MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 5412–5427 doi: 10.1111/mec.13341

A population genomics insight into the Mediterranean origins of wine yeast domestication

PEDRO ALMEIDA,* RAQUEL BARBOSA,* POLONA ZALAR,† YUMI IMANISHI,‡ KIMINORI SHIMIZU,§ BENEDETTA TURCHETTI,¶ JEAN-LUC LEGRAS,** MARTA SERRA,* SYLVIE DEQUIN,** ARNAUD COULOUX,†† JULIE GUY,†† DOUDA BENSASSON,‡‡ PAULA GONÇALVES* and JOSÉ PAULO SAMPAIO*

*UCIBIO@REQUIMTE, Departamento de Ciências da Vida, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal, †Department of Biology, Biotechnical Faculty, University of Ljubljana, Večna pot 111, SI-1000 Ljubljana, Slovenia, ‡Department of Applied Material and Life Science, College of Engineering, Kanto Gakuin University, Mutsuura-higashi 1-50-1, Kanazawa-ku, Yokohama 236-8501, Japan, §Medical Mycology Research Center, Chiba University, Inohana 1-8-1, Chuo-ku, Chiba 260-8673, Japan, ¶Dipartimento di Scienze Agrarie, Alimentari e Ambientali & Industrial Yeasts Collection DBVPG, Università degli Studi di Perugia, Borgo XX Giugno, 74 – 06121 Perugia, Italy, **Institut National de la Recherche Agronomique (INRA), UMR1083 Sciences pour l'Œnologie (SPO) 2, Place Viala 34060, Montpellier, France, ††CEA, Institut de Génomique, Genoscope, Centre National de Séquençage, 2 rue Gaston Crémieux, CP5706 91057 Evry Cedex, France, ‡‡Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK*

Abstract

The domestication of the wine yeast *Saccharomyces cerevisiae* is thought to be contemporary with the development and expansion of viticulture along the Mediterranean basin. Until now, the unavailability of wild lineages prevented the identification of the closest wild relatives of wine yeasts. Here, we enlarge the collection of natural lineages and employ whole-genome data of oak-associated wild isolates to study a balanced number of anthropic and natural *S. cerevisiae* strains. We identified industrial variants and new geographically delimited populations, including a novel Mediterranean oak population. This population is the closest relative of the wine lineage as shown by a weak population structure and further supported by genomewide population analyses. A coalescent model considering partial isolation with asymmetrical migration, mostly from the wild group into the Wine group, and population growth, was found to be best supported by the data. Importantly, divergence time estimates between the two populations agree with historical evidence for winemaking. We show that three horizontally transmitted regions, previously described to contain genes relevant to wine fermentation, are present in the Wine group but not in the Mediterranean oak group. This represents a major discontinuity between the two populations and is likely to denote a domestication fingerprint in wine yeasts. Taken together, these results indicate that Mediterranean oaks harbour the wild genetic stock of domesticated wine yeasts.

Keywords: comparative genomics, domestication fingerprints, microbe domestication, microbe population genomics, yeast molecular ecology

Received 3 January 2015; revision received 24 July 2015; accepted 29 July 2015

Introduction

The production of fermented beverages and foods by humans is contemporary with the onset and expansion of agriculture, with the consequent accumulation of foodstuffs and the need to avoid their deterioration

Correspondence: José Paulo Sampaio, Fax: 351 21 2948530; E-mail: jss@fct.unl.pt

(McGovern 2009). The presence of ethanol and many other metabolites contributed to preserve dietary goods, enhanced their palatability and digestibility and opened the way for the production of a myriad of alcoholic beverages that became important in social habits of many civilizations (Joffe 1998; McGovern 2009). Archaeological and biomolecular evidence indicates that beverages reminiscent of rice wine were produced as far back as 7000 BC in China (McGovern *et al.* 2004), while the forebear of modern beer was consumed in 6000 BC in Sumeria (Michel *et al.* 1992; Hornsey 2003). In Iran's northern Zagros mountains, chemical evidence of wine was dated 5400–5000 BC (McGovern *et al.* 1996) and a contemporary pile of grape pips with skins was found in the Neolithic site of Dikili Tash in northern Greece, thus suggesting that grapes, their juice and probably wine were already present in the Mediterranean region at that time (Valamoti *et al.* 2007). Furthermore, yeast cells were detected in Egyptian leavened bread dough dated 3000 BC (Samuel 1996, 2000) and molecular evidence for the presence of *Saccharomyces cerevisiae* in wine fermentation has been obtained from pottery jars of the same period (Cavalieri *et al.* 2003). From its probable origins in the Near East, viticulture and viniculture gradually disseminated in Europe (Hornsey 2007). Grapevine expansion occurred between 800 and 400 BC, progressed east to west across the Mediterranean sea and involved Phoenicians, Etruscans and Celts (McGovern *et al.* 2013).

Although the study of microbe domestication is incomparably less developed than that of crop and livestock domestication, the understanding of the genetic and functional underpinnings of human-driven microbe selection over millennia are attracting the interest of an increasing number of researchers (Legras *et al.* 2007; Liti *et al.* 2009; Douglas & Klaenhammer 2010; Gibbons *et al.* 2012; Almeida *et al.* 2014; Cheeseman *et al.* 2014). A classic example of the dramatic consequences of domestication in microbes is the recently fully elucidated hybridization of two *Saccharomyces* species that gave rise to the allotetraploid lager-beer yeasts (Libkind *et al.* 2011). Concerning the domestication of the yeast *S. cerevisiae*, a single domesticated lineage has been identified for wine within a complex global population structure, thus suggesting a single wine-related domestication event (Fay & Benavides 2005; Liti *et al.* 2009; Cromie *et al.* 2013). The wine group includes strains isolated from wine must, grapes and vineyard soil in different wine-producing regions around the world and, in spite of geographic heterogeneity, this group was preliminarily found to have less genetic diversity than other groups assumed to represent wild populations (Fay & Benavides 2005; Liti *et al.* 2009). Interestingly, a likely genetic fingerprint of domestication was identified in a large number of wine-associated yeasts, which consists in the presence in the respective genomes of up to three genomic regions (named A, B and C) acquired from other yeasts through horizontal gene transfer and containing a total of 39 genes potentially relevant for the winemaking process (Novo *et al.* 2009; Galeote *et al.* 2010, 2011; Marsit *et al.* 2015). This group of strains is often referred to in the literature as the 'Wine-European' group. Wild and domesticated lineages are likely to coexist in Europe and elsewhere. However, incomplete sampling of wild lineages prevented so far the identification of close wild relatives of the two currently recognized domesticated groups – wine and sake (Fay & Benavides 2005; Liti *et al.* 2009). In both cases, revealing close wild relatives or potential ancestors would be essential for a proper understanding of the genetic basis of domesticated phenotypes.

Different studies support the view that the oak niche (tree bark and soil underneath the trees) is a natural habitat of *S. cerevisiae* in temperate forests of the Northern Hemisphere (Naumov *et al.* 1998; Sniegowski *et al.* 2002; Sampaio & Gonçalves 2008; Wang *et al.* 2012; Hyma & Fay 2013), so that it constitutes a suitable niche in which to search for wild Holarctic *S. cerevisiae* populations. Here, we use whole-genome data from 145 strains and a combination of phylogenomics, population genomics, demographic models and genomic surveys of domestication fingerprints to investigate newly identified oak-associated lineages from Europe and Asia. In particular, we analyse in detail the relationship of the wine group with a presently uncovered oak-associated Mediterranean *S. cerevisiae* population which is its closest wild relative. We propose that this new population contains the wild genetic stock that gave rise to the domesticated wine yeasts.

Materials and methods

Strain isolation, identification and typing

Isolation of *Saccharomyces* yeasts was based on the selective enrichment protocol previously described (Sampaio & Gonçalves 2008). Putative *Saccharomyces* isolates were confirmed by the observation of *Saccharomyces*-type ascospores, and species identifications were performed by sequencing of the ITS and D1/D2 regions of the rDNA.

Microsatellite analysis

Eighty-seven *S. cerevisiae* strains were characterized for their allelic variation at 12 microsatellite loci (Legras *et al.* 2007) and compared to 144 reference genotypes as previously described (Legras *et al.* 2007). The Bruvo dis-

tance among strains was calculated using the package POPPR 1.1.2 (Kamvar *et al.* 2014) of the R statistical environment (R Core Team 2013). For the few aneuploid or tetraploid stains (i.e. bread isolates), two of the scored alleles were chosen randomly per locus (27 strains of 231 had more than two values in at least one locus). As different trees obtained from different genotype data had the same global topology (except inside the clusters containing the bread strains), this step had little impact on the global tree topology. A network was drawn with SPLITSTREE v4.13.1 (Huson & Bryant 2006) using the neighbour-net method (Fig. S1, Supporting information).

Genome sequencing, read alignment and genotype calling

Paired-end or single-end genomic Illumina reads were obtained for a subset of 90 representative new isolates obtained in this study (or their monosporic derivatives) using the Illumina HiSeq2000 system. Genomic data for other isolates were obtained from the NCBI-SRA collection and from *Saccharomyces* Genome Resequencing Project v2 (SGRP2) (Bergström *et al.* 2014) (Table S1, Supporting information). When only finished genome sequences were available in public databases (NCBI), the corresponding error-free Illumina reads were simulated using dwgsim (http://sourceforge.net/apps/mediawiki/dnaa/).

Reads for each isolate were mapped to *Saccharomyces cerevisiae* reference genome (UCSC version sacCer3) using SMALT v0.7.5 aligner (http://www.sanger.ac.uk/resources/software/smalt/). The reference index was built with a word length of 13 and a sampling step size of 2 (−k 13 −s 2). An exhaustive search for alignments (−x) was performed during the mapping step with the random assignment of ambiguous alignments switched off (−r −1) and the base quality threshold for the look-up of the hash index set to 10 (−q 10). With these settings, SMALT v0.7.5 only reports the best unique gapped alignment for each read. Whenever paired-end information was available, the insert size distribution was inferred with the 'sample' command of SMALT prior to mapping. Conversion of SAM format to BAM, sorting, indexing, several mapping statistics and consensus genotype calling were performed using the tools available in the SAMtools package v1.18 (Li *et al.* 2009) as described previously (Almeida *et al.* 2014). The genotype of *S. paradoxus* CBS 432 (Liti *et al.* 2009) was determined using the same approach as above starting with simulated reads. Multiple sequence alignments for each reference chromosome were generated from the resulting fasta files. For downstream analyses, all bases with Phred quality score below Q40 (equivalent to a 99.99% base call accuracy) or ambiguous base calls were converted to an 'N'.

Network, phylogeny and population structure

Chromosomal single nucleotide polymorphisms (SNPs) were extracted from multiple sequence alignments only if the evaluated site was represented by unambiguous high-confidence alleles in all isolates. SNPs were then concatenated to generate a whole-genome SNP alignment.

SplitsTree v4.12.6 (Huson & Bryant 2006) was used to reconstruct a neighbour-net phylogenetic network for *S. cerevisiae* using the Kimura 2-parameter model. Rooted maximum-likelihood phylogenies were estimated using the rapid bootstrap algorithm as implemented in RAxML v7.3.5 (Stamatakis 2006) with GTRGAMMA model of sequence evolution. RAxML was run for 10 times with 1000 rapid bootstraps (100 for the largest data set), and the tree with the highest log likelihood was chosen to represent the most likely phylogenetic reconstruction. Bootstraps from all runs were then combined into this best maximum-likelihood tree. *Saccharomyces paradoxus* was used as outgroup. Population structure of *S. cerevisiae* was explored using the model-based Bayesian clustering method implemented in STRUCTURE v2.3.4 (Pritchard *et al.* 2000; Falush *et al.* 2003) and the chromosome painting algorithm as implemented in fineSTRUCTURE v2.0.2 (Lawson *et al.* 2012). STRUCTURE was run with a subset of 9181 equally spaced parsimony informative sites (mean distance between sites of approximately 1250 bp). The number of Markov chain Monte Carlo (MCMC) iterations was set to an initial burn-in period of 100 000 iterations, followed by 50 000 iterations of sampling. The ancestry model allowed for admixture and allele frequencies were assumed to be correlated among populations. Ten independent simulations were run for each value of K, varying from $K = 1$ to $K = 12$, and stability was assessed by monitoring the standard deviation between simulations. The run with the highest estimated log probability of the data was chosen to represent each value of K. fineSTRUCTURE was run in linked mode using 98 strains with 94 089 informative biallelic SNPs. Strains with identical genotypes or without informative sites were iteratively excluded to allow numerical stability of the expectation-maximization (EM) algorithm. Although several strains from the North American population had to be removed, all groups were represented in this analysis. Strain EXF 7145, which initially presented 31 957 heterozygous sites, was phased with the SAMtools phase (Li *et al.* 2009) command prior to analysis, resolving more than 98% of the heterozygosities in two haplotypes. The genomic profile of each strain was

obtained by copying from every other strain. The 'recombination scaling constant' and the 'per site mutation rate' were set to 35 and 0.0032, respectively, as estimated by 100 iterations of the EM algorithm. The number of 'chunks' per region was set to 50, and all other parameters were left at the default values. The 'c' value inferred by fineSTRUCTURE was 0.41. The recombination map for each chromosome was obtained from 'http://www.yeastgenome.org/pgMaps/pgMap.shtml' as Morgans/bp.

Polymorphism and divergence analyses

Whole-genome levels of polymorphism were estimated using Variscan v2.0 (Hutter *et al.* 2006). For polymorphism analyses within populations, we made use of an additional set of isolates (six from Portugal and fourteen from Japan), chosen randomly from our initial survey. This additional population sampling was found to have identical or very similar genotypes to the representative set but allowed a more detailed description of the estimated population diversity within these regions. To allow for missing data, only sites with valid high-quality alleles (>Q40) in at least 75% of ingroup sequences were used in calculations (defined with the NumNuc parameter together with CompleteDeletion = 0 and FixNum = 0). Divergence estimates between populations were calculated as the mean pairwise divergence between samples from two populations, π_B, using software based on the libsequence library (http://molpopgen.org/) (Thornton 2003). Only sites with valid high-quality alleles in at least 75% of sequences for each population were used in calculations.

Coding and noncoding sequences were extracted from chromosome alignments based on the annotation of *S. cerevisiae* reference genome available at *Saccharomyces* Genome Database (SGD release R64-1-1 of 2011-02-03, same as UCSC sacCer3). Sequences with more than 10% of missing bases in each alignment were excluded from the analyses. After this step, only alignments with more than four sequences were used for calculations. Nucleotide diversity of fourfold degenerate and nonsynonymous sites were calculated with the analysis package from libsequence (http://molpopgen.org/) (Thornton 2003).

Statistics of shared polymorphisms and fixed differences were calculated with the analysis package from libsequence (http://molpopgen.org/) (Thornton 2003). For all comparisons, only positions with at least 75% of valid sites in both populations being compared and excluding singletons were used in calculations.

For these analyses, whenever the wine group was compared, it was represented only by strains isolated from wine environments (i.e. commercial and wine must strains but not vineyard strains). Strains ZP 530, ZP 1050 and UWOPS 83-787.3 were excluded from the North America group because they were isolated in regions outside North America.

Demographic analyses

Two-population demographic inference was performed for the wine population (commercial wine strains and strains isolated from must) and a random sample of 20 strains from the extended Mediterranean oak population using a subset of noncoding regions across the whole genome. These regions were chosen based on the annotation of *S. cerevisiae* reference genome available at *Saccharomyces* Genome Database (SGD release R64-1-1 of 2011-02-03, same as UCSC sacCer3) and have to meet the following criteria: noncoding regions should be separated from each other by at least 3 kb, which approximately corresponds to the decay of linkage disequilibrium to half of its maximum (Liti *et al.* 2009), and should be more than 500 bp long in a tentative to avoid shorter intergenic sequences that can be potentially enriched for regulatory elements. This process resulted in 1247 noncoding regions, totalling 1 286 807 bp length, with 15 857 observed SNPs. The folded joint allele frequency spectrum was calculated from both populations and fitted to different isolation scenarios using a diffusion-based approach as implemented in the program ∂a∂i (Gutenkunst *et al.* 2009). To account for missing data, the allele frequency spectrum for each population was projected down in ∂a∂i to the projection that maximized the number of segregating SNPs, resulting in 13 921 SNPs. Each model was run five times from independent starting values to ensure convergence to the same parameter estimates. The maximum-likelihood estimates of the best-fit demographic model were used to generate 95% confidence intervals from 100 simulated data sets in ms (Hudson 2002). Estimation of the ancestral population size was corrected using an effective sequenced length, as suggested in the study by Gutenkunst *et al.* (2009), of 1 129 699 bp [calculated as 1 286 807*(13 921/15 857)].

De novo assemblies and survey for the horizontally transferred regions A, B and C

For the survey of regions A, B and C, which were horizontally transferred to several *S. cerevisiae* wine strains (Novo *et al.* 2009; Marsit *et al.* 2015) but are not present in the reference genome, we performed *de novo* genome assemblies of the Illumina reads for most of the strains included in this study (and for which there was no genome yet available), using VELVET v.1.2.08 (Zerbino & Birney 2008). Prior to assembly, reads were processed

with Sickle (https://github.com/najoshi/sickle/), based on a quality score threshold of 20 for windowed trimming, discarding reads with length <40 or with any 'Ns' on them. Velvet was run with different kmer values and with coverage mask set to 10×. The final kmer assembly was chosen considering the relationship between the number of final contigs and 'misjoin' errors with the proportion of missing reference bases, as evaluated by GAGE statistics (Salzberg *et al.* 2012).

We set up local BLAST databases for all the genomes available in this study and performed BLASTN searches (1e-4 *E*-value cut-off) using the coding sequences present in each one of the three regions of interest as queries. Blast hits were retained if sequence identity was above 90% and sequence aligned to at least 10% of the query. For the strains where genome assemblies were not available, reads were mapped to a combined reference built with the coding sequences of the three regions using BWA v0.6.2 (Li & Durbin 2009) with default parameters but setting the quality threshold to 10 (-q 10). SAMtools v1.18 (Li *et al.* 2009) was used for the manipulation of the resulting BAM files, following the same approach as described above. Genes showing more than 90% of Q40 bases in more than 10% of the total length were scored as being present in the interrogated strain.

Multi-locus sequence analysis

Thirteen loci previously used to characterize Chinese isolates (Wang *et al.* 2012) were retrieved from the available *de novo* genome sequences using BLASTN (see above) and aligned with FSA v1.15 (Bradley *et al.* 2009). After alignment, loci were concatenated and sequences with <80% of the total length were removed. The phylogenetic history was inferred from the concatenated alignment using the neighbour-joining method in MEGA 5 (Tamura *et al.* 2011). Evolutionary distances were computed with the Kimura 2-parameter model of sequence evolution and are in units of the number of base substitutions per site. All positions with <95% site coverage were eliminated, that is fewer than 5% alignment gaps, missing data and ambiguous bases were allowed at any position. There were a total of 12 680 positions in the final data set. Branch support was estimated from 1000 nonparametric bootstrap replicates.

Results

A geographically and genetically diverse collection of wild isolates

Between 2005 and 2012, we enlarged the number of natural *S. cerevisiae* strains available for study by isolating approximately 120 strains from the oak niche in different regions of the Mediterranean basin (Iberian Peninsula, France, Italy, Slovenia, Greece) and Japan. A preliminary characterization through microsatellite genotyping suggested a separation between isolates obtained in natural and anthropic habitats (Fig. S1, Supporting information). It also brought to light a clear separation of the wild Mediterranean population from oak-associated populations of other geographic origins, thus supporting the view that geography and ecology, rather than ecology alone, contribute to shape the global population structure of *S. cerevisiae*.

Phylogenomic studies performed to date in *S. cerevisiae* have had a strong bias towards anthropic environments (Liti *et al.* 2009; Strope *et al.* 2015), which hinders a good understanding of the relationship between the various lineages. Hence, we carried out a population genomic analysis using a comprehensive and more balanced strain data set that included the novel natural isolates (Table S1, Supporting information). In view of the strong influence of admixture in the genome structure of many *S. cerevisiae* strains (Liti *et al.* 2009; Cromie *et al.* 2013; Strope *et al.* 2015) and its likely negative interference on the reconstruction of population history, we chose to summarize the relationships of the 146 strains studied using a genomewide phylogenetic network (Fig. 1). Most strains were positioned close to the two horizontal extremities of the network, one of which was dominated by strains isolated from wine fermentations or vineyards, as well as most strains isolated from Mediterranean oaks (Table S1, Supporting information). Strains used for bioethanol production and for beer fermentations were placed at the vicinity of the wine-European group. The other extremity of the network was occupied by a complex group of oak isolates, mainly from North America and Japan, but including also six European oak isolates and a diverse group of strains, including natural isolates from Malaysia and the Philippines and from regional fermented beverages in Africa, Japan and the Caribbean. Our split network confirmed the preliminary microsatellite data and showed that the oak-associated strains could be resolved partially by geography, with a clear separation of the Mediterranean population from a complex group that included the North American and Japanese strains.

Population structure and admixture

To delimit populations and infer possible admixture events, we used STRUCTURE (Falush *et al.* 2003) and tested from 2 to 12 ancestral (K) clusters. The addition of more sequence diversity resulted in a larger number of genetic clusters, a tendency already observed (Liti *et al.* 2009; Schacherer *et al.* 2009; Wang *et al.* 2012; Cromie

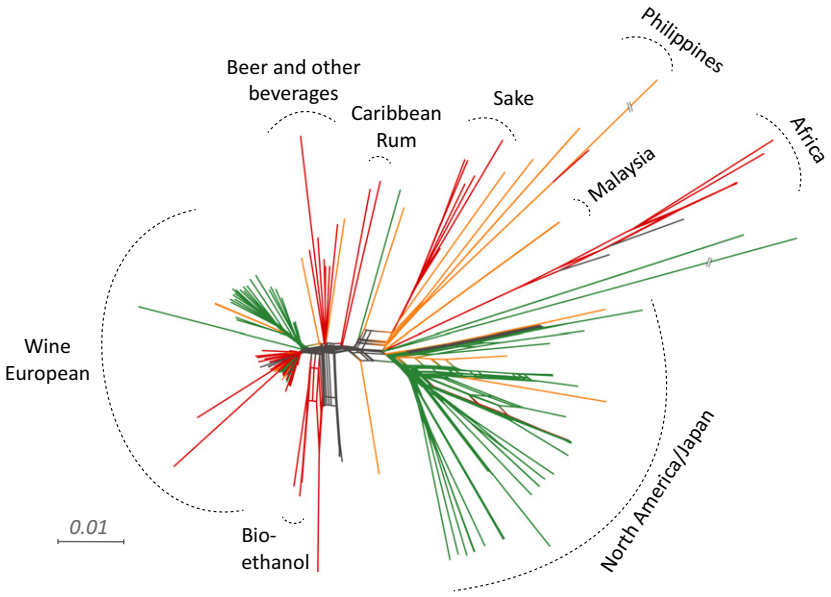

Fig. 1 The global diversity of *Saccharomyces cerevisiae* is shaped by both ecology and geography. Neighbour-net network of 146 strains based on 60 331 SNPs, inferred with Kimura 2-parameter distance. Branches are coloured according to the substrate of isolation of the strain (red – fermentation; green – oak tree; orange – fruit; grey – other or unknown). The scale bar represents the number of substitutions per site.

et al. 2013). The *ad hoc* statistic ΔK (Evanno *et al.* 2005) returned an optimum number of two clusters, but increasing values of *K* resulted in increased information about the actual population structure. The more comprehensive representation of sequence ancestry was achieved with $K = 10$ (Fig. 2a) as analyses using higher *K* values did not reveal new meaningful clusters. The main clusters were assigned to either industrial variants or geographically delimited populations such as Wine (1), Mediterranean oak (2), Sake (3), Philippines (4) Africa (5) North American and/or Japanese populations (6–9). Our analysis revealed a considerable number of strains with admixed genotypes (about 46% at $K = 10$), here arbitrarily defined as <90% ancestry from a single population cluster. This extensive pattern of admixture or 'mosaicism', which has been attributed to anthropic influences (Liti *et al.* 2009), might also have natural causes as several wild North American and Japanese isolates from oak had mixed ancestries that are not likely to be a consequence of human intervention. The Mediterranean oak population was difficult to distinguish from the Wine group, as a separate cluster exclusively associated with Mediterranean oaks was only formed at $K = 9$ or higher. Moreover, even with that number of clusters, several of the Mediterranean oak isolates had a partial ancestry with the wine cluster (Fig. 2a).

We additionally performed the clustering of strains using a haplotype-based approach as implemented in fineSTRUCTURE v2 (Lawson *et al.* 2012). This method differs from that of STRUCTURE because it specifically models the patterns of linkage disequilibrium across blocks of available positions to infer the genealogical information about the local ancestry of an individual. A chromosome 'painting' process is then used to partition the individuals into genetically homogeneous clusters. fineSTRUCTURE resolved the strains in a hierarchical tree representing the relationships among the identified clusters based on a genomewide coancestry matrix. Overall, we observed a broad congruence with the results obtained with STRUCTURE. Industrial variants or geographically delimited populations previously identified, that is Wine, Mediterranean oak, North America – Japan, Sake and Africa were also recovered with fineSTRUCTURE, together with a complex group of mosaic wine strains (Fig. 2b and Fig. S2, Supporting information). Notably, fineSTRUCTURE hinted at a close relationship between wine and Mediterranean oak strains, indicating a high degree of shared haplotype ancestry between the two groups. As already partially captured by STRUCTURE, Mediterranean oak strains were divided into three subclusters and two of them (the two subclusters with less strains) had considerable coancestry, as recipient genotypes, with wine haplotype blocks. It is uncertain if the shared blocks correspond to direct contact between the two groups or to a secondary process that involves wine mosaic strains. Interestingly, the other Mediterranean oak cluster is the only cluster formed by natural isolates that apparently resulted from a stronger effect of drift over admixture, as observed in the coancestry matrix (Fig. 2b). It is also worth noting that the strains not grouped in clusters are probably the outcome of independent and complex admixture events and correspond mostly to the individuals that have been identified as mosaics in STRUCTURE.

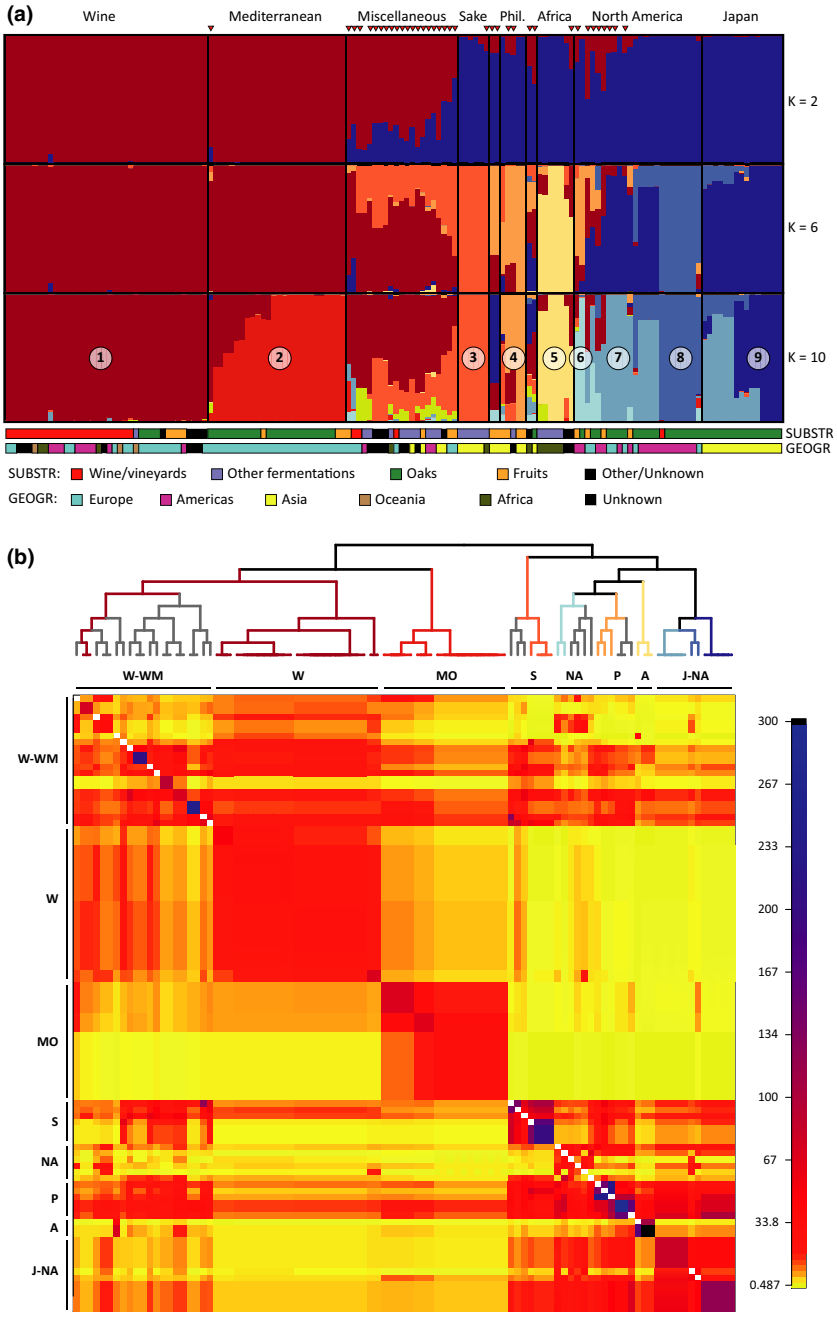

Fig. 2 Population structure and admixture in *Saccharomyces cerevisiae*. (a) STRUCTURE plots based on a subset of 9181 parsimony informative sites for $K = 2$, 6 and 10. Strains identified as mosaics at $K = 2$ are marked with a red triangle. Numbers 1–9 represent the different clusters that capture the maximum representation of population ancestry. The type of substrate of isolation (SUBSTR) and geographic source (GEOGR) are colour-coded. (b) fineSTRUCTURE co-ancestry matrix and population structure of 98 strains using 94 089 informative biallelic SNPs. The colour of each bin in the matrix indicates the expected number of 'chunks' copied from a donor (column) to a recipient strain (row). The dendrogram on the top represents the clustering of strains inferred from the co-ancestry matrix. Branches are coloured according to STRUCTURE clusters for easier comparison. Branches coloured in grey indicate mosaic strains identified by STRUCTURE at $K = 2$. W-WM: Wine and wine mosaics; W: Wine; MO: Mediterranean oak; S: Sake; NA: North America; P: Philippines; A: Africa; J-NA: Japan and North America.

Genetic relationships among populations

As admixed genotypes are likely to hinder the elucidation of the phylogenetic relationships between populations, we discarded all except the two beer and the two Malaysian representatives of the 37 mosaic genotypes detected with STRUCTURE at K = 2, as these strains are likely to represent distinct populations. The six main clades detected in the maximum-likelihood phylogeny of 112 genome sequences (Fig. 3) largely recapitulate the groups found in previous analyses (Liti *et al.* 2009; Schacherer *et al.* 2009; Cromie *et al.* 2013), but with considerably increased genomic diversity, hinting at novel phylogeographic and evolutionary relationships. The clade previously designated 'Wine-European' includes in the present analysis, with high statistical support, the new Mediterranean oak isolates most of which are placed together in a subcluster that does not include wine strains. The upper part of the Wine-European clade groups all wine strains (commercial strains, strains from spontaneous fermentations and from vineyards) and a minority of strains isolated from the oak

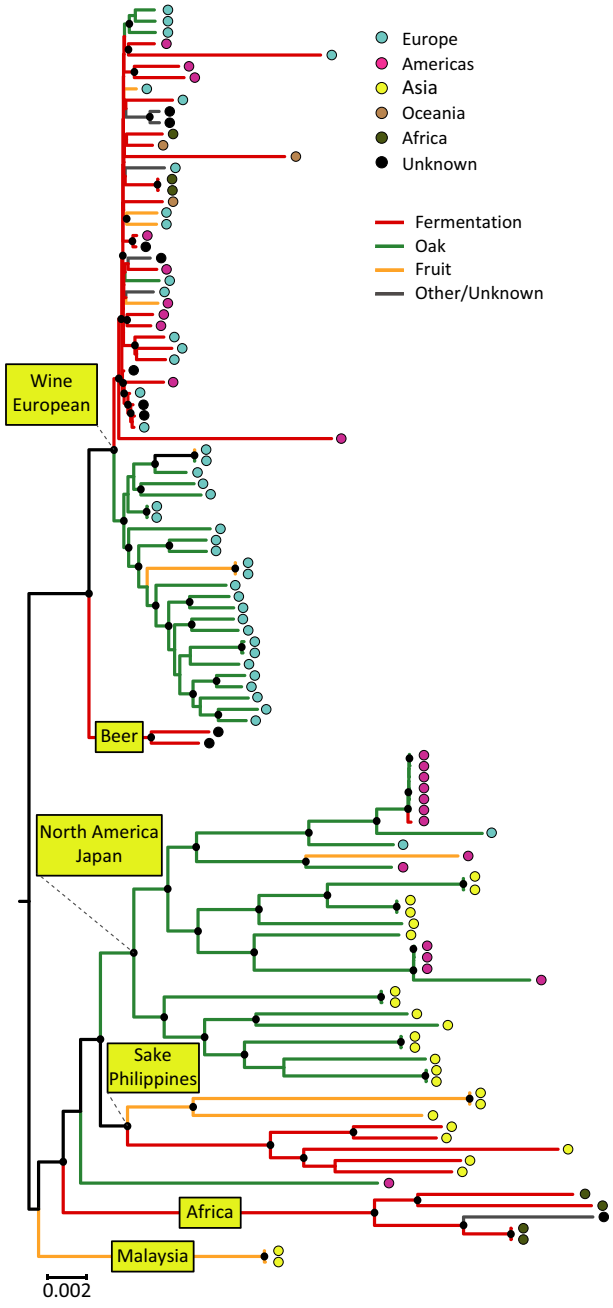

Fig. 3 Whole-genome phylogenetic relationships between *Saccharomyces cerevisiae* strains. Rooted phylogenetic tree of 112 strains that excludes the mosaic strains identified by STRUCTURE at K = 2. The tree was inferred from 193 071 SNPs, using the maximum-likelihood method as implemented in RAxML with the GTRGAMMA model of sequence evolution and was rooted with *S. paradoxus*. Branches are coloured according to the substrate of isolation. Strains are represented by coloured dots indicating the geographic origin. Branch lengths correspond to the expected number of substitutions per site. Support values from bootstrap replicates above 90% are depicted with black dots in the respective tree nodes.

system and from fruits collected in semi-wild systems (e.g. figs, wild apples) (Fig. 3). The 'North America – Japan' clade contains exclusively wild isolates and is sister to a clade formed by strains from sake and similar Asian fermented products and strains from fruits and fermented beverages in the Philippines. The remaining two clades are the previously recognized Malaysian and African lineages. A phylogeny including the complete data set is depicted in Fig. S3 (Supporting information) and shows that most mosaic strains are positioned outside the main clades mentioned above, an expected result given their recombinant nature. To determine the phylogenetic relationships of the new oak-associated lineages with the genetically highly diverse populations from China (Wang *et al.* 2012), for which whole-genome data is not available, we used the published multilocus sequences of those lineages to construct an integrated phylogeny (Fig. S4, Supporting information). This analysis revealed that four of the eight Chinese clades are closely related but not coincident with the groups detected in the present study, whereas the remaining and most divergent Chinese lineages are external to our strain data set. It also highlights that most of the *S. cerevisiae* genetic diversity that we have sampled is contained in the Asian lineages.

The consistent placement of our isolates from Mediterranean oaks as the closest relatives of the wine lineage is noteworthy. Such a strong relationship between the two groups is also supported by the low nucleotide divergence that we observed between them. In fact, whereas these two groups have a divergence of 0.206%, the Wine group is approximately three times more divergent from both the North American and Japanese wild groups (Table 1). Moreover, we observed that the Wine and Mediterranean oak groups exhibited the highest number of shared polymorphisms and the lowest number of fixed differences in pairwise comparisons of the various populations with the wine clade (Table 2). We consider it likely that this was due to retention of shared polymorphisms from a recent common ancestor, rather than to pervasive gene flow between the two groups because the removal of the Mediterranean oak strains presenting mixed ancestry with the wine clade (see Fig. 2a for $K = 10$) did not lower the amount of shared polymorphisms to levels equivalent to those of other comparisons included in Table 2. Additionally, nucleotide divergence, measured in nonoverlapping windows of 20 kb, between wine and Mediterranean oak strains ($\pi_B \times 100 = 0.207$) was significantly lower than the divergence between Japan and North America strains ($\pi_B \times 100 = 0.291$) (2-sided *t*-test, *P*-value < 0.01) indicating a more recent population split for the former populations.

The diversity of domesticated wine yeasts is equivalent to that of wild Mediterranean oak yeasts

We found a higher ratio of polymorphism at nonsynonymous sites relative to fourfold degenerate sites in wine strains and a significant reduction in the population recombination parameter ($t = 2.186$, $P < 0.05$ (2-tailed), Table S2, Supporting information), both of which are consistent with the effects of a domestication bottleneck. However, nucleotide diversity based on pairwise differences (π*100) measured in the wine group and in the Mediterranean oak group are very similar (Table 3, Table S2, Supporting information). Moreover, a comparison of diversity measurements between different wild lineages brings to light a strikingly lower diversity for the Mediterranean oak group. In fact, the genetic diversities of the North America – Japan lineage or of the more geographically restricted Asian population are all higher than that of the Mediterranean oak population (Table 3).

Population demographics

To further test our hypothesis that the Mediterranean oak population is the best approximation for the wild ancestors of domesticated wine strains, we considered different demographic models based on the folded joint site frequency spectra of Wine and Mediterranean oak populations using a diffusion-based approach implemented in ∂a∂i (Gutenkunst *et al.* 2009). These analyses were performed using a subset of noncoding regions across the whole genome to minimize effects of selection that could interfere with demographic inference. First, we tested a model of complete isolation without migration that yielded correlated residuals between the model and the data, with the model predicting too few shared polymorphisms at low frequencies in the Wine population (Fig. 4b). This effect has been shown to result from fitting data having a migration signal in a no-migration model (Gutenkunst *et al.* 2009). Next, we tested the alternative scenario of asymmetrical migration between populations that explained much of the shared variation (Fig. 4c). However, with this simple model of isolation with asymmetrical migration, it was still possible to observe a strong deficit of medium frequency polymorphisms and an excess of singletons in both populations. Considering neutrality for the analysed noncoding loci, these molecular signatures are usually suggestive of recent population growth, in line with the genomewide negative Tajima's D values estimated for Wine and Mediterranean oak populations (Table 3). Therefore, we tested a third demographic model that included population growth after the split from an ancestral population, together with asymmetric migration between populations (Fig. 4d). This population growth model fitted better most of the private polymorphisms in both populations, although we note that not all features of the frequency spectrum have been fully captured. Among the three tested models, that of isolation with asymmetric migration and population growth had a higher maximum-likelihood value and a lower AIC (Akaike information criterion), indicating an increase in the likelihood of this model (Fig. 4 and Table S3, Supporting information).

Table 1 Pairwise whole-genome divergence between populations of *Saccharomyces cerevisiae*. Mean pairwise divergence between 2 alleles drawn from 2 populations (π_B) are estimated per site from pairwise comparisons across the total length of the genome. MO – Mediterranean oak

	Divergence ($\pi_B \times 100$)				
	MO	North America	Japan	Sake	Africa
Wine	0.206	0.614	0.613	0.680	0.688
MO		0.571	0.570	0.641	0.644
North America			0.291	0.422	0.503
Japan				0.414	0.499
Sake					0.568

Table 2 Proportion of shared polymorphisms, fixed differences and private polymorphisms in pairwise comparisons of the Wine group with other populations of *Saccharomyces cerevisiae*. All the comparisons were made to a subset of strains from the Wine group comprising commercial strains or isolates from must (vineyard strains were excluded). Values are given in percentages relative to the total number of SNPs in each comparison. MO – Mediterranean oak; NA – North America; Philip. – Philippines

Population	Shared polymorphisms	Fixed differences	Private in Wine	Private in pop.	Total SNPs
MO	8.04	4.90	41.79	45.27	63 345
MO*	6.15	7.74	50.47	35.64	54 908
NA – Japan	4.41	14.04	18.49	63.06	143 286
Japan	2.64	21.61	22.73	53.02	128 192
North America	1.38	39.25	30.77	28.60	99 596
Sake – Philip.	2.62	25.93	24.73	46.71	118 201
Africa	2.17	43.43	30.31	24.09	97 403

*Strains with mixed ancestry with the wine clade were excluded.

Table 3 Whole-genome diversity within populations of *Saccharomyces cerevisiae*. Diversity values (π, the average pairwise nucleotide differences between strains, and θ_W, the Watterson estimator for the number of segregating sites) are per site estimates calculated for the total length of the genome (the number of analysed sites)

	N° strains	Analysed sites	Segregating sites	π	θ_W	Tajima's D
Wine	19	11 216 288	56 367	0.001116	0.001447	−0.973222
Mediterranean oak	31	11 286 153	56 053	0.000991	0.001250	−0.810951
Wine and Mediterranean oak	50	11 216 436	109 670	0.001532	0.002193	−1.108871
North America and Japan	42	11 348 218	119 184	0.002560	0.002448	0.171544
Japan	29	11 373 293	90 824	0.002349	0.002038	0.602890

The best-fit demographic model was used to obtain converging estimates of demographic parameters (Table 4). Estimated population migration rates were relatively low, but a much higher migration from the Mediterranean oak population into the wine population than in the opposite direction was detected ($M_{MO \rightarrow W} \simeq 0.36$ and $M_{W \rightarrow MO} \simeq 0.02$, Table 4). Interestingly, while the current effective size of the Wine population ($N_e \simeq 5.8 \times 10^6$) (Table 4) was estimated to be higher than that of the Mediterranean oak population ($N_e \simeq 4.3 \times 10^6$), *∂a∂*i inferred a strong bottleneck in the Wine population at the time of the split ($s \simeq 0.22$) as predicted by the classical domestication model. Divergence time was estimated to about 3.8×10^6 generations into the past. Using a known mutation rate for *S. cerevisiae* (Lynch *et al.* 2008) and two generation times, ranging from eight to one generations per day (Fay & Benavides 2005; Liti *et al.* 2006), the split between the two populations could be dated between 1300 years ago (ya) and 10 300 ya, respectively. This divergence time is relatively recent, in agreement with the low number of fixed differences between the two populations (4.9%, Table 2) and with historical evidence for winemaking. The estimated time for the most recent common ancestor of wine and Mediterranean oak strains is compatible with the first biochemical evidence of wine, dated to 5400–5000 BC (McGovern *et al.* 1996).

Domestication fingerprints

We surveyed a set of genetic fingerprints related to wine fermentation in strains from the Wine and Mediterranean oak clades by searching for the presence of three genome portions acquired by horizontal gene transfer (HGT). These regions were designated A, B and C and were found previously to be widespread in wine strains (Novo *et al.* 2009; Galeote *et al.* 2011; Marsit *et al.* 2015). Region A is a 38-kb-long subtelomeric insertion of unknown origin in the left arm of chromosome VI; region B corresponds to a 17-kb insertion into chromosome XIV and was acquired from *Zygosaccharomyces baillii*; and region C, originating from *Torulaspora microellipsoides*, is subtelomeric, is 65 kb long and is located in the right arm of chromosome XV. At least one of these regions was present in 31 of the 40 genomes analysed belonging to the wine lineage, a frequency of 78%, whereas they were completely absent in the 25 genomes of the Mediterranean oak lineage that were surveyed (Fig. S3, Supporting information). In the other lineages, these regions were only rarely found (14% of the strains had at least one of these regions) and most (90%) were associated with mosaic strains. This indicates that the acquisition of regions A, B and C is probably related with the early stages of wine domestication, thus explaining their presence in most domesticated strains and absence in their closest wild relatives. This implies that these regions conferred a selective advantage for winemaking and therefore the strains that harboured them become dominant. The presence of these regions in admixed strains, including a few mosaics isolated from Mediterranean oaks, indicates that genes linked to domestication can in principle be transferred into the wild. Interestingly, six of the eight strains isolated from oaks and fruits that cluster within the wine clade (Fig. S3, Supporting information) also exhibit these regions, which suggests that they are wine strains that have secondarily colonized the oak environment and can therefore be viewed as feral yeasts.

Discussion

Here, we employ genomic data on oak-associated wild isolates of *S. cerevisiae* collected in the Mediterranean region and also in North America and Japan to generate a data set with a balanced number of anthropic and natural strains. We demonstrate that both the North American and Japanese wild populations are polyphyletic. This is consistent with the high genetic diversity previously found in Asia (Wang *et al.* 2012), with China being the most likely radiation centre of the oak-associated *Saccharomyces* lineages (Bing *et al.* 2014). We hypothesize that colonization of North America followed a migration route over the Bering Strait land

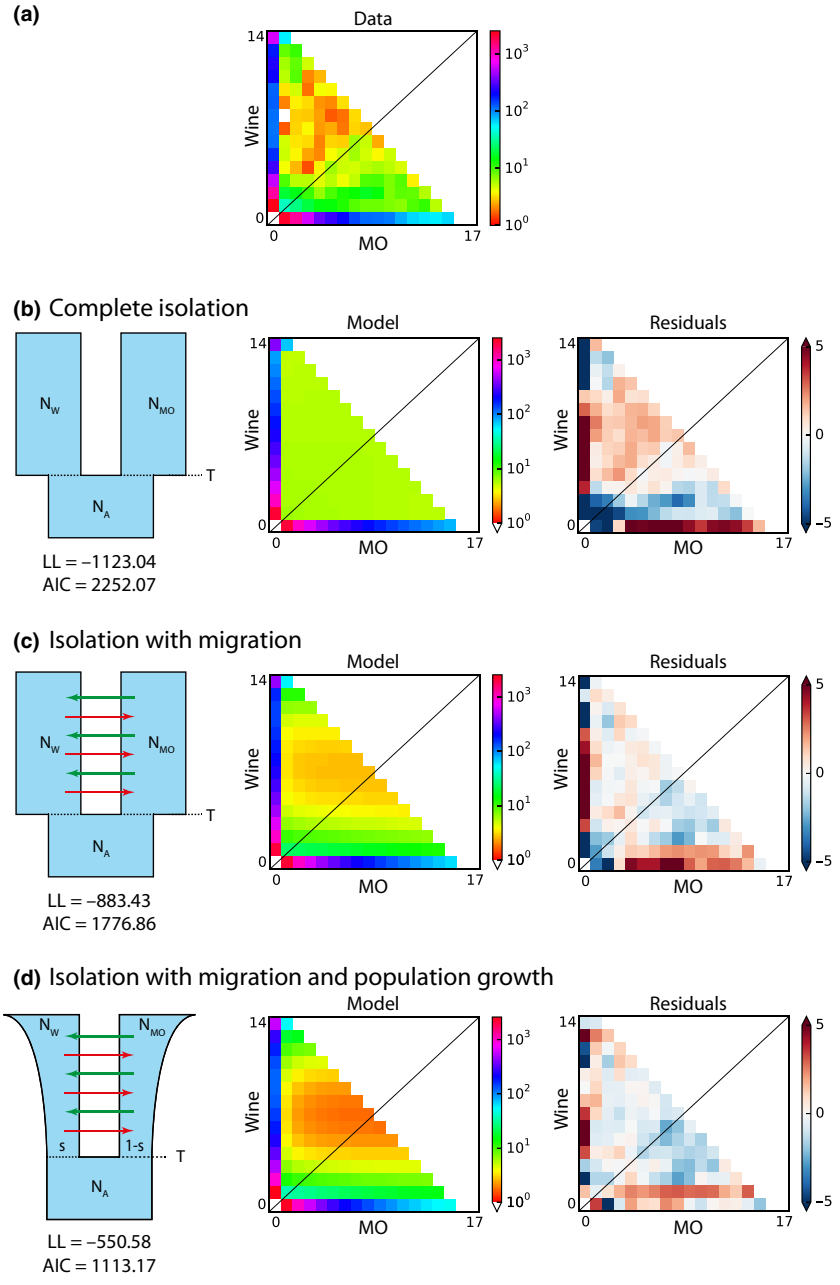

Fig. 4 Joint allele frequency spectrum for the Wine and Mediterranean oak populations of *Saccharomyces cerevisiae* and comparison with the expected frequency spectrum for different demographic scenarios. (a) Representation of the folded joint frequency spectrum, in the form of a heatmap, using 1 286 807 bp of non-coding regions across the genome. X and Y axes represent the number of chromosomes (strains) in the Mediterranean oak (MO) and Wine populations, respectively. (b–d) From left to right, illustrative representation of different demographic scenarios; the joint allele frequency spectrum expected under each model; and the residuals resulting from fitting the data in (a) to the respective model. The residuals represent the normalized difference between model and data for each bin in the spectrum (red indicates that the model predicts too many SNPs in that bin and blue that the model predicts too few). Below each scenario is also shown the log likelihood (LL) for the fitting and the respective Akaike information criterion (AIC). N_W, N_{MO} and N_A represent the effective population sizes for Wine, Mediterranean oak and ancestral populations, respectively. s is the fraction of the ancestral population that goes to the Wine population during the split (1-s goes to the Mediterranean oak population). T is the time of the split.

bridge as documented in other cases (Hewitt 2004). Although more detailed comparisons of Asian and North American populations are needed, our phylogenetic and population genomic analyses are consistent with the interpretation that North American lineages are descendants of Asian populations. By contrast, the newly uncovered natural Mediterranean population is monophyletic and much less diverse. We speculate that this is due to its more recent origin from a small number of migrants, possibly from Asia, perhaps with limited expansion because of competition with the sympatric and more prevalent *S. paradoxus* population.

Previous studies considered up to now the existence of a single lineage known as Wine/European but consisting mostly of strains from the wine environment (Liti *et al.* 2009; Schacherer *et al.* 2009; Wang *et al.* 2012; Cromie *et al.* 2013). Contrary to this, we detected a wild, oak-associated European population that, although closely related to the wine group, can be distinguished from wine strains in phylogenetic and population analyses and in domestication fingerprints. This novel wild population appears confined to Southern Europe because surveys in natural woodland environments in central and northern Europe by our team and others

Table 4 Best-fit population demographic parameters. Maximum-likelihood parameter estimates for an isolation with asymmetrical migration model allowing population growth after the split. The model was fitted to the joint allele frequency spectrum of Wine (W) and Mediterranean oak (MO) populations. Bias-corrected 95% confidence intervals were obtained from 100 simulated data sets using the maximum-likelihood estimates. s is the fraction of the ancestral population that goes to the Wine population during the split (1-s goes to the Mediterranean oak population). N_e is effective population size. Time is given per generation. Migration is the effective number of migrants per generation

	Maximum likelihood	95% confidence interval
Ancestral N_e	1 463 099	1 426 916–1 504 358
s	0.2224	0.2196–0.2249
Wine (W) N_e	5 746 917	5 688 166–5 805 668
Mediterranean oak (MO) N_e	4 294 474	4 248 888–4 340 059
Divergence time (gen.)	3 754 457	3 742 683–3 844 434
Migration MO → W	0.35641	0.3520–0.3586
Migration W → MO	0.01649	0.0137–0.0202

(e.g. Johnson *et al.* 2004) failed to yield *S. cerevisiae*. On the contrary, *S. paradoxus*, possibly an older occupant of the European continent, is distributed over a wider geographic range both in Europe (Johnson *et al.* 2004) and in North America (Charron *et al.* 2014; Leducq *et al.* 2014). Besides the wild Mediterranean population of *S. cerevisiae*, we sporadically isolated pure (nonmosaic) North American genotypes in Western Europe, possibly the result of recent, human-related and episodic migration events. The pervasiveness of wild *S. cerevisiae* yeasts in Southern Europe, along the Mediterranean basin, could have created the opportunity for these yeasts to predominate in the early grape must fermentations carried out in this region. The practice of skimming off the surface of the best musts for use in later fermentations (McGovern 2003) may have fostered the unintended selection of the best strains as wine yeasts have better enological properties than wild strains (Hyma *et al.* 2011).

The close relationship between the Mediterranean oak population and the wine group initially observed in the phylogenetic analyses was subsequently confirmed by the analyses carried out in STRUCTURE and fineSTRUCTURE and by the divergence and shared polymorphisms measurements. Hence, taken together, our analyses are consistent with the hypothesis that the common ancestor between Mediterranean oak and wine strains provided the raw genetic material that participated in early wine fermentations. A competing hypothesis posits that the first wine yeasts belonged to an undetected or extinct wild population close but not coincident with the Mediterranean oak population. We note that there is presently no evidence for such a distinct population. Not only do all wild Mediterranean strains fall within a single clade, but our field surveys yielded no *S. cerevisiae* isolates from North Africa and the Near East (approximately 100 oak samples tested).

Our test of different demographic models using the diffusion-based approach indicated that the best explanation for the observed relationship between Mediterranean oak and wine yeasts contemplated a scenario with partial isolation, asymmetric migration and growth of the two populations. Although migration rates were estimated to be relatively low, thus excluding gene flow as the major drive for the observed closeness of the two populations, a much higher migration from the Mediterranean oak population into the Wine population was detected. These results point to a complex demographic history of domesticated wine yeasts and their wild ancestors, suggesting that further studies are needed to fully capture the population dynamics of wine yeast domestication. Nevertheless, our demographic inference and the weak population structure between wine and Mediterranean oak populations are compatible with a Mediterranean oak population representing the wild genetic stock of wine yeasts.

The strong bottleneck detected for the Wine population at the time of the split fits in the classical domestication scenario (e.g. Doebley *et al.* 2006) and the estimated timing of the divergence of the wine group is generally compatible with available historical evidence. However, nucleotide diversity in the wine group is equivalent to that found in the Mediterranean oak group, which deviates from the norm as typically a loss of diversity in domesticates by comparison with their wild relatives is observed as a consequence of population bottlenecks (Doebley *et al.* 2006). The possible migrant nature of the wild ancestors of wine yeasts could have contributed to this situation due to a reduced genetic diversity of the subpopulation that colonized Europe. We note that the diversity of the other two oak-associated populations from North America and Japan is 2.5 times higher than that of the Mediterranean oak population. Moreover, the detected migration from the wild stock to the domesticated one could also have contributed to increase the diversity of the domesticated group. Also, the expansion of viticulture and winemaking to other continents (America, Asia and Oceania), promoting therefore the dissemination of wine yeasts (Pretorius 2000), might have enlarged the level of admixture with local natural strains, thus increasing the genetic diversity of the domesticated group. It is also possible that variation in preferences for fermentative attributes between regions and wine producers, and differences in wines in different regions,

have selected distinct genotypes thus enhancing diversity. Although less common, equivalent levels of diversity between wild and the corresponding domesticated populations have already been documented like in the case of apple domestication (Cornille *et al.* 2012).

Finally, we observed a marked difference between wine yeasts and their closest wild relatives. Regions A, B and C, acquired independently by HGT, contain genes that enhance sugar and nitrogen metabolism, thus contributing to properties likely to be selected for in wine yeasts. In some cases, the transferred genes increase fitness in wine must, thus supporting the association of these regions with the domestication of wine yeasts (Marsit *et al.* 2015). These regions are pervasive in the wine group but are notoriously absent in the Mediterranean oak group, as well as in other wild groups. Therefore, they are an example of a genomic transformation intrinsically associated with the domestication of wine yeast and consequently a trait that is expected to be absent in their wild relatives.

Although microbe domestication has received much less attention than the study of animal and plant domestication, filamentous fungi of the genera *Aspergillus* and *Penicillium* together with lactic acid bacteria and *Saccharomyces* yeasts have played a key role in the production of foods and beverages since ancient times. Genomic studies are starting to reveal the transformations that originated the domesticated phenotypes (Gibbons *et al.* 2012; Cheeseman *et al.* 2014) and the evolutionary routes that converted natural populations into fined-tuned 'cell factories' (Libkind *et al.* 2011; Almeida *et al.* 2014). Apparently, the markedly different contexts of microbe domestication have driven different organismic responses. In *A. oryzae*, the mould responsible for the saccharification of starch in Asian fermented foods and beverages, a comparison of sequence, gene expression and protein abundance indicated that domestication has led to a restructuring of primary and secondary metabolism (Gibbons *et al.* 2012). Contrastingly, in *Penicillium* used in cheese production such as *P. camemberti* and *P. roqueforti*, a horizontally transferred 575-kb-long genomic island containing genes involved in antagonistic interactions with other microorganisms was detected in strains from food environments. Also, in cider and wine yeast strains of *S. uvarum*, but not in wild isolates of this species, the massive acquisition of foreign genes from the sibling species *S. eubayanus* through introgression has been documented (Almeida *et al.* 2014). Taken together, these examples show that microbe domestication can proceed through multiple routes of genome reorganization that can include subtle reshufflings or dramatic modifications. As in the cases of crop and livestock domestication, linking wild and domesticated microbe genotypes is an essential step for understanding the roots and trajectories of man-driven artificial selection. Our results advance the knowledge of wine yeast domestication by revealing the closest wild relatives of domesticated lineages and the wild genetic stock that underwent domestication.

Acknowledgements

This work was supported by FCT Portugal grants PTDC/BIA-EVF/118618/2010 (JPS, PA, PG), PTDC/AGR-ALI/118590/2010 (JPS, PA, PG, RB), UID/Multi/04378/2013 (JPS, PG) and SFRH/BD/77390/2011 (PA), Infrastructural Centre Mycosmo, MRIC UL, Slovenia (PZ), GIS IBiSA-AO 2010–2011 France (JLL, SD) and the Natural Environment Research Council UK, NE/D008824/1 (DB). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. For kindly providing strains, we thank Ana Pinharanda, Heather Robinson, Eladio Barrio and Stephanie Diezmann.

References

Almeida P, Gonçalves C, Teixeira S *et al.* (2014) A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum. *Nature Communications*, **5**, 4044.

Bergström A, Simpson JT, Salinas F *et al.* (2014) A high-definition view of functional genetic variation from natural yeast genomes. *Molecular Biology and Evolution*, **31**, 872–888.

Bing J, Han P-J, Liu W-Q, Wang Q-M, Bai F-Y (2014) Evidence for a Far East Asian origin of lager beer yeast. *Current Biology*, **24**, R380–R381.

Bradley RK, Roberts A, Smoot M *et al.* (2009) Fast statistical alignment. *PLoS Computational Biology*, **5**, e1000392.

Cavalieri D, McGovern PE, Hartl DL, Mortimer R, Polsinelli M (2003) Evidence for *S. cerevisiae* fermentation in ancient wine. *Journal of Molecular Evolution*, **57**(Suppl 1), S226–S232.

Charron G, Leducq J-B, Bertin C, Dubé AK, Landry CR (2014) Exploring the northern limit of the distribution of *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* in North America. *FEMS Yeast Research*, **14**, 281–288.

Cheeseman K, Ropars J, Renault P *et al.* (2014) Multiple recent horizontal transfers of a large genomic region in cheese making fungi. *Nature Communications*, **5**, 2876.

Cornille A, Gladieux P, Smulders MJM *et al.* (2012) New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genetics*, **8**, e1002703.

Cromie GA, Hyma KE, Ludlow CL *et al.* (2013) Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. *G3 (Bethesda, Md.)*, **3**, 2163–2171.

Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell*, **127**, 1309–1321.

Douglas GL, Klaenhammer TR (2010) Genomic evolution of domesticated microorganisms. *Annual Review of Food Science and Technology*, **1**, 397–414.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fay JC, Benavides JA (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genetics*, **1**, 66–71.

Galeote V, Novo M, Salema-Oom M *et al.* (2010) FSY1, a horizontally transferred gene in the *Saccharomyces cerevisiae* EC1118 wine yeast strain, encodes a high-affinity fructose/H+ symporter. *Microbiology*, **156**, 3754–3761.

Galeote V, Bigey F, Beyne E *et al.* (2011) Amplification of a *Zygosaccharomyces bailii* DNA Segment in wine yeast genomes by extrachromosomal circular DNA formation (N Nikolaidis, Ed,). *PLoS One*, **6**, e17872.

Gibbons JG, Salichos L, Slot JC *et al.* (2012) The evolutionary imprint of domestication on genome variation and function of the filamentous fungus *Aspergillus oryzae*. *Current Biology*, **22**, 1403–1409.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data (G McVean, Ed,). *PLoS Genetics*, **5**, e1000695.

Hewitt GM (2004) The structure of biodiversity - insights from molecular phylogeography. *Frontiers in Zoology*, **1**, 4.

Hornsey IS (2003) *A History of Beer and Brewing*. Royal Society of Chemistry paperbacks, Cambridge, UK.

Hornsey IS (2007) *The Chemistry and Biology of Winemaking*. Royal Society of Chemistry, Cambridge, UK.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**, 254–267.

Hutter S, Vilella AJ, Rozas J (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics*, **7**, 409.

Hyma KE, Fay JC (2013) Mixing of vineyard and oak-tree ecotypes of *Saccharomyces cerevisiae* in North American vineyards. *Molecular Ecology*, **22**, 2917–2930.

Hyma KE, Saerens SM, Verstrepen KJ, Fay JC (2011) Divergence in wine characteristics produced by wild and domesticated strains of *Saccharomyces cerevisiae*. *FEMS Yeast Research*, **11**, 540–551.

Joffe AH (1998) Alcohol and social complexity in ancient Western Asia. *Current Anthropology*, **39**, 297–322.

Johnson LJ, Koufopanou V, Goddard MR *et al.* (2004) Population genetics of the wild yeast Saccharomyces paradoxus. *Genetics*, **166**, 43–52.

Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genetics*, **8**, e1002453.

Leducq J-B, Charron G, Samani P *et al.* (2014) Local climatic adaptation in a widespread microorganism. *Proceedings Biological Sciences/The Royal Society*, **281**, 20132472.

Legras J-L, Merdinoglu D, Cornuet J-M, Karst F (2007) Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Molecular Ecology*, **16**, 2091–2102.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, **25**, 2078–2079.

Libkind D, Hittinger CT, Valerio E *et al.* (2011) Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences*, **108**, 14539–14544.

Liti G, Barton DBH, Louis EJ (2006) Sequence diversity, reproductive isolation and species concepts in saccharomyces. *Genetics*, **174**, 839–850.

Liti G, Carter DM, Moses AM *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–341.

Lynch M, Sung W, Morris K *et al.* (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences*, **105**, 9272–9277.

Marsit S, Mena A, Bigey F *et al.* (2015) Evolutionary advantage conferred by an Eukaryote-to-Eukaryote gene transfer event in wine yeasts. *Molecular Biology and Evolution*, **32**, 1695–1707.

McGovern PE (2003) *Ancient Wine: the Search for the Origins of Viniculture*. Princeton University Press, Princeton, New Jersey.

McGovern PE (2009) *Uncorking the Past: the Quest for Wine, Beer, and Other Alcoholic Beverages*. University of California Press, Berkeley, California.

McGovern PE, Glusker DL, Exner LJ, Voigt MM (1996) Neolithic resinated wine. *Nature*, **381**, 480–481.

McGovern PE, Zhang J, Tang J *et al.* (2004) Fermented beverages of pre- and proto-historic China. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 17593–17598.

McGovern PE, Luley BP, Rovira N *et al.* (2013) Beginning of viniculture in France. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 10147–10152.

Michel RH, McGovern PE, Badler VR (1992) Chemical evidence for ancient beer. *Nature*, **360**, 24.

Naumov GI, Naumova ES, Sniegowski PD (1998) *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* are associated with exudates of North American oaks. *Canadian Journal of Microbiology*, **44**, 1045–1050.

Novo M, Bigey F, Beyne E *et al.* (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast *Saccharomyces cerevisiae* EC1118. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 16333–16338.

Pretorius IS (2000) Tailoring wine yeast for the new millennium: novel approaches to the ancient art of winemaking. *Yeast (Chichester, England)*, **16**, 675–729.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salzberg SL, Phillippy AM, Zimin A *et al.* (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Research*, **22**, 557–567.

Sampaio JP, Gonçalves P (2008) Natural populations of Saccharomyces kudriavzevii in Portugal are associated with oak bark and are sympatric with *S. cerevisiae* and *S. paradoxus*. *Applied and Environmental Microbiology*, **74**, 2144–2152.

Samuel D (1996) Investigation of ancient Egyptian baking and brewing methods by correlative microscopy. *Science*, **273**, 488–490.

Samuel D (2000) Brewing and baking. In: *Ancient Egyptian Materials and Technology*, pp. 537–576. Cambridge University Press, Cambridge, New York.

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature*, **458**, 342–345.

Sniegowski PD, Dombrowski PG, Fingerman E (2002) *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* coexist in a natural woodland site in North America and display different levels of reproductive isolation from European conspecifics. *FEMS Yeast Research*, **1**, 299–306.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, **22**, 2688–2690.

Strope PK, Skelly DA, Kozmin SG *et al.* (2015) The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Research*, **25**, 762–774.

Tamura K, Peterson D, Peterson N *et al.* (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**, 2731–2739.

Thornton K (2003) Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics (Oxford, England)*, **19**, 2325–2327.

Valamoti SM, Mangafa M, Koukouli-Chrysanthaki C, Malamidou D (2007) Grape-pressings from northern Greece: the earliest wine in the Aegean? *Antiquity*, **81**, 54–61.

Wang Q-M, Liu W-Q, Liti G, Wang S-A, Bai F-Y (2012) Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Molecular Ecology*, **21**, 5404–5417.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

D.B., J.P.S., P.A. and P.G. conceived and designed the research; B.T., D.B., J.P.S., K.S., P.A., P.G., P.Z., R.B. and Y.I. isolated and identified new yeast strains; A.C., D.B., J.G., P.A. and R.B. obtained and assembled genomic data; J.L.L. and R.B. performed experiments; D.B., J.L.L., J.P.S., M.S., P.A., P.G., R.B. and S.D. analysed the data; D.B., J.L.L., P.A. and S.D. contributed new analytic tools; and P.A., J.P.S. and P.G. wrote the manuscript with advice and consent from all authors.

Data accessibility

All new and pre-existing sequence reads are available in public repositories and Accession nos are provided in Table S1 (Supporting information). Genome sequencing data generated for this study have been deposited in NCBI's SRA (http://www.ncbi.nlm.nih.gov/sra) as SRP059414 and in EBI's ENA (https://www.ebi.ac.uk/ena) as PRJEB7675 and PRJEB7601. Raw data of microsatelite genotypes, whole-genome SNP alignments, STRUCTURE, fineSTRUCTURE and ∂a∂i input files, raw results from VariScan and libsequence analysis package, newick tree files and *de novo* draft assemblies are available from Dryad Digital Repository doi:10.5061/dryad.hm2jf.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Network analysis showing ecological and geographic separation of *Saccharomyces cerevisiae* lineages. Neighbour-net network inferred from the allelic variation at 12 microsatellite loci using the Bruvo distance. Branches are coloured according to the source of strain isolation.

Fig. S2 fineSTRUCTURE co-ancestry matrix and clustering showing strain information. The clustering was performed on 98 strains using 94 089 informative biallelic SNPs. The colour of each bin in the matrix indicates the expected number of 'chunks' copied from a donor (column) to a recipient strain (row). The dendrogram on the top represents the clustering of strains inferred from the co-ancestry matrix. Branches are coloured according to STRUCTURE clusters for easier comparison. Branches coloured in grey indicate mosaic strains identified by STRUCTURE at $K = 2$. These mosaics are also identified with an 'M' in front of the strain name.

Fig. S3 Whole-genome phylogeny of the large strain data set (145 strains) depicting those identified as mosaics by STRUCTURE at $K = 2$ and the presence/absence of the wine-related regions A, B and C *sensu* Novo *et al.* (27). The tree was inferred from 146 097 SNPs, using the maximum-likelihood method as implemented in RAxML with the GTRGAMMA model of sequence evolution and was rooted with *S. paradoxus*. Branches are coloured according to the substrate of isolation of each strain. Strains are represented by coloured dots indicating the geographic origin. Branch lengths correspond to the expected number of substitutions per site. Support values from bootstrap replicates above 90% are depicted with black dots in the respective tree nodes. Coloured squares illustrate the presence of regions A, B and C. These regions were searched by BLASTing *de novo* assemblies (*) or by read mapping (‡) when only single-end read data was available. Red stars mark the strains identified as mosaics by STRUCTURE at $K = 2$.

Fig. S4 Multilocus phylogeny of *Saccharomyces cerevisiae*, including the Chinese lineages. Neighbour-Joining tree inferred from a concatenated alignment of 13 loci using the Kimura 2-parameter model of sequence evolution. Branch lengths correspond to the expected number of substitutions per site. Support values from bootstrap replicates above 50% are indicated with orange dots. The tree is rooted with *S. paradoxus*. Some branches are collapsed to indicate the major phylogenetic groups. Branches coloured in black represent Chinese strains for which whole-genome data in not available. Branches coloured in blue indicate strains for which genome data is available (strains used in this study).

Table S1 Strains and genomes used in this study and relevant information pertaining to them.

Table S2 Comparison of polymorphisms (mean values) at coding and noncoding regions between the Wine and Mediterranean oak groups. For the Wine group only commercial or strains isolated from wine must were used.

Table S3 Best-fit parameter estimates for three alternative demographic models. Maximum-likelihood parameter estimates for isolation models assuming no migration between populations, asymmetric migration and asymmetric migration with population growth. Each model was fitted to the joint allele frequency spectrum of Wine (W) and Mediterranean oak (MO) populations. Units are reported as in ∂a∂i. Population sizes are reported relative to a reference (ancestral) population set at $N_A = 1$. k is the number of parameters in the model.