The rate of adaptive molecular evolution in wild and domesticated *Saccharomyces cerevisiae* populations

Maximilian W. D. Raas[1,#,*], Julien Y. Dutheil[1,2]

[1] Research Group Molecular Systems Evolution, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306 Plön, Germany

[2] Unité Mixte de Recherche 5554 Institut des Sciences de l'Evolution, CNRS, IRD, EPHE, Université de Montpellier, Montpellier, France.

[#] Present address: Theoretical Biology and Bioinformatics, Biology, Science Faculty, Utrecht University, 3584 CH Utrecht, The Netherlands

[*] Corresponding author

Email: m.w.d.raas@uu.nl

Abstract

Through its fermentative capacities, *Saccharomyces cerevisiae* was central in the development of civilization during the Neolithic period, and the yeast remains of importance in industry and biotechnology giving rise to *bona fide* domesticated populations. Here, we conduct a population genomic study of domesticated and wild populations of *S. cerevisiae*. Using coalescent analyses, we report that the effective population size of yeast populations decreased since the divergence with *S. paradoxus*. We fitted models of distribution of fitness effects to infer the rate of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions in protein-coding genes. We report an overall limited contribution of positive selection to *S. cerevisiae* protein evolution, albeit with higher rates of adaptive evolution in wild compared to domesticated populations. Our analyses revealed the signature of background selection and possibly Hill-Robertson interference, as recombination was found to be negatively correlated with ω_{na} and positively correlated with ω_a. However, the effect of recombination on ω_a was found to be labile, as it is only apparent after removing the impact of codon usage bias on the synonymous site frequency spectrum and disappears if we control for the correlation with ω_{na}, suggesting it could be an artefact of the decreasing population size. Furthermore, the rate of adaptive non-synonymous substitutions is significantly correlated with the residue solvent exposure, a relation that cannot be explained by the population's demography. Together, our results provide a detailed characterization of adaptive mutations in protein-coding genes across *S. cerevisiae* populations.

Keywords: Adaptive evolution, Domestication, *Saccharomyces cerevisiae*, Demographic inference, Population genetics

Introduction

The domestication of crops and livestock is widely regarded as a defining determinant in the rise of permanent human settlement and the development of contemporary civilisation during the Neolithic period (Diamond, 2002; Fuller & Stevens, 2019; Zeder, 2008). The budding yeast *Saccharomyces cerevisiae* is believed to be among the earliest domesticated species, as archaeological evidence has shown that humans were crafting fermented beverages as early as 9,000 years ago in ancient China (McGovern et al., 2004). From its initial use in liquid fermentation reactions, domesticated usage of *S. cerevisiae* has further diversified into bread leavening and myriad biotechnological applications (Lahue, Madden, Dunn, & Smukowski Heil, 2020; Parapouli, Vasileiadis, Afendra, & Hatziloukas, 2020), and it has been adopted as a leading model organism in molecular biology and (evolutionary) genetics (Botstein & Fink, 2011; Duina, Miller, & Keeney, 2014; Marsit et al., 2017), being the first eukaryote to have its full genome sequenced (Goffeau et al., 1996). The close alliance between humankind and yeast has led to the global dispersal of *S. cerevisiae* through trade and migration, leaving a marked impact on the life history of the yeast (Legras, Merdinoglu, Cornuet, & Karst, 2007; Ludlow et al., 2016; Money, 2018; Sicard & Legras, 2011). Consequently, industrially-used strains of *S. cerevisiae* show distinct patterns of demographic history and genome evolution marked by typical domestication signatures, including historic bottlenecks and higher rates of unbalanced rearrangements, polyploidies and aneuploidies in domesticated- compared to wild strains (Duan et al., 2018; Gallone et al., 2016; Liti et al., 2009; Peter et al., 2018; Yue et al., 2017). In contrast to these highly domesticated strains, the relatively recent discovery of truly wild populations of *S. cerevisiae* in Asian primeval forests showed that there are present-day populations that have continued to evolve under purely natural conditions (Liti, 2015; Wang, Liu, Liti, Wang, & Bai, 2012), countering the

long-standing view that *S. cerevisiae* is an exclusively domesticated species. High-resolution phylogenetic analyses have confirmed the historic separation between wild and domesticated populations (Duan et al., 2018; Gallone et al., 2016; Legras et al., 2018; Ludlow et al., 2016; Peter et al., 2018). Therefore, while *S. cerevisiae* isolates routinely used in research and industry represent *bona fide* domesticated populations, there also exist truly wild populations of *S. cerevisiae* in their native Far East Asia. Owing to the unique availability of large amounts of genomic data from both these domesticated and wild populations, *S. cerevisiae* provides an outstanding opportunity to investigate the impact of domestication on genome evolution.

The study of domestication has a long-standing tradition in evolutionary biology, and has been influential in establishing the theory of evolution by natural selection (Darwin, 1868). It is known to impact population demographic dynamics substantially and, at the molecular level, genome diversity and evolution. One open question is how domestication impacts the relative contributions of positive selection and genetic drift to genome evolution (Moyers, Morrell, & McKay, 2017). With the dawn of the (post-)genomic era, novel population genomics approaches towards delineating the relative contributions of these drivers of sequence evolution have become available. Importantly, these methods enable the quantification of adaptive sequence evolution (i.e., resulting from positive selection) from non-adaptive sequence evolution (i.e., attributable to genetic drift and purifying selection). This is made possible by the comparison of intra-specific (polymorphism) and inter-specific (divergence) data, an approached pioneered by (Hudson, Kreitman, & Aguadé, 1987) and extended by (McDonald & Kreitman, 1991). The latter approach makes the distinction between two types of mutations in protein-coding sequences: synonymous mutations, which do not change the encoded amino-acid owing to the redundancy of the genetic code and are thus assumed to be neutral, and non-synonymous mutations, which change the amino-acid sequence,

possibly altering the function of the resulting protein and making these subject to selection. The comparison of the amount of synonymous and non-synonymous polymorphism on the one hand, with the quantity of synonymous and non-synonymous divergence on the other hand allows estimating the proportion of non-synonymous substitutions fixed by positive selection, a metric termed α by (Smith & Eyre-Walker, 2002). More recent developments include the modelling of the distribution of fitness effects (DFE) underlying polymorphism and divergence patterns obtained from population genomics data (Adam Eyre-Walker & Keightley, 2007; A. Eyre-Walker & Keightley, 2009; A. Eyre-Walker, Woolfit, & Phelps, 2006). This approach was later further extended to explicitly account for slightly beneficial mutations (Galtier, 2016; Tataru, Mollion, Glémin, & Bataillon, 2017). These statistical approaches permit the inference of parameters describing historic adaptive evolution: the rate of adaptive non-synonymous substitutions (ω_a), the rate of non-adaptive non-synonymous substitutions (ω_{na}) and the proportion of adaptive non-synonymous substitutions (α).

There are various ways in which domestication could have impacted the contributions of adaptive versus non-adaptive evolution in *S. cerevisiae*. On the one hand, studies of animal and plant domestication have shown that the strong directional selection involved in domestication leaves notable signatures of adaptive evolution on domestication-associated genes (Ahmad et al., 2020; Ross-Ibarra, Morrell, & Gaut, 2007). Therefore, it is possible that similar mechanisms underly niche-adaptation in domesticated yeast populations (Duan et al., 2018; Legras et al., 2018), in turn yielding elevated values for ω_a in genes involved in phenotypic adaptations towards domesticated niches. On the other hand, differences in adaptive evolution across populations may be the result of their distinct effective sizes. The effective population size (N_e) and the rate of adaptive evolution are tightly linked as, in small populations, (i) genetic drift is stronger, preventing advantageous

mutations with small effect (lower than 1/Ne) to get fixed, and (ii) the available supply of standing genetic variation will generally be lower (Charlesworth, 2009; Gossmann, Keightley, & Eyre-Walker, 2012; Lanfear, Kokko, & Eyre-Walker, 2014). Importantly, domestication is associated with strong demographic dynamics and domestication-associated bottlenecks have been identified in domesticated *S. cerevisiae* populations (Gallone et al., 2016). Therefore, it is expected that domesticated populations of *S. cerevisiae* have lower overall rates of adaptive evolution compared to wild populations as a result of the domestication bottleneck, despite the strong directional selection acting on specific cellular processes.

Another major driver of the rate of evolution (both adaptive and non-adaptive) is recombination, which impacts the efficacy of both adaptive and purifying selection as it can combine interfering adaptive mutations and unlink adaptive mutations from deleterious ones (Marais & Charlesworth, 2003). The relation between sexual reproduction and the efficacy of selection is well-documented in yeasts (Connallon & Knowles, 2007; Goddard, Godfray, & Burt, 2005). It has recently been reported that domesticated populations of *S. cerevisiae* largely abolish sexual reproduction (De Chiara et al., 2022) and do not undergo meiotic recombination. Therefore, the suppression of recombination in domesticated populations is also expected to induce a lower rate of adaptive evolution in domesticated populations.

Earlier landmark work by (Liti et al., 2009) identified no evidence for adaptive evolution in a set of ~1,100 *S. cerevisiae* genes from a diverse set of *S. cerevisiae* strains, using the method developed by (McDonald & Kreitman, 1991). However, this method has important limitations when it comes to accounting for slightly deleterious mutations, which can therefore substantially bias inferences of adaptive evolution (A. Eyre-Walker & Keightley, 2009). Moreover, protein structure plays an important role in shaping the rate of substitutions. Specifically, residues with a

high relative solvent accessibility (RSA) show higher rates of both adaptive and non-adaptive substitutions (Moutinho, Trancoso, & Dutheil, 2019). Accounting for RSA will yield higher-resolution insights into the factors shaping substitution rates in *S. cerevisiae* and allow for more sensitive detection of adaptive sites.

Here, we use a population genomics approach to assess the impact of several factors on the rate of adaptive and non-adaptive non-synonymous substitutions in wild and domesticated populations of *S. cerevisiae*. By comparing 17 populations of different origins, we show that domestication resulted in both a decrease of the rate of adaptive – and an increase of the rate of non-adaptive – non-synonymous substitutions, and that the effective population size is a determinant of both rates. We find that only the rate of non-adaptive non-synonymous substitutions is significantly lower in the domestication-associated functional categories of genes. Adaptive non-synonymous substitutions are more frequently found outside the core proteome, that is, in isolate-specific genes. Signatures of linked selection are pervasive, with recombination rate being negatively correlated with ω_{na} and positively correlated with ω_a. This effect, however, is weak and masked by codon usage bias when not accounted for and disappear if we correct for the correlation of recombination with the strength of purifying selection. Finally, the residues' solvent exposure strongly impacts both adaptive and non-adaptive rates, with the majority of adaptive substitutions occurring at the protein surface.

Materials & Methods

Genomic dataset and population definition

The *S. cerevisiae* genome data used in this study was generated by (Peter et al., 2018) as part of the "1,002 Yeast Genomes Project". Unless specified otherwise, the *de novo* genome assembly of each isolate was used in the following analyses. Relevant populations of isolates were defined based on the clades as determined in the original publication, followed by manual filtering of isolates based on ecological- and geographical origins to assure population homogeneity. We included populations with a minimum sample size of five isolates for further analyses. The Italian Wine population had a large sample size and it was randomly split into two subsets of isolates for computational tractability. Given that results were highly similar between the two subsets, we only report the ones for one subset, which we refer to as the "Italian Wine" population in the following. Combined, this yields a total of 17 *S. cerevisiae* populations from both domesticated and wild origins (*Supplementary Table 1*).

Genome alignment and gene extraction

For each population, we generated full genome alignments and extracted the protein-coding DNA sequence (CDS). The genomes in each population and the reference genome R64 (Downloaded from ftp://ftp.ensemblgenomes.org/pub/fungi/release-49/fasta/saccharomyces_cerevisiae/dna/; last accessed December 2020) were aligned using the MultiZ v11.2 alignment program from the Threaded-Blockset Aligner tool suite (Blanchette et al., 2004). The alignments were then projected onto the reference sequence, and the resulting alignment blocks were realigned and filtered using MafFilter v1.3.1 (Dutheil, Gaillard, & Stukenbrock, 2014). Alignment blocks were split into windows with a maximum size of 10 kb, which were subsequently realigned using MAFFT v7.38 (Katoh & Standley, 2013). We then filtered the alignment to only include i) blocks with sequences of all selected isolates in the population and ii) blocks with a minimum length of 10 bp. Ten bp

windows, slid by 1 bp, containing ≥ 30% gap and unresolved characters were discarded. The remaining alignment blocks were merged if not further apart than 100 bp and between-block positions were filled with unresolved characters "N", resulting in a masking of ambiguously aligned regions. A callability mask was also generated, recording the coordinates of the filtered positions. Diversity statistics (average pairwise heterozygosity) were computed in 10 kb non-overlapping windows along the alignment, for each population.

CDSs were extracted according to the General Feature Format (GFF) file of the reference sequence (Downloaded from ftp://ftp.ensemblgenomes.org/pub/fungi/release-49/gff3/saccharomyces_cerevisiae; last accessed December 2020) and the obtained CDSs were concatenated to reconstruct genes from the alignment files. Reconstructed gene-coding sequences were checked for premature stop codons to ensure accurate reconstruction. No genes were removed at this step. In these and subsequent procedures, genomic data was handled with scripts using the Biopython package v1.78 (Cock et al., 2009), available at https://gitlab.gwdg.de/molsysevol/saccharomycespopgen. The total number of nucleotide sites and final number of genes for each population are summarised in *Supplementary Table 2*.

Phylogenetic tree

We used the complete (unaligned) genome sequences of all isolates in the selected populations to compute pairwise distances using *andi* v0.12 (Haubold, Klötzl, & Pfaffelhuber, 2015). A phylogenetic tree was inferred from the resulting distance matrix using FastME v2.0 (Lefort, Desper, & Gascuel, 2015) and plotted using FigTree v1.4.4 (Rambaut, 2018).

Demographic history inference

We used the Multiple Sequentially Markovian Coalescent 2 v2.1.1 (MSMC2) (Malaspinas et al., 2016) to reconstruct the demographic histories of the selected populations. As SMC-based methods have been benchmarked on Primates dataset (Schiffels & Wang, 2020), we conducted a

simulation analysis to assess their accuracy on yeast datasets (see *Supplementary Methods*). We assessed the power of the MSMC2 method by simulating data under a neutral model of sequence evolution, which fulfilled the hypotheses of the inference model (homogeneous recombination and mutation rate, no selection). We simulated 10 datasets of 5 diploid individuals (10 haploid genomes), with the same chromosome structure as the reference *S.cerevisiae* genome (R64). We found that the default parametrisation of the time epochs led to biased estimates in the most recent and ancient times (*Supplementary Figure 1A*). We then fitted a model with fewer parameters. We considered 30 epochs, with the five most recent and the five most ancient epochs sharing a common coalescence rate, leading to 22 parameters, the 20 intermediate epochs being allowed to have a distinct coalescence rate. With this discretisation scheme, no bias was observed (*Supplementary Figure 1B*). However, we observed an increased estimation variance in the most recent and ancient times. Adding a distribution of missing data identical to the one of one of our population alignments (Far East Asia) did not significantly affect the inference (*Supplementary Figure 1C*, repeated in *Figure 3A*). For all remaining MSMC2 inferences, we used the 22 parameters time pattern.

We separated the dataset into isolates for which the *de novo* assembled genome could be considered as a haplotype (haploid and diploid homozygous isolates) and isolates for which the assembled genome represented a consensus genome (diploid heterozygous isolates). In the first case, MSMC-compatible input files were generated from the filtered genome alignments using MafFilter v1.3.1 (Dutheil et al., 2014). In the second case, the corresponding diploid genotypes were extracted from the 1,011 yeast genome variant calls using the bcftools (Danecek et al., 2021; Peter et al., 2018). Positions with missing genotypes were masked. As a callable mask, we used the positions of the *de novo* genome alignments, after filtering. Because the variants are unphased,

only haploid pairs from the same isolate were compared in MSMC2 (n pairs, where n is the number of isolates in a population), while all haplotype pairs were analysed in the case of haploid or homozygous diploid isolates (n*(n-1)/2 pairs). Following our benchmark procedure, MSMC2 was run using a 1*5+20*1+1*5 discretisation scheme. Other options were kept as program defaults. We ran MSMC2 on all isolates in each population, but also separately on several pairs of haplotypes to estimate the robustness of the inferred demography for each population, as well as the homogeneity of the samples. Results are reported in coalescent units (x-axis) and inverse instantaneous coalescence rate (IICR, y-axis).

We estimated the effective population size of the *S. cerevisiae* / *S. paradoxus* ancestor using the IM-CoalHMM model (Mailund et al., 2012). *S. paradoxus* is a close relative of *S. cerevisiae* and was for this reason selected as the outgroup for this analysis. The model considers a speciation process with an initial differentiation phase with symmetric migration between the two populations, followed by a phase where the two populations are genetically isolated. The estimated parameters are the start and end of the migration phase, the average ancestral population size, and the average recombination rate. A seven species genome alignment of yeasts was downloaded from the UCSC genome browser. The alignment was projected on the *S. cerevisiae* (sacCer3) genome using the 'maf_project' program v12 from the TBA package (Blanchette et al., 2004) and the pairwise alignment of *S. cerevisiae* with *S. paradoxus* (sacPar) was extracted and merged, replacing non-aligned positions by missing data. The 16 chromosomes were further concatenated before being input to the IMCoalHMM program (v0.7.0, https://github.com/mailund/IMCoalHMM).

Quantifying adaptive evolution

For reconstructed gene alignments, an outgroup sequence was identified using the closest match of a 'blastn' v2.9.0+ search (Altschul, Gish, Miller, Myers, & Lipman, 1990) on the *S. paradoxus*

genome (downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/Saccharomyces_paradoxus/latest_assembly_versions/GCF_002079055.1_ASM207905v1/). Specifically, the reference sequence of each *S. cerevisiae* gene alignment (R64) was queried against the reference CDS file of *S. paradoxus* (Downloaded from https://yjx1217.github.io/Yeast_PacBio_2016/data/) (Yue et al., 2017). Protein-coding sequences, including the outgroup sequence, were re-aligned using the codon alignment program MACSE v2 (Ranwez, Douzery, Cambon, Chantret, & Delsuc, 2018). We calculated statistics on synonymous and nonsynonymous divergence between the species, transversion per transition ratio, and inferences on ancestral alleles from these alignments using the BppPopStats v2.4.0 program from the Bio++ Program Suite (Guéguen et al., 2013) and used them to produce unfolded site frequency spectra (SFSs). Models of distribution of fitness effects (DFE) were then fitted with Grapes v1.1 (Galtier, 2016), to estimate the rates of adaptive and non-adaptive non-synonymous substitutions (ω_a and ω_{na}), and the proportion of adaptive non-synonymous substitutions (α). Grapes was run with the '-no_div_param' option, as the underlying model here was shown to be more robust in light of historic changes in population size (M. Rousselle, Mollion, Nabholz, Bataillon, & Galtier, 2018). Additionally, the option '-nb_rand_start 5' was specified to test distinct optimization starting conditions and prevent model settling on suboptimal local maximum likelihoods. Three DFE models were fitted onto the data: GammaExponential (GammaExpo), DisplacedGamma (DisplGamma) and ScaledBeta (Galtier, 2016). To determine the quality of model fit, Akaike's Information Criterion (AIC) was calculated for the complete SFSs per population (Akaike, 1973). Here, the GammaExpo and ScaledBeta models performed comparably well, but the ScaledBeta model failed to estimate ω_a in more cases than the GammaExpo model. Therefore, the GammaExpo model was determined as the best

performing model, which is in accordance with previous results obtained from data in animals (Galtier, 2016), plants (Moutinho et al., 2019) and fungi (Grandaubert, Dutheil, & Stukenbrock, 2019). Finally, sites were bootstrapped to create 150 SFS replicates for each dataset enabling a more comprehensive analysis of α, ω_a and ω_{na}. For some replicates, optimisation failed and the corresponding replicates were discarded. The minimum number of replicates for which estimates could be obtained was 90 for the North American Oak population in the Metal Stress gene category. The average number of successful bootstraps across all runs is 149.2. The mean values of the estimates over all bootstrap replicates were used as the final estimates of these parameters in each analysis.

Gene set definition

Gene sets were defined based on "Genes and gene products" annotations obtained from AmiGO 2 (http://amigo.geneontology.org/amigo) using the search terms "Meiosis" and "Sporulation" (corresponding to the gene set "Meiosis"), "Osmotic Stress", "Oxidative Stress", "Glucose OR Xylose OR Maltose OR Fructose OR Sucrose" (corresponding to the gene set "Domestic Carbon"), and "Heat Stress" and "Temperature" (corresponding to the gene set "Heat Stress"). The "Metal Stress" gene set was compiled based on data from genetic screens performed by (Thorsen et al., 2009) for arsenic exposure and by (Jo et al., 2007) for copper exposure. Genes that were found in none or multiple of the aforementioned categories were combined into the "Other Genes" and "Multiple sets" categories. All gene sets are restricted to only include the genes that were recovered from all populations, which totalled 3,215 genes (*Supplementary Figure 2*). Gene overlap analysis was performed with the 'UpSetR' package v.1.4.0 (Conway, Lex, & Gehlenborg, 2017).

Recombination landscape

Recombination rates across the genome for *S. cerevisiae* were taken from (Liu, Maclean, & Zhang, 2018), who quantified crossover (CO) rates in a window-based approach with a window size of 20 kb, and these rates were mapped onto the genes included in our analyses based on the genomic location of the midpoint of these genes in the respective reference genomes. The genes were then binned into 15 approximately equal sized categories, except for the populations Ecuadorian, Malaysian and North American Oak, for which the overall genetic variation was lower and so they were binned into 5 categories with sufficient polymorphic sites.

Prediction of relative solvent accessibility

The relative solvent accessibility (RSA) per site was predicted using NetSurfP-2.0 (Klausen et al., 2019). NetSurfP-2.0 was run on R64 translated CDSs to obtain the RSA for each site. These per-site RSA values as computed for R64 were mapped onto the corresponding sites of the population-specific alignments. The sites were binned into 15 equal-sized bins based on RSA except for the populations Ecuadorian and Malaysian, which were binned into 5 categories, and the North American Oak population, which was binned into 3 categories.

Linear modelling

All statistical analyses were conducted in the R software, version 4.1.0 (R Development Core Team, 2021). We fitted generalised least squares (GLS) linear models to assess the impact of several factors on the rate of adaptive and non-adaptive non-synonymous substitutions while accounting for the phylogenetic relationships between populations. Models were fitted with the maximum likelihood method to compute values of Akaike's information criterion, using the 'gls' method of the 'nlme' package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2021) with phylogenetic correlation structures taken from the 'ape' package (Paradis & Schliep, 2019). We fitted separate models to each rate (response variable) including estimates for all populations. All

models included the life history of the population (Wild, Domesticated-Asian, and Domesticated-Non Asian), the culture method (Environment, Repitching, and Starter-based), and the mean synonymous diversity (π_S) as explanatory variables. We specified a correlation model using the previously estimated population phylogenetic tree and fitting branch lengths using the method of Grafen (Grafen, 1989). Residues were normalised using the Box-Cox transform, modified to allow for non-strictly positive values, as implemented in the 'car' package (Fox & Weisberg, 2018).

To assess the effect of gene sets, recombination rates and RSA, we fitted extended models including each of these variables in addition to life history, culture method, and π_S, together with their interaction. For instance, the model assessing the impact of RSA on ω_a is described by the formula:

$$\omega_a \sim RSA + \text{Life history} + \text{Method} + \pi_S + \text{RSA: Life history} + \text{RSA: Method} + \text{RSA:} \pi_S .$$

After normalisation of the residues, a model selection was conducted using Akaike's information criterion (AIC), as implemented in the 'stepAIC' function of the 'MASS' package (Venables, 2002). Normality and independence of the residues were assessed using the Shapiro and Ljung-Box test, as implemented in the 'stats' package.

In some cases, the non-normality of the residues was still significant after using the Box-Cox transform, owing to the distribution of $\omega_a$being zero-inflated. Therefore, for comparison, we fitted another set of models where all points with $\omega_a < 1e-3$ were discarded. While these models were fitted on reduced datasets, their residues were generally in better agreement with the Gauss-Markov assumption and had possibly more accurate P-values. Results were visualised with the 'ggplot2' (Wickham, 2016), 'cowplot' (Claus O. Wilke, 2020), and 'ggpubr' (Alboukadel Kassambara, 2020) packages.

Codon usage analyses

We computed codon frequencies for the set of genes used to estimate rates of adaptive and non-adaptive substitutions, for all populations. All 61 non-stop codons frequencies were computed for each analysed site in each population. We did not observe any significant difference in codon usage between populations and summed all counts of all populations. Frequencies were then further summed according to site groups, one of gene functional category, recombination rate category, or RSA category. The resulting contingency tables were submitted to a within-group correspondence analysis as described in (D. Charif, Thioulouse, Lobry, & Perrière, 2004). Relative synonymous codon usage (RCSU) values were also computed for each codon. The RSCU value of a codon is defined as the ratio of the observed codon frequency to the expected frequency if all synonymous codons for the encoded amino-acid were equally used. It is a measure of codon usage bias, which takes values above 1 is the codon is preferentially used, and below 1 if the codon is less preferentially used. The mean codon usage bias for a category of sites was thus computed as the average of the absolute values of RSCU – 1. All codon analyses were conducted using the 'seqinr' (Delphine Charif & Lobry, 2007) and 'ade4' (Thioulouse, 2018) packages for R.

Results

We defined 17 population encompassing 329 *S. cerevisiae* isolates from the *1,002 Yeast Genomes Project* genome resource (Peter et al., 2018), based on geography and ecology (*Supplementary Table 1*). A phylogenetic analysis of the selected isolates confirmed that the defined populations represent distinct lineages (*Figure 1*). In agreement with previous reports (Peter et al., 2018), we observed that the domesticated populations formed two distinct groups, based on their Asian or Non-Asian origin (see *Supplementary Results* for a detailed description of the obtained

phylogeny). For each population, we estimated the rate of adaptive non-synonymous substitutions in protein-coding genes. We gained insights into the drivers of adaptive evolution between populations and along the yeast genome by estimating rates of adaptive evolution in gene sets with distinct functional categories or local recombination rates, as well as within sites with different solvent exposure. We then compared these estimates between populations, accounting for their evolutionary relationship and demographic history.

A low rate of adaptive non-synonymous substitutions in domesticated *S. cerevisiae*.

To study the rate of adaptive evolution in both domesticated and wild *S. cerevisiae* populations, we calculated unfolded site frequency spectra (SFSs) capturing within-population segregating frequencies of synonymous and non-synonymous polymorphic sites. We then fitted models of distributions of fitness effects (DFEs) onto the SFSs using the *Grapes* method (Galtier, 2016), allowing for the computation of the proportion of adaptive non-synonymous substitutions (α), the rate of adaptive non-synonymous substitutions (ω_a), and the rate of non-adaptive non-synonymous substitutions (ω_{na}). To ensure that the same genes are compared between groups, we restricted our analyses to genes that we could successfully recover from genome alignments in all populations, totaling 3,215 core genes (see *Supplementary Figure 2* and Materials & Methods).

We classified the 17 populations according to two criteria, which we refer to as “life history” and “method”. The first distinguishes wild populations from domesticated populations and within those, the Asian domesticated from non-Asian domesticated populations, as the latter two are believed to be the result of independent domestication events (Duan et al., 2018; Peter et al., 2018). The second criterion refers to distinct application-specific practices for introducing yeasts to fermentation reactions in domesticated settings. These can be broadly classified into “repitching” (also known as “backslopping”) and “starter-based” approaches. In repitching, yeasts are seeded

into multiple consecutive fermentation cycles, either by purposely isolating yeasts from an ongoing or terminated fermentation reaction and re-seeding these into fresh wort (traditional repitching, characteristic for beer brewing) (Large et al., 2020), or indirectly via contaminated surfaces that carry residual yeast from previous fermentations such as flasks, baskets or tools (De León-Rodríguez, González-Hernández, Barba de la Rosa, Escalante-Minakata, & López, 2006; Djeni et al., 2020; Schwan & Wheals, 2004). In contrast, yeast cells used in starter-based methods are discarded after a single fermentation reaction. Previous studies have shown that repitching practices significantly affects rates of chromosomal rearrangement rates in *S. cerevisiae*, revealing the substantial impact of application-specific practices on yeast genome evolution (Large et al., 2020). Consequently, we hypothesise that these different practices might impact the rate of adaptation, with repitching methods being likely to better facilitate adaptation through the extended multi-generational exposure to environmental pressures, compared to the continuous evolutionary "reset" of populations resulting from the use of fixed starter cultures.

These analyses revealed that the proportion of adaptive substitution, α, is generally low in *S. cerevisiae* (*Table 1*), with some differences between wild and domesticated populations. Focusing on core genes present in all populations, α values range from 0 (several populations) to 8.79% (Alpechin population) in domesticated populations and from 1.65% (Ecuadorian population) to 70.57% (Malaysian population) in wild populations. In addition to the Malaysian population, the Far East Asian wild population also shows a high α value of 40.59%. These two extreme values are the result of a comparatively higher ω_a and lower ω_{na} (*Table 1*), and are in the range of what was measured in several animal species (Galtier, 2016). Accounting for the phylogenetic relationships of the populations, we find that these differences are marginally significant: domesticated populations have a lower α compared to wild populations (Generalised Least Square

linear model, P value = 0.057 for non-Asian populations, 0.037 for Asian populations, see *Table 2*). We also find a marginal negative effect of the starter-based method, these populations having a lower α (P value = 0.075) compared to environmental populations. Repitching populations have no significant difference with the environmental ones (P value = 0.249). We further note a positive effect of π_S on α (P value = 0.023). Including population-specific genes leads to more homogeneous results among populations, as all domesticated populations still have a very low α, while wild populations display higher values. Both α and ω_a are significantly lower in domesticated vs wild populations, while ω_{na} is significantly higher, as expected if domesticated populations have a lower effective population size. In this dataset, however, the remaining effect of population size (π_S) is no longer significant after accounting for the life-history variables (Supplementary table 3).

As opposed to previous work based on the MK-test (Liti et al., 2009), and estimates from the closely-related wild yeast *S. paradoxus* (Gossmann et al., 2012), we found a positive rate of adaptive non-synonymous substitutions in several populations of *S. cerevisiae*. We hypothesise that the generally low rate of adaptive non-synonymous substitutions results from two possible, non-mutually exclusive causes: population size changes since the time of divergence, and intra-genomic variation of the substitution rates. We will examine these two aspects in the following sections.

S. cerevisiae population size decreased since the divergence with *S. paradoxus*.

The historical change of population is an important factor shaping adaptive evolution (Charlesworth, 2009; Gossmann et al., 2012; Lanfear et al., 2014; M. Rousselle et al., 2018). We employed two approaches based on the sequentially Markov coalescent framework to infer the demographic histories of the selected populations. We first used the multiple sequentially Markov

coalescent version 2, MSMC2 (Malaspinas et al., 2016) to get insights into populations sizes at the polymorphism scale (that is, from present time to the time of the most recent common ancestor of the sample). We then employed the IMCoalHMM approach to fit a model of isolation with migration between *S. cerevisiae* and *S. paradoxus* (Mailund et al., 2012), allowing us to estimate the population size at the time of the divergence between the two species. The estimated divergence time was found close to the most ancient time interval of the MSMC analysis, in agreement with the recent speciation between the two species (*Figure 2*). Estimating population sizes in number of individuals and determining coalescence events in number of years requires knowledge of the per-generation mutation rate and generation times, which are difficult to assert in the case of yeast, owing to the existence of an asexual phase. Assessing the impact of demography on the rate of adaptive substitutions, however, only requires information about the relative population sizes at the polymorphism and divergence level (Soni, Moutinho, & Eyre-Walker, 2022). Therefore, we report time in coalescence units and inferred historic variations of the inverse instantaneous coalescent rate (IICR), a measure proportional to N_e when sequences are evolving neutrally (Chikhi et al., 2018; Mazet, Rodríguez, Grusea, Boitard, & Chikhi, 2016). We further conducted PSMC analyses on pairs of individuals in each population (*Figure 2*). This analysis showed no inconsistencies between individuals within each population, as could result, for instance, from population structure. This was also true for the Cocoa population, despite its paraphyly (*Figure 1*). As expected, increasing the sample size led to a higher resolution towards more recent times. All populations have a lower recent population size compared to the one of the *cerevisiae-paradoxus* ancestor. Furthermore, most populations have a decreasing population size. A notable exception is the North American Oak population, which shows a recent expansion, and the Mediterranean oak, which seems to have undergone a recent “burst” followed by another collapse.

Linked selection, and in particular background selection (BgS), was shown to impact demographic inference (Boitard, Arredondo, Chikhi, & Mazet, 2022; Johri et al., 2021). Because the genome of *S. cerevisiae* is small and has a high density of coding regions, BgS can potentially be strong. We assessed the performance of MSMC2 in the presence of background selection in a *S. cerevisiae*-like genome by simulating data with the reference gene set and deleterious mutations (see Methods). We found that compared to simulations under a purely neutral scenario (*Figure 3A*), background selection led to an apparent recent increase in population size despite the real population size being constant (*Figure 3B*), recapping the results of previous studies (Boitard et al., 2022; Johri et al., 2021). When the real demography is declining, the resulting inferred demography shows a combination of decline followed by a recent expansion. In the light of these simulation results, we conclude that the observed decline of IICR in most yeast populations reflects a decrease in population size since the divergence with *S. paradoxus*. Yet any observed recent increase of the IICR might not reflect a population size recovery, as such an increase is expected under a scenario of constant or decreasing population size with BgS.

Historic demographic changes are a known confounding factor in inferring adaptive evolution, as fluctuations in population size can influence segregating frequencies of polymorphisms independent, or even in spite, of selection acting on sites (Olson-Manning, Wagner, & Mitchell-Olds, 2012; M. Rousselle et al., 2018; Soni et al., 2022). In particular, decreasing population sizes since the population divergence leads to underestimations of the rate of adaptive non-synonymous substitutions, an argument that was previously invoked to explain the low rate of adaptive non-synonymous substitutions in Primates and could explain the generally low rate of adaptive mutations in yeast too (Soni et al., 2022). Moreover, a lower present effective population size may lead to artificial positive correlations between the rate of adaptive substitutions and any variable

that is itself positively correlated with the strength of negative selection. In the following, we estimate the rate of adaptive non-synonymous substitutions (ω_a) in all yeast populations and assess the impact of intra-genomic factors on its distribution. Because of the difference in population sizes, we further control for the strength of negative selection.

Domestication-associated gene sets harbour lower rates of non-adaptive evolution

To gain insights into adaptive evolution in wild and domesticated populations of *S. cerevisiae*, and in particular niche-adaptation of domesticated populations of *S. cerevisiae*, we focused on genes involved in known cellular processes that are implicated in the domesticated use of *S. cerevisiae*. Specifically, domesticated yeasts largely abandon meiosis and are exposed to strong osmotic stresses due to high concentrations of mono- and disaccharides such as maltose, xylose, glucose, sucrose and fructose present early-on in fermentation reactions (Betlej et al., 2020; De Chiara et al., 2022), and also use these domestication-associated carbon sources more efficiently (De Chiara et al., 2022). Furthermore, domesticated isolates are oftentimes exposed to toxic metals, such as those present in pesticides, including copper and arsenic (Bencko & Yan Li Foong, 2017; Pietrzak & McPhail, 2004; Pontes, Čadež, Gonçalves, & Sampaio, 2019). These exposures are all typical of domestication-associated environments due to the processes involved in industrial fermentation and agriculture (Legras et al., 2018). Therefore, we expect to detect signatures of relaxed selection on meiosis-associated genes, while for osmotic- and metal stress-associated genes, we expect an increased rate of adaptive non-synonymous substitutions in domesticated populations over the wild populations. Conversely, wild populations are more tolerant to natural stresses such as heat stress and oxidative stress compared to domesticated populations (De Chiara et al., 2022; Gallone et al., 2016). From these different environmental exposures and life cycle characteristics, we have compiled seven gene sets. We will refer to these resulting gene sets as “Meiosis” (N = 125), “Osmotic Stress” (N = 17), “Domestic Carbon” (N = 81), “Metal Stress” (N = 103), “Heat Stress”

(N = 25), and “Oxidative Stress” (N = 48), respectively. Genes that were found to be annotated in several of these processes were combined into a category labelled “Multiple” (N = 59), and genes not annotated in any of these processes were categorised under the label “Other genes” (N =2,739). We hypothesise that genes involved in stress responses and carbon metabolism have a higher rate of adaptive non-synonymous substitutions in domesticated populations, while genes involved in meiotic processes should display signature of relaxed purifying selection, with a higher rate of non-adaptive non-synonymous substitutions.

Linear models showed that the gene set variable only had an impact on ω_{na}; that is, the rate of adaptive substitutions did not differ significantly between the gene categories (*Figure 4 and Table 3*). Conversely, several gene categories had a significantly lower ω_{na} compared to unlabelled genes: “metal stress”, “domestic carbon”, “osmotic stress”, “oxidative stress”, as well as genes with multiple labels. These results indicate that the corresponding genes undergo stronger purifying selection. However, this effect was found to be independent from life history or culture method (no interaction term was found significant), suggesting that this feature is general to *S. cerevisiae* and not affected by domestication. Domestication was found to have a significant positive effect on ω_{na} and a negative effect on ω_a, although the latter should be taken with caution as (1) the model residues did not fulfil the independence and normality assumption, and (2) the effect vanished when only strictly positive $\omega_a$were considered (*Table 3*). Repitching and starter-based populations had a significantly lower ω_a compared to environmental populations. They also had a higher ω_{na}, although the effect vanished when considering only strictly positive ω_a. Finally, π_S had a positive effect on ω_a, and a negative one on ω_{na}, both weakly significant, consistent with the hypothesis that population with larger effective size have more efficient negative and positive selection.

This gene set analysis did not allow to confirm *a priori* defined gene categories as key players in the adaptive process linked to domestication. Their observed higher level of conservation is more likely to be the result of their central functions than the consequence of domestication. Nonetheless, these results suggest that domesticated populations undergo a lower rate of adaptive - and a higher rate of non-adaptive - non-synonymous substitution, and that effective population size, as measured by π_S, affects both rates.

No apparent correlation between the recombination rate and the rate of adaptive non-synonymous substitutions in core yeast genes

We sought to assess how the non-synonymous substitution rates varied along the genome of *S. cerevisiae*. Recombination may affect the substitution rate at positions linked to loci under selection, a mechanism termed linked selection. A lower recombination rate is expected to slow down the fixation of advantageous mutations and reduce ω_a, but also to reduce the efficacy of purifying selection and lead to an increased fixation of deleterious mutations, thus increasing ω_{na}. To investigate the role of recombination in shaping the rate of adaptive evolution across genes in our *S. cerevisiae* populations, we binned genes into categories of approximately equal sizes according to the broad recombination landscape of *S. cerevisiae* (Liu et al., 2018) and quantified the rate of adaptive evolution across categories. We found no relation between the crossing over rate and the rate of adaptive evolution in *S. cerevisiae*, and a weak but significant negative relationship with ω_{na} (*Figure 5 and Table 3*), suggesting that regions with a higher recombination rate tend to have a more efficient purifying selection and fix less deleterious mutations. The absence of a relation between ω_a and the recombination rate indicates a limited role of Hill-Robertson interference in determining the rate of adaptive evolution in yeast. This contrasts with organisms like *Drosophila melanogaster*, where the recombination rate was shown to have a strong positive effect on the rate of adaptive evolution (Castellano, Coronado-Zamora, Campos,

Barbadilla, & Eyre-Walker, 2016), but also *Zymoseptoria tritici*, a fungal pathogen of wheat (Grandaubert et al., 2019). We further note that, as with whole genome estimates of ω_a and ω_{na}, the effect of life history and culture method were significant: domesticated populations have a lower rate of adaptive and a higher rate of non-adaptive non-synonymous substitutions. Population size (as estimated by size π_S) had a significant positive effect on ω_a, and a significant negative effect on ω_{na}, indicating that both positive and negative selection was more efficient in larger populations. We note that, despite variable transformation, the model fit was relatively poor, with significant departure from normality and independence of residues, indicating that the P-values of the model coefficients may be biased (Table 3). Furthermore, we note that the recombination rate is negatively correlated with codon usage bias, which points at the eventuality that codon usage bias masks the effect of recombination; a possibility that we investigate in more detail below.

Protein structure shapes the rate of adaptive evolution within genes

Lastly, we assessed the impact of the residues solvent exposure, as it was recently shown to be a major determinant of substitution rates in *Arabidopsis thaliana*, *Drosophila melanogaster* (Moutinho et al., 2019), and Primates (Soni et al., 2022), with adaptation predominantly occurring at solvent-exposed residues in these species. To gain a more in-depth understanding of adaptive evolution at the protein-structural level in domesticated and wild populations of *S. cerevisiae*, we computed the rate of adaptive non-synonymous substitutions in residue categories with different levels of relative solvent accessibility (RSA) and inferred ω_a and ω_{na} in these categories. We found that RSA had a strong positive, significant effect on both rates, exposed residues having a higher rate of adaptive and non-adaptive substitutions (*Figure 6* and *Table 3*). As before, we find that domesticated populations had a significantly lower ω_a, but culture methods had a significant interaction with RSA (the negative effect was stronger at exposed residues). The effect of

population size (π_S) was also significant as an interaction term with RSA: larger populations tend to undergo more adaptive substitutions at exposed residues.

We investigated whether the correlation between ω_a and RSA could be an artefact of the decreasing population size (Soni et al., 2022). Such a correlation may artificially occur if RSA was positively correlated with the strength of purifying selection. We note, however, that RSA is positively correlated with ω_{na}, implying that the strength of selection is negatively correlated with RSA. Demography is expected to bias the correlation in the opposite direction of the observed one, which is, therefore, genuine, similar to what was reported in Primates (Soni et al., 2022).

The impact of codon usage bias on the observed patterns of adaptive evolution

Finally, we assessed whether selection on codon usage could bias our estimates of adaptive rates. The Mac-Donald-Kreitman formalism involves the use of a neutral reference, for which synonymous mutations are used. Selection for optimal codons on synonymous mutations may, therefore, affect the estimates of adaptive substitution rates. As we are investigating differences in rates between genes and populations, we assessed whether differences in codon usage could explain the correlations we observed, or on the contrary, hide an existing correlation signal. We computed codon usage biases for all genes and populations of our dataset, and we found no evidence for distinct average codon usage between populations (pairwise Kendall's rank correlation tests on relative synonymous codon usage (RSCU) values, minimum correlation > 0.999, maximum P value < 8.408e-131). Pulling all populations together (with the exclusion of Ecuadorian, Malaysian, and North American Oak, for which 15 recombination rates and RSA rates could not be computed), we did not find any difference in codon usage between gene sets (Kruskal-Wallis test, P value = 0.8827). Conversely, we find a significant effect of the recombination rate on codon usage bias (Kendall's tau = -0.85, P Value = 4.073e-7, correlation between the crossing-over rate and the mean deviation of RSCU from 1 for each recombination rate category,

Supplementary Figure 3), suggesting that genes in low recombination rates exhibit strong codon bias than gene in highly recombining regions. An effect of recombination on the efficacy of selection would lead to the opposite pattern, where highly recombining regions should display stronger codon bias. Therefore, the observed correlation between codon usage bias and recombination rate is probably spurious, resulting from a covariate such as gene expression level. To test whether codon usage bias could mask the effect of recombination on the rate of adaptive non-synonymous substitutions, we re-estimated ω_a and ω_{na} in all recombination categories, using the synonymous site frequency spectrum of the highest recombination rate category, as it is the least affected by selection on codon usage. As a result, the crossing over rate has a marginally significant positive effect on the rate of adaptive non-synonymous substitutions (P value = 6.00e-2, Table 4) and a negative effect on the rate of non-adaptive non-synonymous substitutions (P value = 2.33e-3). Domesticated isolates have a significantly lower ω_a and a significantly higher ω_{na}, while the method (repitching vs starter-based), as well as the effective population size (π_S), have no significant effect. In conclusion, correcting for codon usage bias revealed a small effect of the recombination rate, the direction of which is consistent with the action of linked selection. The effect of the domestication variable remains strong after the correction, in contrast to the inoculation method and population size, whose effect proved to be not robust.

Finally, looking at codon usage bias per RSA category, we find that the first axis of a within-class correspondence analysis was well described by RSA (*Supplementary Figure 4A*) and that the mean codon usage bias was significantly negatively correlated with the average RSA (*Supplementary Figure 4B*, Kendall's tau = -0.87, -value = 1.543e-07, see Methods). A strong relation between RSA and codon usage has been previously reported in several species (Ramsey, Scherrer, Zhou, & Wilke, 2011; Zhou, Weems, & Wilke, 2009), suggesting that synonymous codon usage is under

selection in a structure-dependent manner. As the assumption of neutrality of synonymous sites is central in our DFE-based adaptive protein evolution inference methods, we reanalysed the relation between RSA and the rate of adaptive evolution using synonymous sites only from the most exposed RSA category, as these showed the least amount of codon bias. Overall, the results from this reanalysis are strongly consistent with our initial findings (*Table 4*), confirming that the relation between RSA and the rate of adaptive evolution is independent of the stronger codon bias in buried residues.

Discussion

The rate of molecular adaptive evolution was previously reported to be low in *Saccharomyces*, with a rate of adaptive non-synonymous substitutions found to be equal to zero (Gossmann et al., 2012; Liti et al., 2009). This low rate contrasts with the high rate of molecular evolution of other fungal species, in particular fungal pathogens (Grandaubert et al., 2019; Schweizer et al., 2021). Here, we showed that despite this average low rate, signatures of adaptation are present in *S. cerevisiae* populations, most notably in wild populations. We further determined which factors underlie the variation of substitution rates between populations and within genomes, using linear models accounting for the historical relationships between populations. Our phylogenetic models generally reveal a stronger phylogenetic component (parameter ρ) for the rate of non-adaptive non-synonymous substitutions compared to the rate of adaptive non-synonymous substitutions, indicating that adaptive substitutions occur in a population-specific manner.

To search for regions with signature of adaptation, we assessed three factors that may impact the substitution rate: the gene functional category, the local recombination rate, and, at the site level, the relative solvent accessibility. While all these factors proved to have a significant impact on the rate of non-adaptive non-synonymous substitutions, ω_{na}, only solvent accessibility significantly

and consistently impacted the rate of adaptive non-synonymous substitutions, ω_a. This effect of RSA is consistent with previous reports from *Drosophila*, *Arabidopsis* (Moutinho et al., 2019) and human (Soni et al., 2022) emphasizing its generality across organisms with distinct life-history traits. The impact of RSA on ω_{na} can be attributed to mutations having a generally less deleterious effect; exposed residues having on average less contact with other residues and, therefore, having more flexibility to cope with a sidechain change. We further showed that the correlation of ω_a with RSA was not an artefact of its correlation with the strength of purifying selection.

We estimated substitution rates in different gene categories, based on the possible roles the underlying genes may play in the domestication process. However, we did not find any significant differences in ω_a between these gene classes. Our results revealed that domestication is associated with a general reduction of the rate of adaptive substitutions, both in core and dispensable genes, and are consistent with domesticated populations having undergone a reduction in their effective population size, limiting the rate of adaptive – and increasing the rate of non-adaptive – non-synonymous substitutions. While these results do not preclude the occurrence of some adaptive substitutions, it parallels previous reports of gene copy number variations (CNV) associated with domesticated populations (Duan et al., 2018), supporting CNV mutations as a driver of the molecular changes underlying the domestication process. Furthermore, the comparatively constant environment associated with human practices could temper the rate of beneficial mutations. We note that our analysis focused on the genes recovered from all populations, and thus did not assess molecular rates in population-specific genes. Assessing the substitution rate in population-specific duplicated genes offers a promising perspective to dissect the genetic mechanisms of yeast domestication.

While demography inference in all populations revealed a general trend of decreasing population sizes, we did not observe systematic differences between domesticated and wild populations. Historical effective population size, however, as estimated by the neutral diversity π_S proved to be a significant determinant of the rates of both adaptive and non-adaptive substitutions. These results, therefore, suggest that in *S. cerevisiae* populations, adaptation is limited by the supply of beneficial mutations, as observed is several animal taxonomic groups (Marjolaine Rousselle et al., 2020).

Finally, we investigated the evidence for signatures of linked selection on the substitution rate. Using a previously published estimated map of crossing-over rates along the genome of *S. cerevisiae*, we found no correlation between the rate of adaptive substitutions and the recombination rate. We concluded that adaptive evolution in *S. cerevisiae* is only weakly, if at all, impeded by Hill-Robinson interference (Comeron, Williford, & Kliman, 2008; Hill & Robertson, 1966). This result is most likely explained by the combination of a globally low rate of adaptive substitutions and a high recombination rate along the *S. cerevisiae* genome. Conversely, we found evidence for the occurrence of background selection (BgS), as we observed a significant negative correlation between the recombination rate and the rate of non-adaptive non-synonymous substitutions. BgS is expected to be high in *S. cerevisiae*, given its high gene density (more than 73% of the genome encodes an exon). Although significant, the correlation between the recombination rate and ω_{na} was found to be low, probably because of the high recombination rate resulting in a low linkage disequilibrium.

In summary, our analyses revealed the occurrence of adaptive non-synonymous substitutions in several populations of *S. cerevisiae*. While we found the rate of adaptive substitutions to be generally low, we showed that accounting for intra and inter-genic factors unravelled signatures

of molecular adaptation. Adaptive substitutions were notably found in population specific genes, and otherwise at the solvent-exposed surface of proteins, following a general trend observed is other species of animals and plants and suggesting that these mutations were not directly linked to the domestication process. Moreover, we showed that domestication is associated with a reduction of the rate of adaptive non-synonymous substitutions, and that in *S. cerevisiae* wild and domesticated populations, adaptation is generally limited by the supply of mutations.

Acknowledgements

We would like to thank Ana Filipa Moutinho, Adam Eyre-Walker, and Nicolas Galtier for discussions on this project. We thank Ana Filipa Moutinho for help with running the NetSurfP-2.0 software and Haozuan Liu and Jianzhi Zhang for kindly providing data on recombination rates across the *S. cerevisiae* genome, and Fernanda Trancoso, Ana Filipa Moutinho, and Chrats Melkonian for critical reading of the manuscript. J. Y. D. would like to thank Lucie Vivier, wine grower and winemaker in Charente, France, for her explanations on yeast industrial usage in winemaking. J.Y.D. acknowledges funding from the Max Planck Society.

Conflicts of interest

The authors declare no competing interests.

Data availability

All data and scripts are accessible via FigShare at http://doi.org/10.6084/m9.figshare.21725513 (Raas & Dutheil, 2022)

Author Contributions

MWDR: Conceptualization, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review & Editing, Visualization

JYD: Conceptualization, Validation, Formal analysis, Investigation, Writing – Review & Editing, Visualization, Supervision

References

Ahmad, H. I., Ahmad, M. J., Jabbir, F., Ahmar, S., Ahmad, N., Elokil, A. A., & Chen, J. (2020). The domestication makeup: Evolution, survival, and challenges. *Frontiers in Ecology and Evolution, 8*, 103.

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika, 60*(2), 255-265. doi:10.1093/biomet/60.2.255

Alboukadel Kassambara. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. Retrieved from https://CRAN.R-project.org/package=ggpubr

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Bencko, V., & Yan Li Foong, F. (2017). The history of arsenical pesticides and health risks related to the use of Agent Blue. *Ann Agric Environ Med, 24*(2), 312-316. doi:10.26444/aaem/74715

Betlej, G., Bator, E., Oklejewicz, B., Potocki, L., Górka, A., Slowik-Borowiec, M., . . . Kwiatkowska, A. (2020). Long-Term Adaption to High Osmotic Stress as a Tool for Improving Enological Characteristics in Industrial Wine Yeast. *Genes (Basel), 11*(5). doi:10.3390/genes11050576

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., . . . Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res, 14*(4), 708-715. doi:10.1101/gr.1933104

Boitard, S., Arredondo, A., Chikhi, L., & Mazet, O. (2022). Heterogeneity in effective size across the genome: effects on the inverse instantaneous coalescence rate (IICR) and implications for demographic inference under linked selection. *Genetics, 220*(3). doi:10.1093/genetics/iyac008

Botstein, D., & Fink, G. R. (2011). Yeast: an experimental organism for 21st Century biology. *Genetics, 189*(3), 695-704. doi:10.1534/genetics.111.130765

Castellano, D., Coronado-Zamora, M., Campos, J. L., Barbadilla, A., & Eyre-Walker, A. (2016). Adaptive Evolution Is Substantially Impeded by Hill-Robertson Interference in Drosophila. *Mol Biol Evol, 33*(2), 442-455. doi:10.1093/molbev/msv236

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In U. Bastolla, M. Porto, H. E. Roman, & M. Vendruscolo (Eds.), *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (pp. 207-232). Berlin, Heidelberg: Springer Berlin Heidelberg.

Charif, D., Thioulouse, J., Lobry, J. R., & Perrière, G. (2004). Online synonymous codon usage analyses with the ade4 and seqinR packages. *Bioinformatics, 21*(4), 545-547. doi:10.1093/bioinformatics/bti037

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics, 10*(3), 195-205. doi:10.1038/nrg2526

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2018). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice. *Heredity, 120*(1), 13-24. doi:10.1038/s41437-017-0005-6

Claus O. Wilke. (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. Retrieved from https://CRAN.R-project.org/package=cowplot

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., . . . de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423. doi:10.1093/bioinformatics/btp163

Comeron, J. M., Williford, A., & Kliman, R. M. (2008). The Hill–Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. *Heredity, 100*(1), 19-31. doi:10.1038/sj.hdy.6801059

Connallon, T., & Knowles, L. L. (2007). Recombination rate and protein evolution in yeast. *BMC Evolutionary Biology, 7*(1), 235. doi:10.1186/1471-2148-7-235

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics, 33*(18), 2938-2940. doi:10.1093/bioinformatics/btx364

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience, 10*(2). doi:10.1093/gigascience/giab008

Darwin, C. (1868). *The variation of animals and plants under domestication*. London: London : J. Murray.

De Chiara, M., Barré, B. P., Persson, K., Irizar, A., Vischioni, C., Khaiwal, S., . . . Liti, G. (2022). Domestication reprogrammed the budding yeast life cycle. *Nat Ecol Evol, 6*(4), 448-460. doi:10.1038/s41559-022-01671-9

De León-Rodríguez, A., González-Hernández, L., Barba de la Rosa, A. P., Escalante-Minakata, P., & López, M. G. (2006). Characterization of volatile compounds of Mezcal, an ethnic alcoholic beverage obtained from Agave salmiana. *J Agric Food Chem, 54*(4), 1337-1341. doi:10.1021/jf052154+

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature, 418*(6898), 700-707. doi:10.1038/nature01019

Djeni, T. N., Kouame, K. H., Ake, F. D. M., Amoikon, L. S. T., Dje, M. K., & Jeyaram, K. (2020). Microbial Diversity and Metabolite Profiles of Palm Wine Produced From Three Different Palm Tree Species in Côte d'Ivoire. *Scientific Reports, 10*(1), 1715. doi:10.1038/s41598-020-58587-2

Duan, S.-F., Han, P.-J., Wang, Q.-M., Liu, W.-Q., Shi, J.-Y., Li, K., . . . Bai, F.-Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. *Nature Communications, 9*(1), 2690. doi:10.1038/s41467-018-05106-7

Duina, A. A., Miller, M. E., & Keeney, J. B. (2014). Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. *Genetics, 197*(1), 33-48. doi:10.1534/genetics.114.163188

Dutheil, J. Y., Gaillard, S., & Stukenbrock, E. H. (2014). MafFilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics, 15*, 53. doi:10.1186/1471-2164-15-53

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics, 8*(8), 610-618. doi:10.1038/nrg2146

Eyre-Walker, A., & Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol, 26*(9), 2097-2108. doi:10.1093/molbev/msp119

Eyre-Walker, A., Woolfit, M., & Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics, 173*(2), 891-900. doi:10.1534/genetics.106.057570

Fox, J., & Weisberg, S. (2018). *An R Companion to Applied Regression*: SAGE Publications.

Fuller, D. Q., & Stevens, C. J. (2019). Between domestication and civilization: the role of agriculture and arboriculture in the emergence of the first urban societies. *Vegetation History and Archaeobotany, 28*(3), 263-282. doi:10.1007/s00334-019-00727-4

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., . . . Verstrepen, K. J. (2016). Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell, 166*(6), 1397-1410.e1316. doi:10.1016/j.cell.2016.08.020

Galtier, N. (2016). Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS Genet, 12*(1), e1005774. doi:10.1371/journal.pgen.1005774

Goddard, M. R., Godfray, H. C., & Burt, A. (2005). Sex increases the efficacy of natural selection in experimental yeast populations. *Nature, 434*(7033), 636-640. doi:10.1038/nature03405

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., . . . Oliver, S. G. (1996). Life with 6000 genes. *Science, 274*(5287), 546, 563-547. doi:10.1126/science.274.5287.546

Gossmann, T. I., Keightley, P. D., & Eyre-Walker, A. (2012). The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol Evol, 4*(5), 658-667. doi:10.1093/gbe/evs027

Grafen, A. (1989). The phylogenetic regression. *Philos Trans R Soc Lond B Biol Sci, 326*(1233), 119-157. doi:10.1098/rstb.1989.0106

Grandaubert, J., Dutheil, J. Y., & Stukenbrock, E. H. (2019). The genomic determinants of adaptive evolution in a fungal pathogen. *Evol Lett, 3*(3), 299-312. doi:10.1002/evl3.117

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., . . . Dutheil, J. Y. (2013). Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol Biol Evol, 30*(8), 1745-1750. doi:10.1093/molbev/mst097

Haubold, B., Klötzl, F., & Pfaffelhuber, P. (2015). andi: fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics, 31*(8), 1169-1175. doi:10.1093/bioinformatics/btu815

Hill, W. G., & Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet Res, 8*(3), 269-294.

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics, 116*(1), 153-159. doi:10.1093/genetics/116.1.153

Jo, W. J., Loguinov, A., Chang, M., Wintz, H., Nislow, C., Arkin, A. P., . . . Vulpe, C. D. (2007). Identification of Genes Involved in the Toxic Response of Saccharomyces cerevisiae against Iron and Copper Overload by Parallel Analysis of Deletion Mutants. *Toxicological Sciences, 101*(1), 140-151. doi:10.1093/toxsci/kfm226

Johri, P., Riall, K., Becher, H., Excoffier, L., Charlesworth, B., & Jensen, J. D. (2021). The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects. *Molecular Biology and Evolution*, *38*(7), 2986-3003. doi:10.1093/molbev/msab050 %J Molecular Biology and Evolution

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, *30*(4), 772-780. doi:10.1093/molbev/mst010

Klausen, M. S., Jespersen, M. C., Nielsen, H., Jensen, K. K., Jurtz, V. I., Sønderby, C. K., . . . Marcatili, P. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, *87*(6), 520-527. doi:https://doi.org/10.1002/prot.25674

Lahue, C., Madden, A. A., Dunn, R. R., & Smukowski Heil, C. (2020). History and Domestication of Saccharomyces cerevisiae in Bread Baking. *Front Genet*, *11*, 584718. doi:10.3389/fgene.2020.584718

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution*, *29*(1), 33-41. doi:https://doi.org/10.1016/j.tree.2013.09.009

Large, C. R. L., Hanson, N., Tsouris, A., Saada, O. A., Koonthongkaew, J., Toyokawa, Y., . . . Dunham, M. J. (2020). Genomic stability and adaptation of beer brewing yeasts during serial repitching in the brewery. *bioRxiv*, 2020.2006.2026.166157. doi:10.1101/2020.06.26.166157

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol Biol Evol*, *32*(10), 2798-2800. doi:10.1093/molbev/msv150

Legras, J. L., Galeote, V., Bigey, F., Camarasa, C., Marsit, S., Nidelet, T., . . . Dequin, S. (2018). Adaptation of S. cerevisiae to Fermented Food Environments Reveals Remarkable Genome Plasticity and the Footprints of Domestication. *Mol Biol Evol*, *35*(7), 1712-1727. doi:10.1093/molbev/msy066

Legras, J. L., Merdinoglu, D., Cornuet, J. M., & Karst, F. (2007). Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Mol Ecol, 16*(10), 2091-2102. doi:10.1111/j.1365-294X.2007.03266.x

Liti, G. (2015). The fascinating and secret wild life of the budding yeast S. cerevisiae. *Elife, 4*. doi:10.7554/eLife.05835

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., . . . Louis, E. J. (2009). Population genomics of domestic and wild yeasts. *Nature, 458*(7236), 337-341. doi:10.1038/nature07743

Liu, H., Maclean, C. J., & Zhang, J. (2018). Evolution of the Yeast Recombination Landscape. *Molecular Biology and Evolution, 36*(2), 412-422. doi:10.1093/molbev/msy233

Ludlow, C. L., Cromie, G. A., Garmendia-Torres, C., Sirr, A., Hays, M., Field, C., . . . Dudley, A. M. (2016). Independent Origins of Yeast Associated with Coffee and Cacao Fermentation. *Curr Biol, 26*(7), 965-971. doi:10.1016/j.cub.2016.02.012

Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., Andersen, L. N., . . . Schierup, M. H. (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS Genet, 8*(12), e1003125. doi:10.1371/journal.pgen.1003125

Malaspinas, A.-S., Westaway, M. C., Muller, C., Sousa, V. C., Lao, O., Alves, I., . . . Willerslev, E. (2016). A genomic history of Aboriginal Australia. *Nature, 538*(7624), 207-214. doi:10.1038/nature18299

Marais, G., & Charlesworth, B. (2003). Genome evolution: recombination speeds up adaptive evolution. *Curr Biol, 13*(2), R68-70. doi:10.1016/s0960-9822(02)01432-x

Marsit, S., Leducq, J. B., Durand, É., Marchant, A., Filteau, M., & Landry, C. R. (2017). Evolutionary biology through the lens of budding yeast comparative genomics. *Nat Rev Genet, 18*(10), 581-598. doi:10.1038/nrg.2017.49

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity, 116*(4), 362-371. doi:10.1038/hdy.2015.104

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature, 351*(6328), 652-654. doi:10.1038/351652a0

McGovern, P. E., Zhang, J., Tang, J., Zhang, Z., Hall, G. R., Moreau, R. A., . . . Wang, C. (2004). Fermented beverages of pre- and proto-historic China. *Proc Natl Acad Sci U S A, 101*(51), 17593-17598. doi:10.1073/pnas.0407921102

Money, N. P. (2018). *The rise of yeast : how the sugar fungus shaped civilization*.

Moutinho, A. F., Trancoso, F. F., & Dutheil, J. Y. (2019). The Impact of Protein Architecture on Adaptive Evolution. *Mol Biol Evol, 36*(9), 2013-2028. doi:10.1093/molbev/msz134

Moyers, B. T., Morrell, P. L., & McKay, J. K. (2017). Genetic Costs of Domestication and Improvement. *Journal of Heredity, 109*(2), 103-116. doi:10.1093/jhered/esx069

Olson-Manning, C. F., Wagner, M. R., & Mitchell-Olds, T. (2012). Adaptive evolution: evaluating empirical support for theoretical predictions. *Nat Rev Genet, 13*(12), 867-877. doi:10.1038/nrg3322

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*(3), 526-528. doi:10.1093/bioinformatics/bty633

Parapouli, M., Vasileiadis, A., Afendra, A.-S., & Hatziloukas, E. (2020). Saccharomyces cerevisiae and its industrial applications. *AIMS microbiology, 6*(1), 1-31. doi:10.3934/microbiol.2020001

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., . . . Schacherer, J. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature, 556*(7701), 339-344. doi:10.1038/s41586-018-0030-5

Pietrzak, U., & McPhail, D. C. (2004). Copper accumulation, distribution and fractionation in vineyard soils of Victoria, Australia. *Geoderma, 122*(2), 151-166. doi:https://doi.org/10.1016/j.geoderma.2004.01.005

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2021). nlme: Linear and Nonlinear Mixed Effects Models. Retrieved from https://CRAN.R-project.org/package=nlme

Pontes, A., Čadež, N., Gonçalves, P., & Sampaio, J. P. (2019). A Quasi-Domesticate Relic Hybrid Population of Saccharomyces cerevisiae × S. paradoxus Adapted to Olive Brine. *Front Genet, 10*, 449. doi:10.3389/fgene.2019.00449

R Development Core Team. (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://R-project.org

Rambaut, A. (2018). FigTree. Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

Ramsey, D. C., Scherrer, M. P., Zhou, T., & Wilke, C. O. (2011). The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics, 188*(2), 479-488. doi:10.1534/genetics.111.128025

Ranwez, V., Douzery, E. J. P., Cambon, C., Chantret, N., & Delsuc, F. (2018). MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. *Mol Biol Evol, 35*(10), 2582-2584. doi:10.1093/molbev/msy159

Ross-Ibarra, J., Morrell, P. L., & Gaut, B. S. (2007). Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proc Natl Acad Sci U S A, 104 Suppl 1*(Suppl 1), 8641-8648. doi:10.1073/pnas.0700643104

Rousselle, M., Mollion, M., Nabholz, B., Bataillon, T., & Galtier, N. (2018). Overestimation of the adaptive substitution rate in fluctuating populations. *Biol Lett, 14*(5). doi:10.1098/rsbl.2018.0055

Rousselle, M., Simion, P., Tilak, M.-K., Figuet, E., Nabholz, B., & Galtier, N. (2020). Is adaptation limited by mutation? A timescale-dependent effect of genetic diversity on the adaptive substitution rate in animals. *PLOS Genetics, 16*(4), e1008668. doi:10.1371/journal.pgen.1008668

Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (pp. 147-166). New York, NY: Springer US.

Schwan, R. F., & Wheals, A. E. (2004). The microbiology of cocoa fermentation and its role in chocolate quality. *Crit Rev Food Sci Nutr, 44*(4), 205-221. doi:10.1080/10408690490464104

Schweizer, G., Haider, M. B., Barroso, G. V., Rössel, N., Münch, K., Kahmann, R., & Dutheil, J. Y. (2021). Population Genomics of the Maize Pathogen Ustilago maydis: Demographic History and Role of Virulence Clusters in Adaptation. *Genome Biology and Evolution, 13*(5). doi:10.1093/gbe/evab073

Sicard, D., & Legras, J. L. (2011). Bread, beer and wine: yeast domestication in the Saccharomyces sensu stricto complex. *C R Biol, 334*(3), 229-236. doi:10.1016/j.crvi.2010.12.016

Smith, N. G. C., & Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature, 415*(6875), 1022-1024. doi:10.1038/4151022a

Soni, V., Moutinho, A. F., & Eyre-Walker, A. (2022). Changing Population Size in McDonald-Kreitman Style Analyses: Artifactual Correlations and Adaptive Evolution between Humans and Chimpanzees. *Genome Biol Evol, 14*(2). doi:10.1093/gbe/evac022

Tataru, P., Mollion, M., Glémin, S., & Bataillon, T. (2017). Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. *Genetics, 207*(3), 1103-1119. doi:10.1534/genetics.117.300323

Thioulouse, J. D. S. p. D. A.-B. a. S. A. l. J. T. P. S. (2018). *Multivariate Analysis of Ecological Data with ade4*.

Thorsen, M., Perrone, G. G., Kristiansson, E., Traini, M., Ye, T., Dawes, I. W., . . . Tamás, M. J. (2009). Genetic basis of arsenite and cadmium tolerance in Saccharomyces cerevisiae. *BMC Genomics, 10*(1), 105. doi:10.1186/1471-2164-10-105

Venables, W. N. (2002). *Modern applied statistics with S / W.N. Venables, B.D. Ripley*. New York: Springer.

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. *Mol Ecol, 21*(22), 5404-5417. doi:10.1111/j.1365-294X.2012.05732.x

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Yue, J.-X., Li, J., Aigrain, L., Hallin, J., Persson, K., Oliver, K., . . . Liti, G. (2017). Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nature Genetics, 49*(6), 913-924. doi:10.1038/ng.3847

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proc Natl Acad Sci U S A, 105*(33), 11597-11604. doi:10.1073/pnas.0801317105

Zhou, T., Weems, M., & Wilke, C. O. (2009). Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Molecular Biology and Evolution, 26*(7), 1571-1580. doi:10.1093/molbev/msp070

Tables

Table 1: Genome-wide estimates of the rate of adaptive and non-adaptive non-synonymous substitutions in domesticated and wild populations of *S. cerevisiae*.

LifeHistory	Method	Core genes			All genes			Core genes			All genes			Genome
		α	ωa	ωna	α	ωa	ωna	πN	πS	$\pi N/\pi S$	πN	πS	$\pi N/\pi S$	π
Wild	Environmental	2.68%	0.004	0.142	54.58%	0.0942	0.0783	4.75E-06	4.47E-06	1.06	1.73E-05	1.26E-04	0.14	8.49E-04
Wild	Environmental	70.57%	0.105	0.044	74.35%	0.1290	0.0445	4.25E-06	6.24E-05	0.07	1.44E-05	1.17E-04	0.12	1.97E-03
Wild	Environmental	40.59%	0.059	0.087	52.56%	0.0902	0.0813	1.13E-03	7.58E-03	0.15	1.20E-03	7.80E-03	0.15	4.26E-03
Wild	Environmental	1.65%	0.002	0.143	52.45%	0.0900	0.0816	6.69E-06	9.60E-06	0.70	1.61E-05	1.06E-04	0.15	1.50E-03
Wild	Environmental	14.71%	0.022	0.125	25.89%	0.0444	0.1271	1.39E-03	5.72E-03	0.24	1.47E-03	6.03E-03	0.24	3.52E-03
Wild	Environmental	10.07%	0.015	0.132	26.74%	0.0461	0.1260	8.18E-04	4.65E-03	0.18	8.58E-04	4.77E-03	0.18	2.60E-03
Non-Asian & Domesticated	Environmental	0.40%	0.001	0.146	10.97%	0.0189	0.1531	4.85E-04	1.86E-03	0.26	4.98E-04	1.92E-03	0.26	1.73E-03
Non-Asian & Domesticated	Environmental	8.79%	0.013	0.136	27.35%	0.0481	0.1278	7.06E-04	3.94E-03	0.18	8.10E-04	4.46E-03	0.18	3.76E-03
Non-Asian & Domesticated	Repitching	1.84%	0.003	0.146	14.37%	0.0250	0.1489	1.42E-03	6.64E-03	0.21	1.50E-03	7.01E-03	0.21	4.41E-03
Non-Asian & Domesticated	Repitching	0.00%	0.000	0.149	0.03%	0.0001	0.1748	6.33E-04	1.58E-03	0.40	6.67E-04	1.68E-03	0.40	1.52E-03
Non-Asian & Domesticated	Repitching	2.33%	0.003	0.142	17.38%	0.0297	0.1408	1.24E-03	4.78E-03	0.26	1.28E-03	4.91E-03	0.26	2.79E-03
Non-Asian & Domesticated	Starter-based	0.00%	0.000	0.148	0.00%	0.0000	0.1729	6.04E-04	1.46E-03	0.41	6.27E-04	1.47E-03	0.43	1.15E-03
Non-Asian & Domesticated	Starter-based	0.01%	0.000	0.150	0.00%	0.0000	0.1748	1.26E-03	3.55E-03	0.36	1.29E-03	3.68E-03	0.35	2.46E-03
Non-Asian & Domesticated	Starter-based	0.70%	0.001	0.146	12.02%	0.0207	0.1511	1.11E-03	4.99E-03	0.22	1.16E-03	5.24E-03	0.22	2.70E-03
Non-Asian & Domesticated	Starter-based	0.00%	0.000	0.146	0.00%	0.0000	0.1707	1.19E-03	3.22E-03	0.37	1.21E-03	3.25E-03	0.37	2.33E-03
Asian & Domesticated	Environmental	0.00%	0.000	0.147	0.01%	0.0000	0.1728	4.00E-04	1.30E-03	0.31	4.14E-04	1.31E-03	0.32	9.75E-04

Table 2: Effect of life history, culture method and effective population size (as estimated by π_S) on the genome average rate of adaptive and non-adaptive non-synonymous substitutions.

	a			wa			wna		
Variable	Coefficient	P value		Coefficient	P value		Coefficient	P value	
(Intercept)	-78.52	0.002	(**)	-1414.26	0.008	(**)	-0.12	1.11E-78	(***)
Life history, non-Asian & Domesticated	-60.23	0.057	(.)	-954.65	0.145		9.63E-09	0.084	(.)
Life history, Asian & Domesticated	-80.76	0.037	(*)	-347.76	0.307		5.88E-09	0.122	
Method, repitching	-39.84	0.249		-304.33	0.586		4.21E-09	0.401	
Method, starter-based	-54.28	0.075	(.)	-367.73	0.053	(.)	4.86E-09	0.040	(*)
pS	11071.48	0.023	(*)	93421.03	0.030	(*)	-1.33E-06	0.010	(*)
Phylogeny (rho)	1.52E-09			1.02			0.67		
Normality		0.037	(*)		0.093	(.)		0.044	(*)
Independence		0.219			0.467			0.097	(.)

Table 3: Effect of the gene functional category, local recombination rate and solvent exposure on the rate of adaptive and non-adaptive non-synonymous substitutions. Generalised least square models accounting for populations' historical relationships. The reported models are the most parsimonious models according to Akaike's information criterion, after transforming the response variable with a Box-Cox transform. Variable with an empty coefficient were not retained in the final model, for a given response variable.

Dataset	Variable (1)	Wa						Wna					
		All data			Positive Wa			All data			Positive Wa		
		Coefficient	P value		Coefficient	P value		Coefficient	P value		Coefficient	P value	
Gene set	(Intercept)	-1425.6	4.40E-18	(***)	-361.7	5.41E-20	(***)	-0.401	4.84E-249	(***)	-0.477	4.30E-143	(***)
	Gene set, metal stress							-0.002	2.57E-07	(***)	-0.003	4.34E-03	(**)
	Gene set, domestic carbon							-0.002	8.73E-09	(***)	-0.003	3.54E-03	(**)
	Gene set, heat stress							0.000	8.40E-01		0.000	5.85E-01	
	Gene qet, meiosis							0.000	8.58E-01		0.001	4.42E-01	
	Gene set, osmotic stress							-0.001	8.73E-05	(***)	-0.002	6.47E-02	(.)
	Gene set, oxidative stress							-0.002	9.33E-08	(***)	-0.003	4.46E-03	(**)
	Gene set, multiple sets							-0.002	1.16E-12	(***)	-0.004	2.19E-05	(***)
	Life history, non-Asian & domesticated	-411.8	1.86E-02	(*)				0.001	6.57E-03	(**)	0.001	6.99E-02	(.)
	Life history, Asian & domesticated	-684.4	1.33E-03	(**)				0.001	9.51E-04	(***)	0.002	5.23E-02	(.)
	Method, repitching	-747.1	1.68E-04	(***)	-192.2	4.57E-05	(***)	0.001	4.60E-04	(***)	0.002	1.07E-02	(*)
	Method, starter-based	-795.8	2.78E-06	(***)	-185.6	1.22E-05	(***)	0.001	1.55E-04	(***)	0.001	6.77E-02	(.)
	pS	99180.4	2.99E-04	(***)	16730.2	1.71E-02	(*)	-0.131	3.43E-04	(***)	-0.221	4.29E-02	(*)
	Phylogeny (r)	0.0			0.0			0.000			0.000		
	Normality		1.31E-02	(*)		2.99E-01			4.88E-03	(**)		4.39E-02	(*)
	Independence		1.49E-03	(**)		6.13E-02	(.)		7.71E-01			8.09E-01	
ecombinati	(Intercept)	-123.1	5.18E-33	(***)	-41.7	5.30E-30	(***)	-0.443	0.00E+00	(***)	-0.459	1.91E-280	(***)
	Recombination rate							-0.001	7.93E-03	(**)	-0.001	3.37E-02	(*)
	Life history, non-Asian & domesticated	-44.3	2.96E-06	(***)	-11.3	2.41E-04	(***)	0.002	8.77E-07	(***)	0.002	4.55E-05	(***)
	Life history, Asian & Domesticated	-38.9	6.28E-04	(***)	-15.7	1.58E-04	(***)	0.002	9.27E-04	(***)	0.003	2.88E-04	(***)
	Method, repitching	-35.2	3.35E-04	(***)	-10.4	3.08E-03	(**)	0.001	2.86E-03	(**)	0.001	1.35E-02	(*)
	Method, starter-based	-46.4	1.58E-08	(***)	-7.6	1.15E-02	(*)	0.002	9.53E-07	(***)	0.001	2.91E-02	(*)
	pS	10209.3	6.73E-11	(***)	2085.8	1.34E-04	(***)	-0.412	6.53E-10	(***)	-0.323	4.62E-04	(***)
	Phylogeny (r)	0.0			0.1			0.122			0.124		
	Normality		3.99E-07	(***)		2.60E-03	(**)		9.07E-03	(**)		8.58E-02	(.)
	Independence		2.89E-10	(***)		2.50E-05	(***)		4.93E-01			2.84E-01	
RSA	(Intercept)	-191.0	3.26E-32	(***)	-41.0	2.16E-20	(***)	-1.075	2.33E-174	(***)	-0.949	2.41E-124	(***)
	RSA	92.6	4.88E-04	(***)	22.8	1.82E-03	(**)	0.325	7.33E-29	(***)	0.195	2.02E-18	(***)
	Life history, non-Asian & domesticated	-38.0	5.73E-05	(***)	-8.9	5.36E-05	(***)	0.001	9.23E-01		0.002	8.74E-01	
	Life history, Asian & domesticated	-35.5	1.83E-03	(**)	-15.7	9.31E-07	(***)	0.010	4.49E-01		0.031	4.76E-02	(*)
	Method, repitching	-2.2	8.83E-01		-3.4	1.49E-01		0.006	4.15E-01		0.001	9.17E-01	
	Method, starter-based	-5.5	6.63E-01		1.9	3.09E-01		0.009	7.05E-04	(***)	-0.009	4.00E-03	(**)
	pS	2128.8	4.20E-01		-160.7	8.16E-01		-0.546	6.60E-01				
	RSA:Life history, non-Asian & domesticated							0.114	3.92E-05	(***)	0.087	5.37E-04	(***)
	RSA:Life history, Asian & domesticated							0.020	4.96E-01		0.029	3.19E-01	
	RSA:Method, repitching	-73.5	1.51E-02	(*)									
	RSA:Method, starter-based	-93.4	3.03E-04	(***)									
	RSA:pS	22295.7	7.26E-05	(***)	2570.6	6.24E-02	(.)	-14.566	1.32E-07	(***)			
	Phylogeny (r)	0.0			0.3			0.822			0.938		
	Normality		2.11E-05	(***)		2.72E-01			2.72E-16	(***)		3.08E-06	(***)
	Independence		2.16E-01			1.11E-01			8.71E-01			5.07E-05	(***)

(1) The colon ":" symbol stands for the interaction term between the left and right variables

Dataset	Variable	ωa						ωna					
		All data			Positive ωa			All data			Positive ωna		
		Coefficient	P value		Coefficient	P value		Coefficient	P value		Coefficient	P value	
ecombinatic	(Intercept)	-1223.4	3.05E-12	(***)	-253.2	5.87E-14	(***)	-0.394	0.00E+00	(***)	-0.487	1.17E-245	(***)
	Rec	337.3	4.21E-02	(*)	64.7	3.81E-02	(*)	-0.001	2.21E-03	(**)	-0.003	1.51E-03	(**)
	Life history	-1235.9	1.27E-22	(***)	-185.1	3.99E-14	(***)	0.002	4.64E-13	(***)	0.005	1.98E-11	(***)
	Life history	-1196.2	6.80E-14	(***)	-204.9	1.40E-10	(***)	0.002	1.79E-11	(***)	0.005	1.48E-08	(***)
	Method, repitching							0.000	8.17E-01				
	Method, starter-based							0.000	5.42E-02	(.)			
	Phylogeny	0.2			0.2			0.466			0.409	NA	
	Normality		3.25E-07	(***)		1.30E-01			1.63E-01			2.09E-01	
	Independence		5.31E-04	(***)		8.69E-02	(.)		9.91E-04	(***)		3.90E-02	(*)
RSA	(Intercept)	-1713.6	9.72E-37	(***)	-158.4	9.67E-35	(***)	-1.043	5.66E-190	(***)	-0.905	9.50E-117	(***)
	RSA	1082.4	7.99E-07	(***)	145.3	2.23E-17	(***)	0.319	6.40E-38	(***)	0.178	1.07E-19	(***)
	Life history	-385.3	6.72E-07	(***)	-8.3	4.79E-01		-0.003	7.84E-01		-0.008	5.24E-01	
	Life history	-213.0	2.09E-02	(*)	-5.8	7.69E-01		-0.002	8.54E-01		-0.013	3.25E-01	
	Method, re	168.1	1.73E-01		-8.0	2.99E-01		-0.001	9.21E-01				
	Method, st	139.4	1.81E-01		8.3	2.65E-01		-0.002	6.50E-01				
	πS	11740.2	5.86E-01		3705.1	1.28E-03	(**)	0.073	9.34E-01		-2.449	5.43E-03	(**)
	RSA:Life history, non-Asian & domesticated				-50.8	2.13E-02	(*)	0.104	7.88E-05	(***)	0.093	3.02E-04	(***)
	RSA:Life history Asian & domesticated				-74.8	2.22E-02	(*)	-0.002	9.24E-01		0.052	3.93E-02	(*)
	RSA:Metho	-981.8	7.93E-05	(***)				0.016	4.83E-01				
	RSA:Metho	-1018.1	1.83E-06	(***)				0.015	5.16E-02	(.)			
	RSA:πS	201172.2	1.25E-05	(***)				-16.375	1.67E-15	(***)			
	Phylogeny	0.0			0.0			0.949			0.854		
	Normality		2.87E-03	(**)		9.95E-01			3.72E-16	(***)		1.29E-06	(***)
	Independence		5.91E-01			5.64E-02	(.)		2.03E-01			3.21E-01	

Table 4 : Effect of local recombination rate and solvent exposure on the rate of adaptive and non-adaptive non-synonymous substitutions after correction for selection on synonymous codon usage.

Figures

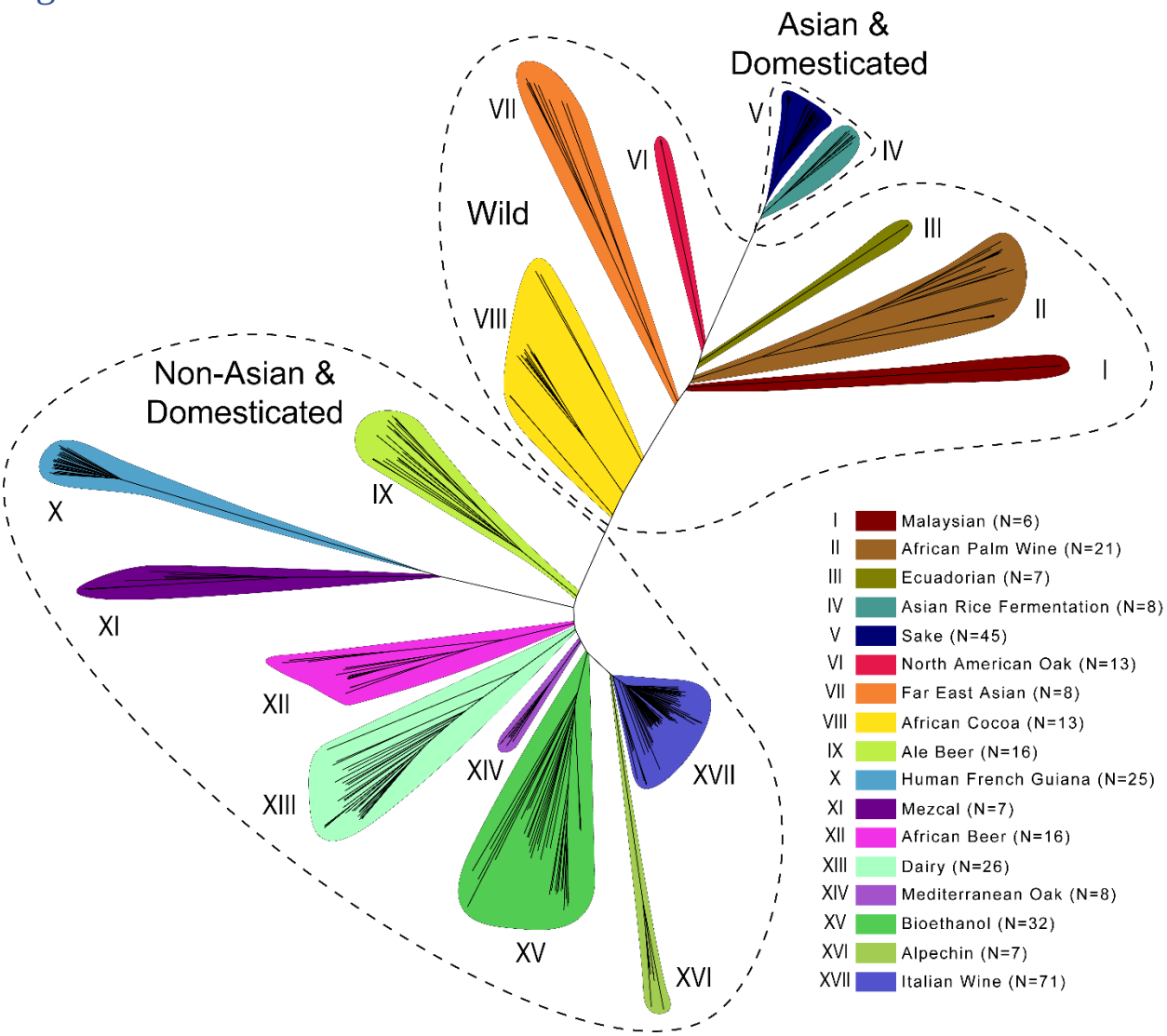

Figure 1: Phylogenetic tree of S. cerevisiae populations. Unrooted phylogenetic tree highlighting the distinct lineages considered as populations in this study. Life history clusters are indicated with dotted lines. Colours indicate population identity as specified in the in-figure legend and numerals.

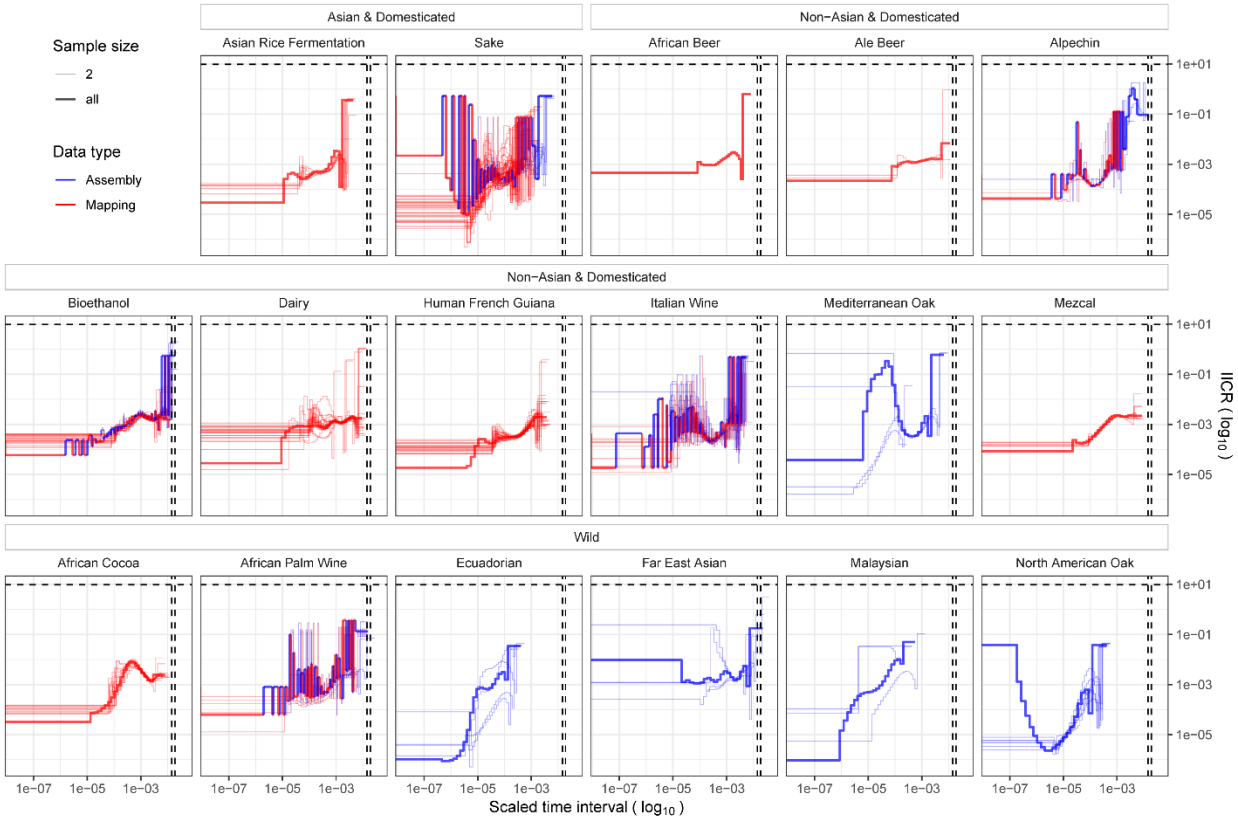

Figure 2: Demographic histories of *S. cerevisiae* populations. The inversed instantaneous coalescent rate (IICR, as inferred with MSMC2) is plotted as a function of time in coalescence units. The demographic inference was conducted either from *de novo* genome alignments (haploid or homozygous diploid isolates, blue lines) or from SNP calling after mapping reads to the *S. cerevisiae* reference genome (heterozygous diploid isolates, red lines). Thick lines represent the demography inferred from all isolates using MSMC. Thin lines represent the demography inferred from pairs of haplotypes using PSMC. Vertical dash lines: divergence time from *S. paradoxus*; horizontal dash lines: average ancestral IICR for the *S. paradoxus* – *S. cerevisiae* ancestor as estimated from IMCoalHMM

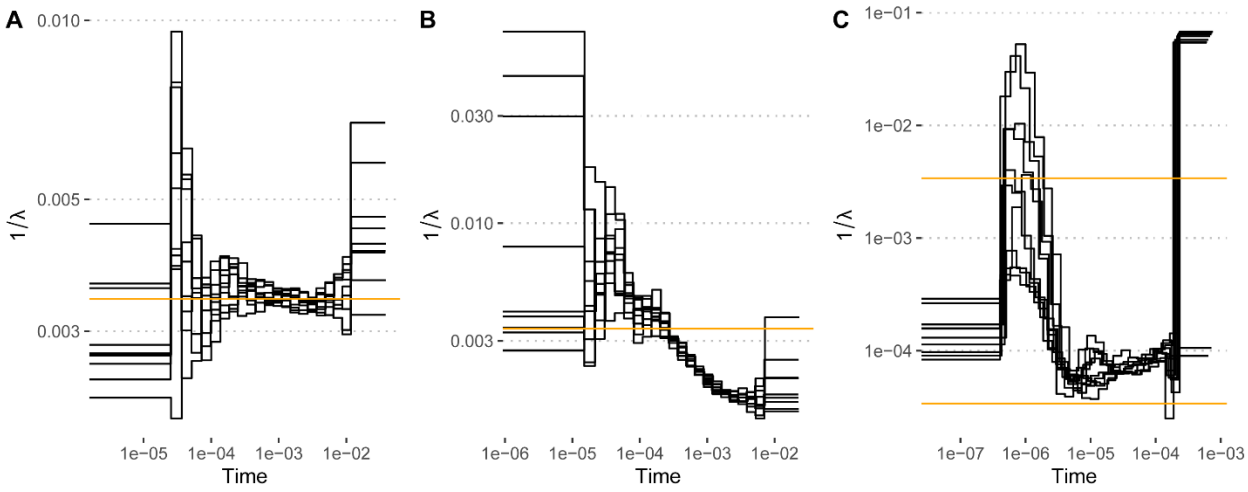

Figure 3: Demography recovery from simulated data with yeast genome characteristics, including chromosome sizes, gene content and callability mask. Ten replicates are plotted with black lines. A) Simulations under a neutral scenario and a flat demography (constant population size, plotted in orange). B) Simulations under a scenario with background selection and a flat demography (constant population size, plotted in orange). C) Simulations under a scenario with background selection and an exponentially declining population (from 10,000 individuals, upper orange line) to 100 individuals (lower orange line).

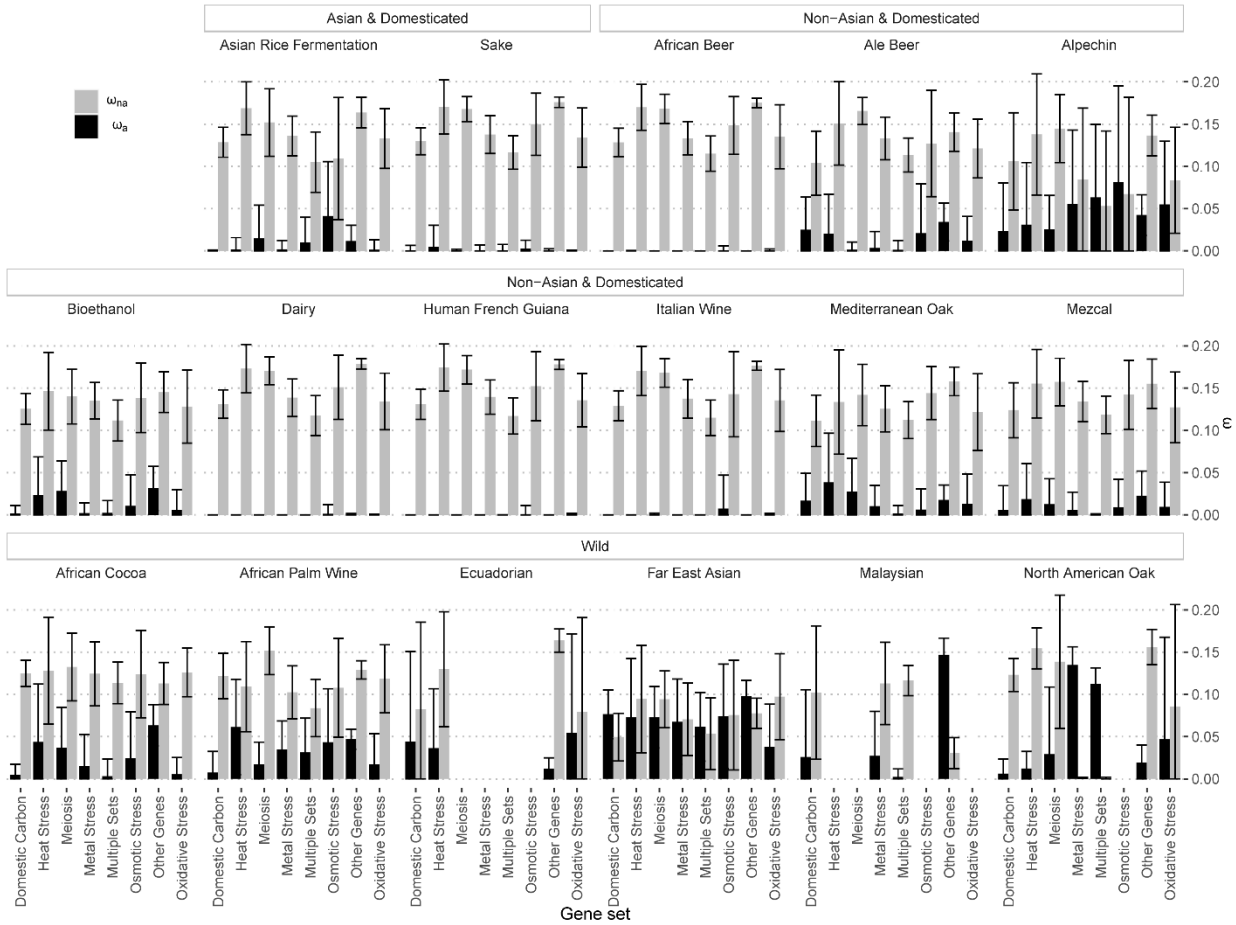

Figure 4: Rates of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions in distinct gene set. Bars indicate the mean over a 150 bootstrap replicates, and the error bars the corresponding 95% confidence interval (1.96 * standard deviation).

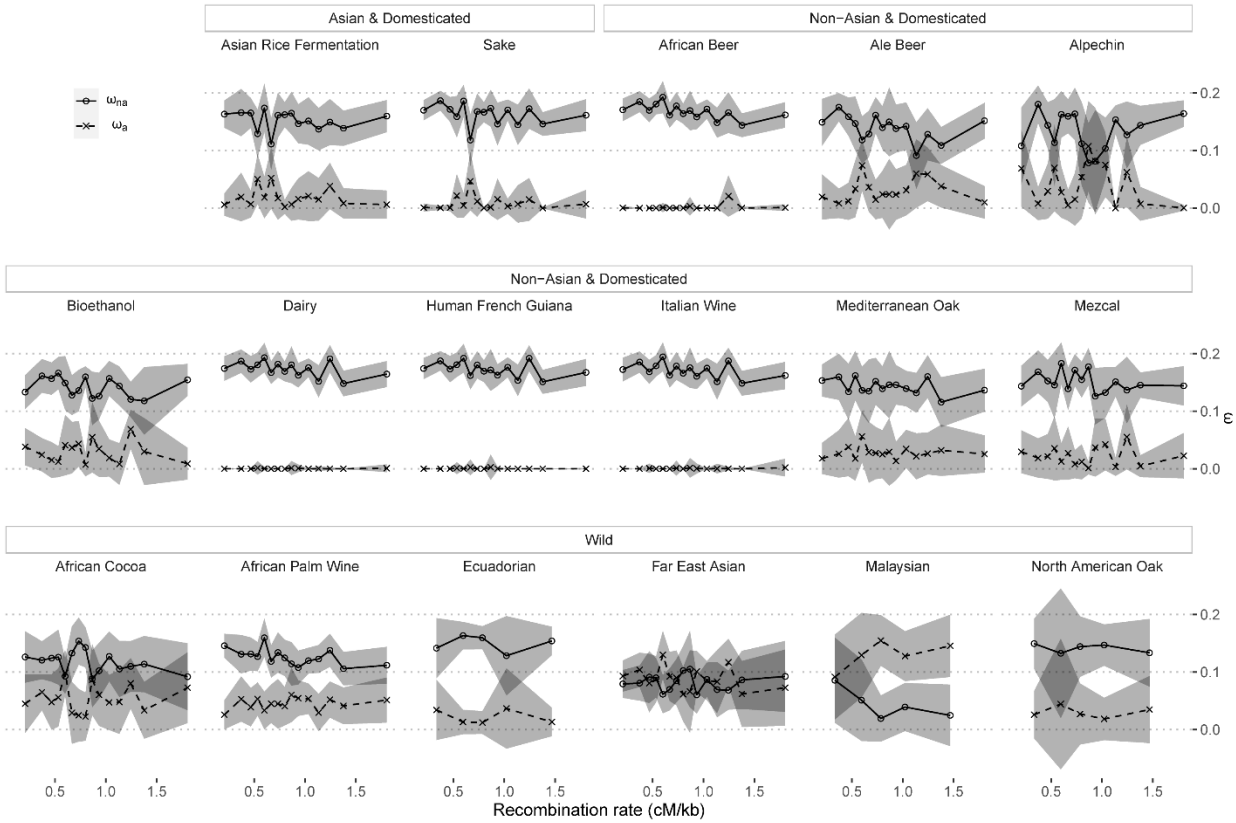

Figure 5: Rates of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions in regions with distinct recombination rates. Points and lines indicate the mean over a 150 bootstrap replicates, and the shaded area the corresponding 95% confidence interval (1.96 * standard deviation).

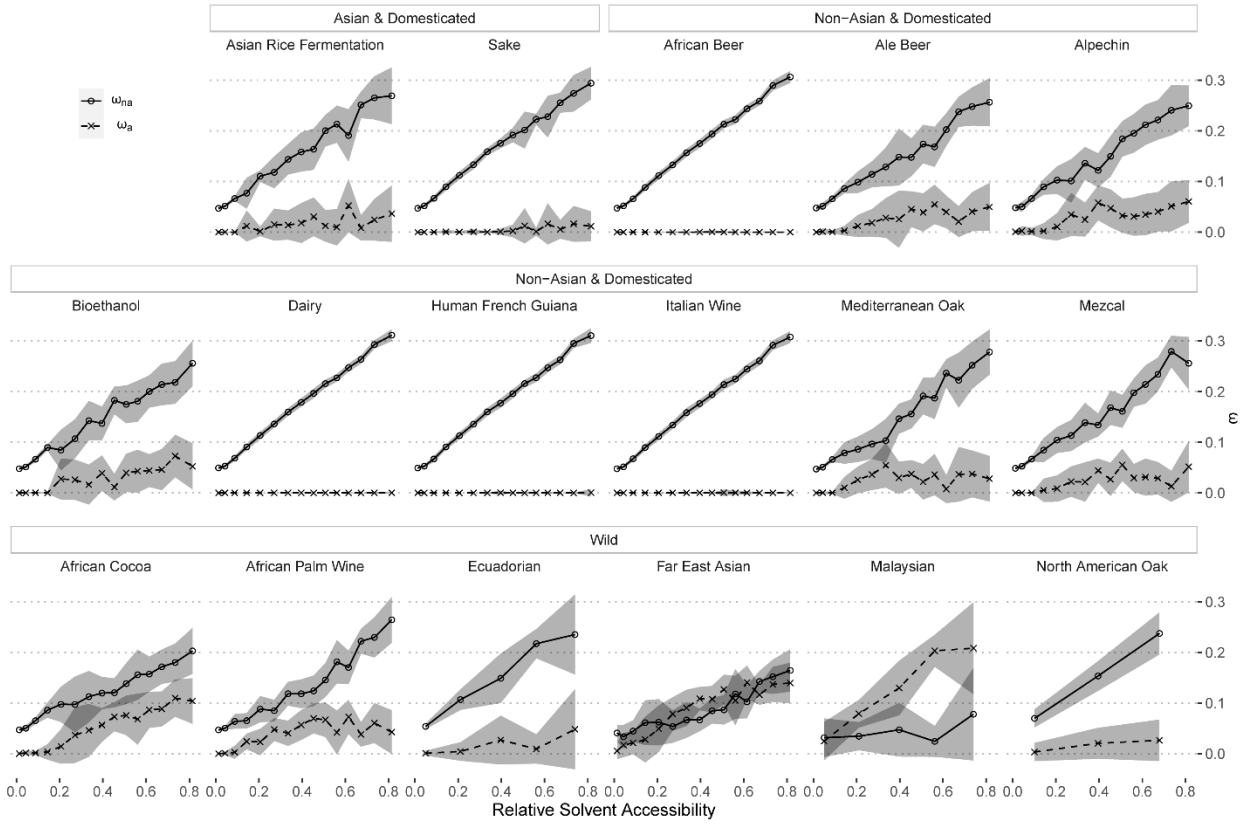

Figure 6: Rates of adaptive (ω_a) and non-adaptive (ω_{na}) non-synonymous substitutions in sites with distinct relative solvent accessibility (RSA). Points and lines indicate the mean over a 150 bootstrap replicates, and the shaded area the corresponding 95% confidence interval (1.96 * standard deviation).

Supplementary tables

Supplementary Table 1: List of all selected isolates, their designated population, assembly identifiers, zygosity, ploidy and aneuploidy status.

Supplementary Table 2: Length of nucleotide alignment for each population after whole genome alignment and after extraction of CDS regions, and the total number of genes recovered from each population.

Supplementary Table 3: Effect of life history, culture method and effective population size (as estimated by π_S) on the genome average rate of adaptive and non-adaptive non-synonymous substitutions when accounting for population-specific genes.

Supplementary figures

Supplementary Figure 1: Demography recovery from simulated data with yeast genome characteristics, under a neutral scenario and flat demography. The true demography (constant population size) is plotted in orange. The inferred demography from ten replicates is plotted with black lines. A) Demography inference using MSCM2's default time discretization. B) Demography inference using a 1*5+20*1+1*5 time discretization scheme. C) Demography inference using a 1*5+20*1+1*5 time discretization scheme and callability mask of the real Far East Asia population.

Supplementary Figure 2: UpSet plot showing the distribution and overlap of genes across different populations. Diagram summarising the number of genes for all unique intersection set. Corresponding intersection sets are represented with the dots connected by lines in the plot underneath the main bar plot, where filled-in dots signify a population is part of the set, and

where a non-included population is shown as a light grey dot with possibly the dark line spanning across it. The set size plotted on the left of the figure represents the total amount of genes recovered from each of the populations. This figure was created using the 'UpSetR' package v.1.4.0 (Conway, Lex, & Gehlenborg, 2017).

Supplementary Figure 3: Codon usage analysis with respect to the crossing-over rate. A) within-group correspondence analysis. Labels correspond to the 15 categories of sites with increasing recombination rate. B) Codon usage bias as a function of recombination rate class. Codon usage bias was measured as the mean relative synonymous codon usage (RSCU) – 1 (absolute value).

Supplementary Figure 4: Codon usage analysis with respect to relative solvent accessibility (RSA). A) within-group correspondence analysis. Labels correspond to the 15 categories of sites with increasing RSA. B) Codon usage bias as a function of RSA class. Codon usage bias was measured as the mean relative synonymous codon usage (RSCU) – 1 (absolute value).