnas.org/doi/10.1073/pnas.1105430108 (Publisher : Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1105430108

LIFE STREAMS project. (2023). Consulté le 2023-09-23, sur https://www.lifestreams.eu/the-project/?lang=en

Liti, G., Barton, D. B. H., & Louis, E. J. (2006, octobre). Sequence Diversity, Reproductive Isolation and Species Concepts in Saccharomyces. *Genetics*, *174*(2), 839–850. Consulté le 2021-06-01, sur https://doi.org/10.1534/genetics.106.062166 doi: 10.1534/genetics.106.062166

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., ... Louis, E. J. (2009, mars). Population genomics of domestic and wild yeasts. *Nature*, *458*(7236), 337–341. Consulté le 2022-02-05, sur https://www.nature.com/articles/nature07743 (Number : 7236 Publisher : Nature Publishing Group) doi: 10.1038/nature07743

Liu, L., Anderson, C., Pearl, D., & Edwards, S. V. (2019). Modern Phylogenomics : Building Phylogenetic Trees Using the Multispecies Coalescent Model. In M. Anisimova (Ed.), *Evolutionary Genomics : Statistical and Computational Methods* (pp. 211–239). New York, NY : Springer. Consulté le 2023-08-10, sur https://doi.org/10.1007/978-1-4939-9074-0_7 doi: 10.1007/978-1-4939-9074-0_7

Maclean, C. J., & Greig, D. (2008, janvier). Prezygotic reproductive isolation between Saccharomyces cerevisiae and Saccharomyces paradoxus. *BMC Evolutionary Biology*, *8*(1), 1. Consulté le 2024-02-18, sur https://doi.org/10.1186/1471-2148-8-1 doi: 10.1186/1471-2148-8-1

Maddison, W. P. (1997, septembre). Gene Trees in Species Trees. *Systematic Biology*, *46*(3), 523–536. Consulté le 2023-06-14, sur https://doi.org/10.1093/sysbio/46.3.523 doi: 10.1093/sysbio/46.3.523

Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., ... Turner, G. F. (2015, décembre). Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science (New York, N.Y.)*, *350*(6267), 1493–1498. Consulté le 2023-09-11, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4700518/ doi: 10.1126/science.aac9927

Malinsky, M., Matschiner, M., & Svardal, H. (2021). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), 584–595. Consulté le 2023-09-11, sur https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13265 (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13265) doi: 10.1111/1755-0998.13265

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., & Steinmetz, L. M. (2008, juillet). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, *454*(7203), 479–485. Consulté le 2022-11-17, sur https://www.nature.com/articles/nature07135 doi: 10.1038/nature07135

Mao, Y., Catacchio, C. R., Hillier, L. W., Porubsky, D., Li, R., Sulovari, A., ... Eichler, E. E. (2021, juin). A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*, *594*(7861), 77–81. Consulté le 2021-10-05, sur http://www.nature.com/articles/s41586-021-03519-x doi: 10.1038/s41586-021-03519-x

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., ... Jiggins, C. D. (2013, novembre). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Research*, *23*(11), 1817–1828. Consulté le 2023-09-25, sur https://genome.cshlp.org/content/23/11/1817 (Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab) doi: 10.1101/gr.159426.113

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015, janvier). Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*, *32*(1), 244–257. Consulté le 2023-09-11, sur https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu269 doi: 10.1093/molbev/msu269

Martin, S. H., & Jiggins, C. D. (2017, décembre). Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, *47*, 69–74. Consulté le 2023-07-17, sur https://www.sciencedirect.com/science/article/pii/S0959437X17300357 doi: 10.1016/j.gde.2017.08.007

Merriam-Webster.com, D. (s. d.). *Saccharomyces*. Consulté le 2023-09-12, sur https://www.merriam-webster.com/dictionary/saccharomyces

Minh, B. Q., Nguyen, M. A. T., & von Haeseler, A. (2013, mai). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195. Consulté le

2023-01-05, sur https://doi.org/10.1093/molbev/mst024 doi: 10.1093/molbev/mst024

Mirarab, S., Nakhleh, L., & Warnow, T. (2021, novembre). Multispecies Coalescent : Theory and Applications in Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *52*(1), 247–268. Consulté le 2023-08-10, sur https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-012121-095340 doi: 10.1146/annurev-ecolsys-012121-095340

Molinet, J., & Cubillos, F. A. (2020). Wild Yeast for the Future : Exploring the Use of Wild Strains for Wine and Beer Fermentation. *Frontiers in Genetics*, *11*. Consulté le 2023-09-23, sur https://www.frontiersin.org/articles/10.3389/fgene.2020.589350

Moran, B. M., Payne, C., Langdon, Q., Powell, D. L., Brandvain, Y., & Schumer, M. (2021, août). The genomic consequences of hybridization. *eLife*, *10*, e69016. Consulté le 2023-07-17, sur https://doi.org/10.7554/eLife.69016 (Publisher : eLife Sciences Publications, Ltd) doi: 10.7554/eLife.69016

Morgan, D. (2007). *The cell cycle : principles of control*. London : Sunderland, MA : Published by New Science Press in association with Oxford University Press ; Distributed inside North America by Sinauer Associates, Publishers. (OCLC : ocm70173205)

Mousseau, T., Koufopanou, V., Lomas, S., Pronina, O., Almeida, P., Sampaio, J., … Burt, A. (2020, juillet). Population Size, Sex and Purifying Selection : Comparative Genomics of Two Sister Taxa of the Wild Yeast Saccharomyces paradoxus. *Genome Biology and Evolution*, *evaa141*. doi: 10.1093/gbe/evaa141

Mozzachiodi, S., Bai, F., Baldrian, P., Bell, G., Boundy-Mills, K., Buzzini, P., … Boynton, P. (2022, janvier). Yeasts from temperate forests. *Yeast*, *39*(1-2), 4–24. Consulté le 2023-09-23, sur https://onlinelibrary.wiley.com/doi/10.1002/yea.3699 doi: 10.1002/yea.3699

Mozzachiodi, S., Tattini, L., Llored, A., Irizar, A., Škofljanc, N., D'Angiolo, M., … Liti, G. (2021, novembre). Aborting meiosis allows recombination in sterile diploid yeast hybrids. *Nature Communications*, *12*, 6564. Consulté le 2023-01-06, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8589840/ doi: 10.1038/s41467-021-26883-8

Naseeb, S., Alsammar, H., Burgis, T., Donaldson, I., Knyazev, N., Knight, C., & Delneri, D. (2018, septembre). Whole Genome Sequencing, *de Novo* Assembly and Phenotypic Profiling for the New Budding Yeast Species *Saccharomyces jurei*. *G3 Genes|Genomes|Genetics*, *8*(9), 2967–2977. Consulté le 2022-11-15, sur https://academic.oup.com/g3journal/article/8/9/2967/6027056 doi: 10.1534/g3.118.200476

Naumov. (1986). Genetic differentiation and ecology of the yeast Saccharomyces paradoxus Batschinskaia. Consulté le 2023-09-23, sur https://scholar.google.com/scholar_lookup?title=Genetic+differentiation+and+ecology+of+the+yeast+Saccharomyces+paradoxus+Batschinskaia&author=G.+I.+Naumov&journal=Dokl.+Biol.+Sci.+%28Engl.+Transl.+Dokl.+Akad.+Nauk+SSSR%29&pages=213-216&publication_year=1986&#d=gs_cit&t=1695489027201&u=%2Fscholar%3Fq%3Dinfo%3At0_PzxxGVXMJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

Naumov, G., Naumova, E., Azbukina, Z. M., Korhola, M., & Gaillardin, C. (1993). Genetic and karyotypic identification of Saccharomyces yeasts from Far East Asia. *Cryptogamie*

Mycologie. Consulté le 2023-09-23, sur https://www.semanticscholar.org/paper/Genetic-and-karyotypic-identification-of-yeasts-Far-Naumov-Naumova/3958c852fef68424bf695bb7d6175d17bf22491a

Naumov, G. I. (1998). Saccharomyces paradoxus and Saccharomyces cerevisiae are associated with exudates of North American oaks. Consulté le 2023-09-23, sur https://cdnsciencepub.com/doi/abs/10.1139/w98-104

Naumov, G. I. (1999). Divergent populations od Saccharomyces paradoxus in Hawaii : species in statu nascendi.

Naumov, G. I., Naumova, E. S., Hagler, A. N., Mendonça-Hagler, L. C., & Louis, E. J. (1995). A new genetically isolated population of the Saccharomyces sensu stricto complex from Brazil. *Antonie Van Leeuwenhoek*, *67*(4), 351–355. doi: 10.1007/BF00872934

Nei, M., Suzuki, Y., & Nozawa, M. (2010, septembre). The Neutral Theory of Molecular Evolution in the Genomic Era. *Annual Review of Genomics and Human Genetics*, *11*(1), 265–289. Consulté le 2021-05-15, sur http://www.annualreviews.org/doi/10.1146/annurev-genom-082908-150129 doi: 10.1146/annurev-genom-082908-150129

Neto, C. (2019). What is a lineage ? *Philosophy of Science*, *86*(5), 1099–1110. (Publisher : Cambridge University Press)

Nguyen, H.-V., Legras, J.-L., Neuvéglise, C., & Gaillardin, C. (2011, octobre). Deciphering the Hybridisation History Leading to the Lager Lineage Based on the Mosaic Genomes of Saccharomyces bayanus Strains NBRC1948 and CBS380T. *PLOS ONE*, *6*(10), e25821. Consulté le 2023-09-12, sur https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0025821 (Publisher : Public Library of Science) doi: 10.1371/journal.pone.0025821

Ohta, T. (1973, novembre). Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, *246*(5428), 96–98. Consulté le 2023-09-26, sur https://www.nature.com/articles/246096a0 (Number : 5428 Publisher : Nature Publishing Group) doi: 10.1038/246096a0

Ono, J., Greig, D., & Boynton, P. J. (2020, septembre). Defining and Disrupting Species Boundaries in *Saccharomyces*. *Annual Review of Microbiology*, *74*(1), 477–495. Consulté le 2024-02-18, sur https://www.annualreviews.org/doi/10.1146/annurev-micro-021320-014036 doi: 10.1146/annurev-micro-021320-014036

Onyema, V. O., Amadi, O. C., Moneke, A. N., & Agu, R. C. (2023, octobre). A Brief Review : Saccharomyces cerevisiae Biodiversity Potential and Promising Cell Factories for Exploitation in Biotechnology and Industry Processes – West African Natural Yeasts Contribution. *Food Chemistry Advances*, *2*, 100162. Consulté le 2023-09-23, sur https://www.sciencedirect.com/science/article/pii/S2772753X22001502 doi: 10.1016/j.focha.2022.100162

O'Donnell, S., Yue, J.-X., Saada, O. A., Agier, N., Caradec, C., Cokelaer, T., … Fischer, G. (2022, octobre). *142 telomere-to-telomere assemblies reveal the genome structural landscape in Saccharomyces cerevisiae* (preprint). Genomics. Consulté le 2022-10-27, sur http://biorxiv.org/lookup/doi/10.1101/2022.10.04.510633 doi: 10.1101/2022.10.04.510633

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H., Tischfield, S., … Keeney, S. (2011, mars). A Hierarchical Combination of Factors Shapes the Genome-wide Topography of Yeast Meiotic Recombination Initiation. *Cell*, *144*(5), 719–731.

Consulté le 2021-03-06, sur https://linkinghub.elsevier.com/retrieve/pii/S0092867411001231 doi: 10.1016/j.cell.2011.02.009

Parapouli, M., Vasileiadi, A., Afendra, A.-S., Hatziloukas, E., 1 Molecular Biology laboratory, Department of Biological applications and Technology, University of Ioannina, Ioannina, Greece, 2 Genetics laboratory, Department of Biological applications and Technology, University of Ioannina, Ioannina, Greece, & # These two authors contributed equally. (2020). Saccharomyces cerevisiae and its industrial applications. *AIMS Microbiology*, *6*(1), 1–32. Consulté le 2023-09-23, sur http://www.aimspress.com/article/10.3934/microbiol.2020001 doi: 10.3934/microbiol.2020001

Paul Moran, & Irv Kornfield. (1993, septembre). Retention of an Ancestral Polymorphism in the Mbuna Species Flock (Teleostei : Cichlidae) of Lake Malawi. *Molecular Biology and Evolution*, *10*(5), 1015. Consulté le 2023-06-15, sur https://academic.oup.com/mbe/article/10/5/1015/1037516/Retention-of-an-Ancestral-Polymorphism-in-the doi: https://doi.org/10.1093/oxfordjournals.molbev.a040063

Payseur, B. A., Krenz, J. G., & Nachman, M. W. (2004, septembre). DIFFERENTIAL PATTERNS OF INTROGRESSION ACROSS THE X CHROMOSOME IN A HYBRID ZONE BETWEEN TWO SPECIES OF HOUSE MICE. *Evolution*, *58*(9), 2064–2078. Consulté le 2023-09-25, sur https://academic.oup.com/evolut/article/58/9/2064/6755720 doi: 10.1111/j.0014-3820.2004.tb00490.x

Pearson, W. R., & Lipman, D. J. (1988, avril). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(8), 2444–2448. Consulté le 2023-01-05, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC280013/

Pease, J. B., & Hahn, M. W. (2015, juillet). Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, *64*(4), 651–662. Consulté le 2023-09-11, sur https://doi.org/10.1093/sysbio/syv023 doi: 10.1093/sysbio/syv023

Perez-Sepulveda, B. M., Heavens, D., Pulford, C. V., Predeus, A. V., Low, R., Webster, H., … Wilson, C. (2021, décembre). An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. *Genome Biology*, *22*(1), 349. Consulté le 2022-11-14, sur https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02536-3 doi: 10.1186/s13059-021-02536-3

Peris, D., Ubbelohde, E. J., Kuang, M. C., Kominek, J., Langdon, Q. K., Adams, M., … Hittinger, C. T. (2023, février). Macroevolutionary diversity of traits and genomes in the model yeast genus Saccharomyces. *Nature Communications*, *14*(1), 690. Consulté le 2023-09-25, sur https://www.nature.com/articles/s41467-023-36139-2 doi: 10.1038/s41467-023-36139-2

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., … Schacherer, J. (2018, avril). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature*, *556*(7701), 339–344. Consulté le 2021-03-06, sur http://www.nature.com/articles/s41586-018-0030-5 doi: 10.1038/s41586-018-0030-5

Pfeifer, B., & Kapan, D. D. (2019, avril). Estimates of introgression as a function of pairwise distances. *BMC Bioinformatics*, *20*(1), 207. Consulté le 2021-07-09, sur https://doi.org/10.1186/s12859-019-2747-z doi: 10.1186/s12859-019-2747-z

Plante, S., & Landry, C. R. (2021). Closely related budding yeast species respond to different ecological signals for spore activation. *Yeast*, *38*(1), 81–89. Consulté le 2024-02-18, sur https://onlinelibrary.wiley.com/doi/abs/10.1002/yea.3538 (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/yea.3538) doi: 10.1002/yea.3538

Pontes, A., Čadež, N., Gonçalves, P., & Sampaio, J. P. (2019, mai). A Quasi-Domesticate Relic Hybrid Population of Saccharomyces cerevisiae × S. paradoxus Adapted to Olive Brine. *Frontiers in Genetics*, *10*, 449. Consulté le 2021-03-06, sur https://www.frontiersin.org/article/10.3389/fgene.2019.00449/full doi: 10.3389/fgene.2019.00449

Porretta, D., & Canestrelli, D. (2023, août). The ecological importance of hybridization. *Trends in Ecology & Evolution*, S0169534723001908. Consulté le 2023-09-25, sur https://linkinghub.elsevier.com/retrieve/pii/S0169534723001908 doi: 10.1016/j.tree.2023.07.003

Presgraves, D. C. (2008, juillet). Sex chromosomes and speciation in Drosophila. *Trends in Genetics*, *24*(7), 336–343. Consulté le 2023-09-25, sur https://www.sciencedirect.com/science/article/pii/S016895250800156X doi: 10.1016/j.tig.2008.04.007

Ramazzotti, M., Stefanini, I., Di Paola, M., De Filippo, C., Rizzetto, L., Berná, L., … Cavalieri, D. (2019, janvier). Population genomics reveals evolution and variation of Saccharomyces cerevisiae in the human and insects gut. *Environmental Microbiology*, *21*(1), 50–71. doi: 10.1111/1462-2920.14422

Ramos-Cormenzana, A., Juárez-Jiménez, B., & Garcia-Pareja, M. (1996, janvier). Antimicrobial activity of olive mill wastewaters (alpechin) and biotransformed olive oil mill wastewater. *International Biodeterioration & Biodegradation*, *38*(3-4), 283–290. Consulté le 2022-02-08, sur https://linkinghub.elsevier.com/retrieve/pii/S0964830596000613 doi: 10.1016/S0964-8305(96)00061-3

Rannala, B., Edwards, S. V., Leaché, A., & Yang, Z. (s. d.). Chapter 3.3 The Multi-species Coalescent Model and Species Tree Inference.

Reilly, P. F., Tjahjadi, A., Miller, S. L., Akey, J. M., & Tucci, S. (2022, septembre). The contribution of Neanderthal introgression to modern human traits. *Current Biology*, *32*(18), R970–R983. Consulté le 2023-09-08, sur https://linkinghub.elsevier.com/retrieve/pii/S0960982222013045 doi: 10.1016/j.cub.2022.08.027

Ridley, M. (2004). *Evolution* (3rd ed éd.). Malden, MA : Blackwell Pub.

Rivas-González, I., Rousselle, M., Li, F., Zhou, L., Dutheil, J. Y., Munch, K., … Zhang, G. (2023, juin). Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science*, *380*(6648), eabn4409. Consulté le 2023-08-07, sur https://www.science.org/doi/10.1126/science.abn4409 doi: 10.1126/science.abn4409

Robinson, H. A., Pinharanda, A., & Bensasson, D. (2016, février). Summer temperature can predict the distribution of wild yeast populations. *Ecology and Evolution*, *6*(4), 1236–1250. Consulté le 2023-09-23, sur https://onlinelibrary.wiley.com/doi/10.1002/ece3.1919 doi: 10.1002/ece3.1919

Rosenberg, N. A., & Nordborg, M. (2002, mai). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*(5), 380–390. Consulté le 2021-08-30, sur http://www.nature.com/articles/nrg795 doi: 10.1038/nrg795

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016, juin). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology*, *25*(11), 2387–2397. Consulté le 2023-09-11, sur https://onlinelibrary.wiley.com/doi/10.1111/mec.13610 doi: 10.1111/mec.13610

Sankararaman, S., Mallick, S., Patterson, N., & Reich, D. (2016, mai). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current biology : CB*, *26*(9), 1241–1247. Consulté le 2023-01-05, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4864120/ doi: 10.1016/j.cub.2016.03.037

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., … Durbin, R. (2012, mars). Insights into hominid evolution from the gorilla genome sequence. *Nature*, *483*(7388), 169–175. Consulté le 2023-08-17, sur https://www.nature.com/articles/nature10842 doi: 10.1038/nature10842

Schacherer, J., Shapiro, J. A., Ruderfer, D. M., & Kruglyak, L. (2009, mars). Comprehensive polymorphism survey elucidates population structure of Saccharomyces cerevisiae. *Nature*, *458*(7236), 342–345. Consulté le 2021-07-13, sur https://www.nature.com/articles/nature07670 (Bandiera_abtest : a Cg_type : Nature Research Journals Number : 7236 Primary_atype : Research Publisher : Nature Publishing Group) doi: 10.1038/nature07670

Scholefield, J., & Greenberg, J. (2007, mai). A common SNP haplotype provides molecular proof of a founder effect of Huntington disease linking two South African populations. *European Journal of Human Genetics*, *15*(5), 590–595. Consulté le 2023-09-27, sur https://www.nature.com/articles/5201796 doi: 10.1038/sj.ejhg.5201796

Schrempf, D., & Szöllösi, G. (2020). The Sources of Phylogenetic Conflicts. *No commercial publisher |Authors open access book,*, 24.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., … Rokas, A. (2018, novembre). Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell*, *175*(6), 1533–1545.e20. doi: 10.1016/j.cell.2018.10.023

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., … Pääbo, S. (2018, septembre). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, *561*(7721), 113–116. Consulté le 2021-10-05, sur http://www.nature.com/articles/s41586-018-0455-x doi: 10.1038/s41586-018-0455-x

Song, Y., Endepols, S., Klemann, N., Richter, D., Matuschka, F.-R., Shih, C.-H., … Kohn, M. (2011, août). Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. *Current Biology*, *21*(15), 1296–1301. Consulté le 2023-09-09, sur https://linkinghub.elsevier.com/retrieve/pii/S0960982211007160 doi: 10.1016/j.cub.2011.06.043

Sousa, V., & Hey, J. (2013, juin). Understanding the origin of species with genome-scale data : modelling gene flow. *Nature Reviews Genetics*, *14*(6), 404–414. Consulté le 2023-11-22, sur https://www.nature.com/articles/nrg3446 (Number : 6 Publisher : Nature Publishing Group) doi: 10.1038/nrg3446

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., & Rokas, A. (2023, juin). Incongruence in the phylogenomics era. *Nature Reviews Genetics*. Consulté le 2023-07-13, sur https://www.nature.com/articles/s41576-023-00620-x doi: 10.1038/s41576-023-00620-x

Stelkens, R. B., Miller, E. L., & Greig, D. (2016, mai). Asynchronous spore germination in isogenic natural isolates of Saccharomyces paradoxus. *FEMS Yeast Research*, *16*(3), fow012. Consulté le 2024-02-18, sur https://doi.org/10.1093/femsyr/fow012 doi: 10.1093/femsyr/fow012

Strope, P. K., Skelly, D. A., Kozmin, S. G., Mahadevan, G., Stone, E. A., Magwene, P. M., … McCusker, J. H. (2015, mai). The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research*, *25*(5), 762–774. Consulté le 2023-09-23, sur http://genome.cshlp.org/lookup/doi/10.1101/gr.185538.114 doi: 10.1101/gr.185538.114

Suarez-Gonzalez, A., Lexer, C., & Cronk, Q. C. B. (2018, mars). Adaptive introgression : a plant perspective. *Biology Letters*, *14*(3), 20170688. Consulté le 2021-03-06, sur https://royalsocietypublishing.org/doi/10.1098/rsbl.2017.0688 doi: 10.1098/rsbl.2017.0688

Tattini, L., Tellini, N., Mozzachiodi, S., D'Angiolo, M., Loeillet, S., Nicolas, A., & Liti, G. (2019, décembre). Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes. *Molecular Biology and Evolution*, *36*(12), 2861–2877. Consulté le 2021-03-07, sur https://doi.org/10.1093/molbev/msz177 doi: 10.1093/molbev/msz177

Taylor, S. A., & Larson, E. L. (2019, janvier). Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, *3*(2), 170–177. Consulté le 2023-09-02, sur https://www.nature.com/articles/s41559-018-0777-y doi: 10.1038/s41559-018-0777-y

Todesco, M., Pascual, M. A., Owens, G. L., Ostevik, K. L., Moyers, B. T., Hübner, S., … Rieseberg, L. H. (2016, février). Hybridization and extinction. *Evolutionary Applications*, *9*(7), 892–908. Consulté le 2023-09-25, sur https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4947151/ doi: 10.1111/eva.12367

Tsai, I. J., Bensasson, D., Burt, A., & Koufopanou, V. (2008, mars). Population genomics of the wild yeast Saccharomyces paradoxus : Quantifying the life cycle. *Proceedings of the National Academy of Sciences*, *105*(12), 4957–4962. Consulté le 2021-07-13, sur https://www.pnas.org/content/105/12/4957 (Publisher : National Academy of Sciences Section : Biological Sciences) doi: 10.1073/pnas.0707314105

Veller, C., Edelman, N. B., Muralidhar, P., & Nowak, M. A. (2023, avril). Recombination and selection against introgressed DNA. *Evolution*, *77*(4), 1131–1144. Consulté le 2023-09-09, sur https://academic.oup.com/evolut/article/77/4/1131/7034890 doi: 10.1093/evolut/qpad021

Vilgalys, T. P., Fogel, A. S., Anderson, J. A., Mututua, R. S., Warutere, J. K., Long, I., … Tung, J. (2022). Selection against admixture and gene regulatory divergence in a long-term primate field study. , 8.

Wakeley, J. (1996, février). The Variance of Pairwise Nucleotide Differences in Two Populations with Migration. *Theoretical Population Biology*, *49*(1), 39–57. Consulté le 2023-09-11, sur https://linkinghub.elsevier.com/retrieve/pii/S0040580996900027 doi: 10.1006/tpbi.1996.0002

Wang, K., Lenstra, J. A., Liu, L., Hu, Q., Ma, T., Qiu, Q., & Liu, J. (2018, octobre). Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. *Com-*

munications Biology, *1*(1), 169. Consulté le 2023-08-11, sur https://www.nature.com/articles/s42003-018-0176-6 doi: 10.1038/s42003-018-0176-6

Wang, Q.-M., Liu, W.-Q., Liti, G., Wang, S.-A., & Bai, F.-Y. (2012). Surprisingly diverged populations of Saccharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology*, *21*(22), 5404–5417. Consulté le 2023-09-21, sur https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-294X.2012.05732.x (_eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-294X.2012.05732.x) doi: 10.1111/j.1365-294X.2012.05732.x

Warren, M. (2018). First ancient-human hybrid.

Wei, W., McCusker, J. H., Hyman, R. W., Jones, T., Ning, Y., Cao, Z., … Steinmetz, L. M. (2007, juillet). Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proceedings of the National Academy of Sciences*, *104*(31), 12825–12830. Consulté le 2023-09-20, sur https://pnas.org/doi/full/10.1073/pnas.0701291104 doi: 10.1073/pnas.0701291104

Wolf, A. B., & Akey, J. M. (2018, mai). Outstanding questions in the study of archaic hominin admixture. *PLOS Genetics*, *14*(5), e1007349. Consulté le 2021-10-06, sur https://dx.plos.org/10.1371/journal.pgen.1007349 doi: 10.1371/journal.pgen.1007349

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., … Liti, G. (2017, juin). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics*, *49*(6), 913–924. Consulté le 2021-05-15, sur https://www.nature.com/articles/ng.3847 (Number : 6 Publisher : Nature Publishing Group) doi: 10.1038/ng.3847

Yue, J.-X., & Liti, G. (2018, juin). Long-read sequencing data analysis for yeasts. *Nature Protocols*, *13*(6), 1213–1231. Consulté le 2023-10-05, sur https://www.nature.com/articles/nprot.2018.025 (Number : 6 Publisher : Nature Publishing Group) doi: 10.1038/nprot.2018.025

apacite biblio

List of Figures

Annexes

A Publications

Accurate Tracking of the Mutational Landscape of Diploid Hybrid Genomes

Lorenzo Tattini[1] **Nicolò Tellini**[1], Simone Mozzachiodi[1], Melania D'Angiolo[1], Sophie Loeillet[2] Alain Nicolas[2] and Gianni Liti[1].

1. CNRS UMR7284, INSERM, IRCAN, Universite Cote d'Azur, Nice, France 2. CNRS UMR3244, Institute Curie, PSL Research University, Paris, France

Abstract

Mutations, recombination, and genome duplications may promote genetic diversity and trigger evolutionary processes. However, quantifying these events in diploid hybrid genomes is challenging. Here, we present an integrated experimental and computational workflow to accurately track the mutational landscape of yeast diploid hybrids (MuLoYDH) in terms of single-nucleotide variants, small insertions/deletions, copy-number variants, aneuploidies, and loss-of-heterozygosity. Pairs of haploid *Saccharomyces* parents were combined to generate ancestor hybrids with phased genomes and varying levels of heterozygosity. These diploids were evolved under different laboratory protocols, in particular mutation accumulation experiments. Variant simulations enabled the efficient integration of competitive and standard mapping of short reads, depending on local levels of heterozygosity. Experimental validations proved the high accuracy and resolution of our computational approach. Finally, applying MuLoYDH to four different diploids revealed striking genetic background effects. Homozygous *Saccharomyces cerevisiae* showed a ∼4-fold higher mutation rate compared with its closely related species *S. paradoxus*. Intraspecies hybrids unveiled that a substantial fraction of the genome (∼250 bp per generation) was shaped by loss-of-heterozygosity, a process strongly inhibited in interspecies hybrids by high levels of sequence divergence between homologous chromosomes. In contrast, interspecies hybrids exhibited higher single-nucleotide mutation rates compared with intraspecies hybrids. MuLoYDH provided an unprecedented quantitative insight into the evolutionary processes that mold diploid yeast genomes and can be generalized to other genetic systems.

Keywords : genome evolution, mutation rate, hybrid genomes, heterozygosity, loss-of-heterozygosity, *Saccharomyces paradoxus*

N.T. Analyzed the data and performed simulations.

B Pipelines

B.1 Intropipeline

An automated computational framework for detecting *Saccharomyces paradoxus* introgressions in Saccharomyces cerevisiae strains from paired-end Illumina sequencing. The pipeline is described in the Supplementary of Chapter 4. Intropipeline is accessible at Intropipeline. Intropipeline is the main pipeline developed during the PhD project. A version 2 is under construction with several improvements in speed and robustness.

B.2 MuLoYDH

MuLoYDH pipeline performs the analysis of paired-end short-read sequencing experiments of clonal samples from yeast diploid hybrids. MuLoYDH is accessible at MuLoYDH. MuLoYDH is the pipeline developed by Lorenzo Tattini published in Tattini et al. 2019. I contribute with proposals for improving the data parsing speed and RAM usage. My journey in R began with MuLoYDH.

B.3 LRSDAY v1.6-Patch

Patch for LRSDAY v1.6 © Jia-Xing Yue for Debian-based OS and SUSE. The debugging was performed with Ubuntu 18.04 and OpenSUSE Leap 15.4. The patch is accessible at LRSDAY v1.6-Patch. LRSDAY is the pipeline developed by Jia-Xing Yue published in Yue & Liti 2018. LRSDAY installs 63 software, I wrote a patch for v. 1.6, simplifying the installation script to make debugging easier in case of installation failures. This allows for individual software installations as needed.

B.4 LICO

LInk COntroller (**LICO**) is a schedulable utility for monitoring the link integrity that allows to promptly identify broken links that crack the installation process of a pipeline. LICO tests the integrity of the links that point to a specific version of a tools integrated into your pipeline. LICO operates silently in the background in a scheduled manner and returns a report to your Telegram account with info concerning the integrity of the links. LICO is accessible at LICO. LICO went out along with LRSDAY patch. It ensures the integrity of links pointing to LRSDAY software but

can be applied to any scheduled activity. It sends a notification to the user via text message on Telegram.

Quantification du tri des lignées incomplètes et de l'introgression au cours de l'histoire évolutive de *Saccharomyces cerevisiae*

Résumé

L'étude de la distribution de la variation génétique au sein des populations et entre elles permet de mieux comprendre l'histoire évolutive d'une espèce. De plus, l'accès à de grands ensembles de données génomiques permet d'étudier les processus évolutifs qui façonnent la variation ségrégative de l'espèce. Dans ce travail, nous retraçons l'histoire évolutive de l'espèce *Saccharomyces cerevisiae* par la détection et la classification de polymorphismes partagés avec son espèce sœur *Saccharomyces paradoxus*. Nous identifions des polymorphismes partagés acquis par hybridation suivie d'introgression et des polymorphismes qui persistent à travers le processus de spéciation et de diversification résultant en des cas de triage incomplet des lignées (ILS).Nous définissons un ensemble de données de polymorphismes nucléotidiques simples diagnostiques bialléliques entre *Saccharomyces cerevisiae* et *Saccharomyces paradoxus* que nous utilisons comme marqueur diagnostique pour décrire la composition génomique de 1,673 *S. cerevisiae*, pour lesquels un séquençage à lecture courte du génome entier était publiquement disponible. Nous développons une méthode basée sur les marqueurs pour la détection et la classification des marqueurs de diagnostic organisés soit en 1) blocs de marqueurs *S. paradoxus* consécutifs, soit en 2) marqueurs *S. paradoxus* isolés à l'échelle du génome. Pour les blocs, nous décrivons les limites, et la distribution dans la collection *S. cerevisiae* et nous retraçons l'origine de *S. paradoxus* par comparaison de séquences avec des assemblages de génomes entiers télomère à télomère des principales populations de *S. paradoxus*. Pour un événement récurrent, nous avons effectué un test pour évaluer l'effet sur la condition physique de porter un haplotype *S. paradoxus* à un locus unique englobant une paire de gènes impliqués dans la dégradation de composés toxiques pour la levure. Nous avons démontré que l'haplotype *S. paradoxus* confère un avantage par rapport à l'haplotype *S. cerevisiae* dans des conditions environnementales caractéristiques de la niche habitée par la population *S. cerevisiae*. Pour les marqueurs isolés, nous appliquons une méthode classique de détection des signatures d'un tri lignager incomplet, qui peut expliquer l'excès de marqueurs S. paradoxus distribués à l'échelle du génome dans certaines populations de *S. cerevisiae*. Nous montrons des preuves convaincantes de la rétention d'allèles ancestraux dans une seule population sauvage de *S. cerevisiae* qui se trouve à la racine de l'espèce. Nous émettons l'hypothèse que la persistance d'une telle variation ancestrale est due à la possibilité réduite de croisement avec d'autres populations de *S. cerevisiae* dans la nature, en raison du nombre réduit de générations et des goulets d'étranglement moins spectaculaires qu'ont connu les autres lignées au cours de la dispersion et de la domestication.Dans l'ensemble, nous avons retracé l'histoire de la divergence et des contacts secondaires entre les populations de *S. cerevisiae* et de *S. paradoxus* et dévoilé un cas convaincant d'introgression interespèces avec un résultat fonctionnel.

Mots-clés : Saccharomyces, triage incomplet des lignées, hybridation, introgression, introgression adaptative, introgression ancestrale.