Genomic signatures of adaptation to wine biological aging conditions in biofilm-forming flor yeasts

Anna Lisa Coi, Frederic Bigey, Sandrine Mallet, Souhir Marsit, G. Zara, Pierre Gladieux, Virginie Galeote, Marilena Budroni, Sylvie Dequin, Jean-Luc Legras

Received Date: 31-Mar-2016

Revised Date: 31-Jan-2017

Accepted Date: 31-Jan-2017

Article Type: Special Issue

Genomic signatures of adaptation to wine biological aging conditions in biofilm-forming flor yeasts

Coi AL[1], Bigey F[2], Mallet S[2], Marsit S[2], Zara G[1], Gladieux P[3], Galeote V[2], Budroni M[1], Dequin S[2], Legras JL[2]

[1]Dipartimento di Agraria, Università di Sassari, 07100 Sassari, Italy

[2]SPO, INRA, SupAgro, Université de Montpellier, 34060, Montpellier, France

[3]INRA, UMR BGPI, 34398 Montpellier, France

Corresponding author: jean-luc.legras@inra.fr

Keywords: Saccharomyces cerevisiae, flor yeast, genome, adaptation, *ZRT1*, biofilm, biological aging, FLO11, domestication

Abstract:

The molecular and evolutionary processes underlying fungal domestication remain largely unknown despite the importance of fungi to bioindustry and for comparative adaptation genomics in eukaryotes. Wine fermentation and biological aging are performed by strains of *S. cerevisiae* with, respectively, pelagic fermentative growth on glucose, and biofilm aerobic growth utilizing ethanol. Here, we use environmental samples of wine and flor yeasts to investigate the genomic basis of yeast adaptation to contrasted anthropogenic environments. Phylogenetic inference and population structure analysis based on single nucleotide polymorphisms (SNPs) revealed a group of flor yeasts separated from wine yeasts. A combination of methods revealed several highly differentiated regions between wine and flor yeasts, and analyses using codon-substitution models for detecting molecular adaptation identified sites under positive selection in the high affinity transporter gene *ZRT1*. The Cross Population Composite Likelihood Ratio (XP-CLR) revealed selective sweeps at three regions, including in the hexose transporter gene *HXT7*, the yapsin gene *YPS6* and the membrane protein coding gene *MTS27*. Our analyses also revealed that the biological aging environment has led to the accumulation of numerous mutations in proteins from several networks, including *Flo11* regulation and divalent metal transport. Together, our findings suggest that the tuning of *FLO11* expression and zinc transport networks are a distinctive feature of the genetic changes underlying the domestication of flor yeasts. Our study highlights the multiplicity of genomic changes underlying yeast adaptation to man-made habitats, and reveals that flor/wine yeast lineage can serve as a useful model for studying the genomics of adaptive divergence.

Introduction

Domestication is a specific case of adaptive divergence in response to man-mediated artificial selection, and an opportunity to study the genetic basis of adaptation (Ross-Ibarra *et al.* 2007; Gladieux *et al.* 2014). Despite the importance of domesticated fungi to bioindustry, and as model eukaryote species, relatively few studies – compared to plants and animals - have investigated the genomic and evolutionary changes underlying fungal domestication (Galagan *et al.* 2003; Gibbons & Rinker 2015; Ropars *et al.* 2015). Domesticated *Saccharomyces cerevisae* yeasts have emerged as a preeminent fungal group for comparative evolutionary genomics of domestication and subsequent strain improvement, due to (1) their close connection with the history of agriculture and human migrations, (2) the existence of multiple domestication events (Fay & Benavides 2005; Baker *et al.* 2015; Gallone *et al.* 2016), and (3) unrivaled set of tools to explore genotype-phenotype-fitness relationships (Liti *et al.* 2009; Scannell *et al.* 2011; Ho & Zhang 2014). For example, the comparison of sake and wine *Saccharomyces cerevisiae* yeasts indicated that the recurrent loss of functional aquaporins was involved in their adaptation to contrasted anthropogenic environments (Will *et al.* 2010). The acquisition by horizontal gene transfer of peptide transporters with high affinity for grape peptides (Marsit *et al.* 2015) or strong ability to expel sulfites from the cell (Pérez-Ortín *et al.* 2002) are other examples of yeast adaptation to the wine-making environment. More generally, wine-making exposes yeast populations to contrasted environmental conditions inside wineries and between wineries and vineyards, providing unique opportunities for investigating the genetic bases of yeast adaptation to man-made habitats.

Wine making was invented more than 7000 years before present and technological developments have continuously transformed the wine-making process to improve wine quality and conservation. The use of sulfites by Egyptians and Romans to clean wine containers (Romano & Suzzi 1993), or the spread of bottles with corks in the 17th century (Estreicher 2006) are examples of technical developments that have drastically improved oxygen management and paved the way to the modern practice of protecting wine aromas from oxygen after alcoholic fermentation. Before the advent of the efficient methods for limiting the impact of oxygen on aromas, winemakers were making use of

the spontaneous oxidative changes occurring during wine aging and this practice has continued in some European countries (Hungary, France, Italy and Spain). The process of wine aging, referred to as biological aging, takes place after the completion of alcoholic fermentation and it leads to the production of specific flavors and aromas along with the progressive oxidation of alcohol and remaining carbohydrates. Biological aging involves a biofilm called velum that forms on the surface of wine and consists in specific *Saccharomyces cerevisiae* yeast strains called "flor yeasts". (Ibeas & Jimenez 1997; Legras *et al.* 2014).

Wine and flor yeasts are intimately associated with wine fermentation and biological aging processes. Wine yeasts, which perform alcoholic fermentation, represent a genetically homogeneous group that experienced a recent demographic expansion (Fay & Benavides 2005; Legras *et al.* 2007; Liti *et al.* 2009; Schacherer *et al.* 2009), preceded by a domestication bottleneck that greatly reduced genetic variability (Fay & Benavides 2005). A recent population genomic study suggests that *S. cerevisiae* associated with Mediterranean oaks is the main progenitor of the wine yeast lineage (Almeida *et al.* 2015). Flor yeasts have the specific ability of weaving a biofilm, in addition to their ability to perform alcoholic fermentation (Fidalgo *et al.* 2006; Legras *et al.* 2014). The close association of flor yeast with the aging process, which has been used for centuries of almost continuous growth, also suggests that flor yeasts represent a domesticated group of microbes and that they may display specific genomic features allowing them to thrive in the harsh conditions of biological aging (Legras *et al.* 2007). Fidalgo et al. (2006) suggested that the process of biological aging impacted biofilm formation and led to the fixation of adaptive mutations in the promoter and coding region of the gene *FLO11* (coding for a flocculin), resulting in a higher hydrophobicity of yeast cells. The Flo11 protein is also central to the mechanism enabling yeasts to invade agar or to produce a pseudohypha under nutritional deprivation (Ryan *et al.* 2012). The regulation of *FLO11* expression is highly complex, involving several pathways, such as the RAS/cAMP-PKA and the MAP-kinase pathways, and also two non-coding RNAs (ICR1 and PWR1)(Brückner & Mösch 2011). To date, no other allelic variation has been associated with the specific habitat of flor yeasts. Here, we used whole genome resequencing to investigate the evolutionary origins of flor yeasts and to identify candidate genomic

regions involved in the adaptation of flor yeasts to their specific habitat. For some of the candidate genomic regions identified by our analyses, we provide evidence for the phenotypic impact of allelic differences between wine and flor yeast through allelic replacement or reciprocal hemizygosity tests.

Materials and methods

Yeast strains, media and growth conditions:

Yeast strains used in this work are listed in Table S1a and b. For some strains, we chose to sequence diploid derivatives from the original wine or flor strain to obtain assemblies from homozygous strains and to improve the phasing of other diploid genomes.

For the phenotypic validation experiments, yeasts were precultivated in YEPD (2% glucose, 1% yeast extract, and 2% peptone), and after 24 hours, they were inoculated at OD_{600} = 0.1 in YNB (Difco) + 4% ethanol (v/v) in 24-well plates (Coi *et al.* 2016). Photographs of plates were taken after 5 to 7 days of growth, and the surfaces of the biofilms were measured using images taken with IMAGEJ V2.0.0 under the FIDJI environment (http://imagej.net).

Construction of mutants:

The primers used to construct the deletion cassette and to verify the deletion are listed in the Supplementary Material (Table S2). The allelic replacement of *RGA2* and *ZRT1* in P3-D5 MATα was performed in two steps, as follows. *RGA2* or *ZRT1* was deleted using a Kanamycin cassette and then replaced with the wine functional allele amplified in tandem with hygromycin for *RGA2* or with the wine functional allele for *ZRT1*. Selection was then performed in SD medium supplemented with 1 mM EDTA to chelate zinc (Gitan *et al.* 1998). The resulting P3-D5*K1-RGA2*:Hyg MATα and P3-D5 *K1-ZRT1*MATα strains were checked by sequencing. For constructing *SFL1* reciprocal hemizygotes, we deleted *SFL1* using a Kanamycin cassette in P3-D5 MATα and in K1-280-2B MATa and crossed each deleted strain with the wild type of the other background. *SLN1* hemizygotes were obtained by the deletion of one copy of *SLN1* in the P3-D5 MATα x K1-280-2B MATa zygote and identification of the remaining allele by sequencing. All cassettes were amplified by PCR using KAPA HiFi HotStart Taq polymerase (KAPA BIOSYSTEMS) or conventional Taq DNA Polymerase (Fermentas).

Genome sequencing, population analysis and SNP detection:

We used the genomic sequences of 9 flor strains from Spain, Italy, France and Hungary and 9 wine strains from Italy and France (Table S1a) recently published by our group (Marsit *et al.* 2015). We added new genome sequences obtained with the same technology for EC1118 and 59A, the haploid derivative of EC1118. Genomic DNA was prepared following a standard phenol chloroform extraction protocol. Autosomal DNA was separated from mitochondrial DNA via isopycnic CsCl ultracentrifugation at 289,000*g* for 20 hours (Barth & Gaillardin 1996) and processed to generate libraries of 300-bp inserts. The 20 libraries were multiplexed in different lanes of Illumina HiSeq 2000. Paired-end (2x100 nucleotides) sequencing resulted in an average theoretical coverage of 77 to 730X. Image analysis and data extraction were performed using Illumina RTA version 1.13.48.0 and CASAVA version 1.8.2. Read sequences were processed to improve their global quality; they were trimmed for the first 9 nucleotides using FASTX Toolkit version 0.0.13 and for low-quality regions using Sickle version 1.000 (quality below 25 in a 20-bp sliding window). Reads shorter than 50 bp were removed. Genomic sequences were assembled with SOAP*denovo* v1.05 (Li et al., 2010) using different Kmers and by keeping the assembly with the highest N50.

We used the Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010), version 2.3-9, for calling SNPs and indels, using Best Practice Variant Detection with GATK v4, release 2.0, available online. In short, the workflow is divided into the following four sequential steps: initial mapping, refinement of the initial reads, multi-sample indel and SNP calling and, finally, variant quality score recalibration. First, reads were aligned to the S288c reference genome (release number R64-1-1, downloaded from the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/) using BWA version 0.6.2 (Li & Durbin, 2009). Second, optical and PCR duplicates were removed using MARKDUPLICATE from the PICARD TOOLS version 1.84 (http://picard.sourceforge.net). Reads encompassing indels were realigned to increase the indel call accuracy using IndelRealigner (GATK). The base quality was recalibrated using BaseRecalibrator/PrintReads (GATK) to bring this probability closer to the probability of mismatching the reference genome. At the end of this step, we obtained the analysis-ready reads. Third, we performed SNP and indel discovery across all the sequences simultaneously using UnifiedGenotyper (GATK). Fourth, we used Variant Quality Score Recalibration

(VQSR) to build an adaptive error model using known variant sites and then applied this model to estimate the probability that each variant in the callset is a true genetic variant or a machine/alignment artifact. This step was performed with the VariantRecalibrator/ApplyRecalibration tools of GATK, using a dataset of known SNPs and indels obtained from 86 available genomes to train the model (including EC1118 and reference genome S288c). This genotyping pipeline provided us a final multi-sample VCF file that contained all the variant sites discovered across samples with which a genotyping quality for each strain was associated. The final set was obtained when keeping biallelic variants corresponding to the truth sensitivity threshold of 100% to limit false positives and filtering genotypes with a minimum genotype quality of 30 and a minimum sequencing depth of 10 (HighQuality-SNP set). The genotyping of the 20 strains revealed 59,973 variant positions (indels and biallelic positions). The potential impact of each of these variants corresponding to nonsynonymous positions was then evaluated with the Sorting Intolerant from Tolerant program (SIFT) (Ng & Henikoff 2001), which predicts the probability that a mutation impacts protein structure from the physical properties and conservation of amino acids.

To compare these sequences with 84 other genomes available, genome sequences were downloaded from SGD (Wei *et al.* 2007; Liti *et al.* 2009; Argueso *et al.* 2009; Borneman *et al.* 2011; Akao *et al.* 2011; Engel & Cherry 2013) and aligned pairwise to the S288C reference genome sequence using MUMmer 3.0 (Kurtz *et al.* 2004). SNPs were extracted with a custom script and combined with the set of SNPs obtained from genotyping, resulting in a set of 427,587 variant positions after filtering out a maximum of 20% missing sites per locus (Allyeast-set). This combined set of variants has been used for the analysis of global phylogeny, population structure analysis, variant differentiating flor strains from other populations, and the subset of wine and flor strains for sweep detection with XP-CLR. These sets of variants were then further phased with BEAGLE 3.3 (Browning and Browning, 2007).

Genetic analyses:

To infer relationships among the strains, the table of SNPs (vcf format) was converted to a fasta file using SNPhylo v20140701 (Lee *et al.* 2014) and a maximum-likelihood genealogy was estimated using RAxML v8.1.3 (Stamatakis 2014) with the GTRGAMMA model of sequence evolution. The

population structure was analyzed using principal component analysis (PCA) and discriminant analysis of the principal component (DPCA) using the R package Adegenet v2.0 (Jombart *et al.* 2010; Jombart & Ahmed 2011) and the Admixture v1.22 software (Alexander *et al.* 2009).

Comparative analyses of wine and flor strains were carried out only on the set of variants resulting from genotyping with GATK. Metrics of absolute divergence (D_{XY}) and relative divergence ($D_a = D_{XY} - (\pi_{wine}+\pi_{flor})/2$ and Fst (Excoffier *et al.* 1992)), as well as a measure of the skewness of the allele frequency spectrum (Tajima's D), were computed using POPGENOME v2.1.6 R package (Pfeifer *et al.* 2014). F_{IS} were calculated using an R-script written by Eva Chan (http://www.evachan.org). Linkage disequilibrium (r^2) was analyzed on phased data for wine and flor populations using PLINK v1.9, and r^2 was averaged in classes of 10 bp distances surrounding pairs of SNPs. Recombination rates were estimated using the PAIRWISE and INTERVAL programs in LDHAT v2.2a (Auton & McVean 2007). The selection of the most divergent regions (D_{XY}, D_a, PCA differentiating flor strains from all other yeast) was performed using a threshold of 2% in order to obtain lists of segments corresponding to at most 5% of the total number of ORFs (330). A threshold of 20% was chosen from the distribution of the contribution of non synonymous potentially impacting SNPs to the PCA 1st component (Fig. S9) for the selection of SNPs differentiating best wine and flor strains.

Tests for detecting positive selection

We used a combination of approaches for detecting positive selection to identify adaptive events that occurred at different time-frames (Oleksyk *et al.* 2010). Positive selection over the long term was investigated based on the rate of accumulation of amino acid changes (approach based on the ratio of nonsynonymous to synonymous substitutions, ω=dN/dS) or based on the contrast between intraspecies polymorphism and interspecies divergence (McDonald-Kreitman test). Recent positive selection was investigated based on distinctive footprints on the site frequency spectrum and patterns of linkage disequilibrium.

Tests for positive selection based on the $\omega=d_N/d_S$ ratio were carried out using the codeml program in PAML v4 (Yang 2007). We used the branch-site test, which is aimed at detecting positive selection that impacted nucleotide substitutions in a specific lineage (the foreground branch; i.e., the flor

yeasts) in comparison to the rest of a phylogeny (the background branches). The branch-site test compares the likelihood of the fit of the alignment to the following two models: model A, which assumes two classes of sites with ω_0 between 0 and 1 and $\omega_1 = 1$ in the background branches and one supplementary class with $\omega_2 > 1$ in the foreground branch; in the second (null) model, the ω_2 parameter is set to 1. The test was performed on a subset of genes identified by at least 2 methods. For each gene, an alignment was obtained with muscle v3.6, including the alleles of P3-D5 and K1-280-2B (obtained previously for the McDonald-Kreitman test, with the exception of *ZRT1,* for which we used the assembly sequence given its too high divergence from the reference genome, leading to missing variant positions in the variant dataset) and the orthologs available at http://sss.genetics.wisc.edu/cgi-bin/s3.cgi for *S. paradoxus* CBS432, *S. uvarum* CBS7001, *S. mikate* IFO1815, and *S. kudriavzevii* IFO1802 or obtained from *S. arboricola* CBS 10644 and *S. eubayanus* CBS 12357 genome sequences (http://www.ncbi.nlm.nih.gov/genome/) when available. For each gene, maximum-likelihood phylogenies were estimated using RAxML v8.1.3 (Stamatakis 2014) with the GTRGAMMA model of sequence evolution.

To carry out the McDonald-Kreitman test, pseudo DNA sequences were generated for all genes using the phased variant file and the reference sequence, and divergence was measured against *S. paradoxus* (www.***saccharomycessensustricto***.org). The coding sequence was then translated into amino acids before alignment using MAFTT version 7 (Katoh & Standley 2013). Significance was computed using the MKtest program in the analysis package version 0.8.4 (Thornton 2003) using a Fischer exact test with correction for multiple testing (Benjamini & Hochberg 1995).

Genome scans for selective sweeps were performed using the Cross-population Composite Likelihood Ratio (XP-CLR) (Chen *et al.* 2010). This method relies on the modeling of the evolutionary trajectory of an allele under linked selection and under neutrality and enables the detection of regions in the genome for which the changes in allelic frequencies occurred too quickly to be due to random drift. The significance threshold was obtained from 1000 neutral simulations performed with ms (Hudson 2002) using the best demographic scenarios inferred for flor yeasts (see Demographic inferences

section) and using the global recombination rate inferred for the flor population with the pairwise module of LDhat. For each simulation, the highest value of the XP-CLR statistic was kept and the threshold was chosen as the 95% quantile of simulations.

Demographic inferences:

We performed demographic inferences separately for the split between wine and jura/Hungarian flor strains on the one hand and for the split between wine Spanish/Italian flor strains on the other hand. Preliminary analyses indicated that the joint demographic history of the two flor yeast lineages could not be inferred, likely owing to the limited sample size of each group. The analysis was run on a subset of SNPs for which ancestral states could be determined based on alignments with *S. paradoxus* CBS432 and *S. mikate* IFO 1815^T genome sequences. The folded joint allele frequency spectrum was calculated and fitted to different scenarios using a diffusion approximation based method as implemented in δaδi (Gutenkunst *et al.* 2009). Demographic models are presented in Supplementary material Fig. S5A and B.

Translocation and Chimeric allele detection

We searched for translocation events from the analysis of discordantly mapped paired-ends with Delly v0.2.2 (Rausch *et al.* 2012) using a cutoff of 50 minimum events per translocation site. Translocations were checked by sequencing candidate translocated alleles and by examining variations in sequencing depth.

Results

Origins of flor yeasts:

To examine the evolutionary history of flor strains in comparison to strains isolated from wine and other origins, we first inferred a maximum-likelihood genome genealogy (Fig. 1A). The genome genealogy revealed the main yeast lineages previously described, such as the lineages associated with wine, Asian beverages, North American oak (Liti *et al.* 2009; Cromie *et al.* 2013) or beer (Gallone *et al.* 2016). Eight of the nine flor strains formed a new cluster that was close but distinct from the cluster including wine yeasts. The flor strain F12-3B did not belong to the main flor lineage and was

instead included in the wine group; F12-3B was thus considered as a wine strain for subsequent analyses. The industrial strains EC1118, QA23, Vin13 and wine strain MJ73 formed a group that was basal to the group of flor strains and included recently described Champagne and fructophilic isolates (Borneman *et al.* 2016). Absolute and relative divergences between flor and wine were D_{XY} = 1.29 per kb and D_a = 0.7, respectively, which is lower than those estimated between wine strains and their putative progenitors isolated from Mediterranean oaks (D_{XY} = 2.0, D_a = 1.5 per kb) (Almeida *et al.* 2015). Nucleotide diversity was 0.94 per kb in wine strains and 0.53 or 0.28 per kb in flor strains with or without EC1118, respectively (Table 1).

Analyses of population subdivision and admixture:

We used a combination of model-based and non-parametric analyses of population structure to analyze more finely the subdivision within the lineage as well as the differentiation between, flor and wine lineages. We first used discriminant analysis of the principal component (DAPC) (Jombart & Ahmed 2011), which does not rely on any population genetic model and is therefore not affected by deviations from Hardy–Weinberg assumptions, such as those caused by selfing, which is relatively frequent in *S. cerevisiae* (Magwene *et al.* 2011)(Fig. 1B). The model with K = 10 clusters, delineating isolates from Africa, US Oaks, Asia, and wine/Europe (Liti *et al.* 2009) as well as flor and Champagne strains (closely related to EC1118), was the best model according to the Bayesian Information Criterion (BIC). For all K values, all individuals had high membership probabilities in a single cluster, suggesting limited admixture. However, because admixture is not specifically addressed in DAPC analyses, a second analysis was performed using the model-based clustering algorithm implemented in the Admixture software (Alexander *et al.* 2009), which jointly infers K clusters at Hardy-Weinberg and linkage equilibrium and estimates membership proportions of all individuals in the K clusters. The BIC suggested optimal clustering at K = 9 clusters (Fig. 1C), separating flor strains from those isolated from wine and other origins. Wine and flor genotypes were not grouped in separate clusters until models with K>5 clusters were used, and in contrast to DAPC, strains related to industrial strain EC1118 were inferred to result from admixtures between the wine and flor yeast gene pools. The mixed ancestry of EC1118 and related strains was confirmed by

analyses of shared haplotype segments implemented in the fineStructure program (Fig. S1) and by gene genealogies of two loci that are well differentiated between wine and flor strains (see section 'Genomic regions of elevated divergence between wine, flor and other yeasts'). Gene genealogies indicated that strains EC1118 and MJ73 harbor two ancestral phases of the high affinity zinc transporter gene *ZRT1* (Fig. S2a) and of the osmosensor gene *SLN1* (Fig. S2b).

Life cycle of flor yeasts:

As a low spore viability has been reported for several flor yeasts (Ibeas & Jimenez 1996; Budroni *et al.* 2000) in comparison to *S. cerevisiae* from other origins, suggesting a mainly asexual life cycle, we compared patterns of linkage disequilibrium and estimates of the recombination rate to gain insight into the contribution of recombination, and thus likely sexual reproduction, to the genomic variability of wine and flor populations. Linkage disequilibrium declined rapidly with distance, reaching half of its maximum value at 1.7 kb and 5.4 kb in flor and wine populations, respectively (Fig. S3). Estimates of the recombination rate also indicated genomic heterogeneity in rates of recombination within the genomes of each population (Fig. S4). Using the inbreeding coefficient F_{IS} (0.69 for flor yeasts), we estimated an outcrossing rate in the flor population that was similar to previous estimates for the wine population (Magwene *et al.* 2011), with 18% of individuals resulting from an outcrossing event. The decay of linkage disequilibrium with distance in the flor population, similar to wine population, and positive value of F_{IS} do not support that the actual flor population is a clonal lineage.

Demographic inference:

To gain insight into the demographic history of wine and flor yeasts and because deviations from the standard neutral model assuming constant population size can bias inferences of positive selection, we used the diffusion-based approach implemented in the δaδi program to infer the historical demography of wine and flor yeasts based on their joint allele frequency spectra. We compared different demographic models, including post-divergence gene flow and population size change (Gutenkunst *et al.* 2009). The most likely scenarios for the divergence between the wine group and the Spanish/Italian and French/Hungarian flor populations, respectively, included demographic

expansion of the wine group constant size followed by a recent bottleneck for the two flor populations, and with asymmetrical gene flow (Fig. S5a and 5b).

Differences in genomic content between wine and flor yeasts:

We previously identified three genomic regions in the genome of EC1118 and other wine strains (called A, B, and C) that have been acquired by horizontal gene transfer from distant yeasts (Novo *et al.* 2009). The latter region, which results from an initial transfer of at least 158 kb from *Torulaspora microellipsoides* into *S. cerevisiae*, provides an ecological advantage in grape must by enhancing the assimilation of oligopeptides from grape musts (Marsit *et al.* 2015). We thus searched for these three regions in flor yeasts. Region A was not found in any flor strain, whereas region B was found in 2 strains, and region C in 5 strains. Compared to EC1118, which contains a region C of 65 kb, truncated forms were found in the wine strains analyzed here, all carrying *FOT1* and *FOT2* genes in addition to 1 to 5 other ORFs. In contrast, the 4 flor strains that contain this fragment (P3-D5, GUF54, TA12-2, 7.7, F25) have the full 65-kb region C described for EC1118 that contains 18 ORFs, including the high-affinity fructose transporter *FSY1* (Galeote *et al.* 2010).

Specific translocations and recombined alleles in flor yeast:

A reciprocal translocation between the sulfite export pump *SSU1* and *ECM34* has been shown to be involved into the adaptation of wine yeast to the adjunct of sulfites in grape must (Pérez-Ortín *et al.* 2002). We thus searched for such translocation events in the genome of flor yeasts. The *ECM34-SSU1* translocation was detected in 4 flor strains. In addition, recombinations between orthologs were detected in flor strains (Table S3), for example, between *FRE2* and *FRE3,* encoding siderophore ferric reductases, and between *HXT3* and *HXT1,* encoding for low affinity glucose transporters (Fig. S6). The flor *HXT3* allele is identical to a fructophile allele of *HXT3* (Guillaume *et al.* 2007), which enables faster fermentation.

Genomic regions of elevated divergence between wine, flor and other yeasts:

Divergent genomic regions between populations exchanging migrants adapted to distinct environments are expected to be enriched in genes involved in adaptive divergence (Akey *et al.* 2002, 2010; Amato *et al.* 2009; Moradi *et al.* 2012) because genomic divergence between populations should be higher in regions resistant to introgression due to divergent selection than in regions homogenized by gene flow. We thus searched for such regions using different methods. As Principal Component Analysis (PCA) has been widely used for the detection of population structure (Patterson *et al.* 2006; McVean 2009) and provides the contribution of each SNP to the axis that differentiate two populations, we applied PCA to the set of genomes representing strains from wine, flor and various other origins (Allyeast-set). Flor strains were differentiated from all other yeast along the sixth axis of the PCA (Fig. 2A). The summed contributions of SNPs to these axis per 500 bp region (Fig. 2B) were used to detect highly differentiated regions between flor and wine strains or between flor and all other strains. Based on the PCA, we identified 319 genes or tRNAs differentiating the flor group from other groups (Table S3). We also computed metrics of absolute and relative genomic divergence (D_{XY} and D_a) between populations (Fig. 2C, Fig. S7) (Cruickshank & Hahn 2014). Genome scans using D_{XY} and D_a revealed 320 and 311 highly divergent genes or tRNAs between wine and flor, and 159 genes out of 507 were common to the sets of genes identified using PCA, D_{XY} and D_a (Fig. 3), including key genes involved in biofilm formation such as the major regulator of cAMP level, the GTPase *IRA1*; *SFG1,* a transcription factor required for growth of superficial pseudohyphae; *HMS2,* a protein with similarity to the heat shock transcription factor; the osmosensor *SLN1;* and three genes of the MAP-kinase pathway (the MAP-kinase *STE7*; *KDX1,* the protein kinase involved in cell wall integrity; and the CDC42 GTPase activating MAP kinase *RGA2*).

Functional impact of genetic variants:

The unique ability of flor strains to produce a biofilm suggests that several key molecular functions are modified in this group compared to the wine group. Genetic drift occurring relatively independently in the wine and flor groups should result in the random accumulation of mutations and

not in mutations enriched in functional categories. Because divergence-based analyses do not account for the synonymy of SNPs, we focused our analysis on nonsynonymous SNPs, differentiating wine and flor strains for which we could predict an impact on protein function. Almost one-third (2086) of the 6607 ORFs of wine and flor *S. cerevisiae* strains (HighQuality-SNP set) contained non-synonymous mutations that may impact their function. However, focusing on SNPs in the top 20% fraction differentiating wine and flor strains, we could detect an enrichment for 19 "biological process" and 15 "molecular functions" GO categories , including the "fungal cell wall organization", "signal transduction" and "zinc ion transport" categories (adj. P value = 0.008, 0.017 and 0.006, respectively) relevant in biological aging. In addition, these categories emcompassed 44.4% of the initial set containing 583 genes.

Signatures of positive selection in flor yeasts:

To identify genes potentially involved in adaptation to the biological aging environment, we used different methods aimed at detecting positive selection. We first carried out the McDonald-Kreitman test, which compares the ratio of nonsynonymous to synonymous polymorphisms of the flor groups to the ratio of nonsynonymous to synonymous divergence with *S. paradoxus*, the closest species to *S. cerevisiae*, and thus searches for positive selection on a wide time frame. When carrying out this test for the 561 genes with enough polymorphism in flor strains, none deviated from neutral expectations after correction for multiple tests.

We then searched for positive selection in the flor branch based on the ratio of non-synonymous to synonymous substitution rates ($\omega = d_N/d_S$). This approach is inaccurate for intrapopulation studies (Kryazhimskiy & Plotkin 2008), but it may be applied here given the isolation of wine and flor populations, which probably occurred after more than *Ne* generations according to our demographic inferences (Kryazhimskiy & Plotkin 2008). The branch-site test of positive selection was shown to be more robust and powerful than branch-based tests that average substitution rates over all codons (Yang & Dos Reis 2011). To reasonably limit the computational burden, the analysis was performed using a subset of 232 genes that were identified as highly divergent in analyses based on PCA or divergence metrics. After correction for multiple tests, positive selection was inferred for the flor

allele of *ZRT1* (adj. P value = 0.013) at the following three residues: S15, E16 and Q306 (Table 2). In addition, three genes with interesting functions displayed P values close to significance levels after correction for multiple tests: *HMS2* can restore pseudohyphal growth in a strain defective for ammonium transport (P value 0.005); *CCS1* (P value = 0.007), a chaperon for superoxide dismutase *SOD1,* is involved in oxidative stress protection; and *ADH6* (P value = 0.003) provides yeast resistance to hydroxymethyl furfural through its reduction into 5-hydroxymethylfurfuryl alcohol (Petersson *et al.* 2006) and might offer detoxification against the various aldehydes produced during biological aging.

For detecting genes under recent selective pressures, we used the Cross Population Composite Likelihood Ratio XP-CLR (Chen *et al.* 2010), which relies on the extent of multilocus genetic differentiation around a selected variant and was suggested recently to be one of the most sensitive methods for detecting recent selective sweeps (Vatsiou *et al.* 2016). We performed a genomic scan for selective sweeps in the Mediterranean flor strains group and the Jura/Hungary flor strains group, using the results of demographic inference to set significance thresholds (see above demographic inferences). Three genomic regions shared by the two groups of flor strains displayed a selective sweep signature (Table S3). These regions encompassed the high affinity glucose transporter gene *HXT7*; the gene *MST27,* which is involved in vesicle formation; and the gene *YPS6,* which codes for a putative GPI-anchored aspartic protease involved in cell wall growth and maintenance.

Phenotypic impact of genetic variation in candidate genes from flor strains:

To assess the importance of genes detected in analyses of divergence or positive selection for biological aging and to examine how they contribute to the adaptation of flor yeasts to their habitat, we performed various physiological assays. We focused on genes known to be associated with biofilm formation, osmotic pressure and zinc transport (*RGA2, SFL1, SLN1,* and *ZRT1*). We found an impact of allelic variation on the two genes (*SFL1* and *RGA2*) participating in the regulation of *FLO11* (Fig. 4). The presence of the flor allele of *SFL1* enhanced velum formation in comparison to a strain carrying a wine allele. Similarly, flor strains present a unique allele of *RGA2* that enhances biofilm formation on liquid media (Fig. 4) in an equivalent manner to a strain carrying an inactivated

allele of *RGA2*. Because Rga2 acts as an inactivator of Cdc42, a GTPase member of the MAPK cascade that regulates *FLO11* expression (Ueno 1999; Sopko *et al.* 2007), the flor allele of *RGA2* could present an attenuated activity in comparison to the wine allele. In addition, a frameshift in the *RGA2* sequence that leads to a premature stop–codon was observed in one Spanish flor strain.

The comparison of hemizygous strains carrying either the flor or wine allele of *SLN1* (Fig. 4) revealed differences in the growth of the velum produced by the two strains, but surprisingly, the growth was better for the strain carrying the wine allele under our test conditions. The two alleles of wine and flor strains differed by 25 amino acid residues (Fig. S8). Finally, introduction of the flor allele of *ZRT1* in a flor strain did not induce clear differences in velum growth (data not shown).

Discussion:

Flor strains represent a specific lineage:

We provide the first phylogenetic picture of flor strains obtained from genome-scale data. Our data show that the group of flor strains is divergent from wine strains and is much less diverse (mean nucleotide diversity represents 60% of that calculated for wine strains). Demographic inferences revealed different demographic histories for wine and flor populations. While the wine yeast population presented a clear pattern of demographic expansion (Almeida *et al.* 2015), flor yeast went through a recent bottleneck. The fact that changes in wine-making technology have led to the stricter protection of wines from oxygen, except for wines exposed to biological aging, is consistent with the inferred demographic histories. We also found that the group of strains related to the Champagne strain EC1118 and related strains (QA23, VIN13, MJ73; Borneman *et al.* 2016) results from a cross between flor and wine strains. Some of the mechanisms underlying the ability of flor strains to grow on nutritionally depleted media and cope with the ethanol stress of wine production may be helpful during the second fermentation of the “prise de mousse” step that imposes a second anaerobic growth on wine.

Genomic changes underlying adaptation to wine biological aging:

We identified 155 loci that are highly divergent between wine and flor yeasts based on different methods. Thirty-seven of the divergent genes also present non-synonymous potentially impacting substitutions and therefore may play a key role in adaptation to the biological aging conditions. Five of the 37 candidate genes (*IRA1*, *SFG1*, *HMS2, IME4* coding for a mRNA N6-adenosine methyltransferase and *PIK1* coding for a Phosphatidylinositol 4-kinase) were shown to impact pseudohyphal growth (Lorenz & Heitman 1998; Fujita *et al.* 2005; Cullen & Sprague 2012; Adhikari & Cullen 2015).

In addition to highly divergent genes, d_N/d_S analyses revealed signatures of repeated positive selection in *ZRT1* and possibly three other genes (*HMS2, ALD6*, and *CCS1)*. The cross-population composite likelihood test (XP-CLR), which is aimed at detecting more recent targets of selection, revealed signatures of selective sweeps in genomic regions surrounding *HXT7* and *YPS6*, two genes with functions of interest for biological aging (high affinity sugar transporter and protease).

Functional changes associated with adaptation to wine biological aging:

Adaptation to biofilm formation. When exposed to limited nutrients, *S. cerevisiae* can induce pseudohyphal growth biofilm or mat formation (Reynolds & Fink 2001; Karunanithi *et al.* 2012). These responses lead to the induction of *FLO11,* the main actor of cell-cell adhesion, and are mediated by the same main pathways, the MAPK and Ras/cAMP/PKA pathways and nutrient signalization pathways (Fig. 5). In agreement with this finding, the adhesin *FLO11* and the transcription factor *SFL1* (Reynolds & Fink 2001; Zara *et al.* 2005; Fidalgo *et al.* 2006) were differentiated between flor and wine lineages. The differentiation of non-coding RNA *ICR1* (Bumgarner *et al.* 2009) was only observed with D_{XY}, but one wine yeast (6320-A7) contains a flor allele, which may explain a slightly lower rank with other metrics. We demonstrated the role of allelic variation of the genes *RGA2* and *SFL1*, which both present deleterious mutations relieving *FLO11* from negative regulation. The role of *HMS2* and *SFG1* remains to be investigated. In addition, 26 other differentiated genes were found that are important for pseudohyphal growth or biofilm formation (Ryan *et al.* 2012), and other genes participating in that mechanism were found to carry

non-synonymous mutations. Allelic divergence was observed in the regulatory pathways associated with *FLO11* regulation, such as the Ras/cAMP/PKA signaling pathway and the MAP kinase signaling pathways (*IRA1 or GPB1*) (Fig. 5).

Adaptation to carbon sources. At the end of alcoholic fermentation, ethanol is the main carbon source, followed by glycerol and the traces of fructose that might remain. We thus expected signatures of adaptive divergence in genes involved in the use of non-glucose carbon sources. Our study reveals the fructophily of flor yeasts, attested by the presence of a fructophile allele of *HXT3* (Guillaume *et al.* 2007) and the presence of a full region C including the high-affinity transporter of fructose *FSY1* (Galeote *et al.* 2010), which may help cells to utilize the traces of fructose that are present in wine. The finding of a selective sweep signature around the *HXT6-7* region adjacent to *HXT3* is also consistent with adaptation to the limited availability of glucose. The exhaustion of glucose also implies active gluconeogenesis. A candidate for adaptive divergence is the gene *MDH2*, which displays a specific allele in the flor lineage and encodes a cytoplasmic malate dehydrogenase that catalyzes the interconversion of malate and oxaloacetate and plays a major role in gluconeogenesis during growth on two-carbon compounds.

Flor yeast and osmotic pressure.

Grapes contain a high content of sugars, but given the conversion of each molecule of glucose into 2 molecules of ethanol, the osmotic pressure in wine is higher than in grape must. Unlike wine yeasts, which have lost functional aquaporins and gained a higher fitness in the grape must environment (Will *et al.* 2010), flor strains contain a functional *AQY2* gene that has also been shown to be involved in the control of cell surface properties and impact the pseudohyphal growth (Furukawa *et al.* 2009). In addition to aquaporins, the homeostasis of osmotic pressure in yeast is ensured by the high osmolarity glycerol response (HOG pathway). *SLN1* functions as an osmosensor that permits the adjustment of intracellular glycerol concentration either after an hyper-osmotic stress via Ssk1 or in response to an hypo-osmotic shock via Skn7 (Saito & Posas 2012). Skn7 participates in the regulation of cell-wall integrity that is critical for growth under hypo-osmotic conditions. The slower velum growth measured for the hemizygote containing a flor allele in comparison to the wine allele indicates that the two alleles are not equivalent in terms of sensing and osmotic response. We cannot rule out

the existence of negative interactions between the flor *SLN1* allele and other genes of the wine background of the hemizygote leading to an apparent lower fitness under biological aging conditions. Interestingly, Kvitek and Sherlock observed that adaptation to a low-glucose culture leads to a high frequency of mutations in the Ras/cAMP/PKA pathway and the high-osmolarity glycerol (HOG) response pathway (Kvitek & Sherlock 2013).

Flor yeast and divalent metal transport. One of our most puzzling findings is the enrichment of divergent genomic regions in genes involved in zinc ion transmembrane transport and signatures of positive selection at *ZRT1* and possibly *CCS1*. Zinc is an essential nutrient for all organisms, playing a structural role in many proteins and being a catalytic component for over 300 enzymes such as superoxide dismutase *SOD1* (Vallee & Auld 1990). In *C. albicans,* zinc has an essential role in biofilm formation (Nobile *et al.* 2009), and *ZRT1* is involved in pathogenicity via its role in zinc absorption from the host (Citiulo *et al.* 2012). *ZRT1* has already been reported to present evidence of balancing selection in *S. cerevisiae* (Engle & Fay 2013) and the alleles of different origins present large variations. We found that the flor allele of *ZRT1* is much closer to the allele found in African palm wine yeasts than to the allele found in wine yeasts, while the topologies of the tree of neighboring genes *ADH4* and *FZF1* are in agreement with the global phylogeny (Fig. S2a). This phylogenetic structure suggests that the flor allele may have been introgressed from a distant group (African or US Oak lineages).

Conclusions

Comparative population genomics revealed the phylogenetic origin of flor yeasts and identified genomic regions putatively involved in the adaptive divergence between pelagic wine yeasts fermenting glucose and biofilm-forming flor yeasts utilizing ethanol. Candidate genomic regions involved in adaptation to biological aging conditions included multiple genes belonging to specific regulatory networks such as genes involved in cell-cell adhesion, zinc transport, hexose transport or signaling pathways. Our results suggest that the tuning of these regulatory networks is a major genomic signature of flor yeast domestication, highlighting the plurality of evolutionary changes underlying adaptation to biological aging conditions. More generally, the high differentiation and

limited gene flow observed between flor and wine lineages make this system an excellent model for studying the early stages of ecological speciation.

Acknowledgments

This work was supported by the grant AIP BioRessources (GENOWINE project) from the French National Institute for Agricultural Research (INRA). Anna Lisa Coi gratefully acknowledges the Sardinia Regional Government for the financial support of her PhD and postdoc scholarships (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).

The authors are also thankful to three anonymous reviewers and to the editor for their insightful comments which helped to improve the manuscript.

References:

Adhikari H, Cullen PJ (2015) Role of phosphatidylinositol phosphate signaling in the regulation of the filamentous-growth mitogen-activated protein kinase pathway. *Eukaryot Cell*, **14**, 427–440.

Akao T, Yashiro I, Hosoyama A *et al.* (2011) Whole-genome sequencing of sake yeast Saccharomyces cerevisiae Kyokai no. 7. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **18**, 423–34.

Akey JM, Ruhe AL, Akey DT *et al.* (2010) Tracking footprints of artificial selection in the dog genome. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 1160–5.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome research*, **12**, 1805–14.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–1664.

Almeida P, Barbosa R, Zalar P *et al.* (2015) A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular Ecology*, **24**, 5412–5427.

Amato R, Pinelli M, Monticelli A *et al.* (2009) Genome-wide scan for signatures of human population differentiation and their relationship with natural selection, functional pathways and diseases. *PloS one*, **4**, e7927.

Argueso JL, Carazzolle MF, Mieczkowski PA *et al.* (2009) Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production. *Genome research*, **19**, 2258–70.

Auton A, McVean G (2007) Recombination rate estimation in the presence of hotspots. *Genome Research*, **17**, 1219–1227.

Baker AE, Wang B, Bellora N *et al.* (2015) The genome sequence of Saccharomyces eubayanus and the domestication of lager-brewing yeasts. *Molecular biology and evolution*, 1–30.

Barth G, Gaillardin C (1996) Non-Conventional Yeasts in Biotechnology. A Handbook. In: (ed Wolf K), pp. 313–388. Springer, Berlin, Heidelberg, New York.

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, **57**, 289–300.

Borneman AR, Desany B a., Riches D *et al.* (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of Saccharomyces cerevisiae (G Sherlock, Ed,). *PLoS Genetics*, **7**, e1001287.

Borneman AR, Forgan AH, Kolouchova R, Fraser JA, Schmidt SA (2016) Whole Genome Comparison Reveals High Levels of Inbreeding and Strain Redundancy Across the Spectrum of Commercial Wine Strains of Saccharomyces cerevisiae. *G3: Genes|Genomes|Genetics*, **6**, g3.115.025692.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, **81**, 1084–97.

Brückner S, Mösch H-U (2011) Choosing the right lifestyle: adhesion and development in Saccharomyces cerevisiae. *FEMS microbiology reviews*, **36**, 25–58.

Budroni M, Giordano G, Pinna G, Farris GA (2000) A genetic study of natural flor strains of Saccharomyces cerevisiae isolated during biological ageing from Sardinian wines. *Journal of applied microbiology*, **89**, 657–662.

Bumgarner SL, Dowell RD, Grisafi P, Gifford DK, Fink GR (2009) Toggle involving cis-interfering noncoding RNAs controls variegated gene expression in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 18321–6.

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. *Genome research*, **20**, 393–402.

Citiulo F, Jacobsen ID, Miramón P *et al.* (2012) Candida albicans scavenges host zinc via Pra1 during endothelial invasion. *PLoS pathogens*, **8**, e1002777.

Coi AL, Legras J-L, Zara G, Dequin S, Budroni M (2016) A set of haploid strains available for genetic studies of Saccharomyces cerevisiae flor yeasts. *FEMS yeast research*, **underpress**, 1–26.

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**, 2163–71.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Cullen PJ, Sprague GF (2012) The Regulation of Filamentous Growth in Yeast. *Genetics*, **190**, 23–49.

Engel SR, Cherry JM (2013) The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple Saccharomyces cerevisiae strains at the Saccharomyces Genome Database. *Database : the journal of biological databases and curation*, **2013**, bat012.

Engle EK, Fay JC (2013) ZRT1 Harbors an Excess of Nonsynonymous Polymorphism and Shows Evidence of Balancing Selection in Saccharomyces cerevisiae. *G3 (Bethesda, Md.)*, **3**, 665–673.

Estreicher SK (2006) *Wine from neolithic times to the 21st century*. Algora publishing, New York, New York, USA.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–91.

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

Fidalgo M, Barrales RR, Ibeas JI, Jimenez J (2006) Adaptive evolution by mutations in the FLO11 gene. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11228–33.

Fujita A, Hiroko T, Hiroko F, Oka C (2005) Enhancement of superficial pseudohyphal growth by overexpression of the SFG1 gene in yeast Saccharomyces cerevisiae. *Gene*, **363**, 97–104.

Furukawa K, Sidoux-Walter F, Hohmann S (2009) Expression of the yeast aquaporin Aqy2 affects cell surface properties under the control of osmoregulatory and morphogenic signalling pathways. *Molecular microbiology*, **74**, 1272–86.

Galagan JE, Calvo SE, Borkovich KA *et al.* (2003) The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, **422**, 859–868.

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 (Reading, England)*, **156**, 3754–61.

Gallone B, Steensels J, Prahl T *et al.* (2016) Domestication and Divergence of Saccharomyces cerevisiae Beer Yeasts. *Cell*, **166**, 1397–1410.e16.

Gibbons JG, Rinker DC (2015) The genomics of microbial domestication in the fermented food environment. *Current Opinion in Genetics & Development*, **35**, 1–8.

Gitan RS, Luo H, Rodgers J, Broderius M, Eide D (1998) Zinc-induced inactivation of the yeast

ZRT1 zinc transporter occurs through endocytosis and vacuolar degradation. *Journal of Biological Chemistry*, **273**, 28617–28624.

Gladieux P, Ropars J, Badouin H *et al.* (2014) Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular ecology*, **23**, 753–73.

Guillaume C, Delobel P, Sablayrolles J-M, Blondin B (2007) Molecular basis of fructose utilization by the wine yeast Saccharomyces cerevisiae: a mutated HXT3 allele enhances fructose fermentation. *Applied and Environmental Microbiology*, **73**, 2432–9.

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

Ho WC, Zhang J (2014) The genotype-phenotype map of yeast complex traits: Basic parameters and the role of natural selection. *Molecular Biology and Evolution*, **31**, 1568–1580.

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

Ibeas JI, Jimenez J (1996) Genomic complexity and chromosomal rearrangements in wine-laboratory yeast hybrids. *Current genetics*, **30**, 410–6.

Ibeas JI, Jimenez J (1997) Mitochondrial DNA Loss Caused by Ethanol in Saccharomyces Flor yeasts. *Applied and Environmental Microbiology*, **63**, 7–12.

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics (Oxford, England)*, **27**, 3070–1.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC genetics*, **11**, 94.

Karunanithi S, Joshi J, Chavel C *et al.* (2012) Regulation of mat responses by a differentiation MAPK pathway in Saccharomyces cerevisiae. *PloS one*, **7**, e32294.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7:

Improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. *PLoS genetics*, **4**, e1000304.

Kurtz S, Phillippy A, Delcher AL *et al.* (2004) Versatile and open software for comparing large genomes. *Genome biology*, **5**, R12.

Kvitek DJ, Sherlock G (2013) Whole Genome, Whole Population Sequencing Reveals That Loss of Signaling Networks Is the Major Adaptive Strategy in a Constant Environment (J Zhang, Ed,). *PLoS Genetics*, **9**, e1003972.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC genomics*, **15**, 162.

Legras J-L, Erny C, Charpentier C (2014) Population Structure and Comparative Genome Hybridization of European Flor Yeast Reveal a Unique Group of Saccharomyces cerevisiae Strains with Few Gene Duplications in Their Genome. *PloS one*, **9**, e108089.

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**, 1754–60.

Li R, Zhu H, Ruan J *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, **20**, 265–72.

Liti G, Carter DM, Moses AM *et al.* (2009) Population genomics of domestic and wild yeasts. *Nature*, **458**, 337–41.

Lorenz MC, Heitman J (1998) The MEP2 ammonium permease regulates pseudohyphal differentiation in Saccharomyces cerevisiae. *EMBO Journal*, **17**, 1236–1247.

Magwene PM, Kayıkçı Ö, Granek J a. *et al.* (2011) Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1987–92.

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.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–303.

McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**.

Moradi MH, Nejati-Javaremi A, Moradi-Shahrbabak M, Dodds KG, McEwan JC (2012) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. *BMC genetics*, **13**, 10.

Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome research*, **11**, 863–74.

Nobile CJ, Nett JE, Hernday AD *et al.* (2009) Biofilm matrix regulation by Candida albicans Zap1. *PLoS Biology*, **7**.

Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **365**, 185–205.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS genetics*, **2**, e190.

Pérez-Ortín JE, Querol A, Puig S, Barrio E (2002) Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome research*, **12**, 1533–9.

Petersson A, Almeida JRM, Modig T *et al.* (2006) A 5-hydroxymethyl furfural reducing enzyme encoded by the Saccharomyces cerevisiae ADH6 gene conveys HMF tolerance. *Yeast (Chichester, England)*, **23**, 455–64.

Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: An Efficient Swiss

Army Knife for Population Genomic Analyses in R. *Molecular biology and evolution*, 1–8.

Rausch T, Zichner T, Schlattl A *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics (Oxford, England)*, **28**, i333–i339.

Reynolds TB, Fink GR (2001) Bakers' yeast, a model for fungal biofilm formation. *Science (New York, N.Y.)*, **291**, 878–81.

Romano P, Suzzi G (1993) Sulfur dioxide and wine micro- organisms. In: *Wine Microbiology and Biotechnology* (ed Fleet GH), pp. 373–393. Harwood Academic Publishers.

Ropars J, Rodríguez De La Vega RC, López-Villavicencio M *et al.* (2015) Adaptive horizontal gene transfers between multiple cheese-associated fungi. *Current Biology*, **25**, 2562–2569.

Ross-Ibarra J, Morrell PL, Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, **104 Suppl**, 8641–8.

Ryan O, Shapiro RS, Kurat CF *et al.* (2012) Global gene deletion analysis exploring yeast filamentous growth. *Science (New York, N.Y.)*, **337**, 1353–6.

Saito H, Posas F (2012) Response to hyperosmotic stress. *Genetics*, **192**, 289–318.

Scannell DR, Zill O a, Rokas A *et al.* (2011) The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. *G3 (Bethesda, Md.)*, **1**, 11–25.

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

Sopko R, Huang D, Smith JC, Figeys D, Andrews BJ (2007) Activation of the Cdc42p GTPase by cyclin-dependent protein kinases in budding yeast. *The EMBO journal*, **26**, 4487–4500.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Thornton K (2003) libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics*, **19**, 2325–2327.

Ueno Y (1999) RGA2, a putative Rho-type GTPase-activating protein, is regulated by the transcription factor Tec1 during filamentous growth of Saccharomyces cerevisiae. Massachusetts Institute of Technology.

Vallee BL, Auld DS (1990) Zinc coordination, function, and structure of zinc enzymes and other proteins. *Biochemistry*, **29**, 5647–5659.

Vatsiou AI, Bazin E, Gaggiotti OE (2016) Detection of selective sweeps in structured populations: a comparison of recent methods. *Molecular ecology*, **25**, 89–103.

Wei W, McCusker JH, Hyman RW *et al.* (2007) Genome sequencing and comparative analysis of Saccharomyces cerevisiae strain YJM789. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 12825–30.

Will JL, Kim HS, Clarke J *et al.* (2010) Incipient balancing selection through adaptive loss of aquaporins in natural Saccharomyces cerevisiae populations. *PLoS genetics*, **6**, e1000893.

Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586–1591.

Yang Z, Dos Reis M (2011) Statistical properties of the branch-site test of positive selection. *Molecular Biology and Evolution*, **28**, 1217–1228.

Zara S, Bakalinsky AT, Zara G *et al.* (2005) FLO11-based model for air-liquid interfacial biofilm formation by Saccharomyces cerevisiae. *Applied and Environmental Microbiology*, **71**, 2934–9.

Data accessibility

All new sequence reads, assemblies and variant files are available in the European Nucleotide Archive (ENA) and accessions are provided in Table S1. These datasets were deposited under the study accessions PRJEB6529, PRJEB7675, and PRJEB6586.
All other data files: including variant files in vcf format, SIFT analysis tables, table of all genes retained by each method, data for recombination rates, data for dadi demographic inferences, scripts for ms simulation and XPCLR calculation were made available from a Dryad Digital Repository. (http://datadryad.org/review?doi=doi:10.5061/dryad.j315n)

Figure captions:

Figure 1: Whole-genome genealogical relationships among flor and wine strains. A: Maximum-likelihood genealogy showing the relationships among the 9 flor strains (triangles) and 9 wine strains (circles) sequenced, in comparison to other groups of yeasts. The genealogy was inferred based on 427,587 SNPs with less than 20% missing. Flor strains are colored in green, wine strains are in bright green, sake and Asian strains are in dark blue, Oak strains are in blue-green, bread and beer strains are in orange, laboratory strains are in red, and African strains are in brown. Bootstrap values greater than 95 are indicated as red dots for the wine and the flor clusters. B: Analyses of population subdivision based on Discriminant Analysis of Principal Components (Jombart *et al.* 2010). C: Ancestry estimations from ADMIXTURE (Alexander *et al.* 2009) assuming K = 3 to K = 9 of ancestral clusters.

Figure 2: A- Principal Component Analysis (PCA) of genomic variation in flor yeasts, wine yeasts and yeasts from multiple origins. The 1st and 6th axes (PC1, PC2) represent 26% and 2.8% of the global variance, respectively. Wine strains are in dark green, flor strains in light green, and sake, palm wine and USA oak strains are in blue. **B-** Genome scan showing the regions that differentiate wine and flor yeasts. A: contribution of individual SNPs to the 6st axis of PCA per 500-bp segment. The blue dashed line indicates the 2% segment that has the highest contribution. C**-** Genome scan of absolute divergence (D_{XY}) between the wine and flor strains, (x1000 per kb), per 1 kb sliding window (500 bp step).

Figure 3: Venn diagram comparing the three sets of highly divergent genes between flor and wine yeasts, identified using Principal Component Analysis, absolute divergence (D_{XY}), and relative divergence (D_a).

Figure 4: Comparison of the impact of wine and flor alleles of different genes on biofilm formation. Strains were grown on YNB with 4% ethanol and evaluated after incubation at 23°C for 3, 5 and 7 days for SLN1, RGA2 and SFL1 respectively, in triplicates. The comparison was performed using the haploid strain P3-D5 mat alpha carrying the wine or flor allele of *RGA2* after allelic exchange, or in hemizygotes of P3-D5 MATα x K1-280-2B MATa zygote in which one copy of the gene has been inactivated for *SFL1* and *SLN1*. Errors bars correspond to one SD.

Figure 5: Schematic representation of the multiple regulatory circuits involved in *FLO11* regulation. Genes differentiated between wine and flor strains are indicated in red. Differentiated genes with unequal velum growth permitted by the two alleles are indicated in bold. Adapted from Brückner and Mösch (2011).

Supplementary figures

Figure S1: FINESTRUCTURE co-ancestry matrix and population structure of wine and flor strains using a set of 54351 SNPs phased with Beagle. Wine and flor yeasts are indicated by an arrow. EC1118 and related strains (59A and MJ73) have a median position because of the medium proportion of haplotype segments shared with flor or wine strains as indicated by the heatmap.

Figure S2: Genealogical relationships among haplotypes of Champagne strain EC1118 and flor and wine strains. A- Maximum-likelihood genealogy of *ZRT1* showing the position of the two alleles of EC1118 in relation with 9 flor strains and 9 wine strains (sequences obtained from the assemblies) and with strains of different origins (data from SGD). B- Maximum-likelihood genealogy of *SLN1*

showing the position of the two alleles of EC1118 in relation to 9 flor strains and 9 wine strains (sequences obtained from the assemblies) and to strains of different origins.

Figure S3: Linkage disequilibrium as a function of physical distance. The r^2 values were averaged in distance classes of 10 bp.

Figure S4: Genomic distribution of recombination rates in wine and flor yeasts. Recombination rates were inferred using INTERVAL in the LDhat package.

Figure S5: Joint allele frequency spectrum of the wine population and two geographic populations of flor yeasts. Data, models and residuals are presented using a heatmap. A- Best model inferred for the dataset including Spain and Italy flor yeasts, B- Best model inferred for the dataset including French and Hungarian flor yeasts. T: time of the lineage splitting, Mw-f migration: from wine yeasts into flor yeasts, Mf-w migration: from flor yeasts into wine yeasts, N: ratio between ancestral and actual population sizes, and Na: ratio between ancestral and actual population sizes during a population size change. LL = log likelihood of the model, AIC: Akaike information criterion. Best model is in bold.

Figure S6: Alignment of the nucleotide sequence of the *HXT3* alleles of wine and flor strains in comparison to the *HXT1* allele of wine strain 20B2. The two red bars indicate the region of the possible recombination between HXT3 and HXT1 in flor strains.

Figure S7: Genome scan of relative divergence (D_a) between the wine and flor strains (x1000 per kb) window.

Figure S8: Alignment of the protein sequence of wine and flor *SLN1* showing the multiple amino acid changes.

Table 1: Summary statistics of genomic variation within and between populations of flor and wine *S. cerevisiae* yeasts.

		Wine population	**Flor population**
Number of analyzed strains		8	9
Segregating sites [1]		34553	8712
Pi (x1000) [1]		0.942	0.278
Watterson theta (x1000) [1]		1.126	0.284
Tajima's D[2]		-0.884	-0.115
F_{IS} [2,3]		0.69	0.61
Outcrossing rate		0.18	0.24
Divergence between Wine and Flor populations (x1000)	D_{XY}[2]		1.31
Relative divergence between wine and flor populations (x1000)	D_a[2]		0.70
	Φst[2]		0.43

[1] Calculated for the whole genome

[2] Values averaged across chromosomes

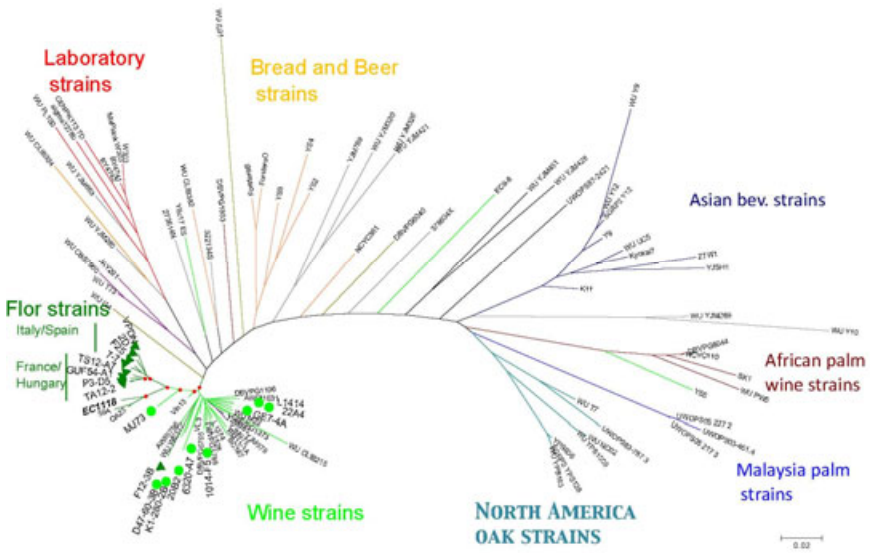
[3] F_{IS} estimates were obtained from diploid individuals

Table 2: Results of tests of positive selection based on non-synonymous to synonymous substitution rates. Significance after correction of multiple testing is indicated in bold. Nsites: number of sites.

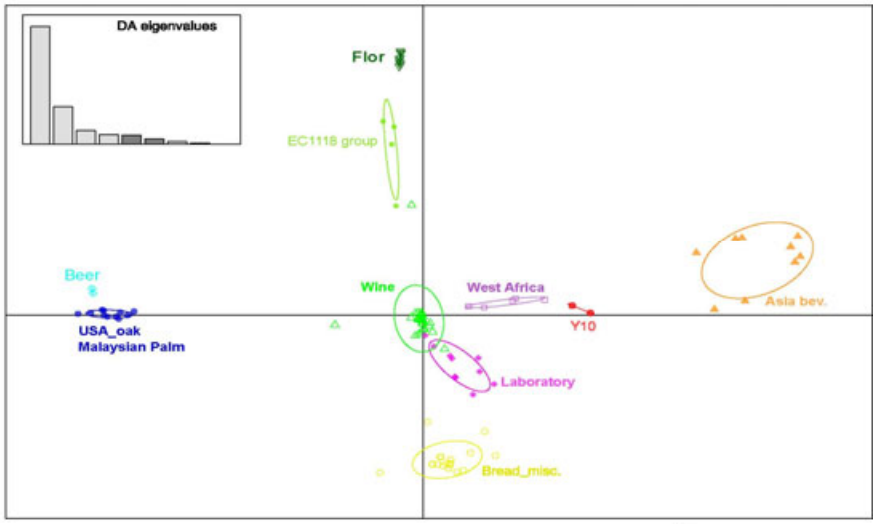
Gene	Gene Name	N. species in alignment	Nsite in the alignment	Likelihood Model A	Likelihood null Model	P value	adj. P value
YJR147W	*HMS2*	8	999	-4695.61	-4699.48	0.00542	0.261
YLR044C	*PDC1*	8	513	-3578.57	-3586.37	0.00008	0.013*
YMR038C	*CCS1*	8	669	-2480.15	-2483.82	0.00679	0.273
YMR318C	*ADH6*	8	813	-2993.65	-2998.14	0.00272	0.164
YGL255W	***ZRT1***	**7**	**1113**	**-3961.97**	**-3969.46**	**0.00011**	**0.013**

* : this pvalue is caused by the homologous recombination between orthologues genes *PDC1* and *PDC5* and cannot be considered as indicative of positive selection.

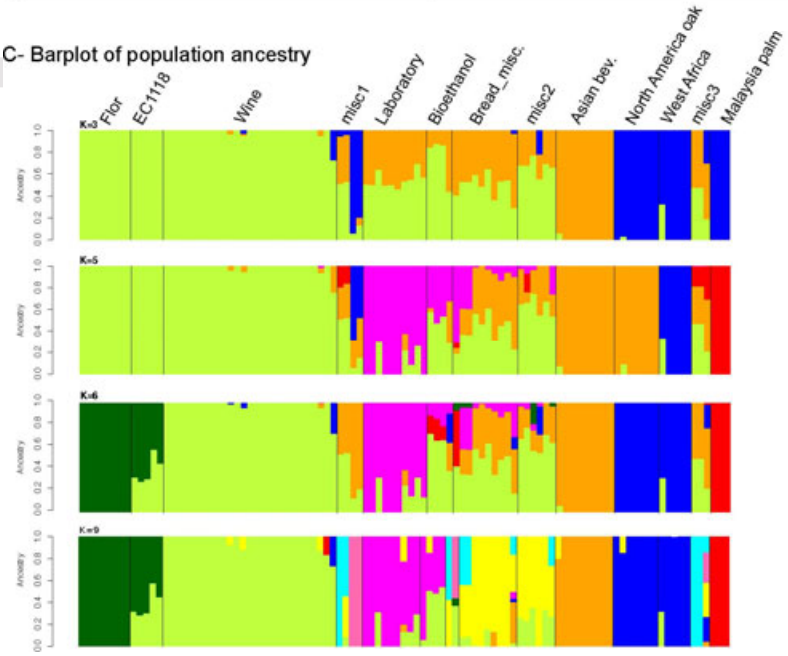
A- Global diversity of *S. cerevisiae*



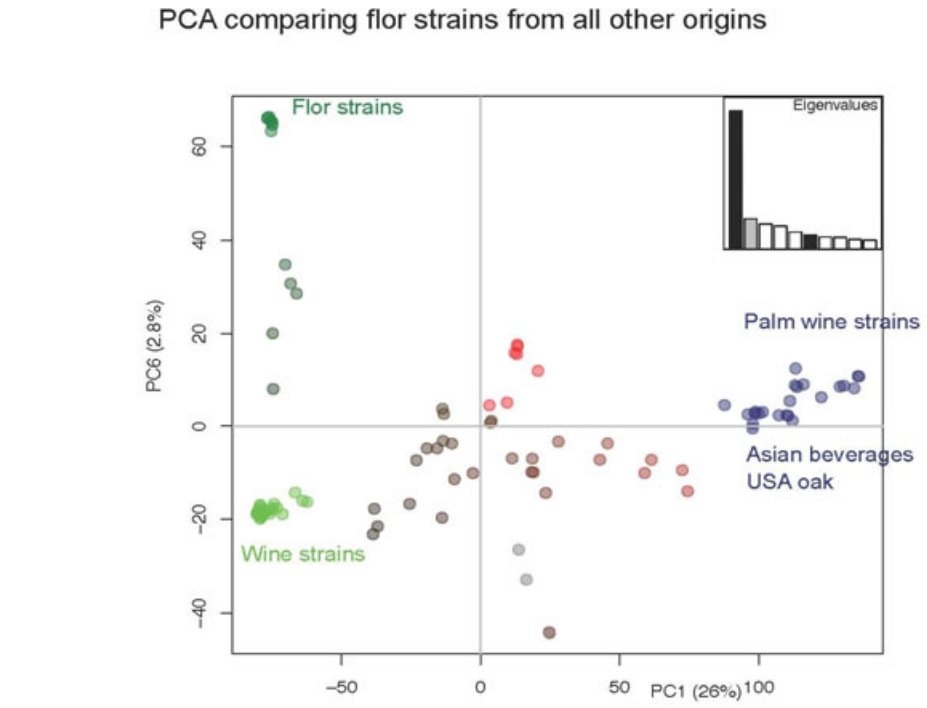
B- Population structure (DAPC axis 5 and 6: 4.7 and 2.7% of total variance)



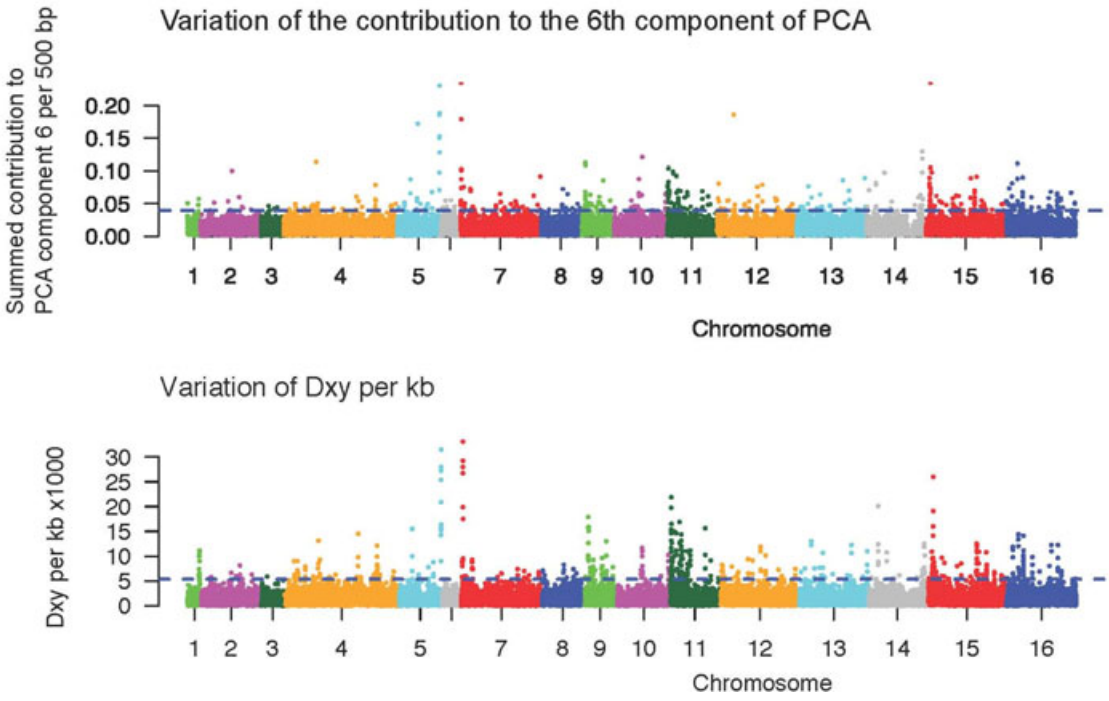
C- Barplot of population ancestry

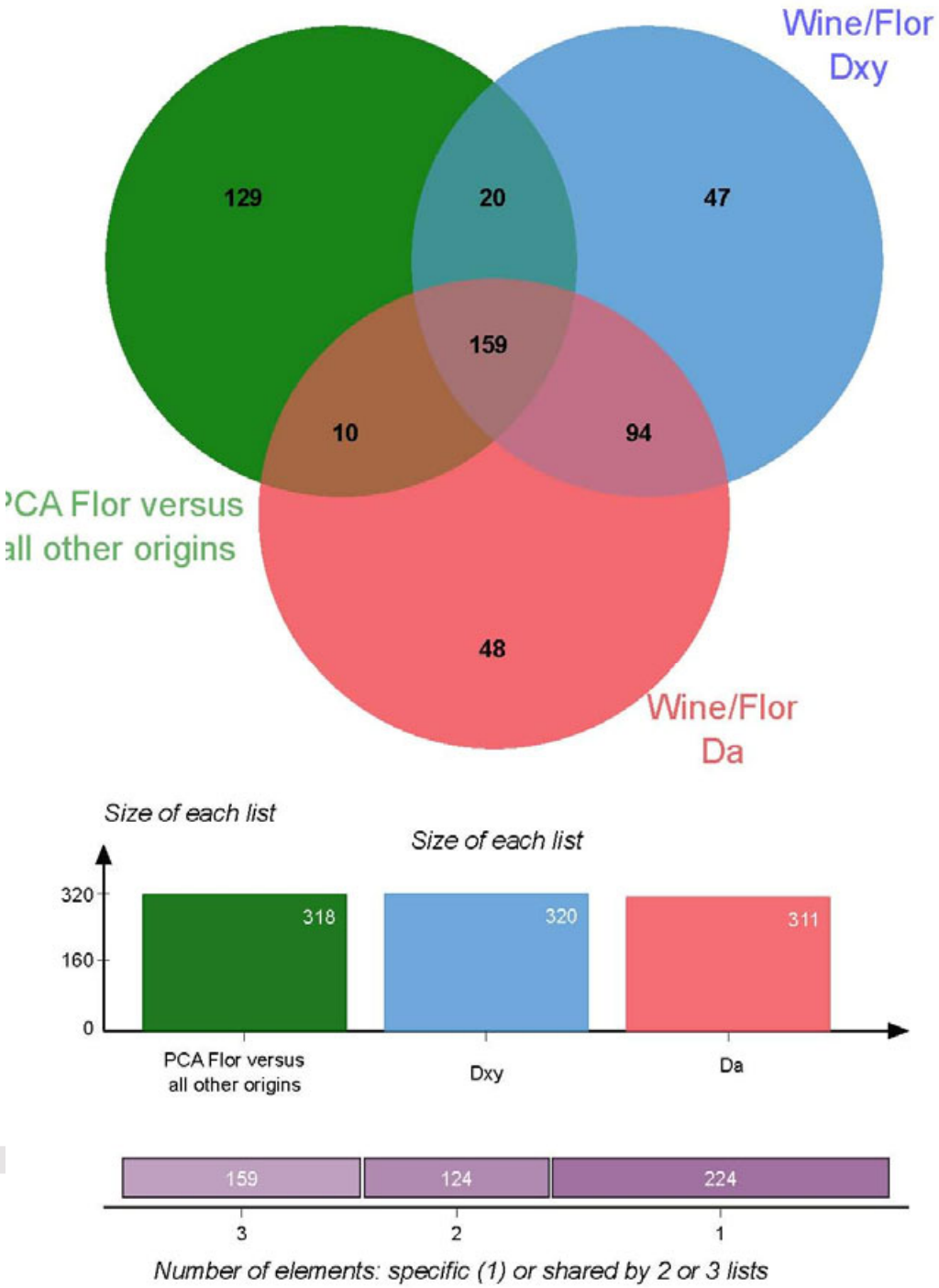


A-

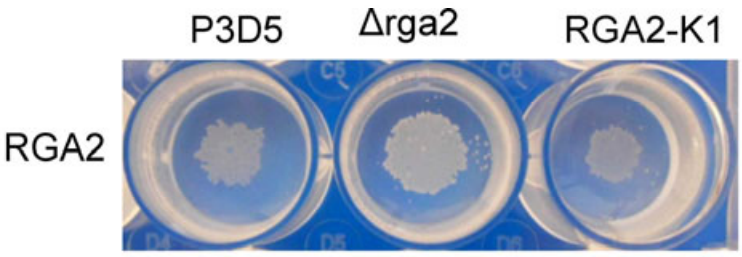


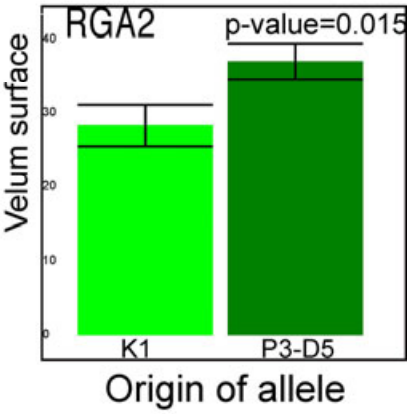
B-





P3-D5 background





P3-D5 x K1 background