Pedro Miguel Coelho de Almeida

Mestre em Microbiologia Aplicada

Microbe Domestication and the Identification of the Wild Genetic Stock of Wine Yeasts

Dissertação para obtenção do Grau de Doutor em Biologia

Orientador: Doutor José Paulo Sampaio, Prof. Associado, FCT/UNL
Co-orientador: Doctor Douda Bensasson, Assistant Prof., University of Georgia

Júri:

Presidente: Prof. Doutor Fernando Manuel Anjos Henriques
Arguentes: Profª. Doutora Maria João Marques Ferreira Sousa Moreira
Doutora Lília de Fátima Gomes Perfeito
Vogais: Prof. Doutor José Paulo Nunes de Sousa Sampaio
Prof. Doutor Manuel José de Carvalho Pimenta Malfeito Ferreira
Doutora Patrícia Filipa Homem de Campos Tavares de Brito

Outubro de 2016

Pedro Miguel Coelho de Almeida

Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa

Microbe Domestication and the Identification of the Wild Genetic Stock of Wine Yeasts

Copyright por Pedro Almeida, FCT/UNL e UNL:

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualguer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios cièntificos e de admitir a sua cópia e distribuição com objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Acknowledgments

I want to express sincere gratitude to my thesis advisors. To Prof. José Paulo Sampaio I am very thankful for the opportunity to take on this project with me. His guidance and thoughtful suggestions shaped the path to this thesis. Without his care and patience nothing of what is presented here would have been possible. To Prof. Douda Bensasson I want to thank for having received me and made me feel so comfortable in her laboratory while at the University of Manchester. I learned much during this period and her continued support and helpful guidance along these years were really important.

I would also like to thank Prof. Paula Gonçalves and Dr. Patrícia Brito, members of the thesis committee, for the knowledge transfer and helpful discussions that helped to improve this work.

To all my colleagues of the Yeast Genomics Lab (formerly Lab. 305…). A special thanks to Carla Gonçalves, Raquel Barbosa, Marco Coelho, Márcia Palma e Margarida Gonçalves for the valuable discussions and support in my scientific work.

To Fundação para a Ciência e Tecnologia thanks for financially supporting this thesis through the PhD scholarship SFRH/BD/77390/2011.

Last, but definitely not least, I thank to my family. I dedicate this thesis to my parents, whom unconditional love always supported my decisions, even if hard for them to accept. To Mara I want to thank you for always being there for me and for having fed me in these last months like if I was a baby. To my son Martim I thank you for fulfilling my days with love and tender; you are big in my heart and I would not have made it this far if it was not you. I love you all!

Resumo

A utilização de leveduras do género *Saccharomyces* na produção de alimentos e bebidas tem uma história milenar, promovendo involuntariamente a sua domesticação. No entanto, pouco se sabe sobre sua ecologia, distribuição e história natural. Esta tese teve como objetivo investigar a distribuição geográfica e estrutura populacional de novas linhagens selvagens e a sua relação com estirpes domesticadas do vinho em duas leveduras industrialmente importantes, *S. cerevisiae* e a espécie próxima criotolerante *S. uvarum*.

A utilização de uma colecção extensa e à escala global de novos isolados selvagens combinada com abordagens genómicas, permitiu mostrar que populações geográficas em *S. cerevisiae* e *S. uvarum* ocupam habitats naturais e têm uma história genética independente dos variantes industriais. Em ambas as espécies foram encontradas linhagens selvagens directamente relacionadas com estirpes domesticadas. No entanto, enquanto que em *S. cerevisiae* as estirpes domesticadas do vinho formam uma linhagem distinta dos seus putativos ancestrais selvagens dos carvalhos Mediterrâneos, em *S. uvarum* as estirpes do vinho e de cidra não puderam ser resolvidas dos seus parentes selvagens do Holárctico, nem ao nível filogenético nem populacional, o que indica diferentes vias na transição de selvagem para domesticado. Em ambos os casos, introgressões e regiões transferidas horizontalmente representaram as maiores descontinuidades entre estirpes do vinho e estirpes selvagens, e possivelmente representam impressões digitais de domesticação relevantes em leveduras do vinho de ambas as espécies. Adicionalmente, as estirpes do vinho e de carvalhos Mediterrâneos de *S. cerevisiae* também apresentaram uma extensa divergência nucleotídica em locais não-codificantes e em proteínas de ligação ao DNA, possivelmente como um subproduto de domesticação.

Os resultados apresentados aqui constituem uma primeira indicação molecular de domesticação em *S. uvarum* e também apontam para uma história populacional mais complexa do que se pensava em *S. cerevisiae*. No seu conjunto, esta investigação oferece novas ideias sobre os mecanismos gerais de adaptação em microrganismos eucariotas e ilumina a evolução de estirpes industriais modernas a partir de variantes selvagens.

Palavras-chave: *Saccharomyces cerevisiae*, *Saccharomyces uvarum*, genómica populacional de microrganismos, ecologia molecular de leveduras, domesticação de microrganismos, biogeografia de microrganismos

Abstract

For millennia that humans have used yeasts of the genus *Saccharomyces* in the production of foods and beverages, unwittingly fostering their domestication. However, little is known about their ecology, distribution and natural history. This thesis aimed to investigate the geographic distribution and population structure of new wild lineages and their relationship with domesticated wine stocks in two industrially important yeasts, *S. cerevisiae* and its cryotolerant relative *S. uvarum*.

Using an extended collection at a global scale of previously uncharacterised wild isolates combined with population genomic approaches, it is shown that geographic populations in *S. cerevisiae* and *S. uvarum* thrive in natural habitats and have genetic histories independent of the industrial variants. Wild lineages closely related to domesticated stocks were found in the two species. However, whereas in *S. cerevisiae* the domesticated wine strains form a distinct lineage from their putative wild ancestors of Mediterranean oaks, in *S. uvarum* wine and cider strains and their wild Holarctic relatives could not be resolved at the phylogenetic and population structure levels, indicating different routes for the transition from wild to domesticate. In both cases, introgressions and horizontally transferred regions represented major discontinuities between wine and wild strains, and are likely to denote relevant domestication fingerprints in wine yeasts of both species. In addition, wine and Mediterranean oaks strains of *S. cerevisiae* also showed extensive sequence divergence at non-coding sites and DNA binding proteins, which is interpreted as a by-product of domestication.

The results presented here constitute a first molecular indication of domestication in *S. uvarum* and also highlight a more complex population history in *S. cerevisiae* than previously thought. Altogether, these findings provide new insights into the general mechanisms of adaptation in eukaryotic microbes and illuminate the emergence of modern industrial microbial strains from wild variants.

Keywords: *Saccharomyces cerevisiae*, *Saccharomyces uvarum*, microbe population genomics, yeast molecular ecology, microbe domestication, microbe biogeography

Index

Acknowledgments i

Resumo iii

Abstract v

Index vii

Index of Figures xi

Index of Tables xiii

Abbreviations xv

CHAPTER 1 General Introduction 1

1.1 Microbe Domestication 3

1.1.1 Lactic Acid Bacteria (LAB) 4

1.1.2 Filamentous Fungi 5

1.2 *Saccharomyces* Species and Their Relevance for Food and Beverage Fermentations 5

1.2.1 The *Saccharomyces* Species Complex 6

1.2.2 Ecology 7

1.2.3 Reproduction and Life Cycle 8

1.2.4 Hybridization and Introgression Among *Saccharomyces* Yeasts 10

1.2.5 The Case of Domestication in Lager Beer Yeasts 11

1.3 *Saccharomyces* Wine Yeasts 11

1.3.1 The Success of *Saccharomyces* in Alcoholic Fermentations 12

1.3.2 Archaeological Evidence of Wine Yeast Domestication 13

1.3.3 Diversity and Natural History of Wine Yeasts 14

1.3.4 Phenotypic Evidence of Wine Yeast Domestication 15

1.3.5 Molecular Evidence of Wine Yeast Domestication 16

1.4 Tracing the Origins of Wine Yeasts 17

1.4.1 The Winery/Vineyard Hypothesis 17

1.4.2 The Wild vs. Domesticated Model 18

1.5 Aims and Structure of the Thesis 20

CHAPTER 2 A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum* 21

2.1 Summary 22

2.2 Introduction 22

2.3 Materials and Methods 22

2.3.1 Yeast isolation, identification and crosses 22
2.3.2 Quality filtering and alignment to reference genome 23
2.3.3 Phylogeny, population structure and polymorphism analyses 23
2.3.4 Bayesian coalescent analysis 25
2.3.5 Phylogeny of the genus *Saccharomyces* 25
2.3.6 Screening for foreign genes in the genomes of *S. uvarum* 26
2.3.7 Screening for introgressions from *S. eubayanus* 26
2.3.8 Gene Ontology (GO) analysis 26
2.3.9 *de-novo* assemblies and analysis of *FZF1* and *ZRT1* sequences 27
2.4 Results **27**
2.4.1 Occurrence of *S. uvarum* in the Southern Hemisphere 27
2.4.2 Population structure and phylogeography of *S. uvarum* in South America 30
2.4.3 Limited genetic diversity of the Holarctic population 31
2.4.4 Introgressions from other *Saccharomyces* species 33
2.5 Discussion **37**
CHAPTER 3 A population genomics insight into the Mediterranean origins of *Saccharomyces cerevisiae* wine yeasts **41**
3.1 Summary **42**
3.2 Introduction **42**
3.3 Materials and Methods **42**
3.3.1 Strain isolation, identification and typing 42
3.3.2 Microsatellite analysis 43
3.3.3 Genome sequencing, read alignment and genotype calling 43
3.3.4 Network, phylogeny and population structure 44
3.3.5 Polymorphism and divergence analyses 44
3.3.6 Demographic analyses 45
3.3.7 *de-novo* assemblies and survey for the horizontally transferred regions A, B and C 46
3.3.8 Multi-locus sequence analysis 46
3.4 Results **46**
3.4.1 A geographically and genetically diverse collection of wild isolates 46
3.4.2 Population structure and admixture 48
3.4.3 Genetic relationships among populations 50
3.4.4 The diversity of domesticated wine yeasts is equivalent to that of wild Mediterranean oaks yeasts 55
3.4.5 Population demographics 56

3.4.6 Domestication fingerprints ... 59
3.5 Discussion ... 60
CHAPTER 4 From Wild to Domesticated – Divergence and Adaptive Evolution Between Wine and Wild Yeasts ... 63
4.1 Summary ... 64
4.2 Introduction ... 64
4.3 Materials and Methods ... 65
4.3.1 Genome Sequencing, Read Alignment, and Genotype Calling ... 65
4.3.2 Phylogenetic Inference ... 65
4.3.3 Data Filtering ... 65
4.3.4 SNP Polymorphisms and Allele Frequency Differences ... 66
4.3.5 Genome Scans of Divergence and Polymorphism ... 66
4.3.6 Gene Ontology Analyses ... 67
4.3.7 Screening for introgressions from other Saccharomyces species ... 67
4.3.8 Analysis of *FZF1*, *SSU1* and *DTR1* introgressions ... 67
4.4 Results ... 68
4.4.1 Population Genomic Data ... 68
4.4.2 Distribution of Single Nucleotide Polymorphisms (SNPs) ... 69
4.4.3 Genomic Scans of Divergence Between Natural and Domesticated Yeast Populations ... 72
4.4.4 Divergence in *FZF1* is Caused by an Introgression in the Mediterranean Oaks Population ... 77
4.4.5 *SSU1*, a regulatory target for *FZF1*, is also introgressed in the MO population ... 81
4.4.6 European *S. paradoxus* and *S. cerevisiae* harbour two divergent copies of *SSU1* ... 81
4.5 Discussion ... 83
4.5.1 Divergence and gene flow ... 83
4.5.2 Patterns of Selection During Wine Yeast Domestication ... 84
4.5.3 Genomic Divergence Between Wine and Wild Yeasts ... 85
4.5.4 New roles for old genes? The case of *FZF1* and *SSU1* ... 86
4.5.5 Adaptation in wild and domesticated populations ... 87
CHAPTER 5 Concluding Remarks and Future Perspectives ... 89
5.1 Characterisation of the global diversity of wild lineages of *S. cerevisiae* and *S. uvarum* 90
5.2 Multiple Domestication trajectories in *Saccharomyces* ... 92
REFERENCES ... 95
APPENDIX I Supporting Information for Chapter 2 ... 109

APPENDIX II Supporting Information for Chapter 3 .... **123**

APPENDIX III Supporting Information for Chapter 4 .... **133**

DIGITAL RESOURCES **141**

Index of Figures

Figure 1.1 | Literature survey by year for terms related with domestication and microbe domestication 3
Figure 1.2 | Schematic cladogram of the genetic relationships among *Saccharomyces* species and their key industrial hybrids 7
Figure 1.3 | *S. cerevisiae* and *S. paradoxus* phylogenomics 14
Figure 2.1 | Geographic distribution, phylogeny and population structure of *Saccharomyces uvarum* .28
Figure 2.2 | Phylogenetic relationships between species and populations of *Saccharomyces* 29
Figure 2.3 | Sliding window analysis showing the mosaic nature of the genome of CRUB 1782, a South American strain of *S. uvarum* 31
Figure 2.4 | Proportion of shared and privately segregating polymorphisms and coalescence analysis in the South American and Holarctic (HOL) populations 32
Figure 2.5 | Sliding window analysis of *S. eubayanus* introgressions in *S. uvarum* genomes....34
Figure 3.1 | The global diversity of *Saccharomyces cerevisiae* is shaped by both ecology and geography 47
Figure 3.2 | Population structure and admixture in *S. cerevisiae*....48
Figure 3.3 | Co-ancestry and population clustering in *S. cerevisiae* 50
Figure 3.4 | Whole-genome phylogenetic relationships between *S. cerevisiae* strains 51
Figure 3.5 | Whole-genome phylogeny of the large strain dataset (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. 2009....52
Figure 3.6 | Multilocus phylogeny of *S. cerevisiae* including the Chinese lineages 53
Figure 3.7 | Joint allele frequency spectrum for the Wine and Mediterranean oaks populations of *S. cerevisiae* and comparison with the expected frequency spectrum for different demographic scenarios 57
Figure 3.8 | Competing models for the demographic scenario between wine and Mediterranean oaks populations incorporating an unsampled population 58
Figure 4.1 | Genetic relationships between *Saccharomyces cerevisiae* strains....68
Figure 4.2 | Distributions of F_{ST} and π_B values and corresponding Z transformations for all 5 kb windows (n=22486) 73
Figure 4.3 | Genomic landscape of divergence between Wine and Mediterranean oaks populations. (a) divergence was estimated in 5 kb sliding windows with 500 bp step....74
Figure 4.4 | Boxplots contrasting nucleotide diversity and Tajima's D statistic between divergence outliers and genomic background in wine and Mediterranean oaks (MO) populations....77
Figure 4.5 | Detection of the introgression of gene *FZF1* from *S. paradoxus* in the *S. cerevisiae* population associated with Mediterranean oaks....78
Figure 4.6 | Diagram illustrating the distribution of *FZF1* and *SSU1-DTR1* introgressions in the Mediterranean oaks (MO) population 79
Figure 4.7 | Sliding window analysis of *FZF1* nuceotide divergence at synonymous (Ks, dotted line) and replacement (Ka, filled line) sites in the wine strains relative to *S. paradoxus*....80

Figure 4.8 | Detection of the introgression of the region encompassing *DTR1-SSU1* from *S. paradoxus* in the *S. cerevisiae* population associated with Mediterranean oaks....83

Figure AI.1 | Phylogeny of the complete set of *Saccharomyces uvarum* strains used in this study.......119

Figure AI.2 | Genetic clusters of *S. uvarum* for K=2 to K=6 inferred with STRUCTURE and considering all but the Australasian sequences. 120

Figure AI.3 | 5′ Alignment of the promoter region of *FZF1* in *S. uvarum* and *S. eubayanus* 121

Figure AII.1 | Network analysis showing ecological and geographic separation of *Saccharomyces cerevisiae* lineages....131

Index of Tables

Table 2.1 | Whole-genome diversity values for *Saccharomyces uvarum* populations 30
Table 2.2 | Pairwise whole-genome divergence comparisons between *S. uvarum* populations and to *S. eubayanus* 30
Table 2.3 | Estimated contributions of *S. uvarum*, *S. eubayanus*, *S. cerevisiae*, and *S. kudriavzevii* to the complex genomes of two triple hybrid strains (CBS 2834 and CID 1), the *S. uvarum* x *S. cerevisiae* hybrid S6U, and to *S. bayanus* 35
Table 2.4 | Detection of horizontally transferred regions A, B and C in *S. uvarum*....37
Table 3.1 | Pairwise whole-genome divergence between populations of *Saccharomyces cerevisiae* 54
Table 3.2 | Proportion of shared polymorphisms, fixed differences and private polymorphisms in pairwise comparisons of the wine group with other populations of *S. cerevisiae* 54
Table 3.3 | Comparison of polymorphisms (mean values) at coding and non-coding regions between the Wine and Mediterranean oaks (MO) groups 55
Table 3.4 | Whole-genome diversity within populations of *S. cerevisiae*....56
Table 3.5 | Best-fit parameter estimates of three alternative demographic models for the joint allele frequency spectrum for the Wine and Mediterranean oaks populations....58
Table 3.6 | Best-fit population demographic parameters....59
Table 4.1 | Fixed differences and polymorphism in coding and non-coding DNA between Wine and Mediterranean oaks (MO) populations 69
Table 4.2 | Significant Gene Ontology (GO) terms in function and process categories for $\Delta AF > 0.8$ SNPs in replacement sites 70
Table 4.3 | List of ORFs within the candidate divergent regions between natural and domesticated populations of *Saccharomyces cerevisiae* 75
Table 4.4 | Summary of DNA variation in the coding sequences of the three genes where an introgression of *S. paradoxus* (Sp) was detected in the Mediterranean oaks (MO) lineage 80
Table 4.5 | Screening for the translocation of *SSU1*-R in wine and wild yeasts 82

Table AI.1 | List of strains used in Chapter 2 and relevant information pertaining to them 110
Table AI.2 | Carbon source assimilation and maximum growth temperature for *Saccharomyces uvarum* strains from different populations....114
Table AI.3 | Growth rates with glucose and melibiose as sole carbon and energy sources of an Australasian (ZP 964) and a Patagonian (CRUB 1595) strain of *S. uvarum* at different temperatures 114
Table AI.4 | Spore viability values from crosses of different lineages of *S. uvarum* and comparison with *S. eubayanus* 115
Table AI.5 | Complete list of *S. cerevisiae* and *S. kudriavzevii* introgressed genes into *S. uvarum* 116
Table AII.1 | List of *S. cerevisiae* strains used in Chapter 3 and relevant information pertaining to them 124
Table AIII.1 | List of strains used in Chapter 4 and relevant information pertaining to them....134

Table AIII.2 | Top 15 Yeast GO Slim Mapper results in Process category for the ORFs with two or more high differentiated (ΔAF > 0.8) SNPs in the 500 bp upstream sequence 136
Table AIII.3 | Top 15 Yeast GO Slim Mapper results in Process category for the ORFs with two or more high differentiated (ΔAF > 0.8) replacement SNPs 137
Table AIII.4 | Yeast GO Slim Mapper results in Process category for the ORFs within the candidate regions 138
Table AIII.5 | Yeast GO Slim Mapper results in Function category for the ORFs within the candidate regions 140

Dataset D1 | Complete list of *Saccharomyces eubayanus* introgressed genes into *S. uvarum* 141
Dataset D2 | Significant Gene Ontology (GO) terms for the introgressed *S. eubayanus* genes found in *S. uvarum*..141
Dataset D3 | Complete list of introgressed genes into *Saccharomyces cerevisiae* 141

Abbreviations

I .. XVI	chromosome names in the reference genome of *Saccharomyces cerevisiae*, strain S288C
AD	anno domini
AIC	Akaike information criterion
AUS	Australasia
BAM	binary SAM format
BC	before Christ
BP	before present
bp	nucleotide base-pair
CNV	copy-number variation
D1D2	domains 1 and 2 of the 26S rDNA
DNA	deoxyribonucleic acid
EM	expectation-maximization algorithm
EUR	Europe
FE	Far East
GO	gene ontology
GTR	general time reversible
HGT	horizontal gene transfer
HOL	Holarctic
ITS	internal transcribed spacer
kb	nucleotide kilobase-pair
LL	log likelihood
Mb	nucleotide megabase-pair
MCMC	Markov chain Monte Carlo
ML	maximum likelihood
MO	Mediterranean oaks
Mya	million years ago
NA	North America
NCBI	National Center for Biotechnology Information
nd	not determined
NJ	neighbour joining
NNI	nearest neighbour interchange
ORF	open reading frame
PCR	polymerase chain reaction
QTL	quantitative trait loci
rDNA	ribosomal DNA
RNA	ribonucleic acid
SA	South America
SAM	sequence alignment/map format
sd	standard deviation of the mean
SGD	Saccharomyces Genome Database
SGRP	Saccharomyces Genome Resequencing Project
SNP	single-nucleotide polymorphism
SPR	subtree pruning and regrafting
UCSC	University of California, Santa Cruz
VCF	variant call format
WGD	whole-genome duplication
ya	years ago

Abbreviations of Relevant Culture Collections

AWRI	Wine Microorganism Culture Collection of the Australian Wine Research Institute, Australia
CBS	Centraalbureau voor Schimmelcultures, The Netherlands
CRUB	Centro Regional Universitario Bariloche yeast culture collection, Argentina
DBVPG	Industrial Yeasts Collection, University of Perugia, Italy
EXF	Culture Collection of Extremophilic Fungi, Department of Biology, Biotechnical Faculty, University of Ljubljana, Slovenia
IFO	Institute for Fermentation, Japan (collection transferred to NBRC)
NBRC	National Institute of Technology and Evaluation Biological Resource Center, Japan
NCAIM	National Collection of Agricultural and Industrial Microorganisms, University of Horticulture and Food Industry, Hungary
NCYC	National Collection of Yeast Cultures, Institute of Food Research, UK
NRRL	Agriculture Research Service (ARS) Culture collection, Peoria, Illinois, USA
PYCC	Portuguese Yeast Culture Collection, UCIBIO, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
UCD	Phaff Yeast Culture Collection, University of California, Davis, USA
UWOPS	Culture collection of the University of Western Ontario, Department of Biology, Canada

Abbreviated species names

A. oryzae	*Aspergillus oryzae*
A. flavus	*Aspergillus flavus*
L. acidophilus	*Lactobacillus acidophilus*
P. camemberti	*Penicillium camemberti*
P. roqueforti	*Penicillium roqueforti*
S. arboricola	*Saccharomyces arboricola*
S. cerevisiae	*Saccharomyces cerevisiae*
S. eubayanus	*Saccharomyces eubayanus*
S. kudriavzevii	*Saccharomyces kudriavzevii*
S. mikatae	*Saccharomyces mikatae*
S. paradoxus	*Saccharomyces paradoxus*
S. uvarum	*Saccharomyces uvarum*

CHAPTER 1

General Introduction

The early domestication of plants and animals represents a critical development in modern human history. It was so important that it has been coined as the "Neolithic Revolution" because it marks a crucial transition in human lifestyle, from hunting and gathering to settled farming communities, triggering the raise of civilizations and the rise of technological innovation (Bar-Yosef 1998; Diamond 2002).

Darwin was one of the first scientists who became interested in the evolutionary perspectives of domestication. He documented several cases of phenotypic variation in domesticates compared with their wild ancestors, which were later used to support his theory of evolution (Darwin 1876). Since then, a large body of literature has accrued, detailing phenotypic as well as genetic modifications via human intervention in almost all domesticated animals and plants (Doebley et al. 2006; Larson & Fuller 2014; Wang et al. 2014). This knowledge provides a better understanding of the evolutionary mechanisms underlying changes in domesticated species, while it also helps to identify specific genes or loci controlling important traits associated with domestication. The genomic basis of domestication therefore represents an important model system for studying the genetic variation that shapes phenotypic diversity. This is directly relevant to improve existing crops and livestock and also to harness new domesticated species (Zeder 2015). In a more applied agricultural perspective, such advances can now be potentially translated into improvement programs targeting enhanced food security and sustainability (Womack 2005; Bevan & Uauy 2013; Ronald 2014).

Microbes were initially left behind in the studies of domestication even though they play important roles in the production, spoilage and safety of many food products. An illustrative example of this problem is depicted in Figure 1.1, in which a keyword search by "topic" on Web of Science using the terms "domestication AND (bacteria OR yeast OR fungi)" resulted in only 7 indexed publications by the turn of the century (year 2000). However, a breath of recent research is increasingly integrating the study of microbe domestication in the same evolutionary context as that of crop and livestock domestication, aiming to understand the genetic and functional variation driven by human selection over millennia on microbes (Legras et al. 2007; Siezen et al. 2008; Liti et al. 2009; Douglas & Klaenhammer 2010; Libkind et al. 2011; Gibbons et al. 2012; Cheeseman et al. 2014). Within this context, the work described in this thesis investigates the domestication of *Saccharomyces* wine yeasts. Focusing on *Saccharomyces cerevisiae* and *S. uvarum*, two lineages of *Saccharomyces* with remarkable value in the fermentation of wine and other alcoholic beverages, the dynamics and biogeography of wild and wine populations are analysed using complete genome sequence data. In addition, genomic variations potentially linked with the domestication process in either species are also explored.

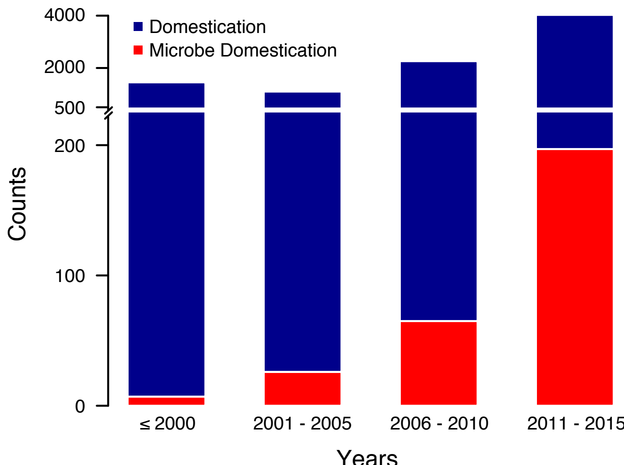


Figure 1.1 | Literature survey by year for terms related with domestication and microbe domestication. The search was made on the Web of Science server, accessed on 15 July 2016, using the keywords "domestication" (blue) and "domestication AND (bacteria OR yeast OR fungi)" (red) in the Topic field. The citation index on the Web of Science server is only available after 1970, and results are only presented until the end of 2015.

In this thesis, the term domestication will be used to define the man-driven changes in an organism, which make it genetically distinct from its wild ancestors in ways that it becomes more useful to humans - who, at least in plants and animals, control its reproduction. Largely derived from a genetic perspective of domestication (Diamond 2002), this definition summarises the direct relationship between humans and target species, without necessarily implying foresight and intentionality (Zeder 2015), and the consequences of this relationship. It is also broad enough to be applied to microorganisms, in which changes in behaviour are difficult, when not impossible, to quantify and the variation at the genetic level is usually the best tool to determine relationships between closely related taxa.

1.1 Microbe Domestication

The production of fermented foods and beverages can be traced back to the onset and expansion of agriculture (McGovern et al. 1996; Ross et al. 2002; Sicard & Legras 2011; Dietrich et al. 2012; Salque et al. 2013), being an ancient practice that is now widespread and can be found among almost every human culture. These fermentations depend on the activities of microorganisms which transform the substrate and inhibit or supress the growth or activity of other microbiota. The reasons for the early adoption of this technology are not entirely clear but it was probably harnessed unwittingly as a consequence for the need to accumulate and maintain increasing amounts of foodstuffs while avoiding their deterioration, increasing their digestibility and supplementing the own diet (Steinkraus 1996). Furthermore, the presence of ethanol and other metabolites in alcoholic beverages also contributed for important social habits of many civilisations (Joffe 1998; McGovern et al. 2013).

In contrast to the majority of fermentations practiced nowadays with controlled starter cultures, traditional fermentations rely on natural occurring microorganims in the fermentation substrate. Either through intentional back-slopping, a method that carries a small portion of a previously fermented batch to start a new batch, or by the simple persistence of these microbial "factories" in the processing environment, traditional artisanal fermentations practiced in many cultures resulted in the long-term,

continuous specialised adaptation and genetic differentiation of isolated microbial populations to specific environmental conditions (Gibbons & Rinker 2015). The best batches were probably propagated for longer as a route to the preservation of a good microbial inoculum (Steinkraus 1996). Intentional back-slopping practices are thought to have been already well established in ancient Egypt for baking and brewing (Spencer & Spencer 2009) and probably also in ancient winemaking practices (McGovern 2007). Thus, it is likely that these ancient practices were on the origin of increased selection for the wine, brewer's and baker's yeast, *S. cerevisiae*, leading this species to its status as today's prime domesticated microorganism. A hypothesis for the progression towards microbial domestication has been recently proposed (Gibbons & Rinker 2015). In this model, wild microbes better pre-adapted to the spontaneous fermentative environment dominate other microbial communities and the long-term passage of the "winners" through multiple rounds of fermentation promotes adaptation and specialisation, ultimately leading to the pure cultures of domesticated microbes used today in more controlled and stable fermentations. In microorganisms, the molecular mechanisms responsible for this progression towards domestication usually include, among others, novel mutations, hybridisation and horizontal gene transfer (HGT) (Douglas & Klaenhammer 2010; Gibbons & Rinker 2015).

The most widely used microorganisms in the food and beverage industries include prokaryotic lactic acid bacteria (LAB) and some eukaryotic filamentous fungi and yeasts (Douglas & Klaenhammer 2010). LAB and filamentous fungi will be briefly introduced in this section as a way of providing a general overview of the diversity of domesticated microbes. Yeast domestication, in particular the domestication of wine yeasts, is the focus of this thesis and therefore will be considered in more detail throughout the following sections.

1.1.1 Lactic Acid Bacteria (LAB)

LAB comprise a diverse ecological group of fermentative bacteria widely associated with animals, occurring naturally in the gastrointestinal tract and other mucosal surfaces, and in wine, milk and meat (Makarova & Koonin 2007). They are defined by their ability to obtain energy from the fermentation of hexoses that are converted in lactic acid. Industrially, LAB are mainly used in the fermentation of dairy products, such as cheese, yogurt and kefir, meat and vegetables. The discovery of ancient pottery sieves in Poland dated from 5200 – 4800 BC to process milk into cheese provides the earliest evidence of LAB utilization for food processing (Salque et al. 2013). The domestication of LAB had profound effects in their genomes with many strains showing genome decay, through pseudogenization and gene loss, and loss of biosynthetic capacity during the adaptation to specific fermentation niches (Makarova et al. 2006; O'Sullivan 2009; Douglas & Klaenhammer 2010). In contrast, the acquisition of new chromosomal elements by HGT and the gain of adaptive mutations are likely to promote key fermentation activities (reviewed in Douglas & Klaenhammer 2010) . Domesticated strains of LAB are distinct from strains not associated with fermentations. For example, a comparative genomic analysis of 100 *Lactobacillus rhamnosus* strains identified a LAB cluster adapted to stable nutrient-rich niches, such as dairy products, that was clearly distinct both at the genotype and phenotype level from strains isolated from ecologically variable environments (Douillard et al. 2013). Similarly, the genome sequencing of 34 isolates of the probiotic bacterium *L. acidophilus* revealed that commercial strains had almost identical

sequences and clustered independently from strains from other sources (Bull et al. 2014), probably reflecting genetic bottlenecks and artificial selection.

1.1.2 Filamentous Fungi

Filamentous fungi are highly valuable for the secretion of a vast number of enzymes and some saprophytic species play an important role in the food and beverage fermentation industry. The most extensively studied filamentous fungi at the genomic level have been *Penicillium* species, commonly associated with the production of some types of cheese and fermented meat, and species of *Aspergillus*, important in the production of miso, soy sauce and also in koji, which is used in the brewing of a variety of traditional alcoholic beverages in Japan. Koji is the saccharification of rice or barley, in which the starch-rich substrate is broken into simple sugars that can be used in subsequent fermentation stages by yeasts such as *S. cerevisiae* or bacteria. Koji prepared from rice is used as a substrate for sake fermentations. The domestication history of *A. oryzae* used in the brewing of sake was investigated recently, with phylogenetic and population genomic analyses supporting a single domestication of *A. oryzae* from the wild species *A. flavus* (Machida et al. 2008; Gibbons et al. 2012). The genetic diversity within *A. oryzae* is only about one fourth of that of *A. flavus* (Gibbons et al. 2012), therefore suggesting a strong genetic bottleneck and possibly strain selection.

Domestication had important effects in the genomes of filamentous fungi. A remarkable example of this is the loss of the ability to produce toxins in *A. oryzae*. While *A. flavus* is the producer of the potent carcinogen aflatoxin, the biosynthetic pathway for aflatoxin is downregulated in *A. oryzae*, rendering this species non toxigenic (Chang et al. 2005; Tominaga et al. 2006; Gibbons et al. 2012). Because *S. cerevisiae* is sensitive to aflatoxin (Keller-Seitz et al. 2004), it has been proposed that the non-toxicity of *A. oryzae* might have been selected by its impact on yeast survival (Gibbons et al. 2012). In contrast to this apparent loss of function in *Aspergillus*, *Penicillium* species used as starter cultures for cheese, namely *P. camemberti* and *P. roqueforti*, acquired recently through horizontal transfer a large 575 kb-long genomic block (Cheeseman et al. 2014). The donor genome of this transfer has not been identified yet but the region contains genes predicted to be involved in the regulation of spore production and in antimicrobial activities, suggesting that it may confer adaptive advantage in competition with other microorganisms (Cheeseman et al. 2014).

1.2 *Saccharomyces* Species and Their Relevance for Food and Beverage Fermentations

The yeasts belonging to the genus *Saccharomyces* are arguably the most important group of microorganisms for biotechnological applications (Johnson & Echavarri-Erasun 2011). They have been explored for millennia, even if inadvertently, for baking, brewing and winemaking, as well as in the fermentation of a myriad of other foods and beverages at a global scale. In particular, the yeast *S. cerevisiae* epitomises the concept of microbe domestication. The influence of *S. cerevisiae* in human societies is such that it has been extensively characterised at the biochemical, molecular and cellular levels, becoming the model organism of choice in almost every field of biological research. Hence, it is no surprise that *S. cerevisiae* became the first eukaryotic organism with its genome completely sequenced (Goffeau et al. 1996). *S. cerevisiae* nuclear genome is only about 12 Mb and contains 6604

almost entirely intron-less open reading frames (ORFs) distributed over 16 nuclear chromosomes. With a highly packed and small genome, fast generation times in the laboratory and powerful methods for genetic manipulation, new genome-wide technologies are usually developed first in *S. cerevisiae* before being applied to any other eukaryotic organism (Cherry et al. 2012). Furthermore, the ease of large scale genetic crosses and screening of hybrids is bringing *S. cerevisiae* to the forefront of quantitative genomic analyses (Ehrenreich et al. 2010; Liti & Louis 2012). With the availability of complete genome sequences for other species of the genus *Saccharomyces* (Liti et al. 2009; Scannell et al. 2011), there is now the opportunity to carry out detailed studies between species of existing and potential industrial relevance. Surprisingly, in spite of the advances in comparative genomics studies involving multiple *Saccharomyces* species, relatively little is known about the mechanisms that contributed to shape their within-species diversity and dynamics of natural populations.

1.2.1 The *Saccharomyces* Species Complex

Saccharomyces yeasts are single-celled fermentative ascomycetous fungi. The genus *Saccharomyces*, historically known as the *Saccharomyces sensu stricto* complex, is currently restricted to seven natural species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. arboricola*, *S. eubayanus* and *S. uvarum* (Hittinger 2013; Boynton & Greig 2014) (Figure 1.2). All *Saccharomyces* species are very similar morphologically and can only be differentiated for a few biochemical traits (Vaughan-Martini & Martini 2011; Warringer et al. 2011). The genus is thought to have originated 10 – 20 million years ago (Mya) (Kellis et al. 2003; Dujon 2006), approximately at the same time as the speciation of the great apes (Raaum et al. 2005; Scally et al. 2012), but sequence divergence between two of the most distant lineages (*S. cerevisiae* – *S. uvarum*) parallels that of human and chicken (Dujon 2006). Increased divergence between yeast species is possibly a consequence of shorter generation times in microorganisms, thus determining a higher number of accumulated mutations (Hittinger 2013). Yet, in contrast to mammals and birds, *Saccharomyces* genomes are largely collinear facilitating broad comparative studies to understand the evolution and the functional and regulatory architecture of the genome (Kellis et al. 2003; Dujon et al. 2004; Borneman et al. 2007; Scannell et al. 2011).

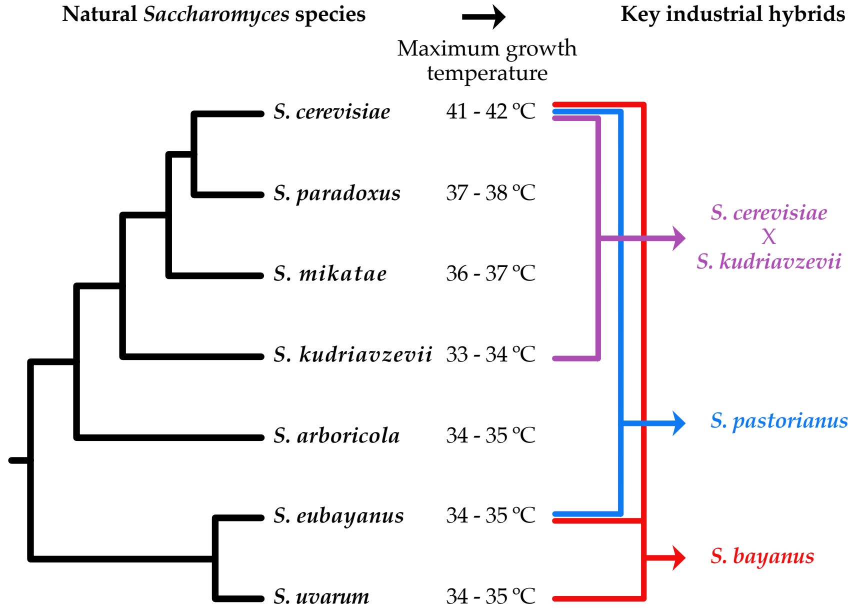


Figure 1.2 | Schematic cladogram of the genetic relationships among *Saccharomyces* species and their key industrial hybrids. The tree topology on the left and the genetic contributions of the hybrids on the right were adapted from Hittinger 2013. Maximum growth temperatures were obtained from Gonçalves et al. 2011 and Libkind et al. 2011. Note that the hybrids *S. pastorianus* and *S. bayanus* were given species names.

1.2.2 Ecology

S. cerevisiae was the first species of the genus to be described: "*Saccharomyces*" meaning sugar mould and "*cerevisiae*" meaning beer. This species is readily isolated from spontaneous and industrial fermentative environments worldwide and associated habitats such as vineyards (Vaughan-Martini & Martini 2011). Outside fermentations, it is found in the intestinal tract of *Drosophila* fruit flies (Phaff et al. 1956; Chandler et al. 2012; Buser et al. 2014), wasps (Stefanini et al. 2012) and birds (Francesca et al. 2012) and in association with various fruits (Naumov et al. 2006; Wang et al. 2012). Moreover, *S. cerevisiae* is also frequently isolated from arboreal habitats in the Northern Hemisphere, especially the bark and surrounding soil of oak trees (*Quercus* spp., Fagaceae family) (Naumov et al. 1998; Sniegowski et al. 2002; Sampaio & Gonçalves 2008; Wang et al. 2012; Hyma & Fay 2013), where it is often found along with *S. paradoxus* (Sniegowski et al. 2002; Sampaio & Gonçalves 2008). However, in contrast to *S. cerevisiae*, its sibling species *S. paradoxus* has only been isolated from natural environments – mostly associated with oak trees – and it is not associated with human fermentations (Replansky et al. 2008; Vaughan-Martini & Martini 2011). *S. mikatae*, *S. kudriavzevii* and *S. arboricola* are represented by only a few number of strains isolated in Eastern Asia, usually from decaying leaves and soil or the bark of Fagaceae trees (Naumov et al. 2000a; Wang & Bai 2008). The geographic distribution of *S. kudriavzevii* has been recently extended with the discovery of a new population on oak trees in Portugal, which is genetically and phenotypically divergent from the Japanese population (Sampaio & Gonçalves 2008; Hittinger et al. 2010). *S. uvarum* is commonly found in human fermentations or related habitats (Naumov et al. 2000b, 2001), sharing the same artificial ecological niches with *S. cerevisiae*. In the wild, it has been sporadically isolated from insects (*Mesophylax adopersus* and *Drosophila*), mushrooms, tree

exudates and oak trees in the Northern Hemisphere (Naumov et al. 2003, 2006; Sampaio & Gonçalves 2008). Interestingly, in the Southern Hemisphere, *Nothofagus* trees (southern beeches) occupy the same ecological niches as oak trees in the Holarctic. A recent ecological survey in Patagonia using an enrichment protocol with incubation at low temperature (10ºC) (Libkind et al. 2011), readily discovered two populations of *Saccharomyces* species in the bark and soil of *Nothofagus* spp. and in the stromata of *Cyttaria hariotii* (an obligate ascomycete parasite of *Nothofagus* spp.), *S. uvarum* and a new described species, *S. eubayanus*, suggesting that these species could be well established in this region.

The dispersal of *Saccharomyces* yeasts between regions and substrates is still poorly understood. It is known that these yeasts are not typically airborne (Mortimer & Polsinelli 1999), therefore relying on some sort of vector to travel. As noted above, *S. cerevisiae* and *S. uvarum* are occasionally isolated from insects, most often from *Drosophila* spp., and *S. paradoxus* has also been isolated from *Drosophila* (Naumov et al. 2000a). *Drosophila* fruit flies can transport yeasts in their body parts, such as in the abdomen (Christiaens et al. 2014), or in their gut (Reuter et al. 2007). Additionally, wasps can harbour *S. cerevisiae* cells during overwintering, suggesting that these insects, apart from the potential role as yeast vectors, can also act as a natural reservoir of yeasts during all seasons. Yeasts can also be disseminated by migratory birds and although living *S. cerevisiae* cells were shown to persist only for up to 12h after ingestion by birds this interval may be sufficient to cover long migration distances (Francesca et al. 2012).

Within *Saccharomyces*, *S. uvarum*, which is commonly associated with low-temperature wine and cider fermentations (Giudici et al. 1998; Naumov et al. 2000b, 2001), *S. kudriavzevii* and *S. eubayanus* are often considered cryotolerant species given their lower maximum growth temperatures (Sampaio & Gonçalves 2008; Gonçalves et al. 2011; Libkind et al. 2011; Salvadó et al. 2011) (Figure 1.2). On the other hand, *S. cerevisiae* and *S. paradoxus* tolerate higher growth temperatures than the former group and for this reason are generally considered thermotolerant (Sampaio & Gonçalves 2008; Gonçalves et al. 2011; Salvadó et al. 2011) (Figure 1.2). Interestingly, variation in temperature preferences might represent an important evolutionary trait allowing species pairs to frequently co-exist in temperate woodland forests in various geographic regions (Sweeney et al. 2004; Sampaio & Gonçalves 2008). The application of an isolation protocol with parallel enrichment at high and low temperature, 30ºC and 10ºC respectively, identified sympatric populations of thermotolerant and cryotolerant *Saccharomyces* species that can even share the bark of the same oak tree (Sampaio & Gonçalves 2008). Within species, temperature-dependent fitness components have also been shown to correlate with the distribution of *S. paradoxus* populations in North America (Leducq et al. 2014). These findings coupled with increased evolutionary rates in glycolytic genes between pairs of co-occurring species suggest that growth temperature profiles could be an important trait that is under divergent selection in the evolution and ecological speciation of *Saccharomyces* yeasts (Gonçalves et al. 2011).

1.2.3 Reproduction and Life Cycle

Saccharomyces yeasts can alternate between ploidy states, having a haplo-diploid life cycle. In the laboratory, these yeasts normally grow as diploids and most isolates collected from the environment are also diploid (Replansky et al. 2008; Cubillos et al. 2009). Growth is predominantly dominated by clonal reproduction but nitrogen starvation normally triggers meiosis which results in the formation of

an ascus with four resistant haploid spores (tetrad), two of each mating type (a and α). When suitable conditions are restored, ascospores germinate and, although they can continue to reproduce clonally as haploids via mitosis, opposite mating types will usually mate to form a new diploid vegetative cell (Herskowitz 1988). Mating often occurs within the ascus between haploids of the same meiosis product (intra-tetrad mating) but spores can also outcross with others from a different tetrad. Moreover, *Saccharomyces* isolates are usually homothallic, meaning that mitotic spores can switch mating types to mate with their own clonemates (haplo-selfing or autodiploidization), producing homozygous diploids. All three modes of reproduction have been documented indirectly in a wild population of *S. paradoxus* found in Europe (Johnson et al. 2004). Furthermore, population genomic analyses of two *S. paradoxus* wild populations estimated the occurrence of a sexual cycle (meiosis) approximately once in every 1000 asexual generations, with a frequency of matings of 94% from within the same tetrad, 5% of haplo-selfing and 1% outcrossed (Tsai et al. 2008). Outcrossing breaks linkage disequilibrium between linked genetic markers. Using whole-genome data, it has been estimated that linkage disequilibrium decays to half its maximum much faster in *S. cerevisiae* than in *S. paradoxus*, indicating more opportunities for mating and recombination in the former (Liti et al. 2009; Schacherer et al. 2009). It has been suggested that human association might create greater chances in *S. cerevisiae* for the movement of yeast strains, bringing them into proximity more often (Liti et al. 2009; Zhang et al. 2010), or for spore dispersal by insect vectors (Reuter et al. 2007). Indeed, digestion of sporulated *S. cerevisiae* yeasts by fruit flies is associated with an increased outbreeding rate (Reuter et al. 2007), and the intestine of social wasps has also been recently shown to favour its outcrossing probability (Stefanini et al. 2016). However, increased rates of outcrossing in *S. cerevisiae* might also be related with differences in life-history traits between species. For example, mate choice assays using strains isolated from sympatric natural populations of *S. cerevisiae* and *S. paradoxus* at a single woodland site in North America observed a marked difference in mating propensity between the two species, with *S. cerevisiae* showing a more promiscuous behaviour than *S. paradoxus* (Murphy et al. 2006).

Saccharomyces species are reproductively isolated from each other. Differences in germination times have been observed between *S. cerevisiae* and *S. paradoxus* (Murphy & Zeyl 2012), which may impose an important pre-zygotic isolation mechanism in nature. However, pre-zygotic isolation is likely to be a relatively weak barrier in *Saccharomyces* because species hybrids (or "alloploids") can readily form when no mate choice is given (Maclean & Greig 2008; Greig 2009). The most widely recognized model of speciation in *Saccharomyces* is explained by post-zygotic barriers: sequence divergence between species prevents the diverged chromosomes to recombine during meiosis of the F1 hybrids, generating highly aneuploid and non-viable offspring (Hunter et al. 1996; Greig et al. 2002, 2003). Hence, crosses between most *Saccharomyces* species usually result in less than 1% of viable spores (Naumov et al. 2000a; Liti et al. 2006; Wang & Bai 2008) demonstrating levels of reproductive isolation that meet the biological species concept. For the most closely related species within the genus, *S. uvarum* and *S. eubayanus*, the percentage of viable spores is about 7%, but it is still low enough to consider them separate species (Libkind et al. 2011). Either way, when alloploid hybrids are formed they can usually propagate by mitotic divisions as efficiently as the parental strains (Sipiczki 2008).

1.2.4 Hybridization and Introgression Among *Saccharomyces* Yeasts

Although natural *Saccharomyces* species other than *S. cerevisiae* and *S. uvarum* are rarely associated with human-driven fermentations, several are often found as contributors to interspecies hybrids in fermentative environments, such as beer, wine and cider (Masneuf et al. 1998; Le Jeune et al. 2007; Lopandic et al. 2007; González et al. 2008; Sipiczki 2008; Libkind et al. 2011) (Figure 1.2). In fact, the most economically important industrial yeast after *S. cerevisiae* is the allotetraploid hybrid *S. pastorianus* which is used to brew lager beer and combines the genomes of *S. cerevisiae* and the cryotolerant *S. eubayanus* (Libkind et al. 2011). Other hybrids harbouring genomic contributions from *S. cerevisiae* and a distinct cryotolerant species, such as *S. uvarum*, *S. eubayanus* or *S. kudriavzevii*, are commonly found among strains used to produce Belgian-style beers and wines fermented at low temperatures (Nguyen & Gaillardin 2005; González et al. 2006; Lopandic et al. 2007; González et al. 2008; Libkind et al. 2011; Nguyen et al. 2011; Erny et al. 2012) (Figure 1.2). Such hybrids often have superior combinations of properties than the parental strains, as shown by an increase tolerance to different stresses found in artificial environments (Belloch et al. 2008; Sipiczki 2008; Morales & Dujon 2012). For example, hybrids between *S. cerevisiae* and *S. kudriavzevii* inherited competitive traits from either parental species to tolerate better both high concentrations of ethanol and lower temperatures, respectively (Belloch et al. 2008). Some inherited traits can also contribute for a greater production of esters or fruity thiols, influencing the aromatic complexity of the final product (González et al. 2007; Lopandic et al. 2007; Swiegers et al. 2009).

Introgressions are the result of rare interspecies hybridization events followed by successive backcrossing with one of the parental species. In addition to hybrids, introgressions between *Saccharomyces* species are also frequently observed in fermentation associated strains (Doniger et al. 2008; Naumova et al. 2011; Dunn et al. 2012; Pérez-Través et al. 2014) and in clinical *S. cerevisiae* samples (Muller & McCusker 2009; Strope et al. 2015). Although hybrid genotypes are only rarely observed in natural strains (Liti et al. 2005), cases of introgression have been detected in natural *Saccharomyces* populations. For example, a subtelomeric segment of 23 kb from the left arm of chromosome XIV has introgressed from *S. cerevisiae* into European *S. paradoxus*, becoming fixed in this population (Liti et al. 2006). More recently, an ecological survey of *S. cerevisiae* from natural habitats in Brazil revealed multiple introgressions from *S. paradoxus* (Barbosa et al. 2016).

It is not known whether these hybridizations/introgressions indeed contributed to adaptation and increased fitness, or if the process is simply triggered by stressful environmental conditions (Morales & Dujon 2012). The introgression in some wine yeasts of *S. paradoxus* genes with relevant functions for the wine environment suggests a possible adaptive role or eventually industrially desirable qualities that could have been selected for. For example, the largest introgressed region shown to occur in four commercial wine yeast includes the *S. paradoxus SUC2* gene, which encodes a sucrose-hydrolyzing invertase, and also a gene similar to *S. cerevisiae HPF1*, a glucan alpha-1,4-glucosidase that can reduce protein haze formation in white wines (Dunn et al. 2012). It was also recently proposed that the multiple introgressions of *S. paradoxus* observed in Brazilian *S. cerevisiae* strains, that are enriched in genes coding for secondary transmembrane transporters, could have facilitated habitat shift and colonization of the tropical ecosystem (Barbosa et al. 2016). Furthermore, experimental evidence shows that hybrid genotypes in *Saccharomyces* tend to have greater environmental ranges relative to their parents (Stelkens et al. 2014), thereby possibly favouring adaptation to changing or new environmental niches.

1.2.5 The Case of Domestication in Lager Beer Yeasts

Lager beer yeasts are a classic example of highly successful interspecies hybrids driven by domestication. Originally from Bavaria, lager-type beers have been brewed since the 15th century (Meußdoerffer 2009). Lager yeasts, commonly referred to as *S. pastorianus* (synonym *S. carlsbergensis*), are specialised strains characterised by their low brewing temperatures. However, although it has been long known that *S. pastorianus* was a hybrid combining the genomes of a *S. cerevisiae* ale yeast with that of a cryotolerant *Saccharomyces* species (Dunn & Sherlock 2008), the identity of this unknown species was only recently clarified; strains of *S. eubayanus* collected in association with *Nothofagus* trees in Patagonia were shown to share ~99.5% sequence identity to the non-*S. cerevisiae* moiety of the *S. pastorianus* genome (Libkind et al. 2011) (Figure 1.2). Recent work has identified other *S. eubayanus* isolates in North America (Peris et al. 2014) and in China (Bing et al. 2014), with phylogenetic and sequence analyses suggesting that lager yeasts originated from East Asia rather than Patagonia (Bing et al. 2014). Lager yeasts consist of two distinct lineages, Saaz and Frohberg, and a comparative genomic study further suggested that hybridization between *S. cerevisiae* and *S. eubayanus* may have occurred independently at least twice generating the different lager lineages (Baker et al. 2015). Interestingly, non-hybrid *S. eubayanus* strains have not yet been found in Europe and are not known to be associated with human-controlled fermentations (Rodríguez et al. 2014; Gibson & Liti 2015).

1.3 *Saccharomyces* Wine Yeasts

Winemaking is the microbiological crafting of varied and complex wines from simple sugars present in the grape juice. In contrast to lager-type beers (see section 1.2.5), the production of wine over millennia is mostly associated with pure lineages of *Saccharomyces*. The primary yeast responsible for this process, *S. cerevisiae*, is very often referred to simply as "the wine yeast"; *S. cerevisiae* is the workhorse of wine production, and many other industrial processes. Spontaneous grape juice fermentations are consistently dominated by *S. cerevisiae* (Fleet 1998, 2003) and since the middle of the 20th century, strains with good oenological properties are available as starter cultures to ensure stable and reproducible fermentations (Borneman et al. 2016). When colonisation of the grape juice or dominance of the fermentation by *S. cerevisiae* does not happen, only low concentrations of ethanol are produced (Goddard 2008) and the grape must is rapidly dominated by spoilage microorganisms (Jolly et al. 2014).

Besides *S. cerevisiae*, the cryotolerant species *S. uvarum* is the only other natural, non-hybrid species of *Saccharomyces* with remarkable value for the wine industry. *S. uvarum* is less common than *S. cerevisiae* in wine fermentations, occurring predominantly in fermentations carried out at lower temperatures where it usually replaces the latter (Naumov et al. 2000b; Fernández-Espinar et al. 2003; Rementeria et al. 2003; Demuyter et al. 2004; Antunovics et al. 2005). It occurs frequently associated with sweet wines such as those produced in the Northern France regions of Alsace, Jurancon and Sauternes, Tokaj wines from Slovakia and Hungary, Valpolicella in Italy, and Yalta in Ukraine (reviewed in Sipiczki 2008). Both *S. cerevisiae* and *S. uvarum* are characterized by a set of traits that makes them well adapted to the oenological conditions, including growth on substrates with high sugar

and ethanol content, low pH and high sulfite concentrations (Sipiczki 2008). However, *S. uvarum* differs from *S. cerevisiae* in a number of oenological properties, with low ethanol tolerance at 24ºC and production of high levels of 2-phenylethanol and its acetate as discriminative technological traits distinguishing the two species (Masneuf-Pomarède et al. 2010). Moreover, the phenotypic differences between *S. uvarum* and *S. cerevisiae* have been associated with pronounced proteomic differences (Blein-Nicolas et al. 2013). Apart from wine, *S. uvarum* is also frequently associated with cider fermentations, which are usually performed at lower temperatures (Naumov et al. 2001; Valles et al. 2007). Nonetheless, evidence of domestication in *S. uvarum* is not yet clear and much of what is know about the molecular and cellular mechanisms behind wine fermentation have so far been learned from *S. cerevisiae*.

1.3.1 The Success of *Saccharomyces* in Alcoholic Fermentations

The success of *Saccharomyces* yeasts in alcoholic fermentations is notorious and surpasses that of any other microorganism. Grape musts naturally contain a rich microbial community with more than forty identified species of yeasts (Jolly et al. 2014). However, the highly selective ecological conditions during must fermentation lead to the succession of this natural microbiota with the rapid decrease in yeast diversity, leaving *Saccharomyces* as the dominant yeast to successfully complete the alcoholic fermentation (Fleet et al. 1984). The main characteristics behind this achievement have been attributed to unique traits that are unusual among other yeast genera: the ability to propagate anaerobically (Pronk et al. 1996; Møller et al. 2001), and the preference to degrade sugars, usually six-carbon carbohydrates such as glucose or fructose, into ethanol and CO_2 via fermentation independently of the presence of oxygen. This last property is called the "Crabtree effect" and is strongly enhanced by a glucose-repression system preventing respiration when sugars are present at sufficiently high concentrations (Johnston 1999; Piškur et al. 2006; Merico et al. 2007). The phylogenetic distribution of this trait among yeasts suggests that it evolved at about 100 Mya, approximately at the same time when the first fruit trees start to evolve, after a whole-genome duplication (WGD) leading to the *Saccharomyces* lineage and other closely related species (Conant & Wolfe 2007; Merico et al. 2007; Hagman & Piškur 2015). The emergence of glucose repression and ethanol accumulation capacity in the post-WGD yeasts appears to be associated with a global rewiring of the yeast transcriptional network with the concomitant loss of regulatory elements from genes involved in respiration (Ihmels et al. 2005) and to the expansion of hexose transporters (*HXT* genes) (Lin & Li 2011). Furthermore, upon depletion of glucose and in the presence of oxygen, the by-product of the fermentation, ethanol, becomes the substrate for aerobic respiration. This metabolic change (or diauxic shift) is known as the "make-accumulate-consume" strategy and restores a source for carbon assimilation from ethanol (Piškur et al. 2006). The "make-accumulate-consume" strategy works with the rapid consumption of glucose and subsequent conversion into ethanol. It has been suggested that this metabolic strategy could have evolved either to sequester the carbon source from competitors while also inhibiting their growth through ethanol production (Conant & Wolfe 2007; van Hoek & Hogeweg 2009; Dashko et al. 2014; Hagman & Piškur 2015) or because it can support faster growth (Pfeiffer & Morley 2014). These innovations are very old (~100 Mya) and occurred far earlier than any human-driven domestication event. The "make-accumulate-consume" strategy is shared by all *Saccharomyces* species and fitness

differences between *S. cerevisiae* and *S. paradoxus* under oenological conditions are mostly not significant (Williams et al. 2015), however *S. paradoxus* is not usually associated with wine fermentations. Thus, it still remains to be elucidated why only some of the species are chief contributors in human fermentations.

1.3.2 Archaeological Evidence of Wine Yeast Domestication

Winemaking was probably one of the first human uses of microbial fermentations because the process did not require the addition of an initial inoculum. The oldest archaeological and biomolecular evidence for a fermented beverage revealed a mixed fermentation made of rice, honey and fruit, dated back to 7000 BC from a Neolithic village in China's Yellow River valley (McGovern et al. 2004). The first evidence of grape wine was found in 5400 – 5000 BC ancient narrow-necked jars, presumably used for liquids, from Iran's Northern Zagros mountains in Mesopotamia (McGovern et al. 1996). The detection of substantial amounts of tartaric acid – which occurs in large amounts only in grapes – and terebinth (*Pistacia*) resin – widely used as a medicine and additive in antiquity – in the amphorae suggests that the fermentation of grapes was intentional. Further indication for the early production of wine in this region was found in southeastern Armenia from artefacts of the Late Chalcolithic dating to around 4000 BC (Barnard et al. 2011). Excavations in this site revealed a grape platform, possibly used to flow liquid into large ceramic jars, and the remains of crushed grapes and grape seeds, likely indicating that it could represent an ancient winery. The earliest evidence for winemaking in the eastern European Mediterranean is from a contemporary Neolithic site in Dikili Tash in northern Greece dated to 4500 – 4000 BC. Besides the chemical detection of tartaric acid in a pot, grape pressing practices were also evident at this site, thus suggesting that wine was already present in the Mediterranean region at the time (Valamoti et al. 2007; Valamoti 2015). These findings can have at least two interpretations regarding the beginnings of wine technology: either it was discovered independently in different areas along the Mediterranean, or it started in a single region and spread rapidly via networking between cultures and regions.

Whether these ancient wines were made from wild or cultivated grapes it is not known. Seeds of domesticated grapes have been discovered in Georgia and Turkey dating to about 8000 BC (This et al. 2006). Indeed, the domestication of grape seems tightly linked with winemaking and is believed to have started around the same geographic area – although it is unclear which process begun first (This et al. 2006). However, viticulture is not a precondition for wine making as wild grapes can naturally ferment, for example when crushed in harvesting pots. Moreover, the distribution of the modern wild grapevine (*Vitis vinifera* subsp. *sylvestris*) occupies a wide area across all the Mediterranean basin, from western Europe to the Trans-Caucasian zone, and wild and domesticated grape remains have both been recovered from Neolithic sites with confirmed wine jars (McGovern 2007).

From its probable origins in the primo-domestication area, viticulture and viniculture gradually disseminated, first to Egypt and Lower Mesopotamia (3500 – 3000 BC) and later into Europe (McGovern 2007). The oldest molecular evidence for the presence of *S. cerevisiae* in human-made fermentations has been obtained from a pottery jar from Egypt dated to 3150 BC (Cavalieri et al. 2003). This jar was found in the tomb of Scorpion I, one of the first kings of Egypt, and had been imported from Palestine and the Jordan Valley (Cavalieri et al. 2003). This discovery has major implications because it seems to indicate

that *S. cerevisiae* yeasts were already part of the microbial consortium utilised in ancient wine fermentations. The European expansion of wine occurred between 800 and 400 BC and progressed through west across the Mediterranean Sea probably following the Phoenicians, Etruscans and Celts (McGovern et al. 2013). By 500 BC wine was already being produced in the region now occupied by Portugal and Spain and in Northern Africa (Pretorius 2000). The spread to Northern Europe occurred only later (100 AD) with the Catholic Church, and only more than one thousand years later it did arrive to the Americas (1500 AD) (Pretorius 2000).

1.3.3 Diversity and Natural History of Wine Yeasts

The recent accumulation of complete genome sequence data and population genomic surveys in *S. cerevisiae* is starting to provide insights into its population structure and evolutionary history. Isolates collected from different ecological sources and geographic locations worldwide have been analysed at the whole-genome level using different molecular methods. These results support a complex global population structure with many strains showing a mosaic genome architecture involving only a few defined lineages that mainly reflect their technological or ecological origin (Figure 1.3a) (Liti et al. 2009; Schacherer et al. 2009; Cromie et al. 2013; Strope et al. 2015; Ludlow et al. 2016). The lineage commonly described as Wine/European contains primarily wine- and vineyard-associated strains from almost every continent but also includes a low number of European soil and clinical isolates (Liti et al. 2009; Schacherer et al. 2009). Similarly, strains used in sake fermentations form a unique lineage that also includes other isolates from Asian food products (Cromie et al. 2013). West Africa fermentation and Malaysia fruit strains have been identified as separate lineages too (Liti et al. 2009), although more strains with more diverse genotypes need to be analysed to strengthen this interpretation.

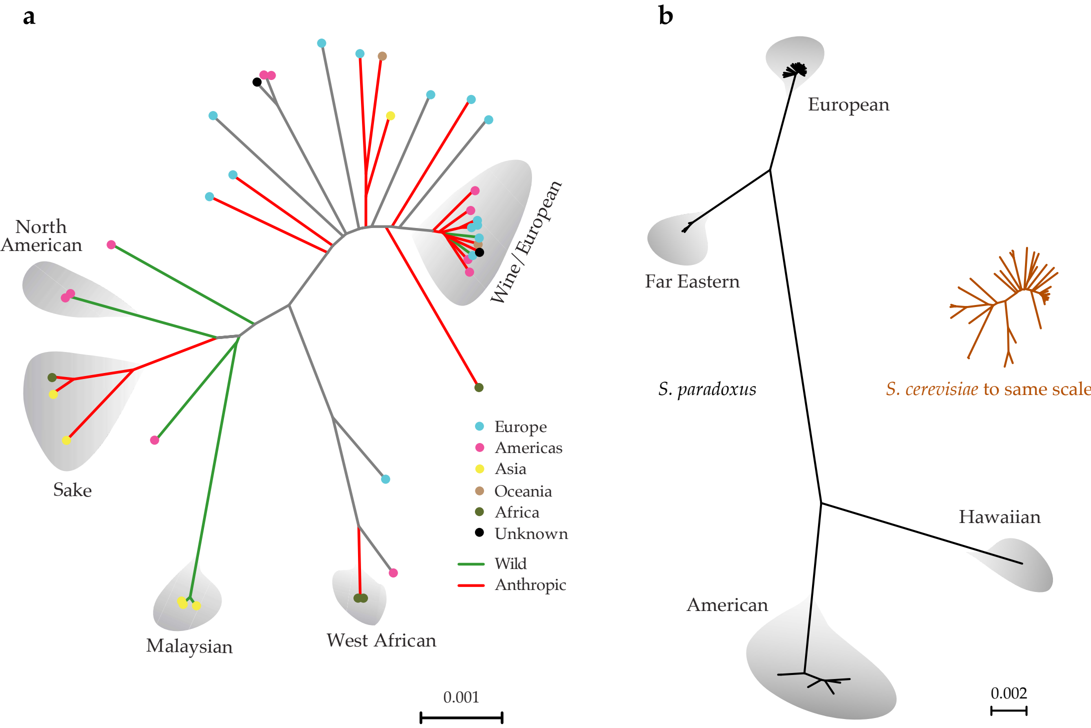


Figure 1.3 | *S. cerevisiae* and *S. paradoxus* phylogenomics. Whole-genome phylogenetic trees of **(a)** *S. cerevisiae*, depicting clean lineages highlighted in grey and admixed strains, and **(b)** *S. paradoxus*, with *S. cerevisiae* tree drawn at the same scale. Figures adapted from Liti et al. 2009.

Similar analyses in *S. paradoxus* showed a different pattern (Figure 1.3b) (Liti et al. 2009). *S. paradoxus* has a global distribution in which sequence divergence increases with increasing geographical distance at different geographical scales and, traditionally, it is considered a truly wild species that is not associated with anthropic environments (reviewed in Replansky et al. 2008). Furthermore, population structure in *S. paradoxus* is markedly shaped by geographical boundaries with independent lineages evolving in Europe, Far East and the North America continents (Figure 1.3b) (Koufopanou et al. 2006; Replansky et al. 2008; Liti et al. 2009). The complex genetic structure observed in *S. cerevisiae* therefore contrasts markedly with the one found in *S. paradoxus*, where in the latter genetic differentiation between populations is high and admixture between them is rarely observed (Liti et al. 2009). In fact, the lower diversity observed in *S. cerevisiae* seems to be unusual for the genus (Liti et al. 2009; Hittinger et al. 2010; Hittinger 2013; Peris et al. 2014). However, this low diversity could reflect the incomplete sampling of natural strains, as illustrated by the recent discovery of an unprecedented genetic diversity in natural Chinese populations (Wang et al. 2012). Thus, it is important to include additional strains collected from natural habitats across the world to improve our understanding of the population structure, biogeography and diversity of natural *S. cerevisiae* populations.

Concerning the domestication of the yeast *S. cerevisiae*, wine yeasts sampled around the world are genetically similar and clearly distinct from wild populations or other domesticated lineages, thus suggesting a single wine-related domestication event that occurred independently of other domestication processes in this species, such as for sake, beer or bread strains (Fay & Benavides 2005; Legras et al. 2007; 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 worldwide and, in spite of geographic heterogeneity, this group was found to have less genetic diversity than other groups assumed to represent wild populations, supporting the model of wine yeast domestication (Fay & Benavides 2005; Liti et al. 2009). Estimates of divergence time within the wine group place the diversification of the wine lineage posterior to the last glaciation era, at approximately 11000 BP indicating a relatively recent origin for the wine group (Fay & Benavides 2005; Legras et al. 2007). Incorporating natural yeast samples collected around the Mediterranean basin in molecular genetic studies could, in principle, shed light on the origins and timing of divergence between wine and wild populations.

1.3.4 Phenotypic Evidence of Wine Yeast Domestication

Directly or indirectly aided by human activity, *S. cerevisiae* wine yeasts evolved a series of adaptations in order to thrive in the stressful and changing environment of grape juice fermentations. Wine fermentation environments expose yeasts to extreme conditions such as low pH, high osmolarity, high sulfite levels, the depletion of important nutrients including nitrogen, lipids and vitamins, ethanol toxicity and competition and interaction with other microorganisms (Bauer & Pretorius 2000). Possibly as a consequence of these adaptations, *S. cerevisiae* wine yeasts are genetically and phenotypically distinct for several traits from other non-wine strains (Kvitek et al. 2008; Camarasa et al. 2011; Hyma et al. 2011; Warringer et al. 2011). At the phenotypic level, several wine yeasts have the unique ability to

utilize xylose as a carbon source (Wenger et al. 2010) and under winemaking fermentative conditions, life-history and metabolic traits show strong divergence between industrial and wild strains. Whereas industrial strains behave as "grasshoppers", reproducing slowly and reaching a small carrying capacity with a big cell size in fermentation, wild strains or "ants" reproduce quickly, reach a large carrying capacity and have a small cell size (Spor et al. 2009). Moreover, strains isolated from sugar-rich environments such as wine yeasts and strains collected from fruits have a better fermentative capacity than natural strains isolated from environments that are poor in sugar content, such as oaks (Camarasa et al. 2011). Wine strains also produce wine with fruitier and more floral characteristics than wild strains, suggesting the evolution of sensorial phenotypes relevant in wine production (Hyma et al. 2011).

1.3.5 Molecular Evidence of Wine Yeast Domestication

Wine yeasts also harbour important variations associated with relevant functions related to winemaking that have been mapped at the genomic level. The genetic variability of wine yeasts has been studied at many different levels, from nucleotide and transcriptional changes to changes in copy-number variation (CNV) of key genes and in ploidy level and also the detection of structural variations, introgressions and HGT (Cavalieri et al. 2000; Dunn et al. 2005; Novo et al. 2009; Rossouw et al. 2010, 2012). Such modifications in wine yeasts are potentially the result of the strong selective pressures that are present in wine environments and may serve as the main genetic adaptive signatures of wine yeast domestication (Querol et al. 2003). Many of these modifications however, such as CNVs, aneuploidies and other structural variations, are observed in only a minority of wine strains, which could indicate either adaptation to particular fermentative environments or transient intermediates to further stable adaptive mechanisms (Yona et al. 2012).

Examples of potential widespread adaptation among wine yeasts are the increase tolerance to sulfite (Park & Bakalinsky 2000; Pérez-Ortín et al. 2002) and copper (Fay et al. 2004; Warringer et al. 2011), two widely used chemicals in winemaking practices. Copper is present in older containers where fermentations took place and is also used as a fungicide to control powdery mildew (Fay et al. 2004). Wine yeasts and sake yeasts both developed copper tolerance associated with a CNV of *CUP1* gene (Fay et al. 2004; Warringer et al. 2011), which encodes for a copper-binding metallothionein, therefore indicating evolutionary convergence due to man-directed selection for industrial production (Warringer et al. 2011).

A well documented case of adaptation to winemaking is the ectopic translocation between chromosomes VIII and XVI in the promoter region of the *SSU1* gene, which encodes for a sulfite extruding pump involved in sulfite metabolism and detoxification. This translocation is only found in wine yeasts and generates a second *SSU1* allele, *SSU1-R*, that it has been shown to increase the transcription rates of *SSU1* and the degree of sulfite resistance (Goto-Yamamoto et al. 1998; Pérez-Ortín et al. 2002; Yuasa et al. 2004). A second translocation involving *SSU1* between chromosomes XV and XVI has also been identified recently in wine yeasts (Zimmer et al. 2014). However, although almost all wine yeasts possess at least one of these translocations, the XV-t-XVI translocation has only been observed in industrial starter strains, suggesting that it occurred more recently than the VIII-t-XVI translocation (Zimmer et al. 2014). Sulfite is widely used as an antioxidant and antimicrobial agent in

winemaking and therefore sulfite resistance is an important characteristic of wine yeasts. Sulfites have been used in wineries since the Egyptians (Pérez-Ortín et al. 2002), and this practice is likely to have led to the selection of strains with increase resistance. The effect of the most common *SSU1* translocation VIII-t-XVI is likely to be more profound as it may also serve as a post-zygotic genetic barrier reducing the spore viability between wine and oak strains (Hou et al. 2014; Clowers et al. 2015), and enforces ecological differentiation between these populations (Clowers et al. 2015). Sulfite resistance is controlled by *FZF1*, the transcriptional activator of *SSU1*. Both genes are highly polymorphic, with evidence of diversifying selection acting on *SSU1* coding sequence, possibly as an adaptation to the exposure of sulfur-based microbicides (Aa et al. 2006). Interestingly, divergence in sulfite resistance between *S. cerevisiae* and *S. paradoxus* has been linked to coding and non-coding nucleotide changes in *FZF1*, suggesting a major role of single transcription factors in driving both gene expression and phenotypic divergence between wine and wild yeasts (Engle & Fay 2012).

Another remarkable domestication fingerprint identified in a large number of wine associated yeasts, but not in other non-wine industrial yeasts, consists in the presence in the genomes of these yeasts of up to three genomic regions (named A, B and C) acquired independently from distant yeast species through horizontal gene transfer and containing a total of 39 genes (120 kb in total) with potentially relevant functions for the winemaking process such as in sugar and nitrogen metabolism (Novo et al. 2009; Galeote et al. 2010, 2011; Marsit et al. 2015; Borneman et al. 2016). The donor species of the region B as been identified to *Zygosaccharomyces bailii*, a major contaminant of wine fermentations (Novo et al. 2009; Borneman et al. 2011; Galeote et al. 2011). Recently, the region C has been identified to have originated from a recent transfer of a 158-kb genomic region from *Torulaspora microellipsoides*, a yeast species that is also present in the wine environment (Marsit et al. 2015). The region C has been characterised in more detail, and includes a high-affinity fructose/H^+ symporter, *FSY1*, possibly advantageous when fructose is the most abundant sugar at the end of the wine fermentation (Galeote et al. 2010; Marsit & Dequin 2015), and also two tandemly duplicated genes, *FOT1* and *FOT2*, that encode for oligopeptide transporters (Marsit et al. 2015). In competition experiments, FOT genes were shown to provide a strong competitive advantage in natural grape must fermentation (Marsit et al. 2015). These results highlight the possible evolutionary advantage of HGT genes in the genomes of *Saccharomyces* yeasts. Indeed, HGT events are known to represent a major source of innovation since these can readily contribute with genetic diversity, novel genes and new functions, eventually facilitating the adaptation to new environments (Wisecaver & Rokas 2015).

1.4 Tracing the Origins of Wine Yeasts

1.4.1 The Winery/Vineyard Hypothesis

People from the antiquity knew the uses of fermentation but, apart from that, nothing else about it. It was not before the seminal works of Louis Pasteur in the mid 19th century that *the* yeast (*S. cerevisiae*) was recognized as the responsible agent for the fermentative behaviour observed and used for millennia. Pasteur demonstrated that the fermentation of grape must, i.e. the transformation of sugars present in the fruit into alcohol, was a physiological process that depended of resident yeast cells on the grape surface (reviewed in Barnett 2000). Grounded on these discoveries and with the

development of pure culture methods, *S. cerevisiae* has been since then repeatedly isolated from diverse fermentative environments (Liti et al. 2009).

However, finding (i.e. isolating) *S. cerevisiae* outside the ecosystem of artificial fermentations is not a trivial task, even for experienced microbial ecologists; it is only rarely present in the surface of grape berries or vineyard soil, or on other wild habitats, and when present, its detection requires the use of enrichment culture methods meaning that it must be present at very low densities (Martini 1993; Martini et al. 1996). The abundance of *Saccharomyces* species from entire bunches of ripe vineyard grapes has been estimated recently using metagenomic technologies. In New Zealand vineyards it was found to comprise only about 1 *Saccharomyces* DNA molecule to about 20000 DNA molecules of other fungi, at least in that studied community (Taylor et al. 2014). In contrast, *S. cerevisiae* invariably dominates alcoholic fermentations and is almost the exclusive species colonising the cellar walls and equipment where it persists and reproduces between vintages (Martini 1993; Ciani et al. 2004; Mercado et al. 2007). These observations gave rise to the hypothesis that wine yeasts are strictly associated with the cellar environment in a closed cycle: they enter the fermentation from the cellar walls and equipment and at the end are transferred back to the cellar surfaces (Vaughan-Martini & Martini 1995; Ciani et al. 2004). This model therefore excludes a natural origin for *S. cerevisiae*. A series of experimental surveys in grape berries has found that while most intact grapes do not have wine yeasts – only about 0.1% is positive – damaged grape berries appear to be rich deposits, with about 24% probability of finding *S. cerevisiae* in them (Török et al. 1996; Mortimer & Polsinelli 1999). Moreover, the detection of different yeast strains in two consecutive years in the same vineyard pointed out to a flow of *S. cerevisiae* between the vineyard and the cellar (Török et al. 1996). These results instead indicated that yeasts exist in the vineyard and are transported to the winery with grapes, some of them certainly damaged. Insects that feed on damaged grape berries in the vineyard, would then carry the yeast cells from the cellars back to the vineyard (Mortimer & Polsinelli 1999). In either way, the common view about *S. cerevisiae* ecology was that the whole species was the product of domestication, represented by synanthropic populations almost only invariably found in close association with human activities (Naumov 1996).

1.4.2 The Wild vs. Domesticated Model

Recent ecological and biogeographic studies have provided evidence for a more complex situation and challenged the model that *S. cerevisiae* is strictly a domesticated organism. The consistent isolation of *S. cerevisiae* from the bark of oak trees and surrounding soil in many different regions of the world is providing convincing evidence for the persistence of this species in environments not associated with human activities (Naumov et al. 1998; Sniegowski et al. 2002; Sampaio & Gonçalves 2008; Wang et al. 2012; Hyma & Fay 2013; Robinson et al. 2016). Molecular studies with natural samples from North America and China have shown that these populations are genetically distinct from the wine/vineyard population and from other fermentation-related strains, hence suggesting independent evolutionary histories (Fay & Benavides 2005; Liti et al. 2009; Wang et al. 2012; Hyma & Fay 2013). Moreover, highly divergent wild lineages found in primary forests of China indicate that *S. cerevisiae* populations have differentiated in nature more than previously assumed (Wang et al. 2012). Together, these studies demonstrate that, as a species, *S. cerevisiae* is not domesticated. Because natural habitats are presumably older, the finding of wild *S. cerevisiae* strains further suggests that this species could exist in nature long

before it was explored for human-driven fermentations (Sampaio & Gonçalves 2008). Surprisingly, evolutionary and population genomic studies performed so far, have been strongly biased towards synanthropic yeast strains, lacking a comprehensive sampling at a global scale of wild isolates. Hence, it is unclear how much of the total genetic diversity has been sampled in natural habitats and how it is distributed between populations. Consequently, detailed knowledge about the genetic relationships between wild and domesticated lineages is lacking.

It was recently disputed whether *S. cerevisiae* is adapted to the oak environment, or any other habitat, or if instead it might not have a natural ecological niche (Goddard & Greig 2015). The reasons for this proposal were that *S. cerevisiae* has been isolated from multiple substrates, such as fruits, insects and clinical samples, and is only present at low densities in the oak substrate. However, as stated above for *S. cerevisiae* and further supported with ecological surveys for other species of the genus, there is now substantial evidence suggesting that the oak system – the bark and soil underneath oak trees – in the Northern Hemisphere are a preferential habitat for *Saccharomyces* yeasts, including *S. cerevisiae* (Naumov et al. 1998; Sniegowski et al. 2002; Johnson et al. 2004; Sampaio & Gonçalves 2008; Libkind et al. 2011; Hyma & Fay 2013; Charron et al. 2014; Peris et al. 2014; Sylvester et al. 2015; Robinson et al. 2016). While it is true that the range of environments where *S. cerevisiae* can be found is usually broader than for other species of the genus, such pan-habitats mostly reflect its association with human activities facilitating yeast dispersal, such as orchard fruits (Wang et al. 2012) or even as human commensal or pathogen (Angebault et al. 2013). Therefore, the presence of *S. cerevisiae* in many of these substrates is probably recent and derived from its anthropological association. In support of this view, *Saccharomyces* species are only rarely found in natural *Drosophila* populations (Chandler et al. 2012; Hoang et al. 2015), except when these are sampled in close proximity to vineyards, in which case *S. cerevisiae* can be recovered (Buser et al. 2014). Social wasps can carry diverse *S. cerevisiae* genotypes during all seasons (Stefanini et al. 2012). However, wasps feed on sugary fruits such as grapes and build their nests from chewed wood fibres and bark, which could represent a potential source of wild yeasts.

Although the number of *Saccharomyces* in oak bark might not be more than 2 cells per cm^2 (Kowallik et al. 2015), the best chance to isolate *Saccharomyces* species – including *S. cerevisiae* – in natural habitats is from the bark and associated soil of oak trees. When positive, the isolation frequency of *Saccharomyces* from oak trees can vary, but success rates as high as 70% (samples of *Q. pyrenaica* in Portugal) have been reported (Sampaio & Gonçalves 2008). More recent systematic isolation surveys in Northern Hemisphere forests provided further support for this association as have found a strong and significant association of *Saccharomyces* with oak trees; *Saccharomyces* yeasts are more likely to be found in oak trees than in any other tree species (Sylvester et al. 2015), and the probability of finding *Saccharomyces* yeasts is higher closer to the trunk of oak trees than further away (Kowallik & Greig 2016). These studies were based on the distribution of *S. paradoxus* but they are likely to apply to *S. cerevisiae* because, although no such studies exist for *S. cerevisiae*, the two species commonly have sympatric distributions in oak forests (Sniegowski et al. 2002; Sampaio & Gonçalves 2008). In one of the broadest sampling efforts performed to date, with 12 different species of trees and 7 different types of fruits sampled in many different locations in China – when considering substrates with more than 10 samples analysed – the bark of *Quercus* trees resulted in the highest frequency of *S. cerevisiae* isolates (61.1% for *Q. griffithii* and 44.4% for *Q. wutaishanica* versus 28.8% for pear which was the fruit with the highest yeast isolation frequency) (Wang et al. 2012). Importantly, the vast majority of forest isolates in

that study grouped in well defined, clean and divergent lineages, whereas most fruit isolates had mixed ancestries between lineages. In summary, although occasionally sampled from diverse substrates, oaks and possibly other trees of the Fagaceae are the best candidates for a primary reservoir of wild *Saccharomyces* yeasts in the Northern Hemisphere. The detection of simple sugars on the bark of oak trees that harbour *Saccharomyces* suggests that the oak system could support active growth of *Saccharomyces* populations, thereby providing a constant habitat all year round less exposed to seasonal fluctuations (Sampaio & Gonçalves 2008).

1.5 Aims and Structure of the Thesis

The overall objective of the research presented in this thesis is to transpose to the study of a group of biotechnologically relevant microorganisms, the yeasts of the genus *Saccharomyces*, the conceptual approaches used to understand crop and livestock domestication. More specifically, the work described here aims to identify the wild genetic stock(s) that can be associated with wine yeast domestication and also to unveil possible evolutionary processes and genetic changes linked with the domestication in *Saccharomyces* wine yeasts. In addition, this work also aims to contribute for a better understanding of the natural ecology and biogeography that have remained obscure in these important microbial model organisms.

Apart from *S. cerevisiae*, *S. uvarum* is the only non-hybrid lineage of *Saccharomyces* with remarkable value for the wine industry. **Chapter 2** describes a global phylogeographic survey in *S. uvarum*, uncovering new geographical ranges and previously unknown genetic diversity in natural populations of this species. Furthermore, comparative genomic analyses of wild and human-associated strains, combined with data from other species of the genus, identifies, for the first time, molecular signatures of domestication in *S. uvarum*. Chapter 3 and Chapter 4 focuses on *S. cerevisiae*. In **Chapter 3**, using a geographical diverse collection of new oak-associated wild isolates and human associated strains, it is shown that the global diversity of *S. cerevisiae* is shaped by both ecology and geography. Genome-wide population analyses further support a novel identified Mediterranean oaks population as the closest relative of domesticated wine yeasts. **Chapter 4** characterises the degree and distribution of nucleotide variation at the whole-genome level between wine yeasts and Mediterranean oaks strains, providing new insights into the general mechanisms of adaptation in eukaryotic microbes and elucidating the transition from wild to anthropic environments. Finally, **Chapter 5** draws general conclusions about the research described in this thesis and also highlights future perspectives and potential lines of research to build upon those presented here.

CHAPTER 2

A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum*

The work presented in this chapter is published in:

Almeida P[&], Gonçalves C[&], Teixeira S, Libkind D, Bontrager M, Masneuf-Pomarède I, Albertin W, Durrens P, Sherman DJ, Marullo P, Hittinger CT, Gonçalves P, Sampaio JP: **A Gondwanan imprint on global diversity and domestication of wine and cider yeast Saccharomyces uvarum.** *Nature Communications* 2014, **5**:4044. DOI: 10.1038/ncomms5044

[&] equally contributing authors

The sequencing data generated in this study have been deposited in the National Center for Biotechnology Information short-read archive (NCBI-SRA) under accession codes PRJNA230139 and PRJEB5133.

Contributions:

Pedro Almeida performed DNA isolation and genomic library preparations together with Carla Gonçalves.
Pedro Almeida developed all scripts and bioinformatics pipelines for the analyses of whole-genome sequence data.
Pedro Almeida executed all whole-genome sequence data analyses with help from Carla Gonçalves.
The contributions of the various authors were as follows:
C.G., C.T.H., D.J.S., D.L., I.M.-P., J.P.S., P.A., P.G. and P.M. conceived and designed the research;
C.T.H., D.L. and J.P.S. isolated new yeast strains;
C.G., D.L., M.B., P.A., P.A., P.D., S.T. and W.A. performed DNA extractions and genomic library preparations;
C.G. and S.T. performed the phenotypic tests and crosses;
C.G. performed sequence analysis of *FZF1* and *ZRT1* genes;
C.G., C.T.H., P.A., P.G. and J.P.S. wrote the manuscript.

2.1 Summary

In addition to *Saccharomyces cerevisiae*, the cryotolerant yeast species *S. uvarum* is also used for wine and cider fermentation but nothing is known about its natural history. Here, a population genomics approach was used to investigate its global phylogeography and domestication fingerprints using a collection of isolates obtained from fermented beverages and from natural environments on five continents. South American isolates contain more genetic diversity than that found in the Northern Hemisphere. Moreover, coalescence analyses suggest that a Patagonian sub-population gave rise to the Holarctic population through a recent bottleneck. Holarctic strains display multiple introgressions from other *Saccharomyces* species with those from *S. eubayanus* being prevalent in European strains associated with human-driven fermentations. These introgressions are absent in the large majority of wild strains and gene ontology analyses indicate that several gene categories relevant for wine fermentation are overrepresented. Such findings constitute a first indication of domestication in *S. uvarum*.

2.2 Introduction

Little information is available on the distribution of *S. uvarum*, and virtually nothing is known concerning possible domestication events. Apart from the appreciable number of strains that have been isolated in Europe from fermentations, only a few strains from natural environments have been obtained in Europe and North America, mostly in association with oak trees (Sampaio & Gonçalves 2008).

Interestingly, a radically different situation was encountered in South America. After extensive fieldwork on *Nothofagus* (southern beech), a tree genus that, along with *Quercus* (oaks), belongs to the order Fagales, a conspicuous population of *S. uvarum* was discovered in sympatry with *S. eubayanus* in Patagonia (Libkind et al. 2011). Based on these initial findings, a more representative strain data set of *S. uvarum*, especially in Australasia, the only other region of the globe where *Nothofagus* are native, was obtained and analysed at the whole-genome level.

This chapter describes the use of a population genomics approach to perform a comprehensive phylogeographic survey of *S. uvarum* with special focus on the detection, for the first time, of fingerprints of domestication in this species.

2.3 Materials and Methods

2.3.1 Yeast isolation, identification and crosses

The selective protocol used for *Saccharomyces* isolations was based on the selective enrichment in a raffinose–ethanol liquid medium as previously described (Sampaio & Gonçalves 2008; Libkind et al. 2011). Putative *Saccharomyces* isolates were confirmed by the observation of *Saccharomyces*-type ascospore production. Species identifications were based on DNA sequencing of the ITS and D1/D2 regions of the rDNA. For each pair of strains tested, ascospores were isolated and crossed to obtain intra- and interspecies hybrids. Hybridization between the two parental strains was confirmed by sequencing of the *MET2* and/or *FSY1* genes and confirmation of the expected heterozygous sites. For

each cross, interspecific spore viability was determined by examining at least 200 ascospores produced by two independent hybrid strains.

2.3.2 Quality filtering and alignment to reference genome

Genomic Illumina sequencing libraries were prepared for most of the isolates (or their monosporic derivatives; Appendix I Table AI.1) and sequenced for 100 cycles (single-end) or 2x100 cycles (paired-end) using the Illumina HiSeq2000 or GAxII systems. Some strains were multiplexed as described previously (Hittinger et al. 2010). To make results directly comparable, all sequencing data were treated as single-end reads. Reads were filtered to include only those with a perfect index match. To optimise downstream analysis, quality control for each set of reads was carried out using FASTX toolkit v0.13.2 (http://hannonlab.cshl.edu/fastx_toolkit). In brief, any adapter sequences were clipped from the reads, and low quality bases from the 3'-end of reads were trimmed based on the Illumina Phred scores. Reads were retained only if read length was above 32. Finally, reads were discarded if >5% of the positions had Phred scores <10.

Filtered reads from each isolate were mapped to the reference genome of *S. uvarum* CBS 7001 (Scannell et al. 2011) using SMALT v0.6.4 (Wellcome Trust Sanger Institute, www.sanger.ac.uk/resources/software/smalt) with default parameters, except that it was allowed a step size of 2 (-k 13 -s 2) and SAM format output. In addition to reporting the best gapped alignment for each read, the default settings of SMALT v0.6.4 only generate uniquely mapped reads. Downstream analyses, such as conversion to BAM format, 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) with default parameters. Consensus genotypes in VCF format were then converted to fastq format by restricting the maximum depth to 100, 250 or 500, depending on the median mapping depth, in order to avoid overrepresented regions that could be copy-number variants or artefacts. A fasta file was generated for each alignment where base calls with a consensus Phred quality score <40 (equivalent to a 99.99% of base call accuracy) were masked to lower case. The final quality of the filtered consensus fasta files was assessed by calculating the proportion of high-quality bases (with Phred quality score >40, Q40 hereafter) relative to the entire mapped genome. For downstream analysis, all bases with Phred quality score <Q40 were converted to an 'N'. The resulting fasta files were pooled together to generate multiple sequence alignments for each reference chromosome.

Reads from three hybrid strains studied and from *S. bayanus* NCAIM 676 (Appendix I Table AI.1) were treated in the same way as described above but were mapped to a combined reference that included the genomes of *S. uvarum* CBS 7001, *S. kudriavzevii* IFO 1802 (Scannell et al. 2011) and *S. cerevisiae* (UCSC version sacCer3). This process provided a reliable alignment of the mappable portion of reads for each hybrid strain to the *S. uvarum* reference genome that were treated in the same way as for the other isolates.

2.3.3 Phylogeny, population structure and polymorphism analyses

Before analyses, regions with evidence for introgression of *S. eubayanus* (see below) were removed from the alignments. For phylogenetic and population structure analyses, unambiguous SNPs were extracted from the chromosome multiple sequence alignments if the evaluated site was represented

only by high-confidence alleles, meaning that for each position information was available for all of the isolates.

Two phylogenetic analyses were carried out for the strain data set. The first included the complete set of isolates and was constructed based on the concatenated SNP alignment of chromosomes 5, 8 and 12 (3113 SNPs; note that chromosome 10 was mislabelled as chromosome 12, and vice-versa, in the annotation of Scannell et al. 2011). A more restricted phylogeny was built using the 54 isolates that exhibited >70% of Q40 bases. Chromosomal SNPs were then concatenated to generate a whole-genome SNP alignment (129096 SNPs). Both phylogenies were inferred by the Neighbour-Joining method with 1000 bootstrap replicates using the p-distance to compute evolutionary distances. *S. eubayanus* CRUB 1568 was used as outgroup. Phylogenetic analyses were performed in MEGA v.5.05 (Tamura et al. 2011). For the restricted data set, a maximum likelihood phylogeny was also performed using the rapid bootstrap algorithm in RAxML (Stamatakis 2006) with the GTRCAT approximation.

Population structure of *S. uvarum* was explored using the model-based Bayesian clustering method implemented in STRUCTURE v2.3.4 (Pritchard et al. 2000; Falush et al. 2003). For the global sampling of *S. uvarum*, STRUCTURE was run with a subset of 10337 equally spaced SNPs. The divergent Australasian population was then removed to facilitate identification of subtle population structure among the remaining clades of *S. uvarum*. A new subset of 9391 equally spaced SNPs was generated for this data set. For both data, the value of K was allowed to vary from K = 1 to 8, and the number of Markov chain Monte Carlo iterations was set to an initial burn-in period of 50000 iterations, followed by 30000 iterations of sampling. Allele frequencies were assumed to be correlated among populations, and the ancestry model allowed for admixture. Ten independent simulations were run for each value of K and stability was assessed by monitoring the sd between simulations. The optimal number of K clusters were estimated from the ad hoc statistic ΔK (Evanno et al. 2005). CLUMPP v1.1.2 (Jakobsson & Rosenberg 2007) was used to compute the similarity coefficient between replicate simulations of STRUCTURE (G′) using the Greedy or LargeKGreedy (K = 7 and 8) algorithms with 10000 random configurations. The highest value of H′ was always found to be above 0.99, indicating high modal similarity between the replicate Q-matrices.

Nucleotide diversity within populations and divergence between populations were estimated using Variscan v2.0 (Hutter et al. 2006). Only positions with valid alleles in at least 90% of the individuals were used for calculations (defined with the NumNuc parameter together with CompleteDeletion = 0 and FixNum = 1). The null distributions of Tajima's D statistic were generated in ms (Hudson 2002) under the standard coalescent neutral model with 10000 independent simulations conditioned on the sample size (as determined by NumNuc in Variscan) and observed estimates of θ_w. Sliding window analyses were performed on non-overlapping windows of 1000 sites, retaining only those windows where at least 500 sites (Q40) were used in calculations.

Statistics of shared polymorphisms, fixed differences and private alleles were calculated using sharedPoly program from the analysis package (http://molpopgen.org) (Thornton 2003) using whole-genome SNP alignments.

2.3.4 Bayesian coalescent analysis

Rooted coalescent phylogenies were estimated using a subset of strains from each identified population. Potential hybrid/mosaic strains were excluded from the original alignments in order to minimize violations of the models' assumptions. High-confidence alignment sites were extracted from the chromosome alignments whenever a site was represented with an unambiguous high-quality allele (Q40) in all strains. These high-confidence alignments were broken into regions with no four-gamete test violations using RminCutter.pl v1.05 (https://github.com/RILAB/rmin_cut/) (Hudson & Kaplan 1985) with the following settings: -v -q -f -m -g -n. For each chromosome, the segment with the highest number of polymorphisms was chosen and used as a different partition in the BEAST v1.7.4 package (Drummond et al. 2012). This approach produced a total of 51159 sites with 2884 segregating positions. BEAST was run with linked partitions using the coalescent constant size tree prior, the GTR + Γ + I model of sequence evolution and a strict molecular clock. Monophyly was assumed for the main clades. Coalescent estimates are given in units of substitution per site. Two independent chains of 30000 Markov chain Monte Carlo generations each were performed, with sampling at every 1000 generations. Traces were combined in LogCombiner, discarding the first 10% of generations as burn-in, and convergence was assessed by examination of the effective sampling size in Tracer v1.5.0. All estimated parameters showed effective sampling size values >600.

2.3.5 Phylogeny of the genus *Saccharomyces*

A recent multi-locus sequence approach (Koufopanou et al. 2013) was deployed to examine the evolutionary relationships among the species and known populations of the genus *Saccharomyces*, including the new Australasian population of *S. uvarum* described here. Complete coding sequences for the 14 genes were extracted from the available annotations of six *Saccharomyces* species (Scannell et al. 2011; Liti et al. 2013) (*Saccharomyces* Genome Database, SGD). Homologous regions for a representative strain of the Far Eastern (IFO 1804) and North American (YPS 138) populations of *S. paradoxus* were retrieved from the BLAST server available at http://www.sanger.ac.uk/cgi-bin/blast/submitblast/s_paradoxus_sgrp (Liti et al. 2009). A local BLAST database was also set to retrieve the same regions from the *de-novo* assemblies (see below) of *S. eubayanus* (CRUB 1568) and from one representative strain of the *S. uvarum* Australasian population (ZP 962), and from the reference genome of *Naumovozyma castellii* CBS 4309 (NCBI assembly ASM23734v1), which was used as an outgroup. The 14 coding sequences were concatenated and then aligned with FSA v1.15.7 (Bradley et al. 2009). Sites with alignment gaps and unknown bases were removed from the alignment before phylogenetic reconstruction. A maximum likelihood phylogeny was estimated in PhyML v3.0 (Guindon & Gascuel 2003) using the GTR + Γ + I model of sequence evolution, estimating the Γ distribution parameter, with five substitution rate categories, and the proportion of invariable sites. Tree topology moves were performed using the best of NNI and SPR searches with five initial random trees. Branch support was estimated with 100 non-parametric bootstrap replicates.

2.3.6 Screening for foreign genes in the genomes of *S. uvarum*

Evidence of introgressions from other *Saccharomyces* species were searched by mapping the reads to a combined reference including all the available annotated coding sequences of six *Saccharomyces* species (Scannell et al. 2011; Liti et al. 2013) (SGD). Reads were quality-filtered as above and mapped to this combined reference using BWA with default parameters (Li & Durbin 2010). 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 >80% of Q40 bases, which provides a good initial measure for the mappability of reads to the corresponding targets, in >100 bases were selected for further analysis. Finally, only genes with orthologues unambiguously annotated in at least four species, including *S. uvarum*, were analysed. Exceptions are mainly owing to the fact that some of the *bona fide* transfers of genes that are not annotated in *S. uvarum* genome seemed to be close neighbours to other genes that are annotated in both the *S. uvarum* and donor genomes. A gene was recognized as having a foreign origin if the number of reads that mapped to the putatively donor genome was higher than the sum of the reads that mapped to the orthologous genes in the other *Saccharomyces* species. Five single-gene transfer events were identified using this process but were excluded from further analysis because they corresponded to short coding sequences (<200 bp long) or had a low number of reads mapped.

Introgressions from other fungal non-*Saccharomyces* species were searched by assembling the reads that failed to map to the reference genome of *S. uvarum* into contigs with Velvet v.1.2.08 (Zerbino & Birney 2008). Assembled contigs with >500 bp length were used as query and 'blasted' against the ORF sequences of the three horizontally transferred regions identified by Novo et al. 2009 and to the NCBI nr database using BLASTN (1e–10 E-value cutoff). Only the best blast hit for each query was retained and hits to the rDNA region were excluded from the analysis.

2.3.7 Screening for introgressions from *S. eubayanus*

S. eubayanus does not yet have a reference assembly against which reads can be mapped. Therefore, additional reads from the type strain of *S. eubayanus* (CRUB 1568) were generated in this study, and then treated with the methods described above for the alignment to the reference genome of *S. uvarum*. Pairwise divergence between the *S. uvarum* and *S. eubayanus* type strains was used as a proxy to search for evidence of DNA segments of *S. eubayanus* in the genomes of *S. uvarum*. Divergence per site, k, (with Jukes–Cantor correction) was calculated in Variscan v2.0 (Hutter et al. 2006) using a non-overlapping sliding window of 1000 sites. Only the windows with at least 500 high-quality sites used in the pairwise comparisons (Q40) were retained for further analyses. For easier visualization and interpretation of these results, each data point in the plots (Figure 2.5) represents the average divergence of itself plus three windows on each side.

2.3.8 Gene Ontology (GO) analysis

Standard GO term discovery was performed with the Generic GO Term Finder (Boyle et al. 2004). This procedure, however, does not consider the observation that the same gene can be found within introgressed regions of multiple strains. To account for this redundancy, the number of times each GO attribute was annotated for the set of introgressed genes in each strain was first determined and then

summed over all the strains considered. Fisher's exact test was used to test the hypothesis that a specific GO attribute was overrepresented in the pooled introgressed gene set relative to the expected number of genes in the reference gene annotation file (retrieved from YeastMine on December 2013). The Bonferroni procedure was applied to correct for multiple comparisons using the total number of tests performed for each ontology term (process, component or function). Significance was assessed for all data with a corrected P-value cutoff of 0.05.

2.3.9 *de-novo* assemblies and analysis of *FZF1* and *ZRT1* sequences

For a more detailed analysis of the introgression on chromosome 7, which harbours *FZF1* and *ZRT1*, it was performed *de-novo* genome assemblies of the Illumina single-end or paired-end reads for all strains included in this study, using Velvet v.1.2.08 (Zerbino & Birney 2008). The expected coverage for each region was calculated, and the contigs with a coverage value <10x were discarded, whereas the regions with a coverage value <15x were masked. To retrieve the gene sequences, we set up a local BLAST database for each genome and searched for the aforementioned genes by a TBLASTX search using *S. uvarum FZF1* and *ZRT1* sequences from strain CBS 7001 as queries (SGD PORF 7762; SGD PORF 7764). The gene sequences were then aligned using ClustalW from within MEGA v.5.05 (Tamura et al. 2011). For the promoter analysis of each gene, a 1000-bp upstream sequence was retrieved from SGD and the aforementioned procedure was followed.

2.4 Results

2.4.1 Occurrence of *S. uvarum* in the Southern Hemisphere

It was recently reported the sympatric occurrence of a wild population of *S. uvarum* and its novel sibling species *S. eubayanus* in Northwestern Patagonia, in association with *Nothofagus* spp. and the sugar-rich fruiting structures (stromata) of its biotrophic fungal parasite, *Cyttaria* hariotii (Libkind et al. 2011). In that study, 47 isolates of *S. uvarum* were recovered from 133 samples obtained from three species of *Nothofagus*, with *N. dombeyi* standing out as the species with the highest frequency of isolation with a global success rate of 35%. A much lower frequency of isolation had been previously obtained for this species in Europe when the *Quercus* habitat was explored (Sampaio & Gonçalves 2008). This difference highlighted the need to study the ecological association of *S. uvarum* with the *Nothofagus*–*Cyttaria* system, not only in Patagonia but also in Australasia. Therefore, it was carried out an expanded survey of isolations that included a southward extension of the Patagonian region previously covered, as well as sampling in Australia and New Zealand. In the South American survey, 218 samples that included *N. dombeyi* and five additional species of *Nothofagus* were collected in the Spring and Summer months of 2008, 2009 and 2010, yielding 59 *S. uvarum* isolates. In Australasia, 113 samples were collected in the Spring of 2009, yielding eight isolates.

To provide insight into the global genetic diversity and phylogeography of *S. uvarum*, a set of strains representing the entire scope of geographic origins and habitats (Figure 2.1a,b and Appendix I Table AI.1) was selected for whole-genome sequencing using the Illumina technology. In addition to a representative collection of isolates from the *Nothofagus* niche, the strain data set also included wild isolates collected in North America, Eurasia and Far East Asia, which was obtained mostly from the

bark of oak trees and associated soil. In addition, European strains isolated from artificial environments, mostly wine and cider, were included. Finally, three confirmed hybrid strains of *S. uvarum* x *S. cerevisiae* (two of them, triple hybrids with contributions also from *S. kudriavzevii*) were also included (Appendix I Table AI.1).

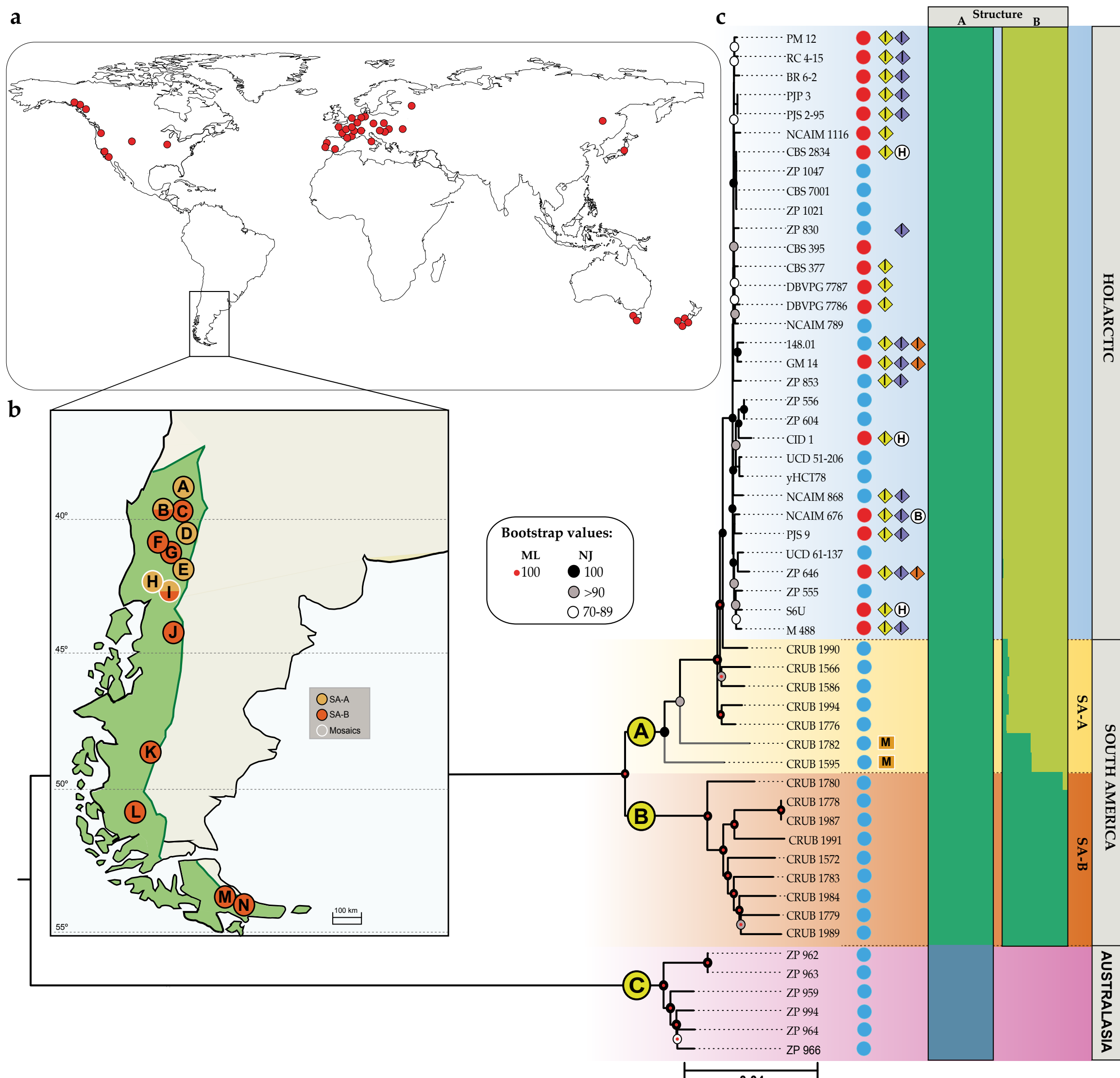


Figure 2.1 | Geographic distribution, phylogeny and population structure of *Saccharomyces uvarum*. **(a)** World map depicting the geographic origin of HOL and AUST *S. uvarum* strains. **(b)** Map of South America depicting the collecting sites. Locations are indicated by capital letters surrounded by light- or dark-orange coloured circles corresponding to populations SA-A and SA-B, respectively. The co-occurrence of strains of the two populations is depicted by light- and dark-orange semi-circles, and the occurrence of mosaic strains is depicted by white contour lines. The distribution of *Nothofagus* is shown in green. **(c)** Whole-genome Neighbor-Joining (NJ) phylogeny of 54 strains based on 129096 SNPs and excluding introgressed regions from *S. eubayanus*, inferred with p-distance and rooted with *S. eubayanus*. The three main clades are marked by letters A, B and C. Support values from 1000 bootstrap NJ and ML trees are included. Incongruent topologies between NJ and ML are denoted by grey lines. Branch lengths correspond to the mean number of base differences per site. Strains isolated from natural environments are marked by blue circles, whereas strains isolated from anthropic environments are marked by red circles. Strains having introgressions are marked with an 'I' in a yellow (introgressions from *S. eubayanus*), blue (introgressions from *S. kudriavzevii*) or orange (introgressions from *S. cerevisiae*) diamond. For the hybrid strains (marked with an 'H' in a white circle) or *S. bayanus* ('B' in a white circle), only introgressions in the *S. uvarum* sub-genome are depicted. South American mosaic strains are marked with an 'M' in an orange square. Clusters inferred with STRUCTURE considering all sequences (A) and all but the Australasian sequences (B) are depicted.

The phylogenetic relationships of 54 strains, based on 129096 high-quality polymorphic sites, are depicted in Figure 2.1c. Appendix I Figure AI.1 depicts a topologically similar phylogeny obtained on an expanded set of 61 strains using 3113 high-quality single-nucleotide polymorphisms (SNPs) from chromosomes 5, 8 and 10 that do not show introgressions with *S. eubayanus* (see below). These broad phylogenetic analyses resolved the various representatives of *S. uvarum* into three main clades. Clade A contained all Holarctic strains and a few from South America; clade B encompassed only strains from South America; and clade C corresponded to the Australasian population and was clearly separated from the other two clades. In fact, this lineage was 4.4% divergent (divergence per site, k, with Jukes–Cantor correction) from the other two main lineages, which is similar to the divergence of the North American population of *S. paradoxus* from both the European and Far Eastern populations (Liti et al. 2006). In spite of their divergence, the Australasian strains are clearly more closely related to *S. uvarum* than to any other *Saccharomyces* species, as depicted in Figure 2.2. The phenotypic profiling of the Australasian isolates showed also some unique features, including delayed or absent growth on maltose and melibiose, maximum growth temperatures 3–4°C lower than those observed for other *S. uvarum* strains (31–32 versus 35°C for the non-Australasian strains), and lower growth rates on melibiose at some temperatures (Appendix I Table AI.2 and Table AI.3).

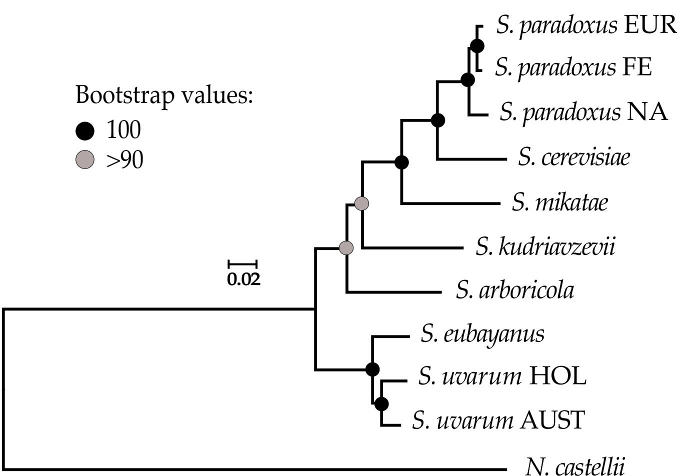


Figure 2.2 | Phylogenetic relationships between species and populations of *Saccharomyces*. Maximum likelihood (ML) phylogeny of the genus *Saccharomyces* based on a concatenated alignment of 14 gene sequences and rooted with *Naumovozyma castellii*. Representatives of populations of *S. paradoxus* (EUR, European; FE, Far Eastern; and NA, North American) and of the Australasian (AUST) and Holarctic (HOL) populations of *S. uvarum* are included. Support values correspond to 100 bootstrap replicates and branch lengths correspond to the mean number of nucleotide substitutions per site.

The observation that the Australasian strains were both genetically and phenotypically divergent from the remaining lineages, combined with their geographic isolation, led to the hypothesis that the observed divergence could be the consequence of allopatric partitioning. To help ascertain this, crosses between Australasian strains and representatives of the other two main lineages represented in Figure 2.1c were performed to determine the viability of the resulting progeny. Interestingly, the four crosses involving an Australasian strain mated with a South American or European strain yielded intermediate levels of spore viability of 27–36% (Appendix I Table AI.4), whereas the progeny of the cross of two Australasian strains, one from Tasmania and the other from New Zealand, had 95.7% spore viability. Another observation worth noting was the relatively high spore viability (18.8%) produced by a cross

between an Australasian *S. uvarum* representative and a South American strain of *S. eubayanus* (Appendix I Table AI.4). This value contrasts with the reported 7% of spore viability from crosses of sympatric (South American) representatives of *S. uvarum* and *S. eubayanus* (Libkind et al. 2011), which could be a sign of the reinforcement of reproductive isolation that deserves further study.

2.4.2 Population structure and phylogeography of *S. uvarum* in South America

The *S. uvarum* phylogeny (Figure 2.1c) revealed that, by far, most of the diversity contained in this species is found in the Southern Hemisphere. In fact, the nucleotide diversity based on pairwise differences (θ_π x 100) contained in the group of South American isolates was 0.689, approximately half of the nucleotide diversity observed globally for the species (1.248) and much higher than that of the Australasian isolates (0.162) and of the Holarctic group (0.141; Table 2.1). Therefore, in addition to the extremely divergent Australasian lineage, the isolates from the Southern Hemisphere encompass more genetic diversity than those from the Northern Hemisphere. This high diversity observed for the Patagonian strains can be partly attributed to the presence of two major well-separated lineages, such as clades A and B. The South American strains of clade A (SA-A) are phylogenetically related to the Holarctic group and were placed at the basal positions of this branch, whereas the Holarctic isolates occupied more nested positions (Figure 2.1c). Clade B was restricted to South American strains only, and is hereafter designated SA-B. The two South American lineages (SA-A and SA-B) had a genetic distance of 1.0% (Table 2.2) and genetic diversities of 0.415 (SA-A) and 0.333 (SA-B) (Table 2.1).

Table 2.1 | Whole-genome diversity values for *Saccharomyces uvarum* populations. Diversity values were calculated for the total length of the genome (excluding introgressions from other *Saccharomyces* species) with at least 90% of high quality alignment coverage

Population	Nº strains	Analyzed sites	Segregating sites	θ_π	θ_W	Tajima's D	P value
S. uvarum	54	5722528	341794	0.01248	0.01340	-0.25018	0.4638
AUST	6	6123420	19874	0.00162	0.00156	0.28113	0.6260
SA	16	5490591	103726	0.00689	0.00594	0.72569	0.8186
HOL	32	6348178	29411	0.00141	0.00118	0.78073	0.8398
SA-A	7	5469941	55976	0.00415	0.00448	-0.48797	0.3767
SA-B	9	5815388	55279	0.00333	0.00367	-0.50626	0.3186
SA + HOL	48	6000052	126087	0.00488	0.00486	0.01847	0.5903

AUST, Australasia; HOL, Holarctic; SA, South America.

Table 2.2 | Pairwise whole-genome divergence comparisons between *S. uvarum* populations and to *S. eubayanus*. Divergence values per site k*100 (Jukes-Cantor corrected) for *S. uvarum* populations (excluding introgressions from other *Saccharomyces* species). Divergence between *S. uvarum* and *S. eubayanus* is also shown for comparison. Only sites with valid alleles in at least 90% of the strains were used for calculations.

Population	HO	SA-A	SA-B	HOL + SA	*S. eubayanus*
SA-A	0.28	-	-	-	-
SA-B	1.11	1.01 (1.13*)	-	-	-
AUST	4.39	4.40	4.36	4.39	-
S. uvarum	-	-	-	-	6.34

* South American mosaic strains excluded.

Population structure inference carried out in STRUCTURE (Figure 2.1c) clearly indicated the presence of two distinct genetic clusters worldwide, as determined by the high value of the *ad hoc* statistic, ΔK, which measures the rate of change between log probabilities for successive K values ($\Delta K_2 = 3682.6$). One of these clusters was assigned to the Australasian lineage and the other consisted of all remaining isolates (South America and Holarctic region). Because STRUCTURE inference is strongly influenced by the sampled genotypes that are used to run the software and because of the marked divergence of the Australasian population, this lineage was excluded from the data set and the analysis repeated. Interestingly, this restricted *S. uvarum* data set was resolved into two new genetic clusters that reflected the phylogenetic inference, one cluster was cosmopolitan and present in both South America and the Northern Hemisphere, whereas the other cluster was endemic to South America (SA-B; Figure 2.1c). Increasing values of K suggested only minor possible ancestry contributions to those two clusters and did not uncover any new clean clusters (Appendix I Figure AI.2).

In addition to the strains clearly belonging to either one of these two populations, there were also two strains in the phylogenetic tree (CRUB 1782 and CRUB 1595) that occupied positions that could be explained by a mosaic structure of their genomes, with contributions from different lineages. In fact, these strains showed mixed ancestries in STRUCTURE (Figure 2.1c), which is consistent with admixture between the two populations. A sliding window analysis comparing the genome of CRUB 1782 with those of strains representing each of the two clean lineages confirmed that it was nearly identical to one or the other of the two South American populations across its genome (Figure 2.3). These analyses suggested that both strains were ~65% SA-A and 35% SA-B, although the exact regions contributed by each parental population were frequently non-overlapping. The existence of strains with mixed ancestries is in line with the high (97%) spore viability found in crosses between strains representing the two populations (Appendix I Table AI.4).

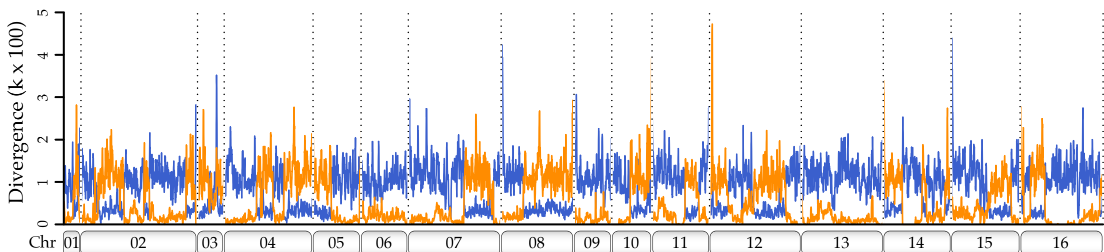


Figure 2.3 | Sliding window analysis showing the mosaic nature of the genome of CRUB 1782, a South American strain of *S. uvarum*. Each plot represents the divergence k (with Jukes-Cantor correction), relative to either one of the two South American clean lineages (divergence to SA-A is shown in blue and divergence to SA-B in orange). Note the changes of highest identity to either one of the two clean lineages in multiple chromosomes.

2.4.3 Limited genetic diversity of the Holarctic population

When compared with the genetic diversity observed for strains from the Southern Hemisphere, strains from a wide range of environments in the Northern Hemisphere show remarkably low diversity across their genomes. This seems to be consistent with the Southern Hemisphere and the *Nothofagus-Cyttaria* system being native geographical ranges and habitats of *S. uvarum*, respectively. It also suggests the hypothesis of a recent migration of the Patagonian *S. uvarum* population SA-A into the

Northern Hemisphere and the consequent habitat shift to oaks and other non-*Nothofagus* trees imposed by the absence of *Nothofagus* north of the equator. Indeed, several lines of evidence support the view that the Northern Hemisphere population might be recently derived from the South American SA-A population. First, the nucleotide diversity (θ_π x 100) is much higher in the SA-A population (0.415) than in the Holarctic population (0.141), as shown in Table 2.1. In addition, the Holarctic group has the highest value of Tajima's D statistic, although it is not significant. The relative excess of intermediate frequency alleles is compatible with a recent population contraction where there has not been sufficient time for many new mutations to accumulate (Table 2.1). Moreover, the analysis of shared polymorphisms involving the South American and Holarctic populations supports the hypothesis that the Northern Hemisphere population is derived from the Patagonian SA-A population. The proportion of shared polymorphisms between the SA-A and Holarctic populations is much higher than that of the comparison involving SA-B and Holarctic populations (Figure 2.4a). In addition, an almost negligible proportion of fixed differences were found between the SA-A and Holarctic populations, contrary to what was observed between SA-A and SA-B and, most notably, between the Holarctic population and SA-B. If the Northern Hemisphere population has derived from the Patagonian SA-A population, the time to coalescence of all Holarctic lineages in a single common ancestor is expected to be more recent than that of the coalescence of lineages of SA-A. In fact, an analysis of relative coalescence times suggested a more recent common ancestor for the Northern Hemisphere population than for either of the South American populations. Specifically, the Holarctic clade is only ~71% as old as the SA-A clade (Figure 2.4b), which fits in a model of colonization of the Northern Hemisphere from South America. Therefore, taken together, our analyses support the view suggested by the phylogeny in Figure 2.1c that the Northern Patagonian population SA-A gave rise to the Holarctic population.

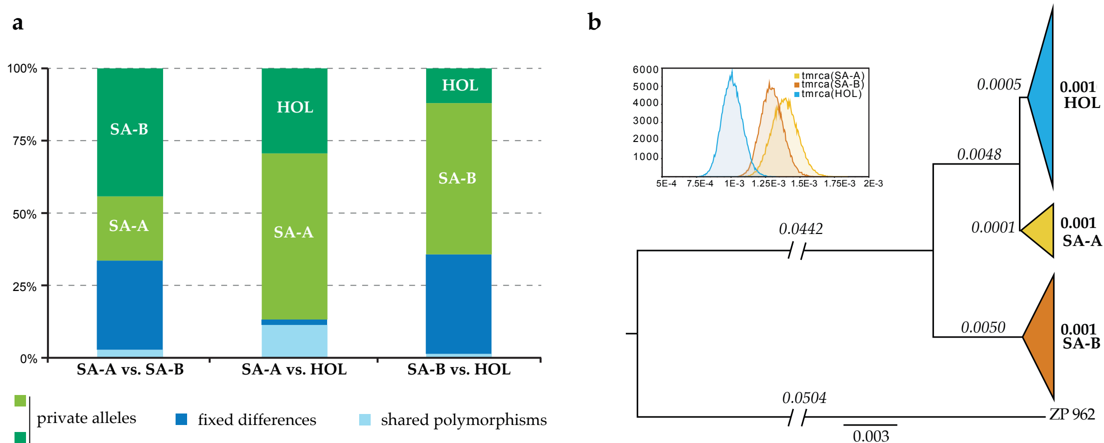


Figure 2.4 | Proportion of shared and privately segregating polymorphisms and coalescence analysis in the South American and Holarctic (HOL) populations. **(a)** Proportion of private alleles, fixed differences and shared polymorphisms among SNPs found in all possible pairwise comparisons involving the HOL and the two South American populations, SA-A and SA-B (South American mosaic strains were excluded from the analysis). **(b)** Genome-wide estimation of relative time to coalescence for the HOL and the two South American populations. The tree was built from an alignment of 51159 high-quality sites (Phred quality score >Q40) partitioned over the 16 chromosomes. Each partition represents regions without evidence for intra-locus recombination. The scale bar depicts estimated substitutions per site. Node ages are printed in bold near the nodes of the population whose coalescence they estimate. Branch lengths are printed in italics above the branches. One representative strain of the divergent Australasian population was used to root the tree. The insert shows the marginal posterior densities of the time to the most recent common ancestor (tmrca) in each population.

2.4.4 Introgressions from other *Saccharomyces* species

Because several introgression events were previously reported between *S. paradoxus* and *S. cerevisiae* (Liti et al. 2006; Doniger et al. 2008; Muller & McCusker 2009), all genome sequences of *S. uvarum* were screened for introgressions from other *Saccharomyces* species. Although signs of foreign genomic DNA were not found in strains isolated in the Southern Hemisphere, 20 out of 34 Holarctic strains displayed multiple introgressions that could be clearly ascribed to various *Saccharomyces* species, including from *S. kudriavzevii* (14 strains), *S. cerevisiae* (3 strains) and most notably from *S. eubayanus* (20 strains; Appendix I Table AI.5 and Digital Resources Dataset D1). These introgressions were widespread across the *S. uvarum* genomes, but their number and size were found to differ between strains. The distribution of the *S. eubayanus* introgressions are summarized in Figure 2.5 (next page). Interestingly, all but one of the 18 strains isolated from anthropic niches were found to have introgressions, either solely from *S. eubayanus* or from up to two additional *Saccharomyces* species (Figure 2.1c and Appendix I Table AI.5).

The most striking and recurrent introgressions are from *S. kudriavzevii* and *S. eubayanus*. The *S. kudriavzevii* introgressions are all partially overlapping and relatively short (encompassing up to seven genes). Sequence comparison readily shows that these fragments of *S. kudriavzevii* DNA were acquired from the European population of the species, in line with the European origin of all but one of the strains harbouring the *S. kudriavzevii* introgressions. Curiously, the sole non-European strain carrying a *S. kudriavzevii* introgression was isolated in Japan, where a genetically distinct *S. kudriavzevii* population has been identified (Hittinger et al. 2010). However, this strain is likely to be phylogenetically related with the European strains since its *S. kudriavzevii* introgression is very similar to those present in European strains, both in gene content and sequence (Appendix I Table AI.5).

Introgressions from *S. eubayanus* are the most prevalent and extensive. The introgressions are, on average, 99.5% identical to the genome of the South American strain of *S. eubayanus* CRUB 1568 (FM1318), but they are 99.9% identical to the *S. eubayanus* portion of the genome of *S. pastorianus* strain Weihenstephan 34/70. This divergence between the *S. eubayanus* type strain and the *S. eubayanus* moiety found in *S. pastorianus* has already been reported (Libkind et al. 2011). Since all the *S. uvarum* strains with *S. eubayanus* introgressions are European, it seems reasonable to assume that the introgressions occurred in Europe. Therefore, these results suggest that the *S. eubayanus* introgressions in *S. uvarum* either originate from *S. pastorianus* or are derived from a yet-unidentified Eurasian lineage of *S. eubayanus* that was also the contributor to the genomes of the hybrid species *S. pastorianus* and *S. bayanus*. The number of *S. eubayanus* introgressions per strain varied between seven (strain DBVPG 7787) and one (strains CBS 377, BR 6-2, NCAIM 868, ZP 646 and CID 1). The genes involved in the introgressions are listed in Digital Resources Dataset D1, excluding those present in strain NCAIM 676, which possesses a much more extensive contribution from *S. eubayanus*.

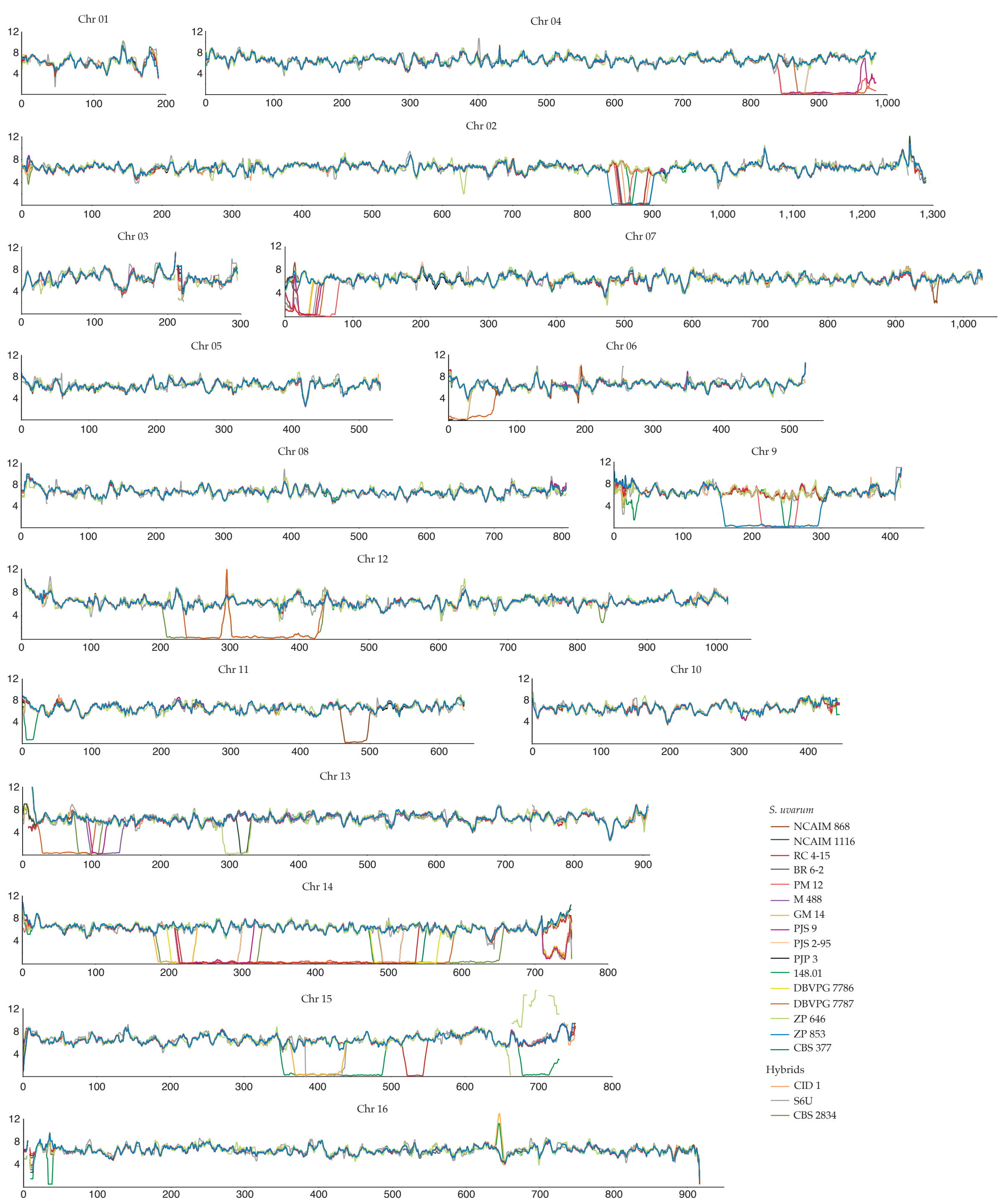


Figure 2.5 | Sliding window analysis of *S. eubayanus* introgressions in *S. uvarum* genomes. Each plot represents the % divergence k (with Jukes–Cantor correction) relative to *S. eubayanus* reference strain CRUB 1568 (x-axis values are in kb). Introgressed strains are colour-coded according to the key.

The strain NCAIM 676 has a genomic composition similar to that of *S. bayanus* strains CBS 380 and NBRC 1948 (Supplementary Table 7). This strain has slightly more genetic material from *S. uvarum* than from *S. eubayanus*, but the extensive genomic contributions from *S. eubayanus* and limited contributions from *S. cerevisiae* suggest that it should be regarded as a representative of the artificial species *S. bayanus*

sensu Libkind et al. 2011. The *S. uvarum* hybrids CBS 2834, CID 1 and S6U contain substantial contributions across their genomes from *S. cerevisiae* and / or *S. kudriavzevii* (Table 2.3).

Table 2.3 | Estimated contributions of *S. uvarum*, *S. eubayanus*, *S. cerevisiae*, and *S. kudriavzevii* to the complex genomes of two triple hybrid strains (CBS 2834 and CID 1), the *S. uvarum* x *S. cerevisiae* hybrid S6U, and to *S. bayanus*. The reads of the strains indicated in the table were mapped to a combined reference meta-genome including all the available annotated coding sequences of 6 *Saccharomyces* species (*S. arboricola*, *S. cerevisiae*, *S kudriavzevii*, *S. mikatae*, *S. paradoxus* and *S. uvarum*). To avoid the counting of spurious read mappings, only ORFs with more than 200 bp and 80% of Q40 were analysed. For estimating the contributions from *S. eubayanus* alone, a threshold of 3% divergence was used to distinguish between *S. uvarum* and *S. eubayanus* in pairwise comparisons at the whole genome level. Note that for *S. bayanus* hybrids, the *S. eubayanus* contribution is more extensive than in any other case and, additionally, genomic regions can be heterozygous for *S. eubayanus* and *S. uvarum* alleles. *S. uvarum* NCAIM 868 and DBVPG 7787 are used for comparison.

	Hybrids			***S. bayanus***			***S. uvarum***	
Species contributions	**CBS 2834**	**CID 1**	**S6U**	**NCAIM 676**	**CBS 380***	**NBRC 1948***	**NCAIM 868**	**DBVPG 7787**
S. uvarum + S. eubayanus	35.20	29.92	45.27	98.97	99.23	98.99	99.87	99.99
S. cerevisiae	36.21	36.52	54.73	0.98	0.68	0.98	0.00	0.00
S. kudriavzevii	28.08	33.48	0.00	0.04	0.06†	0.00	0.07	0.00
S. eubayanus	5.46	0.34	1.32	26.79	36.39	62.58	0.42	7.80

* From Libkind et al. (2011)
† Only one gene detected (YPR051W) with an unusual high number of reads that mapped to this location.

Gene Ontology (GO) analyses were used to search for a signal of function-specific introgression that might suggest concerted selection, the power of which is limited by the unknown but probably considerable proportion of hitchhiking genes. Separate GO analysis of the gene set acquired by each of the strains through introgression failed to yield overrepresented GO categories that were both reasonably specific and previously associated to wine fermentations (Digital Resources Dataset D2). However, when GO analysis was performed using the set of genes introgressed into any of the strains, the categories 'GO:0006807: nitrogen compound metabolic process' and 'GO:0051171: regulation of nitrogen compound metabolic process' were found to be overrepresented (corrected P-values < 0.05; Digital Resources Dataset D2). Moreover, the use of a gene pool assembled so that the redundancy resulting from the presence of an introgressed gene in multiple strains was preserved, highlighted the additional categories 'GO:1900071: regulation of sulfite transport', 'GO:0006808: regulation of nitrogen utilization', 'GO:0019740: nitrogen utilization', and 'GO:0008238: exopeptidase activity', among others (Digital Resources Dataset D2). These categories include genes pertinent to wine fermentation, some of which have been previously shown in *S. cerevisiae* wine strains to be involved in horizontal gene transfer (HGT) events (Novo et al. 2009; League et al. 2012) or genomic rearrangements (Pérez-Ortín et al. 2002) associated with anthropic environments. Some other overrepresented GO categories include one or several genes located close to genes relevant for nitrogen or sulfite metabolism and could possibly appear as overrepresented as a result of hitchhiking.

All the introgressions on chromosome 2 harbour *ASP1*, a gene encoding the cytosolic L-asparaginase used to degrade asparagine to be used as nitrogen source. This locus was previously associated with low acetic acid production in *S. cerevisiae* wine strains by quantitative trait loci (QTL) mapping (Marullo et al. 2007). Curiously, a similar gene induced by nitrogen starvation (*ASP3*) was acquired by some *S. cerevisiae* strains by horizontal gene transfer (League et al. 2012). Also, nine strains

contained variably sized introgressions of subtelomeric regions of chromosome 7 that contained *FZF1* (Figure 2.5 and Digital Resources Dataset D2). This gene has been shown to encode a fast-evolving transcription factor that regulates several genes, including the gene encoding Ssu1p, an efflux pump involved in sulfite resistance (Engle & Fay 2012). All introgressions in this set also encompass the *ZRT1* gene that encodes a high-affinity zinc transporter that was recently reported to exhibit a sequence signature of balancing selection in *S. cerevisiae* (Engle & Fay 2013). Whereas the 900-bp sequence of the coding region of *FZF1* is identical in all introgressed strains, it exhibits 16 and 126 SNPs when compared with reference strains of *S. eubayanus* and *S. uvarum*, respectively. Interestingly, the introgressed *ZRT1* allele is highly diverged from both reference strains (149 and 148 SNPS, respectively). For both *FZF1* and *ZRT1*, there is no evidence of introgression from other known *Saccharomyces* species. In addition to differences in the coding region, the *FZF1*/*ZRT1* intergenic region appears to carry a large insertion when compared with *S. eubayanus* (Appendix I Figure AI.3). Since only the *S. cerevisiae* copy of the *FZF1* gene can be retrieved from the *S. pastorianus* genome, it is not possible to compare these introgressions to the *S. eubayanus* sub-genome of the hybrid species.

The presence of non-*Saccharomyces* fungal DNA was also searched using *de-novo* assemblies constructed with the subset of reads that could not be mapped to the *S. uvarum* reference genome. These searches did not reveal the presence of non-*Saccharomyces* open reading frames (ORFs), except for genes known to be present in three horizontally transferred regions previously identified in *S. cerevisiae* wine yeasts (Novo et al. 2009) and originating from yeasts outside this genus (Table 2.4). Since all the nine strains harbouring genes from these regions also had introgressions for other *Saccharomyces* species and all but one were associated with anthropic environments it is possible that non-*Saccharomyces* DNA was transmitted simultaneously with *Saccharomyces* introgressions.

Table 2.4 | Detection of horizontally transferred regions A, B and C in *S. uvarum*. Reads that did not map to the reference genome were assembled *de-novo* and the resulting contigs were used to search for evidence of exogenous non-*Saccharomyces* DNA segments. Contigs with more than 500 bases length were "blasted" against ORF sequences of regions A, B and C and to the NCBI nucleotide database with blastn. The latter approach did not result in any hit to fungal ORF DNA other than those from *Saccharomyces* spp. Positive results are marked in dark grey and numbers refer to the number of ORFs found relative to the total number of ORFs present. For comparison the strains for which introgressions from other *Saccharomyces* species were detected are also indicated.

		Introgressions*			Regions from Novo et al. 2009		
Strain	**Source†**	**Seub**	**Skud**	**Scer**	**A**	**B**	**C**
PM 12	A						
RC 4-15	A						
BR 6-2	A						17/19
PJP 3	A						10/19
PJS 2-95	A						17/19
NCAIM 1116	A						
CBS 2834	A						
CBS 295	A						
CBS 377	A						
DBVPG 7787	A						
DBVPG 7786	A						
GM 14	A						1/19
CID 1	A						1/19
NCAIM 676	A				3/15	5/5	1/19
PJS 9	A						
ZP 646	A						5/19
S6U	A					4/5	2/19
M 488	A						
148.01	W						1/19
ZP853	W						
NCAIM 868	W						
ZP 830	W						

* Seub: *S. eubayanus*; Skud: *S. kudriavzevii*; Scer: *S. cerevisiae*
† A: anthropic; W: wild

2.5 Discussion

In *Saccharomyces*, several cases of genome modifications through hybridization (Dunn & Sherlock 2008), introgression (Doniger et al. 2008; Muller & McCusker 2009), genome rearrangements (Pérez-Ortín et al. 2002; Zimmer et al. 2014) and HGT (Novo et al. 2009; League et al. 2012) have been documented, suggesting that genomes modified in this manner appear to be selected for in many industrial environments. Eloquent examples are the interspecies hybrids *S. pastorianus* (*S. cerevisiae* x *S. eubayanus*) in lager beer and *S. cerevisiae* x *S. uvarum* and *S. cerevisiae* x *S. kudriavzevii* in certain types of wines and of Belgian-style beers (González et al. 2006; Lopandic et al. 2007; González et al. 2008). HGT events also appear to enhance the attributes of *S. cerevisiae* wine strains (Novo et al. 2009; League et al. 2012). These changes have led to the domestication of *Saccharomyces* lineages as they have become genetically distinct from their wild relatives (Fay & Benavides 2005; Legras et al. 2007; Liti et al. 2009)

in ways that are beneficial to humans. Although the process shares some similarities to what has been documented for plant crops and livestock (Doebley et al. 2006), HGT, interspecies hybridization and introgression provide a broader array of genetic mechanisms that lead to domesticated phenotypes.

Here it is presented for the first time evidence of domestication in *S. uvarum*: the detection of multiple introgressions, mainly of *S. eubayanus* genomic DNA, into the genomes of European wine or cider strains of *S. uvarum*. These introgressions were also found in *S. uvarum* x *S. cerevisiae* hybrids, which suggests that the *S. eubayanus* contribution is advantageous even in 'enriched' hybrid genomes. Based on these observations, it is possible that the anthropic habitats colonized by *S. uvarum* in Europe may have favoured hybridization of *S. uvarum* with *S. eubayanus*, followed by subsequent introgression by backcrossing to *S. uvarum*. Conversely, hybrids or introgressed strains involving these two species may be less fit than the clean parental lineages in their natural South American habitat.

Many *S. cerevisiae* wine strains possess three gene clusters acquired from non-*Saccharomyces* yeasts by HGT (Novo et al. 2009). These regions are 120 kb long and include 34 genes, many of which are involved in key aspects of must fermentation such as the metabolism and transport of sugars and nitrogen. Hence, there seems to be some similarity between the findings in *S. uvarum* presented here and those reported for *S. cerevisiae* because, in both cases, selective pressures in anthropic environments seem to have promoted genome adaptations involving interspecies exchange of genetic material. Nevertheless, the global pattern of domestication in *S. cerevisiae* seems to be different from that observed in *S. uvarum*. The marked reduction of genetic diversity of *S. cerevisiae* wine domesticates relative to their wild relatives has no parallel in *S. uvarum* because the Holarctic wild and domesticated strains are indistinguishable in this respect.

It is not clear if *S. eubayanus* introgressions occurred repeatedly or if they are the result of a single hybridization event that then spread throughout the European population. The latter possibility seems to be more in line with the apparent low abundance of *S. eubayanus* in Europe. The polymorphic pattern of the introgressions when different strains are compared suggests that their fixation or elimination from the population has not yet stabilized. Judging from the widespread occurrence of introgressed strains across Europe and their almost complete absence in wild populations, it can be hypothesised that introgressions are rapidly lost in European natural environments.

This and other recent studies (Libkind et al. 2011; Peris et al. 2014) concur in documenting an ancestral *Nothofagus* association for the two basal *Saccharomyces* species, *S. eubayanus* and *S. uvarum*. For *S. eubayanus*, two populations have also been reported in Patagonia, with a genetic divergence of 1% (Peris et al. 2014), that is similar to the divergence measured between the two Patagonian *S. uvarum* populations. In South America, *S. eubayanus* and *S. uvarum* share the *Nothofagus* niche, but *S. eubayanus* seems to be much more elusive than *S. uvarum* in the Northern Hemisphere. Only three isolates of *S. eubayanus* were reported in North America (Peris et al. 2014), and the evidence presented here for the presence of *S. eubayanus* in Europe remains indirect. For *S. uvarum*, we could expand the association with *Nothofagus* by identifying a highly divergent population in New Zealand and Tasmania, which constitutes a significant addition to the diversity of early diverged *Saccharomyces* lineages found in the Southern Hemisphere. The partial reproductive isolation of the Australasian population is comparable to what is observed for the most divergent populations of *S. paradoxus* (32% spore viability) (Liti et al. 2006) and also fits within the range of results obtained for divergent lineages of *S. cerevisiae* found in China that have 10.2–55% spore viabilities (Wang et al. 2012). Therefore, until more detailed analyses

are carried out, the Australasian population can be conservatively viewed as a considerably divergent *S. uvarum* population, possibly the result of an ongoing process of allopatric speciation.

Although the two South American populations of *S. uvarum* are generally sympatric, their distribution ranges appear to only partially overlap. In fact, the distribution of one of these populations (SA-A) is restricted to the northern part of Patagonia, whereas the other population (SA-B) has a distribution that extends to southern Patagonia, including Tierra del Fuego (Figure 1c). The evidence for admixture that was obtained and the minor contributions from population SA-B in all studied isolates from population SA-A revealed in the STRUCTURE analysis (Figure 1d) suggest that the two populations are presently in contact and that there is little or no relevant reproductive isolation between them, which could be experimentally confirmed. The geographic distribution of the two Patagonian populations can be related to the phylogeography of *Nothofagus* whose species tend to have latitudinally disjunct populations as a consequence of long-lasting vicariance events related to past glaciations and to the presence of an ancient paleobasin at mid-latitudes in Patagonia (Markgraf et al. 1995; Mathiasen & Premoli 2010). Our results fit this model of two, historically isolated clades, one in the north and the other in the south. Favourable climatic and geologic conditions prevailing since 25000 years ago seem to have facilitated secondary contact. In mid-latitudes, northward migration of *Nothofagus* appears to have been more efficient than southward migration (Mathiasen & Premoli 2010), which is concordant with what we have observed for *S. uvarum* (Figure 1c). Phylogenomic and population structure analyses support the view that a restricted subset of population SA-A gave rise to the Holarctic population, although vectors and mechanisms of this migration into the Northern Hemisphere remain to be elucidated. The lack of diversity found in the extant residents of the Northern Hemisphere contrasts with the vast geographical area colonized, which ranges from temperate North America to Europe and Asia, and suggests that the subsequent dispersal across these northern continents occurred rapidly enough to prevent the accumulation of many local mutations.

Australasia and South America harbour the only ecosystems where *Nothofagus* trees can be found presently and were once united in the mega-continent Gondwana. The much higher genetic diversity of both *S. uvarum* and *S. eubayanus* found in the Southern Hemisphere, the high frequency of isolation and the *Nothofagus* association suggest that both species could be native to the Southern Hemisphere and that their primal niche is the *Nothofagus* system. Based on the basal position of the *S. uvarum*/*S. eubayanus* lineage, it seems plausible to hypothesize that the last common *Saccharomyces* ancestor thrived in the Southern Hemisphere in association with the *Nothofagus* system. In this scenario, an early derived lineage may have undergone speciation while associated with *Nothofagus*, originating *S. eubayanus* and *S. uvarum*, whereas another lineage may have migrated to the Northern Hemisphere through Southeast Asia, the only region of the globe where southern beech and oak relatives of the Fagaceae overlap (Hill 1992). The transition to the oak system could have been associated with the formation of other species, giving rise to endemic Asian taxa such as *S. arboricola* and *S. mikatae*, as well as to species with more global Holarctic distributions, such as *S. kudriavzevii*, *S. paradoxus* and *S. cerevisiae*.

CHAPTER 3

A population genomics insight into the Mediterranean origins of *Saccharomyces cerevisiae* wine yeasts

The work presented in this chapter is published in:

Almeida P, Barbosa R, Zalar P, Imanishi Y, Shimizu K, Turchetti B, Legras JL, Serra M, Dequin S, Couloux A, Guy J, Bensasson D, Gonçalves P, Sampaio JP: **A population genomics insight into the Mediterranean origins of wine yeast domestication.** *Molecular Ecology* 2015, **24**:5412-5427. DOI: 10.1111/mec.13341

The sequencing data generated in this study have been deposited in the National Center for Biotechnology Information short-read archive (NCBI-SRA) under accession code SRP059414, and in the European Bioinformatics Institute Nucleotide Archive (EBI-ENA) as PRJEB7675 and PRJEB7601.

Contributions:
Pedro Almeida developed all scripts and bioinformatics pipelines for the analyses of whole-genome sequence data.
Pedro Almeida executed all whole-genome sequence data analyses with help from Raquel Barbosa.
The contributions of the various authors were as follows:

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. performed microsatellite analysis;
P.A., P.G. and J.P.S. wrote the manuscript.

3.1 Summary

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, a balanced number of anthropic and natural *S. cerevisiae* strains is studied using an enlarged collection of natural lineages and employing whole-genome data of oak-associated wild isolates. This approach identified industrial variants and new geographically delimited populations, including a novel Mediterranean oaks population. This population is the closest relative of the wine lineage as shown by a weak population structure and further supported by genome-wide 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. It is also shown 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 oaks 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.

3.2 Introduction

Wild and domesticated lineages of *S. cerevisiae* 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.

This chapter describes the use of 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, it was analysed in detail the relationship of the wine group with a presently uncovered oak-associated Mediterranean *S. cerevisiae* population which is its closest wild relative. It is proposed that this new population contains the wild genetic stock that gave rise to the domesticated wine yeasts.

3.3 Materials and Methods

3.3.1 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.

3.3.2 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 distance 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.

3.3.3 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) (Appendix II Table AII.1). 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 *S. 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 in Chapter 2 (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'.

3.3.4 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. *S. 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 100000 iterations, followed by 50000 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 94089 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 31957 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.

3.3.5 Polymorphism and divergence analyses

Whole-genome levels of polymorphism were estimated using Variscan v2.0 (Hutter et al. 2006). For polymorphism analyses within populations, it was used an additional set of isolates (six from Portugal and fourteen from Japan), chosen randomly from the initial sampling 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 non-coding 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 replacement 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.

3.3.6 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 oaks population using a subset of non-coding 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: non-coding 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 non-coding regions, totalling 1286807 bp length, with 15857 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 13921 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 1129699 bp (calculated as 1286807*(13921/15857)).

3.3.7 *de-novo* assemblies and survey for the horizontally transferred regions A, B and C

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) were performed with Velvet v.1.2.08 (Zerbino & Birney 2008), to survey the 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. 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 10x. 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).

Local BLAST databases were set up for all the genomes available in this study and BLASTN searches were performed (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 sequences 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 2010) 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.

3.3.8 Multi-locus sequence analysis

Thirteen loci previously used to characterise 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 12680 positions in the final data set. Branch support was estimated from 1000 nonparametric bootstrap replicates.

3.4 Results

3.4.1 A geographically and genetically diverse collection of wild isolates

A preliminary characterization through microsatellite genotyping of a diverse collection of strains, including strains collected from the oak niche in different regions of the Mediterranean basin (Iberian Peninsula, France, Italy, Slovenia, Greece) and Japan, suggested a separation between isolates obtained

in natural and anthropic habitats (Appendix II Figure AII.1). 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, it was carried out a population genomic analysis using a comprehensive and more balanced strain data set that included the novel natural isolates (Appendix II Table AII.1). 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, the relationships of the 146 strains studied was summarised using a genome-wide phylogenetic network (Figure 3.1). 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 (Appendix II Table AII.1). 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. This 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.

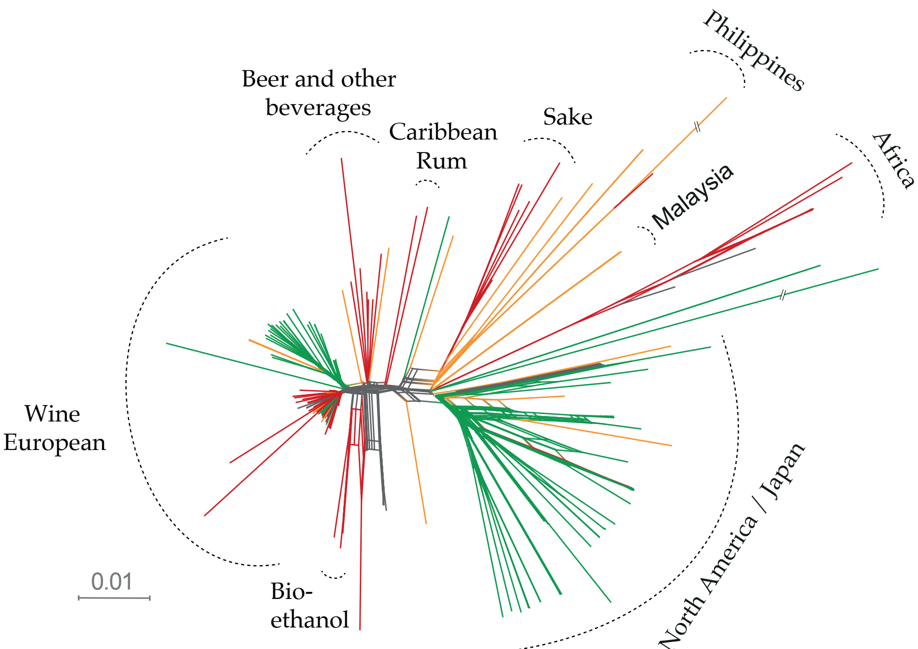


Figure 3.1 | The global diversity of *Saccharomyces cerevisiae* is shaped by both ecology and geography. Neighbour-net network of 146 strains based on 60331 SNPs, inferred with Kimura 2-parameter distance. Branches are coloured according to the substrate of isolation of the strain (red – fermentation; green – oak tree; orange – fruit; grey – other or unknown). The scale bar represents the number of substitutions per site.

3.4.2 Population structure and admixture

To delimit populations and infer possible admixture events, STRUCTURE (Falush et al. 2003) was used to test from 2 to 12 possible 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 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 (Figure 3.2) 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 oaks (2), Sake (3), Philippines (4) Africa (5) North American and/or Japanese populations (6–9). This analysis also 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 oaks 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 oaks isolates had a partial ancestry with the wine cluster (Figure 3.2).

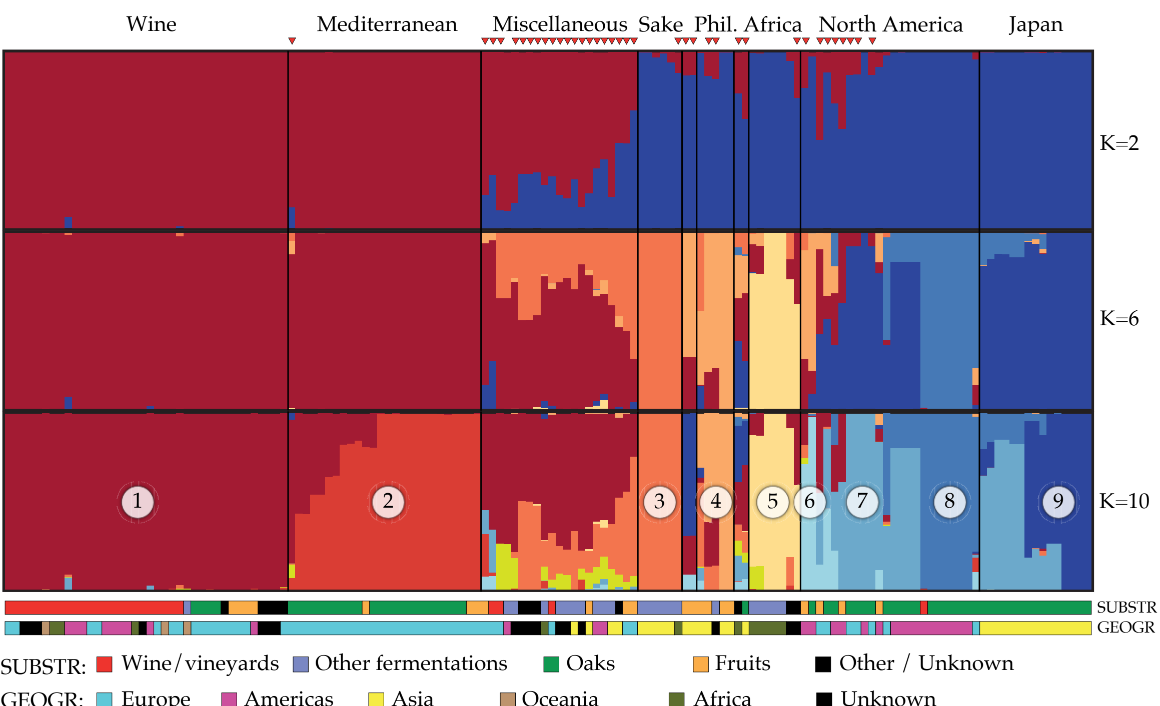


Figure 3.2 | Population structure and admixture in *S. cerevisiae*. STRUCTURE plots based on a subset of 9181 parsimony informative sites for K = 2, 6 and 10. Strains identified as mosaics at K = 2 are marked with a red triangle. Numbers 1–9 represent the different clusters that capture the maximum representation of population ancestry. The type of substrate of isolation (SUBSTR) and geographic source (GEOGR) are colour-coded.

Additionally, the clustering of strains was also performed 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 genome-wide coancestry matrix. Overall, it is 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 (Figure 3.3). Notably, fineSTRUCTURE hinted at a close relationship between wine and Mediterranean oaks strains, indicating a high degree of shared haplotype ancestry between the two groups. As already partially captured by STRUCTURE, Mediterranean oaks 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 oaks 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 (Figure 3.3). 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.

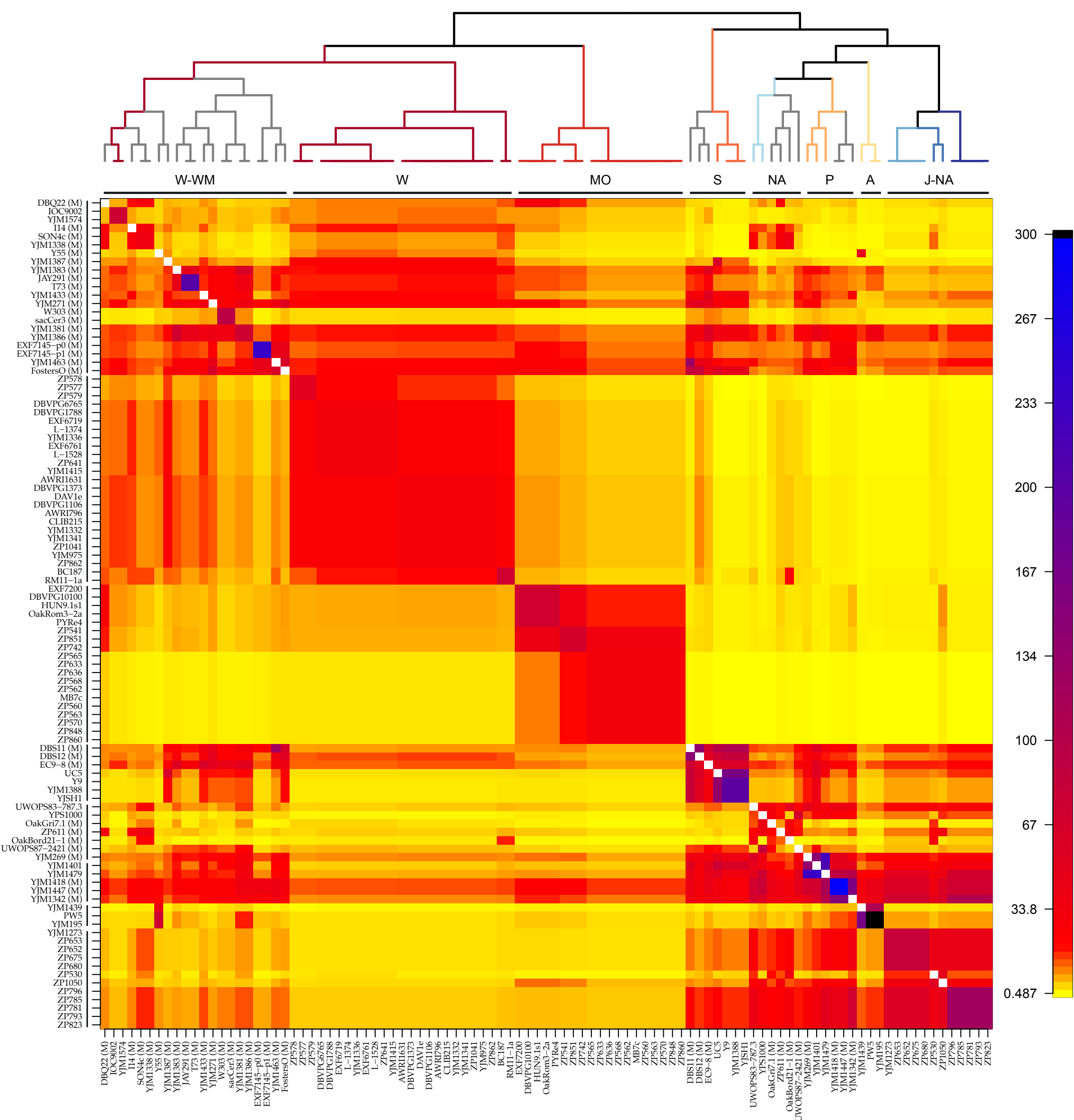


Figure 3.3 | Co-ancestry and population clustering in *S. cerevisiae*. fineSTRUCTURE co-ancestry matrix and population structure of 98 strains using 94089 informative biallelic SNPs. The colour of each bin in the matrix indicates the expected number of 'chunks' copied from a donor (column) to a recipient 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 (Figure 3.2) 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.

3.4.3 Genetic relationships among populations

As admixed genotypes are likely to hinder the elucidation of the phylogenetic relationships between populations, 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, were discarded. The six main clades detected in the maximum-likelihood phylogeny of 112 genome sequences (Figure 3.4) 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 oaks 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 system and from fruits collected in semi-wild systems (e.g. figs, wild apples) (Figure 3.4). 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.

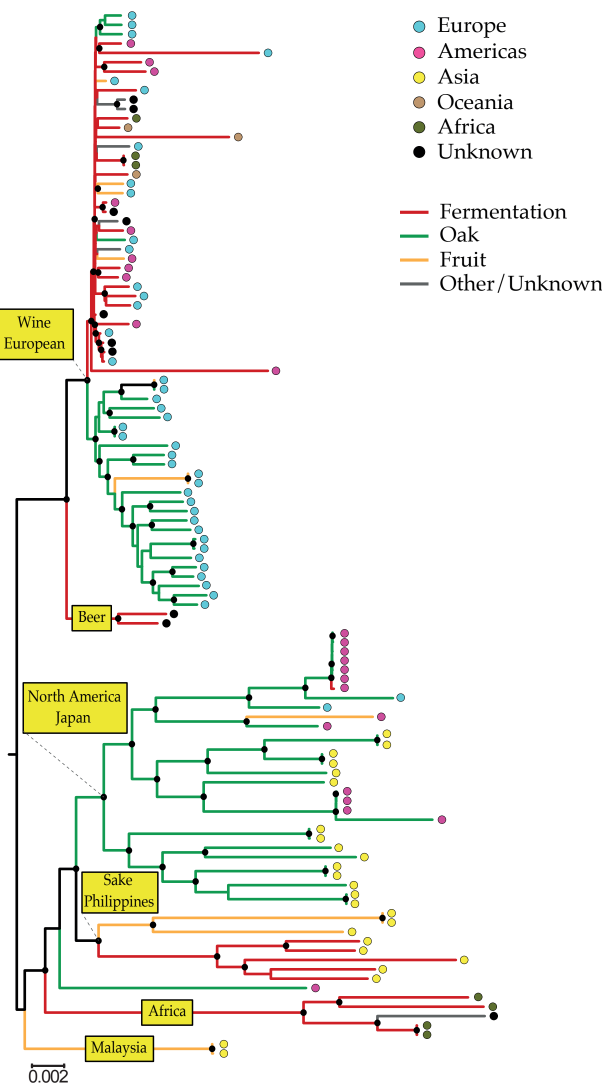


Figure 3.4 | Whole-genome phylogenetic relationships between *S. cerevisiae* strains. Rooted phylogenetic tree of 112 strains that excludes the mosaic strains identified by STRUCTURE at K = 2. The tree was inferred from 193071 SNPs, using the maximum-likelihood method as implemented in RAxML with the 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.

A phylogeny including the complete data set is depicted in Figure 3.5 below and shows that most mosaic strains are positioned outside the main clades mentioned above, an expected result given their recombinant nature.

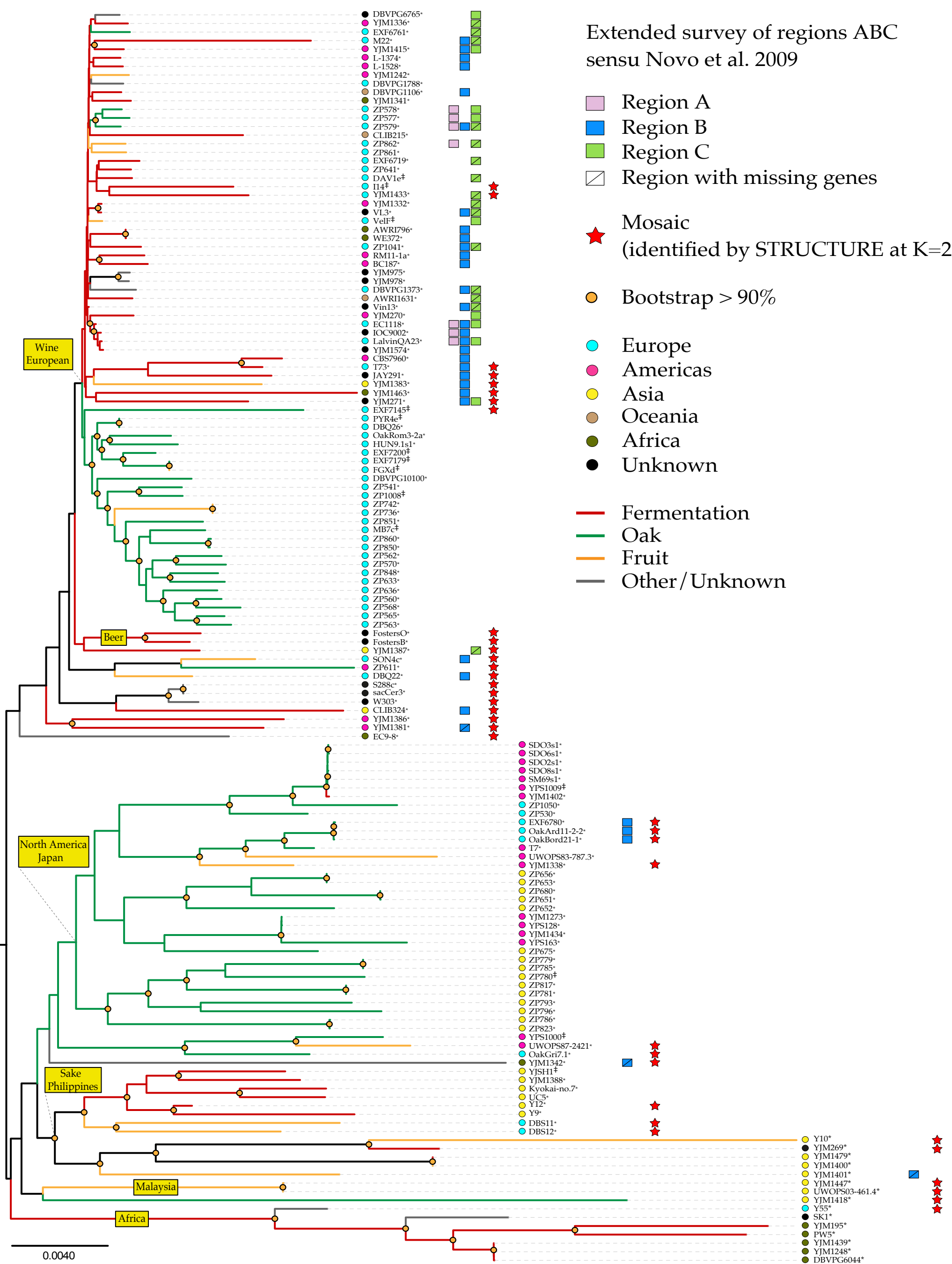


Figure 3.5 | Whole-genome phylogeny of the large strain dataset (145 strains) depicting those identified as mosaics by STRUCTURE at K=2 and the presence / absence of the wine-related regions A, B and C sensu Novo et al. 2009. The tree was inferred from 146097 SNPs, using the maximum likelihood method 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 orange dots in the respective tree nodes. Coloured squares illustrate the presence of regions A, B and C. These regions were search 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.

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, the published multilocus sequences of those lineages were used to construct an integrated phylogeny (Figure 3.6). 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 the strain data set used in this study. It also highlights that most of the *S. cerevisiae* genetic diversity that has been sampled is contained in the Asian lineages.

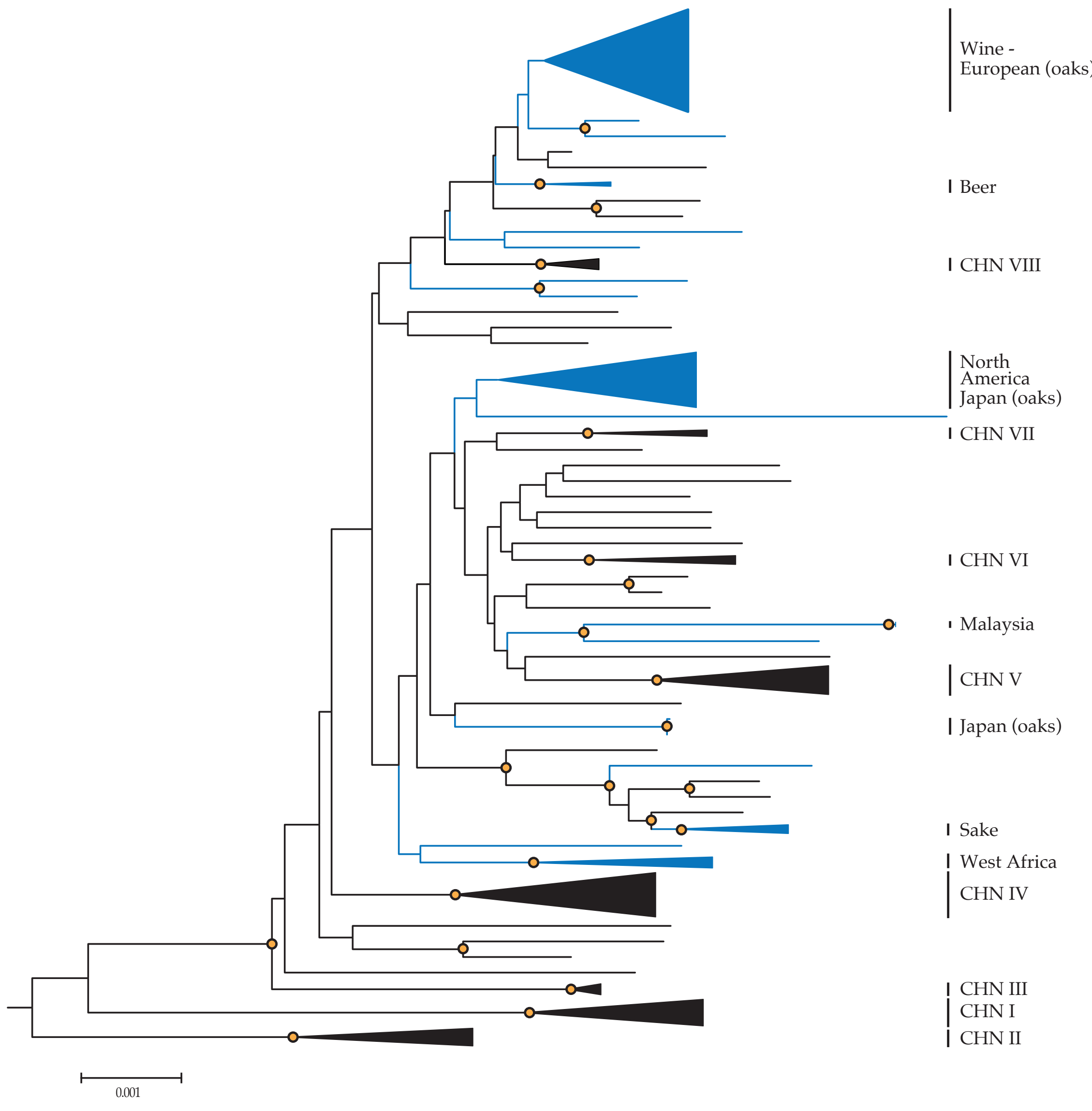


Figure 3.6 | Multilocus phylogeny of *S. cerevisiae* including the Chinese lineages. Neighbour-Joining tree inferred from a concatenated alignment of 13 loci using the Kimura 2-parameter model of sequence evolution. Branch lengths correspond to the expected number of substitutions per site. Support values 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 is not available. Branches coloured in blue indicate strains for which genome data is available and were used in this study.

The consistent placement of the 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 is 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 3.1). Moreover, the Wine and Mediterranean oaks 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 3.2). It is 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 oaks strains presenting mixed ancestry with the wine clade (see Figure 3.2 for K = 10) did not lower the amount of shared polymorphisms to levels equivalent to those of other comparisons included in Table 3.2. Additionally, nucleotide divergence, measured in nonoverlapping windows of 20 kb, between wine and Mediterranean oaks strains (π_B x 100 = 0.207) was significantly lower than the divergence between Japan and North America strains (π_B x 100 = 0.291) (2-sided t-test, P-value < 0.01) indicating a more recent population split for the former populations.

Table 3.1 | Pairwise whole-genome divergence between populations of *Saccharomyces cerevisiae*. Mean pairwise divergence between 2 alleles drawn from 2 populations (π_B x 100) are estimated per site from pairwise comparisons across the total length of the genome.

Population	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

MO – Mediterranean oak

Table 3.2 | Proportion of shared polymorphisms, fixed differences and private polymorphisms in pairwise comparisons of the wine group with other populations of *S. 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).

Population	Shared polymorphisms	Fixed differences	Private in Wine	Private in pop.	Total SNPs
MO	8.04	4.90	41.79	45.27	63345
MO[1]	6.15	7.74	50.47	35.64	54908
NA – Japan	4.41	14.04	18.49	63.06	143286
Japan	2.64	21.61	22.73	53.02	128192
North America	1.38	39.25	30.77	28.60	99596
Sake – Philippines	2.62	25.93	24.73	46.71	118201
Africa	2.17	43.43	30.31	24.09	97403

[1] strains with mixed ancestry with the wine clade were excluded
MO – Mediterranean oak; NA – North America

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

The Wine group has a higher ratio of polymorphism at replacement sites relative to fourfold degenerate sites and it also shows a significant reduction in the population recombination parameter (2-sided t-test, P-value < 0.05, Table 3.3), both of which are consistent with the effects of a domestication bottleneck. However, nucleotide diversity based on pairwise differences (θ_π x 100) measured in the Wine group and in the Mediterranean oaks group are very similar (Table 3.4, Table 3.3). Moreover, a comparison of diversity measurements between different wild lineages brings to light a strikingly lower diversity for the Mediterranean oaks 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 oaks population (Table 3.4).

Table 3.3 | Comparison of polymorphisms (mean values) at coding and non-coding regions between the Wine and Mediterranean oaks (MO) groups. For the Wine group only commercial strains or strains isolated from wine must were used.

	MO	Wine
Coding sequences		
Singletons	2.227	2.793
θ_W	0.001182	0.001357
θ_π	0.0009472	0.001042
Tajima's D	-0.6151	-0.7511
Hudson's C	83.3162	63.4570
Singletons (replacement)	1.078	1.421
θ_W (replacement)	0.0007320	0.0009179
θ_π (replacement)	0.0005652	0.0007029
Singletons (4-fold)	0.5484	0.6248
θ_W (4-fold)	0.002297	0.002391
θ_π (4-fold)	0.001903	0.001863
θ_π (replacement) / θ_π (4-fold)	0.2969	0.3774
Non-coding sequences		
Singletons	1.254	1.47
θ_W	0.001692	0.001837
θ_π	0.001398	0.001457
Tajima's D	-0.4858	-0.5928
Hudson's C	85.7866	61.4463

Table 3.4 | Whole-genome diversity within populations of *S. cerevisiae*. Diversity values (θ_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).

Population	Nº strains	Analysed sites	Segregating sites	θ_π	θ_W	Tajima's D
Wine	19	11216288	50112	0.0011166	0.0014049	-0.9316670
MO	31	11286153	50465	0.0009901	0.0012115	-0.7477315
Wine and MO	50	11216436	99638	0.0015320	0.0021143	-1.0448031
North America and Japan	42	11348218	112086	0.0025590	0.0024525	0.1685470
Japan	29	11373293	86405	0.0023487	0.0020841	0.5236821

MO – Mediterranean oaks

3.4.5 Population demographics

To further test the hypothesis that the Mediterranean oaks population is the best approximation for the wild ancestors of domesticated wine strains, different demographic models based on the folded joint site frequency spectra of Wine and Mediterranean oaks populations were considered using a diffusion-based approach as implemented in ∂a∂i (Gutenkunst et al. 2009). These analyses were performed using a subset of non-coding regions across the whole genome to minimize effects of selection that could interfere with demographic inference. The first model was that of complete isolation without migration, yielding correlated residuals between the model and the data, with the model predicting too few shared polymorphisms at low frequencies in the Wine population (Figure 3.7b). This effect has been shown to result from fitting data having a migration signal in a no-migration model (Gutenkunst et al. 2009). The next alternative scenario was that of asymmetrical migration between populations. Although explaining much of the shared variation (Figure 3.7c), 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 non-coding loci, these molecular signatures are usually suggestive of recent population growth, in line with the genome-wide negative Tajima's D values estimated for Wine and Mediterranean oaks populations (Table 3.3, Table 3.4). Therefore, the third demographic model included population growth after the split from an ancestral population, together with asymmetric migration between populations (Figure 3.7d). This population growth model fitted better most of the private polymorphisms in both populations, although it is possible to 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 (Figure 3.7 and Table 3.5).

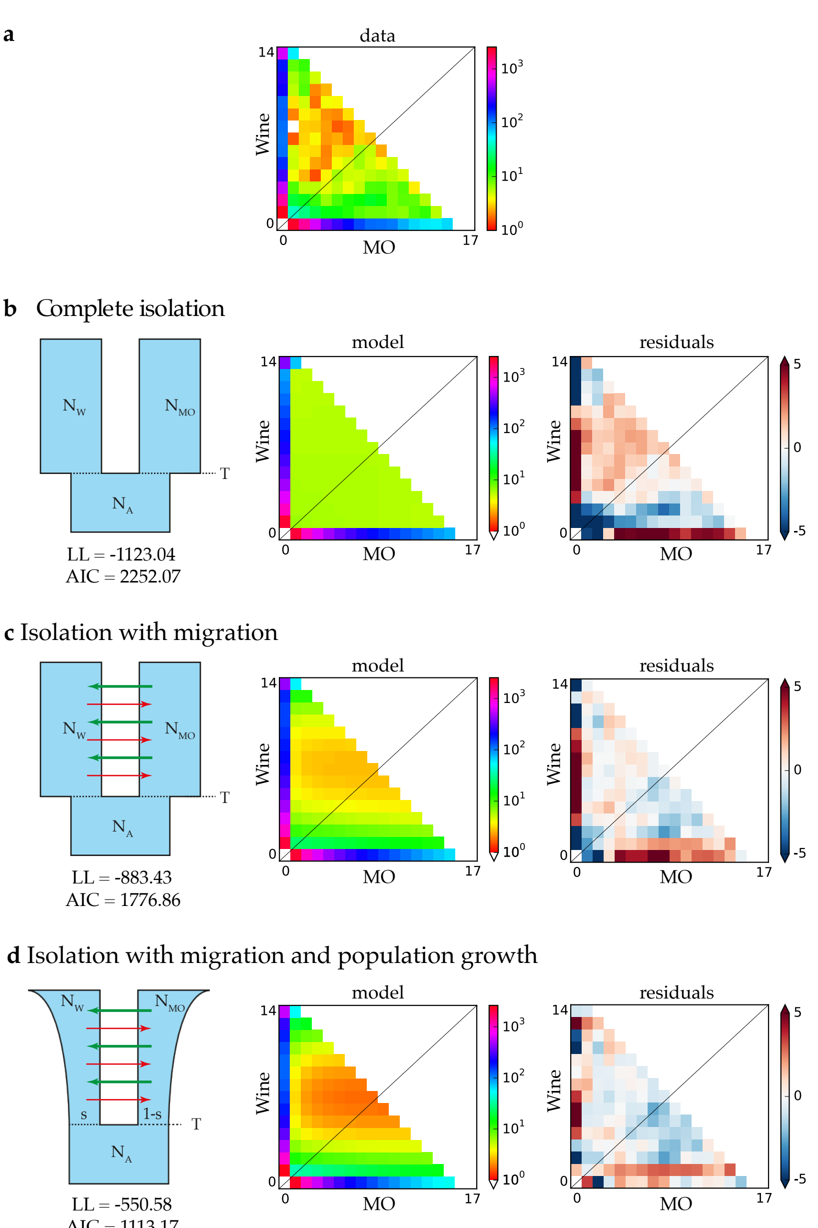
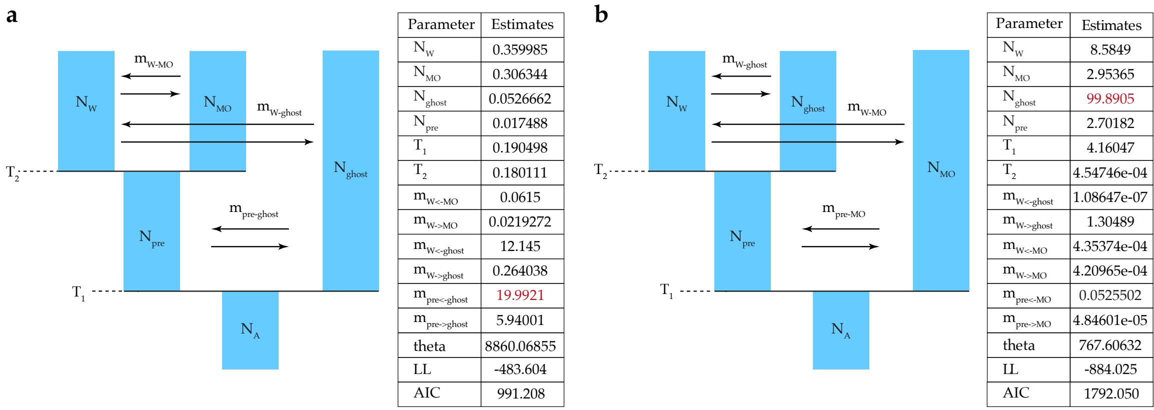


Figure 3.7 | Joint allele frequency spectrum for the Wine and Mediterranean oaks populations of *S. cerevisiae* and comparison with the expected frequency spectrum for different demographic scenarios. (a) Representation of the folded joint frequency spectrum, in the form of a heatmap, using 1286807 bp of non-coding regions across the genome. X and Y axes represent the number of chromosomes (strains) in the Mediterranean oaks (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 normalised 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, MO 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 oaks population). T is the time of the split.

Table 3.5 | Best-fit parameter estimates of three alternative demographic models for the joint allele frequency spectrum for the Wine and Mediterranean oaks populations. 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 oaks (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.

Model	s	N_W	N_{MO}	T_{split}	$M_{MO\text{-}W}$	$M_{W\text{-}MO}$	theta	k	Log likelihood	AIC
No migration	-	1.4147	1.3258	1.2655	-	-	1631.5561	3	-1123.0363	2252.07
Asymmetric migration	-	3.0760	3.3229	4.8954	0.0471	0.0014	676.5082	5	-883.4304	1776.86
Asymmetric migration with growth	0.2224	3.9279	2.9352	2.5661	0.0907	0.0056	1090.8886	6	-550.5826	1113.17

It was also explored the possibility that a yet-unidentified lineage could be more related to the Wine group than the MO population. Three-populations scenarios in which the wine yeasts either derive from the MO population or from an unsampled population were simulated as above (Figure 3.8). In addition, it was also allowed for asymmetrical migration between any two populations present at a given time. The lowest AIC value was obtained for the scenario where wine yeasts derive from the MO population, thus supporting the common ancestry of these two populations. The next lowest AIC value was obtained for the competing scenario, but the divergence time between wine yeasts and the unsampled population was not explained in a biological meaningful interpretation. The split was estimated to be between 0.16 and 1.28 years depending on whether 8 or 1 generations per day were used to translate time from generations to years, respectively. Together, these results do not seem to support the hypothesis of an unsampled population that is the closest relative of the wine yeasts when the Mediterranean oaks population is considered in the model. Moreover, convergence of parameters among five independent runs was never observed for either scenario and the fits for the parameters regarding migration involving the ancestral population or the population size of the unsampled population were always at the bounds of the parameter space. A possible explanation for this behaviour could be the inadequacy of the models employed to describe the data, and for this reason neither model was considered further.



Parameter	Estimates
N_W	0.359985
N_{MO}	0.306344
N_{ghost}	0.0526662
N_{pre}	0.017488
T_1	0.190498
T_2	0.180111
$m_{W<-MO}$	0.0615
$m_{W->MO}$	0.0219272
$m_{W<-ghost}$	12.145
$m_{W->ghost}$	0.264038
$m_{pre<-ghost}$	19.9921
$m_{pre->ghost}$	5.94001
theta	8860.06855
LL	-483.604
AIC	991.208

Parameter	Estimates
N_W	8.5849
N_{MO}	2.95365
N_{ghost}	99.8905
N_{pre}	2.70182
T_1	4.16047
T_2	4.54746e-04
$m_{W<-ghost}$	1.08647e-07
$m_{W->ghost}$	1.30489
$m_{W<-MO}$	4.35374e-04
$m_{W->MO}$	4.20965e-04
$m_{pre<-MO}$	0.0525502
$m_{pre->MO}$	4.84601e-05
theta	767.60632
LL	-884.025
AIC	1792.050

Figure 3.8 | Competing models for the demographic scenario between wine and Mediterranean oaks populations incorporating an unsampled population. The joint allele frequency spectrum of the Wine and Mediteranean oaks (MO) population was compared taking into consideration the presence of an unsample ("ghost") population. Tables show the parameter and respective ∂a∂i estimation. **(a)** the MO population is the closest relative of wine yeasts. **(b)** the "ghost" population is the closest relative of wine yeasts.

The best-fit demographic scenario obtained from the two-populations model was used to obtain converging estimates of demographic parameters (Figure 3.7; Table 3.6). Estimated population migration rates were relatively low, but a much higher migration from the Mediterranean oaks population into the wine population than in the opposite direction was detected ($M_{MO \to W} \simeq 0.36$ and $M_{W \to MO} \simeq 0.02$, Table 3.6). Population effective sizes were on the same order of other estimates for the European and Far Eastern populations of *S. paradoxus* (Tsai et al. 2008). Interestingly, while the current effective size of the Wine population ($Ne \simeq 5.8 \times 10^6$) was estimated to be higher than that of the Mediterranean oaks population ($Ne \simeq 4.3 \times 10^6$), ∂a∂i inferred a strong bottleneck in the Wine population at the time of the split ($s \simeq 0.22$) (Table 3.6), as predicted by a 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 10300 ya, respectively. This divergence time is relatively recent, in agreement with the low number of fixed differences between the two populations (4.9%, Table 3.2) and with historical evidence for winemaking. The estimated time for the most recent common ancestor of wine and Mediterranean oaks strains is compatible with the first biochemical evidence of wine, dated to 5400–5000 BC (McGovern et al. 1996).

Table 3.6 | 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 oaks (MO) populations. Bias-corrected 95% confidence intervals were obtained from 100 simulated datasets 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 oaks population). N_e is effective population size. Time is given per generation. Migration is the effective number of migrants per generation.

Parameter	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 oaks (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

3.4.6 Domestication fingerprints

A set of genetic fingerprints related to wine fermentation was surveyed in all the strain dataset 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 oaks lineage that were surveyed (Figure 3.5). 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 (Figure 3.5) 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.

3.5 Discussion

Here, it was employed 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. It is demonstrated 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). It can be hypothesized that colonization of North America followed a migration route over the Bering Strait land bridge as documented in other cases (Hewitt 2004). Although more detailed comparisons of Asian and North American populations are needed, the phylogenetic and population genomic analyses presented here 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. It is possible 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, it was here identified 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 (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*, it was also sporadically isolated pure (non-mosaic) 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 back-slopping practice of skimming off the surface of the best musts for use in later fermentations (Cavalieri et al. 2003) may have fostered the unintended selection of the best strains as wine yeasts have better oenological properties than wild strains (Hyma et al. 2011).

The close relationship between the Mediterranean oaks 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, these analyses are consistent with the hypothesis that the common ancestor between Mediterranean oaks 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 oaks population. However, 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).

Testing for different demographic models using the diffusion-based approach indicated that the best explanation for the observed relationship between Mediterranean oaks 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 oaks population into the Wine population was detected. These results also 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, the demographic inference and the weak population structure between wine and Mediterranean oaks populations are compatible with a Mediterranean oaks 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 oaks 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. It is noted 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 oaks 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, it was 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 oaks 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). 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 in Chapter 2. 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. The results presented in this chapter advance the knowledge of wine yeast domestication by revealing the closest wild relatives of domesticated lineages and the wild genetic stock that underwent domestication.

CHAPTER 4

From Wild to Domesticated – Divergence and Adaptive Evolution Between Wine and Wild Yeasts

---

The work presented in this chapter is, at the moment of the submission of this thesis, under review in Molecular Ecology:

Almeida P, Barbosa R, Bensasson D, Gonçalves P, Sampaio JP: **From Wild to Domesticated – Divergence and Adaptive Evolution Between Wine and Wild Yeasts.** *Submitted for peer review.*

Contributions:
Pedro Almeida developed all scripts and bioinformatics pipelines for the analyses of whole-genome sequence data. Pedro Almeida executed all whole-genome sequence data analyses with help from Raquel Barbosa.

D.B., J.P.S., P.A. and P.G. conceived and designed the research;
P.A. and R.B. obtained and assembled genomic data;
P.A., P.G. and J.P.S. wrote the manuscript.

4.1 Summary

The fermentation of grape must into wine epitomizes the use of microbes for the production of foods and beverages. Yet, the genetic changes associated with the domestication of wine yeasts are still poorly understood and a comprehensive survey of signatures of divergent selection between wine yeasts and natural populations is lacking. Here, population genomics was used to estimate the degree and distribution of nucleotide variation between wine yeasts and their closest wild relatives, a recently uncovered population associated with Mediterranean oaks. It was found widespread genome-wide divergence which was unexpectedly higher at non-coding sites. These findings are discussed in the light of adaptive evolution and relaxed selection in the Wine population, presumably as by-products of domestication. Extensive divergence was also found in trans-acting DNA binding proteins, likely indicating an important influence on gene expression. Furthermore, it is shown that the genetic nature of divergence between wine and wild yeast affects multiple genes and identify nine outlier regions putatively under strong divergent selection. It was also identified two cases of introgression from the sibling species *S. paradoxus*, involving the genes *FZF1* and *SSU1*. Both genes are mostly known for their role in sulphite resistance, a trait typically relevant for wine yeasts. However, because the introgressions have not been transmitted to wine strains, it is hypothesised that divergent ecological selection segregated the two forms between the different niches. These results provide new insights into the general mechanisms of adaptation in a eukaryotic microbe and illuminate the transition from wild to domesticate of the most relevant biotechnology workhorse.

4.2 Introduction

In spite of the advances made to understand the functional and genomic variations in wine yeasts, a comprehensive genome-wide examination for signatures of divergent selection between wine yeasts and natural populations is lacking. Among the complex population structure of *Saccharomyces cerevisiae*, overrepresented by lineages associated with distinct fermentations and by admixed strains that are inter-lineage recombinants (Strope et al. 2015; Ludlow et al. 2016), natural lineages exclusively associated with habitats not dominantly affected by human activities are underrepresented. In spite of a provocative proposal arguing that *S. cerevisiae* does not have a natural niche (Goddard & Greig 2015), evidence has gradually accumulated implicating oaks and other trees of the Fagaceae as the most likely natural habitats of *S. cerevisiae* in temperate regions of the Northern Hemisphere (Naumov et al. 1998; Sampaio & Gonçalves 2008; Wang et al. 2012; Hyma & Fay 2013). Moreover, the work presented in Chapter 3 suggests that a newly found wild *S. cerevisiae* population associated with oaks in the Mediterranean region, hereafter MO population, is the closest natural relative of domesticated wine yeasts, hereafter Wine population (Almeida et al. 2015). Here, this novel finding and associated unprecedented population-scale genomic data was used to study the evolutionary changes coupled with wine yeast domestication. It was found that the divergence between wild and domesticated groups was genome-wide and likely to impact gene expression. This study also identified putative regions of divergent selection between the two populations and cases of introgression in the MO population that, interestingly, have not been transmitted to the wine strains. These results provide new

insights into the general mechanisms of adaptation in a eukaryotic microbe and illuminate the transition from wild to domesticate of the most relevant biotechnology workhorse.

4.3 Materials and Methods

4.3.1 Genome Sequencing, Read Alignment, and Genotype Calling

Paired-end whole-genome data were obtained for a subset of fourteen new isolates used in this study. DNA was extracted from overnight cultures of monosporic derivatives and pair-end sequenced using the Illumina MiSeq system. Genomic information for other isolates was obtained from public databases (Appendix III Table AIII.1). Where only finished genome sequences were available, the corresponding error- free Illumina reads were simulated using dwgsim (http://sourceforge.net/apps/mediawiki/dnaa/).

Reads for each isolate were mapped to *S. cerevisiae* reference genome (UCSC version sacCer3) using SMALT v0.7.5 aligner (http://www.sanger.ac.uk/science/tools/smalt-0). 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 in Chapter 2 (Almeida et al. 2014). 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".

4.3.2 Phylogenetic Inference

Chromosomal parsimony informative single nucleotide polymorphisms (SNPs) were extracted from multiple sequence alignments of chromosomes only if the evaluated site was represented by unambiguous high-confidence alleles in more than 85% of isolates. The evolutionary history was inferred from an unrooted phylogenetic tree constructed by the Neighbour-Joining method with 100 bootstrap replicates and using the p-distance to compute evolutionary distances as implemented in MEGA7 (Kumar et al. 2016).

4.3.3 Data Filtering

For polymorphism and selection analyses at the whole genome level, strains with more than 10% of missing data were excluded in order to maximize the total number of sites considered. Furthermore, strains with nearly identical genotypes, assessed visually by the presence of very short branches in the phylogenetic tree and later confirmed to be separated by only ~10 SNPs, were also removed since these would shift the allele frequencies without reflecting a potential signature of selection. In this process, the genome with the highest proportion of Q40 bases was retained. Unless otherwise stated, only

positions with at least 85% of valid sites in both populations were used in calculations. Singleton sites were also excluded because these are uninformative to capture signatures of drift and hitchhiking in genome scans (Roesti et al. 2012). In addition, transposable elements and low complexity sequences were masked with RepeatMasker v4.0.6 (http://www.repeatmasker.org) for each chromosome individually using sensitive settings: -engine crossmatch -species 'Saccharomyces cerevisiae' -s -no_is -cutoff 225 -frag 20000.

4.3.4 SNP Polymorphisms and Allele Frequency Differences

Levels of shared and private polymorphisms and fixed differences across the whole-genome were estimated using software based on the libsequence library (Thornton 2003). The per-SNP absolute allele frequency difference (ΔAF) between domestic and wild *S. cerevisiae* populations was calculated using the formula: $\Delta AF = abs(Wine_{AFminor} - Wild_{AFminor})$. AFminor refers to the observed frequency of the minor allele in the wild population as reference, although we note that the result would be the same otherwise because only the absolute value is considered. Using this formula, $\Delta AF = 1$ for any fixed difference. The online Variant Effect Predictor tool (McLaren et al. 2010) from Ensembl release 84 (http://www.ensembl.org/info/docs/tools/vep/index.html) was used to classify the genomic variants as coding, non-coding (intergenic and intronic), synonymous and replacement, based on the annotations of *S. cerevisiae* reference genome assembly R64-1-1 (same as sacCer3).

4.3.5 Genome Scans of Divergence and Polymorphism

Sliding window analyses of divergence were performed using software based on the libsequence library (Thornton 2003). The window size was set to 5 kb with 500 bp step increments, retaining only those windows where at least 2500 sites were used in the calculations after filtering for positions with less than 85% of high quality bases (Q40). FST, the fixation index, and πB, the mean pairwise divergence between 2 alleles drawn from 2 populations, were estimated according to the methods of Hudson (1992) and Charlesworth (1998) , respectively, without weighting for sample sizes. The combination of both measures can, in principle, leverage the effects of random genetic drift and variable recombination rate across the genome (Noor & Bennett 2009; Cruickshank & Hahn 2014). Given the low diversity within each population being compared, spurious fixation or near- fixation signals are more likely to be captured if the windows have insufficient numbers of polymorphic loci. We note that of all the valid 22486 5kb windows with at least one SNP, all but 297 (1.32%) contained at least 10 SNPs, and none mapped within the candidate divergent regions identified in this study. Z-scores, in units of standard deviations from the mean, were calculated for each window and a threshold of Z-score > 3 was set to define candidate divergent regions, representing the extreme high end of the distribution. Contiguous windows above the threshold were merged in a single divergent region. Measures of nucleotide diversity within populations (π and Tajima's D) were calculated in sliding windows as above using Variscan v2.0 (Hutter et al. 2006). All statistical analyses were performed in R v3.2.2 (https://www.r-project.org).

4.3.6 Gene Ontology Analyses

Standard GO (gene ontology) term discoveries was performed with the GO Term Finder tool v0.83, available at Saccharomyces Genome Database (SGD, http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl), using a p-value cutoff of 0.05. P values obtained from the GO Term Finder are calculated using a hypergeometric distribution with multiple hypothesis Bonferroni correction.

4.3.7 Screening for introgressions from other Saccharomyces species

Evidence of introgressions from other *Saccharomyces* species were searched by mapping the reads to a combined reference that includes all the available annotated coding sequences of six *Saccharomyces* species (*S. arboricola*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus* and *S. uvarum*) (Scannell *et al.* 2011; Liti *et al.* 2013, Saccharomyces Genome Database) . Reads were mapped to this combined reference using BWA v0.6.2 (Li & Durbin 2010) with default parameters but setting the quality threshold to 10 (-q 10). SAMtools v1.1852 (Li et al. 2009) was used for the manipulation of the resulting BAM files. Only genes with orthologs unambiguously annotated in all six species were analysed. An ORF was considered to have a foreign origin to *S. cerevisiae* if its coverage was at least higher than one-fourth of the median whole-genome coverage for the analysed strain. The ORF coverage was defined as the product of the total number of mapped reads to the orthologous ORFs by the read size, dividing by the sum of the length of each ORF, considering only the ones with more than 25% of reads mapped (relative to the orthologous ORF with the highest number of reads) to control for spurious alignment counts. This coverage threshold allowed for some heterogeneity in the read counts and for the eventual presence of a foreign ORF together with the native *S. cerevisiae* ORF.

4.3.8 Analysis of *FZF1*, *SSU1* and *DTR1* introgressions

Sequence information for the coding sequence of *FZF1*, *SSU1* and *DTR1* was collected from the de-novo genome assemblies available (Appendix III Table AIII.1 and references therein) and from other *Saccharomyces* species genomes (Scannell et al. 2011). For the new strains used in this study we also performed de-novo assemblies using SPAdes v.3.1.0 (Bankevich et al. 2012). Prior to assembly, reads were trimmed based on a quality score threshold of 20 and discarded if ended up with less than 100 bp of length or with any "Ns" on them. A local BLAST database was then set up for each genome and the introgressed ORFS were retrieved by BLASTN, using the correspondent *S. cerevisiae* ORF sequences available at SGD as queries. Strains where the coding sequence was incomplete were excluded from the analyses individually for each locus. Sequences were first aligned in nucleotide space in order to identify and remove single-base insertions that could be the result of sequencing errors, and then realigned in protein space using the MUSCLE algorithm (Edgar 2004). Where heterozygous sites were found, one of the two observed nucleotides was randomly chosen to represent the position. Maximum likelihood phylogenies were estimated in RAxML v8.2.4 (Stamatakis 2006) using the rapid bootstrap algorithm with 100 replicates and the GTR+Γ model of sequence evolution. Typical estimations of polymorphism and divergence were calculated using software based on the libsequence library (Thornton 2003). Simulations of the neutral evolutionary model were performed in *ms* (Hudson 2002) using 10,000 independent samples and fixing theta to the observed values in the data. Sliding window

analyses of divergence at synonymous and replacement sites were performed in DnaSP v5.10.1 (Librado & Rozas 2009) using a window size of 25 bp with 1 bp step, relative to the sites being considered. In order to delimit the extent of the introgression observed for *FZF1*, we analysed *FZF1* and its flanking genes: 14 kb from *ADH4* through *RTG2*. Sequences were aligned with FSA v1.15 (Bradley et al. 2009) and sliding window analyses of divergence were performed on overlapping windows of 200 sites with a step size of 50 sites using Variscan v2.0 (Hutter et al. 2006).

4.4 Results

4.4.1 Population Genomic Data

Publicly available whole-genome data for *S. cerevisiae* wine yeasts was analysed together with the recently described population isolated from Mediterranean oaks (MO) (Almeida et al. 2015), which so far represents the best proxy for a shared ancestral history with wine yeasts. Furthermore, it was also included in the present analysis an additional set of fourteen strains isolated from wine or available as commercial wine yeast strains (Appendix III Table AIII.1), which were confirmed to cluster within the Wine yeast group (Figure 4.1). After filtering and removing strains with more than 10% of missing SNP data and redundant genotypes (see Methods section), the final dataset comprised a total of 51 genomes, with 27 strains from Wine and 24 strains from MO populations. This joint analysis identified a total of 75351 high quality (>Q40) polymorphic SNPs, excluding singletons, distributed on all nuclear chromosomes and across the combined sample, of which 63210 SNPs were called in more than 85% of genomes in each population. Of these, 5713 (about 9%) SNPs represented shared polymorphisms and 2537 (about 4%) were fixed differences between the populations. 32695 SNPs were identified as being private to the Wine population and the remaining 22265 SNPs were only observed in the MO population in line with previous observations (Chapter 3; Almeida et al. 2015).

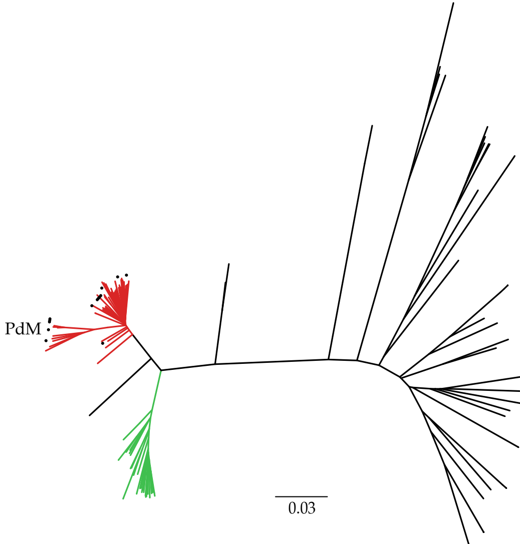


Figure 4.1 | Genetic relationships between *Saccharomyces cerevisiae* strains. The evolutionary history was inferred using the Neighbour-Joining method based on 257,163 high quality parsimony informative SNPs with less than 15% of missing data. Genetic distances were estimated by the p-distance method. Branch lengths correspond to the number of nuceotide differences per site. Highlighted strains were assigned to either the Wine (red) or Mediterranean oaks (green) populations. The Prise de Mousse (PdM) collection of champagne yeasts is also identified.

4.4.2 Distribution of Single Nucleotide Polymorphisms (SNPs)

The distribution of SNPs was further explored by comparing different classes of sites to synonymous sites, assuming the latter should provide a reasonable approximation to neutral evolution, in a modified form of the McDonald-Kreitman test (McDonald & Kreitman 1991). In this way, fixed differences between populations were taken as a measure of divergence that could be compared to polymorphisms. When analysing the combined polymorphism data in both populations, it was found a significant difference between divergence and polymorphism when considering either non-coding sites or replacement sites compared to synonymous sites ($P = 0.013$ and $P < 10$-4 respectively, Fisher's exact test) (Table 4.1). The departure from neutrality, as measured by the direction of selection (DoS) index (Stoletzki & Eyre-Walker 2011), was however different for the two classes of sites. Whereas non-coding sites showed an excess of divergence (DoS > 0), fixed differences were underrepresented at replacement sites (DoS < 0). Overall, it was possible to estimate that approximately 12% more substitutions, 95% CI [3%, 20%], have been fixed between the Wine and MO populations in non-coding sites relative to synonymous sites (Table 4.1). The ratio of polymorphism to divergence at non-coding and replacement sites, relative to synonymous sites, was also different when each population was analysed separately. In the MO population but not in Wine ($P < 10^{-4}$ and $P = 0.095$ respectively), unexpectedly few polymorphisms were found at non-coding sites (Table 4.1). In contrast, when considering replacement sites, the results were only significant in the Wine population, with a considerable skew towards the enrichment of polymorphisms at replacement sites (DoS = -0.103). Interestingly, while the total number of polymorphisms in the MO population is lower in replacement sites than in synonymous sites (pN/pS < 1), this pattern is reversed in the Wine population (pN/pS > 1).

Table 4.1 | Fixed differences and polymorphism in coding and non-coding DNA between Wine and Mediterranean oaks (MO) populations.

Class	Fixed	Polymorphic	*P* value[a]	DoS[b]	α[c]
Wine + MO					
Synonymous	854	19659	-	-	-
Replacement	597	17117	$<10^{-4}$	-0.054	-
Non-coding	959	19383	0.007	0.032	0.122
Wine					
Synonymous	854	9784	-	-	-
Replacement	597	10360	$<10^{-13}$	-0.103	-
Non-coding	959	10117	0.095	0.021	0.079
MO					
Synonymous	854	7895	-	-	-
Replacement	597	5595	0.823	-0.003	-
Non-coding	959	7300	$<10^{-4}$	0.048	0.177

[a] All *P* values are from two-tailed Fisher's exact tests in comparison with synonymous sites
[b] Direction of selection (DoS) is calculated according to Stoletzki and Eyre-Walker 2011.
[c] α estimates the proportion of extra nucleotide fixed substitutions

Highly differentiating SNPs are more likely to have been targets of selection or to occur in the neighbourhood of other SNPs under selection. The absolute allele frequency difference between wild

and domesticated yeast populations was calculated for each individual SNP and classified the most divergent ones (ΔAF > 0.8) as replacement or as found within 500 bp upstream of the closest open reading frame (ORF). The results were consistent when considering 1000 bp of upstream region (data not shown). These most divergent SNPs were then binned into the respective ORF and a Gene Ontology (GO) search was performed. Given the possibility of population bottlenecks in the Wine and MO populations (Almeida et al. 2015), which can increase the likelihood of random highly differentiating SNPs throughout the genome as a consequence of random genetic drift, only those ORFs with more than one associated SNP were considered. In total, 298 ORFs had two or more SNPs in their 500 bp upstream region and 303 ORFs had 2 or more replacement SNPs. Considering the top 15 GO Slim process terms found in higher frequency than the background set, the response to chemical, meiotic cell cycle, carbohydrate metabolic process, cellular response to DNA damage stimulus, DNA repair and signalling categories were common in both sets of ORFs (Appendix III Table AIII.2 and Table AIII.3). Overall, although there were no significant GO categories for the 500 bp upstream SNPs with ΔAF > 0.8, it was found an enrichment of many GO categories when considering the replacement SNPs. All significantly over-represented categories were related with DNA binding and transcription functions (Table 4.2), and included a variety of different processes such as "GO:0006355 regulation of transcription, DNA-templated", "GO:0050896 response to stimulus", "GO:0007154 cell communication", "GO:0016049 cell growth", "GO:0000003 reproduction" and "GO:0051173 positive regulation of nitrogen compound metabolic process", among others (Table 4.2).

Table 4.2 | Significant Gene Ontology (GO) terms in function and process categories for ΔAF > 0.8 SNPs in replacement sites.

GO ID	GO category	Frequency (%)	P value	FDR
Function				
GO:0001071	nucleic acid binding transcription factor activity	8.9	2.67E-07	0
GO:0003700	transcription factor activity, sequence-specific DNA binding	8.9	2.67E-07	0
GO:0003677	DNA binding	14.2	1.33E-06	0
GO:0043565	sequence-specific DNA binding	10.6	3.35E-06	0
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	5.9	0.00018	0
GO:0000976	transcription regulatory region sequence-specific DNA binding	5.0	0.00072	0
GO:0000987	core promoter proximal region sequence-specific DNA binding	4.3	0.00074	0
GO:0001159	core promoter proximal region DNA binding	4.3	0.00088	0
GO:0000982	transcription factor activity, RNA polymerase II core promoter proximal region sequence-specific binding	4.3	0.00123	0
GO:0044212	transcription regulatory region DNA binding	5.3	0.00136	0
GO:0000975	regulatory region DNA binding	5.3	0.00136	0
GO:0001067	regulatory region nucleic acid binding	5.3	0.00175	0
GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	3.6	0.01374	0
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	3.3	0.01534	0
GO:0003690	double-stranded DNA binding	6.3	0.02339	0
GO:0001012	RNA polymerase II regulatory region DNA binding	3.6	0.02436	0

Table 4.2 (continued)

GO ID	GO category	Frequency (%)	P value	FDR
GO:0001077	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	3.0	0.03196	0
GO:1990837	sequence-specific double-stranded DNA binding	5.3	0.03602	0
GO:0001228	transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding	3.0	0.03772	0
Process				
GO:0065007	biological regulation	35.6	2.91E-05	0
GO:0044699	single-organism process	56.8	0.00018	0
GO:0035556	intracellular signal transduction	6.9	0.00023	0
GO:0050789	regulation of biological process	30.0	0.00051	0
GO:0050794	regulation of cellular process	29.0	0.0008	0
GO:0045935	positive regulation of nucleobase-containing compound metabolic process	11.6	0.00087	0
GO:0048585	negative regulation of response to stimulus	3.6	0.00159	0
GO:0050896	response to stimulus	23.8	0.00215	0
GO:0031325	positive regulation of cellular metabolic process	13.2	0.00279	0
GO:0048522	positive regulation of cellular process	15.2	0.00301	0
GO:0009893	positive regulation of metabolic process	13.2	0.00438	0
GO:0016049	cell growth	5.0	0.00439	0
GO:0019219	regulation of nucleobase-containing compound metabolic process	17.2	0.0049	0
GO:0007154	cell communication	10.9	0.00521	0
GO:1902680	positive regulation of RNA biosynthetic process	10.2	0.00548	0
GO:1903508	positive regulation of nucleic acid-templated transcription	10.2	0.00548	0
GO:0045893	positive regulation of transcription, DNA-templated	10.2	0.00548	0
GO:0023057	negative regulation of signaling	3.0	0.00886	0
GO:0009968	negative regulation of signal transduction	3.0	0.00886	0
GO:0010604	positive regulation of macromolecule metabolic process	12.5	0.00891	0
GO:0051254	positive regulation of RNA metabolic process	10.2	0.00994	0
GO:0023052	signaling	8.6	0.011	0
GO:0051173	positive regulation of nitrogen compound metabolic process	11.6	0.01105	0
GO:0010648	negative regulation of cell communication	3.0	0.01149	0
GO:0007124	pseudohyphal growth	4.3	0.01424	0
GO:0044763	single-organism cellular process	49.8	0.01473	0
GO:0006351	transcription, DNA-templated	17.2	0.01784	0
GO:0097659	nucleic acid-templated transcription	17.2	0.01784	0
GO:0032774	RNA biosynthetic process	17.2	0.01936	0
GO:0044700	single organism signaling	8.3	0.02254	0
GO:0007165	signal transduction	8.3	0.02254	0
GO:1903506	regulation of nucleic acid-templated transcription	14.9	0.02382	0

Table 4.2 (continued)

GO ID	GO category	Frequency (%)	P value	FDR
GO:2001141	regulation of RNA biosynthetic process	14.9	0.02382	0
GO:0006355	regulation of transcription, DNA-templated	14.9	0.02382	0
GO:0051252	regulation of RNA metabolic process	15.2	0.02533	0
GO:0080090	regulation of primary metabolic process	20.5	0.02596	0
GO:0060255	regulation of macromolecule metabolic process	20.1	0.02778	0
GO:0031328	positive regulation of cellular biosynthetic process	10.9	0.02807	0
GO:0009891	positive regulation of biosynthetic process	10.9	0.03318	0
GO:0010557	positive regulation of macromolecule biosynthetic process	10.6	0.03365	0
GO:0031323	regulation of cellular metabolic process	20.5	0.03685	0
GO:0000003	reproduction	12.9	0.03793	0
GO:0019222	regulation of metabolic process	20.8	0.04073	0
GO:0023051	regulation of signaling	4.6	0.0448	0
GO:0048523	negative regulation of cellular process	13.9	0.04573	0

[a] P values are corrected for multiple tests with Bonferroni correction.
[b] FDR, False Discovery Rate.

4.4.3 Genomic Scans of Divergence Between Natural and Domesticated Yeast Populations

In order to focus the attention in the most divergent regions, the levels of differentiation between populations across the genome were investigated using a 5 kb sliding window approach. For each window, it was calculated a relative and absolute measure of divergence, in the form of F_{ST} and π_B respectively, and the number of standard deviations each metric was away from its mean (Z-score). The mean value of F_{ST} between the two populations was $0.34 \pm 1.9 \times 10^{-3}$ (95% C.I., n=22486) and the overall levels of genetic differentiation and Z-scores are summarized in Figure 4.2.

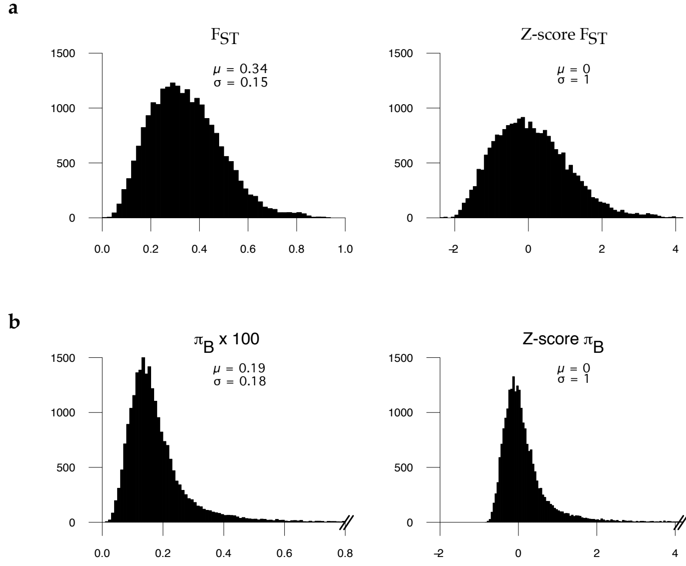


Figure 4.2 | Distributions of F_{ST} and π_B values and corresponding Z transformations for all 5 kb windows (n=22486). **(a)** F_{ST}. **(b)** π_B. Bins are represented in the x axis and the number of windows in the y axis. μ, mean; σ, standard deviation.

A Z-score of three or higher, corresponding to the extreme ends of the distribution, was used to define outlier windows, equivalent to an F_{ST} value of 0.768 and a π_B of 0.00681 (Figure 4.3a). Note that this F_{ST} value to define outlier regions is very close to the previous empirical threshold of 0.8 defined by ΔAF. In total, 34 regions spread across all chromosomes were identified for F_{ST} and also for π_B. A fraction (15%) of these regions, consisting of 9 outliers spanning ~78 kb (0.6% of the genome), was identified by both metrics, therefore standing as strong candidates for divergent selection between the two populations (Figure 4.3). These candidate regions were distributed over seven chromosomes, were well demarcated from the genomic background and 6 of them mapped close to subtelomeric regions.

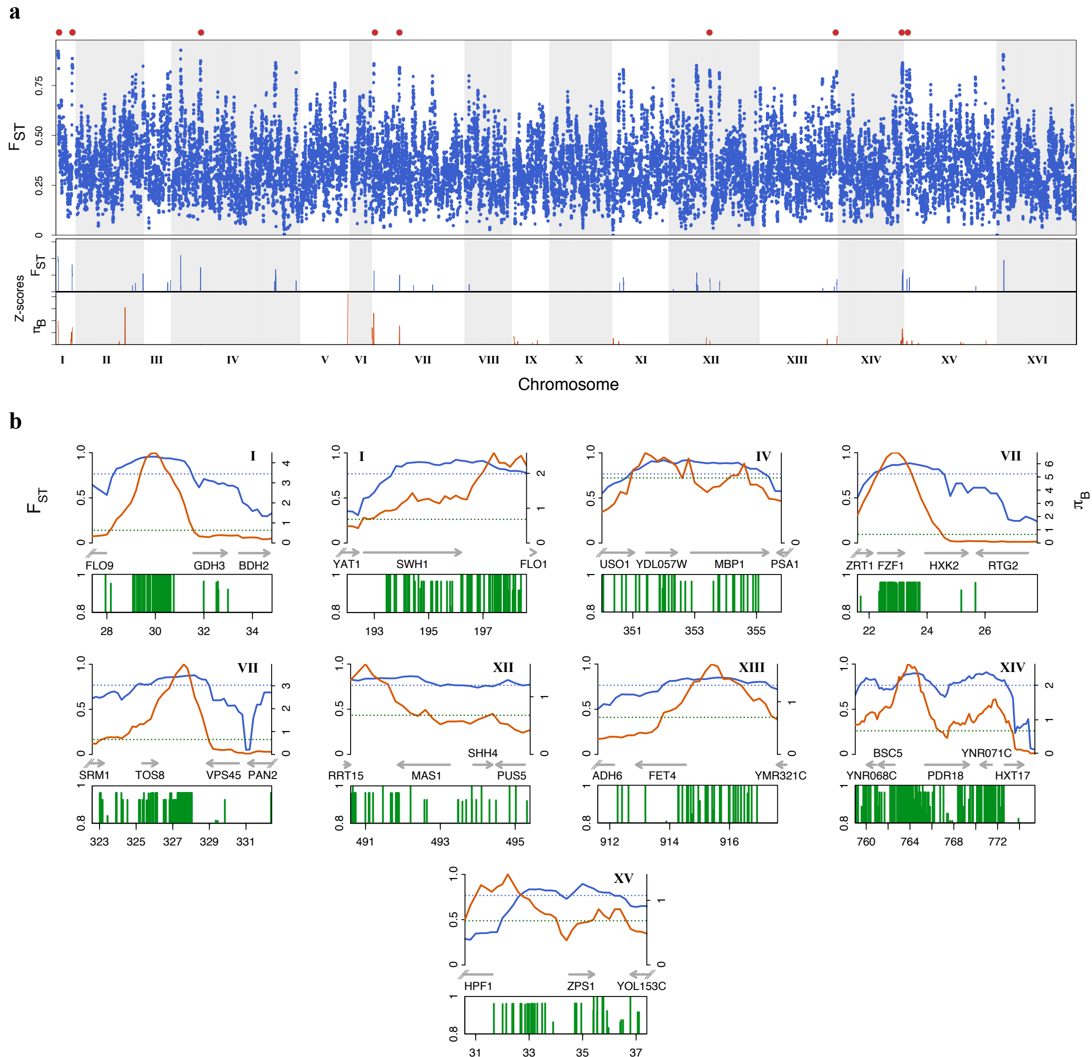


Figure 4.3 | Genomic landscape of divergence between Wine and Mediterranean oaks populations. **(a)** divergence was estimated in 5 kb sliding windows with 500 bp step. The upper panel shows the distribution of F_{ST}. The lower panel shows the Z-scores for F_{ST} (blue) and π_B (dark orange) above 3. Red dots indicate regions of overlapping high divergence in both F_{ST} and π_B. **(b)** divergence within candidate regions was estimated in 2 kb sliding windows with 200 bp step. The relative position and orientation of the genes are depicted below the plot. The upper panels show the distribution of F_{ST} (blue) and π_B (dark orange). Threshold values for a Z-score higher than 3 are denoted with dotted lines. The lower panels show the distribution of SNPs with $\Delta AF > 0.8$. Genome coordinates are in kb.

Many of the regions previously identified as outliers by F_{ST} and π_B contain multiple genes, making it challenging to identify specific targets of selection. Using a gene level resolution, by reducing the window size in the previously identified candidate regions to 2 kb, approximately the mean gene length in *S. cerevisiae*, it was possible to identify a total of 21 ORFs putatively under the effects of divergent selection (Figure 4.3b and Table 4.3).

Table 4.3 | List of ORFs within the candidate divergent regions between natural and domesticated populations of *Saccharomyces cerevisiae*.

Chr	ORF[a]	Type	Description
I	*FLO9*	Ver.	Lectin-like protein with similarity to Flo1p; thought to be expressed and involved in flocculation
I	*GDH3*	Ver.	NADP(+)-dependent glutamate dehydrogenase; synthesizes glutamate from ammonia and alpha-ketoglutarate
I	*SWH1*	Ver.	Protein similar to mammalian oxysterol-binding protein; regulated by sterol binding
I	FLO1	Ver.	Lectin-like protein involved in flocculation; cell wall protein that binds mannose chains on the surface of other cells; important for co-flocculation with other yeasts, mediating interaction with specific species
IV	YDL057W	Unc.	Putative protein of unknown function; YDL057W is not an essential gene
IV	*MBP1*	Ver.	Transcription factor; involved in regulation of cell cycle progression from G1 to S phase
VII	*FZF1*	Ver.	Transcription factor involved in sulfite metabolism; sole identified regulatory target is SSU1
VII	*TOS8*	Ver.	Homeodomain-containing protein and putative transcription factor; target of SBF transcription factor; induced during meiosis and under cell-damaging conditions
VII	*VPS45*	Ver.	Protein of the Sec1p/Munc-18 family; essential for vacuolar protein sorting; required for the function of Pep12p and the early endosome/late Golgi SNARE Tlg2p; essential for fusion of Golgi-derived vesicles with the prevacuolar compartment
XII	*RRT15*	Unc.	Putative protein of unknown function; identified by fungal homology comparisons and RT-PCR; identified in a screen for mutants with decreased levels of rDNA transcription
XII	*MAS1*	Ver.	Beta subunit of the mitochondrial processing protease (MPP); essential processing enzyme that cleaves the N-terminal targeting sequences from mitochondrially imported proteins
XII	*SHH4*	Ver.	Putative alternate subunit of succinate dehydrogenase (SDH); expression induced by nitrogen limitation in a GLN3, GAT1-dependent manner
XIII	*FET4*	Ver.	Low-affinity Fe(II) transporter of the plasma membrane
XIII	YMR321C	Unc.	Putative protein of unknown function; proposed to be a palmitoylated membrane protein; YMR321C has a paralog, SAM4, that arose from a single-locus duplication
XIV	YNR068C	Unc.	Putative protein of unknown function; exhibits homology to C-terminal end of Bul1p; expressed as a readthrough product of BSC5, the readthrough locus being termed BUL3
XIV	*BSC5*	Ver.	Protein of unknown function; shows homology with N-terminal end of Bul1p; ORF exhibits genomic organization compatible with a translational readthrough-dependent mode of expression. Readthrough expression includes YNR068C
XIV	*PDR18*	Ver.	Putative transporter of the ATP-binding cassette (ABC) family; role in plasma membrane sterol incorporation; implicated in pleiotropic drug resistance; provides resistance to ethanol stress and contributes to a decreased intracellular accumulation of ethanol
XIV	YNR071C	Unc.	Putative aldose 1-epimerase
XIV	*HXT17*	Ver.	Putative transmembrane polyol transporter; supports growth on and uptake of mannitol and sorbitol with moderate affinity; minor hexose transport activity; induced by raffinose and galactose at pH 7.7 versus pH 4.7, repressed by high levels of glucose
XV	*HPF1*	Ver.	Haze-protective mannoprotein; reduces the particle size of aggregated proteins in white wines
XV	*ZPS1*	Ver.	Putative GPI-anchored protein; transcription is induced under low-zinc conditions and at alkaline pH

[a] Dubious ORFs were excluded.

For some candidate regions, e.g. chrI *FLO9*- *GDH3*, chrXIV *BSC5-PDR18* and chrXV *HPF1-ZPS1*, the highest levels of population divergence were observed within the intergenic sequence upstream two bidirectional genes. Whereas transcription regulatory elements are likely to be involved in these cases, with this data it is not possible to distinguish whether the observed divergence influences the expression of one transcript but not the other, or if both genes are affected. Thus, the approach was to consider all bidirectional ORFs, aiming for an unbiased perspective of putative genes under selection. Analysis of gene ontology (GO) terms revealed a significant overrepresentation for categories related with flocculation (GO:0000128 "flocculation" and GO:0000501 "flocculation via cell wall protein-carbohydrate interaction"; $P < 0.05$), that were associated with the presence of *FLO1* and *FLO9* genes. These two genes encode cell wall lectin-like proteins involved in flocculation, therefore likely to influence yeast behaviour during the wine-making process, with *FLO1* being also important for co-flocculation with other yeasts and in response to heat and oxidative stress (Table 4.3 and Appendix III Table AIII.4). Despite the presence of 19 other ORFs in the gene set, it was not found any additional GO enrichment. Nevertheless, several ontology categories were represented at high frequency (Appendix III Table AIII.4 and Table AIII.5). These included a variety of cellular functions such as transmembrane transport of ions and carbohydrates (genes *FZF1*, *FET4*, *PDR18* and *HXT17*), hydrolase (*MAS1*, *PDR18* and *HPF1*), peptidase (*MAS1*) and oxidoreductase (*GDH3* and *SHH4*) activity, lipid and ion binding (*SWH1*) and DNA transcription (*FZF1*, *MBP1*), among others. Both *GDH3* and *SHH4* are related to the metabolism of nitrogen. The former is a NADP(+)-dependent glutamate dehydrogenase induced by ethanol and repressed by glucose, and the latter is an alternative subunit of succinate dehydrogenase (SDH), whose expression is induced by nitrogen limitation. Furthermore, *PDR18*, a multidrug resistance ABC transporter, was recently associated to an increased ethanol tolerance and ethanol production in alcoholic fermentations (Teixeira et al. 2012).

The signatures of divergent selection detected above could arise as a result of asymmetric selective pressures acting in either one of the lineages tested. In an attempt to distinguish between population-specific signals of selection, nucleotide diversity (θ_π) and Tajima's D statistics were calculated for each of the previously identified set of candidate regions in each population and compared those to the genomic background. Whereas nucleotide diversity was not significantly lower in outlier regions in comparison to the genomic background in either population ($P > 0.05$, Mann-Whitney U), Tajima's D was significantly lower for the outlier regions in the Wine population but not in the MO population ($P < 0.001$, Mann-Whitney U test) (Figure 4.4).

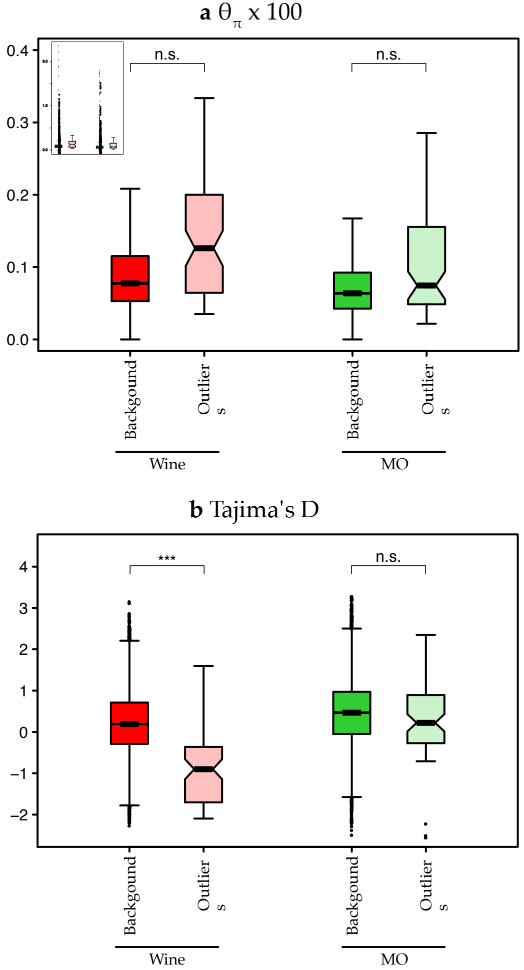


Figure 4.4 | Boxplots contrasting nucleotide diversity and Tajima's D statistic between divergence outliers and genomic background in wine and Mediterranean oaks (MO) populations. (a), nucleotide diversity ($\theta\pi$). For visualisation purposes, quantile outliers were not plotted, but are depicted in the inset. **(b)**, Tajima's D. Boxplots coloured in red represent the distribution of the estimates in the wine population and those coloured in green the distribution of the estimates in the MO population. For each population and each estimate, the distribution of divergence outliers and genomic background are always plotted side by side. n.s., non significant test statistic. ***, P value < 0.001 for one sided Mann-Whitney U test.

4.4.4 Divergence in *FZF1* is Caused by an Introgression in the Mediterranean Oaks Population

The divergent region encompassing *FZF1* was analysed in more detail because this gene has already been identified as a target of recent adaptive evolution in *S. cerevisiae* and its closest relative, *S. paradoxus* (Engle & Fay 2012). *FZF1* is involved in sulphite metabolism by regulating the transcription of *SSU1*, a sulphite efflux pump conferring resistance to sulphite (Avram et al. 1999; Park & Bakalinsky 2000). After a preliminary analysis it was readily apparent that a group of alleles very divergent to those of the *S. cerevisiae* reference genome were causing the high levels of divergence observed in our previous analyses (Figure 4.3). Therefore, *de-novo* assemblies were used to study *FZF1* in more detail. A phylogeny based on the coding sequence of *FZF1* showed that the large majority of alleles from the MO population are likely the result of a single introgression with *S. paradoxus* as donor (Figure 4.5a). In particular, these results identified the introgressed *FZF1* as more related to the European lineage of *S. paradoxus,* than to the North American or Far Eastern ones. Notably, the *S. paradoxus* native and the

introgressed versions in the MO population exhibited considerable nucleotide divergence, ~4.6% (Jukes-Cantor corrected, JC), corresponding to approximately half the divergence between two clearly differentiated lineages of *S. paradoxus*, the European and North American populations (~9.6%, JC). The latter divergence estimate for *FZF1*, is higher than that observed for other loci (Koufopanou et al. 2006), but is in line with recent evidence for an accelerated rate of evolution of *FZF1* (Aa et al. 2006; Engle & Fay 2012). A sliding window analysis of the region encompassing *FZF1* and the neighbouring upstream and downstream genes showed that this transfer is restricted to *FZF1* and does not include the flanking genes (Figure 4.3b, Figure 4.5b). Moreover, this introgression, although prevalent, is not fixed because the native *S. cerevisiae* alleles are still present in two MO strains: one isolated in Greece, and another one in Montenegro (Figure 4.6, Appendix III Table AIII.6).

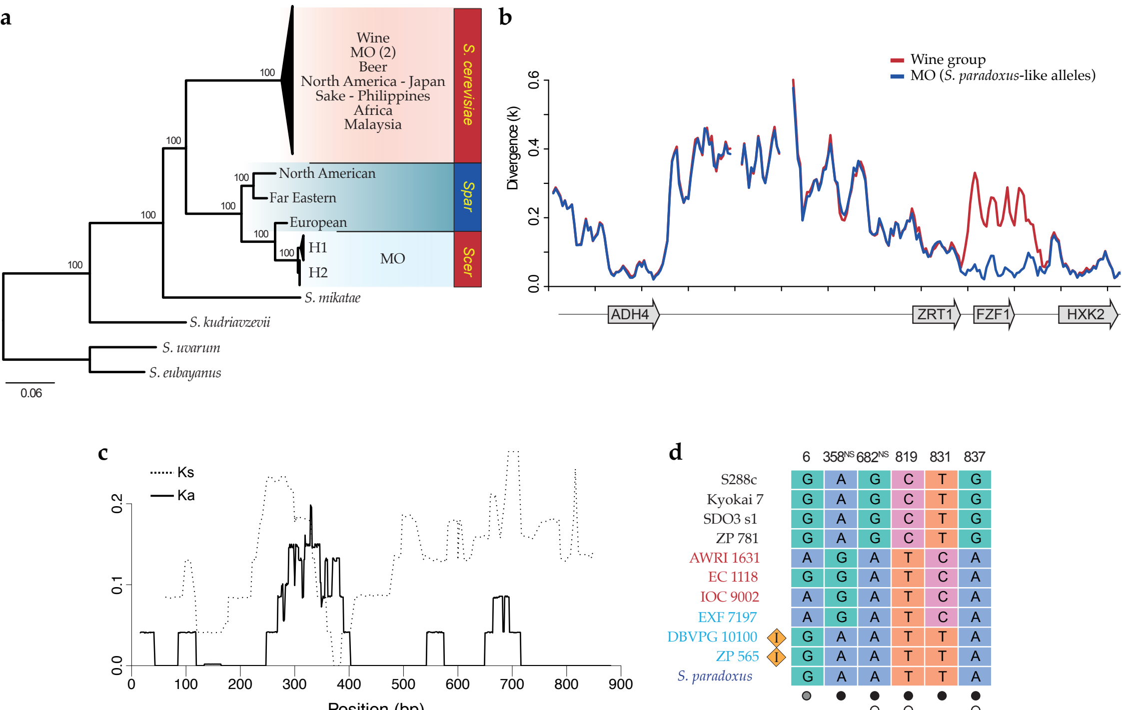


Figure 4.5 | Detection of the introgression of gene *FZF1* from *S. paradoxus* in the *S. cerevisiae* population associated with Mediterranean oaks. (a), maximum likelihood phylogenetic tree of the *FZF1* alignment inferred with the GTR + Γ model of sequence evolution as implemented in RAxML (906 positions in the final dataset). Bootstrap values (100 replicates) are shown next to the branches. Branch lengths correspond to the expected number of substitutions per site. Representative sequences from the three known populations of *S. paradoxus* were used to highlight the phylogenetic position of the Mediterranean oaks (MO) *S. cerevisiae* isolates. **(b)**, sliding window analysis showing divergence from *S. paradoxus* in the chromosomal region of *FZF1* including four neighboring genes. The relative position and orientation of the genes are depicted below the plot. Hash marks on x axis represent 1 kb of sequence in the alignment. Window size was set to 200 bp and step size to 50 bp. Divergence (Jukes-Cantor correction) is represented for *S. cerevisiae* Mediterranean strains showing introgression in *FZF1* (blue) and for the wine strains (red). **(c)**, nuceotide divergence in *FZF1* coding sequence of the introgressed alleles at synonymous (Ks, dotted line) and replacement (Ka, filled line) sites in the Mediterranean strains relative to *S. paradoxus*. Window size was set to 25 bp with a step size of 1 bp. **(d)**, sample of nucleotide differences found in six positions within the *FZF1* coding sequence. Only representative strains from each group are shown. Strain names are colored according to their phylogeny (red – Wine; blue – Mediterranean oaks; black – North America, Japan, Sake). Strains having introgressions from *S. paradoxus* are marked with an “I” in an orange diamond. Numbers represent the alignment position where each polymorphism was found. NS denotes a non-synonymous site. Black dots denote positions in which the wine allele differs from the reference (grey dot denotes a variable site) whereas open dots denote positions identical to *S. paradoxus*.

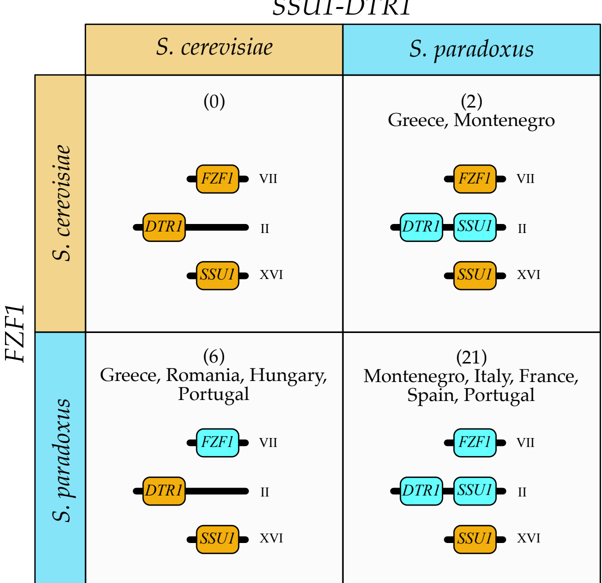


Figure 4.6 | Diagram illustrating the distribution of *FZF1* and *SSU1-DTR1* introgressions in the Mediterranean oaks (MO) population. Rows and columns represent the observed versions of *S. cerevisiae* (orange) and *S. paradoxus* (blue) for each introgression. Species names are abbreviated to a four letters code. Numbers in parenthesis indicate the total number of observed genotypes for the respective combination. All MO strains with the *SSU1* introgression of *S. paradoxus* also harboured the *S. cerevisiae* version of this locus. Note that the native *S. cerevisiae* version of both loci was never observed together in a single strain from this sampling.

Adaptive evolution acting on coding sequences predicts an excess of non-synonymous changes (Ka) relative to synonymous ones (Ks). Therefore, we analysed the changes in *S. paradoxus*-like alleles occurring in the MO lineage relative to the reference *S. paradoxus* allele. A small portion of the coding region between 320 bp and 400 bp exhibited Ka > Ks, however this could be caused to a large extent by a low synonymous divergence and not a high non-synonymous substitution rate (Figure 4.5c). Likewise, a similar pattern was also observed in the wine lineage (Figure 4.7). Furthermore, no significant differences were found in the ratio of mutations between the MO population and *S. paradoxus* as determined by a McDonald-Kreitman test (Table 4.4). Simulations under a neutral evolutionary model using the observed diversity within populations resulted in a significantly higher ratio of replacement to 4-fold polymorphisms in the Wine population relative to MO ($P < 10^{-15}$, two-tailed t-test). Remarkably, our *FZF1* data indicated the opposite trend, with a much higher ratio of replacement to 4-fold polymorphisms in the MO population than in the Wine population (0.43 and 0.17, respectively, Table 4.4), suggesting contrasting selection pressures in this gene between the two populations. These results were reinforced (0.48 and 0.12) after excluding rare substitutions (singletons), indicating that they cannot be simply explained by an excess of slightly deleterious non-synonymous mutations.

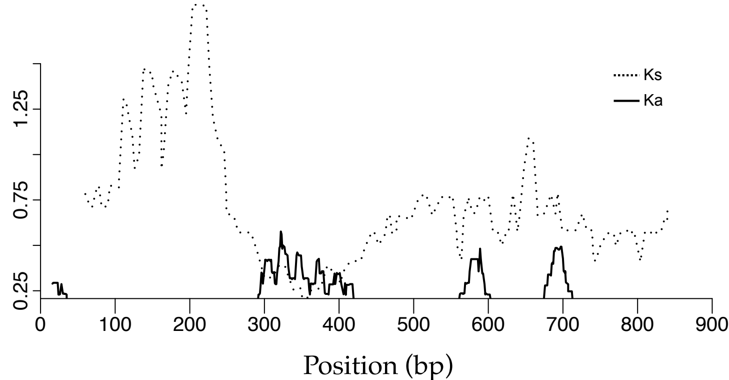


Figure 4.7 | Sliding window analysis of *FZF1* nuceotide divergence at synonymous (Ks, dotted line) and replacement (Ka, filled line) sites in the wine strains relative to *S. paradoxus*. The window size was set to 25 bp with a step of 1 bp.

Table 4.4 | Summary of DNA variation in the coding sequences of the three genes where an introgression of *S. paradoxus* (Sp) was detected in the Mediterranean oaks (MO) lineage. Diversity values (θ_π, the average pairwise nucleotide differences between strains, and θ_W, the Watterson estimator for the number of segregating sites) are per site estimates. Diversity was calculated at replacement (rep) and 4-fold sites.

***FZF1* McDonald-Kreitman test results for the MO lineage**	**Rep.**	**Syn.**	**P value[a]**	**DoS[b]**
Fixed substitutions in Sp-like lineage of MO	16	21	-	-
All polymorphisms	4	5	0.289	-0.012
Polymorphisms, excluding singletons	3	4	0.319	0.004

Nucleotide diversity within populations

	Alleles	**S**	**$\theta\pi$**	**Tajima's D**	**$\theta\pi$ (rep)**	**$\theta\pi$ (4-fold)**	**$\theta\pi$ (rep)/ $\theta\pi$ (4-fold)**
Genome-wide[c]	MO	-	0.0947	-0.6151	0.0565	0.1904	0.2969
	Wine	-	0.1042	-0.7511	0.0703	0.1863	0.3774
FZF1	MO-Sp	9	0.3461	0.8183	0.2037	0.4735	0.4301
	Wine	4	0.0929	-0.4885	0.0620	0.3628	0.1708
DTR1	MO-Sp	2	0.0273	-0.4402	0.0342	0.0000	-
	Wine	9	0.0657	-1.7248[d]	0.0681	0.0269	2.5312
SSU1	MO-Sc	8	0.2194	1.0271	0.0879	0.6694	0.1313
	MO-Sp	5	0.0864	-0.7288	0.0408	0.1497	0.2726
	Wine	11	0.2290	0.0158	0.2107	0.0366	5.7551

[a] All *P* values are from Fisher's exact tests
[b] Direction of selection (DoS) is calculated according to Stoletzki and Eyre-Walker 2011
[c] data from Almeida et al. 2015 (Chapter 3)
[d] significant at $P < 0.05$ level

Surprisingly, nucleotide diversity (θ_π) in the *FZF1* coding region within the MO population was more than three-fold higher than the gene-wide estimate across the genome, and Tajima's D statistic was positive, although not significant, indicating an excess of polymorphisms at intermediate

frequency (Table 4.4). Both estimates contrast with previous analyses using whole genomes (Almeida et al. 2015), and are likely explained by the existence of two main segregating haplotypes within the introgressed MO version of *FZF1* (Figure 4.5a), which lack a significant association with the geographic origin of the strains (P > 0.1, Fisher's exact test). Another interesting observation of polymorphism, this time among the *S. cerevisiae*-like *FZF1* alleles, is that the wine allele not only differs at some positions from the reference genome S288c, but is identical, in those sites, to the sequence of *S. paradoxus*, either due to homoplasy, recombination or positive selection (Figure 4.5d).

4.4.5 *SSU1*, a regulatory target for *FZF1*, is also introgressed in the MO population

The complete set of genomes investigated in this study was also surveyed for the presence of additional introgressions from other *Saccharomyces* species. The cases that could be identified unambiguously always implicated *S. paradoxus* as the donor (Digital Resources Dataset D3). Apart from a few introgressions specific to some wine or industrial strains (Digital Resources Dataset D3), two additional introgressions were found to be exclusive to the MO lineage. One introgression encompasses *DTR1* that encodes for a putative dityrosine transporter that is expressed during sporulation and functions in spore wall synthesis and the second is *SSU1*, that is implicated in the metabolism of sulphite and is regulated by *FZF1*. Like for *FZF1*, introgressions of *DTR1* and *SSU1* were not fixed in the MO population (Figure 4.6, Digital Resources Dataset D3). Twenty-three out of 29 (79%) MO strains have both *DTR1* and *SSU1* introgressions, comparing with a frequency of 93% for *FZF1*. Furthermore, nucleotide diversity within the MO population for each of the *S. paradoxus DTR1* and *SSU1* alleles was lower than the genome wide estimate for coding sequences (Almeida et al. 2015), and also lower than our estimate for *FZF1* (Table 4.4). Nucleotide divergence of the introgressed alleles to the homologous *S. paradoxus* sequence was 2.2% in the case of *DTR1* and 5.3% for *SSU1*.

4.4.6 European *S. paradoxus* and *S. cerevisiae* harbour two divergent copies of *SSU1*

While analysing the results above, two remarkable patterns were noted. First, foreign alleles of *DTR1* and *SSU1* in MO strains always occurred alongside with each other, i.e. whenever a *S. paradoxus* allele for *DTR1* (or *SSU1*) is found, the corresponding allele for *SSU1* (or *DTR1*) is also present. Second, all MO strains with a positive signal for the *S. paradoxus* sequence of *SSU1* also harboured its native (*S. cerevisiae*) allele, indicating the presence of both versions in these genomes (Digital Resources Dataset D3). This fact suggests that the transfer of a second *SSU1* allele might not have involved a homologous replacement.

Previous studies have reported a reciprocal chromosomal translocation involving the upstream region of *SSU1* in some wine yeasts (Pérez-Ortín et al. 2002). Therefore, in a first approach, this translocation was also experimentally screened in a subset of strains and, although it was found to be present in most of the wine strains tested, it was absent in all tested strains from the oak system (Table 4.5). Another circumstance that could explain both the presence of a native *S. cerevisiae SSU1* allele together with a *S. paradoxus* allele, and the association between the transfers of *DTR1* and *SSU1* in MO strains, would be if *SSU1* already had a second copy in the *S. paradoxus* genome with one of the versions in tandem arrangement with *DTR1*. Surprisingly, we found this configuration in European *S. paradoxus* but not in the North American nor in the Far Eastern populations, where only one copy of *SSU1* was

found (Figure 4.8). Whereas *S. cerevisiae* and *S. paradoxus SSU1* homologs are found on chromosome XVI, the extra copy in European *S. paradoxus* is located on chromosome II next to *DTR1* (hereafter designated as *SSU1-2* for brevity). *SSU1-2* is phylogenetic distinct from its European homolog and also from other close sequences, but the low bootstrap support (<70%) does not allow for a confident assertion of its placement (Figure 4.8a,b). Finally, *SSU1-2* was found to be located within a small region of synteny conservation of ~5.5 kb, with a variable but high degree of sequence identity between *S. paradoxus* ORFs *NOG1* (truncated), *SSU1*, *GLR1* and *RPS6B* on chromosome II with the paralogous ORFs on chromosome XVI (Figure 4.8c).

Table 4.5 | Screening for the translocation of *SSU1*-R in wine and wild yeasts. The presence of *SSU1*/*SSU1*-R alleles was investigated using PCR as in Pérez-Ortín *et al*. 2002.

		SSU1 alleles*	
Strain	**Group**	**SSU1**	**SSU1-R**
DBVPG 1106	Wine	NP	P
ZP 641	Wine	P	NP
VL 3	Wine	P	nd
AWRI 796	Wine	P	nd
ZP 1041	Wine	NP	P
AWRI 1631	Wine	NP	P
EC 1118	Wine	P	P
IOC 9002	Wine	NP	P
Lalvin QA23	Wine	P	P
EXF 7200	MO	P	NP
FGX d	MO	P	NP
ZP 1008	MO	P	NP
ZP 742	MO	P	NP
ZP 736	MO	P	NP
ZP 851	MO	P	NP
MB 7c	MO	P	NP
ZP 848	MO	P	NP
YPS 128	Japan /NA	P	NP

* P indicates the allele was detected by PCR screening, NP that it was not detected and nd means that its presence was not tested.

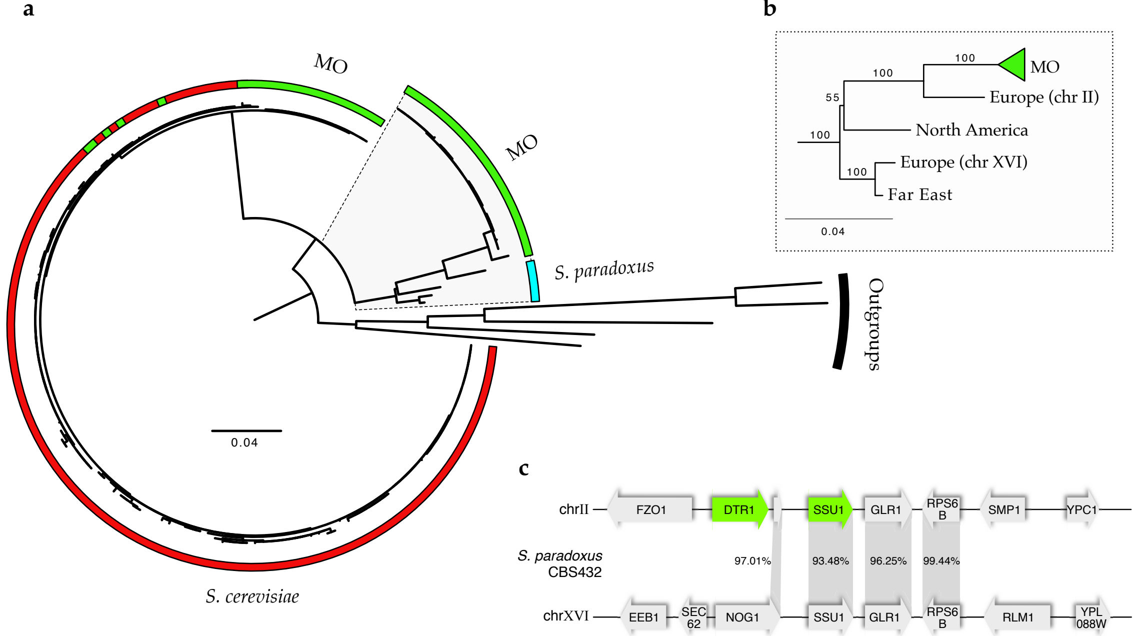


Figure 4.8 | Detection of the introgression of the region encompassing *DTR1-SSU1* from *S. paradoxus* in the *S. cerevisiae* population associated with Mediterranean oaks. (a) Maximum likelihood phylogenetic tree of the *SSU1* alignment inferred with the GTR + Γ model of sequence evolution as implemented in RAxML (1386 positions in the final dataset). Branch lengths correspond to the expected number of substitutions per site. Representative sequences from the three known populations of *S. paradoxus* were used to highlight the phylogenetic position of the Mediterranean *S. cerevisiae* isolates (MO), coloured in green in the side bar. **(b)** Close up view at the genetic relationships between populations of *S. paradoxus* populations and MO strains with the *SSU1* introgression. **(c)** Chromosomal regions of chromosome II and chromosome XVI in the reference strain of *S. paradoxus* showing a small region of synteny and homology around *SSU1*. BLAST percent identities are indicated between the paralogous ORFs. The detected introgressed ORFs in MO strains are coloured in green.

4.5 Discussion

4.5.1 Divergence and gene flow

In this chapter, available genomic data from the closest natural relatives of wine yeasts, the Mediterranean oaks (MO) population, and an improved data set of wine yeasts were used in order to identify genomic signatures separating wild and domesticated forms. Here, it is documented a substantial divergence between Wine and MO populations with a genome-wide F_{ST} considerably high (~0.34) given the likely recent population split. In Chapter 3, it was estimated relatively low levels of gene flow between the two populations, indicating that they could be partially isolated with only limited genetic admixture (Almeida et al. 2015). Barriers to gene flow are the main mechanism by which reproductive isolation can spread through the genome. Within *S. cerevisiae*, the already mentioned reciprocal chromosomal rearrangement in the *SSU1* locus, prevalent in wine strains but absent in all MO strains, has been shown to act as a significant driver of post-zygotic isolation (Hou et al. 2014; Clowers et al. 2015). Furthermore, wine and oak strains are phenotypically distinct showing strong ecological divergence (Clowers et al. 2015), hinting at ecological selection as a driver of divergence.

4.5.2 Patterns of Selection During Wine Yeast Domestication

The observation of increased divergence at non-coding sites is noteworthy since in eukaryotes a number of functional non-coding elements (e.g. promoters, enhancers, flanking sequences, introns and noncoding RNA) affect the expression of genes contributing to phenotypic diversity. Indeed, regulatory divergence can occur through the gain and loss of transcription factor binding sites (Doniger & Fay 2007) or changes in nucleosome position (Tsankov et al. 2010). Furthermore, many transcriptional differences specifically associated with polymorphic non-coding DNA sequence motifs have been identified between functionally diverse *S. cerevisiae* strains (Connelly et al. 2013). One potential source of bias in the data shown here could result from the recent population bottlenecks that both the Wine and MO populations are likely to have experienced (Chapter 3), which can increase the number of neutral alleles fixed by random drift. However, it would be expected that this bias affected equally all regions of the genome or that it would be more frequent in coding regions since these represent more than 70% of the *S. cerevisiae* genome. As this was not observed, population bottlenecks alone do not seem to justify the results obtained here.

The effect of increased divergence at non-coding sequences was most noticeable in the MO population but almost undetectable in the Wine population. Given that the number of fixed differences between Wine and MO populations is unchanged, these contrasting signals can only be attributable to differences in the ratio of synonymous to non-coding polymorphisms. One possible explanation would be that of stronger purifying selection at non-coding sites in the MO population, relative to Wine, in eliminating slightly deleterious polymorphisms. However, one would expect purifying selection to be more effective in the Wine population since it has a much larger effective population size (Chapter 3; Almeida *et al.* 2015). Alternatively, a more relaxed purifying selection in potentially constrained sites genome-wide could increase the number of polymorphisms in non-coding regions in the Wine population. Compatible with this scenario, the ratio of replacement to synonymous polymorphisms was higher in the Wine population relative to the wild lineage, indicating a relaxed efficiency of purifying selection in eliminating slightly deleterious alleles from the former population. In view of these different possibilities, the increased number of fixed substitutions between the Wine and MO populations in non-coding sites relative to synonymous sites (12% more substitutions, 95% CI [3%, 20%]), are likely the result of a combination of both adaptive and relaxed selection pressures driving the segregation of alleles in different environments. As generally predicted for domesticated species (Gepts 2004), it is likely that the domestication process of wine yeasts entailed a shift in selection pressures, in which severe genetic bottlenecks are usually associated with a relaxation of selective constraints.

Further supporting evidence for the role of adaptive selection in non-coding regions is given, indirectly, by an impressive enrichment of Gene Ontology (GO) functions related with transcription activity or binding to DNA regulatory regions, on which were found multiple highly differentiated replacement polymorphisms (ΔAF> 0.8). Therefore, these results seem to indicate a major effect of variations affecting transcriptional regulation between domesticated wine yeasts and their wild relatives. It can be hypothesised that divergence in non-coding sequence is concurrent with divergence in genes encoding DNA binding proteins, or vice-versa, thus, implying the turnover of nucleotide sequence in cis-regulatory regions along with changes in the coding sequence for trans-acting DNA binding proteins. Compatible with this hypothesis, binding sites and transcription factors have been

shown to be primary targets of diversification in mammals (Dermitzakis & Clark 2002). Furthermore, extensive divergence in binding sites has been proposed to be a major cause of rapid ecological specialization between *Saccharomyces* yeast species (Borneman et al. 2007; Doniger & Fay 2007). Interestingly, recent studies in domesticated animals have also found extensive adaptive divergence in non-coding DNA (Carneiro et al. 2014; Boitard et al. 2016).

In this chapter it is provided evidence for the accumulation of widespread small-scale genetic variation segregating wine and wild yeasts from the ancestral population. This variation was found to affect 601 ORFs (298 ORFs with SNPs in their 500 bp upstream region and 303 ORFs with replacement SNPs). Biological processes related with primary metabolism, life history traits, regulation of nitrogen compounds and response to stimulus were enriched with multiple replacement SNPs. Despite virtually nothing is still known about the resources that MO yeasts explore in their natural oak system many of these traits are potential candidates for adaptive evolution between wine and oak yeasts. For example, life history traits such as growth and reproduction, are key players during niche-driven adaptation and have been adopted differently in wild and industrial yeast populations (Spor et al. 2009). Altogether, these results suggest that wine yeast domestication ensued through multiple targets of selection, through the additive influence of multiple alleles across different genes, rather than a single metabolic or development pathway.

4.5.3 Genomic Divergence Between Wine and Wild Yeasts

Selective sweeps, i.e. the increase in frequency of a new positively selected allele that hitchhikes linked neutral variants, usually result in genomic islands of increased divergence, skewed frequency spectrum for rare variants and reduced diversity. The first two patterns were clearly identified in the 9 outlier regions putatively under strong divergent selection. However, nucleotide diversity for the Wine population within these candidate regions was not statistically lower than the genomic background. It is possible that the variety of wines produced in different regions together with variation in fermentation conditions between regions and wine producers have introduced subsequent layers of selection during strain improvement, thus eroding the signal of ancestral sweeps. Moreover, genetic admixture with local strains could also have the same effect, as already documented in maize (Hufford et al. 2012). Since the sampling used here included wine strains from different geographic origins and sources it is also possible that the identified loci under putative selection originated from signatures of either incomplete or soft sweeps, in which case the selected allele may not be fixed in the entire population and the reduction in nucleotide diversity can be weak (Innan & Kim 2004). The observed increased diversity in the candidate regions could also result from elevated mutation rates. In *S. cerevisiae*, subtelomeric sequences are usually more susceptible to higher mutation rate and therefore are more variable than other regions of the genome (Winzeler et al. 2003; Brown et al. 2010), providing an ideal substrate for the rapid evolution of genes and adaptation in response to selection. Interestingly, 6 out of 9 divergence outliers identified in this study localised in subtelomeric regions.

Among the genes present in the regions under putative directional selection it was found a glutamate dehydrogenase gene (*GDH3*) and a succinate dehydrogenase gene (*SHH4*), two genes involved in nitrogen metabolism. Since nitrogen availability during wine fermentations is usually low it is likely that directional selection has played an important role in adaptation to limiting nitrogen

sources. ORFs related with carbohydrate metabolism (*YNR071C*, *HXT17*) and with increase ethanol tolerance (*PDR18*) were also found within these regions, suggesting an important role of adaptation to alcoholic fermentations. Furthermore, the presence of *FLO1* and *FLO9* genes resulted in a significant overrepresentation of two flocculation-related GO categories. Interestingly, non-flocculent *S. cerevisiae* wine yeasts can interact with non-*Saccharomyces* yeasts, resulting in a "co-flocculation" phenotype that is *FLO*-specific, suggesting that *FLO* genes mediate ecological interactions involving *Saccharomyces* wine yeasts and other wine-associated microbes (Rossouw et al. 2015). It is therefore reasonable to speculate that divergence in the upstream regions of *FLO1* and *FLO9* between Wine and MO yeasts is the result of adaptation to the distinct biological interactions occurring in their respective environments.

The above results showing divergent regions between the two populations can be seen as conservative, because only regions identified as outliers for both F_{ST} and π_B were considered as the best candidates for directional selection. Many of the outlier regions identified by a single measure could result from either the segregation of ancestral polymorphisms (high π_B) or reduced nucleotide diversity in both populations (high F_{ST}) (Noor & Bennett 2009; Cruickshank & Hahn 2014). Nonetheless, it is worth noting that many windows showed high F_{ST} (e.g. > 0.6) and although these regions were not considered in the current approach it is likely that many deserve further research.

4.5.4 New roles for old genes? The case of *FZF1* and *SSU1*

The observation of apparently widespread introgressions of *FZF1* and *DTR1*-*SSU1* from European-like *S. paradoxus* in the MO population contrasts with their complete absence in the closely related Wine clade. *S. cerevisiae* and *S. paradoxus* can be found in sympatry (Sniegowski et al. 2002; Sampaio & Gonçalves 2008), which provides the opportunity for hybridization between the two species. Two distinct explanations can be put forward for the distribution of *FZF1* and *DTR1*-*SSU1* alleles. The first assumes that the introgressions are not recent but did not reach fixation and therefore native and introgressed alleles may have coexisted in the most recent common ancestor population of the Wine and MO lineages. Domestication may have subsequently selected against the *S. paradoxus*-like alleles in the winemaking environment while in the MO population they were probably advantageous since they are present in the large majority of the strains. An alternative explanation would be that the introgressions are recent and occurred only in the MO lineage, after divergence of the Wine lineage, in which case it would be necessary to assume that it originated in a so far unidentified, highly divergent *S. paradoxus* lineage. For *FZF1*, other results presented here apparently support the former hypothesis given the surprisingly high nucleotide diversity of the introgressed allele, its wide distribution over the geographic range of *S. cerevisiae* in the Mediterranean region and the stronger evolutionary constraint of *FZF1* in the wine population.

DTR1 and *SSU1* are likely to have been transferred as a single, contiguous genomic tract since both genes sit next to each other on chromosome 2 in *S. paradoxus*. However, the results are less clear to whether these foreign genes result from a single introgression involving also *FZF1*. Although the nucleotide diversity of introgressed alleles in *S. cerevisiae* is much higher for *FZF1* than for either *DTR1* or *SSU1*, raising the possibility of a more recent origin than that of *FZF1*, the nucleotide divergence of *SSU1* to its homologous *S. paradoxus* sequence is similar to the divergence observed for *FZF1*. The role

of *DTR1* in spore maturation could explain its higher degree of sequence conservation while, in contrast, *FZF1* and *SSU1* are genes already known for their unusually high evolution rates (Aa et al. 2006; Engle & Fay 2012). The most parsimonious explanation corresponds to a single introgression. This scenario is supported by the close association of the two genes, being *FZF1* the sole known transcription regulator of *SSU1* (Avram et al. 1999; Park & Bakalinsky 2000). However, the alternative hypothesis of a *DTR1*- *SSU1* introgression independently from that of *FZF1* cannot be discarded because hybridizations and introgressions are commonly observed in *S. cerevisiae* (Dunn et al. 2012; Barbosa et al. 2016).

The frequency of the *S. paradoxus*-like *SSU1* allele in the MO population is considerable (79%). Such high frequency would be expected if the introgressed variant confers some selective advantage in natural populations. However, a population bottleneck in the ancestral MO population could have also provided the opportunity for this frequency to increase by drift. Surprisingly, the genome of European *S. paradoxus* carries two copies of *SSU1*, the *S. cerevisiae* ortholog version on chromosome XVI and a second divergent copy on chromosome II. Interestingly, it was the latter that had introgressed in MO strains of *S. cerevisiae*. Altogether, these observations provide a strong case for selection in maintaining two (divergent) copies of *SSU1* in European wild yeast populations, hinting for some functional relevance of the introgressed allele, likely in combination with *FZF1*. Although the origin of the second *SSU1* copy (*SSU1-2*) on chromosome II of *S. paradoxus* is uncertain, it is likely to have originated from a small translocation between chromosomes II and XVI. *SSU1* and *FZF1* are well known to be involved in sulphite resistance (Avram et al. 1999; Park & Bakalinsky 2000; Engle & Fay 2012), an important trait selected in wine yeasts. However, the complex dynamics of these genes in oak-associated strains suggest that they have alternative and perhaps ancestral roles in natural habitats where sulphite is absent. Interestingly, both *SSU1* and *FZF1* are also involved in the response to nitrosative stress (Sarver & DeRisi 2005; Cabrera et al. 2014), but the impact of this system is unknown in wild yeast populations.

4.5.5 Adaptation in wild and domesticated populations

In line with the hypothesis that the reservoir of wild *S. cerevisiae* strains originated in Asia, introgressions with local *S. paradoxus* genomes during the geographic diversification of wild *S. cerevisiae* lineages could foster the adaptation of migrants to the new environments (Barbosa et al. 2016). Compatible with this hypothesis, the low diversity found within the MO population indicates a relatively recent origin of wild *S. cerevisiae* in Europe (Chapter 3; Almeida et al. 2015). The introgressions of *SSU1-DTR1* and *FZF1* from *S. paradoxus* that are observed only in the MO *S. cerevisiae* population also suggest the acquisition of locally adapted alleles possibly during the early stages of adaptation to the Mediterranean niche. Furthermore, these introgressions were not transmitted to the wine yeasts which could be attributable to either low levels of gene flow between populations, increase selection against these variants in either environment or a combination of both factors. Given the regulatory interaction between *SSU1* and *FZF1* and the different niches occupied by domesticated and wild yeasts, it is likely that divergent ecological selection played an important role in segregating the native and introgressed forms between populations.

The analyses in this chapter have shown that wine yeast domestication involved extensive nucleotide divergence in cis-regulatory regions along with changes in the coding sequence for trans-

acting DNA binding proteins. Evolution across all kingdoms of life can be largely driven by changes in the genetic regulatory architecture (Carroll 2005) and in fungi such regulatory modifications rewired, for example, catabolic processes (Martchenko et al. 2007) and ribosomal transcriptional modules (Ihmels et al. 2005). In *S. cerevisiae*, despite considerable evidence for intraspecific variability in transcriptional regulation in modulating phenotypic differences between strains (Brion et al. 2013; Treu et al. 2014) and powering adaptation to environmental stress in experimental populations (Voordeckers et al. 2015), it is unclear how and at which pace such variation occur in natural populations. The results presented here are likely to help addressing these questions by demonstrating that the divergence between closely related populations appears to be determined by a complex accumulation of cis- and trans-regulatory highly divergent polymorphisms. It is hypothesised that this divergence established rather rapidly, in view of the recent population split at the onset of domestication.

CHAPTER 5

Concluding Remarks and Future Perspectives

In ancient times human civilisations started to use microorganisms unwittingly to carry out fermentations of a variety of foods and beverages worldwide. Continued artificial selection of the most suited microbes promoted domestication through adaptation and specialisation to the artificial environments created by man. However, in spite of microbe domestication being as old as plant and animal domestication, the study of the mechanisms and consequences of artificial selection of microorganisms is still in its early stages. Indeed, a detailed understanding of the origins of microbe domestication, identifying the source wild populations and its impact on genome architecture and function, lags far behind that of crops and livestock domestication. Providing answers to these questions will not only improve our knowledge about the microbial transformations of foods and beverages, but will have important applications that can vary from the exploration of the extant biological diversity to the rational optimisation of already established industrial variants and to the development of new products.

In the light of this, the work described herein aimed at understanding better two yeast species of the genus *Saccharomyces* with remarkable value for alcoholic fermentations, *S. cerevisiae* and the cryotolerant yeast *S. uvarum*. Considering that for the detailed identification of domestication events and their mechanisms it is essential to have a good knowledge of the wild populations from which they have evolved, considerable attention was given to the study of wild populations that are likely candidates of being the closest relatives of domesticates. Indeed, up until the present study, both species have only been scarcely isolated from natural habitats not related with human activities. Using population genomic approaches, it was here investigated (i) the geographic distribution and population structure of new wild isolates and (ii) their relationship with the domesticated lineages. The findings presented in the course of this study provide new insights on the role of wild lineages in the likely multiple domestication events that lead to the emergence of modern industrial strains.

5.1 Characterisation of the global diversity of wild lineages of *S. cerevisiae* and *S. uvarum*

The yeasts of the genus *Saccharomyces* are viewed as model organisms in many areas of biology with emphasis on molecular genetics and cellular biology and more recently on comparative genomics (Scannell et al. 2011; Hittinger 2013). Paradoxically, basic aspects of their natural history are not known and their ecological niches and biogeography have remained elusive. In the present study, the genome-wide characterisation of geographically diverse natural populations of *S. cerevisiae* and *S. uvarum* provided the first representation of their genetic diversity found in the wild at a global scale. The results presented in Chapter 2 and Chapter 3 suggest that geographic populations in *S. cerevisiae* and *S. uvarum* can persist in natural habitats while maintaining independent genetic histories.

In *S. uvarum*, phylogenetic and population structure analyses readily identified three distinct lineages: one containing all Holarctic strains and also some from South America, a second one including only strains from South America and a more divergent cluster containing Australasian isolates (Chapter 2). The Holarctic clade, despite being composed by strains isolated both in wild and anthropic environments, has the lowest genetic diversity, indicating a recent colonization of the Northern Hemisphere possibly originating in South America. The finding of a considerable divergent population in Australasia, with levels of reproductive isolation comparable to those found in the most divergent populations of *S. paradoxus* (Liti et al. 2006) and also for the divergent lineages of *S. cerevisiae* found in

China (Wang et al. 2012), suggests a possible ongoing process of allopatric speciation. Interestingly, the two Patagonian populations are less divergent and showed evidence of secondary contact and genetic admixture, which could have happened after a period of isolation possibly related to past glaciations events. These observations highlight the potential of *S. uvarum* for the genetic dissection of reproductive barriers in natural *Saccharomyces* populations, which in this genus are thought to be largely driven by sequence divergence (Hittinger 2013).

In spite of the status of *S. cerevisiae* as a prime model organism, its natural ecology and population history have remained largely unknown. In Chapter 3, using an extended collection of natural isolates from different geographical regions in the Northern Hemisphere, it was possible to identify new oak associated lineages in the Mediterranean, North America and Japan. In view of these novel findings, it can be proposed that the population history of *S. cerevisiae* is shaped by both ecology and geography. Interestingly, wild isolates found in North America and Japan are polyphyletic and have mixed ancestries. It is hypothesised that North American wild isolates of *S. cerevisiae* originated from the more diverse and possibly ancestral Asian lineages, although more detailed comparisons between wild lineages will be necessary to address further this possibility. These findings are relevant as they indicate that, in contrast to the common view up until now, admixture in *S. cerevisiae* is not only driven by the human traffic of strains but can also have natural causes. In addition, they strengthen the claim that *S. cerevisiae*, as whole, is not a domesticated species, and that natural populations can thrive in environments not related to human activities. They also open the opportunity for comprehensive population and comparative genomic studies of natural populations in *S. cerevisiae* and for their comparison with *S. paradoxus*, which is considered by many a truly wild species not associated with human environments. In *S. paradoxus* there is little evidence of gene flow between continents and it will be interesting to understand the underlying demographic, genetic and life history mechanisms that shaped differently the population structure of wild lineages in the two species.

Considering the Mediterranean oaks (MO) population, its distribution appears to be confined to Southern Europe, which is in contrast to the wider geographical range of *S. paradoxus* in this continent. Furthermore, the genetic diversity observed in the MO population was similar to the diversity in the domesticated wine group and surprisingly lower than for the other wild lineages identified in this study. It is possible that the MO lineage arrived only recently into Europe and that competition with the more prevalent sympatric *S. paradoxus* population, environmental constraints posed by the new conditions, or a combination of both factors, limited its expansion to more northern latitudes. A recent study showed a marked association between the frequency of isolation of *S. paradoxus* and *S. cerevisiae* with summer temperatures (Robinson et al. 2016). As already suggested by the authors, this could be further tested with additional ecological surveys in both species at the upper limits of their predicted optimum summer temperature (Robinson et al. 2016). Although not directly demonstrating a causal effect of temperature variation, that and other studies (Sweeney et al. 2004; Gonçalves et al. 2011; Leducq et al. 2014) seem to suggest that genetic determinants for temperature preferences could constraint the geographic distribution of *Saccharomyces* species. At present, divergence of glycolytic genes are the best candidates to explain the distinct growth temperature preferences of *Saccharomyces* species (Gonçalves et al. 2011), and it would be interesting to map clines of molecular and phenotypic variations of these alleles in both species. In addition, it was shown in Chapter 4 that introgressed copies of genes *FZF1* and *SSU1* from *S. paradoxus* are likely to be maintained by selection in the *S. cerevisiae*

MO population but eliminated from the wine population. Although these genes are best known by their involvement in sulphite resistance, their complex dynamics in oak-associated strains suggest that they might have alternative and possibly ancestral roles in natural habitats where sulphite is absent. Follow-up experimental data, for example by swapping alleles between populations, could help to elucidate the phenotypic outcome of these introgressions in wild and winemaking environments.

The wide phylogenetic scope captured in this work, also allowed new insights into the possible evolution of the genus. The hypothesis put forward in Chapter 2 suggests that the last common ancestor of the *Saccharomyces* lineage could have arisen in the Southern Hemisphere, possibly in association with the *Nothofagus* system. According to this model, the ancestral *Saccharomyces* lineage then expanded north through Southeast Asia which is in line with the currently known geographic range of the more basal *S. uvarum*/*S. eubayanus* lineage and with the apparent endemism of other *Saccharomyces* species in Asia. Nevertheless, much of what is known about the natural ecology of *Saccharomyces* is only relatively recent (Naumov et al. 2000a; Sampaio & Gonçalves 2008; Hittinger et al. 2010; Libkind et al. 2011; Wang et al. 2012; Hyma & Fay 2013; Peris et al. 2014; Sylvester et al. 2015) and it will be necessary to explore further its global distribution, natural habitats and diversity to test these ideas. This is becoming ever more feasible with the advancement of Next-Generation sequencing technologies (Levy & Myers 2016). For example, the development of new powerful metagenomic methods (Logares et al. 2014) are likely to offer researchers interested in this subject an indispensable molecular tool for future ecological studies, bypassing or complementing the more traditional cultivation techniques.

5.2 Multiple Domestication trajectories in *Saccharomyces*

It is known that wine yeasts (*S. cerevisiae*) sampled around the world are genetically similar and clearly distinct from wild populations (Fay & Benavides 2005; Legras et al. 2007). However, studies performed up until now have been critically affected by a strong sampling bias towards domesticated strains with a clear underrepresentation of wild lineages. In the current study, the integration of geographically diverse samples of natural isolates with industrial variants provided the first glance at the natural history of wine yeast domestication. Wild lineages closely related to domesticated stocks were found in both *S. cerevisiae* and *S. uvarum*.

The results presented in Chapter 3 are compatible with a MO lineage representing the wild genetic stock of wine yeasts in *S. cerevisiae*. The newly uncovered MO population proved to be closely related to the Wine group in phylogenetic and population structure analyses and divergence time estimates also agree with historical evidence for winemaking. Interestingly, in spite of a recent separation, population divergence is already substantial, highlighting the need to investigate the potential role of barriers to gene flow and ecological selection between wine and wild populations in *S. cerevisiae* (Chapter 4). Tests of different demographic models indicated that wine yeast domestication must have ensued a genetic bottleneck at the initial stages of domestication and then this population might have expanded through human-related activities, as also suggested by the wide geographical distribution of wine yeasts (Chapter 3). The demographic history of domesticated wine yeasts and their wild ancestors might, however, be more complex than that depicted here and additional studies particularly focused on the exploration of population dynamics during wine yeast domestication are needed.

Whereas in *S. cerevisiae* the domesticated wine strains form a single lineage distinct from their putative wild ancestors of Mediterranean oaks, the wine and cider strains and their wild Holarctic relatives in *S. uvarum* could not be resolved at the phylogenetic and population structure levels (Chapter 2). At least two hypotheses, not mutually exclusive, can be put forward to accommodate the apparently more recent domestication in *S. uvarum* in comparison to that in *S. cerevisiae*. First, colonisation of the Northern Hemisphere by *S. uvarum* is recent, as suggested by the lower genetic diversity measured in the whole Holarctic population, and domestication could have been concurrent with the movement of *S. uvarum* into Europe (Chapter 2). In addition, while *S. cerevisiae* typically dominates most wine fermentations, *S. uvarum* is mostly prevalent in wine fermentations carried out at lower temperatures where usually replaces *S. cerevisiae*. Therefore, it is also likely that fermentations at relatively high temperatures are closer to the ancient practices and that low temperature fermentations developed more recently, which in turn suggests that domestication in *S. uvarum* could also reflect this technological distinction.

Holarctic strains of *S. uvarum* from fermentative environments could be distinguished from the wild strains by the presence of multiple introgressions from *S. eubayanus* (Chapter 2). The *S. eubayanus* lineage that contributed to these introgressions is not yet known and could have originated either from *S. pastorianus*, an entirely domesticated species that is a hybrid of *S. eubayanus* and *S. cerevisiae*, or from a yet-unidentified population. It was suggested recently that *S. pastorianus* might have an East Asian origin, because the *S. eubayanus* genome moiety in this hybrid is more closely related to *S. eubayanus* lineages found in Tibet than to those found in South America (Bing et al. 2014), although more recent evidence challenges this view (Baker et al. 2015). In order to elucidate the origin and apparent pervasiveness of *S. eubayanus* genomes in domesticated *Saccharomyces* strains, it will be indispensable to perform detailed genomic comparisons of East Asian genotypes with the introgressions detected in this study. Similarly, the detection of at least one of three HGT regions with important functions in wine fermentation in wine strains of *S. cerevisiae* is interpreted as a genomic signature of wine yeast domestication in this species (Chapter 3), while also present in some *S. uvarum* domesticated strains. Together, the results presented in this thesis further highlight the evolutionary potential of hybridization and HGT in the evolution of the domesticated phenotype in multiple *Saccharomyces* lineages.

In addition, nucleotide divergence between wine and wild populations of *S. cerevisiae* was surprisingly higher in noncoding regions and in genes related with DNA binding and transcription activities (Chapter 4). Although these results provide only indirect evidence, they support extensive divergence in the transcriptional regulatory architecture between domesticated wine yeasts and their wild relatives. It is also reasonable to consider that the multiple introgressions from *S. eubayanus* in wine and cider *S. uvarum* strains disturbed the original, wild type transcription network (Chapter 2). Therefore, future research addressing specifically expression divergence between the two populations should be carried out to test the effect of these variations in transcription activity and their potential phenotypic role during adaptation to the wine must environment.

Lastly, one key question that, if elucidated, could contribute to the understanding of yeast domestication is why *S. paradoxus*, the closest relative of *S. cerevisiae*, was never domesticated. *S. paradoxus* has a much wider biogeographic distribution and in sugar rich substrates like grape juice does not seem to have fitness differences in co-cultures with *S. cerevisiae* (Williams et al. 2015). Thus, it

would be interesting to investigate the causes of the non-domestication of *S. paradoxus* using detailed evolutionary analyses, including population genomics.

In summary, the multiple trajectories observed in the domestication of *Saccharomyces* lineages do not seem to have parallel in other domesticated microbe. Identified variants with potential success in human artificial environments include domesticated strains within a lineage (*S. uvarum*; Chapter 2), domesticated lineages within a species (*S. cerevisiae*; Chapter 3; Chapter 4) and domesticated hybrid species (e.g. *S. pastorianus;* Baker et al. 2015). Moreover, the identification for the first time in this study of new wild lineages closely related to the domesticated stocks place the whole *Saccharomyces* genus as a model system in the forefront of microbe domestication research and facilitate further studies aiming at the detailed understanding of the evolution of particular domesticated traits.

REFERENCES

Aa E, Townsend JP, Adams RI, Nielsen KM, Taylor JW (2006). Population Structure and Gene Evolution in *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 6(5):702-715.

Almeida P, Gonçalves C, Teixeira S, Libkind D, Bontrager M, Masneuf-Pomarède I, Albertin W, Durrens P, Sherman DJ, Marullo P, Hittinger CT, Gonçalves P, Sampaio JP (2014). A Gondwanan Imprint on Global Diversity and Domestication of Wine and Cider Yeast *Saccharomyces uvarum*. *Nat Commun*, 5:4044.

Almeida P, Barbosa R, Zalar P, Imanishi Y, Shimizu K, Turchetti B, Legras JL, Serra M, Dequin S, Couloux A, Guy J, Bensasson D, Gonçalves P, Sampaio JP (2015). A Population Genomics Insight Into the Mediterranean Origins of Wine Yeast Domestication. *Mol Ecol*, 24(21):5412-5427.

Angebault C, Djossou F, Abélanet S, Permal E, Ben Soltana M, Diancourt L, Bouchier C, Woerther PL, Catzeflis F, Andremont A, d'Enfert C, Bougnoux ME (2013). *Candida albicans* is Not Always the Preferential Yeast Colonizing Humans: A Study in Wayampi Amerindians. *J Infect Dis*, 208(10):1705-1716.

Antunovics Z, Irinyi L, Sipiczki M (2005). Combined Application of Methods to Taxonomic Identification of *Saccharomyces* Strains in Fermenting Botrytized Grape Must. *J Appl Microbiol*, 98(4):971-979.

Avram D, Leid M, Bakalinsky AT (1999). Fzf1p of *Saccharomyces cerevisiae* is a Positive Regulator of Ssu1 Transcription and Its First Zinc Finger Region is Required for DNA Binding. *Yeast*, 15(6):473-480.

Baker E, Wang B, Bellora N, Peris D, Hulfachor AB, Koshalek JA, Adams M, Libkind D, Hittinger CT (2015). The Genome Sequence of *Saccharomyces eubayanus* and the Domestication of Lager-Brewing Yeasts. *Mol Biol Evol*, 32(11):2818-2831.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012). Spades: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*, 19(5):455-477.

Bar-Yosef O (1998). The Natufian Culture in the Levant. *Evolutionary Anthropology*, 6(5):167-168.

Barbosa R, Almeida P, Safar SV, Santos RO, Morais PB, Nielly-Thibault L, Leducq JB, Landry CR, Gonçalves P, Rosa CA, Sampaio JP (2016). Evidence of Natural Hybridization in Brazilian Wild Lineages of *Saccharomyces cerevisiae*. *Genome Biol Evol*, 8(2):317-329.

Barnard H, Dooley AN, Areshian G, Gasparyan B, Faull KF (2011). Chemical Evidence for Wine Production Around 4000 Bce in the Late Chalcolithic Near Eastern Highlands. *Journal of Archaeological Science*, 38(5):977-984.

Barnett JA (2000). A History of Research on Yeasts 2: Louis Pasteur and His Contemporaries, 1850-1880. *Yeast*, 16:755-771.

Bauer FF, Pretorius IS (2000). Yeast Stress Response and Fermentation Efficiency: How to Survive the Making of Wine-a Review. *South African Journal for Enology and Viticulture*, 21:27-51.

Belloch C, Orlic S, Barrio E, Querol A (2008). Fermentative Stress Adaptation of Hybrids Within the *Saccharomyces* sensu stricto Complex. *Int J Food Microbiol*, 122(1-2):188-195.

Bergström A, Simpson JT, Salinas F, Barré B, Parts L, Zia A, Nguyen Ba AN, Moses AM, Louis EJ, Mustonen V, Warringer J, Durbin R, Liti G (2014). A High-Definition View of Functional Genetic Variation From Natural Yeast Genomes. *Mol Biol Evol*, 31(4):872-888.

Bevan MW, Uauy C (2013). Genomics Reveals New Landscapes for Crop Improvement. *Genome Biol*, 14(6):206.

Bing J, Han PJ, Liu WQ, Wang QM, Bai FY (2014). Evidence for a Far East Asian Origin of Lager Beer Yeast. *Curr Biol*, 24(10):R380-1.

Blein-Nicolas M, Albertin W, Valot B, Marullo P, Sicard D, Giraud C, Huet S, Bourgais A, Dillmann C, de Vienne D, Zivy M (2013). Yeast Proteome Variations Reveal Different Adaptive Responses to Grape Must Fermentation. *Mol Biol Evol*, 30(6):1368-1383.

Boitard S, Boussaha M, Capitan A, Rocha D, Servin B (2016). Uncovering Adaptation From Sequence Data: Lessons From Genome Resequencing of Four Cattle Breeds. *Genetics*, 203(1):433-450.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007). Divergence of Transcription Factor Binding Sites Across Related Yeast Species. *Science*, 317(5839):815-819.

Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, Chambers PJ (2011). Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of *Saccharomyces cerevisiae*. *PLoS Genet*, 7(2):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 (Bethesda)*,

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004). Go::termfinder--open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated With a List of Genes. *Bioinformatics*, 20(18):3710-3715.

Boynton PJ, Greig D (2014). The Ecology and Evolution of Non-Domesticated *Saccharomyces* Species. *Yeast*, 31(12):449-462.

Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, Dewey C, Holmes I, Pachter L (2009). Fast Statistical Alignment. *PLoS Comput Biol*, 5:e1000392.

Brion C, Ambroset C, Sanchez I, Legras JL, Blondin B (2013). Differential Adaptation to Multi-Stressed Conditions of Wine Fermentation Revealed By Variations in Yeast Regulatory Networks. *BMC Genomics*, 14:681.

Brown CA, Murray AW, Verstrepen KJ (2010). Rapid Expansion and Functional Divergence of Subtelomeric Gene Families in Yeasts. *Curr Biol*, 20(10):895-903.

Bull MJ, Jolley KA, Bray JE, Aerts M, Vandamme P, Maiden MC, Marchesi JR, Mahenthiralingam E (2014). The Domestication of the Probiotic Bacterium *Lactobacillus acidophilus*. *Sci Rep*, 4:7202.

Buser CC, Newcomb RD, Gaskett AC, Goddard MR (2014). Niche Construction Initiates the Evolution of Mutualistic Interactions. *Ecol Lett*, 17(10):1257-1264.

Cabrera E, González-Montelongo R, Giraldez T, Alvarez de la Rosa D, Siverio JM (2014). Molecular Components of Nitrate and Nitrite Efflux in Yeast. *Eukaryot Cell*, 13(2):267-278.

Camarasa C, Sanchez I, Brial P, Bigey F, Dequin S (2011). Phenotypic Landscape of *Saccharomyces cerevisiae* During Wine Fermentation: Evidence for Origin-Dependent Metabolic Traits. *PLoS One*, 6(9):e25147.

Carneiro M, Rubin CJ, Di Palma F, Albert FW, Alföldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Turner-Maier J, Younis S, Afonso S, Aken B, Alves JM, Barrell D, Bolet G, Boucher S, Burbano HA, Campos R, Chang JL, Duranthon V, Fontanesi L, Garreau H, Heiman D, Johnson J, Mage RG, Peng Z, Queney G, Rogel-Gaillard C, Ruffier M, Searle S, Villafuerte R, Xiong A, Young S, Forsberg-Nilsson K, Good JM, Lander ES, Ferrand N, Lindblad-Toh K, Andersson L (2014). Rabbit Genome Analysis Reveals a Polygenic Basis for Phenotypic Change During Domestication. *Science*, 345(6200):1074-1079.

Carroll SB (2005). Evolution At Two Levels: On Genes and Form. *PLoS Biol*, 3(7):e245.

Cavalieri D, Townsend JP, Hartl DL (2000). Manifold Anomalies in Gene Expression in a Vineyard Isolate of *Saccharomyces cerevisiae* Revealed By DNA Microarray Analysis. *Proc Natl Acad Sci U S A*, 97(22):12369-12374.

Cavalieri D, McGovern PE, Hartl DL, Mortimer R, Polsinelli M (2003). Evidence for *S. cerevisiae* Fermentation in Ancient Wine. *J Mol Evol*, 57 Suppl 1:S226-32.

Chandler JA, Eisen JA, Kopp A (2012). Yeast Communities of Diverse *Drosophila* Species: Comparison of Two Symbiont Groups in the Same Hosts. *Appl Environ Microbiol*, 78(20):7327-7336.

Chang PK, Horn BW, Dorner JW (2005). Sequence Breakpoints in the Aflatoxin Biosynthesis Gene Cluster and Flanking Regions in Nonaflatoxigenic *Aspergillus flavus* Isolates. *Fungal Genet Biol*, 42(11):914-923.

Charlesworth B (1998). Measures of Divergence Between Populations and the Effect of Forces That Reduce Variability. *Mol Biol Evol*, 15(5):538-543.

Charron G, Leducq JB, 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 Res*, 14(2):281-288.

Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, Abraham AL, Ceppi M, Conseiller E, Debuchy R, Malagnac F, Goarin A, Silar P, Lacoste S, Sallet E, Bensimon A, Giraud T, Brygoo Y (2014). Multiple Recent Horizontal Transfers of a Large Genomic Region in Cheese Making Fungi. *Nat Commun*, 5:2876.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED (2012). *Saccharomyces* Genome Database: The Genomics Resource of Budding Yeast. *Nucleic Acids Res*, 40(Database issue):D700-5.

Christiaens JF, Franco LM, Cools TL, De Meester L, Michiels J, Wenseleers T, Hassan BA, Yaksi E, Verstrepen KJ (2014). The Fungal Aroma Gene Atf1 Promotes Dispersal of Yeast Cells Through Insect Vectors. *Cell Rep*, 9(2):425-432.

Ciani M, Mannazzu I, Marinangeli P, Clementi F, Martini A (2004). Contribution of Winery-Resident *Saccharomyces cerevisiae* Strains to Spontaneous Grape Must Fermentation. *Antonie Van Leeuwenhoek*, 85(2):159-164.

Clowers KJ, Heilberger J, Piotrowski JS, Will JL, Gasch AP (2015). Ecological and Genetic Barriers Differentiate Natural Populations of *Saccharomyces cerevisiae*. *Mol Biol Evol*, 32(9):2317-2327.

Conant GC, Wolfe KH (2007). Increased Glycolytic Flux as an Outcome of Whole-Genome Duplication in Yeast. *Mol Syst Biol*, 3:129.

Connelly CF, Skelly DA, Dunham MJ, Akey JM (2013). Population Genomics and Transcriptional Consequences of Regulatory Motif Variation in Globally Diverse *Saccharomyces cerevisiae* Strains. *Mol Biol Evol*, 30(7):1605-1613.

Cornille A, Gladieux P, Smulders MJ, Roldán-Ruiz I, Laurens F, Le Cam B, Nersesyan A, Clavel J, Olonova M, Feugey L, Gabrielyan I, Zhang XG, Tenaillon MI, Giraud T (2012). New Insight Into the History of Domesticated Apple: Secondary Contribution of the European Wild Apple to the Genome of Cultivated Varieties. *PLoS Genet*, 8(5):e1002703.

Cromie GA, Hyma KE, Ludlow CL, Garmendia-Torres C, Gilbert TL, May P, Huang AA, Dudley AM, Fay JC (2013). Genomic Sequence Diversity and Population Structure of *Saccharomyces cerevisiae* Assessed By Rad-Seq. *G3 (Bethesda)*, 3(12):2163-2171.

Cruickshank TE, Hahn MW (2014). Reanalysis Suggests That Genomic Islands of Speciation Are Due to Reduced Diversity, Not Reduced Gene Flow. *Mol Ecol*, 23(13):3133-3157.

Cubillos FA, Louis EJ, Liti G (2009). Generation of a Large Set of Genetically Tractable Haploid and Diploid *Saccharomyces* Strains. *FEMS Yeast Res*, 9(8):1217-1225.

Darwin C (1876). The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London, UK: John Murray.

Dashko S, Zhou N, Compagno C, Piškur J (2014). Why, When, and How Did Yeast Evolve Alcoholic Fermentation. *FEMS Yeast Res*, 14(6):826-832.

Demuyter C, Lollier M, Legras JL, Le Jeune C (2004). Predominance of *Saccharomyces uvarum* During Spontaneous Alcoholic Fermentation, for Three Consecutive Years, in an Alsatian Winery. *J Appl Microbiol*, 97(6):1140-1148.

Dermitzakis ET, Clark AG (2002). Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Mol Biol Evol*, 19(7):1114-1121.

Diamond J (2002). Evolution, Consequences and Future of Plant and Animal Domestication. *Nature*, 418(6898):700-707.

Dietrich O, Heun M, Notroff J, Schmidt K, Zarnkow M (2012). The Role of Cult and Feasting in the Emergence of Neolithic Communities. New Evidence From Göbekli Tepe, South-Eastern Turkey. *Antiquity*, 86:674-695.

Doebley JF, Gaut BS, Smith BD (2006). The Molecular Genetics of Crop Domestication. *Cell*, 127(7):1309-1321.

Doniger SW, Fay JC (2007). Frequent Gain and Loss of Functional Transcription Factor Binding Sites. *PLoS Comput Biol*, 3(5):e99.

Doniger SW, Kim HS, Swain D, Corcuera D, Williams M, Yang SP, Fay JC (2008). A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLoS Genet*, 4(8):e1000183.

Douglas GL, Klaenhammer TR (2010). Genomic Evolution of Domesticated Microorganisms. *Annu Rev Food Sci Technol*, 1:397-414.

Douillard FP, Ribbera A, Kant R, Pietilä TE, Järvinen HM, Messing M, Randazzo CL, Paulin L, Laine P, Ritari J, Caggia C, Lähteinen T, Brouns SJ, Satokari R, von Ossowski I, Reunanen J, Palva A, de Vos WM (2013). Comparative Genomic and Functional Analysis of 100 *Lactobacillus rhamnosus* Strains and Their Comparison With Strain Gg. *PLoS Genet*, 9(8):e1003683.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012). Bayesian Phylogenetics With Beauti and the Beast 1.7. *Mol Biol Evol*, 29(8):1969-1973.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL (2004). Genome Evolution in Yeasts. *Nature*, 430(6995):35-44.

Dujon B (2006). Yeasts Illustrate the Molecular Mechanisms of Eukaryotic Genome Evolution. *Trends Genet*, 22(7):375-387.

Dunn B, Levine RP, Sherlock G (2005). Microarray Karyotyping of Commercial Wine Yeast Strains Reveals Shared, as Well as Unique, Genomic Signatures. *BMC Genomics*, 6:53.

Dunn B, Sherlock G (2008). Reconstruction of the Genome Origins and Evolution of the Hybrid Lager Yeast *Saccharomyces pastorianus*. *Genome Res*, 18(10):1610-1623.

Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012). Analysis of the *Saccharomyces cerevisiae* Pan-Genome Reveals a Pool of Copy Number Variants Distributed in Diverse Yeast Strains From Differing Industrial Environments. *Genome Res*, 22(5):908-924.

Edgar RC (2004). Muscle: Multiple Sequence Alignment With High Accuracy and High Throughput. *Nucleic Acids Res*, 32(5):1792-1797.

Ehrenreich IM, Torabi N, Jia Y, Kent J, Martis S, Shapiro JA, Gresham D, Caudy AA, Kruglyak L (2010). Dissection of Genetically Complex Traits With Extremely Large Pools of Yeast Segregants. *Nature*, 464(7291):1039-1042.

Engle EK, Fay JC (2012). Divergence of the Yeast Transcription Factor Fzf1 Affects Sulfite Resistance. *PLoS Genet*, 8(6):e1002763.

Engle EK, Fay JC (2013). Zrt1 Harbors an Excess of Nonsynonymous Polymorphism and Shows Evidence of Balancing Selection in *Saccharomyces cerevisiae*. *G3 (Bethesda)*,

Erny C, Raoult P, Alais A, Butterlin G, Delobel P, Matei-Radoi F, Casaregola S, Legras JL (2012). Ecological Success of a Group of *Saccharomyces cerevisiae*/*Saccharomyces kudriavzevii* Hybrids in the Northern European Wine-Making Environment. *Appl Environ Microbiol*, 78(9):3256-3265.

Evanno G, Regnaut S, Goudet J (2005). Detecting the Number of Clusters of Individuals Using the Software Structure: A Simulation Study. *Mol Ecol*, 14(8):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(4):1567-1587.

Fay JC, McCullough HL, Sniegowski PD, Eisen MB (2004). Population Genetic Variation in Gene Expression is Associated With Phenotypic Variation in *Saccharomyces cerevisiae*. *Genome Biol*, 5(4):R26.

Fay JC, Benavides JA (2005). Evidence for Domesticated and Wild Populations of *Saccharomyces cerevisiae*. *PLoS Genet*, 1:66-71.

Fernández-Espinar TM, Barrio E, Querol A (2003). Analysis of the Genetic Variability in the Species of the *Saccharomyces* sensu stricto Complex. *Yeast*, 20(14):1213-1226.

Fleet G (1998). Microbiology of alcoholic beverages. In: Wood BJ, editor. Microbiology of Fermented Foods. Volume1. London: Blackie Academic & Professional. p. 217-262.

Fleet G (2003). Yeast Interactions and Wine Flavour. *International Journal of Food Microbiology*, 86(1-2):11-22.

Fleet GH, Lafon-Lafourcade S, Ribéreau-Gayon P (1984). Evolution of Yeasts and Lactic Acid Bacteria During Fermentation and Storage of Bordeaux Wines. *Appl Environ Microbiol*, 48(5):1034-1038.

Francesca N, Canale DE, Settanni L, Moschetti G (2012). Dissemination of Wine-Related Yeasts By Migratory Birds. *Environ Microbiol Rep*, 4(1):105-112.

Galeote V, Novo M, Salema-Oom M, Brion C, Valério E, Gonçalves P, Dequin S (2010). Fsy1, a Horizontally Transferred Gene in the *Saccharomyces cerevisiae* Ec1118 Wine Yeast Strain, Encodes a High-Affinity Fructose/h+ Symporter. *Microbiology*, 156(Pt 12):3754-3761.

Galeote V, Bigey F, Beyne E, Novo M, Legras JL, Casaregola S, Dequin S (2011). Amplification of a *Zygosaccharomyces bailii* DNA Segment in Wine Yeast Genomes By Extrachromosomal Circular DNA Formation. *PLoS One*, 6(3):e17872.

Gepts P (2004). Crop domestication as a long-term selection experiment. In: Janick J, editor. Plant Breeding Reviews, Volume 24, Part 2: Long-term Selection: Crops, Animals, and Bacteria. Wiley. p. 1-44.

Gibbons JG, Salichos L, Slot JC, Rinker DC, McGary KL, King JG, Klich MA, Tabb DL, McDonald WH, Rokas A (2012). The Evolutionary Imprint of Domestication on Genome Variation and Function of the Filamentous Fungus *Aspergillus oryzae*. *Curr Biol*, 22(15):1403-1409.

Gibbons JG, Rinker DC (2015). The Genomics of Microbial Domestication in the Fermented Food Environment. *Curr Opin Genet Dev*, 35:1-8.

Gibson B, Liti G (2015). *Saccharomyces pastorianus*: Genomic Insights Inspiring Innovation for Industry. *Yeast*, 32(1):17-27.

Giudici P, Caggia C, Pulvirenti A, Zambonelli C, Rainieri S (1998). Electrophoretic Profile of Hybrids Between Cryotolerant and Non-Cryotolerant *Saccharomyces* Strains. *Lett Appl Microbiol*, 27(1):31-34.

Goddard MR (2008). Quantifying the Complexities of *Saccharomyces cerevisiae*'s Ecosystem Engineering Via Fermentation. *Ecology*, 89(8):2077-2082.

Goddard MR, Greig D (2015). *Saccharomyces cerevisiae*: A Nomadic Yeast With No Niche. *FEMS Yeast Res*, 15(3)

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996). Life With 6000 Genes. *Science*, 274(5287):546, 563-7.

Gonçalves P, Valério E, Correia C, de Almeida JM, Sampaio JP (2011). Evidence for Divergent Evolution of Growth Temperature Preference in Sympatric *Saccharomyces* Species. *PLoS One*, 6:e20739.

González SS, Barrio E, Gafner J, Querol A (2006). Natural Hybrids from *Saccharomyces cerevisiae*, *Saccharomyces bayanus* and *Saccharomyces kudriavzevii* in Wine Fermentations. *FEMS Yeast Res*, 6(8):1221-1234.

González SS, Gallo L, Climent MA, Barrio E, Querol A (2007). Enological Characterization of Natural Hybrids From *Saccharomyces cerevisiae* and *S. kudriavzevii*. *Int J Food Microbiol*, 116(1):11-18.

González SS, Barrio E, Querol A (2008). Molecular Characterization of New Natural Hybrids of *Saccharomyces cerevisiae* and *S. kudriavzevii* in Brewing. *Appl Environ Microbiol*, 74(8):2314-2320.

Goto-Yamamoto N, Kitano K, Shiki K, Yoshida Y, Suzuki T, Iwata T, Yamani Y, Hara S (1998). Ssu1-R, a Sulfite Resistance Gene of Wine Yeast, is an Allele of Ssu1 With a Different Upstream Sequence. *Journal of Fermentation and Bioengineering*, 86(5):427-433.

Greig D, Louis EJ, Borts RH, Travisano M (2002). Hybrid Speciation in Experimental Populations of Yeast. *Science*, 298(5599):1773-1775.

Greig D, Travisano M, Louis EJ, Borts RH (2003). A Role for the Mismatch Repair System During Incipient Speciation in *Saccharomyces*. *J Evol Biol*, 16(3):429-437.

Greig D (2009). Reproductive Isolation in *Saccharomyces*. *Heredity (Edinb)*, 102(1):39-44.

Guindon S, Gascuel O (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies By Maximum Likelihood. *Syst Biol*, 52(5):696-704.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009). Inferring the Joint Demographic History of Multiple Populations From Multidimensional Snp Frequency Data. *PLoS Genet*, 5(10):e1000695.

Hagman A, Piškur J (2015). A Study on the Fundamental Mechanism and the Evolutionary Driving Forces Behind Aerobic Fermentation in Yeast. *PLoS One*, 10(1):e0116942.

Herskowitz I (1988). Life Cycle of the Budding Yeast *Saccharomyces cerevisiae*. *Microbiol Rev*, 52(4):536-553.

Hewitt GM (2004). The Structure of Biodiversity - Insights From Molecular Phylogeography. *Front Zool*, 1:4.

Hill RS (1992). *Nothofagus*: Evolution from a Southern Perspective. *Trends Ecol Evol*, 7(6):190-194.

Hittinger CT, Gonçalves P, Sampaio JP, Dover J, Johnston M, Rokas A (2010). Remarkably Ancient Balanced Polymorphisms in a Multi-Locus Gene Network. *Nature*, 464(7285):54-58.

Hittinger CT (2013). *Saccharomyces* Diversity and Evolution: A Budding Model Genus. *Trends Genet*, 29(5):309-317.

Hoang D, Kopp A, Chandler JA (2015). Interactions Between *Drosophila* and Its Natural Yeast Symbionts-is *Saccharomyces cerevisiae* a Good Model for Studying the Fly-Yeast Relationship. *PeerJ*, 3:e1116.

Hou J, Friedrich A, de Montigny J, Schacherer J (2014). Chromosomal Rearrangements as a Major Mechanism in the Onset of Reproductive Isolation in *Saccharomyces cerevisiae*. *Curr Biol*, 24(10):1153-1159.

Hudson RR (2002). Generating Samples Under a Wright-Fisher Neutral Model of Genetic Variation. *Bioinformatics*, 18(2):337-338.

Hudson RR, Kaplan NL (1985). Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences. *Genetics*, 111(1):147-164.

Hudson RR, Boos DD, Kaplan NL (1992). A Statistical Test for Detecting Geographic Subdivision. *Mol Biol Evol*, 9(1):138-151.

Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia JM, Cartwright RA, Elshire RJ, Glaubitz JC, Guill KE, Kaeppler SM, Lai J, Morrell PL, Shannon LM, Song C, Springer NM, Swanson-Wagner RA, Tiffin P, Wang J, Zhang G, Doebley J, McMullen MD, Ware D, Buckler ES, Yang S, Ross-Ibarra J (2012). Comparative Population Genomics of Maize Domestication and Improvement. *Nat Genet*, 44(7):808-811.

Hunter N, Chambers SR, Louis EJ, Borts RH (1996). The Mismatch Repair System Contributes to Meiotic Sterility in an Interspecific Yeast Hybrid. *EMBO J*, 15(7):1726-1733.

Huson DH, Bryant D (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol*, 23(2):254-267.

Hutter S, Vilella AJ, Rozas J (2006). Genome-Wide DNA Polymorphism Analyses Using Variscan. *BMC Bioinformatics*, 7:409.

Hyma KE, Saerens SM, Verstrepen KJ, Fay JC (2011). Divergence in Wine Characteristics Produced By Wild and Domesticated Strains of *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 11(7):540-551.

Hyma KE, Fay JC (2013). Mixing of Vineyard and Oak-Tree Ecotypes of *Saccharomyces cerevisiae* in North American Vineyards. *Mol Ecol*, 22(11):2917-2930.

Ihmels J, Bergmann S, Gerami-Nejad M, Yanai I, McClellan M, Berman J, Barkai N (2005). Rewiring of the Yeast Transcriptional Network Through the Evolution of Motif Usage. *Science*, 309(5736):938-940.

Innan H, Kim Y (2004). Pattern of Polymorphism After Strong Artificial Selection in a Domestication Event. *Proc Natl Acad Sci U S A*, 101(29):10667-10672.

Jakobsson M, Rosenberg NA (2007). Clumpp: A Cluster Matching and Permutation Program for Dealing With Label Switching and Multimodality in Analysis of Population Structure. *Bioinformatics*, 23(14):1801-1806.

Joffe AH (1998). Alcohol and Social Complexity in Ancient Western Asia. *Current Anthropology*, 39(3):297-322.

Johnson EA, Echavarri-Erasun C (2011). Yeast Biotechnology. In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts: A Taxonomic Study. Elsevier. p. 21-44.

Johnson LJ, Koufopanou V, Goddard MR, Hetherington R, Schäfer SM, Burt A (2004). Population *Genetics of the Wild Yeast Saccharomyces paradoxus*. *Genetics*, 166(1):43-52.

Johnston M (1999). Feasting, Fasting and Fermenting. Glucose Sensing in Yeast and Other Cells. *Trends Genet*, 15(1):29-33.

Jolly NP, Varela C, Pretorius IS (2014). Not Your Ordinary Yeast: Non-*Saccharomyces* Yeasts in Wine Production Uncovered. *FEMS Yeast Res*, 14(2):215-237.

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.

Keller-Seitz MU, Certa U, Sengstag C, Würgler FE, Sun M, Fasullo M (2004). Transcriptional Response of Yeast to Aflatoxin B1: Recombinational Repair Involving Rad51 and Rad1. *Mol Biol Cell*, 15(9):4321-4336.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003). Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements. *Nature*, 423(6937):241-254.

Koufopanou V, Hughes J, Bell G, Burt A (2006). The Spatial Scale of Genetic Differentiation in a Model Organism: The Wild Yeast *Saccharomyces paradoxus*. *Philos Trans R Soc Lond B Biol Sci*, 361(1475):1941-1946.

Koufopanou V, Swire J, Lomas S, Burt A (2013). Primers for Fourteen Protein-Coding Genes and the Deep Phylogeny of the True Yeasts. *FEMS Yeast Res*, 13(6):574-584.

Kowallik V, Miller E, Greig D (2015). The Interaction of *Saccharomyces paradoxus* with Its Natural Competitors on Oak Bark. *Mol Ecol*, 24(7):1596-1610.

Kowallik V, Greig D (2016). A Systematic Forest Survey Showing an Association of *Saccharomyces* with Oak Leaf Litter. *Environ Microbiol Rep*,

Kumar S, Stecher G, Tamura K (2016). Mega7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*,

Kvitek DJ, Will JL, Gasch AP (2008). Variations in Stress Sensitivity and Genomic Expression in Diverse *S. cerevisiae* Isolates. *PLoS Genet*, 4(10):e1000223.

Larson G, Fuller DQ (2014). The Evolution of Animal Domestication. *Annual Review of Ecology, Evolution, and Systematics*, 45:115-136.

Lawson DJ, Hellenthal G, Myers S, Falush D (2012). Inference of Population Structure Using Dense Haplotype Data. *PLoS Genet*, 8(1):e1002453.

Le Jeune C, Lollier M, Demuyter C, Erny C, Legras JL, Aigle M, Masneuf-Pomarède I (2007). Characterization of Natural Hybrids of *Saccharomyces cerevisiae* and *Saccharomyces bayanus* Var. *uvarum*. *FEMS Yeast Res*, 7(4):540-549.

League GP, Slot JC, Rokas A (2012). The Asp3 Locus in *Saccharomyces cerevisiae* Originated By Horizontal Gene Transfer From *Wickerhamomyces*. *FEMS Yeast Res*, 12(7):859-863.

Leducq JB, Charron G, Samani P, Dubé AK, Sylvester K, James B, Almeida P, Sampaio JP, Hittinger CT, Bell G, Landry CR (2014). Local Climatic Adaptation in a Widespread Microorganism. *Proc Biol Sci*, 281(1777):20132472.

Legras JL, Merdinoglu D, Cornuet JM, Karst F (2007). Bread, Beer and Wine: *Saccharomyces cerevisiae* Diversity Reflects Human History. *Mol Ecol*, 16(10):2091-2102.

Levy SE, Myers RM (2016). Advancements in Next-Generation Sequencing. *Annu Rev Genom Hum Genet*, 17(1):95-115.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 GPDPS (2009). The Sequence Alignment/map Format and Samtools. *Bioinformatics*, 25(16):2078-2079.

Li H, Durbin R (2010). Fast and Accurate Long-Read Alignment With Burrows-Wheeler Transform. *Bioinformatics*, 26(5):589-595.

Libkind D, Hittinger CT, Valério E, Gonçalves C, Dover J, Johnston M, Gonçalves P, Sampaio JP (2011). Microbe Domestication and the Identification of the Wild Genetic Stock of Lager-Brewing Yeast. *Proc Natl Acad Sci U S A*, 108(35):14539-14544.

Librado P, Rozas J (2009). Dnasp V5: A Software for Comprehensive Analysis of DNA Polymorphism Data. *Bioinformatics*, 25(11):1451-1452.

Lin Z, Li WH (2011). Expansion of Hexose Transporter Genes Was Associated With the Evolution of Aerobic Fermentation in Yeasts. *Mol Biol Evol*, 28(1):131-142.

Liti G, Barton DB, Louis EJ (2006). Sequence Diversity, Reproductive Isolation and Species Concepts in *Saccharomyces*. *Genetics*, 174(2):839-850.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V, Tsai IJ, Bergman CM, Bensasson D, O'Kelly MJ, van Oudenaarden A, Barton DB, Bailes E, Nguyen AN, Jones M, Quail MA, Goodhead I, Sims S, Smith F, Blomberg A, Durbin R, Louis EJ (2009). Population Genomics of Domestic and Wild Yeasts. *Nature*, 458(7236):337-341.

Liti G, Louis EJ (2012). Advances in Quantitative Trait Analysis in Yeast. *PLoS Genet*, 8:e1002912.

Liti G, Nguyen Ba AN, Blythe M, Müller CA, Bergström A, Cubillos FA, Dafhnis-Calas F, Khoshraftar S, Malla S, Mehta N, Siow CC, Warringer J, Moses AM, Louis EJ, Nieduszynski CA (2013). High Quality De Novo Sequencing and Assembly of the *Saccharomyces arboricolus* Genome. *BMC Genomics*, 14:69.

Liti G, Peruffo A, James SA, Roberts IN, Louis EJ (2005). Inferences of Evolutionary Relationships from a Population Survey of Ltr-Retrotransposons and Telomeric-Associated Sequences in the *Saccharomyces* sensu stricto Complex. *Yeast*, 22(3):177-192.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG (2014). Metagenomic 16s Rdna Illumina Tags Are a Powerful Alternative to Amplicon Sequencing to Explore Diversity and Structure of Microbial Communities. *Environ Microbiol*, 16(9):2659-2671.

Lopandic K, Gangl H, Wallner E, Tscheik G, Leitner G, Querol A, Borth N, Breitenbach M, Prillinger H, Tiefenbrunner W (2007). Genetically Different Wine Yeasts Isolated From Austrian Vine-Growing Regions Influence Wine Aroma Differently and Contain Putative Hybrids Between *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*. *FEMS Yeast Res*, 7(6):953-965.

Ludlow CL, Cromie GA, Garmendia-Torres C, Sirr A, Hays M, Field C, Jeffery EW, Fay JC, Dudley AM (2016). Independent Origins of Yeast Associated With Coffee and Cacao Fermentation. *Curr Biol*, 26(7):965-971.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK (2008). A Genome-Wide View of the Spectrum of Spontaneous Mutations in Yeast. *Proc Natl Acad Sci U S A*, 105(27):9272-9277.

Machida M, Yamada O, Gomi K (2008). Genomics of *Aspergillus oryzae*: Learning from the History of Koji Mold and Exploration of Its Future. *DNA Res*, 15(4):173-183.

Maclean CJ, Greig D (2008). Prezygotic Reproductive Isolation Between *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. *BMC Evol Biol*, 8:1.

Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee JH, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W, Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, Mills D (2006). Comparative Genomics of the Lactic Acid Bacteria. *Proc Natl Acad Sci U S A*, 103(42):15611-15616.

Makarova KS, Koonin EV (2007). Evolutionary Genomics of Lactic Acid Bacteria. *J Bacteriol*, 189(4):1199-1208.

Markgraf V, McGlone M, Hope G (1995). Neogene Paleoenvironmental and Paleoclimatic Change in Southern Temperate Ecosystems - a Southern Perspective. *Trends Ecol Evol*, 10(4):143-147.

Marsit S, Mena A, Bigey F, Sauvage FX, Couloux A, Guy J, Legras JL, Barrio E, Dequin S, Galeote V (2015). Evolutionary Advantage Conferred By an Eukaryote-to-eukaryote Gene Transfer Event in Wine Yeasts. *Mol Biol Evol*, 32(7):1695-1707.

Marsit S, Dequin S (2015). Diversity and Adaptive Evolution of *Saccharomyces* Wine Yeast: A Review. *FEMS Yeast Res*, 15(7)

Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M (2007). Transcriptional Rewiring of Fungal Galactose-Metabolism Circuitry. *Curr Biol*, 17(12):1007-1013.

Martini A (1993). Origin and Domestication of the Wine Yeast *Saccharomyces cerevisiae*. *Journal of Wine Research*, 4(3):165-176.

Martini A, Ciani M, Scorzetti G (1996). Direct Enumeration and Isolation of Wine Yeasts From Grape Surfaces. *Am J Enol Vitic*, 47(4):435-440.

Marullo P, Aigle M, Bely M, Masneuf-Pomarède I, Durrens P, Dubourdieu D, Yvert G (2007). Single Qtl Mapping and Nucleotide-Level Resolution of a Physiologic Trait in Wine *Saccharomyces cerevisiae* Strains. *FEMS Yeast Res*, 7(6):941-952.

Masneuf-Pomarède I, Bely M, Marullo P, Lonvaud-Funel A, Dubourdieu D (2010). Reassessment of Phenotypic Traits for *Saccharomyces bayanus* Var. *uvarum* Wine Yeast Strains. *Int J Food Microbiol*, 139(1-2):79-86.

Masneuf I, Hansen J, Groth C, Piskur J, Dubourdieu D (1998). New Hybrids Between *Saccharomyces* sensu stricto Yeast Species Found Among Wine and Cider Production Strains. *Appl Environ Microbiol*, 64(10):3887-3892.

Mathiasen P, Premoli AC (2010). Out in the Cold: Genetic Variation of *Nothofagus pumilio* (Nothofagaceae) Provides Evidence for Latitudinally Distinct Evolutionary Histories in Austral South America. *Mol Ecol*, 19(2):371-385.

McDonald JH, Kreitman M (1991). Adaptive Protein Evolution At the Adh Locus in *Drosophila*. *Nature*, 351(6328):652-654.

McGovern PE (2007). Ancient Wine: The Search for the Origins of Viniculture. Princeton University Press.

McGovern PE, Glusker DL, Exner LJ (1996). Neolithic Resinated Wine. *Nature*, 381:480-481.

McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, Nuñez A, Butrym ED, Richards MP, Wang CS, Cheng G, Zhao Z, Wang C (2004). Fermented Beverages of Pre- and Proto-Historic China. *Proc Natl Acad Sci U S A*, 101(51):17593-17598.

McGovern PE, Luley BP, Rovira N, Mirzoian A, Callahan MP, Smith KE, Hall GR, Davidson T, Henkin JM (2013). Beginning of Viniculture in France. *Proc Natl Acad Sci U S A*, 110(25):10147-10152.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010). Deriving the Consequences of Genomic Variants With the Ensembl Api and Snp Effect Predictor. *Bioinformatics*, 26(16):2069-2070.

Mercado L, Dalcero A, Masuelli R, Combina M (2007). Diversity of *Saccharomyces* Strains on Grapes and Winery Surfaces: Analysis of Their Contribution to Fermentative Flora of Malbec Wine from Mendoza (Argentina) During Two Consecutive Years. *Food Microbiology*, 24(4):403-412.

Merico A, Sulo P, Piškur J, Compagno C (2007). Fermentative Lifestyle in Yeasts Belonging to the *Saccharomyces* Complex. *FEBS J*, 274(4):976-989.

Meußdoerffer FG (2009). A comprehensive history of beer brewing. In: Eßlinger HM, editor. Handbook of Brewing: Processes, Technology, Markets. Weinheim (Germany): Wiley-VCH. p. 1-42.

Møller K, Olsson L, Piskur J (2001). Ability for Anaerobic Growth is Not Sufficient for Development of the Petite Phenotype in *Saccharomyces kluyveri*. *J Bacteriol*, 183(8):2485-2489.

Morales L, Dujon B (2012). Evolutionary Role of Interspecies Hybridization and Genetic Exchanges in Yeasts. *Microbiol Mol Biol Rev*, 76(4):721-739.

Mortimer R, Polsinelli M (1999). On the Origins of Wine Yeast. *Res Microbiol*, 150(3):199-204.

Muller LA, McCusker JH (2009). A Multispecies-Based Taxonomic Microarray Reveals Interspecies Hybridization and Introgression in *Saccharomyces cerevisiae*. *FEMS Yeast Res*, 9(1):143-152.

Murphy HA, Kuehne HA, Francis CA, Sniegowski PD (2006). Mate Choice Assays and Mating Propensity Differences in Natural Yeast Populations. *Biol Lett*, 2(4):553-556.

Murphy HA, Zeyl CW (2012). Prezygotic Isolation Between *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* Through Differences in Mating Speed and Germination Timing. *Evolution*, 66(4):1196-1209.

Naumov GI, Naumova ES, Sniegowski PD (1998). *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* Are Associated With Exudates of North American Oaks. *Can J Microbiol*, 44:1045-1050.

Naumov GI, Gazdiev DO, Naumova ES (2003). The Finding of the Yeast Species *Saccharomyces bayanus* in Far East Asia. *Microbiology*, 72(6):834-839.

Naumov GI, Serpova EV, Naumova ES (2006). A Genetically Isolated Population of *Saccharomyces cerevisiae* in Malaysia. *Microbiology*, 75(2):201-205.

Naumov GI (1996). Genetic Identification of Biological Species in the *Saccharomyces* sensu stricto Complex. *Journal of Industrial Microbiology*, 17:295-302.

Naumov GI, James SA, Naumova ES, Louis EJ, Roberts IN (2000a). Three New Species in the *Saccharomyces* sensu stricto Complex: *Saccharomyces cariocanus*, *Saccharomyces kudriavzevii* and *Saccharomyces mikatae*. *International Journal of Systematic and Evolutionary Microbiology*, 50:1931-1942.

Naumov GI, Masneuf I, Naumova ES, Aigle M, Dubourdieu D (2000b). Association of *Saccharomyces bayanus* Var. *uvarum* with Some French Wines: Genetic Analysis of Yeast Populations. *Res Microbiol*, 151(8):683-691.

Naumov GI, Nguyen HV, Naumova ES, Michel A, Aigle M, Gaillardin C (2001). Genetic Identification of *Saccharomyces bayanus* Var. *uvarum*, a Cider-Fermenting Yeast. *Int J Food Microbiol*, 65(3):163-171.

Naumova ES, Naumov GI, Michailova YV, Martynenko NN, Masneuf-Pomarède I (2011). Genetic Diversity Study of the Yeast *Saccharomyces bayanus* Var. *uvarum* Reveals Introgressed Subtelomeric *Saccharomyces cerevisiae* Genes. *Res Microbiol*, 162(2):204-213.

Nguyen HV, Gaillardin C (2005). Evolutionary Relationships Between the Former Species *Saccharomyces uvarum* and the Hybrids *Saccharomyces bayanus* and *Saccharomyces pastorianus*; Reinstatement of *Saccharomyces uvarum* (Beijerinck) as a Distinct Species. *FEMS Yeast Res*, 5(4-5):471-483.

Nguyen HV, Legras JL, Neuvéglise C, Gaillardin C (2011). Deciphering the Hybridisation History Leading to the Lager Lineage Based on the Mosaic Genomes of *Saccharomyces bayanus* Strains Nbrc1948 and Cbs380. *PLoS One*, 6(10):e25821.

Noor MA, Bennett SM (2009). Islands of Speciation or Mirages in the Desert? Examining the Role of Restricted Recombination in Maintaining Species. *Heredity (Edinb)*, 103(6):439-444.

Novo M, Bigey F, Beyne E, Galeote V, Gavory F, Mallet S, Cambon B, Legras JL, Wincker P, Casaregola S, Dequin S (2009). Eukaryote-to-eukaryote Gene Transfer Events Revealed By the Genome Sequence of the Wine Yeast *Saccharomyces cerevisiae* Ec1118. *Proc Natl Acad Sci U S A*, 106(38):16333-16338.

O'Sullivan O, O'Callaghan J, Sangrador-Vegas A, McAuliffe O, Slattery L, Kaleta P, Callanan M, Fitzgerald GF, Ross RP, Beresford T (2009). Comparative Genomics of Lactic Acid Bacteria Reveals a Niche-Specific Gene Set. BMC Microbiol, 9:50.

Park H, Bakalinsky AT (2000). Ssu1 Mediates Sulphite Efflux in *Saccharomyces cerevisiae*. *Yeast*, 16(10):881-888.

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 Res*, 12(10):1533-1539.

Pérez-Través L, Lopes CA, Querol A, Barrio E (2014). On the Complexity of the *Saccharomyces bayanus* Taxon: Hybridization and Potential Hybrid Speciation. *PLoS One*, 9(4):e93729.

Peris D, Sylvester K, Libkind D, Gonçalves P, Sampaio JP, Alexander WG, Hittinger CT (2014). Population Structure and Reticulate Evolution of *Saccharomyces eubayanus* and Its Lager-Brewing Hybrids. *Mol Ecol*, 23(8):2031-2045.

Pfeiffer T, Morley A (2014). An Evolutionary Perspective on the Crabtree Effect. *Front Mol Biosci*, 1:17.

Phaff HJ, Miller MWM, Recca JA, Shifrine M, Mrak EM (1956). Yeasts Found in the Alimentary Canal of *Drosophila*. *Ecology*, 37(3):533-538.

Piškur J, Rozpedowska E, Polakova S, Merico A, Compagno C (2006). How Did *Saccharomyces* Evolve to Become a Good Brewer. *Trends Genet*, 22(4):183-186.

Pretorius IS (2000). Tailoring Wine Yeast for the New Millennium: Novel Approaches to the Ancient Art of Winemaking. *Yeast*, 16(8):675-729.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945-959.

Pronk JT, Yde Steensma H, Van Dijken JP (1996). Pyruvate Metabolism in *Saccharomyces cerevisiae*. *Yeast*, 12(16):1607-1633.

Querol A, Fernández-Espinar MT, del Olmo M, Barrio E (2003). Adaptive Evolution of Wine Yeast. *Int J Food Microbiol*, 86(1-2):3-10.

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

Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR (2005). Catarrhine Primate Divergence Dates Estimated From Complete Mitochondrial Genomes: Concordance With Fossil and Nuclear DNA Evidence. *J Hum Evol*, 48(3):237-257.

Rementeria A, Rodriguez JA, Cadaval A, Amenabar R, Muguruza JR, Hernando FL, Sevilla MJ (2003). Yeast Associated With Spontaneous Fermentations of White Wines From the “txakoli De Bizkaia” Region (Basque Country, North Spain). *Int J Food Microbiol*, 86(1-2):201-207.

Replansky T, Koufopanou V, Greig D, Bell G (2008). *Saccharomyces* sensu stricto as a Model System for Evolution and Ecology. *Trends Ecol Evol*, 23(9):494-501.

Reuter M, Bell G, Greig D (2007). Increased Outbreeding in Yeast in Response to Dispersal By an Insect Vector. *Curr Biol*, 17(3):R81-3.

Robinson HA, Pinharanda A, Bensasson D (2016). Summer Temperature Can Predict the Distribution of Wild Yeast Populations. *Ecol Evol*, 6(4):1236-1250.

Rodríguez ME, Pérez-Través L, Sangorrín MP, Barrio E, Lopes CA (2014). *Saccharomyces eubayanus* and *Saccharomyces uvarum* Associated with the Fermentation of Araucaria Araucana Seeds in Patagonia. *FEMS Yeast Res*, 14(6):948-965.

Roesti M, Salzburger W, Berner D (2012). Uninformative Polymorphisms Bias Genome Scans for Signatures of Selection. *BMC Evol Biol*, 12:94.

Ronald PC (2014). Lab to Farm: Applying Research on Plant Genetics and Genomics to Crop Improvement. *PLoS Biol*, 12(6):e1001878.

Ross RP, Morgan S, Hill C (2002). Preservation and Fermentation: Past, Present and Future. *Int J Food Microbiol*, 79(1-2):3-16.

Rossouw D, van den Dool AH, Jacobson D, Bauer FF (2010). Comparative Transcriptomic and Proteomic Profiling of Industrial Wine Yeast Strains. *Appl Environ Microbiol*, 76(12):3911-3923.

Rossouw D, Jacobson D, Bauer FF (2012). Transcriptional Regulation and the Diversification of Metabolism in Wine Yeast Strains. *Genetics*, 190(1):251-261.

Rossouw D, Bagheri B, Setati ME, Bauer FF (2015). Co-Flocculation of Yeast Species, a New Mechanism to Govern Population Dynamics in Microbial Ecosystems. *PLoS One*, 10(8):e0136249.

Salque M, Bogucki PI, Pyzel J, Sobkowiak-Tabaka I, Grygiel R, Szmyt M, Evershed RP (2013). Earliest Evidence for Cheese Making in the Sixth Millennium Bc in Northern Europe. *Nature*, 493(7433):522-525.

Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, Treangen TJ, Schatz MC, Delcher AL, Roberts M, Marçais G, Pop M, Yorke JA (2012). Gage: A Critical Evaluation of Genome Assemblies and Assembly Algorithms. *Genome Res*, 22(3):557-567.

Salvadó Z, Arroyo-López FN, Guillamón JM, Salazar G, Querol A, Barrio E (2011). Temperature Adaptation Markedly Determines Evolution Within the Genus *Saccharomyces*. *Appl Environ Microbiol*, 77(7):2292-2302.

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*. *Appl Environ Microbiol*, 74(7):2144-2152.

Sarver A, DeRisi J (2005). Fzf1p Regulates an Inducible Response to Nitrosative Stress in *Saccharomyces cerevisiae*. *Mol Biol Cell*, 16(10):4781-4791.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, McCarthy S, Montgomery SH, Schwalie PC, Tang YA, Ward MC, Xue Y, Yngvadottir B, Alkan C, Andersen LN, Ayub Q, Ball EV, Beal K, Bradley BJ, Chen Y, Clee CM, Fitzgerald S, Graves TA, Gu Y, Heath P, Heger A, Karakoc E, Kolb-Kokocinski A, Laird GK, Lunter G, Meader S, Mort M, Mullikin JC, Munch K, O'Connor TD, Phillips AD, Prado-Martinez J, Rogers AS, Sajjadian S, Schmidt D, Shaw K, Simpson JT, Stenson PD, Turner DJ, Vigilant L, Vilella AJ, Whitener W, Zhu B, Cooper DN, de Jong P, Dermitzakis ET, Eichler EE, Flicek P, Goldman N, Mundy NI, Ning Z, Odom DT, Ponting CP, Quail MA, Ryder OA, Searle SM, Warren WC, Wilson RK, Schierup MH, Rogers J, Tyler-Smith C, Durbin R (2012). Insights Into Hominid Evolution From the Gorilla Genome Sequence. *Nature*, 483(7388):169-175.

Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT (2011). The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the *Saccharomyces* sensu stricto Genus. *G3 (Bethesda)*, 1:11-25.

Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009). Comprehensive Polymorphism Survey Elucidates Population Structure of *Saccharomyces cerevisiae*. *Nature*, 458(7236):342-345.

Sicard D, Legras JL (2011). Bread, Beer and Wine: Yeast Domestication in the *Saccharomyces* sensu stricto Complex. *C R Biol*, 334(3):229-236.

Siezen RJ, Starrenburg MJ, Boekhorst J, Renckens B, Molenaar D, van Hylckama Vlieg JE (2008). Genome-Scale Genotype-Phenotype Matching of Two Lactococcus Lactis Isolates From Plants Identifies Mechanisms of Adaptation to the Plant Niche. *Appl Environ Microbiol*, 74(2):424-436.

Sipiczki M (2008). Interspecies Hybridization and Recombination in *Saccharomyces* Wine Yeasts. *FEMS Yeast Res*, 8(7):996-1007.

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 Res*, 1(4):299-306.

Spencer JFT, Spencer DM (2009). Yeasts in Natural and Artificial Habitats. Berlin Heidelberg: Springer-Verlag.

Spor A, Nidelet T, Simon J, Bourgais A, de Vienne D, Sicard D (2009). Niche-Driven Evolution of Metabolic and Life-History Strategies in Natural and Domesticated Populations of *Saccharomyces cerevisiae*. *BMC Evol Biol*, 9:296.

Stamatakis A (2006). Raxml-Vi-hpc: Maximum Likelihood-Based Phylogenetic Analyses With Thousands of Taxa and Mixed Models. *Bioinformatics*, 22(21):2688-2690.

Stefanini I, Dapporto L, Legras JL, Calabretta A, Di Paola M, De Filippo C, Viola R, Capretti P, Polsinelli M, Turillazzi S, Cavalieri D (2012). Role of Social Wasps in *Saccharomyces cerevisiae* Ecology and Evolution. *Proc Natl Acad Sci U S A*, 109(33):13398-13403.

Stefanini I, Dapporto L, Berná L, Polsinelli M, Turillazzi S, Cavalieri D (2016). Social Wasps Are a *Saccharomyces* Mating Nest. *Proc Natl Acad Sci U S A*, 113(8):2247-2251.

Steinkraus KH (1996). Handbook of indigenous fermented foods. New York: Marcel Dekker.

Stelkens RB, Brockhurst MA, Hurst GD, Miller EL, Greig D (2014). The Effect of Hybrid Transgression on Environmental Tolerance in Experimental Yeast Crosses. *J Evol Biol*, 27(11):2507-2519.

Stoletzki N, Eyre-Walker A (2011). Estimation of the Neutrality Index. *Mol Biol Evol*, 28(1):63-70.

Strope PK, Skelly DA, Kozmin SG, Mahadevan G, Stone EA, Magwene PM, Dietrich FS, McCusker JH (2015). The 100-Genomes Strains, an *S. cerevisiae* Resource That Illuminates Its Natural Phenotypic and Genotypic Variation and Emergence as an Opportunistic Pathogen. *Genome Res*,

25(5):762-774.

Sweeney JY, Kuehne HA, Sniegowski PD (2004). Sympatric Natural *Saccharomyces cerevisiae* and *S. paradoxus* Populations Have Different Thermal Growth Profiles. *FEMS Yeast Res*, 4(4-5):521-525.

Swiegers JH, Kievit RL, Siebert T, Lattey KA, Bramley BR, Francis IL, King ES, Pretorius IS (2009). The Influence of Yeast on the Aroma of Sauvignon Blanc Wine. *Food Microbiol*, 26(2):204-211.

Sylvester K, Wang QM, James B, Mendez R, Hulfachor AB, Hittinger CT (2015). Temperature and Host Preferences Drive the Diversification of *Saccharomyces* and Other Yeasts: A Survey and the Discovery of Eight New Yeast Species. *FEMS Yeast Res*, 15(3)

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). Mega5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol*, 28(10):2731-2739.

Taylor MW, Tsai P, Anfang N, Ross HA, Goddard MR (2014). Pyrosequencing Reveals Regional Differences in Fruit-Associated Fungal Communities. *Environ Microbiol*, 16(9):2848-2858.

Teixeira MC, Godinho CP, Cabrito TR, Mira NP, Sá-Correia I (2012). Increased Expression of the Yeast Multidrug Resistance Abc Transporter Pdr18 Leads to Increased Ethanol Tolerance and Ethanol Production in High Gravity Alcoholic Fermentation. *Microb Cell Fact*, 11:98.

This P, Lacombe T, Thomas M (2006). Historical Origins and Genetic Diversity of Wine Grapes. *Trends in Genetics*, 22(9):511-519.

Thornton K (2003). Libsequence: A C++ Class Library for Evolutionary Genetic Analysis. *Bioinformatics*, 19(17):2325-2327.

Tominaga M, Lee YH, Hayashi R, Suzuki Y, Yamada O, Sakamoto K, Gotoh K, Akita O (2006). Molecular Analysis of an Inactive Aflatoxin Biosynthesis Gene Cluster in *Aspergillus oryzae* Rib Strains. *Appl Environ Microbiol*, 72(1):484-490.

Török T, Mortimer RK, Romano P, Suzzi G, Polsinelli M (1996). Quest for Wine Yeasts - an Old Story Revisited. *Journal of Industrial Microbiology*, 17:303-313.

Treu L, Toniolo C, Nadai C, Sardu A, Giacomini A, Corich V, Campanaro S (2014). The Impact of Genomic Variability on Gene Expression in Environmental *Saccharomyces cerevisiae* Strains. *Environ Microbiol*, 16(5):1378-1397.

Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008). Population Genomics of the Wild Yeast *Saccharomyces paradoxus*: Quantifying the Life Cycle. *Proc Natl Acad Sci U S A*, 105(12):4957-4962.

Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ (2010). The Role of Nucleosome Positioning in the Evolution of Gene Regulation. *PLoS Biol*, 8(7):e1000414.

Valamoti SM, Mangafa M, Koukouli-Chrysanthaki C, Malamidou D (2007). Grape-Pressings From Northern Greece: The Earliest Wine in the Aegean? *Antiquity*, 81(311):54-61.

Valamoti SM (2015). Harvesting the 'Wild'? Exploring the Context of Fruit and Nut Exploitation At Neolithic Dikili Tash, With Special Reference to Wine. *Veget Hist Archaeobot*, 24(1):35-46.

Valles BS, Bedriñana RP, Tascón NF, Simón AQ, Madrera RR (2007). Yeast Species Associated With the Spontaneous Fermentation of Cider. *Food Microbiol*, 24(1):25-31.

van Hoek MJ, Hogeweg P (2009). Metabolic Adaptation After Whole Genome Duplication. *Mol Biol Evol*, 26(11):2441-2453.

Vaughan-Martini A, Martini A (1995). Facts, Myths and Legends on the Prime Industrial Microorganism. *J Ind Microbiol*, 14(6):514-522.

Vaughan-Martini A, Martini A (2011). *Saccharomyces* Meyen ex Reess (1870). In: Kurtzman CP, Fell JW, Boekhout T, editors. The Yeasts: A Taxonomic Study. Elsevier. p. 733-746.

Voordeckers K, Kominek J, Das A, Espinosa-Cantú A, De Maeyer D, Arslan A, Van Pee M, van der Zande E, Meert W, Yang Y, Zhu B, Marchal K, DeLuna A, Van Noort V, Jelier R, Verstrepen KJ (2015). Adaptation to High Ethanol Reveals Complex Evolutionary Pathways. *PLoS Genet*, 11:e1005635.

Wang GD, Xie HB, Peng MS, Irwin D, Zhang YP (2014). Domestication Genomics: Evidence From Animals. *Annu Rev Anim Biosci*, 2:65-84.

Wang QM, Liu WQ, Liti G, Wang SA, Bai FY (2012). Surprisingly Diverged Populations of

Saccharomyces cerevisiae in Natural Environments Remote from Human Activity. *Mol Ecol*, 21(22):5404-5417.

Wang SA, Bai FY (2008). *Saccharomyces arboricolus* Sp. Nov., a Yeast Species from Tree Bark. *Int J Syst Evol Microbiol*, 58(Pt 2):510-514.

Warringer J, Zörgö E, Cubillos FA, Zia A, Gjuvsland A, Simpson JT, Forsmark A, Durbin R, Omholt SW, Louis EJ, Liti G, Moses A, Blomberg A (2011). Trait Variation in Yeast is Defined By Population History. *PLoS Genet*, 7(6):e1002111.

Wenger JW, Schwartz K, Sherlock G (2010). Bulk Segregant Analysis by High-Throughput Sequencing Reveals a Novel Xylose Utilization Gene From *Saccharomyces cerevisiae*. *PLoS Genet*, 6(5):e1000942.

Williams KM, Liu P, Fay JC (2015). Evolution of Ecological Dominance of Yeast Species in High-Sugar Environments. *Evolution*, 69(8):2079-2093.

Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, Hartl DL (2003). Genetic Diversity in Yeast Assessed With Whole-Genome Oligonucleotide Arrays. *Genetics*, 163(1):79-89.

Wisecaver JH, Rokas A (2015). Fungal Metabolic Gene Clusters-Caravans Traveling Across Genomes and Environments. *Front Microbiol*, 6:161.

Womack JE (2005). Advances in Livestock Genomics: Opening the Barn Door. *Genome Res*, 15(12):1699-1705.

Yona AH, Manor YS, Herbst RH, Romano GH, Mitchell A, Kupiec M, Pilpel Y, Dahan O (2012). Chromosomal Duplication is a Transient Evolutionary Solution to Stress. *Proc Natl Acad Sci U S A*, 109(51):21010-21015.

Yuasa N, Nakagawa Y, Hayakawa M, Iimura Y (2004). Distribution of the Sulfite Resistance Gene Ssu1-R and the Variation in Its Promoter Region in Wine Yeasts. *J Biosci Bioeng*, 98(5):394-397.

Zeder MA (2015). Core Questions in Domestication Research. *Proc Natl Acad Sci U S A*, 112(11):3191-3198.

Zerbino DR, Birney E (2008). Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. *Genome Res*, 18(5):821-829.

Zhang H, Skelton A, Gardner RC, Goddard MR (2010). *Saccharomyces paradoxus* and *Saccharomyces cerevisiae* Reside on Oak Trees in New Zealand: Evidence for Migration From Europe and Interspecies Hybrids. *FEMS Yeast Res*, 10(7):941-947.

Zimmer A, Durand C, Loira N, Durrens P, Sherman DJ, Marullo P (2014). Qtl Dissection of Lag Phase in Wine Fermentation Reveals a New Translocation Responsible for *Saccharomyces cerevisia*e Adaptation to Sulfite. *PLoS One*, 9(1):e86298.

APPENDIX I
Supporting Information for Chapter 2

Table AI.1 | List of strains used in Chapter 2 and relevant information pertaining to them. The approximate location for the strains collected in South America (SA location) is illustrated in Figure 2.1. The restricted dataset was used to reconstruct the phylogeny in Figure 2.1. Strains belonging to the Large Dataset were also used for the phylogeny in Figure AI.1.

Species	Strain	Population*	Substrate of isolation	Geographic location	SA location	Dataset	Other strain designations
S. uvarum	CRUB 1776	SA-A	Bark of *Nothofagus obliqua*	Hua-Hum, San Martín de los Andes, Neuquén, Patagonia, Argentina	B	Restricted	yHCT 74
S. uvarum	yHCT 77	HOL	Soil next to *Quercus garryana*	Benton County, Oregon, USA		Large	FM 1277
S. uvarum	yHCT 78	HOL	Bark of *Quercus acutissima*	Chaumette Vineyard, Ste. Genevieve, Missouri, USA		Restricted	FM 1258
S. uvarum	CRUB 1984	SA-B	*Cyttaria hariotii* on *Nothofagus dombeyi*	Colonia Suiza, Río Negro, Patagonia, Argentina,	G	Restricted	SR10-M1, yHCT 118
S. uvarum	CRUB 1987	SA-B	*Bark* of *Nothofagus betuloides*	Perito Moreno glacier, Calafate, Santa Cruz, Patagonia, Argentina	L	Restricted	B15-1, yHCT 121
S. uvarum	CRUB 1988	SA-B	*Bark* of *Nothofagus pumilio*	Usuahia, Tierra del Fuego, Patagonia, Argentina	M	Large	B22-1, yHCT 122
S. uvarum	CRUB 1989	SA-B	*Bark* of *Nothofagus dombeyi*	Chubut, Arrayanes river, Patagonia, Argentina	J	Restricted	B26-1, yHCT 123
S. uvarum	CRUB 1990	SA-A	*Bark* of *Nothofagus dombeyi*	Villa Pehuenia, Neuquén, Patagonia, Argentina	A	Restricted	B5-2, yHCT 124
S. uvarum	CRUB 1991	SA-B	*Cyttaria hariotii* on *Nothofagus betuloides*	Martial glacier, Usuahia, Tierra del Fuego, Patagonia, Argentina	N	Restricted	C43, yHCT 125
S. uvarum	CRUB 1993	SA-A	Soil next to *Nothofagus nervosa*	Hua-Hum, San Martín de los Andes, Neuquén, Patagonia, Argentina	B	Large	yHCT 127
S. uvarum	CRUB 1994	SA-A	Soil under Nothofagus *dombeyi*	Huemul arm, Neuquén, Argentina, Patagonia	D	Restricted	S8-4, yHCT 128
S. uvarum	CRUB 1595	SA-A	*Cyttaria* sp. on *Nothofagus dombeyi*	Rucahuenu, Tronador area, Rio Negro, Patagonia, Argentina	I	Restricted	FM 1317, 4-1, 4-4R10
S. uvarum	CRUB 1780	SA-B	Bark of *Nothofagus alpina*	Yuco, San Martín de los Andes, Neuquén, Patagonia, Argentina	C	Restricted	CR10-9
S. uvarum	CRUB 1778	SA-B	Bark of *Nothofagus pumillio*	Perito Moreno National Park, Santa Cruz, Patagonia, Argentina	K	Restricted	CR10-15
S. uvarum	CRUB 1779	SA-B	Bark of *Nothofagus antarctica*	Frias lake, Blest, Rio Negro, Patagonia, Argentina	F	Restricted	CR10-25
S. uvarum	CRUB 1783	SA-B	Bark of *Nothofagus antarctica*	Frias lake, Blest, Rio Negro, Patagonia, Argentina	F	Restricted	CR30-23A
S. uvarum	CRUB 1784	SA-B	Bark of *Nothofagus dombeyi*	Frias lake, Blest, Rio Negro, Patagonia, Argentina	F	Large	CR30-24B
S. uvarum	CRUB 1781	SA-A	*Cyttaria* sp. on *Nothofagus pumillio*	Cerro Otto, Bariloche, Rio Negro, Patagonia, Argentina	E	Large	CR10-A1
S. uvarum	CRUB 1782	SA-A	*Cyttaria* sp. on *Nothofagus dombeyi*	Tronador area, Los Rapidos, Rio Negro, Patagonia, Argentina	H	Restricted	CR10-H1

Table AI.1 (continued)

Species	Strain	Population*	Substrate of isolation	Geographic location	SA location	Dataset	Other strain designations
S. uvarum	CRUB 1586	SA-A	*Cyttaria* sp. on *Nothofagus dombeyi*	Rucahuenu, Tronador area, Rio Negro, Patagonia, Argentina	I	Restricted	2-1, 4-2R10
S. uvarum	CRUB 1572	SA-B	*Cyttaria* sp. on *Nothofagus dombeyi*	Rucahuenu, Tronador area, Rio Negro, Patagonia, Argentina	I	Restricted	3a, 2-3R10
S. uvarum	CRUB 1566	SA-A	*Cyttaria* sp. on *Nothofagus dombeyi*	Tronador area, Los Rapidos, Rio Negro, Patagonia, Argentina	H	Restricted	1a, 1-1R10
S. uvarum	CRUB 1777	SA-B	Bark of *Nothofagus obliqua*	Hua-Hum, San Martín de los Andes, Neuquén, Patagonia, Argentina	B	Large	B9-3, yHAB 488
S. uvarum	ZP 555	HOL	*Bark* of *Quercus garryana*	Hornby Island, Canada		Restricted	
S. uvarum	ZP 556	HOL	Bark of *Quercus garryana*	Hornby Island, Canada		Restricted	
S. uvarum	ZP 604	HOL	Bark of *Quercus garryana*	Hornby Island, Canada		Restricted	
S. uvarum	ZP 646	HOL	Cider	Tübingen, Germany		Restricted	
S. uvarum	ZP 664	HOL	Bark of *Quercus robur*	Bochum, Germany		Large	
S. uvarum	ZP 830	HOL	Bark of *Quercus glauca*	Takamatsu-shi, Kagawa Perfecture, Japan		Restricted	
S. uvarum	ZP 853	HOL	Bark of *Fagus sylvatica*	Bochum, Germany		Restricted	
S. uvarum	ZP 959	AUST	*Cyttaria gunni* on *Nothofagus menziesii*	Lewis Pass, New Zealand		Restricted	
S. uvarum	ZP 962	AUST	Bark of *Nothofagus cunninghamii*	Mount Field National Park, Tasmania		Restricted	
S. uvarum	ZP 963	AUST	Bark of *Nothofagus cunninghamii*	Mount Field National Park, Tasmania		Restricted	
S. uvarum	ZP 964	AUST	Bark of *Nothofagus solandri var. solandri*	Lewis Pass, Klondyke valley route, New Zealand		Restricted	
S. uvarum	ZP 966	AUST	Bark of *Nothofagus solandri var. solandri*	Lewis Pass, Klondyke valley route, New Zealand		Restricted	
S. uvarum	ZP 994	AUST	*Bark of Nothofagus solandri var. solandri*	Lewis Pass, Klondyke valley route, New Zealand		Restricted	
S. uvarum	ZP 1021	HOL	Soil next to *Castanea sativa*	Mêda, Guarda, Portugal		Restricted	
S. uvarum	ZP 1047	HOL	Soil next to orange tree	Mêda, Guarda, Portugal		Restricted	
S. uvarum	NCAIM 00789	HOL	Exudate of *Carpinus betulus*	Babat, Hungary		Restricted	

Table AI.1 (continued)

Species	**Strain**	**Population***	**Substrate of isolation**	**Geographic location**	**SA location**	**Dataset**	**Other strain designations**
S. uvarum	NCAIM 00868	HOL	Slimy material on a stump	Dorog, Hungary		Restricted	
S. uvarum	NCAIM 01116	HOL	Grape berries	Russia		Restricted	
S. uvarum	DBVPG 7786	HOL	Wine	Czech Republic		Restricted	
S. uvarum	DBVPG 7787	HOL	Wine	Slovakia		Restricted	
S. uvarum	CBS 377	HOL	Pear Wine	Germany		Restricted	
S. uvarum	CBS 395[T]	HOL	Juice of *Ribes nigrum* (Black currant)	Netherlands		Restricted	
S. uvarum	CBS 7001	HOL	*Mesophylax adopersus*	Cadiz, Spain		Restricted	
S. uvarum	UCD 51-206	HOL	*Drosophila persimilis*	Porcupine flat, California		Restricted	
S. uvarum	UCD 61-137	HOL	*Drosophila pseudoobscura*	Berryessa Hills, California		Restricted	
S. uvarum	M 488	HOL	Grapes	Moldova		Restricted	
S. uvarum	PJS 9	HOL	Wine/ fermenting grape	Sancerre, France		Restricted	
S. uvarum	PJS 2-95	HOL	Fermented Must	Sancerre, France		Restricted	
S. uvarum	RC 4-15	HOL	Wine/ fermenting grape	Alsace, France		Restricted	
S. uvarum	148.01	HOL	Exudate of *Ulmus pumila*	Blagoveshchensk, Far East, Russia		Restricted	
S. uvarum	PM 12	HOL	Wine/Botrytized grape	Southwest of France		Restricted	
S. uvarum	BR 6-2	HOL	Cider/Fruit juice	Brittany, France		Restricted	
S. uvarum	PJP 3	HOL	Wine/Fermenting Grape	Sancerre, France		Restricted	
S. uvarum	GM 14	HOL	Fermenting grape must	Jurançon, France		Restricted	
triple hybrid†	CBS 2834	HOL	Fendant starter	Wadenswill, Switzerland		Restricted	
triple hybrid†	CID 1	HOL	Cider yeast isolate from a home brewery	Brittany, France		Restricted	CBS 8614

Table AI.1 (continued)

Species	Strain	Population*	Substrate of isolation	Geographic location	SA location	Dataset	Other strain designations
double hybrid‡	S6U	HOL	Wine	Italy		Restricted	CBS 8615
S. bayanus	NCAIM 00676	HOL	Fermented drink	Hungary		Restricted	
S. eubayanus	CRUB 1568T		*Cyttaria* sp. on *Nothofagus dombeyi*	Rucahuenu, Tronador area, Rio Negro, Patagonia, Argentina	Q	Restricted	FM 1318, 2a, 2-2R10

* AUST, Australasia; EU, European; FE, Far East; HOL, Holarctic; SA, South America. T Type strains

† Triple hydrid *S. uvarum* x *S. cerevisiae* x *S. kudriavzevii*

‡ Double hydrid *S. uvarum* x *S. cerevisiae*

Table AI.2 | Carbon source assimilation and maximum growth temperature for *Saccharomyces uvarum* strains from different populations. Ability to use carbon sources was tested at 25ºC in 5 mL of liquid Yeast Nitrogen Base (YNB) medium supplemented with 1% (w/v) of the carbon source tested. The maximum temperature growth assay was performed in 5 mL of liquid YM medium (0.3% (w/v) yeast extract, 0.3% (w/v)

		Carbon Source*			Temperature*							
Population	**Strain**	**Glu‡**	**Mal‡**	**Mel‡**	**25ºC**	**30ºC**	**31ºC**	**32ºC**	**33ºC**	**34ºC**	**35ºC**	**36ºC**
Australasia	ZP 959	+	D	D	+	+	W	W	-	-	-	-
Australasia	ZP 962	+	-	DW	+	+	W	-	-	-	-	-
Australasia	ZP 964	+	D	D	+	+	W	-	-	-	-	-
Australasia	ZP 965	+	D	D	+	+	W	W	-	-	-	-
Australasia	ZP 983	+	D	D	+	+	W	W	-	-	-	-
Holarctic	ZP 555	+	nd	+	+	+	+	+	+	+	W	W
Holarctic	ZP 646	+	+	D	+	+	+	+	+	+	+	-
Holarctic	ZP 663	+	nd	+	+	+	+	+	+	+	+	W
Holarctic	ZP 830	+	+	+	+	+	+	+	+	+	+	-
Holarctic	ZP 853	+	+	+	+	+	+	+	+	+	+	-
Holarctic	ZP 1021	+	+	+	+	+	+	+	+	+	+	-
Holarctic	NCAIM 868	+	+	DW	+	+	+	+	+	+	W	-
Holarctic	NCAIM 1107	+	+	+	+	+	+	+	+	+	W	-
Holarctic	DBVPG 7786	+	+	+	+	+	+	+	+	+	+	-
Holarctic	UCD 51-206	+	+	+	+	+	+	+	+	+	+	-
South America	CRUB 1595	+	nd	+	+	+	+	+	+	W	W	W
South America	CRUB 1783	+	nd	+	+	+	+	+	+	+	+	W

* +, growth ($OD_{640nm} > 2.0$) after two days; -, absence of growth after three weeks; W, weak growth ($OD_{640nm} < 0.5$ after one week); D/D^{W}, delayed growth after two or more weeks (D); nd, not determined

‡ Glu, Glucose; Mal, Maltose; Mel, Melibiose

Table AI.3 | Growth rates with glucose and melibiose as sole carbon and energy sources of an Australasian (ZP 964) and a Patagonian (CRUB 1595) strain of *S. uvarum* at different temperatures. Strains were grown in 100 mL of YNB liquid medium supplemented with 2% (w/v) glucose or melibiose.

Strain	Carbon Source	Temperature	Growth rate (h^{-1})
ZP 964	Glucose	10 ºC	0.036
		18ºC	0.157
		25ºC	0.240
	Melibiose	10ºC	0.018
		18ºC	0.048
		25ºC	0.064
CRUB 1595	Glucose	10ºC	0.037
		18ºC	0.170
		25ºC	0.270
	Melibiose	10ºC	0.044
		18ºC	0.091
		25ºC	0.114

Table AI.4 | Spore viability values from crosses of different lineages of *S. uvarum* and comparison with *S. eubayanus*. At least 200 spores (50 tetrads) were analysed for each independent cross.

Cross*	
Clade C vs Clade A and B	**Spore viability (%)**
ZP 964 (AUT) – ZP 963 (AUT)	95.7
ZP 964 (AUT) – CRUB 1595 (SA-A)	35.8
ZP 964 (AUT) – CRUB 1778 (SA-B)	27.2
ZP 964 (AUT) – CRUB 1586 (SA-A)	29.8
ZP 964 (AUT) – ZP 1021 (HO-EU)	29.6
SA-A vs SA-B populations	
CRUB 1586 (SA-A) – CRUB 1778 (SA-B)	97.2
South American vs. Holarctic (NA/EU)	
CRUB 1778 (SA-B) – ZP 555 (HO-NA)	93.7
CRUB 1778 (SA-B) – ZP 1021 (HO-EU)	96.7
CRUB 1586 (SA-A) – ZP 1021 (HO-EU)	98.8
CRUB 1595 (SA-A) – ZP 1021 (HO-EU)	96.8
S. uvarum* vs. *S. eubayanus	
ZP 964 (AUT) – CRUB 1568 (*S. eubayanus*, SA)	18.8
ZP 1021 (HO-EU) – CRUB 1568 (*S. eubayanus*, SA)	15.6
CRUB 1586 (SA-A) – CRUB 1568 (*S. eubayanus*, SA)	8.5
CRUB 1595 (SA-A) – CRUB 1568 (*S. eubayanus*, SA)	7.3[§]
CRUB 1778 (SA-B) – CRUB 1568 (*S. eubayanus*, SA)	10.3

* AUT; Australasia; EU, Europe; HO, Holarctic; SA, South America; NA, North America

§ Libkind et al. 2011

Table AI.5 | Complete list of *Saccharomyces cerevisiae* and *S. kudriavzevii* introgressed genes into *S. uvarum*. Reads were mapped to a combined reference including all the annotated coding sequences of the six *Saccharomyces* species. Genes showing more than 80% of Q40 bases in more than 100 bases, which provides a good initial measure for the mappability of reads to the corresponding targets, were selected for further analysis. The number of reads mapping to the coding portion of each gene is shown for each species. Only genes with orthologs unambiguously annotated in at least four species, including S. uvarum, were analysed (see notes for exceptions). Species names are shown in three letter designations (for example, *S. uvarum* is shown as uva).

			Mapped reads									Differences to *S. kudriavzevii*	
Strain	**Gene**	**%Q40**	**uva**	**cer**	**par**	**mik**	**kud**	**arb**	**total**	**donor**	**Notes**	**IFO 1802**	**ZP 591**
148.01	YML 131W	96.35	2	0	0	0	332	0	334	kud		8	3
	YML 130C	96.74	2	0	0	0	450	0	452	kud		13	4
	YML 129C	93.20	0	0	0	0	39	0	39	kud		2	0
	YML 128C	98.43	3	0	3	0	483	0	489	kud		9	3
	YHR 209W	94.38	1	191	0	0	0	0	192	cer			
	YHR 210C	95.41	-	349	0	0	0	0	349	cer	not annotated in Suva genome		
	YPR 033C	98.35	3	486	0	0	0	0	489	cer	blast result in annotation of Suva matches Kwal_27.10565		
	YPR 034W	97.98	0	533	2	0	0	0	535	cer			
GM 14	YPR 200C	94.92	-	0	35	-	226	0	261	kud	not annotated in Suva neither in Smik genomes		
	YDR 541C	94.87	-	0	0	1	29	0	30	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	90.16	0	0	0	0	609	0	609	kud	several copies in Skud annotation		
	YML 131W	99.82	0	0	0	0	1381	0	1381	kud			
	YML 130C	99.88	2	0	0	0	2353	0	2355	kud			
	YML 129C	97.61	0	0	0	0	130	0	130	kud			
	YML 128C	99.67	0	4	0	0	1542	0	1546	kud			
	YPR 033C	99.57	0	2049	0	0	0	0	2049	cer			
	YPR 034W	99.58	4	1694	0	0	0	0	1698	cer			
M 488	YDR 541C	85.22	-	0	1	0	8	0	9	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	81.34	0	0	1	0	194	0	195	kud	several copies in Skud annotation		
	YML 131W	98.45	0	0	1	0	518	0	519	kud		8	3
	YML 130C	98.29	1	0	0	0	666	0	667	kud		13	4
	YML 129C	87.86	0	0	0	0	49	0	49	kud		2	0
	YML 128C	98.50	0	0	1	0	688	0	689	kud		9	3
NCAIM 676	YPR 200C	82.82	-	0	18	-	113	0	131	kud	not annotated in Suva neither in Smik genomes		
	YDR 541C	89.47	-	0	0	0	13	0	13	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	85.43	83	0	0	0	291	0	374	kud	several copies in Skud annotation		
	YML 131W	99.09	626	0	1	0	637	0	1264	kud	prossibly with 2 copies, one from Suva and other from Skud		
	YML 130C	99.53	1050	3	0	0	1062	0	2115	kud	prossibly with 2 copies, one from Suva and other from Skud		
	YML 129C	94.74	73	0	0	0	63	0	136	kud	prossibly with 2 copies, one from Suva and other from Skud		
	YML 128C	98.76	715	0	0	0	735	0	1450	kud	prossibly with 2 copies, one from Suva and other from Skud		

Table AI.5 (continued)

Strain	Gene	%Q40	uva	cer	par	mik	kud	arb	total	Possible donor	Notes	IFO 1802	ZP 591
NCAIM	YPR 200C	81.77	-	1	14	-	60	0	75	kud	not annotated in Suva neither in Smik genomes		
00868	YDR 541C	86.96	-	3	2	0	15	0	20	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	81.74	0	3	10	0	250	0	263	kud	several copies in Skud annotation		
	YML 131W	98.72	0	1	1	0	428	0	430	kud			
	YML 130C	98.99	0	6	3	0	1034	0	1043	kud			
	YML 129C	93.81	0	6	0	0	62	0	68	kud			
	YML 128C	98.63	0	3	2	0	600	0	605	kud			
PJS 2-95	YML 131W	94.89	0	0	0	0	156	0	156	kud		8	3
	YML 130C	96.04	0	0	0	0	270	0	270	kud		13	4
	YML 129C	87.08	0	0	0	0	29	0	29	kud		2	0
	YML 128C	95.92	0	0	0	0	217	0	217	kud		9	3
PJS 9	YPR 200C	80.74	-	0	12	-	59	0	71	kud	not annotated in Suva neither in Smik genomes		
	YML 131W	97.35	0	0	0	0	359	0	359	kud		8	3
	YML 130C	99.11	0	0	0	0	449	0	449	kud		13	4
	YML 129C	88.04	0	0	0	0	42	0	42	kud		2	0
	YML 128C	98.76	0	0	0	0	466	0	466	kud		9	3
ZP 646	YML 129C	85.51	0	0	0	0	27	0	27	kud		2	0
	YML 128C	96.87	0	0	0	0	202	0	202	kud		9	3
	YHR 178W	91.06	105	239	0	0	0	0	344	cer	prossibly with 2 copies, one from Suva and other from Scer		
	YHR 179W	81.92	174	124	0	-	0	0	298	cer	prossibly with 2 copies, one from Suva and other from Scer		
	YHR 187W	80.30	0	84	0	0	0	0	84	cer			
	YHR 188C	94.90	2	229	0	0	0	0	231	cer			
	YHR 189W	90.14	0	116	0	0	0	0	116	cer			
	YHR 195W	85.59	1	99	0	0	0	0	100	cer			
	YHR 200W	85.34	0	120	0	0	0	0	120	cer			
	YHR 202W	88.39	0	144	0	0	0	0	144	cer			
	YHR 206W	90.89	0	252	0	0	0	0	252	cer			
	YHR 208W	93.15	0	148	0	0	0	0	148	cer			
ZP 830	YDR 541C	82.30	-	0	1	0	6	0	7	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	80.35	0	0	0	0	222	0	222	kud	several copies in Skud annotation		
	YML 131W	98.81	0	0	0	0	517	0	517	kud		8	3
	YML 130C	99.23	0	0	0	0	626	0	626	kud		13	4
	YML 129C	96.19	0	0	0	0	75	0	75	kud		2	0
	YML 128C	98.43	0	0	0	0	552	0	552	kud		9	3
ZP 853	YPR 200C	81.53	-	0	14	0	86	0	100	kud	not annotated in Suva neither in Smik genomes		

Table AI.5 (continued)

Strain	Gene	%Q40	uva	cer	par	mik	kud	arb	total	Possible donor	Notes	IFO 1802	ZP 591
ZP 853	YDR 541C	90.60	-	0	0	0	13	0	13	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	88.02	0	0	0	0	335	0	335	kud	several copies in Skud annotation		
	YML 131W	99.54	0	0	0	0	900	0	900	kud		8	3
	YML 130C	99.29	0	0	0	0	1210	0	1210	kud		13	4
	YML 129C	93.78	0	0	0	0	92	0	92	kud		2	0
	YML 128C	99.61	0	0	0	0	1540	0	1540	kud		9	3
BR 6-2	YDR 541C	81.58	-	4	0	1	38	0	43	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	87.82	0	0	0	0	404	0	404	kud	several copies in Skud annotation		
	YML 131W	98.27	0	0	0	0	989	0	989	kud			
	YML 130C	98.52	0	2	0	0	1673	0	1675	kud			
	YML 129C	97.14	0	0	0	0	142	0	142	kud			
	YML 128C	98.37	0	1	0	0	1417	0	1418	kud			
PJP 3	YDR 541C	91.23	-	2	0	0	43	0	45	kud	not annotated in Suva genome		
	YHL 044W	88.18	0	8	0	0	439	0	447	kud	several copies in Skud annotation		
	YML 131W	97.90	0	12	0	0	1029	0	1041	kud			
	YML 130C	98.64	0	19	0	0	1687	1	1707	kud			
	YML 129C	96.65	0	1	0	0	132	0	133	kud			
	YML 128C	98.11	0	19	0	0	1438	0	1457	kud			
PM 12	YDR 541C	83.93	-	15	0	2	21	0	38	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	86.21	0	9	0	0	286	0	295	kud	several copies in Skud annotation		
	YML 131W	97.72	0	17	0	0	643	0	660	kud			
	YML 130C	97.34	0	20	0	0	1030	0	1050	kud			
	YML 129C	94.74	0	3	0	0	125	0	128	kud			
	YML 128C	97.92	0	16	0	0	1143	0	1159	kud			
RC 4-15	YPR 200C	90.45	-	9	76	-	178	0	263	kud	not annotated in Suva neither in Smik genomes		
	YDR 541C	84.35	-	6	0	0	51	0	57	kud	not annotated in Suva genome, 2 copies in Skud annotation		
	YHL 044W	89.46	0	3	0	0	566	0	569	kud	several copies in Skud annotation		
	YML 131W	98.36	0	1	0	0	1333	0	1334	kud			
	YML 130C	98.58	0	6	0	0	2188	0	2194	kud			
	YML 129C	97.61	0	1	0	0	191	0	192	kud			
	YML 128C	98.57	0	14	0	0	1971	0	1985	kud			

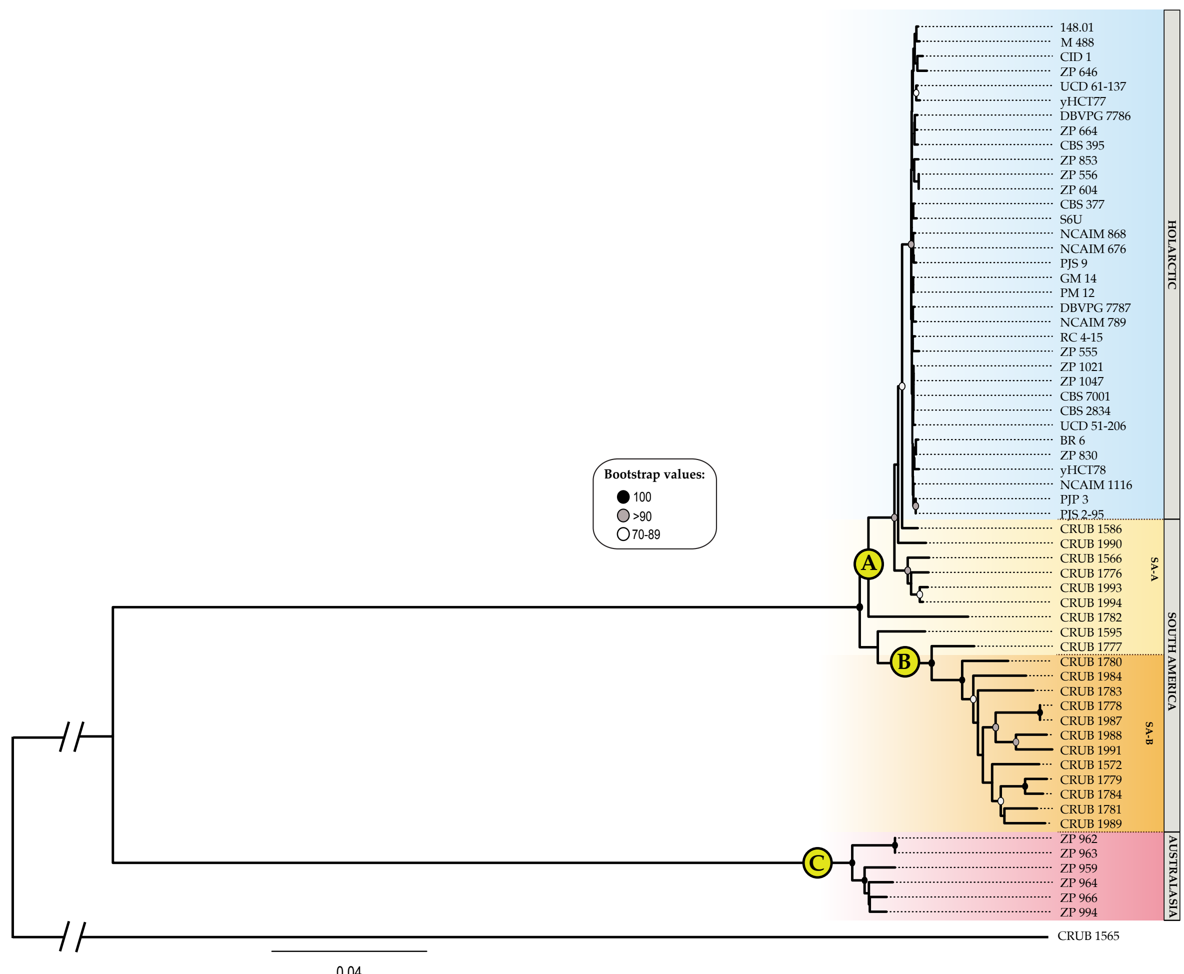


Figure AI.1 | Phylogeny of the complete set of *Saccharomyces uvarum* strains used in this study. Neighbor-Joining phylogeny of 61 strains (large strain dataset, see Table S1) inferred with p-distance based on 3113 high quality SNPs found in the concatenated alignment of chromosomes 5, 8 and 10 (note that chromosome 10 was mislabeled as chromosome 12, and vice-versa, in the annotation of Scannell et al. 2011). The three main clades are marked by letters A, B and C as in Figure 1D. Support values were calculated from 1000 bootstrap replicates. Branch lengths correspond to the mean number of base differences per site. The tree was rooted with *S. eubayanus*.

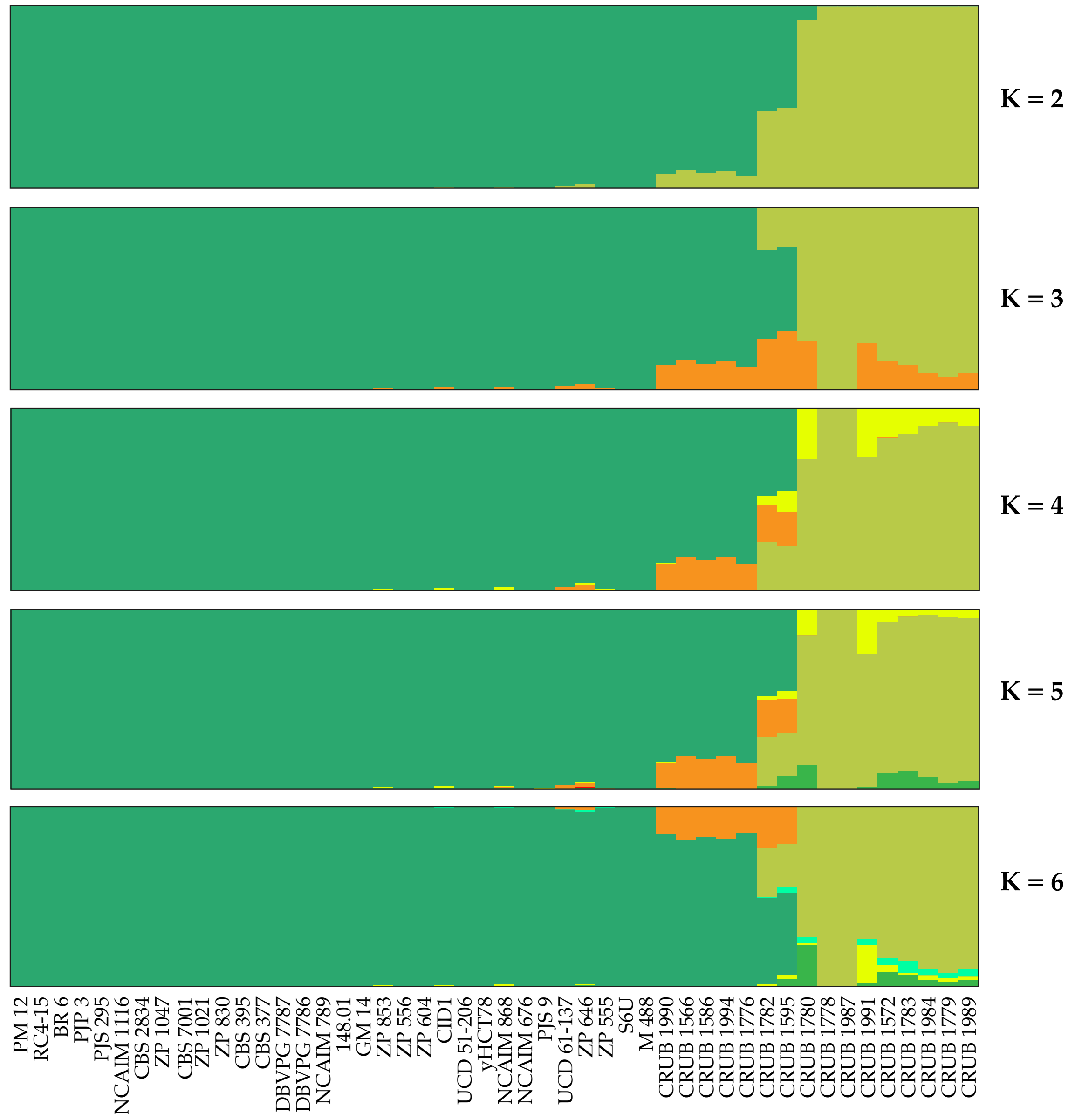


Figure AI.2 | Genetic clusters of *S. uvarum* for K=2 to K=6 inferred with STRUCTURE and considering all but the Australasian sequences.

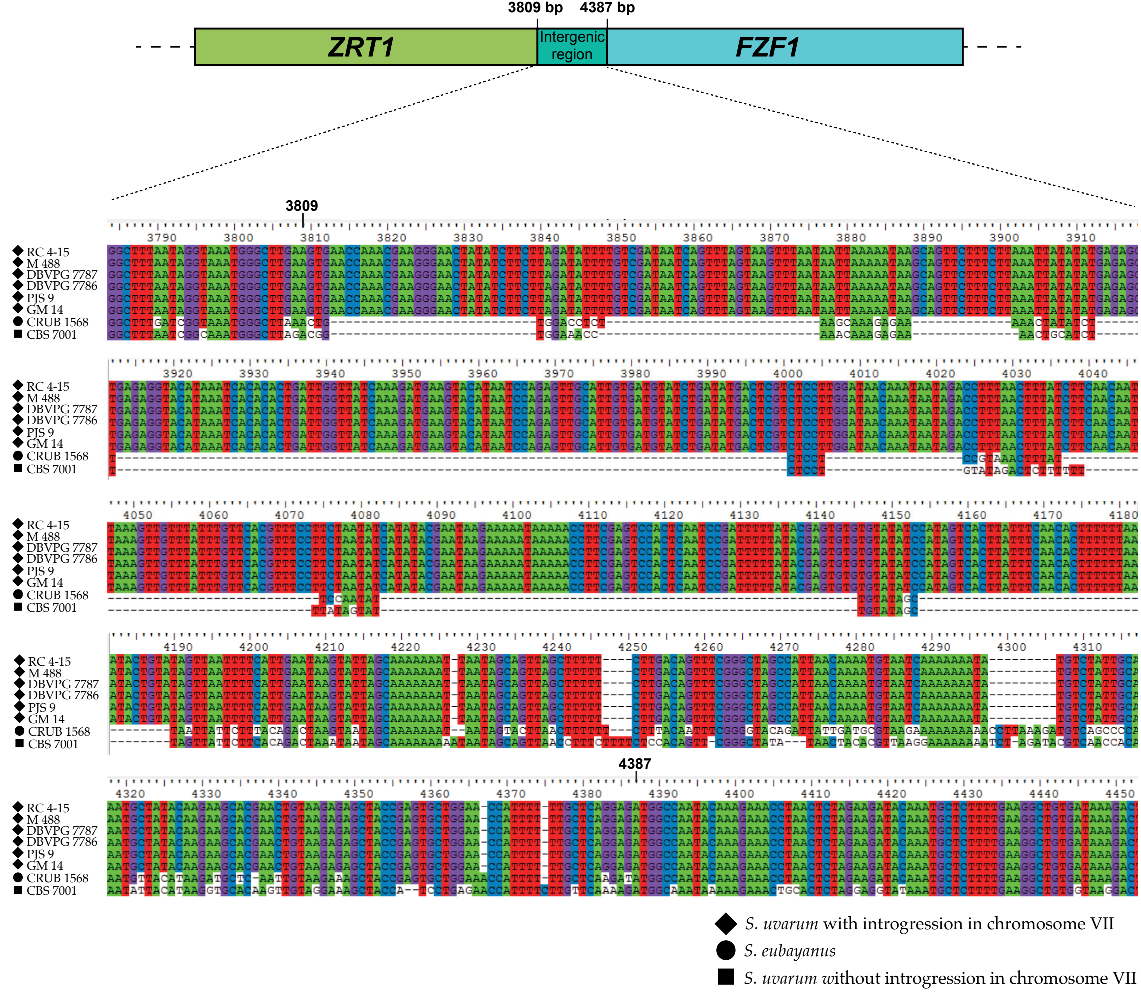


Figure AI.3 | 5′ Alignment of the promoter region of *FZF1* in *S. uvarum* and *S. eubayanus*. The upstream region of *FZF1* gene is very similar in all introgressed strains (diamond) but is strikingly different from e *S. eubayanus* (circle) and non-introgressed *S. uvarum* (square).

APPENDIX II

Supporting Information for Chapter 3

Table AII.1 | List of *Saccharomyces cerevisiae* strains used in Chapter 3 and relevant information pertaining to them. Colour codes indicate substrate of isolation: red, fermentation; green, oak; orange, fruit; and grey, other or unkwon. The restricted dataset was used to reconstruct the phylogeny in Figure 3.4. Strains belonging to the Large Dataset were used for population structure analyses, network and for the phylogeny in Figure 3.5. Strains marked with a ‡ symbol are additional strains used for within population diversity calculations in Table 3.3 and Table 3.4.

Strain	Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
DBVPG6765	-	-	Restricted		SGRP2	[3]
YJM1336	Wine	Italy	Restricted	M28s2	PRJNA189897	[9]
EXF6761	*Quercus petraea*	Cerkvenjak, Slovenia	Restricted		PRJEB7601	This study
M22	Vineyard	Italy	Restricted		PRJNA28815	[7]
YJM1415	Wine	France	Restricted	NRRL Y-268	PRJNA189914	[9]
L-1374	Wine (must from País variety)	Cauquenes, Chile	Restricted		SGRP2	[3]
L-1528	Wine (must from Cabernet variety)	Cauquenes, Chile	Restricted		SGRP2	[3]
YJM1242	Fruit	Illinois, USA	Restricted	NRRL Y-35	PRJNA189886	[9]
DBVPG1788	Soil	Finland	Restricted		SGRP2	[3]
DBVPG1106	Grapes	Australia	Restricted	NCYC3447	SGRP2	[3]
YJM1341	Grape must	South Africa	Restricted	NRRL Y-12637	PRJNA189899	[9]
ZP578	*Quercus faginea*	Aldeia das Dez, Portugal	Restricted		PRJEB7601	This study
ZP577	*Quercus faginea*	Aldeia das Dez, Portugal	Restricted		PRJEB7601	This study
ZP579	*Quercus pyrenaica*	Aldeia das Dez, Portugal	Restricted		PRJEB7601	This study
CLIB215	Baker's yeast	New Zealand	Restricted		PRJNA60143	
ZP862	Apple with insect holes	Caratão, Abrantes, Portugal	Restricted		PRJEB7601	This study
ZP861	Sun-dryed apple with ants	Caratão, Abrantes, Portugal	Restricted		PRJEB7601	This study
EXF6719	Wine must	Cuber, Slovenia	Restricted		PRJEB7601	This study
ZP641	Spontaneous red wine fermentation	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
DAV1e	*Quercus robur*	Davenport Vineyards, Southern UK	Restricted		PRJNA264372	This study
I14	Vineyard soil	Petrina, Italy	Large		PRJNA60149	
YJM1433	Wine	Sauternes, France	Large	YIIc17-E5	PRJNA189918	[9]
YJM1332	Wine	Italy	Restricted	M1-2	PRJNA189896	[9]
VL3	Commercial wine yeast	-	Restricted	B6	PRJNA48565	[5]
VelF	*Ficus carica*	Velanidias, Southern Greece	Restricted		PRJNA264372	This study
AWRI796	Commercial wine yeast	South Africa	Restricted	Active Dry Wine Yeast	PRJNA48559	[5]

Table AII.1 (continued)

Strain		Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
WE372		Commercial wine yeast	Cape Town, South Africa	Restricted		PRJNA60199	
ZP1041		Vineyard soil	Meda, Guarda, Portugal	Restricted		PRJNA264372	This study
RM11-1a		Vineyard	California, USA	Restricted	Bb32	PRJNA13674	
BC187		Barrel fermentation	Napa Valley, California, USA	Restricted	UC10, UCD2120	SGRP2	[3]
YJM975		Clinical	Italy	Restricted		SGRP2	[3]
YJM978		Clinical	Italy	Restricted		SGRP2	[3]
DBVPG1373		Soil	Netherlands	Restricted		SGRP2	[3]
AWRI1631		Wine	Australia	Restricted	N96	PRJNA30553	[4]
Vin13		Commercial wine yeast	-	Restricted		PRJNA48563	[5]
YJM270		Wine	Slovenia	Restricted	CBS2807	PRJNA189852	[9]
EC1118		Industrial; isolated from Champagne	France	Restricted	Prise de Mousse, 59A	PRJEA37863	[8]
IOC9002		Commercial wine yeast	-	Restricted		PRJNA264372	This study
LalvinQA23		Commercial wine yeast	Portugal	Restricted		PRJNA48561	[5]
YJM1574		Wine	-	Restricted	AWRI1775	PRJNA189934	[9]
CBS7960		Industril; ethanol from cane-sugar	Sao Paulo, Brazil	Restricted	DBVPG7960	PRJNA60391	
T73		Red wine from Monastrel grapes	Alicante, Spain	Large		PRJNA60195	
JAY291		Bioethanol-production strain	-	Large		PRJNA32809	[2]
YJM1383		Coconut	Philippines	Large	NRRL Y-5511	PRJNA189904	[9]
YJM1463		White Tecc	Ethiopia	Large	DBVPG1853	PRJNA189926	[9]
YJM271		Brewery	-	Large	CBS1782	PRJNA189853	[9]
EXF7145		*Quercus robur*	Fruska Gora National Park, Novi Sad, Serbia	Large		PRJNA264372	This study
PYR4e		*Quercus pubescens*	Agios Ioannis, Pyrgadikia, Northern Greece	Restricted		PRJNA264372	This study
DBQ26		*Quercus pubescens*	Halkidiki, Greece	Restricted	PYR4b1s1	PRJEB7601	This study
OakRom3-2a		Oak	near Bucarest, Romania	Restricted		PRJEB7675	This study
HUN9.1s1		Oak (isolated by Eladio Barrio)	Hungary	Restricted	DBS14	PRJEB7601	This study
EXF7200		*Quercus robur*	Jasenovo polje, Montenegro	Restricted		PRJNA264372	This study
EXF7197		*Quercus robur*	Jasenovo polje, Montenegro	Restricted		PRJNA264372	This study
FGXd		*Ficus carica*	Hotel Anixi, Anixi, Southern Greece	Restricted		PRJNA264372	This study

Table AII.1 (continued)

Strain	Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
DBVPG10100	*Quercus cerris*	Parco del Monte Subasio, Italy	Restricted	3.10	PRJEB7601	This study
ZP541	*Fagus sylvatica*	Adagoi, Portugal	Restricted		PRJEB7601	This study
ZP1008	*Quercus faginea*	Eja, Melres, Douro, Portugal	Restricted		PRJNA264372	This study
ZP742	Rotten figs	Caratão, Abrantes, Portugal	Restricted		PRJEB7601	This study
ZP736	Rotten figs	Caratão, Abrantes, Portugal	Restricted		PRJEB7601	This study
ZP851	*Quercus ilex*	Alconorales Natural Park, Andaluzia, Spain	Restricted		PRJEB7675	This study
MB7c	*Quercus pubescens*	Montbarri, Southern France	Restricted		PRJNA264372	This study
ZP860	*Quercus ilex*	Alconorales Natural Park, Andaluzia, Spain	Restricted		PRJEB7601	This study
ZP850	*Quercus ilex*	Alconorales Natural Park, Andaluzia, Spain	Restricted		PRJEB7601	This study
ZP562	*Quercus ilex*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP570	*Fraxinus* sp.	Paul Boquilobo, Portugal	Restricted		PRJEB7601	This study
ZP848	*Quercus ilex*	Alter do Chão, Portugal	Restricted		PRJEB7675	This study
ZP633	*Quercus pyrenaica*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP636	*Castanea sativa*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP560	*Quercus pyrenaica*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP568	*Quecus pyrenaica*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP565	*Castanea sativa*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP563	*Castanea sativa*	Castelo de Vide, Portugal	Restricted		PRJEB7601	This study
ZP561	*Quercus pyrenaica*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
ZP564	*Quercus pyrenaica*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
ZP566	*Quercus faginea*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
ZP567	*Quercus pyrenaica*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
ZP575	*Quercus pyrenaica*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
ZP590	*Quercus pyrenaica*	Castelo de Vide, Portugal	‡		PRJEB7601	This study
FostersB	Commercial brewing strain (ale)	-	Restricted		PRJNA48569	[5]
FostersO	Commercial brewing strain (ale)	-	Restricted		PRJNA48567	[5]
YJM1387	Wine	Japan	Large	NRRL Y-12758	PRJNA189907	[9]
SON4c	*Ficus carica*	Sonim, Portugal	Large		PRJNA264372	This study

Table AII.1 (continued)

Strain		Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
ZP611		*Quercus robur*	Vancouver, Canada	Large		PRJEB7675	This study
DBQ22		*Ficus carica*	Halkidiki, Greece	Large	AN3e1s1	PRJEB7601	This study
S288c		Laboratory strain	California, USA	Large		Reference	
W303		Laboratory strain	-	Large		SGRP2	[3]
CLIB324		Baker's yeast	Saigon, Vietnam	Large		PRJNA60415	
YJM1386		Sugar cane	Jamaica	Large	NRRL Y-11878	PRJNA189906	[9]
YJM1381		Rum fermentation	Trinidad	Large	NRRL YB-427	PRJNA189903	[9]
EC9-8		Soil	Evolution Canyon (valley bottom), Israel	Large		PRJNA73985	[6]
SDO3s1		Oak	North Carolina, USA	Restricted		PRJEB7601	This study
SDO6s1		Oak	North Carolina, USA	Restricted		PRJEB7601	This study
SDO2s1		Oak	North Carolina, USA	Restricted		PRJEB7601	This study
SDO8s1		Oak	North Carolina, USA	Restricted		PRJEB7601	This study
SM69s1		Oak	North Carolina, USA	Restricted		PRJEB7601	This study
YPS1009		Oak exudate	New Jersey, USA	Restricted		PRJNA60223	
YJM1402		Grape vine slime flux	-	Restricted	NRRL YB-4449	PRJNA189913	[9]
ZP1050		*Quercus ilex*	Vendinha, Reguengos de Monsaraz, Portugal	Restricted		PRJEB7675	This study
ZP530		*Castanea sativa*	Marão, Campeã, Portugal	Restricted		PRJEB7601	This study
EXF6780		*Quercus petraea*	Velike Lašče, Kobila hill, Slovenia	Large		PRJEB7601	This study
OakArd11-2-2		Oak	Ardéche, Lyon, France	Large		PRJEB7675	This study
OakBor21-1		Oak	Bordeaux, France	Large		PRJEB7675	This study
T7		Oak exudate	Babler State Park, Missouri, USA	Restricted	T7 Fay	PRJNA60387	
UWOPS83-787.3		Fruit of *Opuntia stricta*	Bahamas	Restricted		SGRP2	[3]
YJM1338		Sour fig	Maryland, USA	Large	NRRL Y-963	PRJNA189898	[9]
ZP656		*Quercus acuta*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study
ZP653		*Quercus acutissima*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study
ZP680		*Quercus acutissima*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study
ZP651		*Quercus acutissima*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study
ZP652		*Quercus acutissima*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study

Table AII.1 (continued)

Strain		Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
ZP649		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP650		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP654		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP655		*Quercus acuta*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP657		*Quercus acuta*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP658		*Quercus acuta*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP674		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP676		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP677		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
ZP678		*Quercus acutissima*	Chiba Perfecture, Japan	‡		PRJEB7601	This study
YJM1273		Oak	Pennsylvania, USA	Restricted	YPS134	PRJNA189891	[9]
YPS128		*Quercus alba*	Pennsylvania, USA	Restricted		SGRP2	[3]
YJM1434		Oak	Pennsylvania, USA	Restricted	YPS606	PRJNA189919	[9]
YPS163		*Quercus rubra*	Pennsylvania, USA	Restricted		PRJNA28813	[7]
ZP675		*Quercus acutissima*	Chiba Perfecture, Japan	Restricted		PRJEB7601	This study
ZP779		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP785		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP780		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJNA264372	This study
ZP817		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP781		*Quercus serrata*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP793		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP796		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP786		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP823		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	Restricted		PRJEB7601	This study
ZP778		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	‡		PRJEB7601	This study
ZP783		*Quercus acutissima*	Hirusen highland, Okayama Perfecture, Japan	‡		PRJEB7601	This study
ZP784		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	‡		PRJEB7601	This study
ZP794		*Quercus dentata*	Hirusen highland, Okayama Perfecture, Japan	‡		PRJEB7601	This study

Table AII.1 (continued)

Strain		Substrate of isolation	Geographic location	Dataset	Other designations	Genome data	Reference
YPS1000		Oak exudate	New Jersey, USA	Restricted		PRJNA264372	This study
UWOPS87-2421		cladode of *Opuntia megacantha*	Maui, Hawaii	Large		SGRP2	[3]
OakGri7.1		Oak	Grinon, Paris, France	Large		PRJEB7675	This study
YJM1342		Soil	South Africa	Large	NRRL Y-12638	PRJNA189900	[9]
YJSH1		Bioethanol-producing strain	China	Restricted		PRJNA72403	[10]
YJM1388		Fermented tapioca	Malaysia	Restricted	NRRL Y-12769	PRJNA189908	[9]
Kyokai-no.7		Japanese sake brewerie	Japan	Restricted	K7	PRJNA45827	[1]
UC5		Sene sake	Kurashi, Japan	Restricted	UCD612	PRJNA60197	
Y12		Fermentation (palm wine)	Ivory Coast	Large	NRRL Y-12663, CBS400	SGRP2	[3]
Y9		Fermentation (ragi)	Indonesia	Restricted	NRRL Y-5997	WashU	
DBS11		*Ficus carica*	Halkidiki, Greece	Large	AN1f2s1	PRJEB7601	This study
DBS12		*Ficus carica*	Halkidiki, Greece	Large	AN1g2s1	PRJEB7601	This study
Y10		Coconut water	Philippines	Large	NRRL Y-7567	PRJNA60201	
YJM269		Fermentation (apple juice)	Unknown	Large		PRJNA60389	
YJM1479		Coconut tuba (palm wine)	Philippines	Restricted	NRRL Y-6297	PRJNA189929	[9]
YJM1400		Guava (fruit)	Philippines	Restricted	NRRL YB-4081	PRJNA189911	[9]
YJM1401		Papaya (fruit)	Philippines	Restricted	NRRL YB-4082	PRJNA189912	[9]
YJM1447		Bertram palm	Malaysia	Restricted	UWOPS05-227.2	PRJNA189923	[9]
UWOPS03-461.4		Nectar of Bertram palm	Malaysia	Restricted		SGRP2	[3]
YJM1418		*Quercus variabilis* bark	Asia (country unknown)	Large	NRRL YB-4506	PRJNA189916	[9]
Y55		Laboratory strain	France	Large	YJM627 (seg. Y55)	SGRP2	[3]
SK1		Laboratory strain	USA	Restricted		SGRP2	[3]
YJM195		Palm wine	Nigeria	Restricted	NCYC762	PRJNA189849	[9]
PW5		Raphia palm wine	Aba, Abia state, Nigeria	Restricted		PRJNA60181	
YJM1439		Ginger beer from *Z. officinale*	West Africa	Restricted		PRJNA189920	[9]
YJM1248		Bili wine from *Osbeckia grandiflora*	West Africa	Restricted	NRRL Y-1546	PRJNA189888	[9]
DBVPG6044		Bili wine from *Osbeckia grandiflora*	West Africa	Restricted	Y5, NRRL Y-1546, CBS405	SGRP2	[3]

Genomic data provenance referenced in Table AII.1:

WashU - Washington University (http://www.genetics.wustl.edu/jflab/data4.html)

SGRP2 - The Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/pub/users/dmc/yeast/SGRP2)

References for Table AII.1:

[1] Akao T et al. (2011) Whole-Genome Sequencing of Sake Yeast Saccharomyces cerevisiae Kyokai no. 7. DNA Res. 18 (6), 423-434.

[2] Argueso JL et al. (2009) Genome structure of a Saccharomyces cerevisiae strain widely used in bioethanol production. Genome Res. 19 (12), 2258-2270.

[3] Bergström A et al. (2014) A high-definition view of functional genetic variation from natural yeast genomes. Mol Biol Evol 31:872–888.

[4] Borneman AR et al. (2008) Comparative genome analysis of a Saccharomyces cerevisiae wine strain. FEMS Yeast Res. 8 (7), 1185-1195.

[5] Borneman AR et al. (2011) Whole-Genome Comparison Reveals Novel Genetic Elements That Characterize the Genome of Industrial Strains of Saccharomyces cerevisiae. PLoS Genet. 7 (2), E1001287.

[6] Chang SL et al. (2013) Dynamic Large-Scale Chromosomal Rearrangements Fuel Rapid Adaptation in Yeast Populations. PLoS Genet 9(1): e1003232.

[7] Doniger SW et al. (2008) A Catalog of Neutral and Deleterious Polymorphism in Yeast. PLoS Genet 4(8): e1000183.

[8] Novo M et al. (2009) Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast Saccharomyces cerevisiae EC1118. PNAS 106 (38), 16333-16338.

[9] Strope PK 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 Res. 25(5):762-74

[10] Zheng DQ et al. (2012) Genome sequencing and genetic breeding of a bioethanol Saccharomyces cerevisiae strain YJS329. BMC Genomics 13 (1), 479.

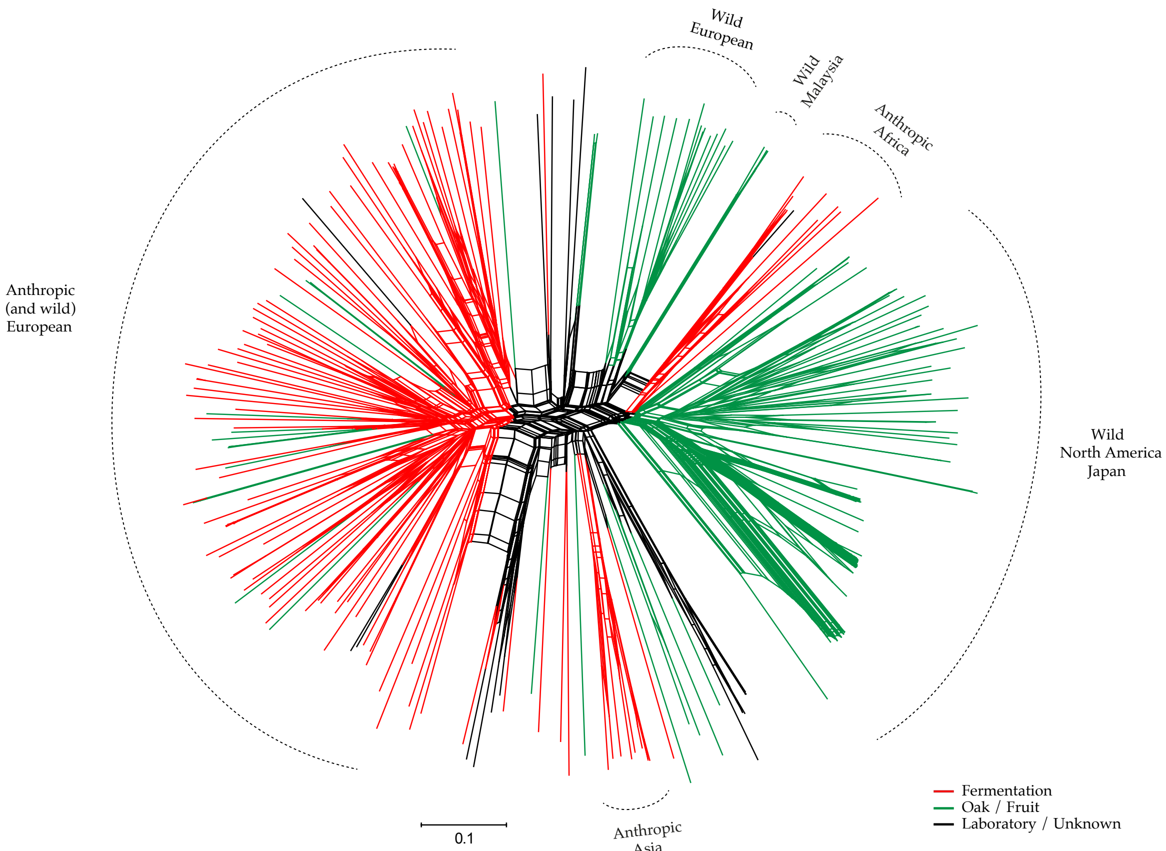


Figure AII.1 | Network analysis showing ecological and geographic separation of *Saccharomyces cerevisiae* lineages. Neigbour-net network inferred from the allelic variation at 12 microsatellite loci using the Bruvo distance. Branches are colored according to the source of strain isolation.

APPENDIX III

Supporting Information for Chapter 4

Table AIII.1 | List of strains used in Chapter 4 and relevant information pertaining to them. Detailed information about strain origin is depicted in Table AII.1. Wine strains highlighted in bold were only used in the analyses presented in Chapter 4. PdM, Prize de Mousse; MO, Mediterranean oaks.

Strain	Substrate	Geographic location	Genome data	Genotype
AWRI1631	Wine	Australia	PRJNA30553	Wine
EC1118	Industrial, isolated from Champagne	France	PRJEA37863	PdM
EXF6719	Wine must	Slovenia	PRJEB7601	Wine
IOC9002	Commercial wine yeast	-	SRP059414	PdM
L-1374	Wine (must from País variety)	Chile	SGRP2	Wine
L-1528	Wine (must from Cabernet variety)	Chile	SGRP2	Wine
YJM1332	Wine	Italy	PRJNA189896	Wine
YJM1336	Wine	Italy	PRJNA189897	Wine
YJM1341	Grape must	South Africa	PRJNA189899	Wine
YJM1415	Wine	France	PRJNA189914	Wine
YJM1574	Wine	-	PRJNA189934	PdM
YJM270	Wine	Slovenia	PRJNA189852	PdM
ZP641	Spontaneous red wine fermentation	Portugal	PRJEB7601	Wine
Zymaflore F15	Commercial yeast (red wine)	-	This study	Wine
IOC 18-2007	Commercial yeast	-	This study	PdM
Lalvin BM45	Commercial yeast (white wine)	-	This study	Wine
Lalvin CY-3079	Commercial yeast (red wine)	-	This study	Wine
Lalvin W15	Commercial yeast (white and red wine)	-	This study	Wine
PR	-	-	This study	Wine
PYCC 4072	commercial wine yeast (Fermivin)	Portugal	This study	Wine
PYCC 4074	commercial wine yeast (Fermichamp)	Portugal	This study	PdM
TUM V1	Bordeaux wine	-	This study	Wine
Uvaferm SGV	Commercial yeast (red wine)	-	This study	Wine
Uvaferm VRB	Commercial yeast (red wine)	-	This study	Wine
PYCC 6722	wine	South Armenia	This study	PdM
PYCC 6726	Jerez-wine	Spain	This study	PdM
PYCC 6729	Jerez-wine	Armenia	This study	PdM
AWRI796	Commercial wine yeast	South Africa	PRJNA48559	~ Lalvin W15
LalvinQA23	Commercial wine yeast	Portugal	PRJNA48561	PdM*
VL3	Commercial wine yeast	-	PRJNA48565	Wine*
Vin13	Commercial wine yeast	-	PRJNA48563	Wine*
WE372	Commercial wine yeast	South Africa	PRJNA60199	Wine*
DBQ26	*Quercus pubescens*	Greece	This study	MO
OakRom3-2a	Oak	Romania	This study	MO
HUN9.1s1	Oak (isolated by Eladio Barrio)	Hungary	This study	MO
EXF7200	*Quercus robur*	Montenegro	This study	MO
FGXd	*Ficus carica*	Greece	This study	MO
DBVPG10100	*Quercus cerris*	Italy	This study	MO
ZP541	*Fagus sylvatica*	Portugal	This study	MO
ZP742	Rotten figs	Portugal	This study	MO
ZP851	*Quercus ilex*	Spain	This study	MO
ZP860	*Quercus ilex*	Spain	This study	MO
ZP562	*Quercus ilex*	Portugal	This study	MO
ZP570	*Fraxinus* sp.	Portugal	This study	MO

Table AIII.1 (continued)

Strain	Substrate	Geographic location	Genome data	Genotype
ZP848	*Quercus ilex*	Portugal	This study	MO
ZP633	*Quercus pyrenaica*	Portugal	This study	MO
ZP636	*Castanea sativa*	Portugal	This study	MO
ZP560	*Quercus pyrenaica*	Portugal	This study	MO
ZP568	*Quecus pyrenaica*	Portugal	This study	MO
ZP565	*Castanea sativa*	Portugal	This study	MO
ZP563	*Castanea sativa*	Portugal	This study	MO
ZP561	*Quercus pyrenaica*	Portugal	This study	MO
ZP564	*Quercus pyrenaica*	Portugal	This study	MO
ZP566	*Quercus faginea*	Portugal	This study	MO
ZP575	*Quercus pyrenaica*	Portugal	This study	MO
ZP590	*Quercus pyrenaica*	Portugal	This study	MO
PYR4e	*Quercus pubescens*	Greece	This study	MO*
ZP1008	*Quercus faginea*	Portugal	This study	MO*
ZP567	*Quercus pyrenaica*	Portugal	This study	MO*
EXF7197	*Quercus robur*	Montenegro	This study	~ FGXd
ZP736	Rotten figs	Portugal	This study	~ ZP742
MB7c	*Quercus pubescens*	France	This study	~ ZP566
ZP850	*Quercus ilex*	Spain	This study	~ ZP860

* genomes with more than 10% of missing data

Table AIII.2 | Top 15 Yeast GO Slim Mapper results in Process category for the ORFs with two or more high differentiated (ΔAF > 0.8) SNPs in the 500 bp upstream sequence. Only GO terms with a frequency higher than the background gene set are shown.

GO term	Frequency	ORF(s)
response to chemical (GO:0042221)	7.0%	OAF1, FLO1, UGA2, KIN82, YCR102C, SIT4, CDC36, PPH3, YCF1, DSE1, ULI1, YHB1, ELM1, HRT3, CDC42, CMP2, ALO1, PDR18, CIN5, CRS5, RDR1
meiotic cell cycle (GO:0051321)	7.0%	GIP1, FMP45, PPH3, MSH6, BMH2, MAM1, SPR6, MSH4, SPR3, DSN1, VPS13, SMC4, SPO77, RIM9, PDS5, FKS3, SPS19, MSH2, RTS1, GAC1, TGS1
cellular response to DNA damage stimulus (GO:0006974)	5.0%	FUN30, SIT4, LYS20, PPH3, UBC13, MSH6, BMH2, RSC8, SPT4, RAD2, REV7, SML1, PDS5, VPS75, MSH2
protein complex biogenesis (GO:0070271)	5.0%	ISW1, UBC13, SLY1, TCA17, OXA1, CNN1, COX13, SHY1, DAM1, ELM1, POM33, COX14, VPS75, OST3, RPT5
cell wall organization or biogenesis (GO:0071554)	5.0%	GIP1, SIT4, PSA1, FMP45, BMH2, MNN10, BIT61, BCH2, SPO77, DFG5, FKS3, PKH2, HPF1, MKK1, FLC1
organelle fission (GO:0048285)	4.7%	PPH3, MSH6, MAM1, MSH4, DSN1, EBP2, SMC4, CDC42, PDS5, PSE1, MSH2, RTS1, GAC1, TGS1
DNA repair (GO:0006281)	4.4%	FUN30, SIT4, LYS20, PPH3, UBC13, MSH6, RSC8, SPT4, RAD2, REV7, PDS5, VPS75, MSH2
signalling (GO:0023052)	4.4%	SIT4, PPH3, BMH2, DSE1, ULI1, BIT61, TPK3, CDC42, CMP2, SRV2, PKH2, RTC1, MKK1
carbohydrate metabolic process (GO:0005975)	3.7%	ADH5, BMH2, MNN1, PCL6, FLO8, HXK2, ELM1, BCH2, PFK2, PFK27, GAC1
pseudohyphal growth (GO:0007124)	3.7%	BMH2, TMN2, ITR1, FLO8, OSH3, PHD1, ELM1, DBR1, CDC42, DFG5, DIA1
regulation of cell cycle (GO:0051726)	3.4%	CDC36, PPH3, DAM1, ECM27, CDC42, ZDS1, PSE1, CDC33, RTS1, GAC1
generation of precursor metabolites and energy (GO:0006091)	3.4%	ADH5, BMH2, PCL6, COX13, HXK2, MDH1, SHH4, PFK2, PFK27, GAC1
chromosome segregation (GO:0007059)	3.4%	AME1, MAM1, CNN1, SPT4, DAM1, DSN1, SMC4, PDS5, RTS1, GAC1
sporulation (GO:0043934)	3.4%	GIP1, FMP45, BMH2, SPR6, SPR3, VPS13, SPO77, RIM9, FKS3, SPS19
ion transport (GO:0006811)	3.4%	YDL183C, YCF1, ATP17, ATO3, JEN1, DIC1, PFK2, FET4, SMF1, PUT4

Table AIII.3 | Top 15 Yeast GO Slim Mapper results in Process category for the ORFs with two or more high differentiated (ΔAF > 0.8) replacement SNPs. Only GO terms with a frequency higher than the background gene set are shown.

GO term	Frequency	ORF(s)
transcription from RNA polymerase II promoter (GO:0006366)	12.2%	CCR4, GPB2, TAF2, CDC39, PRR2, RAD9, HDA2, SUM1, GLN3, FLO8, SPT2, RPH1, SNT2, SUT1, FZF1, STE12, ZAP1, SWI3, PUT3, PHD1, MSN4, ABF1, NUP133, ACE2, RFX1, RSC2, LEU3, TAF13, CRZ1, FKH2, RTT106, CAF120, YNG1, YRR1, PIP2, GCR1, HDA3
response to chemical (GO:0042221)	11.6%	GPB2, FLR1, FIG2, CDC39, PRR2, VMS1, PPH3, AFR1, PDR15, GLN3, PTP3, SNT2, SUT1, VMR1, STE12, BCK1, SSY5, SWI3, MSN4, SPH1, DCR2, STE11, ECM5, SKY1, CRZ1, MLF3, YCK2, YNR064C, PDR18, RGA1, PDR5, YRR1, PIP2, RDR1, GCR1
chromatin organization (GO:0006325)	8.9%	TEL1, SPT7, SWC5, SNT1, NGG1, HDA2, SUM1, EAF1, SPT2, RPH1, NUP145, SNT2, SWI3, MSN4, ABF1, TOF2, IRS4, UTH1, NUP133, RSC2, FKH2, RTT106, YNG1, NPT1, GCR1, DPB2, HDA3
signaling (GO:0023052)	8.6%	GPB2, SNT1, SSK22, AHK1, PPH3, AFR1, HKR1, PTP3, BEM2, RGD2, MDS3, NPR3, KOG1, BCK1, TOR1, PEX1, IRS4, GIS3, DCR2, STE11, CRZ1, WSC3, RTC1, RGA1, PTP2, BEM3
mitotic cell cycle (GO:0000278)	7.6%	CCR4, LTE1, BUD3, MBP1, RAD9, SUM1, HKR1, YBP2, SAP4, IRR1, PAN1, BUD4, HSL1, RAX2, ACE2, SPH1, DCR2, HOF1, PDS5, FKH2, BNI1, GAC1, RFC1
meiotic cell cycle (GO:0051321)	6.3%	GPB2, SPS22, PPH3, ECM11, MAM1, SHE10, ZIP2, NPR3, RIM4, SPO22, IME2, TOR1, VPS13, MSC1, PDS5, FKS3, POL1, GAC1, SMA1
cellular response to DNA damage stimulus (GO:0006974)	5.6%	TEL1, PPH3, RAD9, EAF1, YEN1, NUP145, SRS2, TOR1, ABF1, NUP133, RSC2, OGG1, HOF1, PDS5, POL1, RFC1, DPB2
protein targeting (GO:0006605)	5.6%	SYO1, AST2, COG3, BLM10, NUP145, VPS45, MON1, PEX21, NUP100, PEX1, NUP133, VPS13, SKY1, ATG2, BOR1, NUP1, SRP68
carbohydrate metabolic process (GO:0005975)	5.3%	NTH2, ROT2, MAL33, RBK1, HKR1, FLO8, SAK1, AMS1, GUT1, SWI3, RBH2, CRZ1, GAC1, TYE7, GCR1, GDB1
protein phosphorylation (GO:0006468)	5.3%	TEL1, SNT1, SSK22, AHK1, PRR2, PTP3, SAK1, BCK1, IME2, TOR1, HSL1, KIN2, STE11, SKY1, YCK2, PTP2
DNA repair (GO:0006281)	5%	TEL1, PPH3, RAD9, EAF1, YEN1, NUP145, SRS2, ABF1, NUP133, RSC2, OGG1, PDS5, POL1, RFC1, DPB2
conjugation (GO:0000746)	5%	SPT7, BUD3, FIG2, CDC39, PRR2, AFR1, PTP3, SUT1, STE12, PRM2, SPH1, STE11, STE23, SCW10, RGA1
protein complex biogenesis (GO:0070271)	5%	STU1, SPT7, SED4, USO1, EAF1, NUP157, BLM10, VPS45, DAM1, TRS65, UFD4, RTT106, BNI1, SEC12, SEC16
organelle fission (GO:0048285)	5%	LTE1, PPH3, SUM1, ECM11, MAM1, ZIP2, NPR3, RIM4, IRR1, SPO22, IME2, TOR1, MSC1, PDS5, GAC1
regulation of cell cycle (GO:0051726)	4.6%	CCR4, LTE1, CDC39, PPH3, RAD9, DAM1, SPO22, IME2, TOR1, ECM27, HSL1, DCR2, FKH2, GAC1

Table AIII.4 | Yeast GO Slim Mapper results in Process category for the ORFs within the candidate regions. Only GO terms with a frequency higher than the background gene set are shown.

GO term	Frequency	ORF(s)
biological process unknown (GO:0008150)	38.1%	YDL057W, TOS8, RRT15, YMR321C, YNR068C, BSC5, YNR071C, ZPS1
response to chemical (GO:0042221)	9.5%	FLO1, PDR18
transcription from RNA polymerase II promoter (GO:0006366)	9.5%	MBP1, FZF1
protein targeting (GO:0006605)	9.5%	VPS45, MAS1
ion transport (GO:0006811)	9.5%	FZF1, FET4
regulation of transport (GO:0051049)	9.5%	VPS45, FZF1
cell wall organization or biogenesis (GO:0071554)	4.8%	HPF1
cellular respiration (GO:0045333)	4.8%	SHH4
vacuole organization (GO:0007033)	4.8%	VPS45
generation of precursor metabolites and energy (GO:0006091)	4.8%	SHH4
regulation of organelle organization (GO:0033043)	4.8%	VPS45
lipid transport (GO:0006869)	4.8%	SWH1
transmembrane transport (GO:0055085)	4.8%	FET4
vesicle organization (GO:0016050)	4.8%	VPS45
carbohydrate transport (GO:0008643)	4.8%	HXT17
cellular amino acid metabolic process (GO:0006520)	4.8%	GDH3
endosomal transport (GO:0016197)	4.8%	VPS45
protein maturation (GO:0051604)	4.8%	MAS1
membrane invagination (GO:0010324)	4.8%	SWH1
Golgi vesicle transport (GO:0048193)	4.8%	VPS45
nucleobase-containing small molecule metabolic process (GO:0055086)	4.8%	SHH4

Table AIII.4 (continued)

GO term	Frequency	ORF(s)
response to oxidative stress (GO:0006979)	4.8%	FLO1
endocytosis (GO:0006897)	4.8%	SWH1
response to heat (GO:0009408)	4.8%	FLO1
organelle inheritance (GO:0048308)	4.8%	VPS45
cellular ion homeostasis (GO:0006873)	4.8%	VPS45
exocytosis (GO:0006887)	4.8%	SWH1
organelle fusion (GO:0048284)	4.8%	VPS45
membrane fusion (GO:0061025)	4.8%	VPS45
other	6.3%	FLO9

Table AIII.5 | Yeast GO Slim Mapper results in Function category for the ORFs within the candidate regions. Only GO terms with a frequency higher than the background gene set are shown.

GO term	Frequency	ORF(s)
molecular function unknown (GO:0003674)	33.3%	YDL057W, RRT15, YMR321C, YNR068C, BSC5, YNR071C, ZPS1
DNA binding (GO:0003677)	14.3%	MBP1, TOS8, FZF1
transmembrane transporter activity (GO:0022857)	14.3%	FET4, PDR18, HXT17
hydrolase activity (GO:0016787)	14.3%	MAS1, PDR18, HPF1
nucleic acid binding transcription factor activity (GO:0001071)	9.5%	MBP1, FZF1
oxidoreductase activity (GO:0016491)	9.5%	GDH3, SHH4
chromatin binding (GO:0003682)	4.8%	TOS8
hydrolase activity, acting on glycosyl bonds (GO:0016798)	4.8%	HPF1
peptidase activity (GO:0008233)	4.8%	MAS1
unfolded protein binding (GO:0051082)	4.8%	VPS45
lipid binding (GO:0008289)	4.8%	SWH1
ion binding (GO:0043167)	4.8%	SWH1
ATPase activity (GO:0016887)	4.8%	PDR18
other	9.5%	FLO9, FLO1

DIGITAL RESOURCES

Dataset D1 | Complete list of *Saccharomyces eubayanus* introgressed genes into *S. uvarum*. This dataset is available with the digital copy of this thesis. It is also available as open access with the supplementary information of the publication on Nature Communications with the results of Chapter 2, which can be found directly as Dataset S2 on http://www.nature.com/articles/ncomms5044#supplementary-information.

Dataset D2 | Significant Gene Ontology (GO) terms for the introgressed *S. eubayanus* genes found in *S. uvarum*. This dataset is available with the digital copy of this thesis. It is also available as open access with the supplementary information of the publication on Nature Communications with the results of Chapter 2, which can be found directly as Dataset S3 on http://www.nature.com/articles/ncomms5044#supplementary-information.

Dataset D3 | Complete list of introgressed genes into *Saccharomyces cerevisiae*. This dataset is available with the digital copy of this thesis.