Genotype-phenotype relationship exploration by genome-wide association studies in yeast

Jackson Peter

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

UMR7156

THÈSE présentée par :

Jackson Peter

soutenue le : **25 Septembre 2017**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/ Spécialité : Bioinformatique

Dissection de la relation génotype-phénotype par des études d'association chez *Saccharomyces cerevisiae*

THÈSE dirigée par :

Dr. SCHACHERER Joseph	Maître de Conférences, HDR, Université de Strasbourg

RAPPORTEURS :

Pr. BÄHLER Jürg	Professeur, University College London, UK
Pr. DUJON Bernard	Professeur, Institut Pierre et Marie Curie & Institut Pasteur, Membre de l'Académie de Sciences

---

AUTRES MEMBRES DU JURY :

Dr. LECOMPTE Odile	Maître de Conférences, HDR, Université de Strasbourg
Dr. WINCKER Patrick	Directeur de Recherches, Génoscope

Acknowledgments

This work was completed at the department of genetics, genomics and microbiology, UMR7156/CNRS, University of Strasbourg, under the supervision of Dr. Joseph Schacherer in the group intraspecific variation and genome evolution.

Firstly, I would like to thank all the members of the committee, Pr. Jürg Bähler, Pr. Bernard Dujon, Dr. Odile Lecompte and Dr. Patrick Wincker for accepting to evaluate this work. Special thanks go to Dr. Odile Lecompte and Pr. Bernard Dujon for already being members of my mid-thesis committee.

I would like to express my sincere gratitude to my advisor Dr. Joseph Schacherer for supporting me during these past four years. I appreciate the way you guided me, and the trust you placed in me. Trying to meet the impressive thoroughness and effectiveness you apply to yourself pushed me forward and made me grow.

I have been lucky or well inspired when I first came in this lab for some weeks in summer. I did not imagine a second what an adventure this would be. Thank you Anne for having sparked my interest in genomics during your conferences. Thank you also for always being here, for always offering your help, advices and support. Indeed, you were never really far, yet sending a mail before walking five meters to go to your office was always the way to go!

Thanks as well to all the members of the group for the good atmosphere. Jean-Seb and Téo, thanks for never failing to listen to my doubts, my worries, my anger, and above all, for enduring all stupid jokes that come to my mind when I am tired. It was good to have you around. Best of luck to finish your thesis! Elodie, best of luck to start yours! David, playing chess with you (and losing!) was a nice way to make a break. We're not done playing! Many thanks to Jing, Kelle and Christian, your presence was such an asset for the group! To the ones that I did not mention yet: Claudine, Anastasie, Marion, Cyrielle, Paul and Arnaud. Special thanks to Claudia, for your impressive patience and kindness.

I would like to thank Cécile Fairhead for the discussions we had. Whether or not they were about science, they have always been pleasant and insightful.

I would like to thank my friends for their support, for making my time out of the lab so awesome. For the fun we had at concerts we organized or attended, and for the drinks we shared (one does not get rid of yeast so easily!). To my roommates, former or current, I have always been glad when there was somebody home as I came back from the lab!

Last but not least, I heartily thank my family, especially my parents for the unconditional love and support. Thank you Jonathan, for being such a close friend as well as my brother. The little Aaron and his mom also deserve a special thought!

State of the art
Population genomics within yeast natural populations: Insights on genome evolution and phenotypic variation

Introduction

In 1998, the term population genomics was introduced for the first time to describe the generation of large-scale polymorphism data on a large number of individuals from the same species[1]. The goal of this shift from family studies to the investigation of polymorphisms in populations was to overcome the impossibility to map the genetic determinants of complex traits, and more precisely of human genetic diseases. The establishment of an exhaustive catalog of genetic variants within a species is necessary to obtain precious insights into evolution history or migration patterns. It is a crucial step to have a better view of the genomic landscape and variance of allelic diversity within and between populations. Such resources greatly help to understand how trait variation is generated and maintained within a species. Almost 20 years later, with the advent of high throughput technologies as well as the reduction of their price, population genomic studies have been completed for many model organisms such as human[2–4], *Arabidopsis thaliana*[5,6], *Caenorabditis elegans*[7], and *Drosophila melanogaster*[8]. In human, these datasets helped to find the genetic determinants of human traits, such as the implication of 163 susceptibility loci for inflammatory bowel disease[9] or human height for which 697 Single Nucleotide Polymorphisms (SNPs) have been found in 423 loci[10]. In other species, the genetic basis of industrially relevant traits were investigated, such as yield, agronomic and quality traits of the potato[11], the grain yield under water deficit for *Oryza sativa*[12], and oil composition in *Arabidopsis thaliana*[13], for example.

In the context of population genomics, yeast species and more precisely the members of the Saccharomycotina subphylum constitute powerful model organisms[14]. Indeed, yeasts harbor small genomes (usually inferior to 20 Mb), with few introns and small intergenic regions, making the sequencing cheaper and facilitating the downstream analyses. Moreover, they are unicellular organisms that form clonal colonies that can easily be cultivated in the lab, allowing multiple phenotyping experiments. Up to date, resequencing projects have therefore been performed for many yeast species[14]. For example, *Candida albicans*, an opportunistic pathogen of human, for which molecular mechanisms underlying the evolution of drug resistance has been described at the genetic or phenotypic variation levels[15]. The fission yeast *Schizosaccharomyces pombe* has also been investigated and various signatures of selection (balancing selection and selective sweeps) have been detected[16] as well as several genotype-phenotype associations using SNPs and indels for 89 out of 223 measured traits[17]. Other examples include the investigation of elements suggesting the domestication of the cider producing yeast *S. uvarum*[18], the description of the balanced unlinked gene network polymorphism for galactose utilization[19], the differential evolution pattern of an entire chromosome arm[20], or the evolution of mitochondrial genomes within a species, namely *L. kluyveri* and *L. thermotolerans*[21,22].

Among the Saccharomycotina species, specific attention has been given to the model budding yeast *S. cerevisiae*. Here, I briefly review the main conclusions that have been drawn from the many population genomic studies of *S. cerevisiae*[23–33]. This species presents the particularity to have been associated with human activity to produce fermented food and beverages (wine

and beer) since Neolithic times[31,34,35]. This association of *S. cerevisiae* with human activity led to a partial domestication of this species. The different domestication processes shaped the genomes of the human associated isolates in various ways and consequently this species presents an interesting view of differential population evolutionary histories that have been now explored with different collections of natural isolates coming from a wide range of geographical and environmental origins. A large number of genomes (N > 1,100) have been sequenced over the last years. I will focus more precisely on the main genetic characteristics of this species, the extent of genetic variation and the differential evolution of domesticated isolates compared to feral strains of *S. cerevisiae*. I will then review what has been learnt so far on the genotype-phenotype relationship and argue that the constitution of large datasets able to capture large portions of the genetic diversity within this species is the major bottleneck in performing Genome-Wide Association Studies (GWAS) in *S. cerevisiae*.

Evolutionary history of the *Saccharomyces cerevisiae* species

Population genomic studies have provided a deep insight into the population structure and evolutionary history of *S. cerevisiae*, allowing us to reconstruct with an always-increasing precision the events that led to the adaptation to the many ecological niches it is able to live in since the origin of this species. These studies led to a better understanding of the forces that shaped the genome evolution of *S. cerevisiae* as well as the consequences of its long association with human, leading to the partial domestication of this species.

Support of a single out-of-China origin of the *S. cerevisiae* species

The origins of *S. cerevisiae* have been unclear until the recent discovery of wild Chinese lineages from primeval forest showing highly divergent genomes[28]. Further sampling in Far-East Asia confirmed the presence of highly divergent strains in this region with a Taiwanese wild lineage whose sequence divergence to non-Taiwanese strains go up to 1.1%. There is now a couple of elements that contribute to an out-of-China origin of *S. cerevisiae*. First, when a rooted phylogenetic tree of the *Saccharomyces* species complex is built, they all branch off near the Taiwanese and Chinese isolates, indicating that the most recent common ancestor of the outgroup species and *S. cerevisiae* is closer to the Chinese and Taiwanese isolates than any other strains isolated so far[36]. Principal component comparison of Far-East Asian strains compared to other isolates suggests that this out-of-China origin constitutes a single and shared event for all non-Chinese strains of *S. cerevisiae*[36]. Also, closely related species of the *Saccharomyces* genus have been isolated in Far East Asia such as *S. paradoxus*, *S. cariocanus*, *S. mikatae*, *S. arboricola*, which is coherent with the hypothesis of Far-East Asia being a reservoir of natural variation from which several species of yeast have radiated from the same common ancestor at this place[37].

Since its Chinese origin, *S. cerevisiae* has colonized a wide variety of ecological niches, and has been associated with human, therefore resulting in a radiation of genetic diversity forming multiple distinct subpopulations that constitutes the species as we know it today. The investigation of the differences between subpopulations could give us precious insights on the evolutionary forces that led to the extremely diversified habitats of *S. cerevisiae*.

Subpopulations of *S. cerevisiae*

Relationships among natural isolates can be elucidated by looking at the single nucleotide variants (SNVs) within a large number of individuals. In order to test the influence of factors such as ecology or geography on the genetic diversity, the population can be divided into groups of individuals, and proportion of ancestry of each group can be computed for each individual to represent its phylogenetic relationship with the other individuals of the population. Such genetic structure analysis[26,27,29–31] showed that *S. cerevisiae* as a whole consists of both domesticated and wild isolates, and that several independent domestication events have occurred. It was first discussed whether the genetic differentiation among strains of *S. cerevisiae* was explained by geography origins or ecological niches[23,24]. Further sampling revealed more complex pattern of genetic differentiation correlating with both geographical and ecological influence, as well as with the degree of human association[26,27,29–31,36]. Similarly, the number of populations composing the species was also debated, with five clean lineages defined in early studies (Malaysian, West African, sake, North America, Wine/European strains)[24], and up to twelve more recently[29]. However, structure and phylogeny analyses in over a thousand isolates revealed that the number of populations is obviously a matter of sampling as strains of *S. cerevisiae* are spread all over the diversity of the species[36]. Population genomic studies also revealed that some strains do not fall into clean lineages and their genomes present various levels of ancestry from other subpopulations. The genomes of these mosaic isolates correspond to an admixture of two or more clean lineages derived by outbreeding and further recombination (Figure 1). It is interesting to note that there is a significant enrichment of clinical isolates amongst mosaic strains[23,26,36]. These isolates are found in immune-deficient individuals and can cause infections, making *S. cerevisiae* an opportunistic pathogen. It is therefore likely that clinical isolates do not derive from a common ancestor or any one type of strain, but rather represent multiple events in which strains present in the environment opportunistically colonize human tissues[23]. A genomic screen of 144 strains containing 132 clinical isolates has recently been performed and showed that these strains present many large-scale genomic variations such as polyploidy level variation (3n/4n isolates) for 32% of the isolates, aneuploidies affecting 36% of the strains and identified several chromosomal events, underlining the potential importance of large-scale genomic copy variation in clinical yeast adaptation[25]. After several analyses of the human-related subpopulations, it is now clear that *S. cerevisiae* domestication happened independently several times. For example, sake and wine strains form genetically distinct groups that remained isolated from each other over time[30] or the existence of several genetically differentiated populations of beer strains, strongly suggesting multiple domestication events[31].

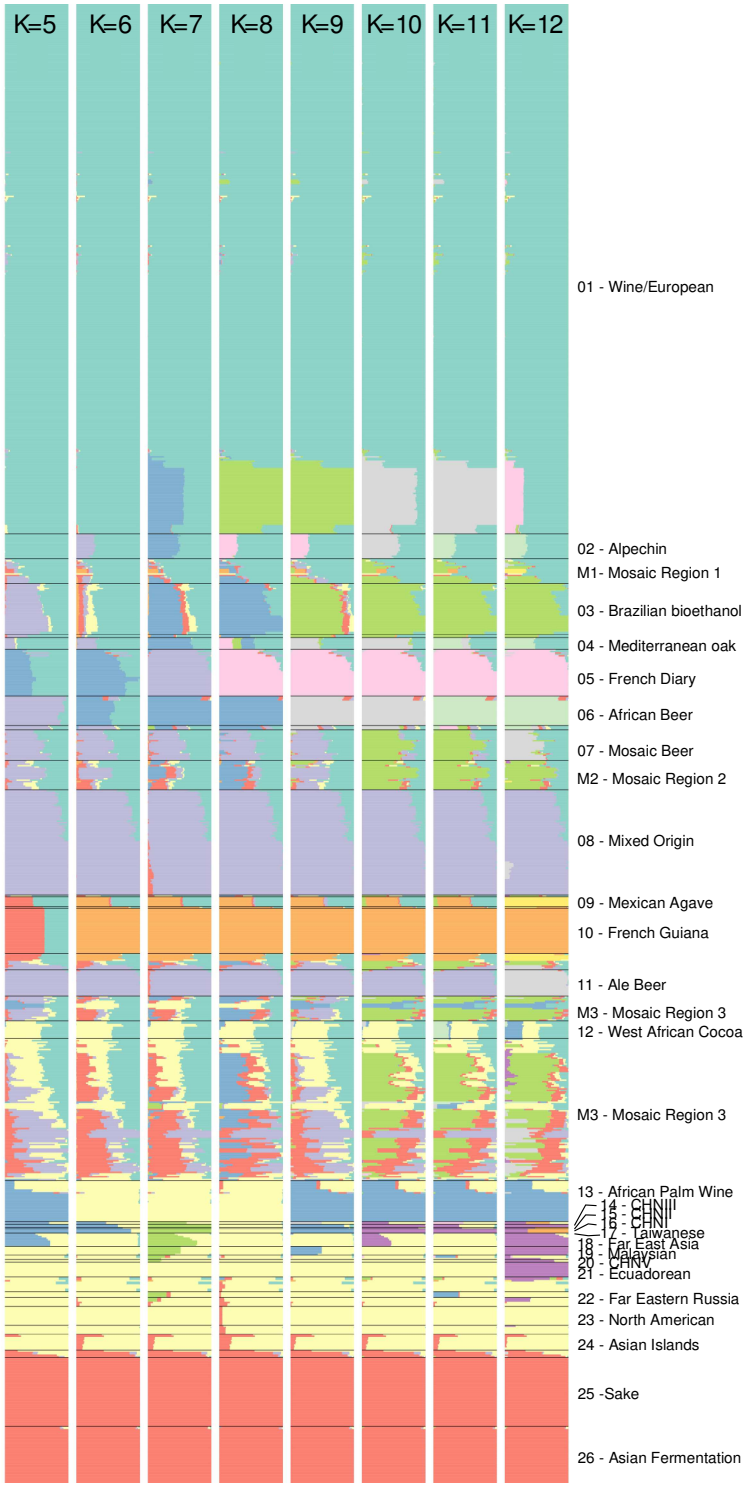


Figure 1: Population structure of 1,011 *S. cerevisiae* isolates. The underlying population structure inferred using the software ADMIXTURE using a varying number of subpopulations (from 5 to 12).

Evolutionary history of *S. cerevisiae* is punctuated by multiple domestication events

Due to its long association with human activities such as the production of fermented beverages or bread and to the fact that no natural individuals have been isolated at the time[38], the genome of *S. cerevisiae* was long thought to be only shaped by human selection. The isolation of strains from oak trees, insects or soil proved that the species is only partially domesticated and has a long history before its association with humans[24,28]. More recently, wild isolates of *S. cerevisiae* were also found in the rainforest in Brazil, suggesting a new natural habitat for this species[32]. These findings made *S. cerevisiae* a very interesting organism for population genomics, as it gives the opportunity to study the impact of domestication on this species by comparing natural isolates and domesticated strains and retrace the evolutionary histories of the populations. Indeed, human played an important role for the worldwide spreading of *S. cerevisiae* and domesticated strains show some unique characteristics.

First, as yeast species are highly inbred, with an outcrossing event occurring approximately once every 50,000 – 100,000 mitotic generations[39–41], the patterns of heterozygosity are severely impacted. A survey of genome-wide heterozygous sites over 794 natural isolates showed that entirely homozygous isolates represent approximately 40% of the strains, but there is a prevalence of heterozygous isolates in some, not all, domesticated clades[36], in agreement with a previous study[42]. Surprisingly, the computation of the average number of heterozygous sites in the different subpopulations of over more than 900 isolates shows that this number varies widely between them. For example, it is among the lowest for the wine and the sake subpopulation (with an average of 2,515 heterozygous sites by individual over the two clades), while it is very high for the three beer clades (average of 31,374 heterozygous sites by individuals over the three clades), probably due to the many polyploid strains composing these clusters[36].

Second, domesticated clades usually show a lower diversity than natural isolates, and they usually cluster together inside a monophyletic group of mostly diploid individuals, suggesting that they all derive from a single domestication event. This is the case for the wine strains being mostly diploid, having a nucleotide diversity value of $\pi = 1 \times 10^{-3}$ and all deriving from the same common ancestor, a Mediterranean oak strain[27]. This is also true for the sake isolates for which the nucleotide diversity is $\pi = 8.06 \times 10^{-4}$, forming a monophyletic group of mostly diploids strains from which a large part of them derive from the Kyokai no. 7 strain[43], being the closest known strain to the common ancestor of the clade. On the other hand, the strains used for the beer fermentation are not following this trend. It is likely that these strains are the consequence of several independent domestication events[30,36], as they are clustered into several independent groups. Moreover, the genetic diversity of beer clades is roughly 3-fold higher than the one observed in the sake and wine clusters[36]. In total, 183 beer isolates were sequenced so far. By building a neighbor-joining tree of a set of 319 strains containing the 183 beer strains publicly available as well as 136 strains covering the diversity of the species, it is

clear that top-fermenting beer yeasts are polyphyletic (Figure 2). The 'Beer1' clade is the major one and contains the previously described subpopulations of "Wheat beer", "English-Irish Ale" and "German Alt-Kölsch"[30]. Also, this tree reveals one more clade compared to previous studies, namely the "African beer" subpopulation, containing many polyploid isolates underlining once more the highly polyphyletic nature of the strains used to produce beer caused by several independent domestication events.

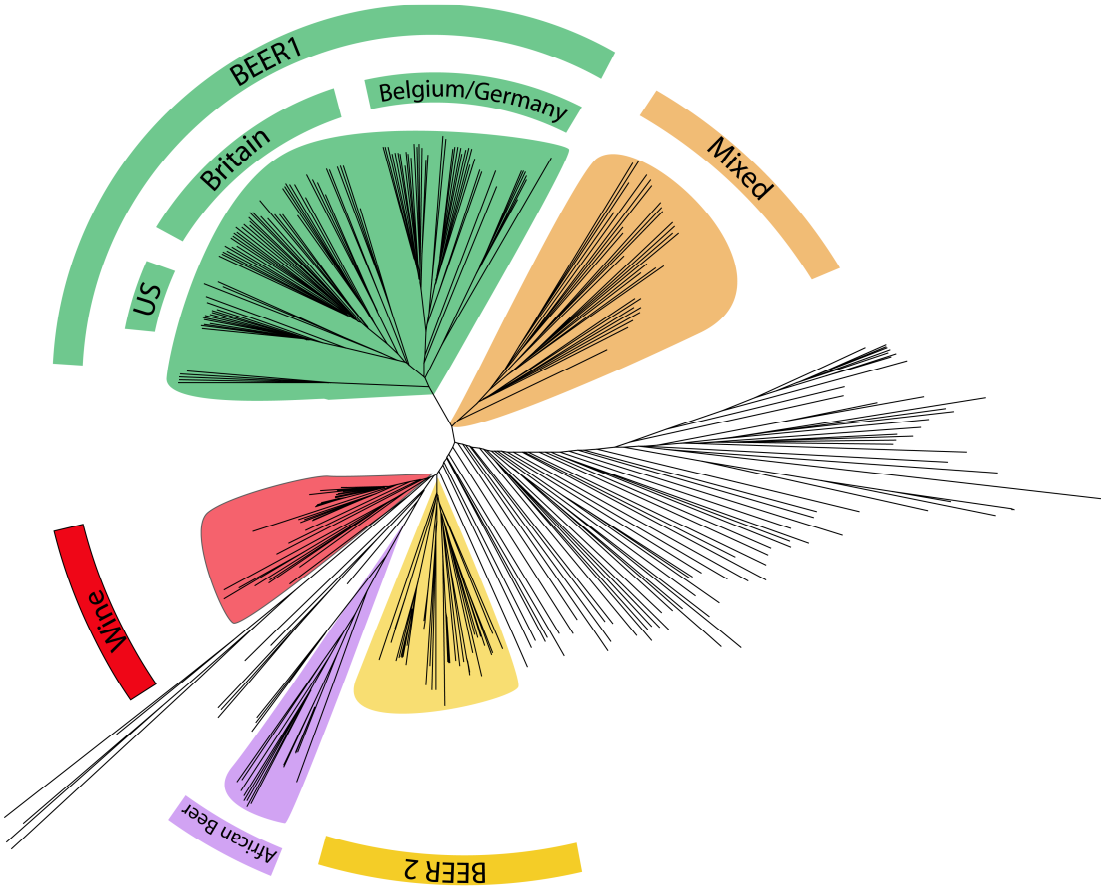


Figure 2: Neighbor-joining tree of the different beer subpopulations. Distances between strains were computed based on the biallelic SNPs

Genome evolution variation across the *S. cerevisiae* species

A catalog "raisonné" of single nucleotide variants

Resequencing studies aim at investigating the genetic diversity within a species to gain insight into the genetic characteristics and evolutionary histories of the different subpopulations. The most studied genetic variants have historically been Single Nucleotide Polymorphisms (SNPs). Indeed, these variants have historically been used as markers to identify genomic

regions associated with Quantitative Trait Loci (QTL), and are an abundant form of genetic variation as demonstrated by large-scale resequencing studies in model organisms such as human[44], *Arabidopsis thaliana*[5,6] or *Drosophila melanogaster*[8]. In *S. cerevisiae*, a total of 1,625,809 SNPs across 1,011 genomes have been identified, with a strong skew towards low-frequency variants (Figure 3a). Indeed, 31.3% of these SNPs are singletons, meaning they are present in only one individual in the population, and almost 93% of all the sites present a minor allele frequency under 5%[36]. The examination of the MAF for the different subpopulations showed different levels of skew towards low frequency polymorphisms. Indeed, while the wine subpopulation is characterized by a strong bias toward low frequency polymorphisms, resulting in an extremely negative Tajima's D value (-2.02), which is expected under the model of a selection bottleneck due to domestication, the population that contains bakery isolates displays a more uniform distribution of the MAF.

Small insertions and deletions (indels) – up to 50 bp – have also been studied, though to a much lesser extent. The results show that their locations are not random, with subtelomeric regions strongly enriched in indels, reinforcing that the subtelomeric regions are a major source for divergence of genome sequences and gene content, contributing to adaptive processes[45]. Moreover, most of these variants are not located in coding regions, and when they are, they are often located in the C-terminal region or their length tends to be a multiple of 3 base-pair, indicating a purifying selection to prevent deleterious frameshifts (Figure 3b).

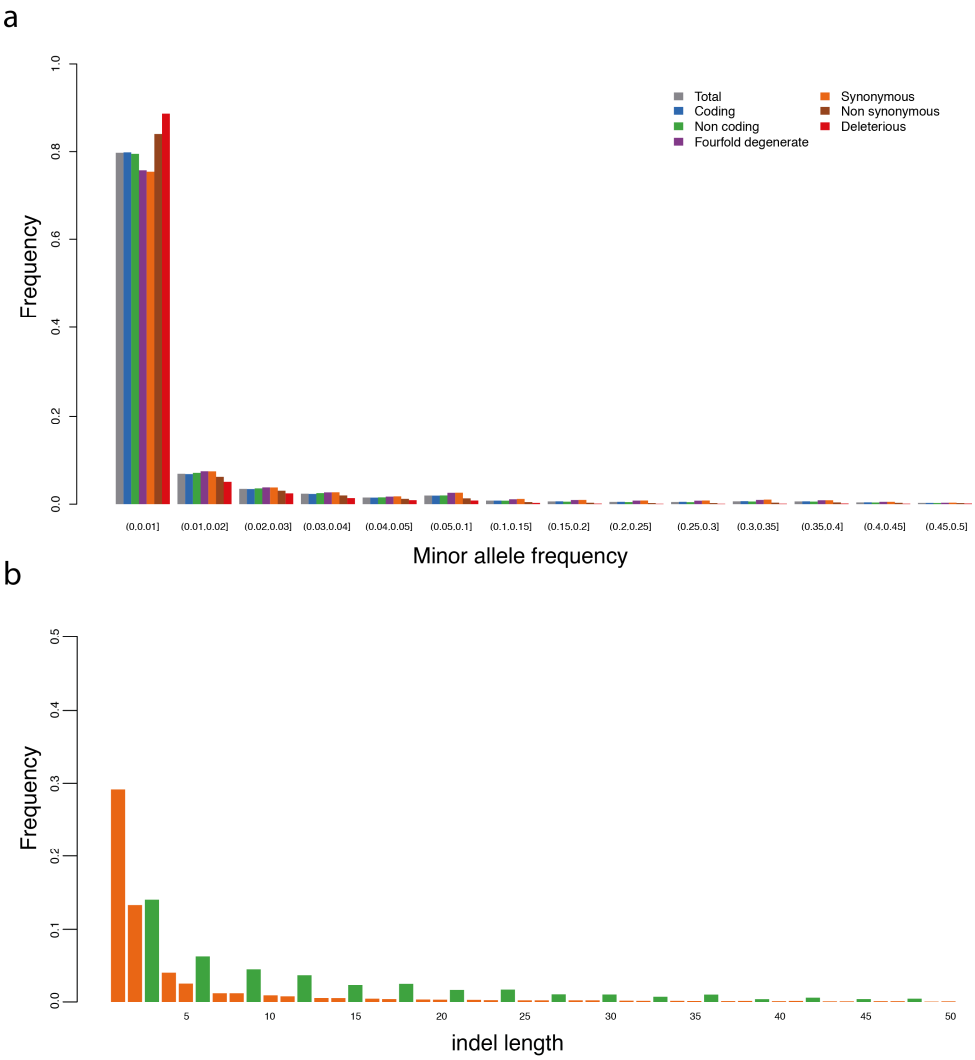


Figure 3: Frequency spectrum of the SNPs in the 1,011 genomes and coding indel length distribution.

Ploidy and aneuploidy levels significantly vary between subpopulations

With the improvements in sequencing technologies and the subsequent high number of *S. cerevisiae* genomes sequenced, it was also possible to have a better overview of other genetic variants, which are also of great importance such as the level of ploidy and aneuploidy. Indeed, these large-scale genomic events are often thought to be a way of rapid adaptation to stress in the environment in yeast[46,47]. Consequently, it is important to investigate the variation of ploidy level as well as chromosomal copy numbers (aneuploidies) across subpopulations.
The haplo-diploid life cycle of the budding yeast allows its propagation as haploid or diploid but the observation of the ploidy among a large number of strains showed that most isolates of *S. cerevisiae* are diploid (77.8% out of a sample of 1,011 isolates) as vegetative cells and that the rare polyploid isolates fell into given subpopulations such as the beer lineages, suggesting that these polyploidies might be a consequence of human selection for beer production. Moreover, taken globally, diploid isolates seem to have a fitness advantage over other ploidies[36,48], comforting the hypothesis of a possible burden to maintain higher ploidies.

Concerning the variation in chromosomal copy number, it is the consequence of chromosome segregation and replication errors. Defect in biological processes such as the proper functioning of the mitotic spindle apparatus or the spindle assembly checkpoint might compromise the proper functioning of chromosome segregation, resulting in an abnormal number of chromosomes in the progeny[49]. It has also been shown that environmental stress can induce chromosome missegregation as a transient way to overcome strong selective pressure allowing the propagation of the population for the emergence of adaptive mutations with a smaller fitness impact[50]. A species-wide scale survey of aneuploidies revealed that they are relatively well tolerated and observed in a quarter of the wild isolates[36]. Aneuploidy events are therefore not uncommon in the *S. cerevisiae* species. Some chromosomes are more often observed in an aneuploid state than others, partially correlated with chromosomal size. Interestingly, strains carrying aneuploid chromosomes are not randomly distributed within the species. In fact, a species-wide survey revealed a significant enrichment of aneuploid strains in the sake (p-value = $2.9e^{-08}$), ale beer (p-value = $5.9e^{-06}$) and mixed population containing the baker isolates (p-value = $3.6e^{-09}$) (Figure 4). The two latest categories showing also an enrichment for strains with a higher ploidy (3n, 4n and 5n). Indeed, haploids and diploids isolates are more stable than strains with ploidy states that are thought to be unstable (≥ 3n).

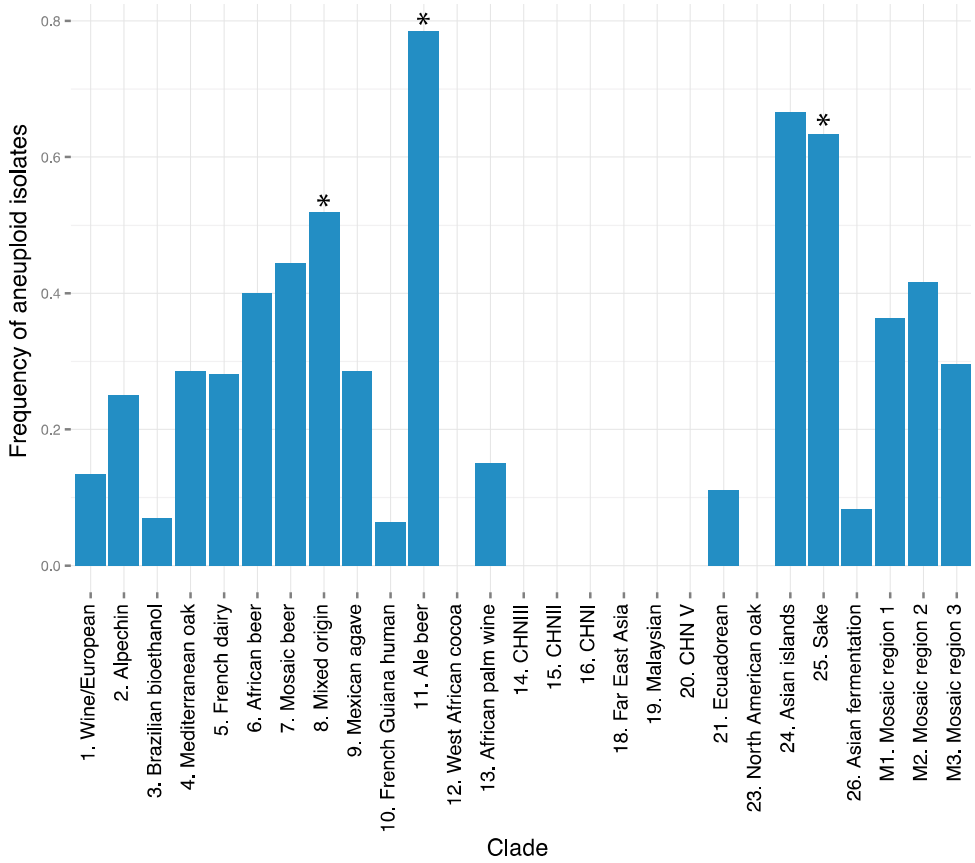


Figure 4: Proportion of aneuploid isolates by clades defined using a set of 1,011 *S. cerevisiae* genomes

Aneuploidy is frequently observed in laboratory evolution studies and in drug-resistant fungal pathogens[50–53]. Such large-scale events play a role in gene amplification but the expression of genes is not always linked to the copy number of the genes, due to the phenomenon of dosage

compensation[47,54]. In fact, it seems that the link between fitness and aneuploidies cuts two ways: they are associated with a decrease of cellular fitness but it has been shown that aneuploidies can be selected under several environments as a response to stress, showing that these events represent a rapid route to phenotypic evolution[50,55].

Copy number variation landscape

Gene duplication is a major force driving biological adaptation in plants[56] (see [57] for a review) and animals[58,59]. The understanding of how gene duplication might give origin to new genes and adaptation is a question of great importance in evolutionary biology. Theory predicted three possible fates for duplicated genes, other than the loss of the duplicated copy: (i) neofunctunalization, which is the acquisition of a new function for one of the duplicated genes which is relieved from purifying selection, (ii) subfunctionalization, splitting the multiple activities of the ancestral gene between the different child copies, and (iii) the two versions of the ancestral gene keep the same function, possibly resulting in increased activity of the gene[60].

As *S. cerevisiae* belongs to the lineage that underwent the Whole Genome Duplication (WGD) around 100 million years ago[61,62], it constitutes a precious resource to study the evolutionary impact of copy number variation. Yet, one of the main drawbacks is the lack of a fossil record of DNA that would allow to link mutations to phenotypic changes. To overcome this limitation, experimentally evolved populations have been constituted to investigate the way mutations arise and spread through a population, how new phenotypes emerge and therefore, how can an organism adapt to its environment. For example, experimental evolution under nitrogen limiting conditions allowed to detect CNVs of the specific nitrogen transporter genes *PUT4*, *DUR3* and *DAL4*, demonstrating that adaptive gene duplication is a significant contributor to adaptation in *S. cerevisiae*[63].

Natural populations of the budding yeast exist multiple ecological niches, implying that they had to adapt to diverse constraints. Investigation using different natural populations of *S. cerevisiae* demonstrated the association between gene duplication and domestication events (Figure 5)[36]. Examples include CNVs of phenotypically relevant genes, like gene families involved in fermentation processes such as copper resistance (*CUP*), flocculation (*FLO*) and the metabolism of glucose (*HXT*)[64]. Another example clearly illustrating the role of gene duplication in adaptation to human-related environments is the duplication of the *MAL1* and *MAL3* loci in strains used for beer fermentation, allowing an efficient maltose metabolism, which is the most abundant sugar available in beer wort[30,31]. Finally, adaptive duplications have also been detected on clinical strains, for which there are more copies of the *CUP1* gene, increasing the resistance of these strains to copper, providing fitness to human host[26].

Recent Genome-Wide Association Studies (GWAS) of fitness traits have shown that CNVs explain a larger proportion of the phenotypic variance and have higher effects compared to the

single nucleotide variants[36]. Moreover, the biological interpretation of the phenotypic impact of CNVs is more straightforward, as they are often linked with the level of expression of the genes they affect. Taken together, those observations are stressing the fact that integrating CNVs in genotype-phenotype relationship scans might lead to a decrease in missing heritability.

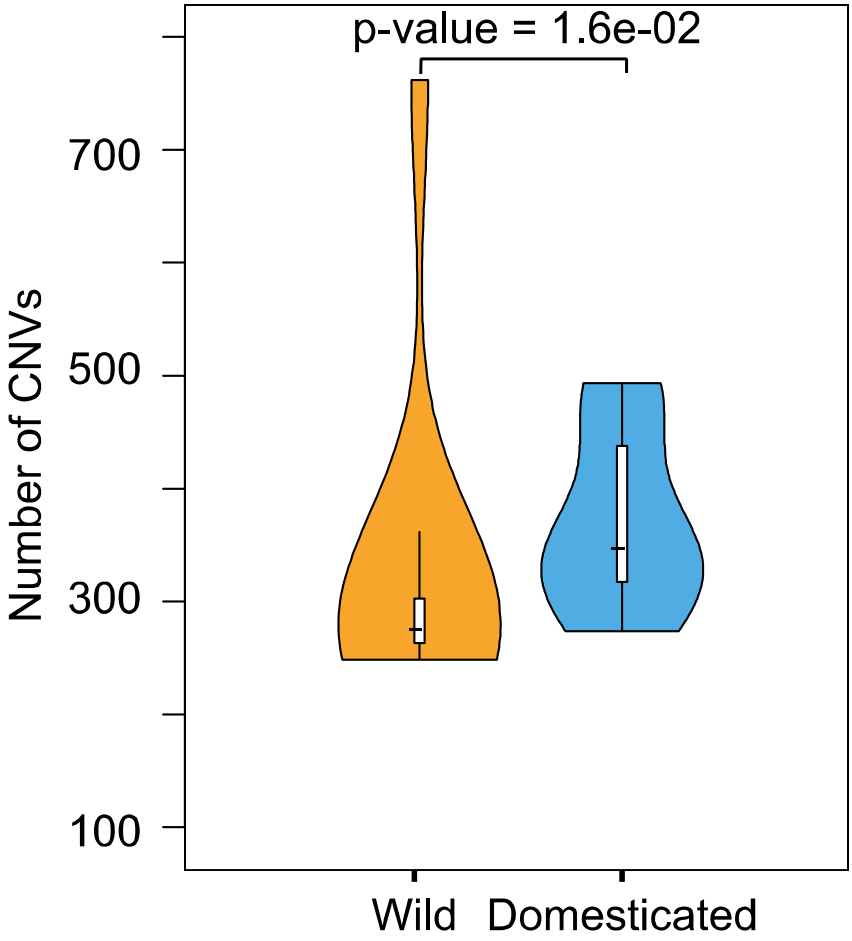


Figure 5: Number of CNVs in wild or domesticated isolates of *S. cerevisiae*. CNVs have been determined using the pangenome of *S. cerevisiae* based on 1,011 isolates.

Introgressions and horizontal genes transfers in *S. cerevisiae*

Investigation of the genome-wide variation within *S. cerevisiae* was limited for a long time due to the low number of sequenced genomes. Some attempts used multi-species micro-arrays to explore the gene-content and retrace the origins of the ORFs and revealed interspecies hybrids and several introgressed *S. paradoxus* DNA fragments[65,66]. Genome sequencing of the wine strain EC1118 allowed the identification of 3 introgressed regions, one of them originating from *Zyggosaccharomyces bailli* and another traced back to the *Saccharomyces* genus[67]. Introgressed ORFs are often replacing *S. cerevisiae* orthologous ORFs, suggesting that their integration happens through homologous recombination[32]. With the investigation of

bigger genomic datasets, it appears that introgressions from *S. paradoxus* are ubiquitous in *S. cerevisiae* isolates, suggesting that the formation of interspecific hybrid and gene flow with this wild sister species is common and that these events constitute major evolutionary processes that shaped the genome of *S. cerevisiae*[36]. Horizontal Gene Transfers (HGT) are mostly affecting the human-related strains that are used in fermentative environments, like wine or beer strains[36,68]. These environments often imply the coexistence of *Torulaspora* and *Zygosaccharomyces* species, which have been identified to be the donor of more than 30% of the introgressions detected in 1,011 individuals. It has also been demonstrated that some events affect clinical strains, with the introgression of *PDR5*, conferring resistance to the translation inhibitor cycloheximide and the ergosterol synthesis inhibitor, representing an adaptation of these strains to their environment[26]. The HGT regions retained are generally short, as multiple rearrangements leading to partial deletions of the ancestral HGT. Events like introgressions or HGT in specific strains lead to a variation of the gene content that is also interesting to assess over the population. To that end, the assemblies of more than a thousand genomes of *S. cerevisiae* allowed the determination of the pangenome, *i.e.* all the open reading frames present within the species. This led to the identification of 7,797 ORFs in the whole species, therefore adding 1,716 ORFs to the 6,081 non-redundant ORFs present in the reference strain S288C[36]. Among these genes, 4,940 are present in all strains, forming the core genome, and 2,857 genes are present with a variable frequency in the population, representing the accessory genome[36]. These accessory genes are found preferentially in subtelomeres, which are thought to be hotspots of gene content variation and translocations[33,69].

New insight into the genetic basis of phenotypic variation

Elucidating the causes of the awesome phenotypic diversity observed in natural populations is a major challenge in biology. It is now clear that the understanding of traits is not only hampered by non-heritable factors such as the environment and epigenetic variation, but also confounded by the lack of complete knowledge concerning the genetic components of complex traits. More than a century after the rediscovery of Mendel's law, the genetic architecture of traits still resists generalization. In fact, dissection of the genotype-phenotype correlation in humans and other higher model eukaryotes such as *A. thaliana* and *C. elegans*, while extremely important, is very difficult not only due to genetic complexity, pleiotropy and gene-environment interactions but also to large and complex genomes.

Linkage mapping in *S. cerevisiae*

Besides population genomics, yeast is also a model organism to dissect the genetic origin of the phenotypic diversity. In human, many genetic diseases are quantitative traits. Studies of quantitative traits in humans and other higher eukaryotes, while extremely important, are hampered not only by genetic complexity, pleiotropy and gene-environment interactions but

also by large genomes. The basic understanding of phenotypic diversity is facilitated by the use of microbial models, particularly *S. cerevisiae* with its physically small (12 Mb), genetically large (4,500 cM)[70] and precisely annotated genome. Phenotypic diversity among yeast isolates is significant, and variation is apparent among the surveyed strains at different levels.
The genetic basis of a number of relevant phenotypes has been studied in yeast, including growth at high temperature, sporulation efficiency, telomere length, gene expression, and response to drugs using linkage mapping[71–77] (see [78] for review).

To date, genetic linkage mapping analyses using *S. cerevisiae* have provided basic and valuable insights into genetic architecture. Over 100 quantitative trait genes (QTGs) and half as many quantitative trait nucleotides (QTNs) have been identified[78]. These studies have yielded a general description of quantitative traits and led to preliminary conclusions such as: (i) some QTGs affect multiple traits (ii) multiple QTGs are linked, (iii) an appreciable fraction of phenotypic variation is not caused by SNPs, and (iv) most QTNs are in conserved protein coding sequences.

Nevertheless, these trends might be biased because all the causal loci were identified via a linkage mapping strategy. In addition, this dataset might also be inappropriate to infer general statements due to the low number of functional alleles identified.

Genome-wide association studies in *S. cerevisiae*

The increasing number of fully sequenced genomes of the same species made genome-wide association studies a very attractive approach in genetics as they allow the scoring of thousands of genetic variants at once from a population of unrelated individuals instead of segregants from a cross. This point is important because it takes advantage of the historical recombination in the species, which promises to increase resolution of the detected loci to small regions, or even to the quantitative trait nucleotide (QTN). A particular interest has been shown for GWAS in human, due to the motivation in understanding the genetic determinants of diseases. Many studies led to near 2,500 publications and described more than 250,000 significant trait/disease associated SNPs[79]. Genome resequencing projects using model organisms such as *Drosophila melanogaster*[80,81] or *Arabidopsis thaliana*[6] were also initiated to identify genetic variation that is correlated with quantitative traits and laid the basis for genome-wide association studies in non-human organisms.

In the budding yeast, the first association study was performed to gain insight into the pathogenicity of *S. cerevisiae*. To that end, a population based genome-wide environmental association analysis of 44 clinical *vs.* 44 nonclinical strains was performed using tiling arrays. Several polymorphisms significantly associated with clinical isolates were identified, located in the coding sequences of genes intervening in processes such as pseudohyphal formation, cell wall maintenance and cellular detoxification. These results suggest that these functions are

likely to play an important role in the pathogenicity of clinical strains[82]. One year later, a study questioning the feasibility of GWAS approach using quantitative phenotypes has been performed on the 36 strains of the *Saccharomyces* genetic resource panel (SGRP), which genomes were partially sequenced[24] and the 63 strains, which were genotyped using tiling microarrays[23]. This study revealed that the type-1 error rate was too high in simulated phenotypes to draw any conclusion, whether one significant SNP located in the *HPR1* coding region was identified using the mitochondrial DNA copy number as a phenotype. They conclude that GWAS can be an efficient approach in *S. cerevisiae*, but due to the high type-1 error rate, these studies might be limited to the mapping of Mendelian phenotypes or *cis*-acting QTL[83]. Shortly after this study, another attempt at performing GWAS has been performed also using the SGRP dataset and data from 201 phenotypic traits. The results highlighted the importance of correcting for local ancestry with, for example, the inclusion of structure as covariates in the model. The few significant results were biased towards SNPs with a low minor allele frequency. They also simulated phenotypes with different architecture on 63 natural strains[23] and hypothesized that the reduction of noise observed might be due to the larger sample size of the latter dataset[84]. More recently, genomic sequences of 100 strains and 49 phenotypes were used to perform GWAS and allowed to find associations representing proof of principles, such as the association of SNPs and CNV of the Li^+ Na^+ pump-encoding *ENA* gene with lithium resistance variation, or the mapping of the chromosome 8-16 translocation previously described in wine strains with the resistance to sulfite[85]. Overall, this study pointed out that the peaks are still wide with 100 genomes, few or no significant association found for multiple phenotypes and the variants detected are always variants of large effects, whereas GWAS aims at identifying multiple loci underlying phenotypic variation. These limitations are pointing to the fact that 100 genomes might still not be sufficient to have sufficient power for GWAS[26].

The sequencing of over a thousand isolates of *S. cerevisiae* together with their phenotyping on 35 fitness conditions gave for the first time the opportunity to perform GWAS on datasets scaling up to the ones used in other organisms. In total, 35 variants were found significantly associated with 14 conditions. Among these variants, 22 were CNVs, usually gathering larger effects than SNPs[36]. Taken together, these results reveal the potential of *S. cerevisiae* as a model organism for GWAS. Even though the population is highly stratified, it is possible to overcome this confounding factor and to limit the high type-1 error rate by using an appropriate model that corrects for structure together with a large enough sample size.

Conclusion

The picture drawn by these yeast population genomic studies is now giving us appreciable insights into the budding yeast *S. cerevisiae* genome evolution. The multiple population genomic studies led up to date represent valuable data that allows us to make sound analyses. An out-of-China origin has been hypothesized, as well as several independent domestication events due to its long association with human-related environments to produce bread or fermented products such as beer or wine. Further sequencing of strains allowed to describe subpopulation formed by geography, ecology or human associated environments, showing differences in genetic variation ranging from single nucleotide variants to polyploidy. This species has been widely used to detect variants underlying phenotypic traits using linkage mapping and more recently genome-wide association studies. This latter strategy was thought to be problematic due to the highly structured population of *S. cerevisiae* but it seems that a large enough sample size can overcome this problem. Performing association studies using a model organism presenting a less complex genome than the human genome might be useful to improve this methodology and might help to characterizing the genotype-phenotype relationship.

References

1. Gulcher, J. & Stefansson, K. Population genomics: laying the groundwork for genetic disease modeling and targeting. *Clin. Chem. Lab. Med.* **36,** 523–7 (1998).
2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–91 (2016).
4. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
5. 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana. Cell* **166,** 481–91 (2016).
6. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43,** 956–63 (2011).
7. Andersen, E. C. *et al.* Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat. Genet.* **44,** 285–90 (2012).
8. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* genetic reference panel. *Nature* **482,** 173–78 (2012).
9. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).
10. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–86 (2014).
11. Mosquera, T. *et al.* Targeted and untargeted approaches unravel novel candidate genes and diagnostic SNPs for quantitative resistance of the potato (*Solanum tuberosum L.*) to *Phytophthora infestans* causing the late blight disease. *PLoS One* **11,** 1–36 (2016).
12. Pantalião, G. F. *et al.* Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica* **144,** 651–64 (2016).
13. Branham, S. E., Wright, S. J., Reba, A., Linder, C. R. & Branham, R. Genome-wide association study of *Arabidopsis thaliana* identifies determinants of natural variation in seed oil composition. *J. Hered.* **1,** 1–9 (2015).
14. Peter, J. & Schacherer, J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33**, 73–81 (2015).
15. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *eLife* **4,** 1–27 (2015).
16. Fawcett, J. a. *et al.* Population genomics of the fission yeast *Schizosaccharomyces pombe*. *PLoS One* **9,** (2014).
17. Jeffares, D. C. *et al.* The genomic and phenotypic diversity of *Schizosaccharomyces pombe*. (2015).
18. Almeida, P. *et al.* A Gondwanan imprint on global diversity and domestication of wine and cider yeast *Saccharomyces uvarum. Mol. Ecol.* **24**, 5412–27(2015).
19. Hittinger, C. T. *et al.* Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464,** 54–8 (2010).
20. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32,** 184–92 (2015).
21. Jung, P. P., Friedrich, A., Reisser, C., Hou, J. & Schacherer, J. Mitochondrial genome evolution in a single protoploid yeast species. *G3 (Bethesda).* **2,** 1103–11 (2012).
22. Freel, K. C., Friedrich, A., Hou, J. & Schacherer, J. Population genomic analysis reveals highly conserved mitochondrial genomes in the yeast species *Lachancea thermotolerans*. *Genome Biol. Evol.* **6,** 2586–94 (2014).
23. Schacherer, J., Shapiro, J. a, Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458,** 342–45 (2009).

24. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–41 (2009).
25. Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda).* **6,** 2421–34 (2016).
26. Strope, P. K. *et al.* The 100-genomes strains , an *S* . *cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* 125, 762–74 (2015).
27. Almeida, P. *et al.* A Population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol. Ecol.* **24**, 5412-27 (2015)
28. Wang, Q. M., Liu, W. Q., Liti, G., Wang, S. A. & Bai, F. Y. Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol. Ecol.* **21,** 5404–17 (2012).
29. Ludlow, C. L. *et al.* Independent origins of yeast associated with coffee and cacao fermentation. *Curr. Biol.* **26,** 965–71 (2016).
30. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–61 (2016).
31. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–410 (2016).
32. Barbosa, R. *et al.* Evidence of natural hybridization in Brazilian wild lineages of *Saccharomyces cerevisiae*. *Genome Biol. Evol.* **8**, 317–29(2016).
33. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–88 (2014).
34. Legras, J. L., Merdinoglu, D., Cornuet, J. M. & Karst, F. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Mol. Ecol.* **16,** 2091–102 (2007).
35. Sicard, D. & Legras, J. L. Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex. *Comptes Rendus - Biol.* **334,** 229–36 (2011).
36. Peter, J. *et al.* Yeast evolutionary history and natural variation revealed by 1,011 genomes. *in revision* (2017).
37. Naumov, G. I., Gazdiev, D. O. & Naumova, E. S. [Identification of the yeast species *Saccharomyces bayanus* in far east Asia]. *Mikrobiologiia* **72,** 834–39 (2003).
38. Vaughan-Martini, A. Martini, A. Facts, myths and legends on the prime industrial microorganism. *J. Ind. Microbiol. Biotechnol.* 514–22 (1995).
39. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38,** 1077–81 (2006).
40. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4957–62 (2008).
41. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32**, 184–92 (2014).
42. Magwene, P. M. *et al.* Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1987–92 (2011).
43. Ohnuki, S. *et al.* Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. *G3 (Bethesda)* **7**, 2807–20 (2017).
44. 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).
45. Li, Y. *et al.* Genomic evolution of *Saccharomyces cerevisiae* under chinese rice wine fermentation. *Genome Biol. Evol.* **6,** 2516–26 (2014).
46. Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519,** 349–52 (2015).
47. Hose, J. *et al.* Dosage compensation can buffer copy-number variation in wild yeast. *eLife* **4**, (2015).
48. Zörgö, E. *et al.* Ancient evolutionary trade-offs between yeast ploidy states. *PLoS Genet.* **9,** (2013).

49. Mulla, W., Zhu, J. & Li, R. Yeast: A simple model system to study complex phenomena of aneuploidy. *FEMS Microbiol. Rev.* **38,** 201–12 (2014).
50. Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 21010–5 (2012).
51. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nat. Genet.* **25**, 333–337 (2000).
52. Dunham, M. J. *et al.* Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16144–9 (2002).
53. Pavelka, N. *et al.* Aneuploidy confers quantitative proteome changes and phenotypic variation in budding yeast. *Nature* **468,** 321–5 (2010).
54. Gasch, A. P. *et al.* Further support for aneuploidy tolerance in wild yeast and effects of dosage compensation on gene copy-number evolution. *eLife* **5,** (2016).
55. Tan, Z. *et al.* Aneuploidy underlies a multicellular phenotypic switch. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12367–72 (2013).
56. Carretero-Paulet, L. & Fares, M. A. Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. *Mol. Biol. Evol.* **29,** 3541–3551 (2012).
57. Panchy, N., Lehti-Shiu, M. D. & Shiu, S.-H. Evolution of gene duplication in plants. *Plant Physiol.* **171,** 2294–316 (2016).
58. Hoegg, S., Brinkmann, H., Taylor, J. S. & Meyer, A. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* **59,** 190–203 (2004).
59. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
60. Ohno, S. *Evolution by gene duplication*. (1970).
61. Wolfe, K. H. Origin of the yeast whole-genome duplication. *PLoS Biol.* **13,** 1–7 (2015).
62. Marcet-Houben, M. & Gabaldón, T. Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLOS Biol.* **13,** e1002220 (2015).
63. Hong, J. & Gresham, D. Molecular specificity, convergence and constraint shape adaptive evolution in nutrient-poor environments. *PLoS Genet.* **10,** (2014).
64. Steenwyk, J. & Rokas, A. Extensive copy number variation in fermentation-related genes among *Saccharomyces cerevisiae* wine strains. *G3 (Bethesda)* **7,** 1475–85 (2017).
65. Muller, L. A. H. & McCusker, J. H. A multispecies-based taxonomic microarray reveals interspecies hybridization and introgression in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **9,** 143–52 (2009).
66. Dunn, B. & Sherlock, G. Reconstruction of the genome origins and evolution of the hybrid lager yeast *Saccharomyces pastorianus*. *Genome Res.* **18,** 1610–1623 (2008).
67. Novo, M. *et al.* 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,** 16333–8 (2009).
68. Marsit, S. & Dequin, S. Diversity and adaptive evolution of *Saccharomyces* wine yeast: a review. *FEMS Yeast Res.***15,** (2015).
69. Brown, C. a, Murray, A. W. & Verstrepen, K. J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* **20,** 895–903 (2010).
70. Mortimer, R. K., Contopoulou, C. R. & King, J. S. Genetic and physical maps of *Saccharomyces cerevisiae*, Edition 11. *Yeast* **8,** 817–902 (1992).
71. Askree, S. H. *et al.* A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 8658–63 (2004).
72. Deutschbauer, A. M. & Davis, R. W. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat. Genet.* **37,** 1333–40 (2005).
73. Gatbonton, T. *et al.* Telomere length as a quantitative trait: Genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS Genet.* **2,** 0304–15 (2006).
74. Gerke, J. P., Chen, C. T. L. & Cohen, B. A. Natural isolates of *Saccharomyces cerevisiae* display complex genetic variation in sporulation efficiency. *Genetics* **174,** 985–97 (2006).

75. Kim, H. S. & Fay, J. C. A combined-cross analysis reveals genes with drug-specific and background-dependent effects on drug sensitivity in *Saccharomyces cerevisiae*. *Genetics* **183,** 1141–51 (2009).
76. Ronald, J. & Akey, J. M. The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS One* **2,** (2007).
77. Steinmetz, L. M. *et al.* Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416,** 326–30 (2002).
78. Fay, J. C. The molecular basis of phenotypic variation in yeast. *Curr. Opin. Genet. Dev.* **23,** 672–7 (2013).
79. Li, M. J. *et al.* GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44,** 869–76 (2016).
80. Mackay, T. F. C., Stone, E. a & Ayroles, J. F. The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* **10,** 565–577 (2009).
81. Lack, J. B. *et al.* The drosophila genome nexus: A population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. *Genetics* **199,** 1229–41 (2015).
82. Muller, L. A. H., Lucas, J. E., Georgianna, D. R. & McCusker, J. H. Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol. Ecol.* **20,** 4085–97 (2011).
83. Connelly, C. F. & Akey, J. M. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191,** 1345–53 (2012).
84. Diao, L. & Chen, K. C. Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics* **192,** 1503–11 (2012).
85. Pérez-Ortín, J. E., Querol, A., Puig, S. & Barrio, E. Molecular characterization of a chromosomal rearrangement involved in the adaptive evolution of yeast strains. *Genome Res.* **12,** 1533–39 (2002).

Project Summary:
Species-wide investigation of the genotype-phenotype relationship in *Saccharomyces cerevisiae*

Overview of the project

One of the challenges in modern genetics is to be able to identify genetic variants underlying phenotypic diversity or even being able to predict an individual's phenotype based on its genotype. It is therefore of great importance to have a precise view on the genetic variability within a species and to establish a precise catalog of genetic variation. To that end, large-scale population genomics projects have been initiated and aimed at increasing our knowledge on population evolutionary history and being able to detect genetic variants impacting traits, with a strong focus on the detection of disease-related risk alleles. One elegant way to access this knowledge is by performing genome-wide association studies. However, the genotype-phenotype relationship remains unclear and hard to explore in human. In this context, we sought to have a better insight on the forces acting on this relation using a well-known model organism, whose genome is much simpler than in human, namely the budding yeast *Saccharomyces cerevisiae*. Their small and compact genome greatly facilitates data generation, processing and interpretation Moreover, the considerable levels of genetic and phenotypic diversity displayed in natural populations of yeast allows them to adapt to changing environments and to new ecological niches, which makes yeast well-suited organisms for population genomics studies.

In order to have a deeper insight into the genetic origin of phenotypic variation within the budding yeast species *Saccharomyces cerevisiae*, it is necessary to use a large number of natural isolates originating from various diverse ecological niches across the world. In the **first chapter**, we present a global description of the species-wide genotypic and phenotypic variation. To that end, we obtained deep coverage sequenced genomes for more than 1,000 natural isolates of *S. cerevisiae*, which constitutes a comprehensive survey of differential genome evolution patterns. We investigated several types of genetic variation, such as Single Nucleotide Polymorphisms (SNPs), Copy Number Variation (CNV), ploidy and aneuploidy levels across 26 subpopulations having their own evolutionary history (Figure 1a). In parallel, we performed high-throughput phenotyping of fitness traits measured on different conditions impacting various physiological and cellular responses. The high SNP density allowed us to perform Genome-Wide Association Studies (GWAS) with an unprecedented statistical power in this species.

In the **second chapter**, we used 5 subsets as well as the total population of 1,011 genomes and performed extensive runs of simulation of GWAS, by generating phenotypes based on randomly chosen SNPs. We then evaluated the ability to detect the causal variants using a linear mixed model approach for both Mendelian and complex traits (Figure 1b). These analyses showed that GWAS are possible in *S. cerevisiae*, and underlined the importance of population size to overcome the high type-1 error brought by population stratification, as well as other parameters such as relatedness, minor allele frequency and effect size.

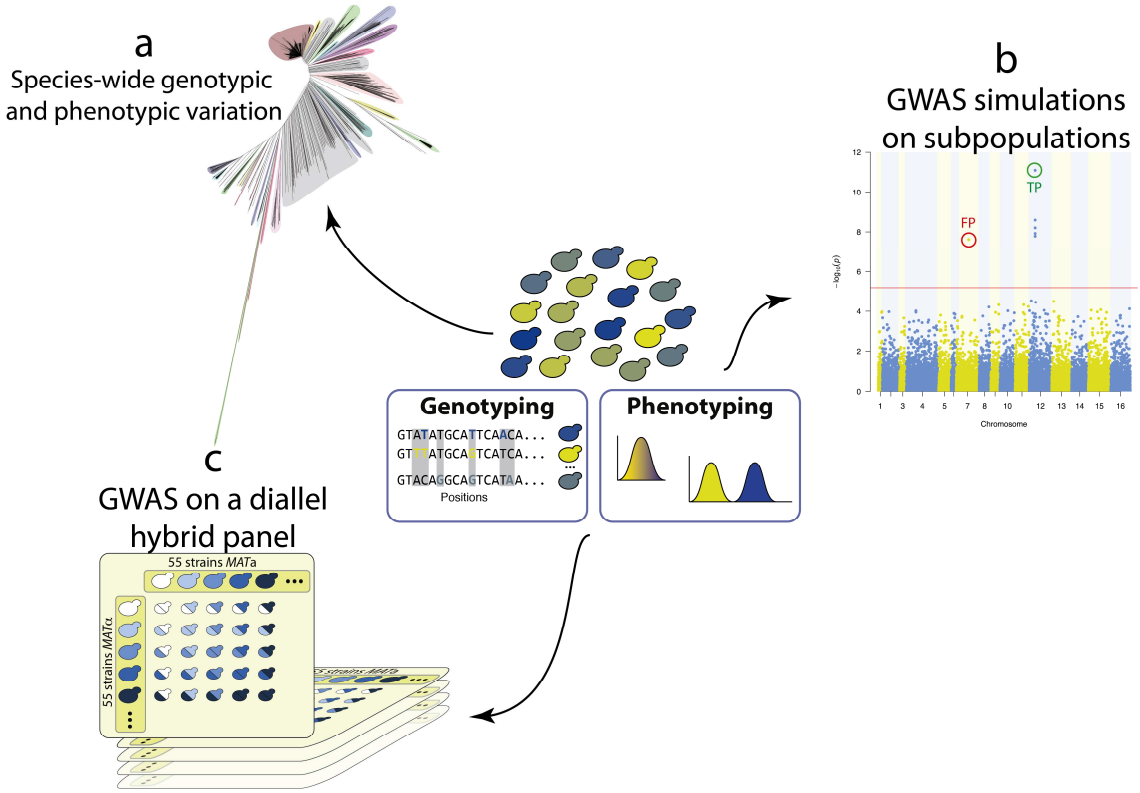


Figure 1: Schematic description of the project. **a.** Species-wide investigation of genotypic and phenotypic variation of *Saccharomyces cerevisiae* based on the genomes and the phenotypic data obtained from 1,011 individuals. **b.** Simulation of genome-wide association studies on subpopulations. **c.** Genome-wide association studies on a diallel hybrid panel of pairwise crosses.

Nevertheless, as it has been observed in many species, there is a strong bias towards rare alleles in *S. cerevisiae*, constituting a possible source of missing heritability. In the **third chapter**, we tried to overcome this limitation by performing GWAS using a population resulting from a diallel cross panel (Figure 1c). Indeed, the shuffling of the parental genomes allows an overrepresentation of some alleles compared to their frequency in the species. This study led to identification of a large set of functional polymorphisms that underlie phenotypic variation, including a rare variant located in the coding sequence of the *GAL2* gene, thus demonstrating that such a panel could be an efficient way to reduce the missing heritability.

Publications related to this work

1. Peter J., De Chiara M., Friedrich A., Yue J-X., Pflieger D., Bergström A., Sigwalt A., Freel K., Llored A., Cruaud C., Labadie, K., Aury J-M., Istace B., Lebrigand K., Barbry P., Engelen S., Lemainque A., Wincker P., Liti G. and Schacherer J. Yeast evolutionary history and natural variation revealed by 1,011 genomes. **submitted** (2017)

2. Peter J. & Schacherer J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast*. **3**:73-81 (2015).

Chapter 1:
Saccharomyces cerevisiae evolutionary history and natural variation revealed by 1,011 genomes

Introduction

Within all species, genetic variation constitutes the raw material for phenotypic diversity between individuals. The elucidation of the genotype-phenotype relationship is one of the major challenges in modern genetics.

The first step to address it is to have a better understanding of the genetic variation across a large number of individuals from the same species by establishing a catalog of genetic variants. Due to the emergence of cost-effective sequencing technologies, the generation of such catalogs no longer represents a bottleneck and it is the logical step after having a reference genome is sequenced, assembled and annotated. Up to date, several large-scale resequencing projects have been initiated in the last decade that enabled researchers to gather enough sequences to make high-throughput approaches possible. Examples of these projects include the 1000 genomes project[1] or the UK10K project[2] for human, the 1001 project for *Arabidopsis thaliana*[3], or the *Drosophila* Genetics Reference Panel (DGRP)[4,5]. Their goal is to gather enough sequences to have the most complete view of genetic variation throughout a species. Motivations for such projects are multiple, and complementary. First, they were designed to enable a better understanding of demographics and the evolutionary histories of populations. Second, such datasets allow researchers to understand the processes by which genetic diversity is generated and maintained. Third, they can help to distinguish between effects acting on the whole genome (such as drift, migration, inbreeding) and those acting on individual loci (selection, mutation, recombination). Finally, one of the major incentives of resequencing projects is to obtain a better insight into the relationship between genotypes and phenotypes, and more precisely to build genetic datasets that would allow the mapping of allelic variants responsible for phenotypic diversity. Indeed, being able to detect genetic variants responsible for a trait of interest is a major challenge in modern biology. The dissection of the genetic architectures of trait is a crucial step to be able at some point to predict a phenotype based on the genotypic information. In the budding yeast *Saccharomyces cerevisiae*, even though several genomes have been sequenced, there was a bias towards human associated strains, such as wine strains, beer strains or clinical strains. As a consequence, our view of the genetic diversity of this species remained incomplete. Moreover, the number of sequenced genomes stood in contrast with the one from other model organism such as human, *Arabidopsis thaliana*, or *drosophila*, preventing us from sufficient statistical power when performing population genomic studies. Increasing the number of sequenced genomes and the diversity of the geographical and ecological origins was therefore necessary.

Here, I present the genome sequencing of more than 1,000 natural isolates of *S. cerevisiae*, as well as the high-throughput phenotyping of those isolates. This study constitutes the first comprehensive view of genome evolution at different levels (*e.g.* accounting for differences among ploidy, aneuploidy, genetic variants, hybridization, and introgressions), which is challenging to obtain at that scale and accuracy for any other model organisms. Also, coupling of the phenotypic characterization with the genome sequencing provided the opportunity to

perform genome-wide association studies with an unprecedented sample size, which was the major issue for such studies in *S. cerevisiae*. With the exhaustive characterization of genetic variants such as single nucleotide polymorphisms (SNPs) and copy number variation (CNV), our study brings new insights into the genetic architecture of traits as well as the missing heritability.

Species-wide overview of the genetic and phenotypic diversity

We assembled a collection of 1,011 *S. cerevisiae* isolates that maximize their ecological and geographical origins. The samples were isolated worldwide (Figure 1a) from human-associated environments (*e.g.* wine fermentation, brewing, baking, dairy products), natural environments (*e.g.* plants, soil, insects) as well as from healthy people and immuno-compromised patients (Supplementary table 1). The isolates were then categorized into 23 ecological groups (Figure 1b).

We deeply sequenced the genomes of 918 isolates using an Illumina HiSeq 100-bp paired-end strategy with a median 226-fold coverage (Supplementary figure 1). We also have included for further analysis 93 strains, which were previously sequenced with a median coverage of 106-fold[6–8]. The reads associated with each sample were mapped to the S288C reference genome and *de novo* assembled. A total of 1,625,809 high-quality reference based SNPs were detected across the 1,011 genomes, of which 69.7% are in coding regions (with coding regions representing a total of 72.9% of the *S. cerevisiae* genome). Most of the detected SNPs are very low-frequency variants, which is in part driven by the sampling scheme. Indeed, 509,011 singletons were detected, representing 31.3% of the total number of polymorphic positions identified and almost 93% of the polymorphic sites are associated with a minor allele frequency (MAF) < 0.05 (Supplementary figure 2). A stronger bias towards rare SNPs can be observed for coding sequences compared to non-coding regions as well as four-fold degenerate sites (Supplementary figure 2). Deleterious mutations, as predicted by SIFT[9] for missense mutations as well as nonsense mutations, show the strongest bias to rare alleles, consistent with the idea that selection prevents such SNPs from spreading in a population (Supplementary figure 2).

In addition, we detected 125,701 small-scale indels (up to 50 bp) and in contrast to what we observed for SNPs, indels positions are mainly located in non-coding regions (76%). The indel length distribution demonstrates that the majority of indels are short (42.6% are 1-bp indel) and that those found in coding regions are longer and strongly biased to multiples-of-three lengths, reflecting the influence of purifying selection (Supplementary figure 3c).

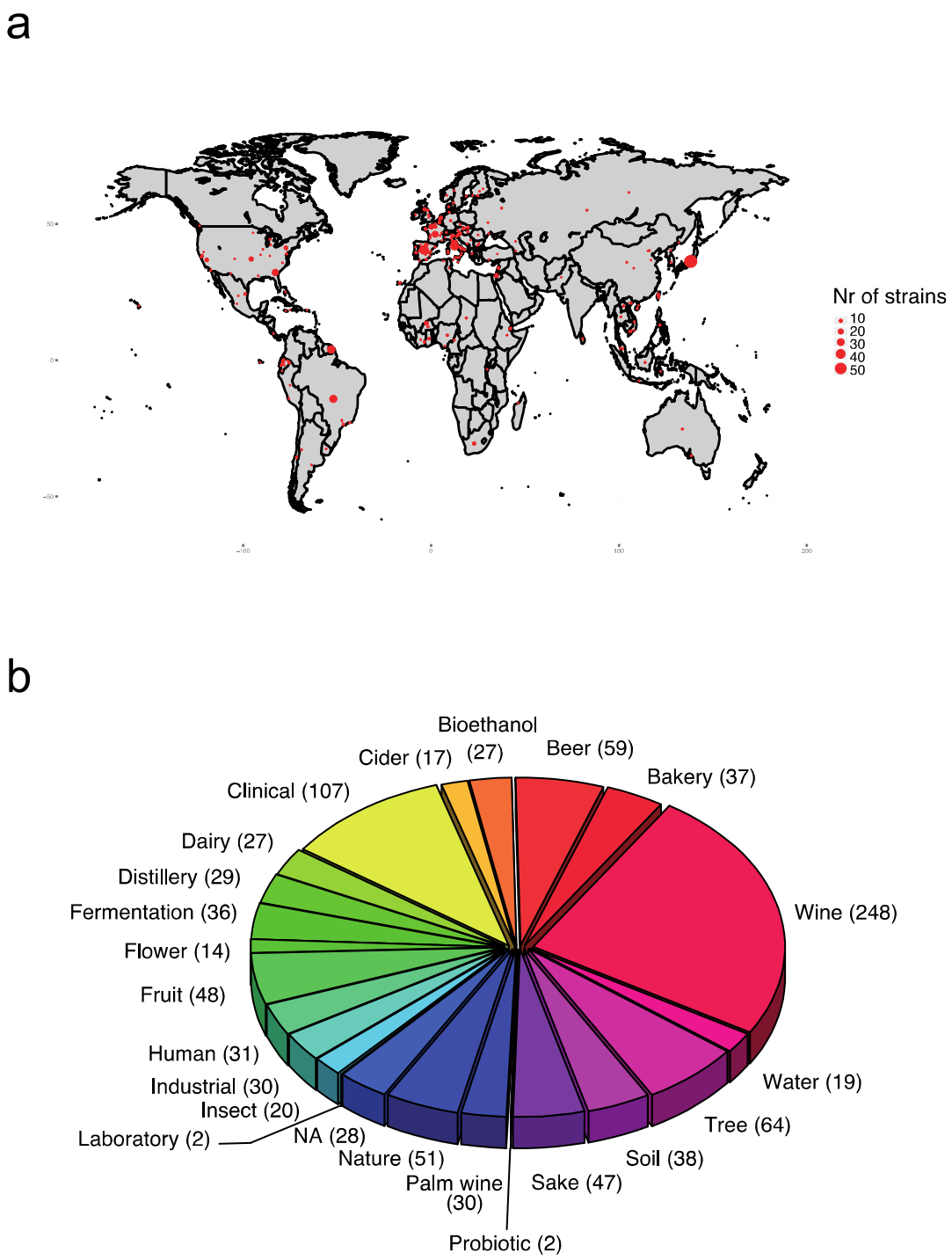


Figure 1: Overview of the 1,011 *S. cerevisiae* sequenced isolates. **a.** Geographical origins of the isolates. The circle size is representative of the number of isolates obtained from each location. **b.** Distribution of the isolates ecological origins. The total number of isolates per category is specified in parentheses.

We also characterized population copy number variants (CNVs), both multi-copies and hemizygosis, by measuring the coverage ratio of every individual pangenomic ORF (see below) normalised for each respective strain's genome (see Methods). Nearly all ORFs have at least one strain with CNV and are heavily enriched in subtelomeric regions, whereas internal

chromosomal regions are largely copy number (CN) stable (Supplementary figure 4b). The majority of CNVs associated with individual ORFs are rare in the population (median minor allele count is 11, Supplementary figure 4c). Variants with high copy numbers only affect a small fraction of ORFs (Supplementary figure 4a) and extreme cases (>20 copies) include ORFs present in the natural 2µ plasmid, mitochondrial genome, ribosomal DNA (RDN), and repetitive elements such as Ty and Y'.

In parallel, 971 strains were phenotyped on different conditions impacting various physiological and cellular responses, including different carbon sources, membrane and protein stability, signal transduction, sterol biosynthesis, transcription, translation, as well as osmotic and oxidative stress (Supplementary table 2, see Methods). In total, we analyzed 34,956 phenotypic measurements covering 36 traits providing a comprehensive analysis of the inheritance patterns of traits. Most of the traits vary continuously across the population, showing the genetic complexity of these phenotypes (Supplementary figure 5). However, some traits follow a bimodal distribution model and therefore a clear Mendelian inheritance pattern such as traits related to $CuSO_4$, 6-azauracil and LiCl resistance[10]. We also have estimated the narrow-sense heritability h^2 (SNP-based h^2 estimation) of each trait from genome-wide SNPs genotyped[11]. Across all the traits, there was a substantial amount of variance explained by all SNPs with a mean of 69%, ranging from 47% to 90%, suggesting the feasibility of performing GWAS on these data (Supplementary figure 6).

Population structure supports a single out-of-China origin

The neighbour-joining phylogenetic tree of 1,011 *S. cerevisiae* strains shows a large variety of well-defined clades, loose clusters and isolated branches (Figure 2). The majority of the strains (813 in total) fall into 26 clades with a broad range of sample sizes (up to 362 for the Wine/European subpopulation), while other 150 strains belong to three groups of poorly related strains (see Supplementary note). In addition, 48 strains do not fall within any of these major clusters and are scattered across the tree. Our data revealed a complex pattern of differentiation that includes distinctly identified lineages, which correlate with geography, environmental niche, and the degree of human association, which is in agreement with previous reports[8,12–15]. Interestingly, the global analysis of a massive number of lineages revealed that domesticated and wild clades largely fall into two well-delineated sides of the phylogenetic tree separated by a large group of mosaic strains. The main exceptions are the wild Mediterranean oak grouping among the domesticated clades and the sake lineage grouping among the wild clades. The Mediterranean oak lineage however shares highly similar genome content with the other wild clades. Strain clustering based on variable ORF content by discriminant analysis of principal component (DAPC) with varying number of groups (up to 35), unequivocally placed the Mediterranean oak with the other wild lineages (Supplementary figure 7).

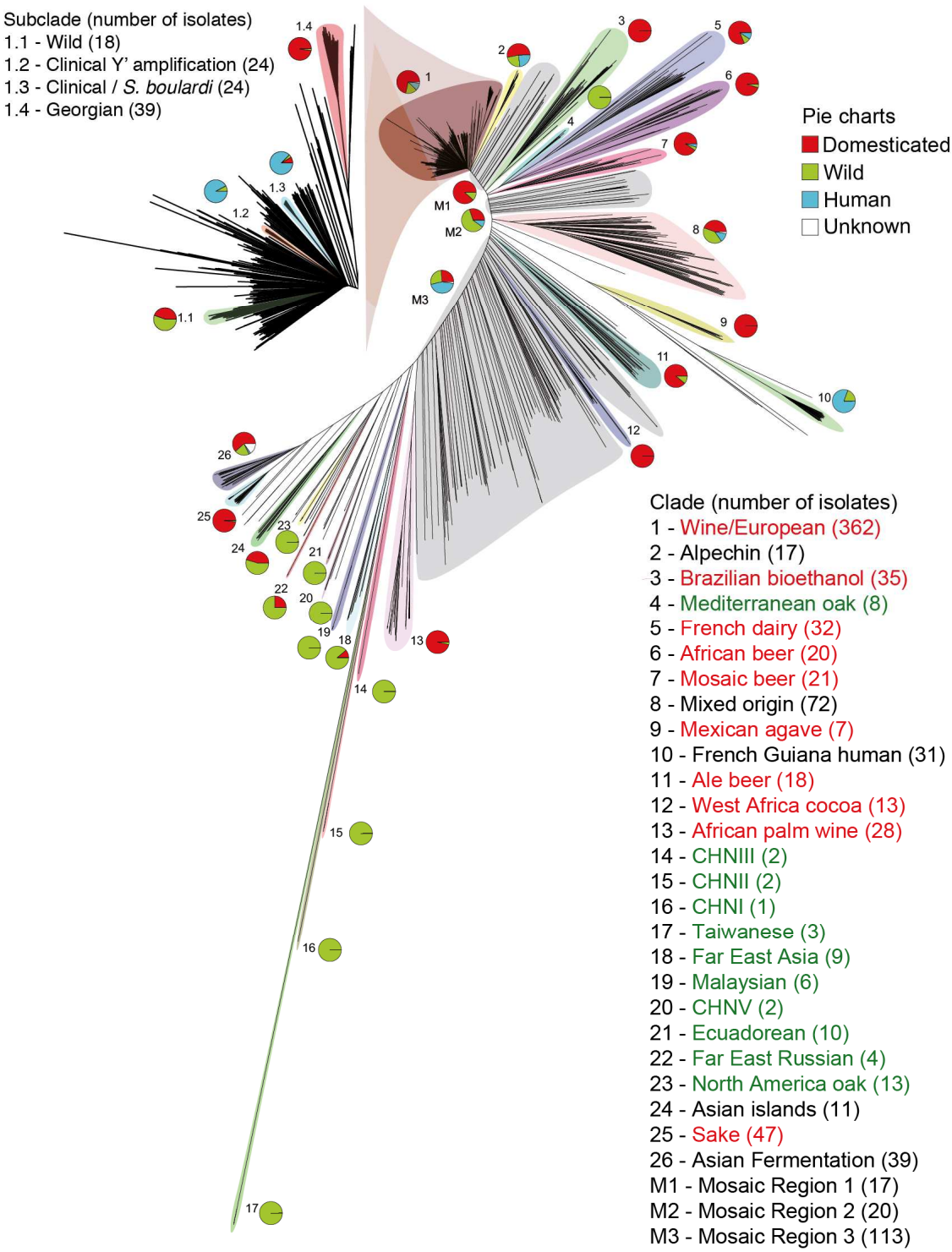


Figure 2: Neighbor-joining tree of the 1,011 *S. cerevisiae* collection built using the biallelic SNPs. We identified 26 clades (numbered clock-wise from 1 to 26) and the three mosaic groups (M1-M3). The pie charts represent the origin of the strains for each clade: domesticated (red), wild (green), human (both clinical and from healthy people, cyan), and with no pie chart for those of unassigned or unknown origin. The clade name color indicates their assignments: domesticated (red) and wild (green). The number of isolates is indicated in parenthesis. The top left inset represents a magnification of the Wine/European clade with four major subclades highlighted.

We further compared domesticated and wild clades for levels of SNP and genome content variation (Supplementary figure 8). Wild clades display higher SNP density than domesticated clades (median 0.55% versus 0.41%), in contrast to lower genome content variation (median 115 of not shared ORFs versus 161 respectively) (Supplementary figure 8, Supplementary table 3-4). Taken together, these findings strongly suggest that a change occurred in the evolutionary mechanisms that shaped the yeast genomes during the domestication process. The wild clades share similar genome content and genomic variation is mainly driven by the accumulation of SNPs. The specific artificial environments colonized by the domesticated clades likely lead to the expansion or loss of ORFs with specific functions resulting in rampant variation in genome content and CNVs.

We used ADMIXTURE[16] to investigate mosaic ancestry in individual strain genomes. Mosaic strains are characterized by admixture from two or more lineages derived by outbreeding[17,18] and frequently manifest as isolated branches in the phylogenetic tree. Our analyses identified at least three main groups of mosaic strains mostly associated with human-related environments (M1-M3, Figure 2). Population structure analysis revealed different sources of ancestry and various level of mosaicism consistent with multiple hybridization events giving rise to the mosaic groups (Supplementary figure 9). These findings underscore the role of human-driven admixture in shaping the *S. cerevisiae* population structure.

The recent discovery of highly diverged wild Chinese lineages suggests East Asia as the possible geographic origin of *S. cerevisiae*[19]. The genome sequences from the Taiwanese wild lineage reported here represent the most divergent population ever described (average of 1.1% sequence divergence to non-Taiwanese strains). This lineage also contains an extremely divergent 2µ plasmid sharing only 80% of identity with the known plasmid variants and was probably inherited from an undescribed *Saccharomyces* species (Supplementary figure 10). We used a subset of highly contiguous *de novo* assemblies from long read technologies that sample the main *S. cerevisiae* lineages and a group of closely related *Saccharomyces* species[20,21]. We identified 2,018 conserved one-to-one orthologs and generated a rooted phylogenetic tree (Figure 3a). The outgroup species branched off near the Taiwanese and Chinese lineages, strongly supporting the *S. cerevisiae* Chinese origin. This scenario is also consistent with the exclusive isolation of *Saccharomyces* species such as *Saccharomyces mikatae* and *Saccharomyces arboricola*[22,23] in East Asia and the broad genetic diversity of the Japanese *Saccharomyces kudriavzevii* populations[24]. Altogether, these observations suggest an Asian origin of the whole *Saccharomyces* species complex.

We then tested the number of “out-of-China” events by investigating the relationship of non-Chinese strains to the Chinese genetic structure. We performed a PCA on SNPs including only the CHN I-V and Taiwanese strains and then projected the rest of the strains onto the PC space defined by these highly diverged lineages (Figure 3b). PC1 defines the separation of the Taiwanese strains from all other strains, consistent with the deep divergence of this lineage. PCs 2-4 then define differentiation between the different Chinese lineages (Figure 3c). Strikingly, the non-Chinese strains all project onto the same part of the space, implying that they all are essentially equally related to the different Chinese lineages (though a slight

gradient still can be observed). In other words, it does not appear that different non-Chinese strains derive from different Chinese lineages. Instead, these results imply not only an East Asian origin of all *S. cerevisiae* strains, but also a single, shared out-of-China origin for all non-Chinese *S. cerevisiae* strains.

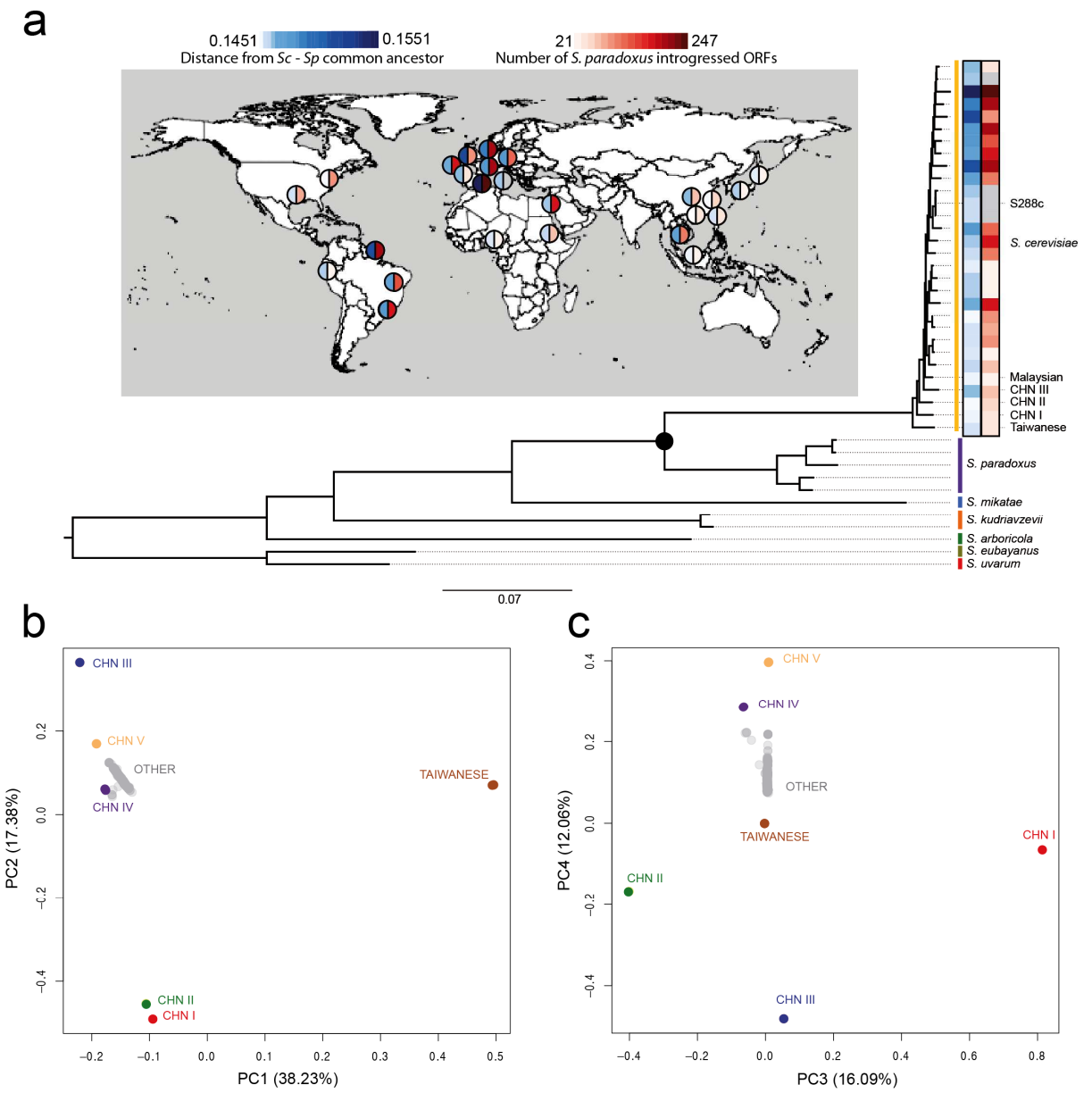


Figure 3: A single, shared Chinese origin of non-Chinese *S. cerevisiae* strains. **a.** ML rooted tree of the Saccharomyces sensu stricto complex, based on the alignment of 2,018 concatenated conserved genes. Heat-maps display the distance from the last *S. cerevisiae* - *S. paradoxus* common ancestor (white to blue) and the number of introgressed *S. paradoxus* ORFs (white to red). The map shows the geographical origin of the *S. cerevisiae* strains. **b-c.** Principal components calculated using only SNP genotypes of the highly diverged CHN I-V and Taiwanese isolates. The plot shows the position of all other sequenced isolates projected onto the resulting space. The non-Chinese lineages all project onto the same part of the space. This suggests a scenario with a single, shared Chinese origin of all non-Chinese *S. cerevisiae* strains, rather than a scenario where multiple, different Chinese lineages contributed to non-Chinese strains in different parts of the world.

Ploidy and level of aneuploidy vary across ecological origins

Ploidy changes and aneuploidy are not uncommon in yeast and this genomic plasticity is often described as a strategy for rapid adaptation to stress conditions in the environment[25–27]. A comprehensive survey of such genomic changes has never been conducted at a species-wide scale. We therefore assigned relative ploidy state to 794 isolates, to investigate the true extent of genomic variation within *S. cerevisiae*. We excluded 217 strains, which were genetically manipulated and no longer in their natural ploidy states. By measuring cell DNA content using high-throughput flow cytometry as well as examining the SNP frequency distributions (see Methods), we found 9 haploids, 694 diploids, and 91 isolates with a higher ploidy level (Figure 4a). *S. cerevisiae* has a haplo-diploid life cycle and therefore can propagate asexually as either haploids or diploids, but our results reveal that most of the natural *S. cerevisiae* isolates are in a diploid state (~87%). Polyploid isolates (3n, 4n or 5n) are not frequent (~11.5%) and fell within specific domesticated subpopulations as underlined by an enrichment in the 3 phylogenetically distinct beer clades (African, mosaic and ale beers; p-value < 4.2e-16), the mixed subpopulation containing the baker isolates (p-value = 2.7e-14) and the African palm wine (p-value = 0.0002), strongly suggesting an impact among some but not all strains of human-related environments on the ploidy level (Supplementary figure 11).

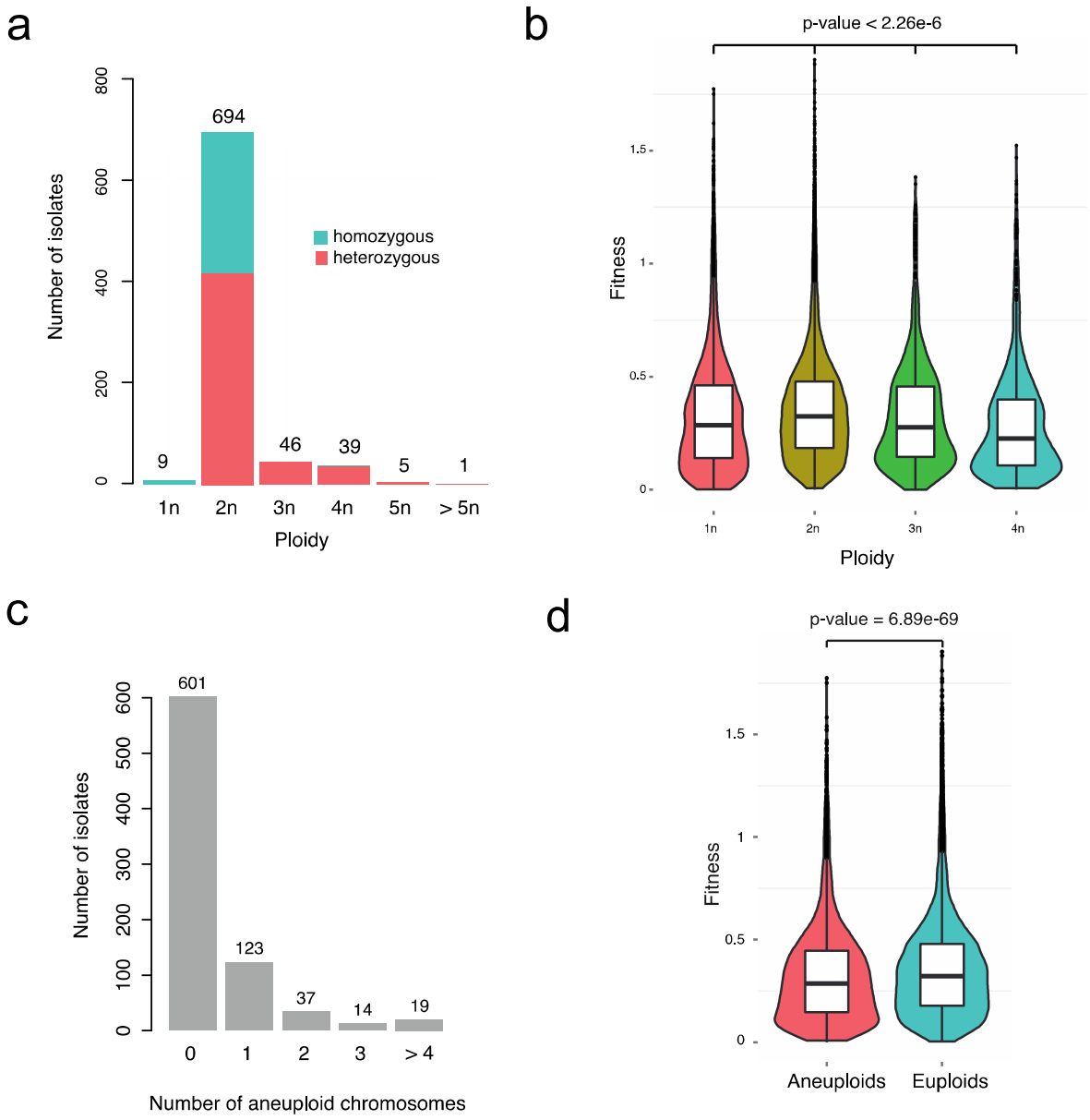


Figure 4: Ploidy and aneuploidy natural variation and growth fitness. **a.** Distribution of ploidy and fraction heterozygous isolates. **b.** Violin plot of the global growth fitness by ploidy. The p-value was calculated using a Mann Whitney Wilcoxon test and indicates that diploid isolates are globally fitter than individuals with other ploidy levels. **c.** Distribution of aneuploid chromosomes per individual across the sequenced isolates. **d.** Violin plot of the global fitness of aneuploid individuals with the global fitness of euploids. The p-value corresponds to a Mann Whitney Wilcoxon test and shows that there is a significant difference in fitness between the two categories.

We tested the impact of the ploidy on growth fitness across the genetic (971 isolates) and the phenotypic (34,956 fitness measurements) spaces of the species (Figure 4b). Collectively, a general fitness advantage was observed for diploid isolates compared to those characterized as haploid, triploid, or as higher ploidy (p-value ≤ 2.26e-6) (Figure 4b). This result supports a general mitotic fitness advantage of diploidy in yeast. In addition, by assessing individually in each condition, we revealed that there is a tendency towards better performance of diploids compared to the other ploidy levels for every tested trait, showing no major environment-specific effects (Supplementary figure 12).

We then assigned the presence of aneuploidies by combining coverage analysis and allele frequency distributions (see Methods). We precisely determined the copy number of each chromosome together with instances of segmental aneuploidies (Figure 4c, Supplementary figure 13a, Supplementary tables 1, 16). By examining the coverage for 1-kb non-overlapping windows across the 16 *S. cerevisiae* chromosomes, we identified a total of 342 aneuploidies affecting 193 isolates *i.e.* 24.3% of natural isolates. The most observed whole chromosomal aneuploidies are chromosome I, III and IX and aneuploidies are only weakly correlated with chromosomal size (Supplementary figure 13b). Interestingly, there is a strong enrichment of aneuploid strains in the sake (p-value = 2.9e-08), ale beer (p-value = 5.9e-06) and the mixed subpopulation containing the baker isolates (p-value = 3.6e-09) (Supplementary figure 14a). The latter categories also showing an enrichment for strains with a higher ploidy (3n, 4n and 5n).

Aneuploidy is therefore not uncommon in the *S. cerevisiae* species. However, the relationship between aneuploidy and fitness is paradoxical. Indeed, it is selected under a variety of environments but globally leads to a decrease of cellular fitness[27–30]. Based on the overall phenotypic landscape, we tested the general impact of aneuploidy (Figure 4d). There is a general mitotic fitness advantage of euploid versus aneuploid strains. We demonstrate and confirm that whole chromosomal aneuploidy has a general fitness cost.

A portrait of the *S. cerevisiae* pangenome

The extensive number of sequenced genomes provided an opportunity to determine the pangenome[31] *i.e.* the global set of ORFs present within the *S. cerevisiae* species, to derive their frequency and to infer their origin. Each of the 1,011 *de novo* assembled genomes was aligned to the S288C reference sequence and the material that could not be aligned (*i.e.* the non-reference material) was retained and annotated (see Methods). In addition to the 6,081 non-redundant ORFs present in the reference genome, an additional 1,716 ORFs were found in the non-reference material. Out of the 7,797 ORFs of the pangenome (Supplementary table 5), 4,940 are invariably present in all the 1,011 isolates and hence correspond to the core genome. By contrast, the remaining 2,857 ORFs have a variable frequency within the population and represent the accessory genome.

We then performed a comparison of the properties of the core and accessory genomes. Strikingly, copy number analysis highlighted a sharp distribution difference, with the majority of core ORFs present in a single copy per haploid genome, whereas variable ORFs show a higher frequency among both hemizygous (*i.e.* one copy per diploid genome) and multi-allelic (Figure 5a). We further investigated core and dispensable ORF differences by restricting the analysis to the subset of 6,056 non-redundant ORFs (corresponding to 4,940 core and 1,116 variable ORFs) present in the S288C reference genome, for which many layers of genomic and functional information are available. First, the distribution of the variable ORFs across the chromosomes is biased towards the subtelomeric regions (Figure 5b). This observation reinforces the idea that subtelomeres are hotspots of gene content variation[7,32]. In addition, this difference in the genomic spatial distribution is also related to a strong functional enrichment of variable ORFs for cell-cell interactions, secondary metabolisms, and stress responses (Supplementary table 6). Second, the ratio of non-synonymous (dN) to synonymous (dS) substitution rates (ω = dN/dS) are higher for the variable ORFs (p-value = 5.89e-15), indicating stronger selective constraints on the core ORFs (Figure 5c). Finally, the core genome is also characterized by much fewer loss-of-function mutations (LOF, which includes either nonsense mutations or frameshifts) compared to the variable ORFs (p-value = 6.45e-78, variance among set non significantly different) (Figure 5d), again reflecting differences in functional constraints.

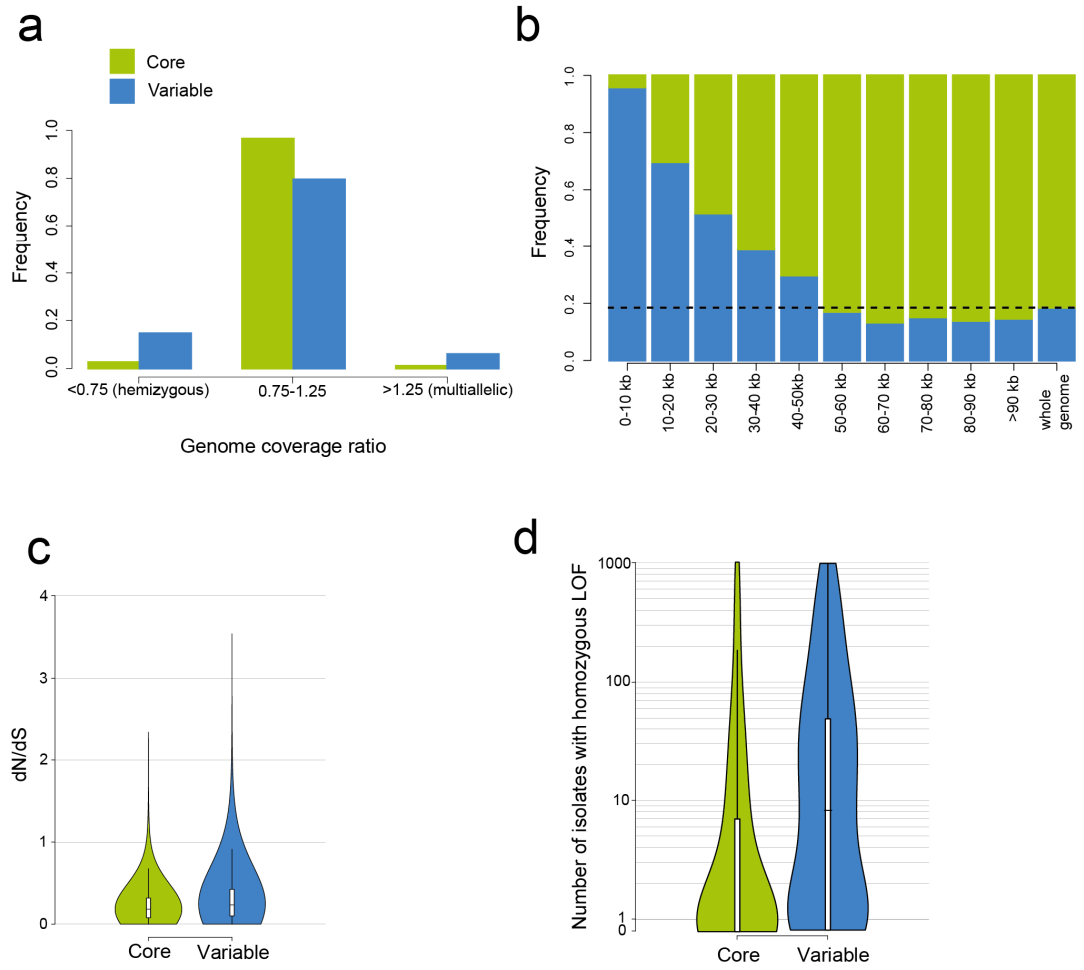


Figure 5: The *S. cerevisiae* pangenome. **a.** Copy number distribution for core and variable ORFs based on mapping coverage. Variable ORFs have a broader distribution with a greater percentage of both hemizygous and multiallelic ORFs. **b.** Frequencies of core and variable ORFs related to the distance from the nearest telomere (x-axis). The horizontal dashed line indicates the whole genome percentage of dispensable ORFs. The terminal 50 kb of the chromosomes are enriched in variable ORFs. **c**. dN/dS shows an overall stronger adaptive signature in the variable genes (p-value = $5.89e^{-15}$, 2-sided Mann-Whitney test). The essential genes have been removed to avoid bias caused by their increased frequency among the core ORFs. **d**. Logarithmic scale distribution of the number of isolates having Loss Of Function (LOF) mutations for the core and variable ORFs. The core genome is characterized by much fewer loss-of-function mutations compared to the variable ORFs (p-value = 6.45e-78, 2-sided Mann-Whitney test).

To trace the origin of the *S. cerevisiae* variable ORFs, we implemented a phylogenetic approach by comparing the evolution of each individual ORF with the genome phylogeny using 57 additional yeast species (Supplementary figure 15, see Methods). We first defined ORFs as ancestral segregating when their sequence similarity levels are consistent with the overall species phylogeny. This is the largest class of dispensable ORFs, representing almost 50% of the total (1,396 out of 2,857). We then identified 899 ancestral introgressed ORFs with the best identity matching other *Saccharomyces* species. Finally, 185 variable ORFs are the result of horizontal gene transfer (HTG) events from highly divergent Saccharomycotina species or even more distantly related species. Introgressed ORFs tend to replace *S. cerevisiae* orthologous ORFs suggesting that they integrated by homologous recombination[33], whereas HGT segments localize mainly within subtelomeric regions. These three classes have distinct

population frequency distribution (Figure 6a). The ancestral segregating ORFs are the most abundant within the population and usually missing in a few isolates or subpopulations. A large number of ancestral segregating ORFs (380) are dispensable due to allelic replacement by introgressed ORFs. By contrast, both introgressed and HGT ORFs are very rare and only present in given isolates or subpopulations. Interestingly, the vast majority of introgressed ORFs (885 out of 899) can be unambiguously traced to a *S. paradoxus* origin. All *S. cerevisiae* isolates carry at least one introgressed *S. paradoxus* ORF (median 26). Ubiquitous *S. paradoxus* introgressions indicate abundant gene flow via interspecific hybridization between these two closely related species. The remaining 14 ORFs were derived from *S. mikatae*, with most of them present in a single Japanese sake strain. Interestingly, no ORFs can be traced to *S. kudriavzevii* or *S. eubayanus*, despite hybrids with *S. cerevisiae* being frequently described in wine and beer fermentation. These results imply that lack of introgression from more divergent species is either not occurring because of higher sequence divergence or is selected against because of biological incompatibilities.

The amount of introgressed content is highly variable between the different clades (Figure 6b). Massive enrichment was found in the alpechin, bioethanol, Mexican agave and French Guiana subpopulations (p-value ≤ 1.34e-5), *i.e.* human associated niches where the two species might coexist and consequently represent interspecific hybrid zones (Figure 5b). There is a striking match between the geographic origins of the four *S. cerevisiae* clades and the ancestry of *S. paradoxus* introgressed ORFs (Supplementary figure 16). Introgressions from the American *S. paradoxus* subpopulation were found in the French Guiana, Brazilian bioethanol and Mexican agave clades whereas the alpechin lineage, mainly isolated from Europe, carries introgressions from the European *S. paradoxus* subpopulation (Supplementary figure 16). In contrast, the highly diverged wild Asian lineages are among the clades with fewer introgressed ORFs, consistent with secondary contacts with *S. paradoxus* occurring mainly after the out-of-China dispersal (Figure 3a).

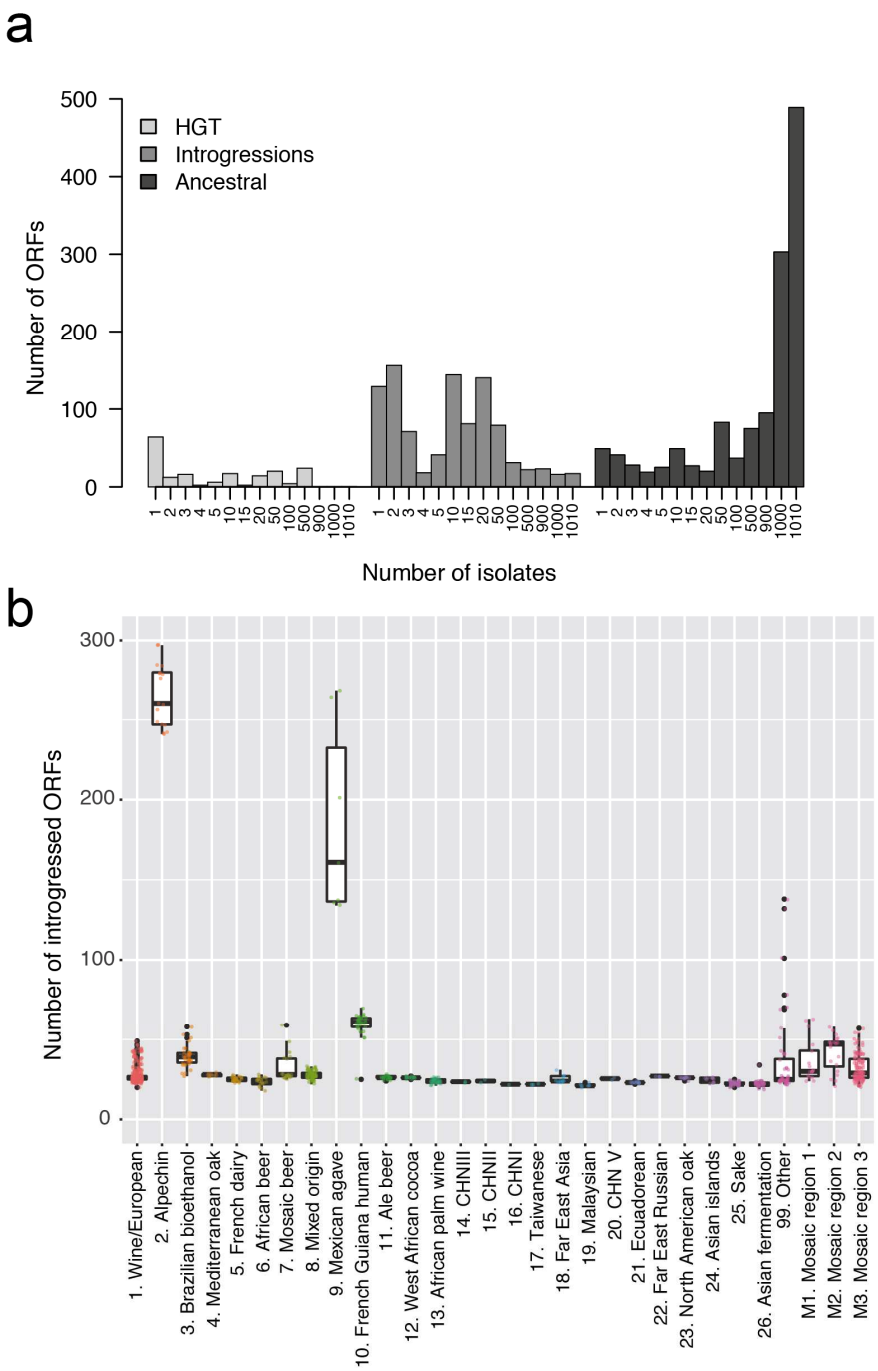


Figure 6: The *S. cerevisiae* pangenome. **a.** Different types of variable ORFs have sharp distribution differences. HGT ORFs are mostly found in single isolates. Introgressed ORFs are conserved within clades and therefore their distribution mimic the clade size. Ancestral segregating are the most abundant and their distribution reflects the loss, due to different mechanisms, in a small subset of strains. **b.** Distribution of introgressed ORFs in the different *S. cerevisiae* clades. Enrichment was found in the alpechin (median of 257 introgressed ORFs, p-value = 2.00e-12, variance among set non significantly different), bioethanol (median 39, 2-sided Mann-Whitney test p-value = 1.02e-09), Mexican agave (median 159, 2-sided Mann-Whitney test p-value = 1.34e-5) and French Guiana (median 61, 2-sided Mann-Whitney test p-value = 7.22e-16) clades.

We unequivocally traced the donor for 60 of the 185 HGT ORFs and found a clear enrichment for *Torulaspora* and *Zygosaccharomyces* species, which coexist with *S. cerevisiae* in fermentative environments. Consequently, some specific clades such as the Wine/European (p-value = 8.50e-05), Brazilian bioethanol (p-value = 2.73e-05), mosaic beer (p-value = 1.33e-05), and mixed origin (p-value = 8.24e-07) subpopulations are enriched for HGT events

(Supplementary figure 17). These specialized environments might select for the retention of HGT events. We identified 6 major HGT events (from 38 to 165 kb in size), among which three of them were partially characterized and likely adaptive for winemaking traits (Supplementary figure 18, Supplementary table 7)[34,35]. Analysis of these regions revealed that very few isolates retained large ancestral HGT events, instead most of the isolates retained smaller segments in complex patterns, consistent with multiple independent rearrangements leading to partial deletions of the ancestral HGT events (Supplementary figure 18)**.** For example, we discovered a strain isolated from the Carlsberg brewery that contains a massive ~165 kb HGT with at least 41 ORFs encompassing the previously described 65 kb region transferred by *T. microellipsoides*. Small relics of this larger event are detected in 186 additional strains derived from multiple independent rearrangements. Altogether, our data revealed that a large fraction of the yeast accessory genome has resulted from both introgression and HGT events. The widespread and pervasive nature of these events shows that they correspond to major evolutionary processes that continuously shape the genome of *S. cerevisiae* isolates.

Low levels of heterozygosity and extensive LOH are main features of *S. cerevisiae* genomes

S. cerevisiae is considered to be a highly inbred organism characterized by rare sexual cycles. Outcrossing is estimated to occur only about once every 50,000 - 100,000 mitotic generations in yeast species[36–38]. The frequency of outcrossing *vs.* inbreeding has a large effect on genome variation and evolution, particularly on the patterns of heterozygosity. Among the 794 natural isolates, a total of 505 isolates (~63%) are heterozygous and this proportion is similar when restricted to diploid isolates (415 out of 694, corresponding to ~60%) (Figure 7a), showing that entirely homozygous isolates are relatively common. However, the proportion of heterozygous *vs.* homozygous strains is very variable across the different subpopulations (Supplementary figure 19a). In addition, our data clearly show that heterozygous isolates are more prevalent in domesticated than in non-domesticated clades (Supplementary figure 19b), as previously observed at a smaller scale[39]. This difference seems to a general trend with some exceptions such as the wine clade (Supplementary figure 19a).

Heterozygous sites are distributed genome-wide across the genomes but they also exhibit large regions of loss-of-heterozygosity (LOH), *i.e.* regions that are completely homozygous. We generated an accurate genome-wide map of natural LOH regions (see Methods) (Figure 7a). LOH events range from 2 to 56 regions per strain and represent up to 80% of the genome in the sake isolates, for example. While variable across subpopulations, we observed an overall high LOH level with 25 regions covering 5.8 Mb per genome on average, *i.e.* almost half of its size (Figure 7c-d, Supplementary figure 20, Supplementary table 8). However, LOH events are not evenly distributed along the genome (Figure 7b). First, centromere-proximal regions exhibit low level of recombination initiation and consistently show very low rate of LOH

events. Second, some regions appear to be especially prone to them, such as the region upstream of the rDNA repeats on chromosome XII known to be sensitive to DNA damage[40].

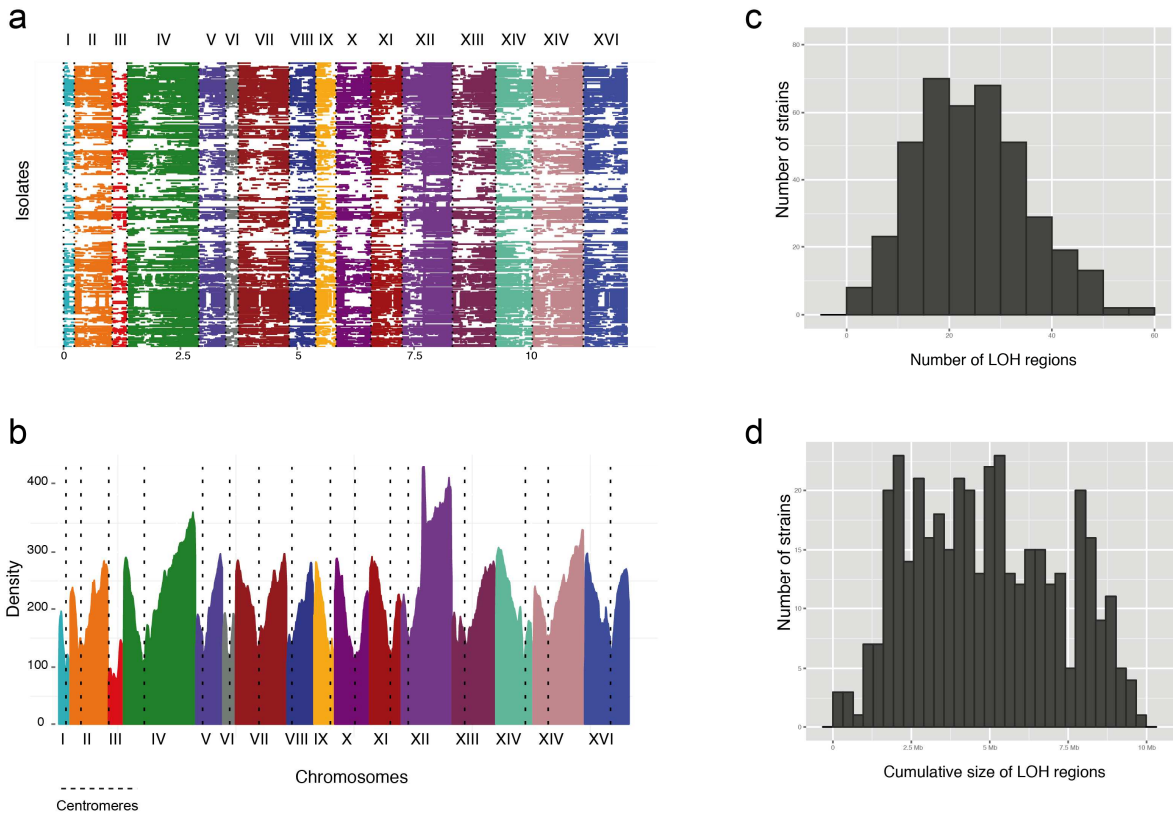


Figure 7: Landscape of loss-of-heterozygosity (LOH) events. **a.** Distribution of the LOH regions within the heterozygous natural isolates across the 16 nuclear chromosomes (I to XVI). Colored regions correspond to LOH events whereas white regions represent heterozygous parts of the genomes. **b.** Density of the regions under LOH across the genome within our population. Each color corresponds to a chromosome, and centromere locations are represented by dotted lines. **c-d.** Distribution of the number of regions under LOH as well as the cumulative size of the LOH regions per genome in our sample, respectively.

By masking the LOH regions, we precisely determined levels of heterozygosity across all of the genomes. This level is variable and ranges from 0.63 to 6.56 heterozygous sites per kb (Supplementary table 9). The distribution of the heterozygosity level in the population is bimodal with two clusters centered around 1 and 3.5 heterozygous sites per kb, respectively (Supplementary figure 21). This distribution reflects the variability observed across subpopulations with some of them showing a very low level (*e.g.* wine, sake, French Guiana, South East Asia) and some others a high level of heterozygosity (*e.g.* dairy, African beers, baker, ale beers) (Supplementary figure 21b-c). In fact, these observed patterns are most likely related to the outcrossing rate variation as a correlation between LOH and heterozygosity levels was observed (Supplementary figure 22).

The extensive and pervasive presence of LOH events across the sequenced collection is completely in agreement with the overall low outcrossing rate determined in *S. cerevisiae*[36]. Most of the detected events are thought to be the result of mitotic recombination initiated by double-strand DNA breaks[41]. However, recent reports showed that cells entering the meiotic program and then returning to mitotic growth can also lead to LOH[42,43]. Overall, our data support the idea that *S. cerevisiae,* which is mostly asexual, is characterized by clonal

expansion followed by diversification via LOH events. These events, underappreciated thus far, play a key role in *S. cerevisiae* life cycle and genome evolution, allowing for the expression of recessive alleles, as well as the production of novel allele combinations with a potential impact on the genetic and phenotypic diversity.

Genetic diversity, genome evolution, and selection across subpopulations

Our dataset allows for the first time to investigate evolutionary patterns across multiple subpopulations. To further characterize the differential evolutionary histories, we examined each of the 12 groups that contained at least 15 isolates. By determining the ploidy level and measuring the genome-wide levels of genetic diversity (π and θ_w), the minor allele frequency (MAF) spectrum and Tajima's D values, our results revealed distinct evolutionary histories at the nucleotide level across the various subpopulations (Supplementary table 10). Interestingly, analyses of the major human-related subpopulations provide strong evidences for multiple independent and specific domestication events for wine, beer, and sake isolates. Although *S. cerevisiae* beer isolates are polyphyletic as shown previously, they are mostly characterized by genomes with a higher ploidy level ($\geq$ 3n) as well as aneuploidy events. In addition, the genetic diversity ($\pi = 2.8 \times 10^{-3}$ on average for the three beer subpopulations) is very high and the number of heterozygous sites is elevated in these subpopulations (ranging from 17,807 to 34,203 heterozygous sites on average). Finally, LOH regions represent a small proportion of the beer genomes (11% on average). The independent but convergent domestication processes undergone by beer isolates are hence marked by genome level modification and high nucleotide variation.

By contrast, the wine and sake isolates are primarily diploids and monophyletic. Their genetic diversity is very limited (π values of 1×10^{-3} and 0.8×10^{-3}, respectively) compared to the other subpopulations, caused by distinct domestication events. Indeed, the genetic diversity is roughly 3-fold lower than the one found in the beer clusters. Interestingly, there is a larger reduction in π compared to θ_w in the wine subpopulation, resulting in an extremely negative Tajima's D value (-2.02). The wine cluster is characterized by a strong bias toward low-frequency polymorphisms. In fact, more than 95% of the polymorphic sites are associated with a MAF < 0.1. This observation clearly indicates a genome-wide excess of rare variants likely consistent with a history of population expansion. In addition, wine isolates harbour a low heterozygosity level (2,540 heterozygous sites on average) and extensive LOH regions (55% of the genome on average), strongly suggesting a low outcrossing rate. All these observations indicate that wine isolates experienced a population expansion after a domestication bottleneck. The impact of the domestication event on the sake genomes is very similar. Indeed, they have a low genetic diversity ($\pi = 0.8 \times 10^{-3}$), a low level of heterozygosity (2,322 heterozygous sites on average) and extensive LOH regions (82% of the genome on average). However, the sake subpopulation does not exhibit a bias toward low-frequency alleles

(Tajima's D value is equal 0.0481) reflecting their more recent origin[44] (see Supplementary note).

To detect a signature of selection within the different clades, we also calculated the divergence index F_{ST} along the genome in 2-kb non-overlapping windows, between each clade and its complementary part of the population. Genes located in the 0.25% right tail of the F_{ST} empirical distribution were considered as associated with significantly high values and were investigated as candidates for genes under selection (Supplementary figure 23, Supplementary table 11). Gene ontology (GO) term analysis revealed a slight enrichment (p-value = 0.024) in "substrate-specific transmembrane transporter activity" within the complete candidate gene sets, in line with the relevance of multiple compounds uptake such as sugar, polyamine or drugs during fermentation processes. Interestingly, the hexose transporters *HXT3* and *HXT7*, already described as playing a major role in alcoholic fermentation, were highlighted as under selection in some fermentation-associated subpopulations (sake, African beer, Asian fermentation and Brazilian bioethanol). Moreover, several genes involved in response to oxidative stress were identified (*HOR2*, *BLM10*, *YAP1*, *GRE2*, *AAD4* and *UTH1*) in the Mosaic beer, Wine/European, Brazilian bioethanol and French Guiana clades, consistent with the fact that early stages of fermentation process can lead to oxidative stress[45,46], against which the cells will activate protective mechanisms.

New insight into the genotype-phenotype relationship in *S. cerevisiae*

A major motivation for sequencing a large number of *S. cerevisiae* isolates was also to obtain new and deep insight into the genetic architecture of traits in yeast. For more than a decade, natural *S. cerevisiae* isolates have been a powerful model for the investigation of the genetic basis of natural variation using linkage mapping[47]. An impressive number of quantitative trait locus (QTL) mapping experiments were performed on a myriad of phenotypes leading to the identification of over 100 quantitative trait genes (QTG) and approximately 50 quantitative trait nucleotides (QTN)[47] (Supplementary table 12). We investigated the allele frequency associated with these polymorphisms underlying quantitative trait variation. Among a total of 36 well-characterized QTNs (32 missense mutations and 4 intergenic substitutions), 25 were found with a frequency lower than 5%, with 16 below 1%, and consequently undetectable using a genome-wide association strategy (Figure 8a). This strong bias toward rare alleles for these polymorphims is in line with the overall MAF spectrum. However, it is striking and interesting to note that 78% of the functional variants characterized in yeast correspond to rare variants with a large phenotypic effect. Consequently, this result highlights the fact that a significant fraction of the missing heritability of complex trait is likely due to rare variants in species with very low-frequency variants such as in yeast.

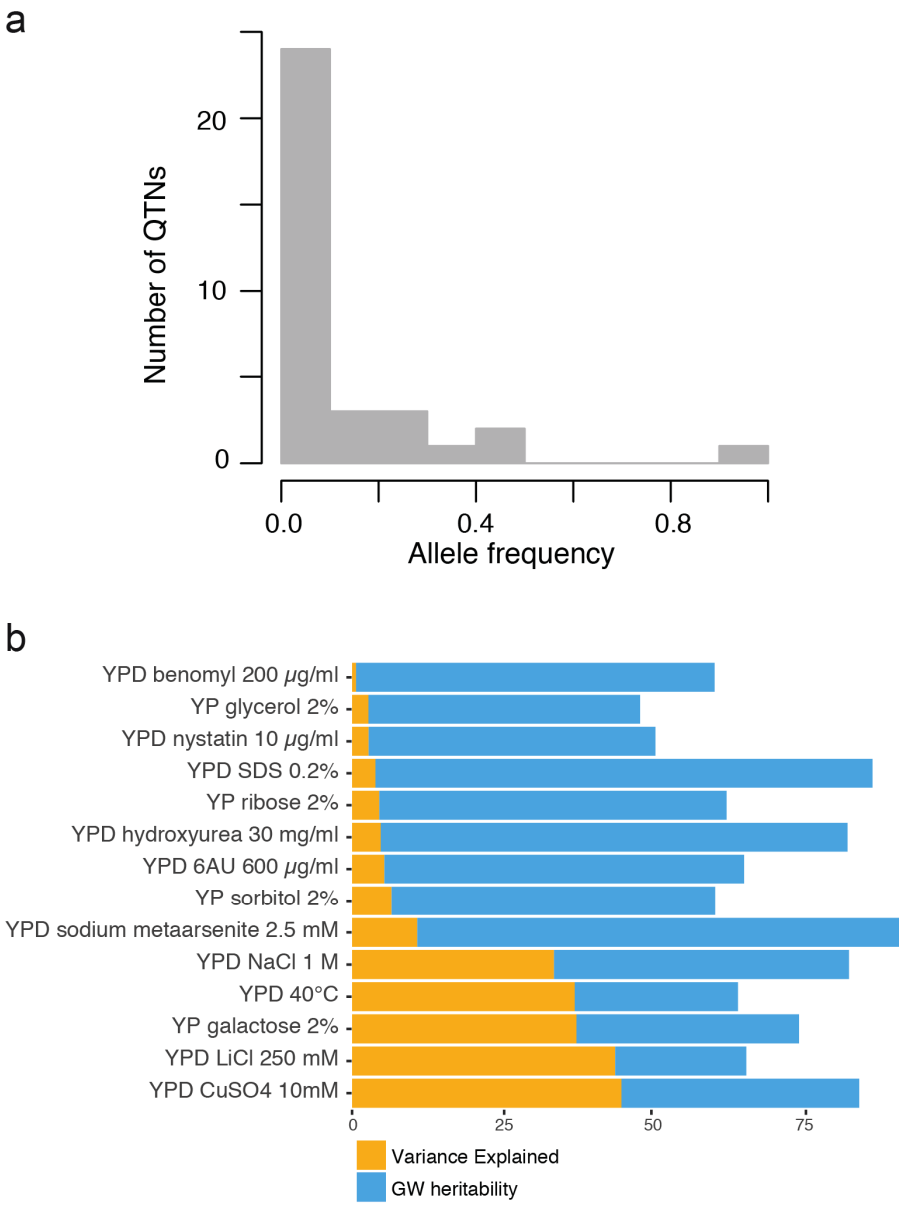


Figure 8: Genotype-phenotype relationship in *S. cerevisiae*. **a.** Allele frequency across the 1,011 genome sequences for mutations underlying quantitative trait variation from the literature. Notably, 78% of the QTNs have a frequency lower than 0.05. **b.** Narrow-sense heritability (blue) and phenotypic variance explained (yellow) for each phenotype showing a significantly associated variant.

The high genetic diversity (π = 3 x 10^{-3}), as well as the low linkage disequilibrium ($LD_{1/2}$ = 500 bp) (Supplementary figure 24) among *S. cerevisiae* isolates indicates that this species could represent a powerful resource for fine-scale mapping of associated loci. Our genomic dataset as well as the fitness values measured allowed us to assess the feasibility of genome-wide association studies (GWAS) on this widely studied model organism. We built a matrix of genetic variants that included comprehensive sets of SNPs, CNVs, and accessory ORFs not present in the reference genome. Successive quality control filtering was applied to retain only biallelic SNPs as well as CNVs with information for more than 1,000 strains (see Methods). Our matrix contains a total of 82,869 SNPs and 925 CNVs, which represents a dense map with one marker every 143 bp on average. In parallel, genome-wide heritability for each phenotype was estimated using FaST-LMM[11]. The narrow-sense heritability of the traits varies between

0.47 (glycerol 2%) to 0.90 (sodium arsenite), with an average value of 0.69 (Supplementary table 13). In addition, a principal components analysis of the phenotypes reveals that there is no clustering of the isolates according to the defined clades (Supplementary figure 25-26, Supplementary table 14), showing that traits are usually not stratified by the defined subpopulations that could lead to a high false positive rate. Interestingly, this analysis also points out that trait variation is not defined by population history contrary to what was previously observed using a smaller number of isolates[48].

We performed mixed-model association analysis of the growth traits with FaST-LMM[11], using the isolates growth fitness traits. We quantified the genomic inflation factor (λ) for each association, to determine if our model was accurately accounting for population structure in our dataset. Our λ values were all close to 1 with the highest (1.013) for growth using ethanol as carbon source (Supplementary figure 27). We also estimated trait-specific p-value thresholds by randomly permuting the phenotypes. In total, 35 variants were significantly associated with 14 conditions, with an enrichment and high association scores for CNVs (22 CNVs *vs.* 13 SNPs) (Supplementary figure 28, Figure 9a and Supplementary table 15). In addition, 4 of the detected variants are linked to variable ORFs, which are not present in the reference genome. Phenotypic variance explained was estimated by running a new association with a similarity matrix containing the significantly associated markers. The variation explained by the null model is then the estimation of the proportion of phenotypic variance explained by these markers (Supplementary table 15 and Figure 8b). For five of the tested traits, the phenotypic variation explained is surprisingly greater than 25% (Figure 8b) and CNVs explained larger proportions of trait variance compared to SNPs (Figure 9b). The high number of associated CNVs is likely due to the nature of these variants. It is known that these variants can contribute to a large amount of phenotypic variation, and that they can interfere with regulating regions or coding sequences of diverse genes[49]. These results emphasize the need to consider all possible types of genetic variants. Moreover, the biological interpretation is more straightforward for CNVs, as they can increase or decrease expression of sensitive genes. As an example, we found the *CUP1-2* gene strongly associated with resistance to copper sulfate (p-value = 4.85e-44) (Supplementary figure 28 and Supplementary table 15). Amplification of this locus plays a role in the resistance to high concentrations of copper and cadmium[50]. The variation in gene copy number explains alone 44.5% of phenotypic variation. Another example consists of the more complex phenotype of resistance to sodium arsenite showing an associated region with multiple CNVs (Supplementary figure 28 and Supplementary table 15). This region contains the three *ARR* genes (*ARR1*, *ARR2* and *ARR3*), which are known to play a key role in the resistance to arsenate, and explains 10.78% of the phenotypic variance[51].

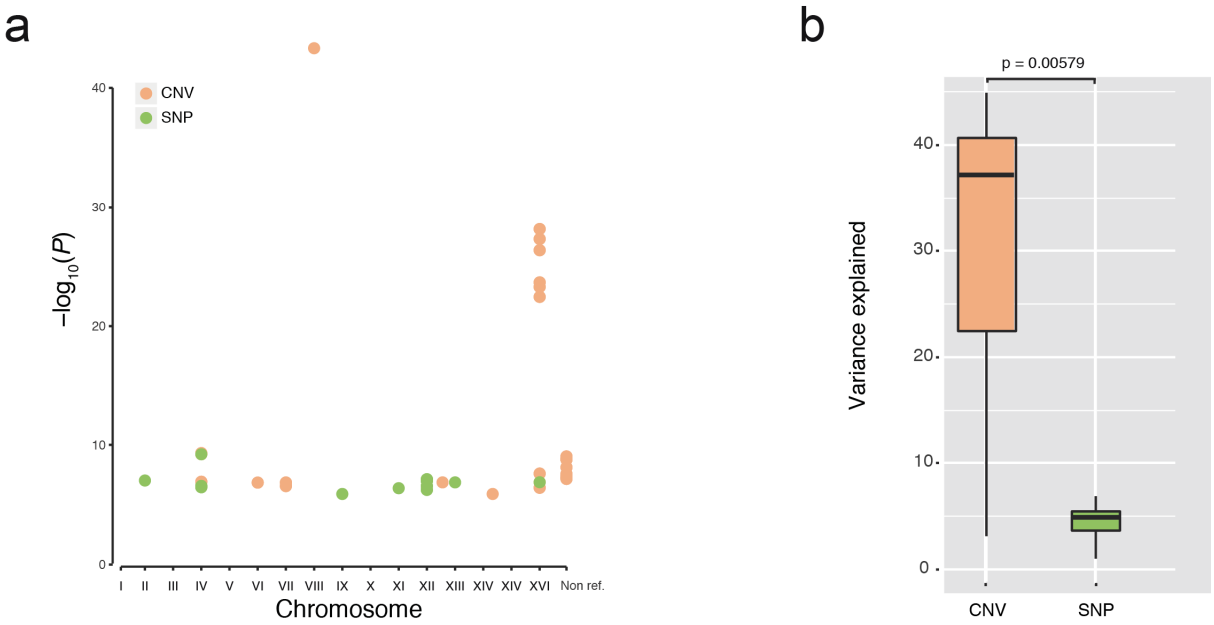


Figure 9: Characteristics of the identified variants. **a.** P-values of variants that were significantly associated with a condition across the 16 chromosomes and the variable ORFs not present in the reference genome. Variants are colored by types *i.e.* SNP or CNV. **b.** Variance explained by CNVs and SNPs associated with traits. Interestingly, association scores as well as variance explained are higher for CNVs compared to SNPs. The p-value corresponds to a Mann Whitney Wilcoxon test.

Conclusion

With the completion of the whole genome sequencing of 1,011 natural isolates, plus the accompanying phenotyping efforts, we have currently obtained one of the best understanding of the natural genetic and phenotypic variation of any eukaryote model system to date. This resource revealed undescribed evolutionary history, driving forces of genome evolution as well as new insights into the genotype-phenotype relationship. Our study constitutes the first attempt to perform GWAS in *S. cerevisiae* and provide a population genomics resource to a scale matching other model organisms[1,3]. Interestingly, the difference between the estimations of genome-wide heritability and phenotypic variance explained by associated variants gives an overview of the extent of the missing heritability for each trait[52,53]. In this context, our genome-wide association analyses, including an exhaustive catalog of genome content and CNVs present in the 1,011 genomes, highlighted the overall importance of these genetic variants on the phenotypic diversity. In fact, CNVs explain a larger proportion of the phenotypic variance and have higher effects compared to the SNVs. Moreover, much of the SNVs are very low-frequency variants with a trend like the one observed in the human population[54], raising the question of the impact of rare variants on the phenotypic landscape within a population and consequently on the missing heritability. By examining the allele frequency of a large number of previously reported QTNs, we found that a majority of them are rare (78% of the them have a frequency lower than 0.05), with a large phenotypic effect. However, an overview of the impact of the rare variants on the phenotypic landscape of a natural population would be now essential. This collection of genetic and phenotypic variants will hence enable new functional approaches in a powerful model system.

Supplementary material

Supplementary tables

Supplementary tables are available at **http://bit.ly/2urEBrO**

Supplementary figures

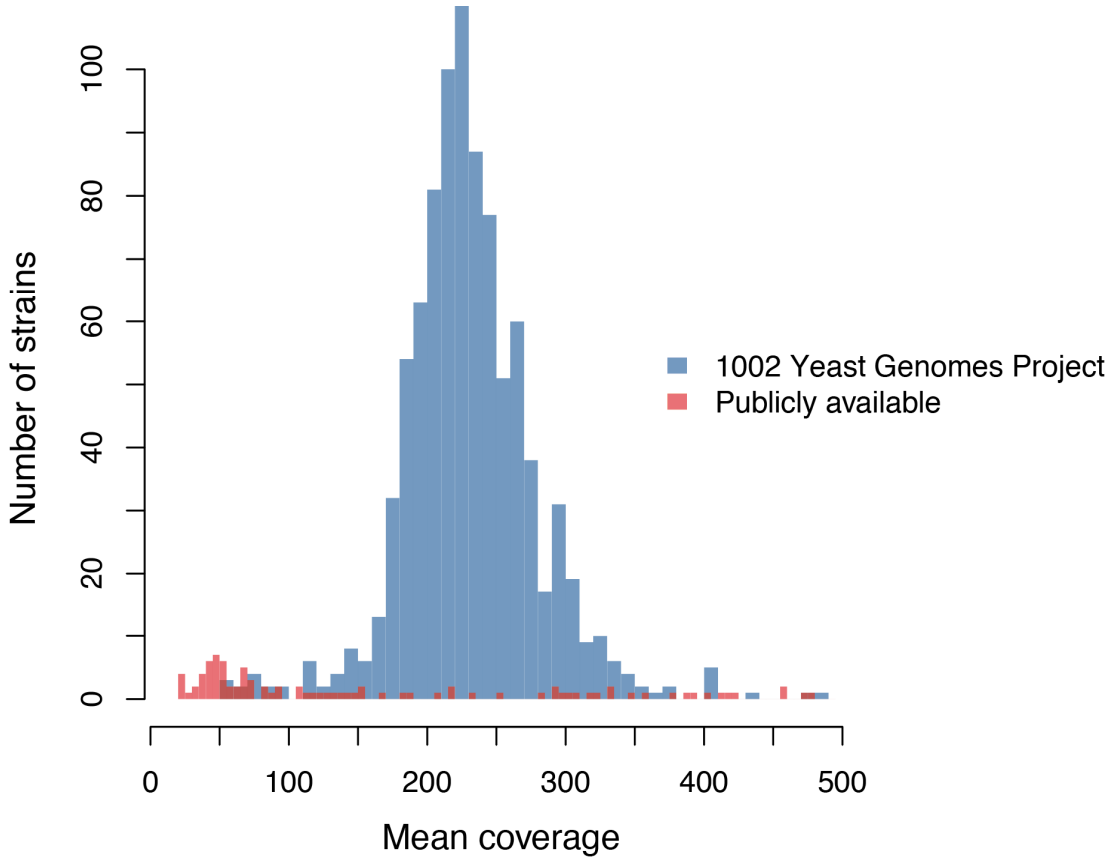


Figure S1: Genome sequencing coverage of the 1,011 *S. cerevisiae* isolates. Sequencing coverage distribution of the 918 sequenced isolates in the frame of this project (blue) and for the 93 external strains (red)

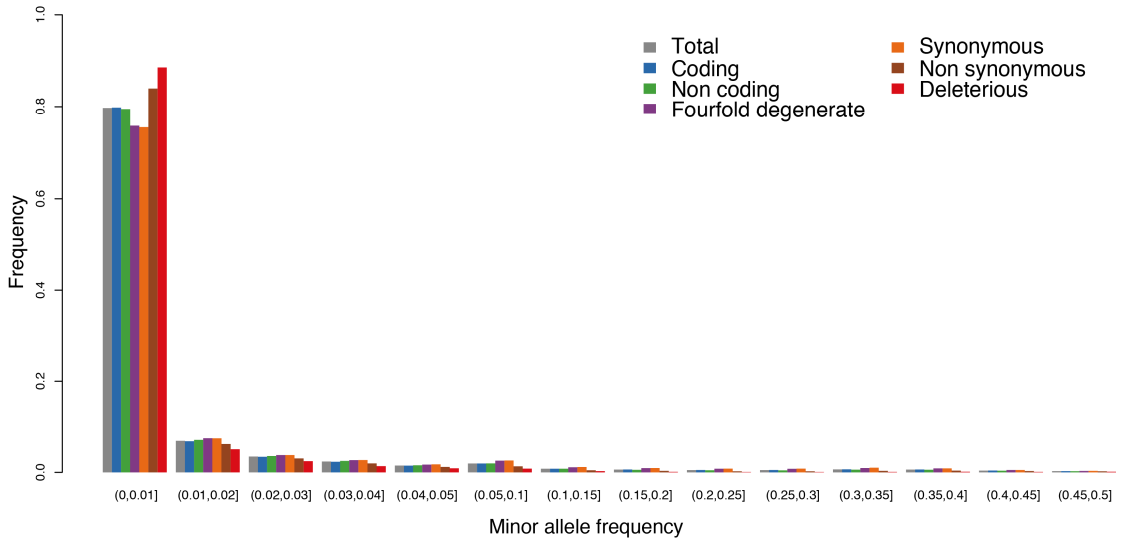


Figure S2: Frequency spectrum if SNPs in the 1,011 genomes. Minor allele frequency (MAF) of polymorphisms was determined for SNPs with different genomic location and functional annotation.

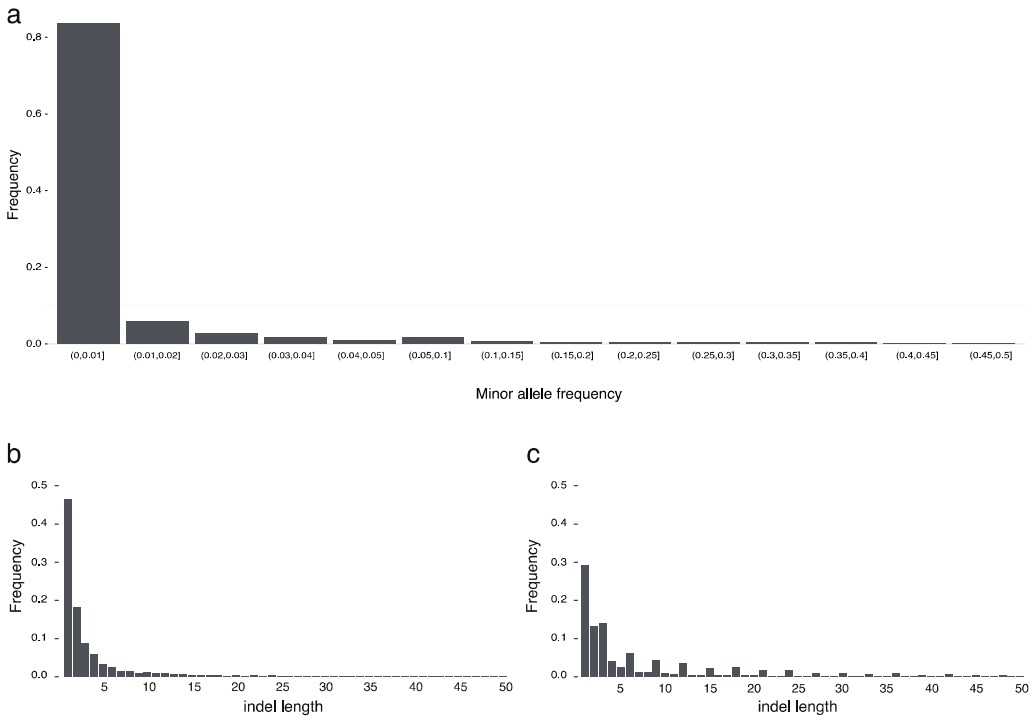


Figure S3: Small indels across the 1,011 genomes. A total of 5,438,463 indels (up to 50 bp) were detected across the 1,011 genomes. a. Frequency spectrum of small-scale indels. Distribution of the indels size in non-coding (b) and coding (c) regions.

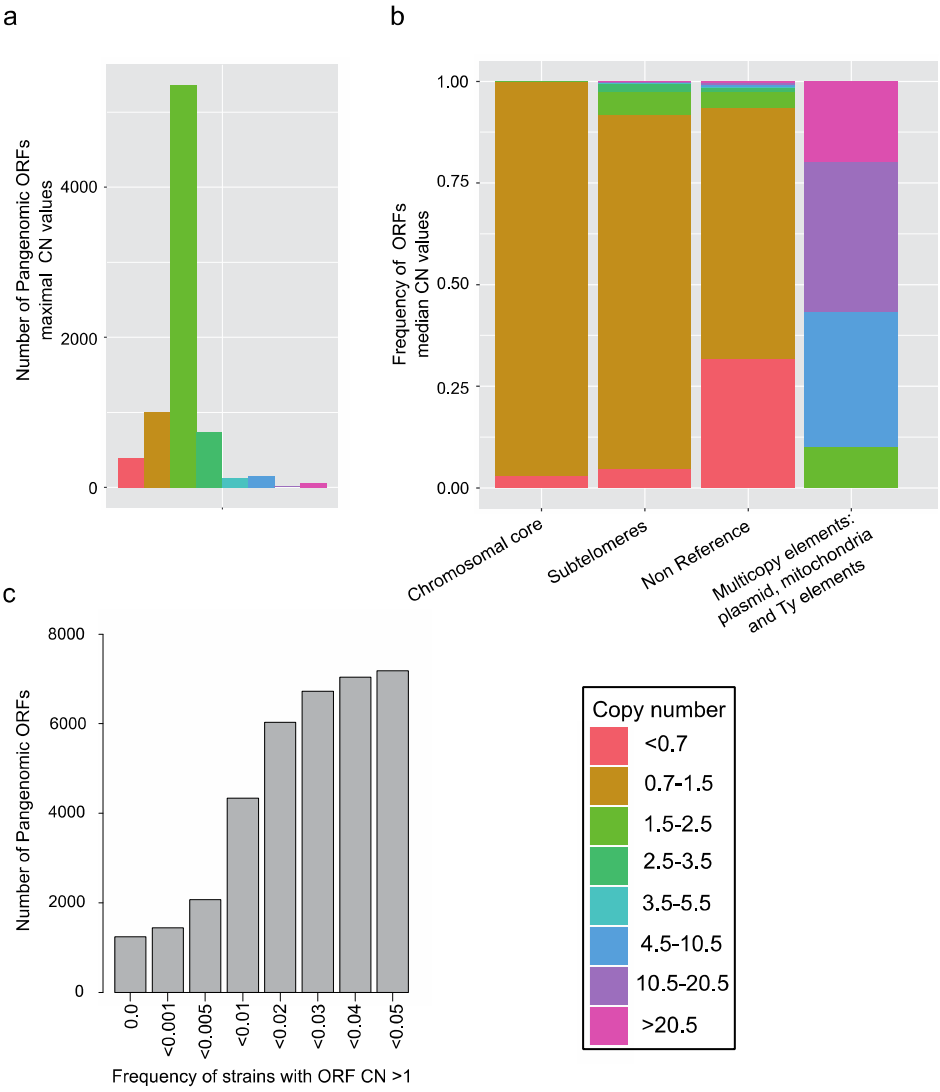


Figure S4: CNV distribution across isolates and ORFs. a. Histogram of the maximal CN for each pangenomic ORF. Most of the ORFs never exceed CN=2 (N=7,042). b. Distribution of median CN values across a different set of ORFs. Plasmids, mitochondria and repetitive elements show the highest CN, while ORFs located in the core chromosome have virtually no CNV. c. CNV occurs in the majority of the ORFs, with only 1,242 ORFs demonstrating no increase of CN in any strain. However, most of ORF CNVs (N:7,182) occur at low frequencies (less than 5% among the isolates)

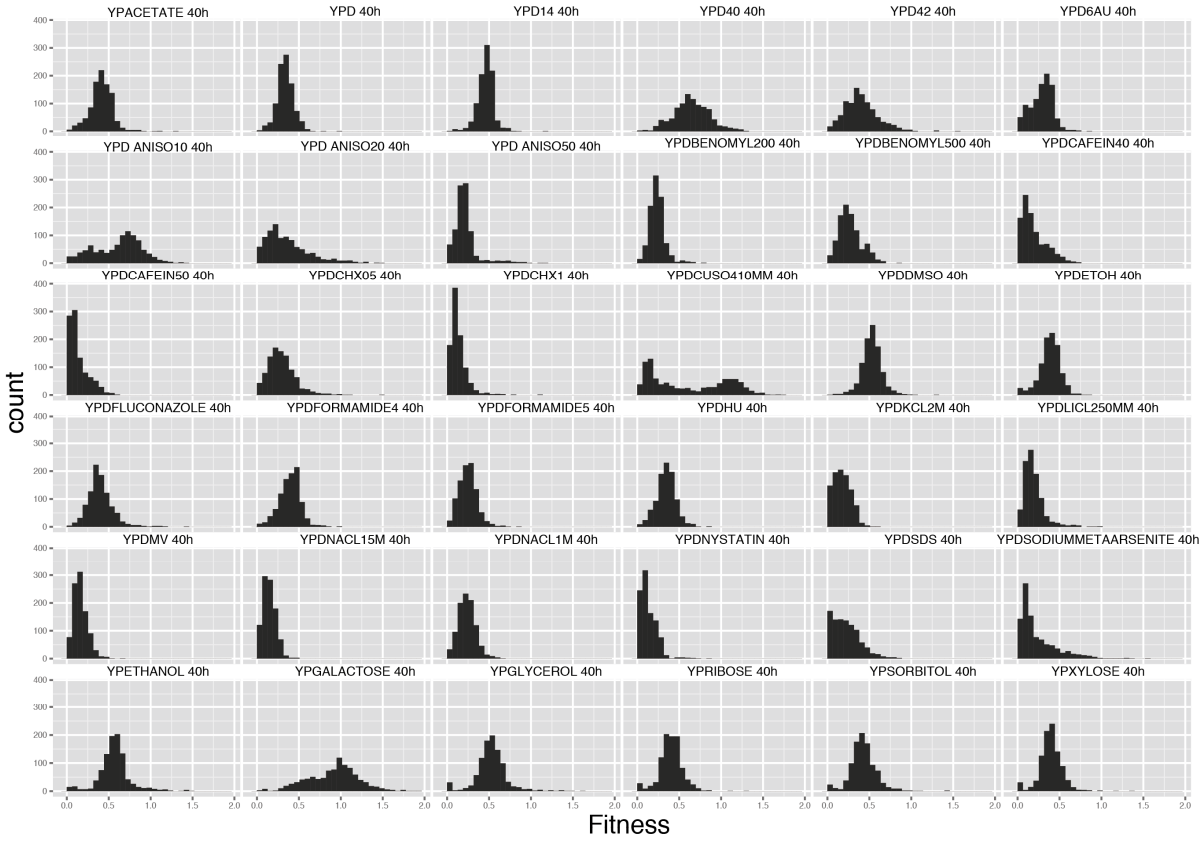


Figure S5: Fitness distribution patterns. Distribution of the 971 phenotyped isolates for each trait. Most of the traits vary continuously across the population and are complex, whereas a small number of them show a bimodal distribution characteristic of a Mendelian inheritance (*e.g.* $CuSO_4$).

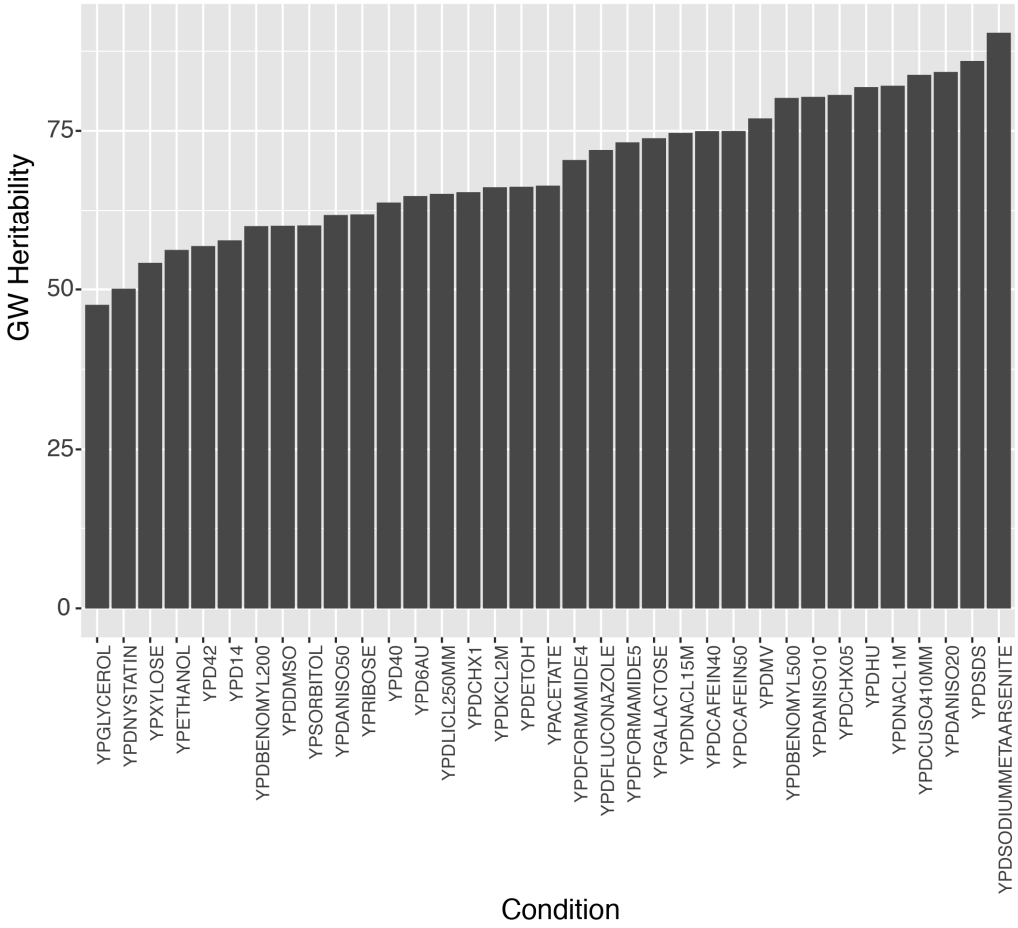


Figure S6: Narrow-sense heritability of growth trait. The genome-wide heritability (GW heritability) ranges from 0.47 to 0.9 with an average of 0.69

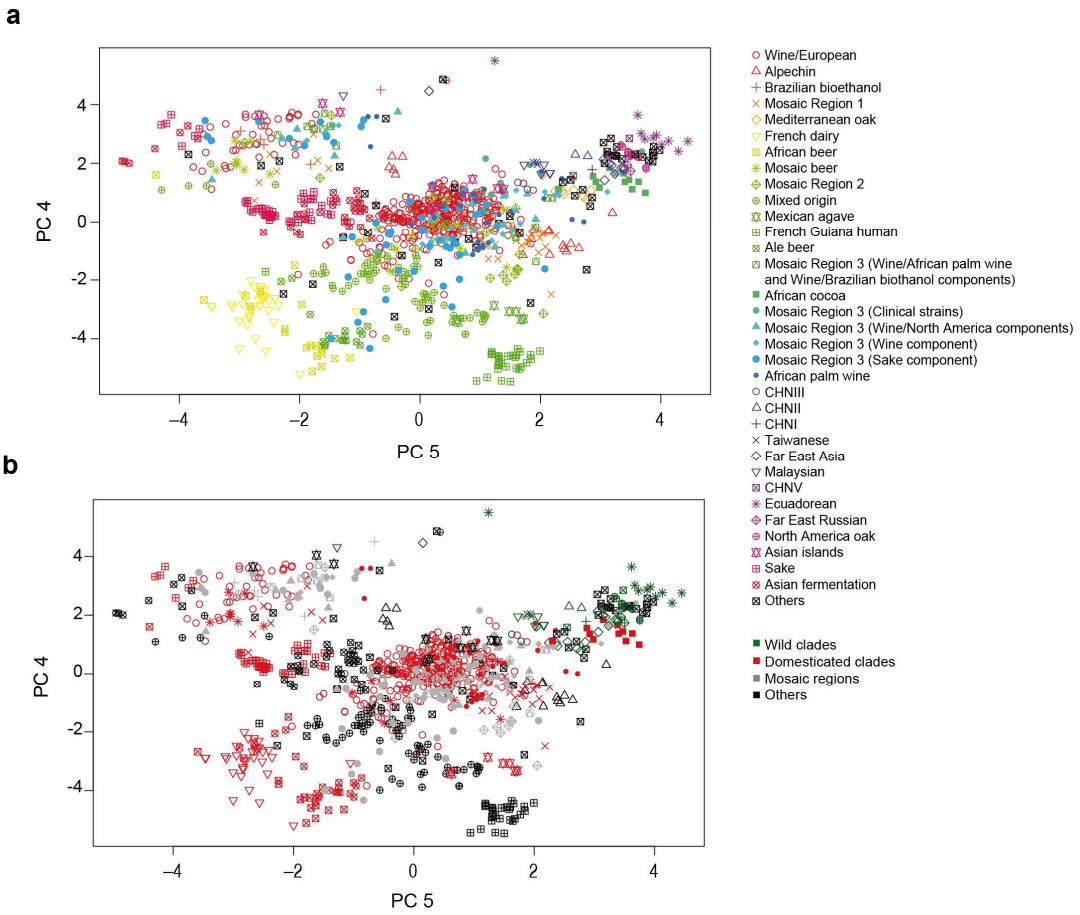


Figure S7: PCA based on presence of variable ORFs. The plots show the position of each strain onto the 4th and 5th component of the PACA based on presence/absence of variable ORFs. The 1st, 2nd and 3rd components (not shown) describe, respectively, the strain order in the phylogenetic tree, the introgressions in the Alpechin clade and the introgressions in the Mexican agave and French Guiana clades. **a.** Colors and symbols indicate the strain clades. **b.** Colors indicate wild and domesticated clades while the symbols are the same as in the panel a. All of the wild clades, including the Mediterranean oak, group together and are in the top right corner. The only domesticated clade that clusters with the wild isolates is the mosaic African coca beans fermentation.

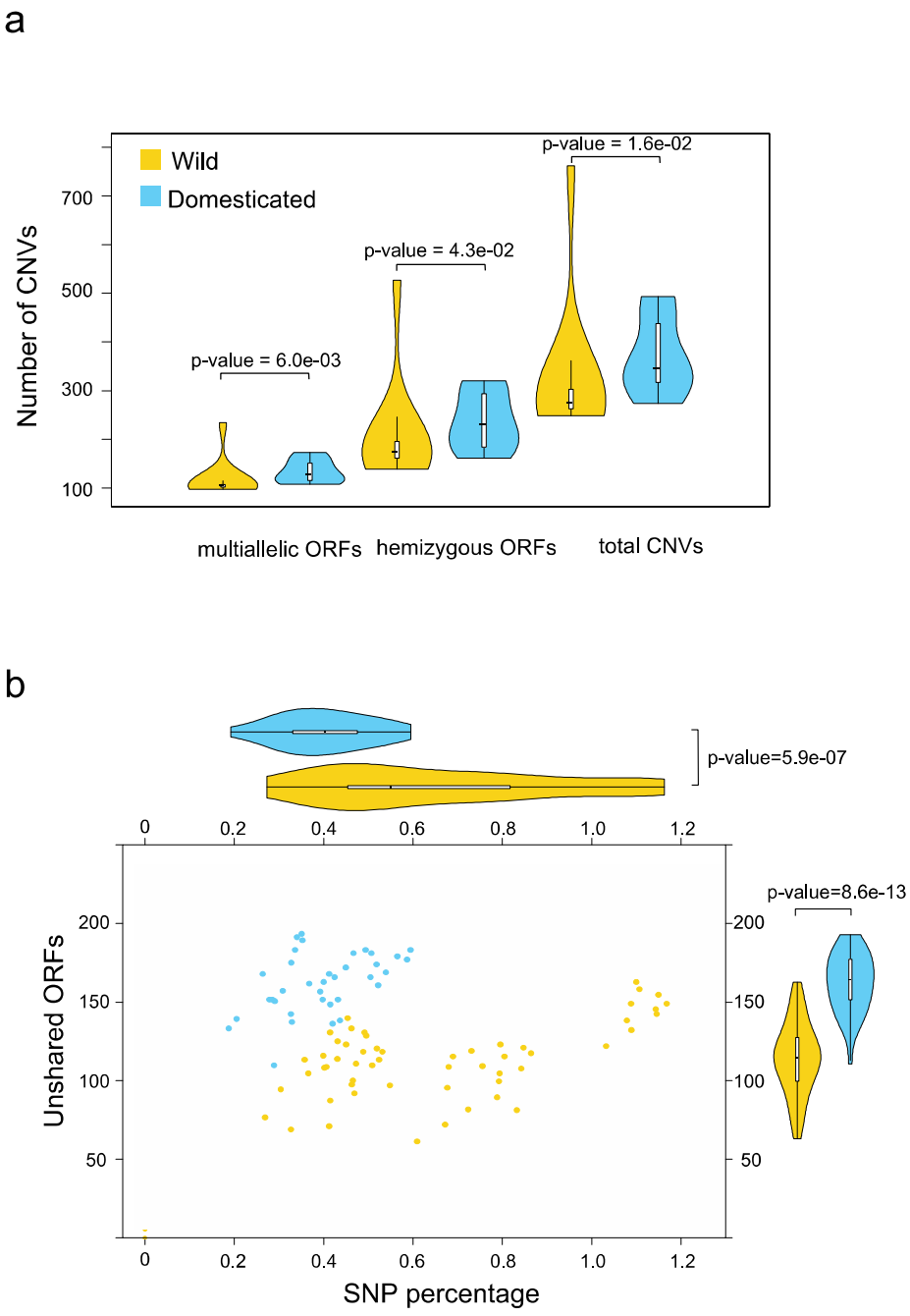


Figure S8: Differential genome evolution between wild and domesticated clades. a. Wild clades have fewer CNVs than domesticated clades (median 275 versus 346, 2-sided Mann-Whitney test p-value = 1.6e-02), either multiallelic (105 versus 127.5, 2-sided Mann-Whitney test p-value = 6.0e-03) or hemizygous ORFs (173 versus 230, 2-sided Mann-Whitney test p-value = 4.3e-02). To avoid bias due to aneuploidy or polyploidy, only euploid diploid isolates have been taken into account, although analyzing all of the strains give similar results. b. Pairwise comparisons between domesticated (blue) and wild (yellow) clade. The domesticated lineages have more pairwise differences in their ORFs content than the wild lineages, despite minor pairwise SNP differences. The violin plots of the distributions of the distance are shown on top for the SNPs distance (2-sided Mann-Whitney test p-value = 5.86e-07) and on the right for the number of unshared ORFs (2-sided Mann-Whitney test p-value = 8.59e-13).

Figure S9: *S. cerevisiae* population structure. The underlying population structure inferred using the software ADMIXTURE using a varying number of subpopulation (from 2 to 17). Strains are ordered based on the phylogenetic tree. The diversity among the mosaic groups of strains is clear with a much higher ancestral complexity of the Mosaic region 3.

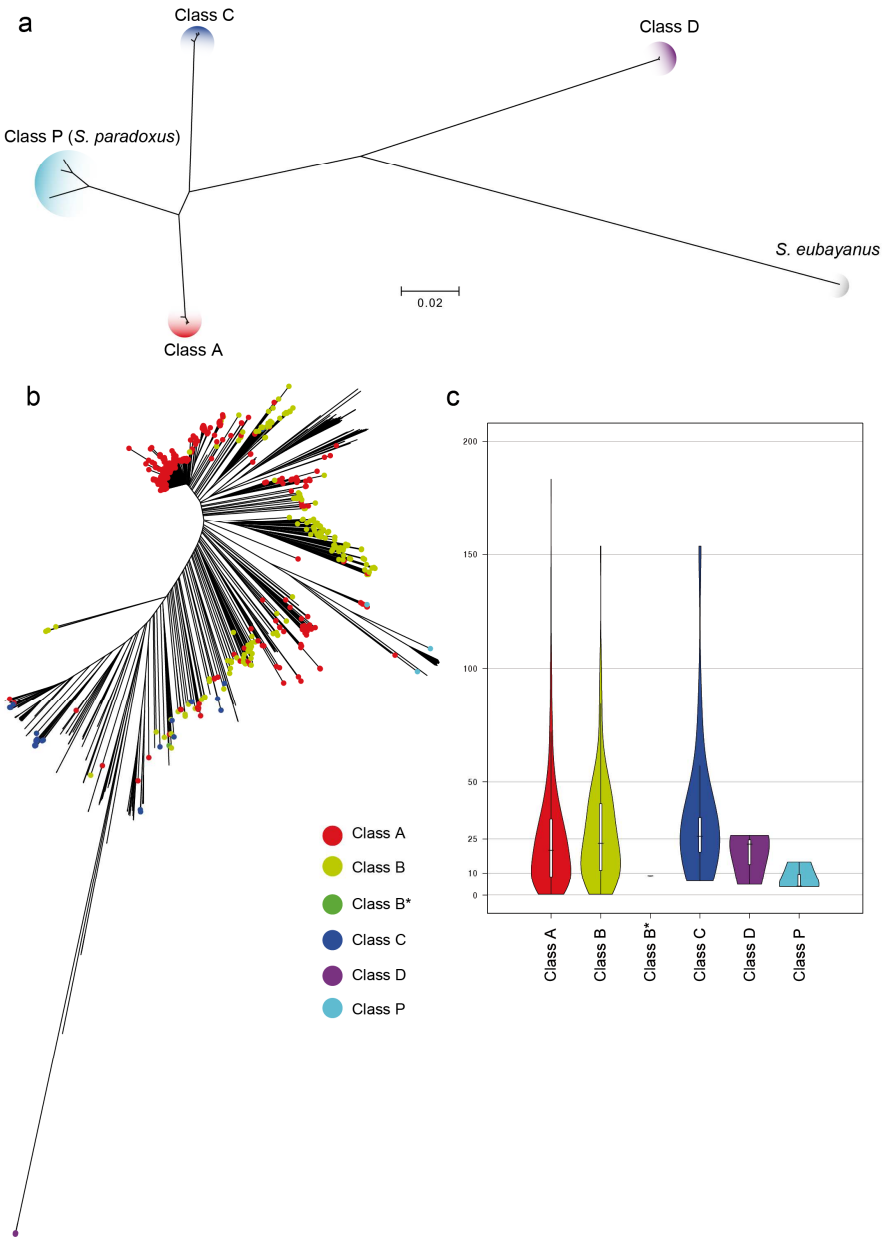


Figure S10: Population genomics of the natural 2µ plasmid. **a.** A NJ tree built using representative versions of the plasmid sequence variants. The three most similar classes (A,C and P) have about 10% sequence divergence from each other (pairwise difference A – C 8%, A – P 10%, C – P 11%). Class D is only found in the most diverged lineage (Taiwanese), separated by about 20% from all other classes, and might be derived from a horizontal transfer from a not-yet characterized *Saccharomyces sensu stricto* species. The classes B and B* are not included since they are recombinant forms of class A and C. In particular, class B retains *FLP1* and REP2 genes from class A and *RAF1* and *REP* genes from class C, while class B* is the reciprocal recombinant form, which we observed for the first time. **b.** Distribution of the different plasmid classes in the sequenced strains. Class A, B, B*, C, D and P are present respectively in 463, 171, 1, 26, 3, and 3 isolates. **c.** Broad range of plasmid copy number in the different sequence variants.

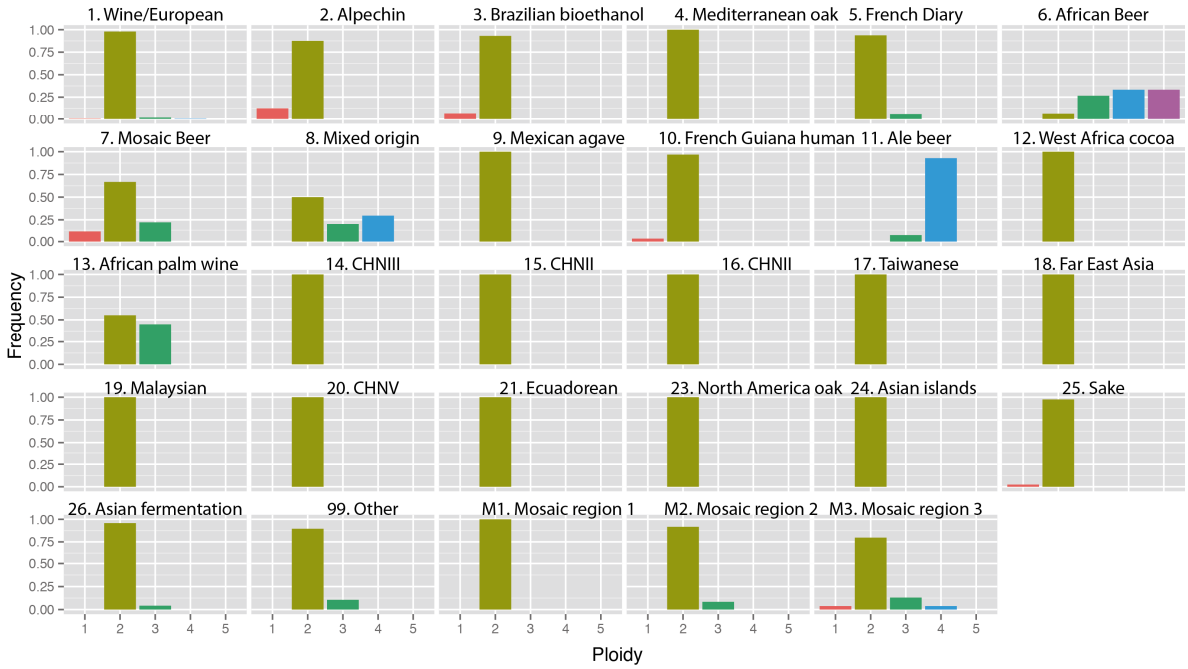


Figure S11: Ploidy level variation across the identified subpopulations. The frequency of the ploidy level was represented for each of the 26 subpopulations as well as for the 3 mosaic groups. Most of the subpopulations have only diploid isolates. However, an enrichment for isolates with higher ploidy (>2n) was found for the 3 phylogenetically distinct beer clades (African, mosaic and ale beers, χ^2 test, p-value < 4.2e-16), the mixed subpopulation containing the baker isolates (χ^2 test, p-value = 2.7e-14) and the African pal wine (χ^2 test, p-value = 0.0002).

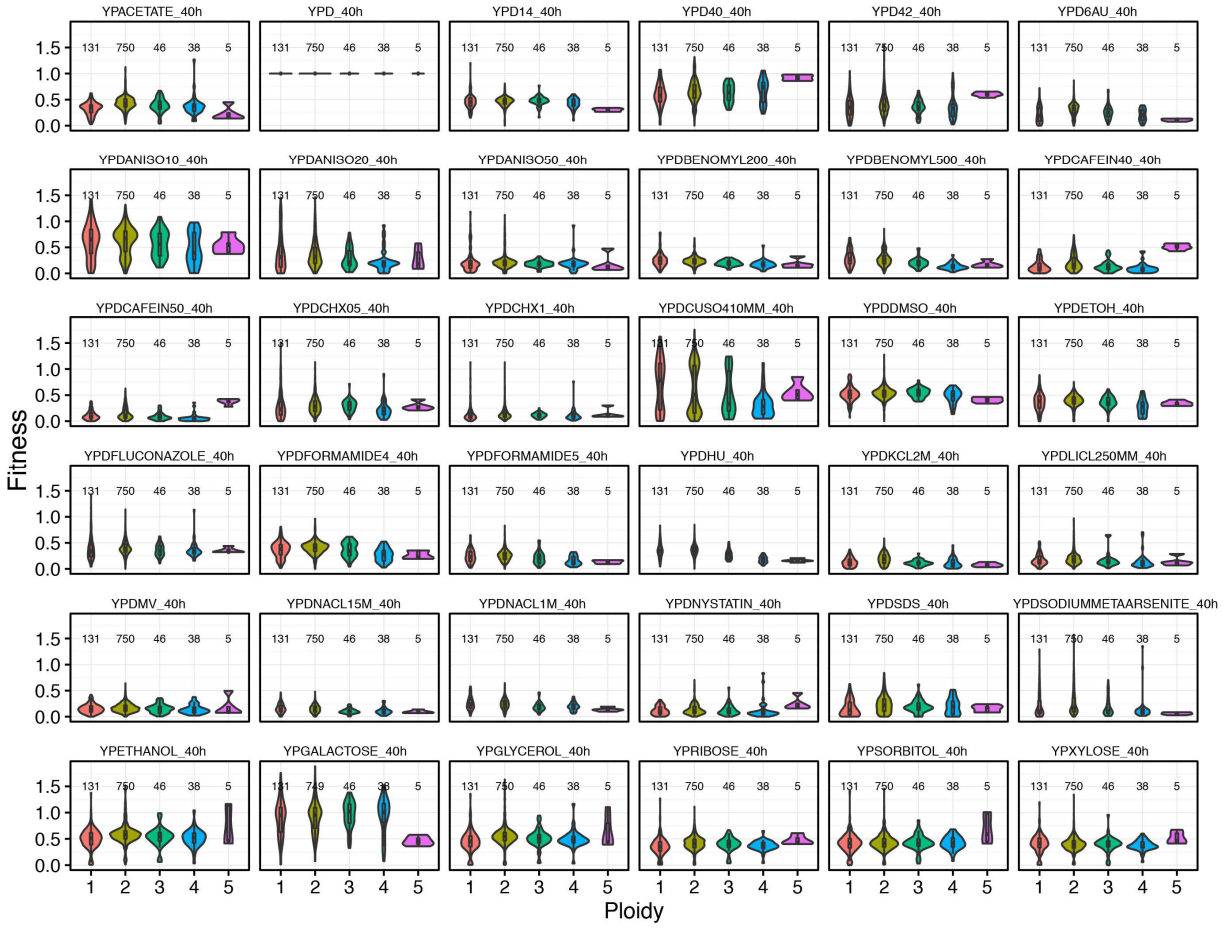


Figure S12: Fitness and ploidy level by condition. For each condition, the fitness distributionwas represented by ploidy level. Most conditins follow the general trend *i.e.* demonstrate mitotic advantage of diploidy.

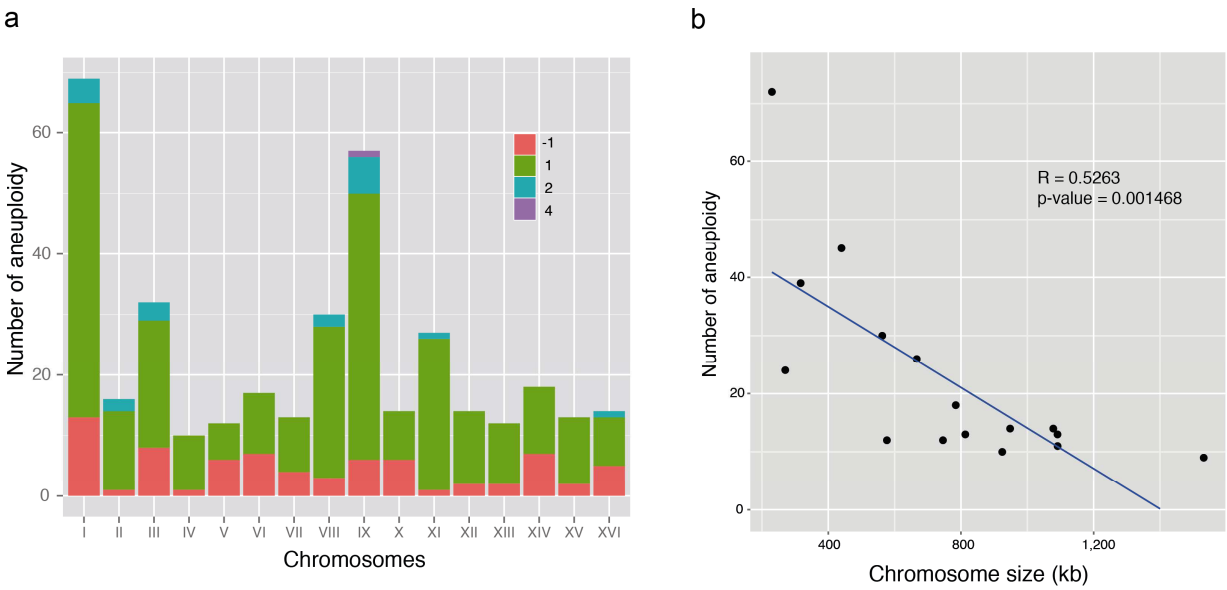


Figure S13: Genome-wide distribution of aneuploidy events. **a.** Distribution of the aneuploidy events by chromosome. The color of the bar plots refers to the number of sub/supernumerary chromosomes. **b.** Number of aneuploidy events affecting each chromosome plotted against the size of the chromosome.

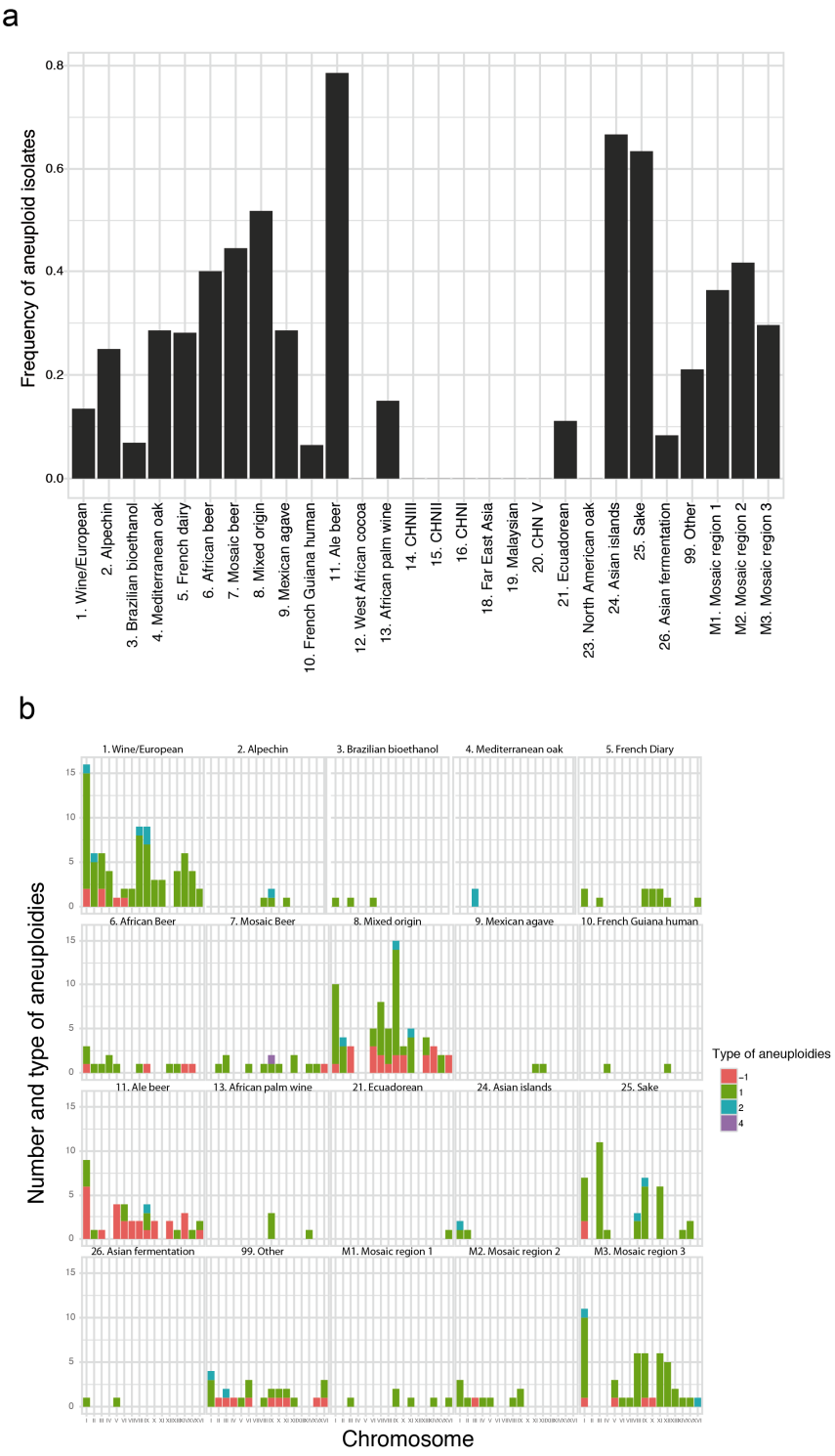


Figure S14: Aneuploidy level and subpopulations. **a.** Proportion of aneuploid isolates for each of the defined subpopulations. An enrichment of aneuploid strains is observed in the sake containing (χ^2 test, p-value = 5.9e-06) and the mixed subpopulation containing the baker isolates (χ^2 test, p-value = 3.6e-09) subpopulations. **b.** Genome-wide distribution of the aneuploidies by clade. The color of the bar plots refers to the number of sub/supernumerary chromosomes.

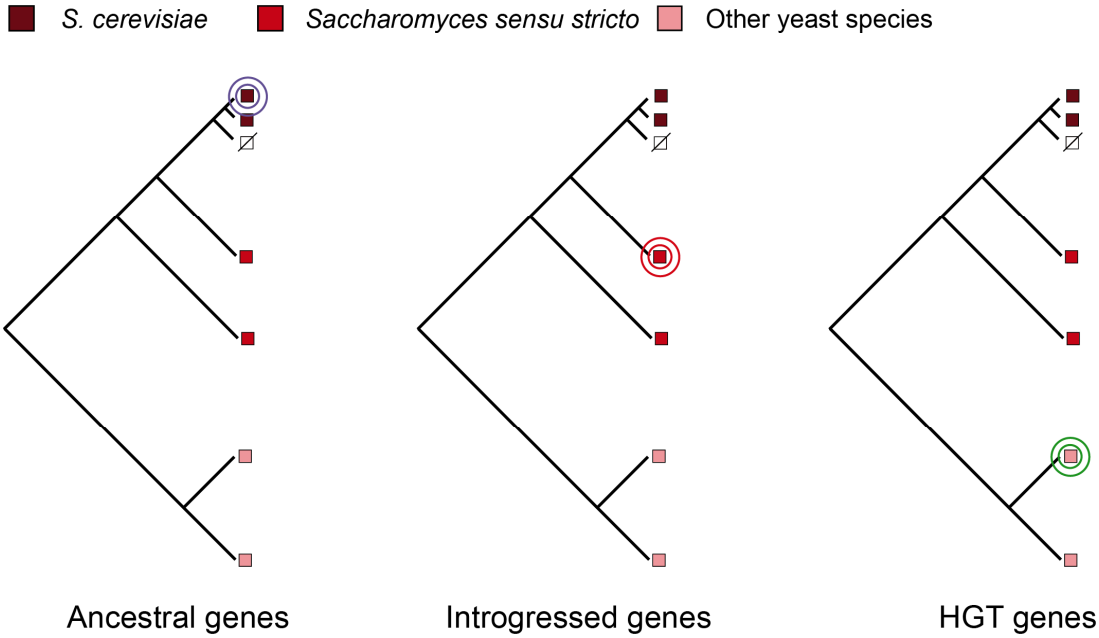


Figure S15: Origin of variable ORFs. We catalogued variable ORFs as ancestral segregating, introgressed or horizontal gene transfer (HGT) according to their phylogeny. ORFs with their closest orthologs in other *S. cerevisiae* isolates (purple circles) and consistent with genome phylogeny are defined as ancestral segregating. ORFs that show the best match with orthologs belonging to a *Saccharomyces sensu stricto* species (red circles) were considered introgressed. ORFs having their best match with orthologs in other less related species were catalogued as HGTs.

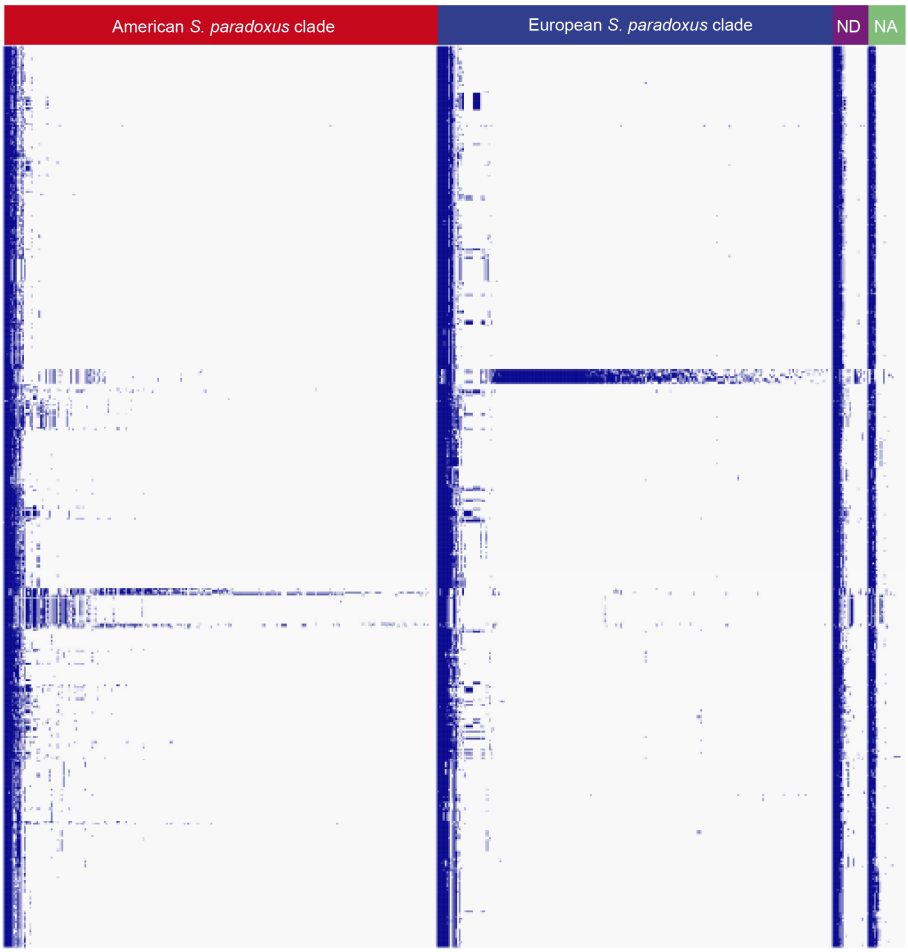


Figure S16: Population of introgressed ORFs. Introgressions from *S. paradoxus* (columns) are divided according to their ancestry (American, European, ND not determined, NA not applicable since derived from different species). There is a striking correlation between the geographical origin of *S. cerevisiae* clades and the ORF ancestry. Spanish Alpechin have mainly Eurasian *S. paradoxus* ORFs, while Brazilian bioethanol, Mexican agave and French Guiana clades have *S. paradoxus* ORFs with American ancestry.

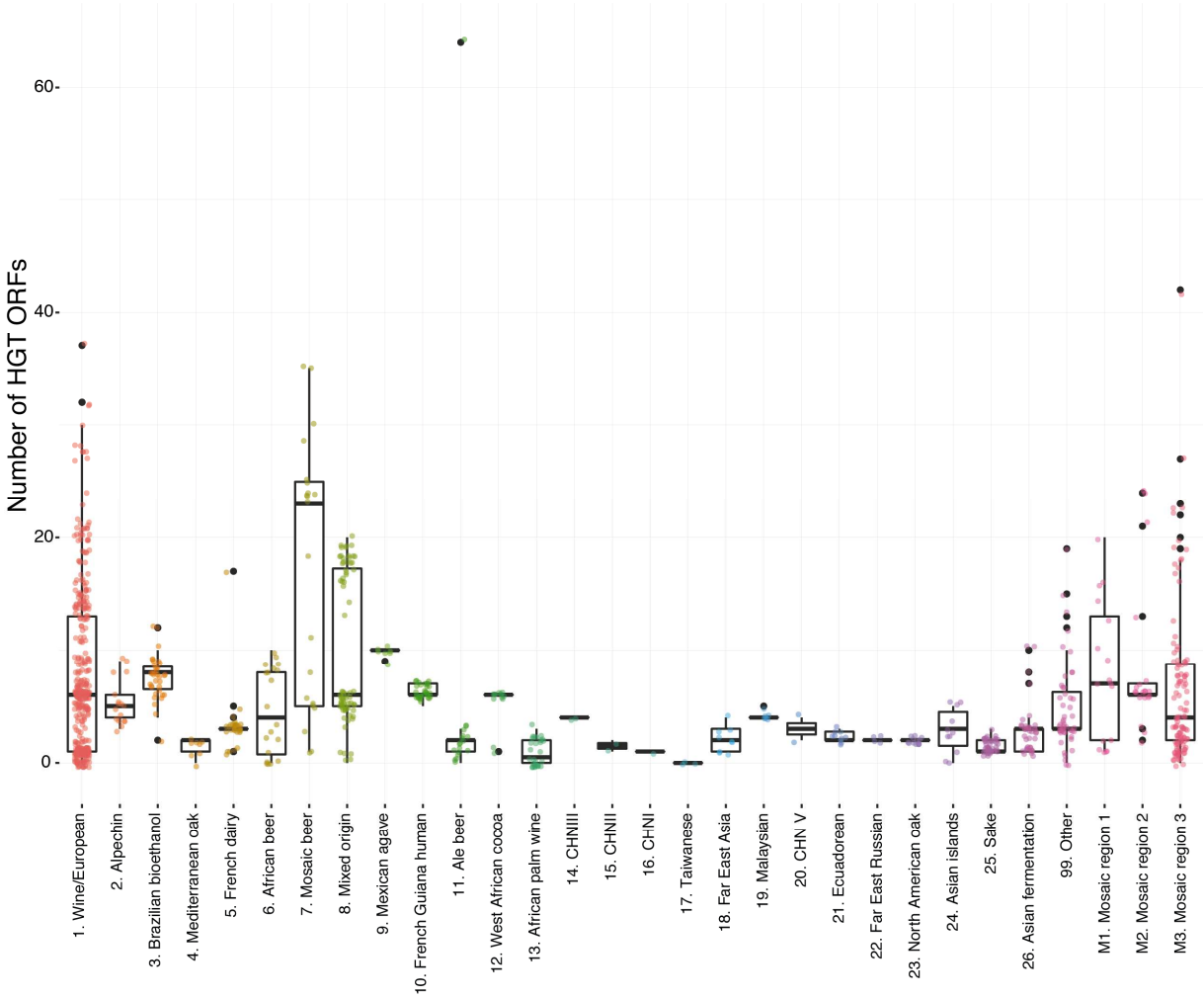


Figure S17: Variation of HGT ORFs across subpopulations. Boxplot representing the HGT ORFs distributions per strain for each subpopulation. Notably, an enrichment of HGT ORFs was found in some human associated subpopulations, such as the Wine/European, Brazilian bioethanol, mosaic beer and mixed origin (2-sided Mann-Whitney test p-values = 5.50e-05, 2.75e-05, 1.33e-05, 8.24e-07, respectively).

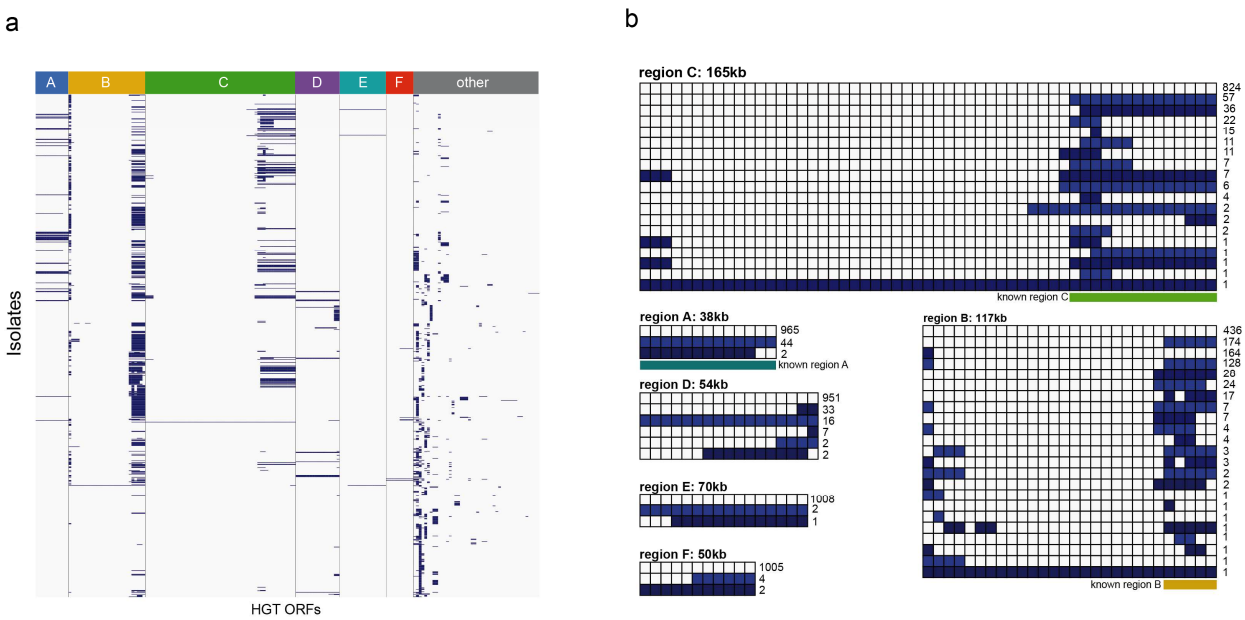


Figure S18: Global view of HGT ORFs. **a.** Presence (dark blue) of laterally transferred ORFs is illustrated in the heat map. The six largest events are labeled as regions A-F and isolates are ordered according to the tree (see isolates in Supplementary table 1). In addition to the HGT events originated from yeast, six inter-kingdom HGT have been detected (five bacterial ORFs and a single viral one, coding for a killer protein partial prepropetide, integrated in the genome and found only in three African isolates). **b.** Patterns of ORFs for the six large HGT events (regions A-F). The profiles are ordered according to the number of strain occurrences (number on the right, starting with the number of strains lacking an HGT event). The extent of the previously characterized regions (A-C) is indicated by the underneath colored bars.

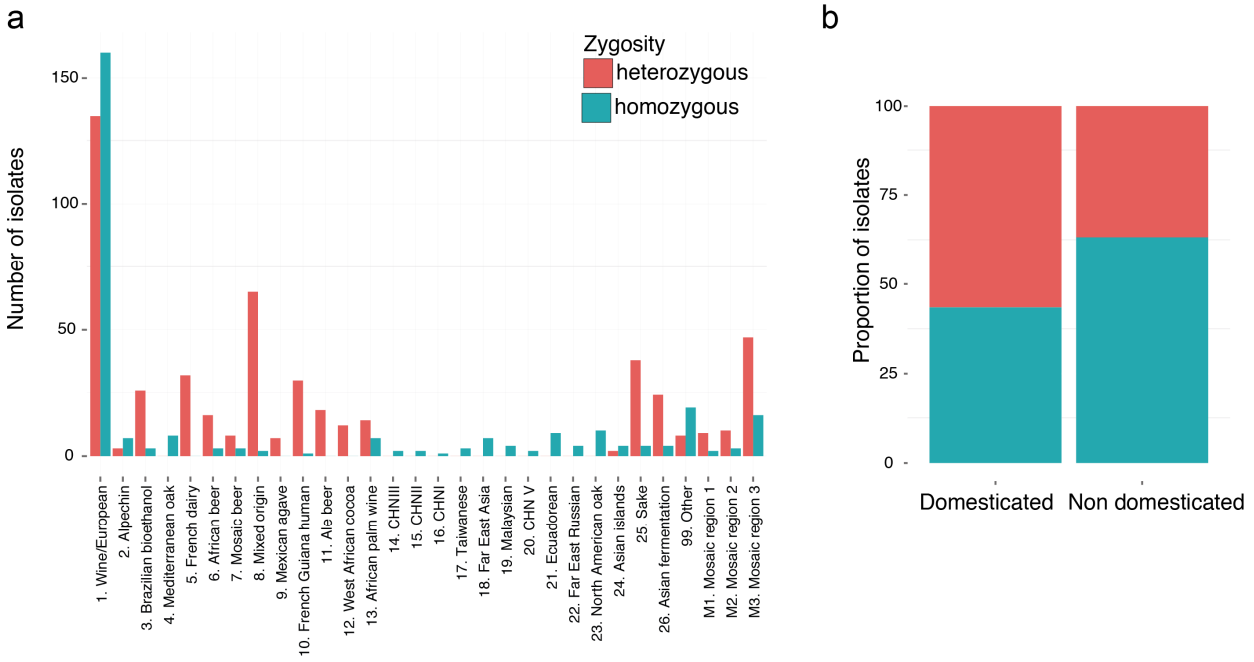


Figure S19: Heterozygous versus homozygous isolates. **a.** Number of heterozygous and homozygous isolates for each subpopulation. **b.** Proportion of heterozygous isolates for domesticated and non-domesticated subpopulations.

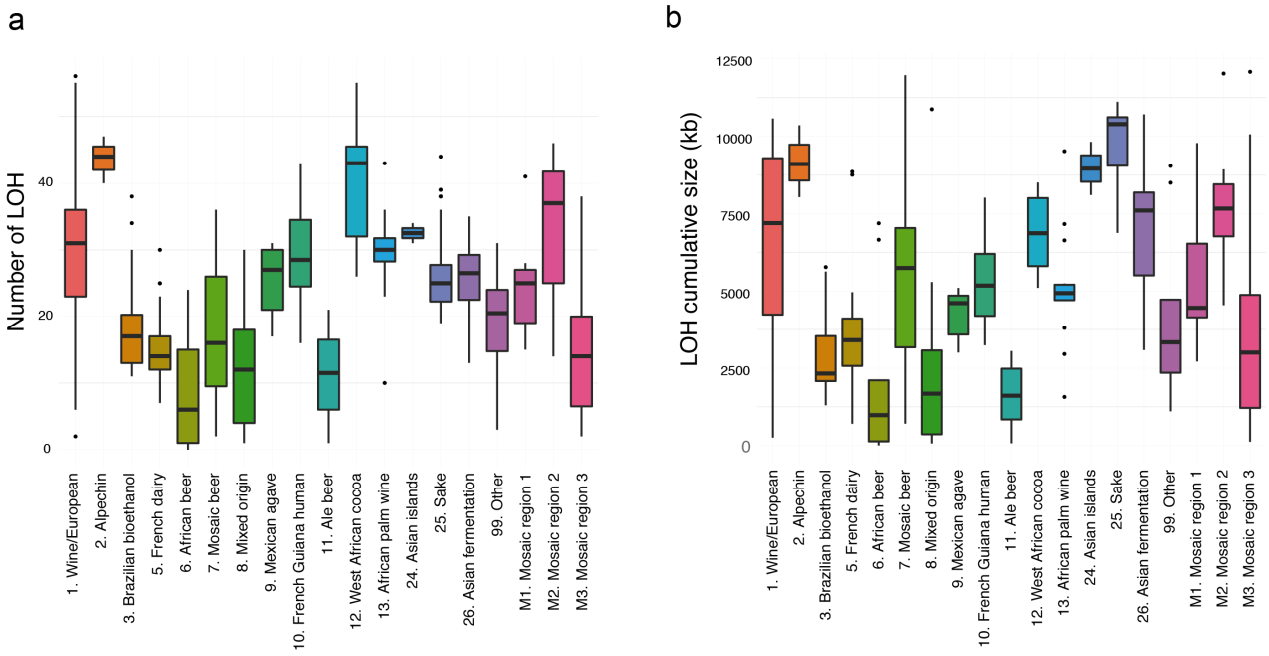


Figure S20: Loss-of-heterozygosity level by subpopulations. **a.** Distribution of the number of region under LOH per isolates. **b.** Cumulative size of the LOH regions per genome in each subpopulation.

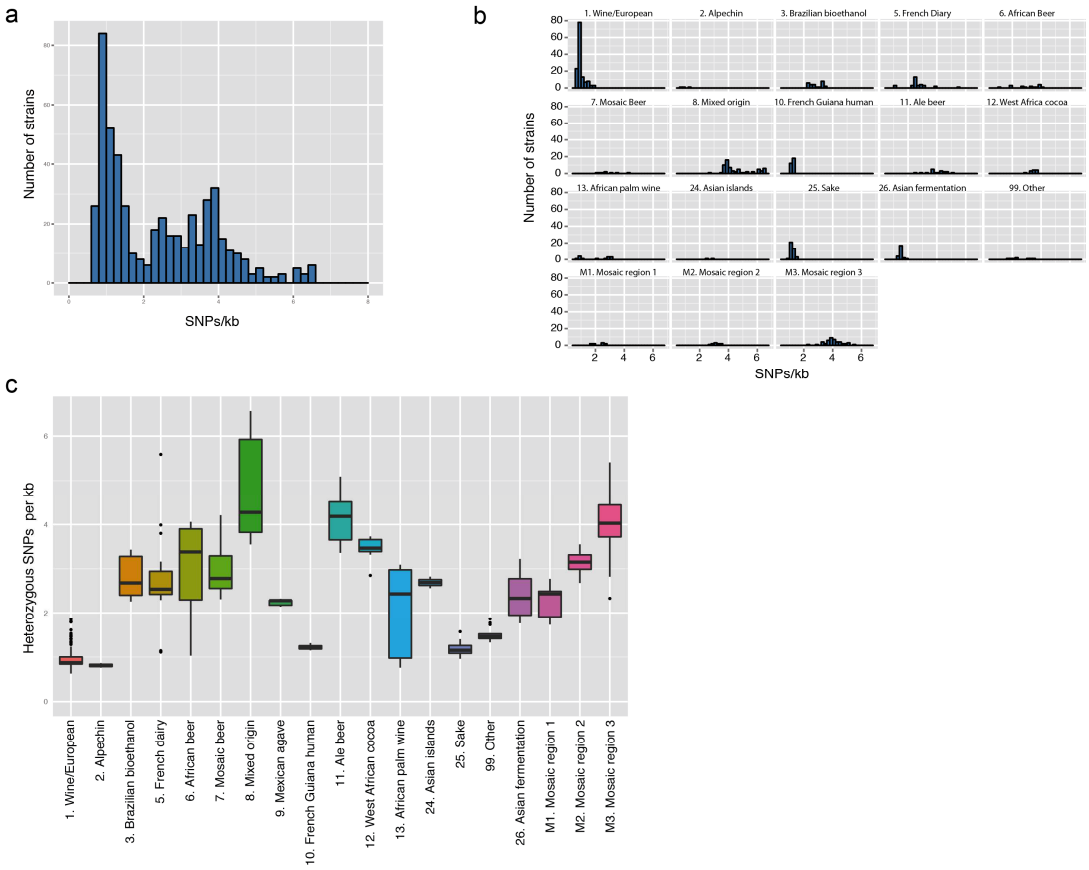


Figure S21: Heterozygous level within the *S. cerevisiae* species. **a.** Distribution of the number of heterozygous site per kb in the natural isolates. This number was determined by removing LOH regions. **b.** Distribution of the number of heterozygous site per kb for each subpopulation. **c.** Boxplots depicting the variation of the number of heterozygous sites per kb in each and between subpopulations.

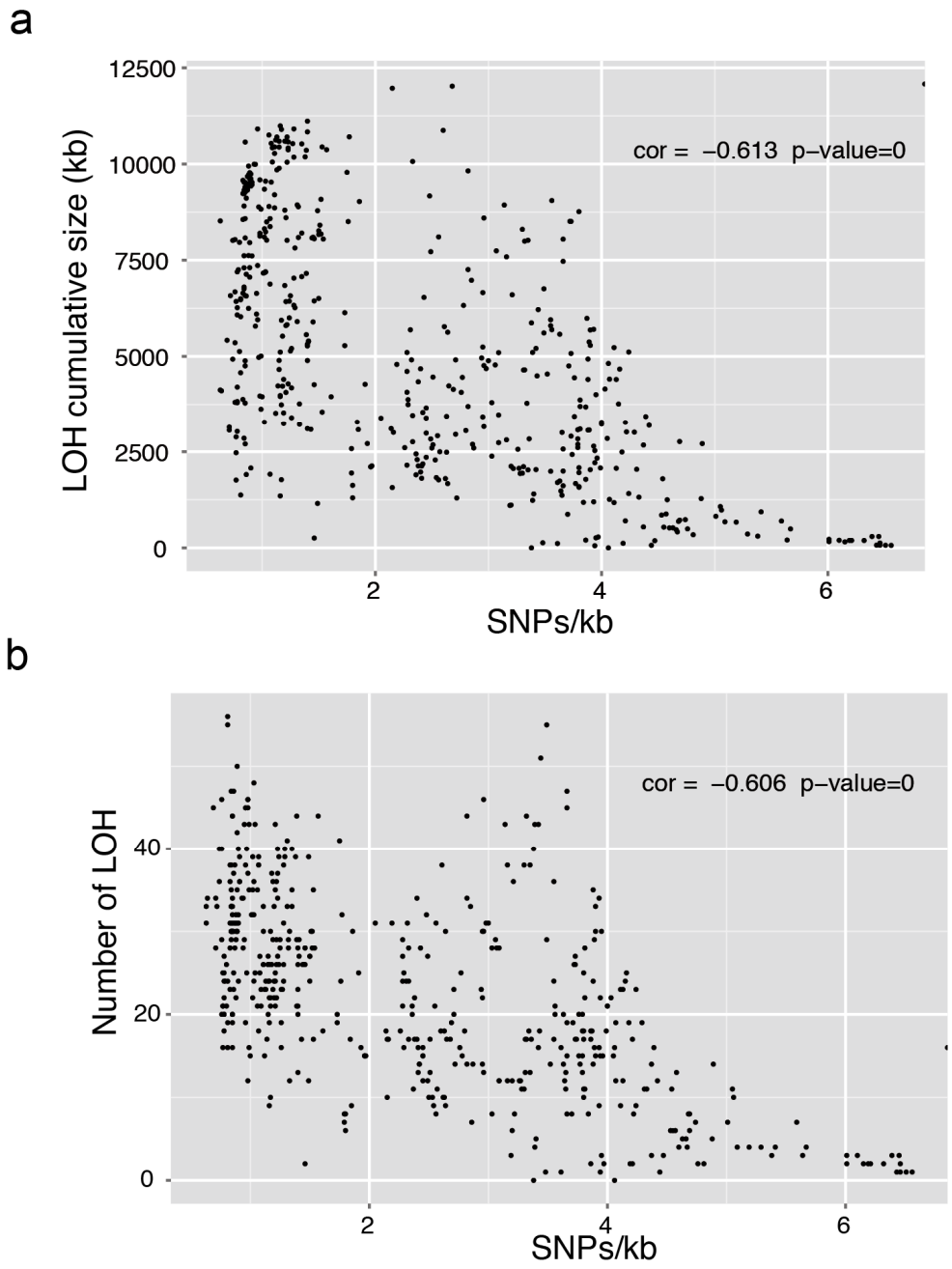


Figure S22: Heterozygosity and LOH level. An anti-correlation is observed between the cumulative size (**a**) as well as the number of LOH regions (**b**) and the number of heterozygous sites per kb. The heterozygosity level is determined by removing the LOH regions.

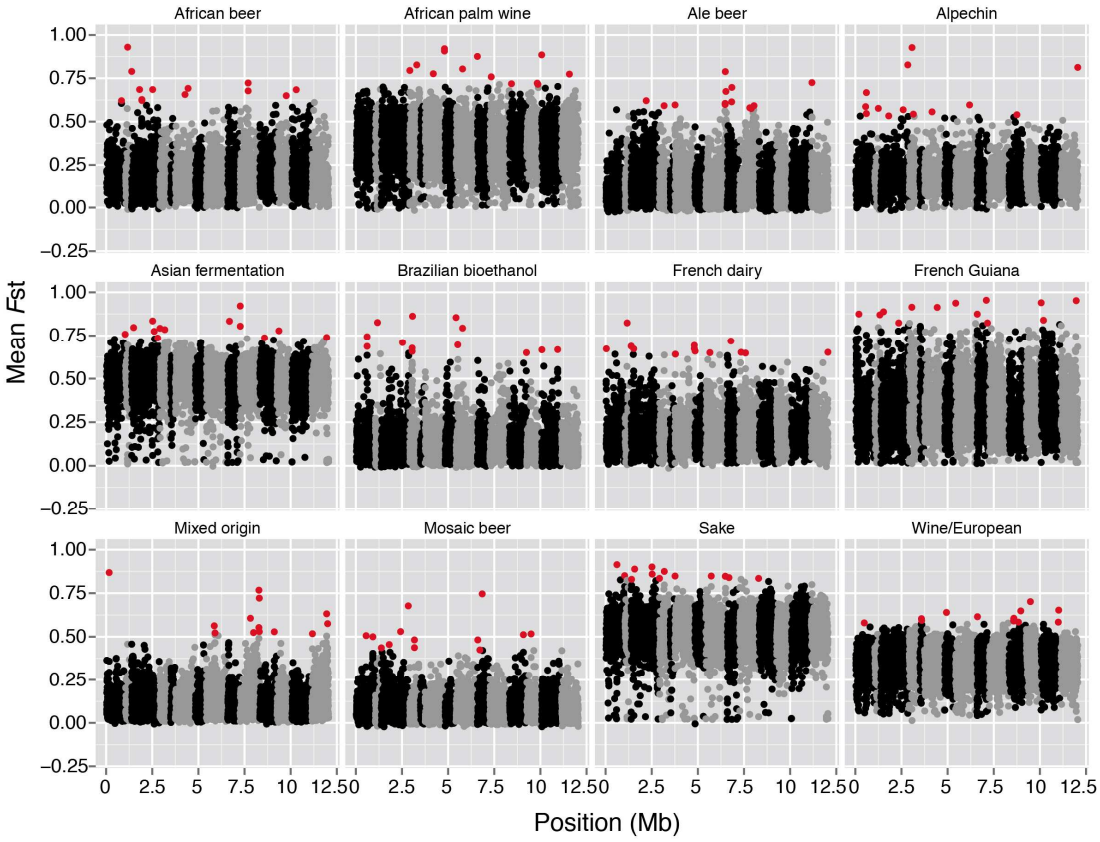


Figure S23: Fst along the genome by clade. Alterative colors (black/grey) represent alternative chromosomes. Regions with high Fst values associated with are highlighted in red, and were considered as candidate for regions under selection.

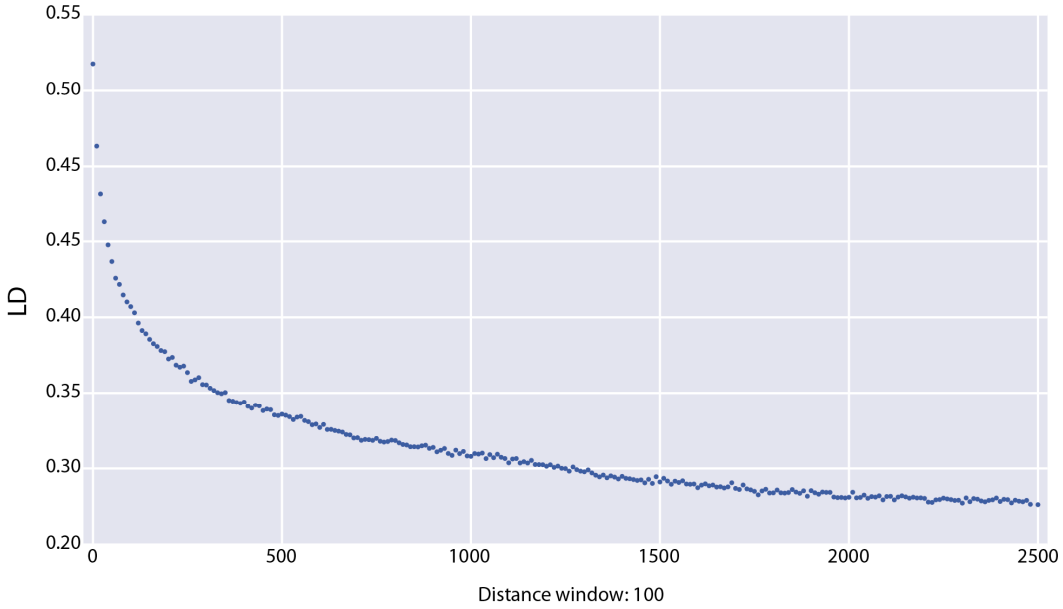


Figure S24: Linkage disequilibrium in the 1,011 genomes. LD decays to half of its maximum value around 500 pb.

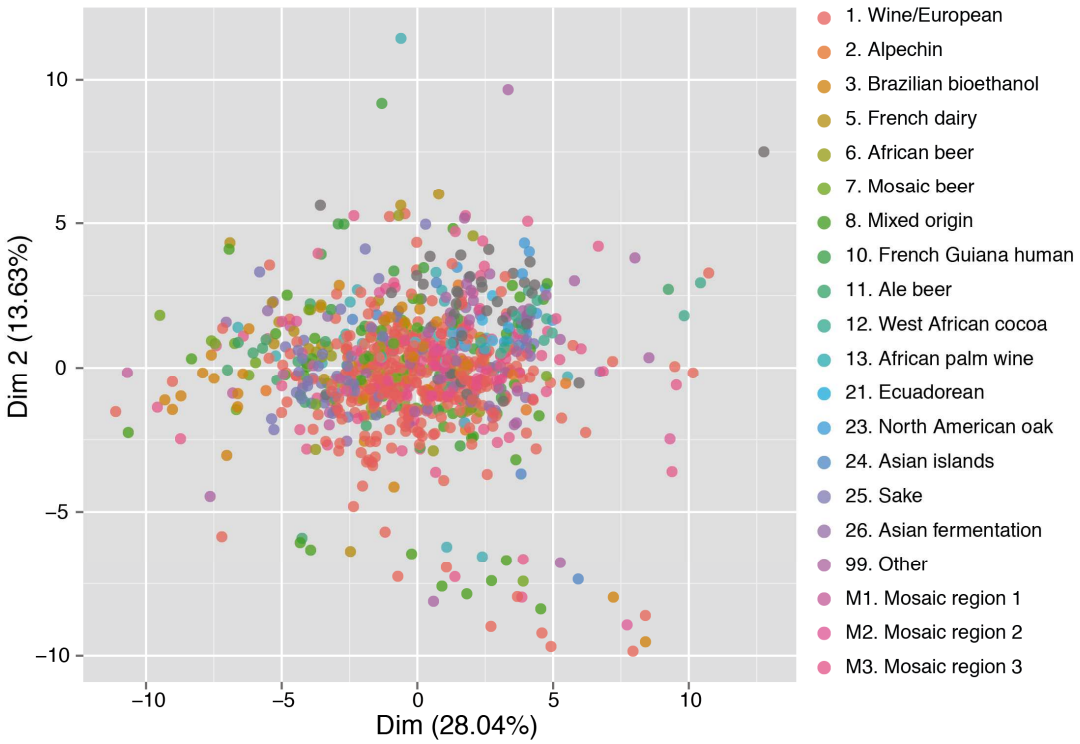


Figure S25: Phenotypic diversity and subpopulations. Principal component analysis (PCA) using growth ratio under all phenotypic conditions as markers. Isolates are colored according to clade. The PCA analysis does not divide individuals into the identified subpopulations.

Figure S26: Principal component projection (PCA) using growth ratio as in Fig. S25. Here, each graph is meant to highlight a clade.

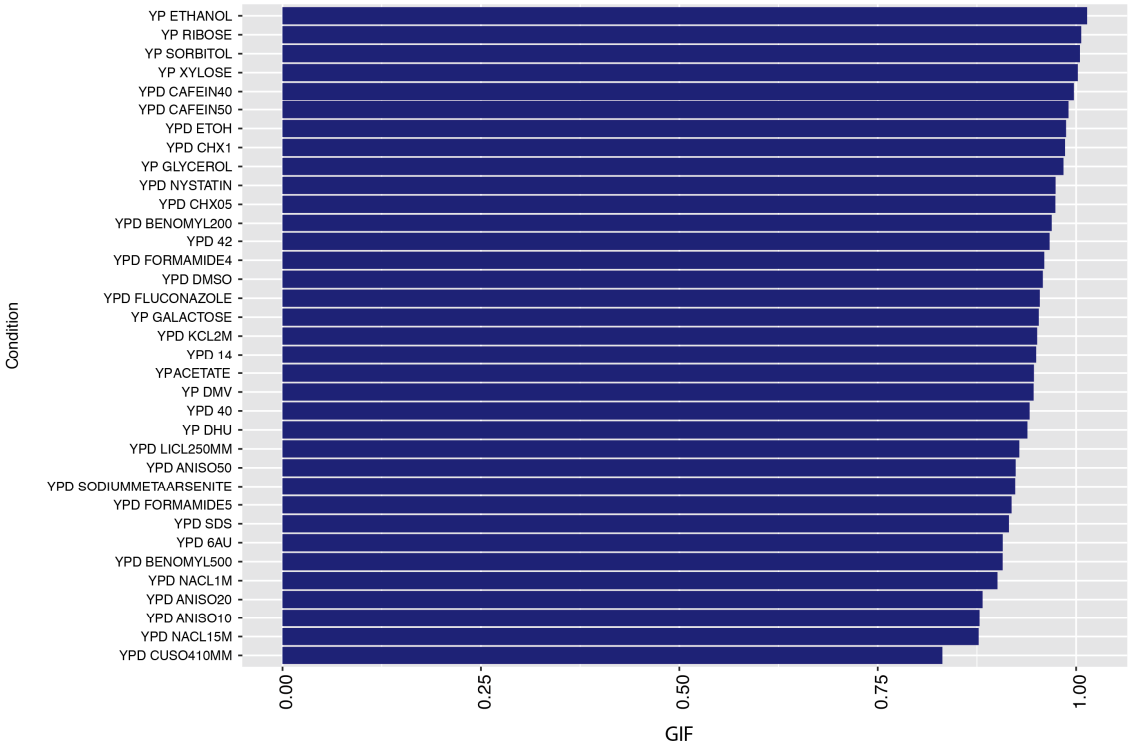
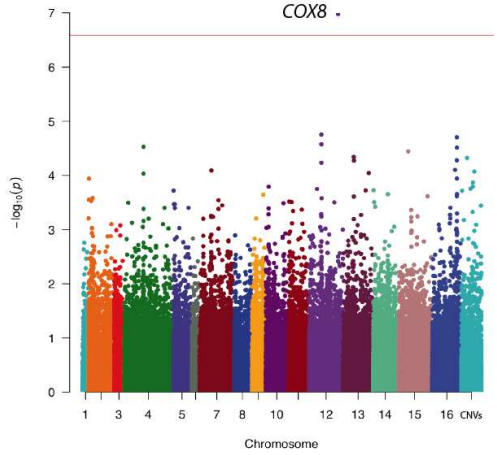
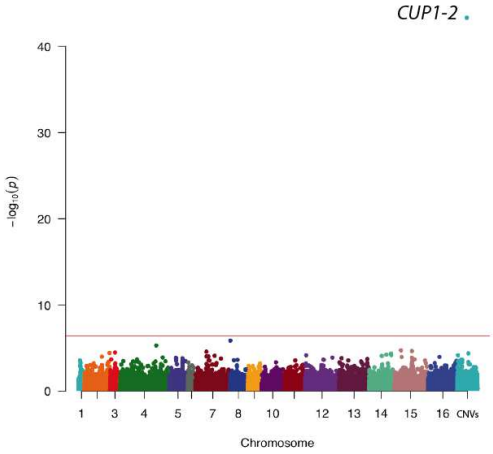


Figure S27: Genome-wide inflation factor of the $\chi 2$ test statistics for our GWAS.

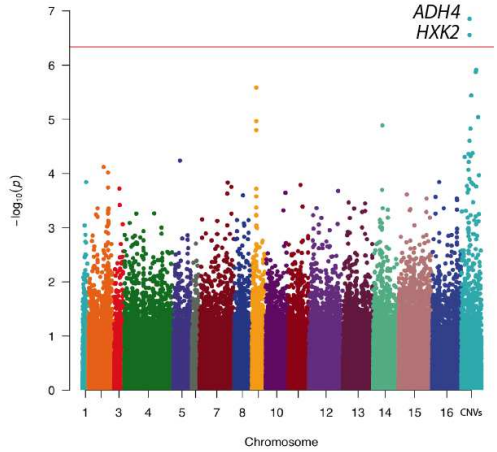
YPD BENOMYL 200



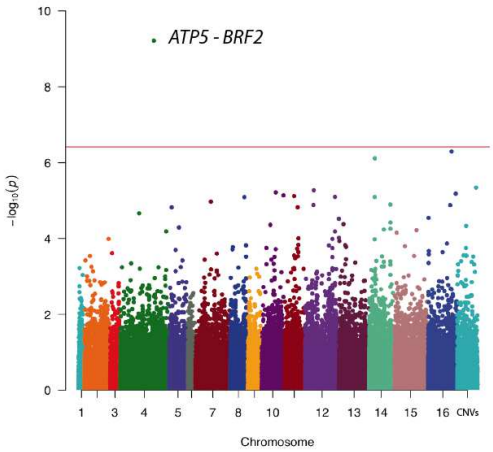
YPD CuSO4 10mM



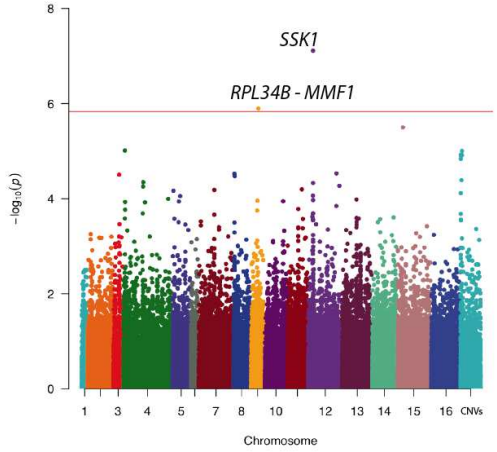
YPD NYSTATIN



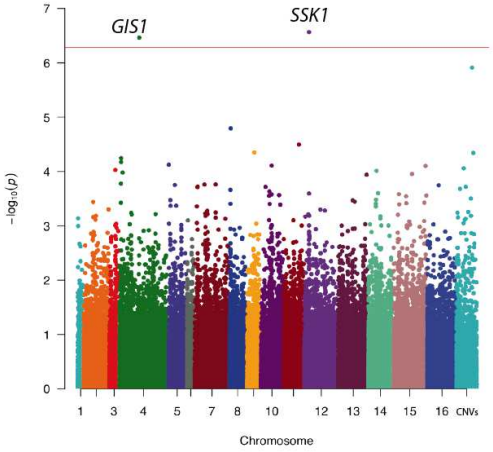
YPD SDS



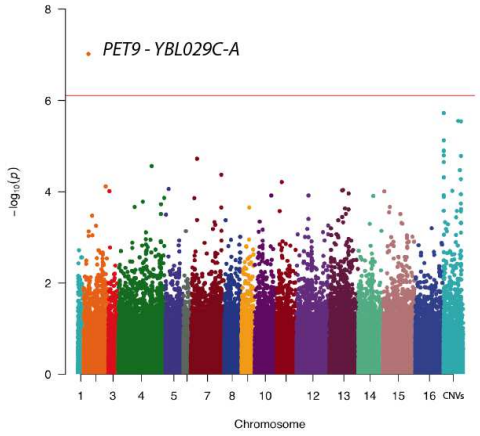
YP RIBOSE



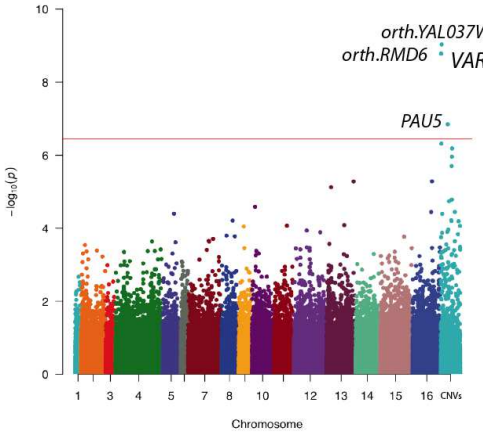
YP SORBITOL



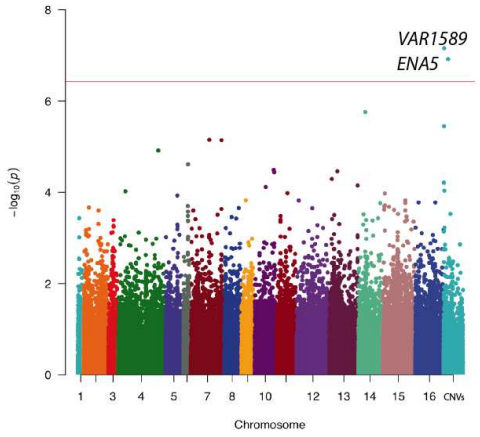
YPD 6AU



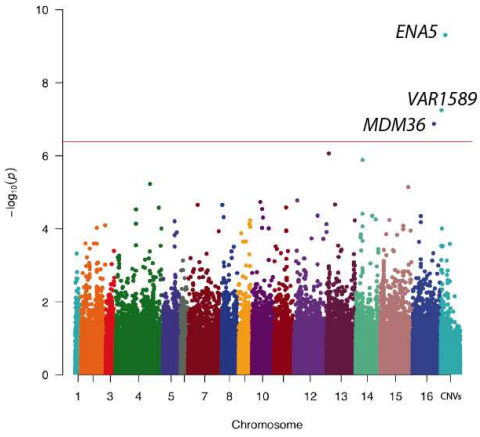
YPD 40



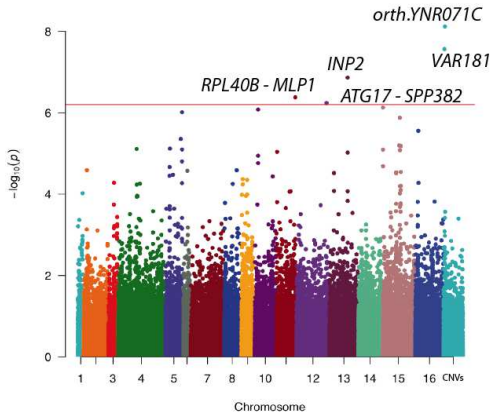
YPD LICL 250mM



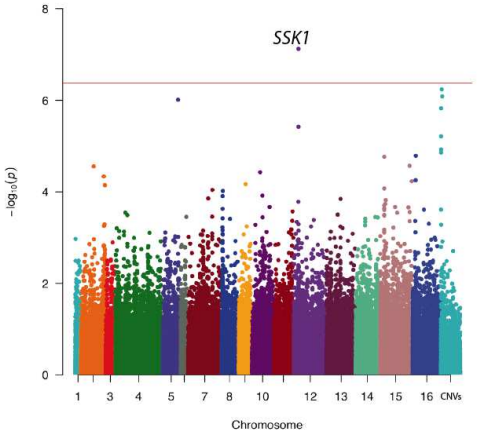
YPD NACL 1M



YP GALACTOSE



YP GLYCEROL



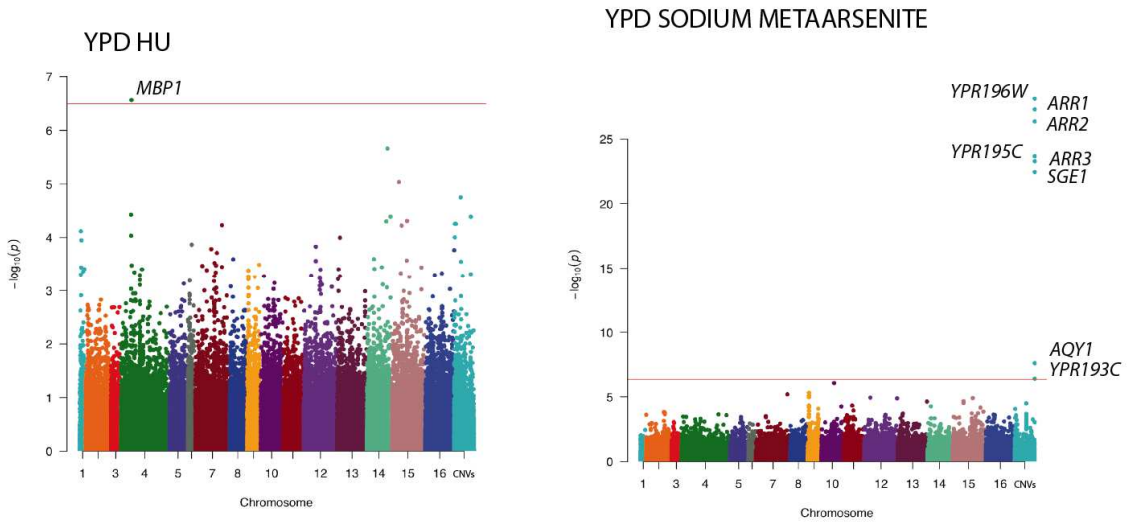


Figure S28: Genome-wide association results in *S. cerevisiae*. Manhattan plots for all conditions that reached significance in our genome-wide association analysis. The chromosomes are represented on the x-axis from 1 to 16, and the CNVs are represented as an extra chromosome. The threshold in red is obtained with 100 permutations of the phenotypes.

References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
2. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
3. 1001 Genomes Consortium. 1,135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166,** 481–91 (2016).
4. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* genetic reference panel. *Nature* **482,** 173–78 (2012).
5. Huang, W. *et al.* Natural variation in genome architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res.* **24,** 1193–208 (2014).
6. Skelly, D. a. *et al.* Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23,** 1496–504 (2013).
7. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–88 (2014).
8. Strope, P. K. *et al.* 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,** 1–13 (2015).
9. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
10. Hou, J., Sigwalt, A., Pflieger, D., Peter, J. & De Montigny, J. The hidden complexity of Mendelian traits across yeast natural populations. *Cell Rep.* **16**, 1106–14 (2016).
11. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–5 (2011).
12. Almeida, P. *et al.* A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol. Ecol.* **24,** 5412–27(2015).
13. Gallone, B. *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166,** 1397–1410 (2016).
14. Gonçalves, M. *et al.* Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26,** 2750–61 (2016).
15. Ludlow, C. L. *et al.* Independent origins of yeast associated with coffee and cacao fermentation. *Curr. Biol.* **26,** 965–71 (2016).
16. Alexander, D. H. & Novembre, J. Fast model-based estimation of ancestry in unrelated individuals. 1655–64 (2009).
17. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458,** 337–41 (2009).
18. Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458,** 342–5 (2009).
19. Wang, Q. M., Liu, W. Q., Liti, G., Wang, S. A. & Bai, F. Y. Surprisingly diverged populations of *Saccharomyces cerevisiae* in natural environments remote from human activity. *Mol. Ecol.* **21,** 5404–17 (2012).
20. Borneman, A. R. & Pretorius, I. S. Genomic insights into the *Saccharomyces sensu stricto* complex. *Genetics* **199,** 281–91 (2015).
21. Yue, J.-X. *et al.* Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49,** 913–24 (2017).
22. Boynton, P. J. & Greig, D. The ecology and evolution of non-domesticated *Saccharomyces species*. *Yeast* **31,** 449–62 (2014).
23. Liti, G. *et al.* High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome. *BMC Genomics* **14,** 69 (2013).
24. Hittinger, C. T. *et al.* Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464,** 54–58 (2010).
25. Hose, J. *et al.* Dosage compensation can buffer copy-number variation in wild yeast. *eLife* **4** (2015).
26. Sunshine, A. B. *et al.* The fitness consequences of aneuploidy are driven by condition-dependent gene effects. *PLoS Biol.* **13,** 1–34 (2015).
27. Cromie, G. a. & Dudley, A. M. Aneuploidy: tolerating tolerance. *Curr. Biol.* **25,** 771–73 (2015).
28. Torres, E. M. *et al.* Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317,** 916–24 (2007).
29. Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl.*

Acad. Sci. U. S. A. **109,** 21010–5 (2012).
30. Tan, Z. *et al.* Aneuploidy underlies a multicellular phenotypic switch. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12367–72 (2013).
31. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 13950–5 (2005).
32. Brown, C. a, Murray, A. W. & Verstrepen, K. J. Rapid expanion and functional divergence of subtelomeric gene families in yeasts. *Curr. Biol.* **20,** 895–903 (2010).
33. Barbosa, R. *et al.* Evidence of natural hybridization in Brazilian wild lineages of *Saccharomyces cerevisiae*. *Genome Biol. Evol.* **8**, 317–29 (2016).
34. Novo, M. *et al.* 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,** 16333–8 (2009).
35. Marsit, S. *et al.* Evolutionary advantage conferred by an eukaryote-to-eukaryote gene transfer event in wine yeasts. *Mol. Biol. Evol.* **32,** 1695–707 (2015).
36. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38,** 1077–81 (2006).
37. Tsai, I. J., Bensasson, D., Burt, A. & Koufopanou, V. Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4957–62 (2008).
38. Friedrich, A., Jung, P., Reisser, C., Fischer, G. & Schacherer, J. Population genomics reveals chromosome-scale heterogeneous evolution in a protoploid yeast. *Mol. Biol. Evol.* **32**, 184–92 (2014).
39. Magwene, P. M. *et al.* Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 1987–92 (2011).
40. Ide, S. *et al.* Abnormality in initiation program of DNA replication is monitored by the highly repetitive rRNA gene array on chromosome XII in budding yeast. *Mol. Cell. Biol.* **27,** 568–78 (2007).
41. Llorente, B., Smith, C. E. & Symington, L. S. Break-induced replication: What is it and what is it for? *Cell Cycle* **7,** 859–64 (2008).
42. Laureau, R. *et al.* Extensive recombination of a yeast diploid hybrid through meiotic reversion. *PLoS Genet.* **12,** 1–30 (2016).
43. Brion, C., Legrand S., Peter J., Caradec C., Pflieger D., Hou, J., Friedrich, A., Llorente, B. & Schacherer, J. Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. *Plos Genet.* **13** (2017).
44. Ohnuki, S. *et al.* Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. *G3 (Bethesda)* **7,** 2807–20 (2017).
45. Attfield, P. V. Stress tolerance: the key to effective strains of industrial baker's yeast. *Nat. Biotechnol.* **15,** 1351–1357 (1997).
46. Higgins, V. J. *et al.* Yeast genome-wide expression analysis identifies a strong ergosterol and oxidative stress response during the initial stages of an industrial lager fermentation. *Appl. Environ. Microbiol.* **69,** 4777–87 (2003).
47. Fay, J. C. The molecular basis of phenotypic variation in yeast. *Curr. Opin. Genet. Dev.* **23,** 672–7 (2013).
48. Warringer, J. *et al.* Trait variation in yeast is defined by population history. *PLoS Genet.* **7,** (2011).
49. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–12 (2010).
50. Fogel, S. & Welch, J. W. Tandem gene amplification mediates copper resistance in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **79,** 5342–6 (1982).
51. Bobrowicz, P. *et al.* Isolation of three contiguous genes, *ACR1*, *ACR2* and *ACR3*, involved in resistance to arsenic compounds in the yeast *Saccharomyces cerevisiae*. *Yeast* **13,** 819–28 (1997).
52. Manolio, T. a *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (2009).
53. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109,** 1193–8 (2012).
54. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–91 (2016).

Chapter 2: Evaluation of the parameters influencing GWAS through extensive simulation in several subpopulations

Introduction

Genome-wide association studies have become an elegant approach to identify genetic variants underlying phenotypic variation. This mapping approach gained importance in several non-human species such as *Oryza sativa*[1], *Arabidopsis thaliana*[2], *Drosophila*[3] or mouse[4]. Due to the ease by which crosses can be made, the budding yeast *S. cerevisiae* represents an historical model organism that has been used to investigate eukaryotic molecular biology, cell biology and genetics. It is also a crucial organism in industrial biotechnology, but wild isolates can also be found in nature. For these reasons it has been the first eukaryotic organism whose genome has been sequenced[5], allowing further research such as the systematic deletion of all genes[6], the identification of many quantitative trait loci (QTL) by linkage mapping (see [7] for a review), and more recently the generation of a global genetic interaction network for this species[8]. However, the strong population stratification in this species, *i.e.* the presence of systematic difference in allele frequencies between subpopulations in a population due to ancestry, can lead to spurious associations and represents the major bottleneck for GWAS in *S. cerevisiae*[9,10]. Even though first studies in yeast laid the foundation for genome-wide association, they all suffered from the lack of statistical power due to the limited size of the cohort used for association. More recently, genomic sequences of 100 strains and 49 phenotypes were used to perform GWAS and allowed to find associations representing proof of principles, with a high probability to detect loci of large effect, despite pointing out that 100 genomic sequences is still limiting the power of the analyses with wide peaks, and few or no significant association found for multiple phenotypes[11]. It is therefore crucial to investigate further the possibility to perform GWAS on *S. cerevisiae*, to see if a large dataset allows the mapping of genetic variation responsible for trait variation, and if some subsets will increase the performance by reducing the bias induced by confounding factors.

In this study, we tested GWAS performance in various subpopulations taken from our 1,011 sequenced isolates of *S. cerevisiae*[12]. Here, we simulated phenotypes based on different genetic architectures, *i.e.* governed by one single nucleotide polymorphism (SNP) for a Mendelian trait and by 10 SNPs for a complex trait. We then measured our capacity to identify the causal variants using GWAS. Our results suggest that associations are easily found for Mendelian traits, with the exception of one dataset (the sake subpopulation) displaying a very high type-I error rate due to cryptic relatedness. Concerning the association with a more complex phenotype, performance varies widely between the cohorts used, with some causal variants left unidentified due to their small effect size. Together, these results emphasize the importance of carefully selecting the individuals that are composing a dataset when performing GWAS. Population size is of great importance, as predicted by previous studies, but the presence of confounding factors sometimes leads to untrustworthy results. Some phenotypic properties also modify the power of detection such as the effect size or the phenotype complexity. Finally, we present some examples of association using real phenotypes to see if the results follow our simulations.

Results

Selection and characteristics of the used populations to perform genome-wide associations

Genome-wide association studies performance varies a lot depending on the characteristics of the population used to run the experiment. In order to evaluate the influence of parameters such as minor allele frequency, sample size and relatedness, we used 6 different subsets of individuals from the 1,011 sequenced isolates of *S. cerevisiae*. Among them, four subsets correspond to subpopulations with more than 40 individuals defined in the frame of the population genomic analysis, namely the wine, sake, mosaic 3 (MR3) subpopulations, and the mixed cluster containing several bakery strains (Figure 1). We also included the total dataset with 1,011 individuals and one dataset composed of 133 individuals spread all along the diversity of the *S. cerevisiae* population (named sampled diversity). Strains from this latest dataset have been chosen to avoid an overrepresentation of some clusters of this species. It is expected that the performance of these datasets will not only be linked to the sample size, as there are multiple factors that can influence the results of association. As for other species like human or *A. thaliana*, there is a bias towards low frequency variants in *S. cerevisiae*.

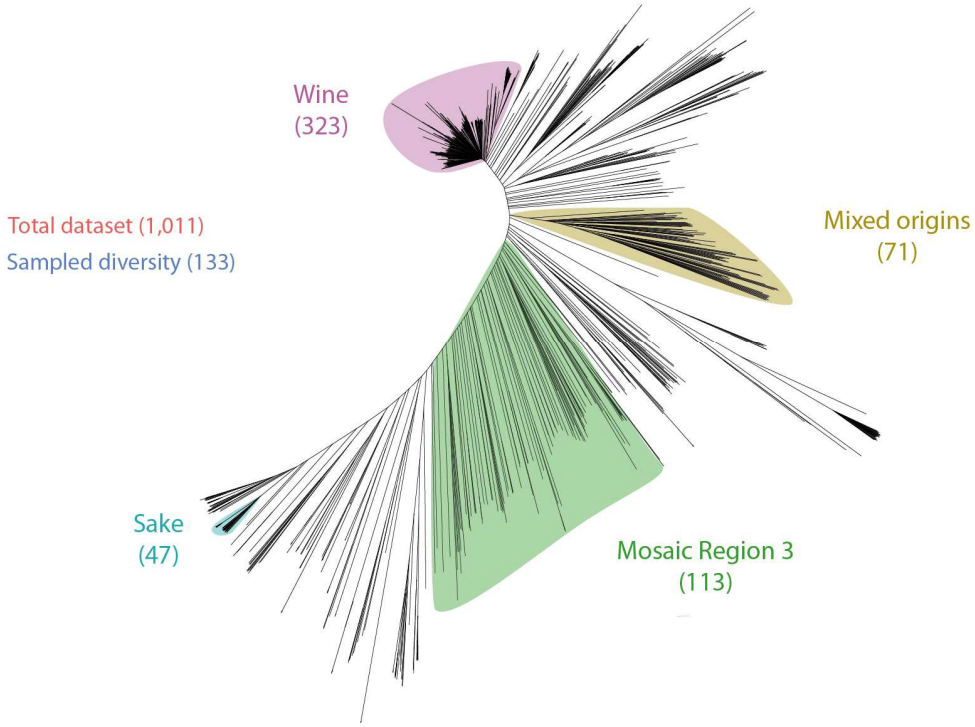


Figure 1: Subpopulations used to simulate GWAS in *S. cerevisiae*, with sample size indicated by the number in parentheses. The sampled diversity dataset is composed of individuals spanning the whole genetic diversity.

In order to give an idea of how this bias affects each dataset, we computed the percentage of singletons, *i.e.* the markers having a minor allele count (MAC) equal to 1 (Table1). These

values indicate how strong is the skew towards rare allele for each dataset. While this percentage is quite high for the sampled diversity and the MR3 dataset (19.24% and 20.01%, respectively), it is lower for the sake and the mixed origins cluster (7.89% and 2.42%, respectively). The total dataset stands in the middle with 11.41% of singletons. In order to avoid spurious associations, successive quality control filtering steps have to be applied. For each population, we therefore built a matrix of single nucleotide polymorphisms (SNPs), keeping only biallelic SNPs with a MAF superior to 5%. We also filtered out SNPs with missing genotypes for all our matrices, except for the total dataset, for which we kept SNPs with at least 1,000 present genotypes. After these filtering steps, the datasets obtained contained between 14,164 SNPs for the wine cluster and 82,869 SNPs for the 1,011 strains (Table1). For each of them, the SNPs are uniformly distributed across the whole genome.

Matrix	# indiv.	%sites MAF>5%	% sites MAC=1
Wine	323	14.16	12.76
Sake	47	21.49	7.89
Sampled Diversity	133	66.30	19.24
MR3	113	72.81	20.01
Mixed origins	71	81.03	2.427
Total Dataset	1011	82.87	11.41

Table 1: Sample size, number of sites with MAF<5% and proportion of sites with MAC=1 of the datasets used in this study

We looked at the distribution of the MAF for each dataset after the pruning steps (Figure 2). These graphs show a decreasing number of SNPs as the MAF grows, with again, a strong bias towards low frequency variants except for the mix bakery dataset, showing more 'uniformly' distributed MAF values, and the sake cluster, that exhibits almost 13% of remaining SNPs with a MAF of 0.489362. The locations of these latter SNPs are distributed across the whole genome. This phenomenon is explained by the fact that strains from the sake cluster are very closely related due to the fact that they share a recent common ancestor[13]. For this reason, it is expected that this might lead to spurious associations, as it would be impossible to distinguish the variation in allelic frequency due to this phenomenon from the one due to association between the variant and the tested phenotype.

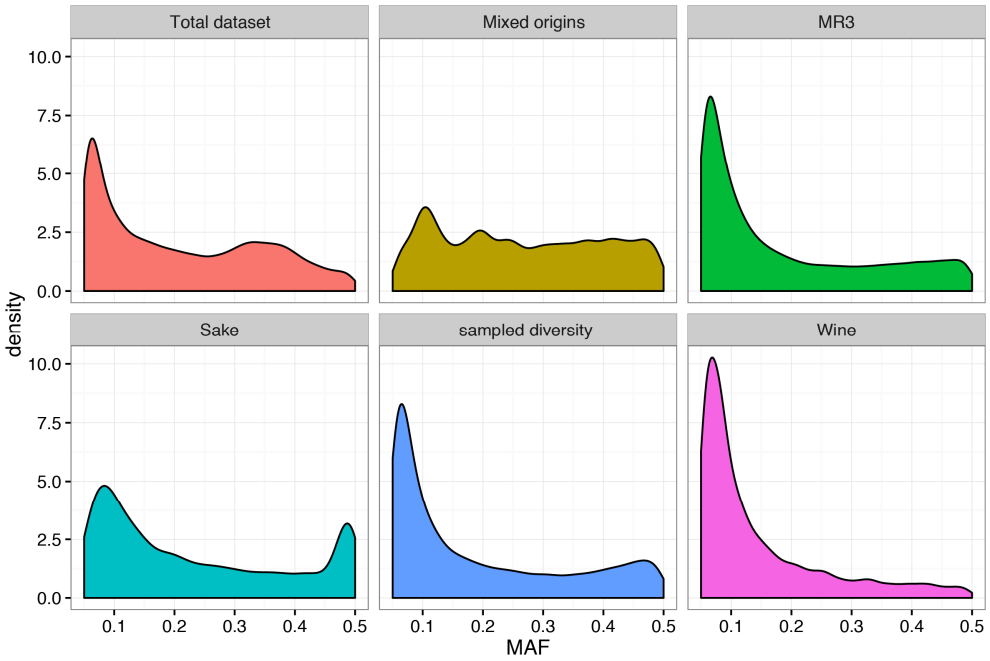


Figure 2: MAF distribution of the datasets used in this study.

Detection and mapping of Mendelian traits across populations

In order to test the ability of each dataset to map the causal variant(s) responsible for trait variation, we randomly selected the "causal" genetic determinants and generated phenotypic data in accordance with the genotype of each strain and association was run using FaST-LMM, which adds a polygenic term to the standard linear regression designed to circumvent the effects of relatedness and population stratification[14]. These steps have been repeated 1,000 times for each dataset with the simulation of a Mendelian trait (*i.e.* governed by 1 SNP), and again 1,000 times for each dataset with the simulation of a complex trait (here governed by 10 SNPs of varying effect sizes). For each of the runs, we then permuted the phenotype data a hundred times and recorder the lowest P-value from each genome-wide test and determined a specific genome-wide significance threshold of 5% family-wise error rate (FWER). The first step was to evaluate the capacity to detect association between the different matrices and the simplest genetic architecture for a trait, *i.e.* a trait governed by a monogenic mutation. For all runs for the 6 datasets, the causal SNP has been identified as significant except for 7 out of the 1000 runs of the sake subpopulation. This result is promising as it shows that in the case of a Mendelian trait, the causal SNP can be detected using GWAS.

Mendelian traits, false positives and evolutionary history

Once the ability to identify the causal SNP is evaluated, it is important to check if the experiment is specific enough, so that the true positive detected here is well differentiated from the other markers. In order to look at the extent to which spurious associations has been

generated by our datasets and whether there is variation between the datasets for this factor, we evaluated the genomic inflation factor (λ) for all runs by dataset. This value quantifies the extent of the bulk inflation and the excess of false positive rate. While some datasets show value inferior than one, indicating a certain lack of power to detect associations, the mix bakery, the mosaic, the total dataset and the sampled diversity datasets show λ values for most runs that are close to the expected value of 1, with a median value ranging from 0.90 to 1.11 for the wine and sake clusters (Figure 3). These values indicate that the dataset are well suited for the detection and mapping of a Mendelian trait.

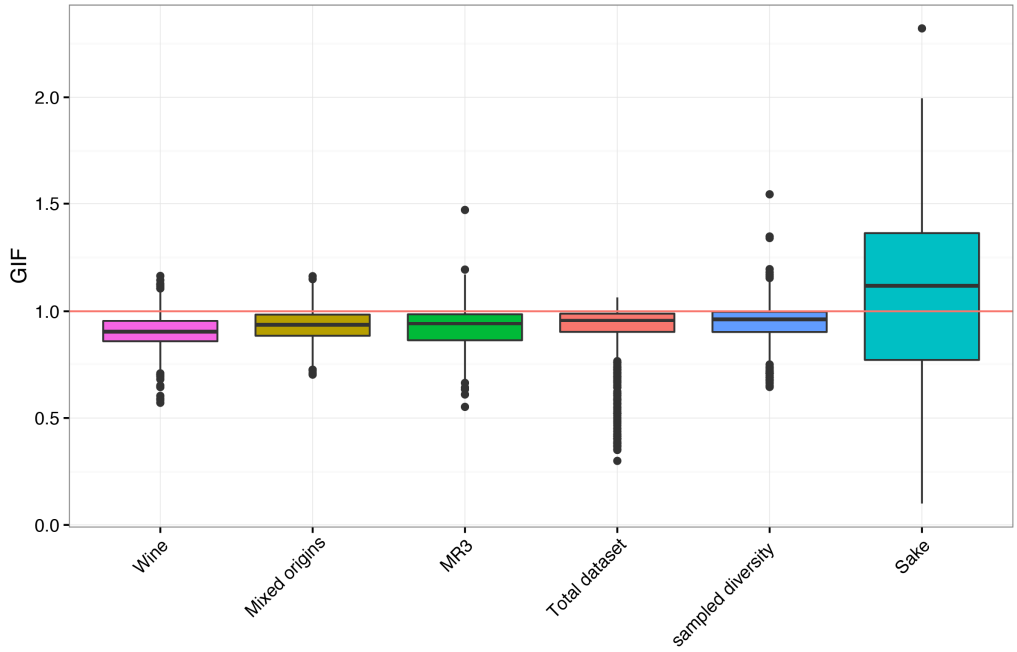


Figure 3: GIF by dataset for the simulations of a Mendelian trait. The red horizontal line corresponds to a GIF of 1, which is the expected value.

We also computed and represented the false positive rates (FPR) for each run by dataset (Figure 4). This value represents the proportion of false positives among the variants that are not causal. For all datasets, the median of the FPR is quite low and ranges from 1.36e-04 for the sampled diversity dataset up to 1.86e-03 for the sake subpopulation.

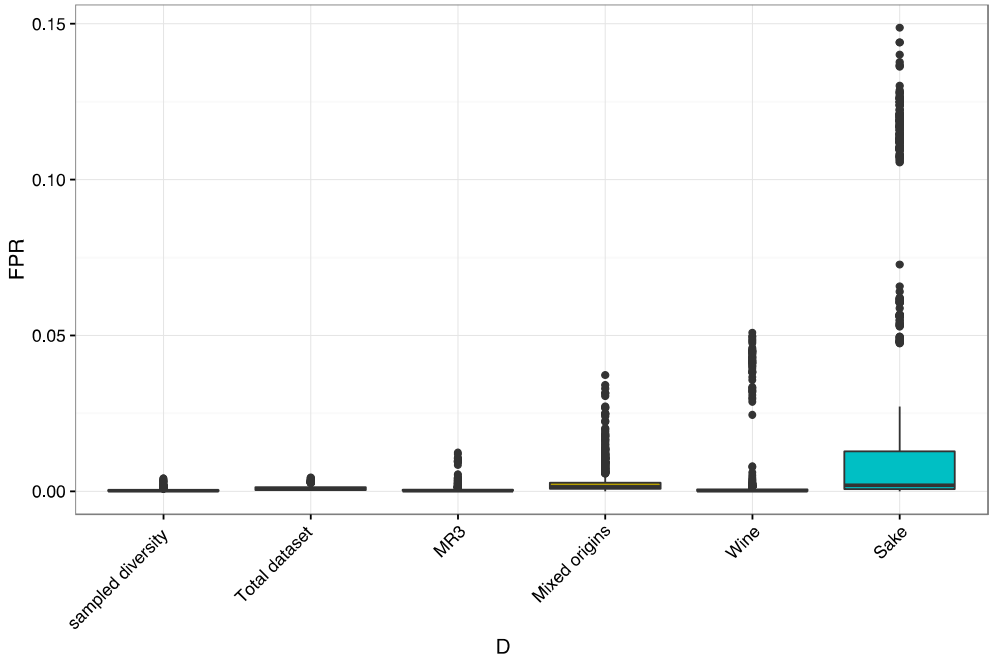


Figure 4. False positive rate by dataset for the simulations of a Mendelian trait.

It is important to mention that the number of false positives is overestimated, as the SNPs in linkage with the causal SNP due to their proximity are considered as false positives. This effect can be seen by looking at the distribution of the distance between false positive SNPs and the causal SNP when they are located on the same chromosome. Indeed, the ratio between the medians of the most extreme datasets (sampled diversity and sake) equals 44.7 (Figure 5).

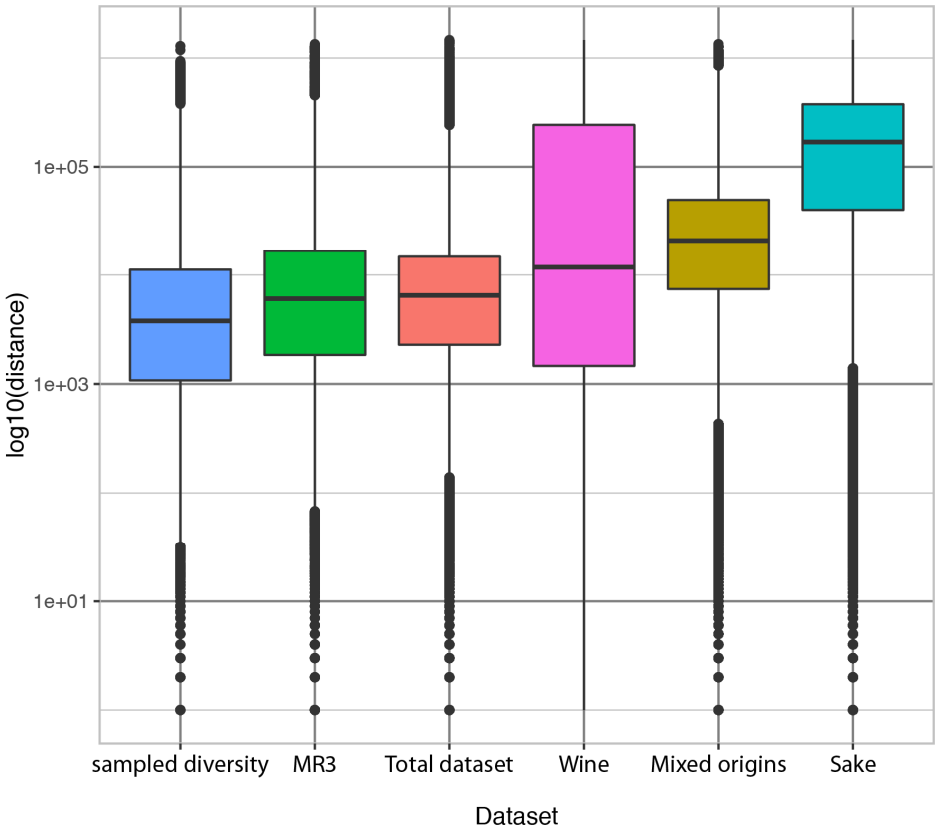


Figure 5: Distribution of the distances between the detected markers and the causal SNP.

It is also interesting to note that a large number of GWA runs present a high the false positive rate, and these extreme values go high for specific subpopulations, namely mixed origin, sake and wine (Figure 4). Indeed, a total of 145 runs using the sake subpopulation show a FPR of more than 10%, making the results of this dataset unreliable. On the other side, we can observe that some cohorts present few outliers in terms of false positives and an overall low FPR rate (Figure 4). For example, the sampled diversity cluster has the lowest FPR median of all datasets, and the GWA run with the highest FPR is 4.09e-03. Similarly, the total dataset with 1,011 strains has an overall low false positive rate, with the highest FPR beeing 4.3e-03. For that reason, we ordered the plots by the variance of the FPR, which is here more indicative of the performance of the dataset (Figure 4). Although the causal SNP governing a Mendelian trait is always found except for a few runs, we already notice a discrepancy between the dataset in terms of false positives. These results suggest that most of the datasets seem appropriate to perform GWAS as we do not notice inflated p-values due to structure, except for the sake subpopulation, displaying as a result a very large number of false positives. Sample size is not the only factor influencing the FPR, as the sampled diversity shows best performance for this parameter, and wine cluster (2nd highest sample size) performs poorly, meaning that a careful selection of the strains to avoid the overrepresentation of some closely related strains might be a good way to build an efficient dataset for GWAS.

Relatedness and the mapping of Mendelian traits: the example of the sake subpopulation

The sake subpopulation shows bad performance for GWAS, even for a Mendelian trait. First, it fails to identify the causal SNP in 7 out of the 1,000 runs, but more importantly it shows for many runs a FPR too high to allow any conclusion to be drawn from these association studies. One reason explaining the high FPR we notice is that this dataset is only composed of 47 strains, which is of comparable size to what has been already used by the previous studies[9,10]. But as stated previously, the sample size is not the only factor explaining the observed discrepancy between the datasets. Our hypothesis is that the strains in this dataset share a very recent common ancestor and this is the source of a systematic bias in allele frequency. It is therefore impossible to distinguish between a correlation of allele frequencies across individuals due to relatedness and one due to biological adaptation. To investigate this, we looked at the 7 cases where the causal SNP was not detected in the 1,000 runs of simulation, and we noticed that 6 of them have a MAF of 0.489362, which is the overrepresented MAF value shared by almost 13% of the SNPs in the sake dataset. Another interesting observation concerns the MAF of the false positives for these 7 runs. Indeed, we can see a clear link between the MAF of those falsely identified SNPs and the causal variant that has not been found (Figure 6a). Despite that, there are 129 runs for which the causal SNP has a MAF of 0.489362 and has been significantly associated with the phenotype, but all these runs seem to lead to more false positives (Figure 6b) than runs with another value of MAF associated to the causal SNP. As a general rule for this dataset, it seems that the enrichment of SNPs having a MAF of 0.489362 will cause problems. If they are causal, it will be harder to detect them by association or it will lead to an increased number of false positives. If they are not causal, we will be observed as wrongly associated to the phenotype. To illustrate this statement, we represented on boxplots the minor allele frequencies of the false positive, false negative and true positive hits (Figure 6a). We see that these SNPs are overrepresented among false positive and false negatives (Figure 6a). This example allowed us to measure the impact of relatedness (*i.e.* sharing a recent common ancestor) on the mapping of traits with GWAS, as it will result in a MAF bias among the SNPs.

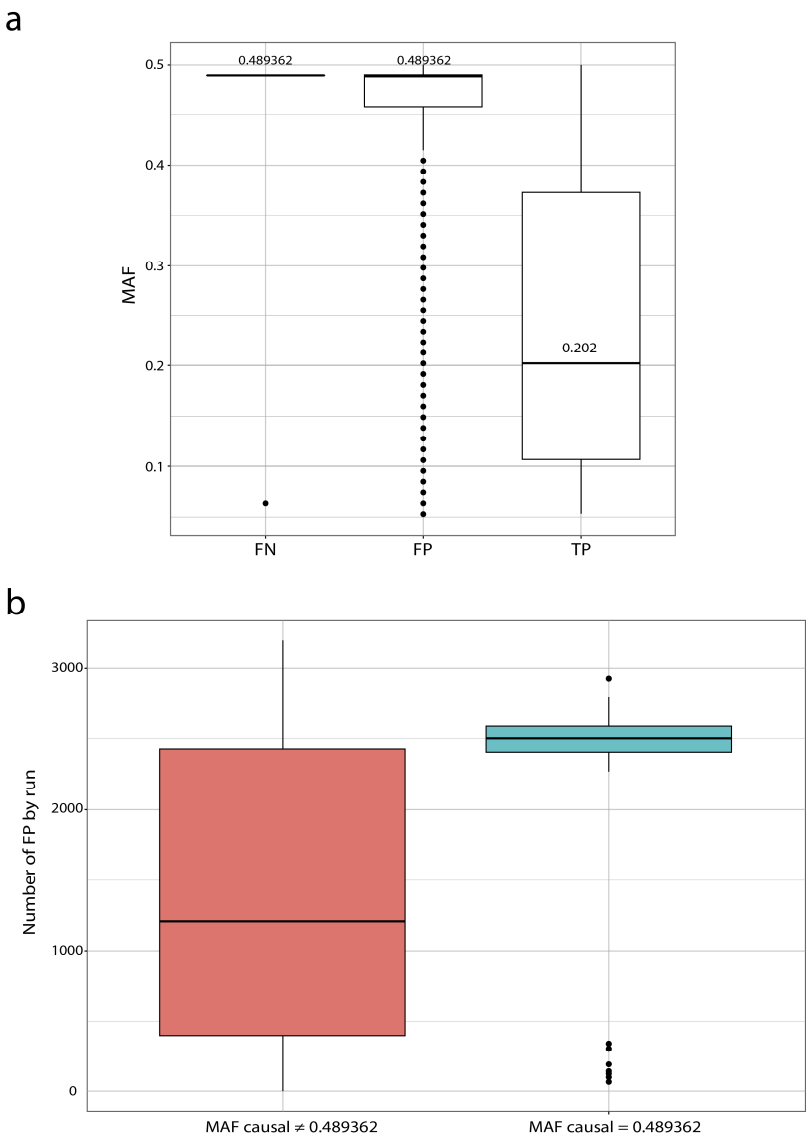


Figure 6: MAF bias and simulation results. **a.** MAF of FN, FP and TP. The MAF for the FN fall in the overrepresented MAF value of 0.489362 for 6 out of 7 cases. FP are also enriched for this MAF value. **b.** Boxplots of the number of FP by run whether the causal SNP has or not a MAF of 0.489362.

Mapping of complex traits: the importance of sample size and effect size

The detection and the mapping of genetic variants underlying a complex trait is one of the main goals for association studies, as Mendelian traits are more the exception than the rule. The last decade, large genome resequencing projects were launched with very optimistic forecasts that larger datasets would be able to map the large genetic contributions to many traits, including diseases[15,16]. While many traits are known to be determined by a large genetic contribution, these studies with extremely large datasets failed to explain a large part of the heritability. For example it has been shown recently that using genomic data from over 700,000 individuals, around 700 variants have been significantly identified as playing a role in the adult human height, but they only explain 20% of the heritability for this trait, which is estimated to be around 80%, depending on the estimates[17,18]. The unexplained part of genetic

variation is what we call the missing heritability. This result underlines one possible explanation to the problem of missing heritability, as the use of very large cohort to perform association allowed to map numerous rare variants of large effect, which is not possible with smaller datasets, but is far from being sufficient to overcome the missing heritability problem. Other possible reasons include non-additive genetic effects, multiple variants of small effects, epigenetics.

To further evaluate the performance of our datasets, we wanted to see which are the ones able to recover the largest part of heritability by mapping the most causal SNPs possible. We therefore simulated complex traits governed by 10 SNPs. The effect sizes of the causal SNPs followed a normal distribution, offering the possibility to see the influence of this factor on the detection of causal variants. Indeed, as most traits are complex, it is possible that the contribution of causal variants to the phenotypic variation will span a wide range of effect sizes. Obviously, it is expected that SNPs with low effect sizes will be harder to detect.

No dataset was able to recover all the causal SNPs that were used to simulate the phenotypes. The best results obtained being the detection of 8 causal SNPs out of 10 for 15 runs of the dataset containing all 1,011 individuals. The median of the TPR for this dataset is also the best compared to the other ones, with a value of 5 true positive SNPs out 10. Expectedly, the worst dataset is again the sake cluster, with a median of 0 causal SNPs detected. Between these two, the median TPR ranges from 0.1 to 0.3 (Figure 7). We notice here that the TPR is growing together with the sample size, indicating that this parameter is probably playing an important role in the power of detection for a complex trait (Figure 7).

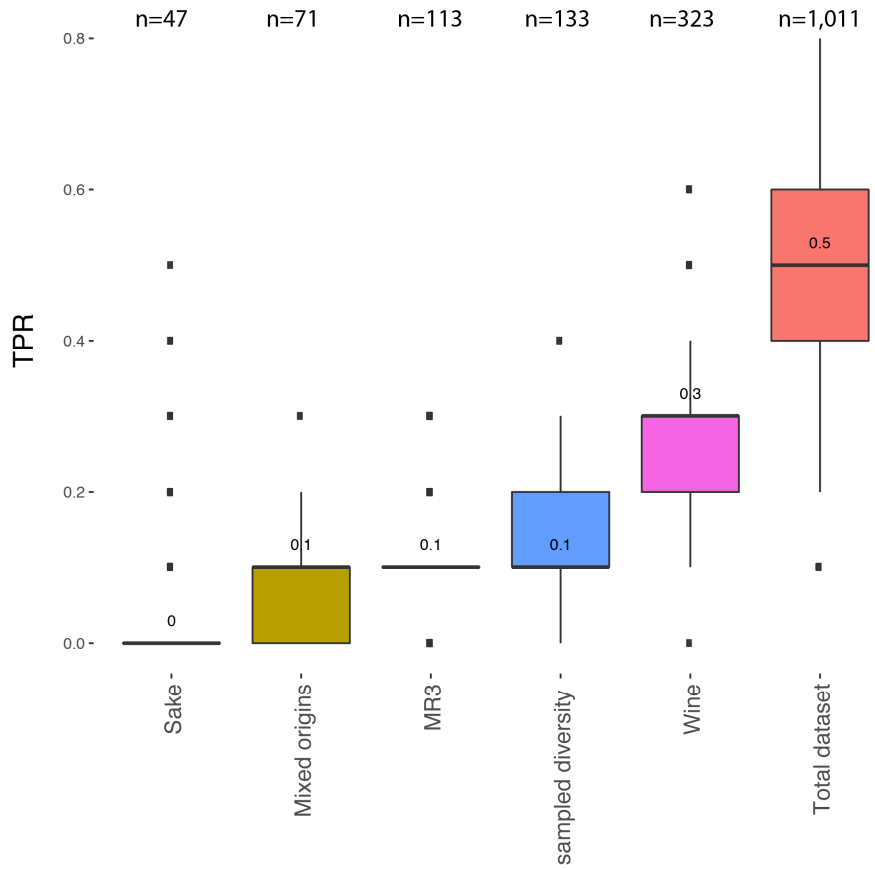


Figure 7: True positive rate by dataset for the simulation of a complex trait. The power of detection grows together with the sample size (written on top).

It seems that there is no influence of the MAF on the detection (Figure 8a), whereas we can observe a significant impact of the effect size on all datasets, except for the wine and the Mixed origins datasets (Figure 8b. Indeed, the genetic contribution on the phenotype being split on several SNPs, the lower the effect size of a SNP, the harder it will be to detect it (Figure 8b).

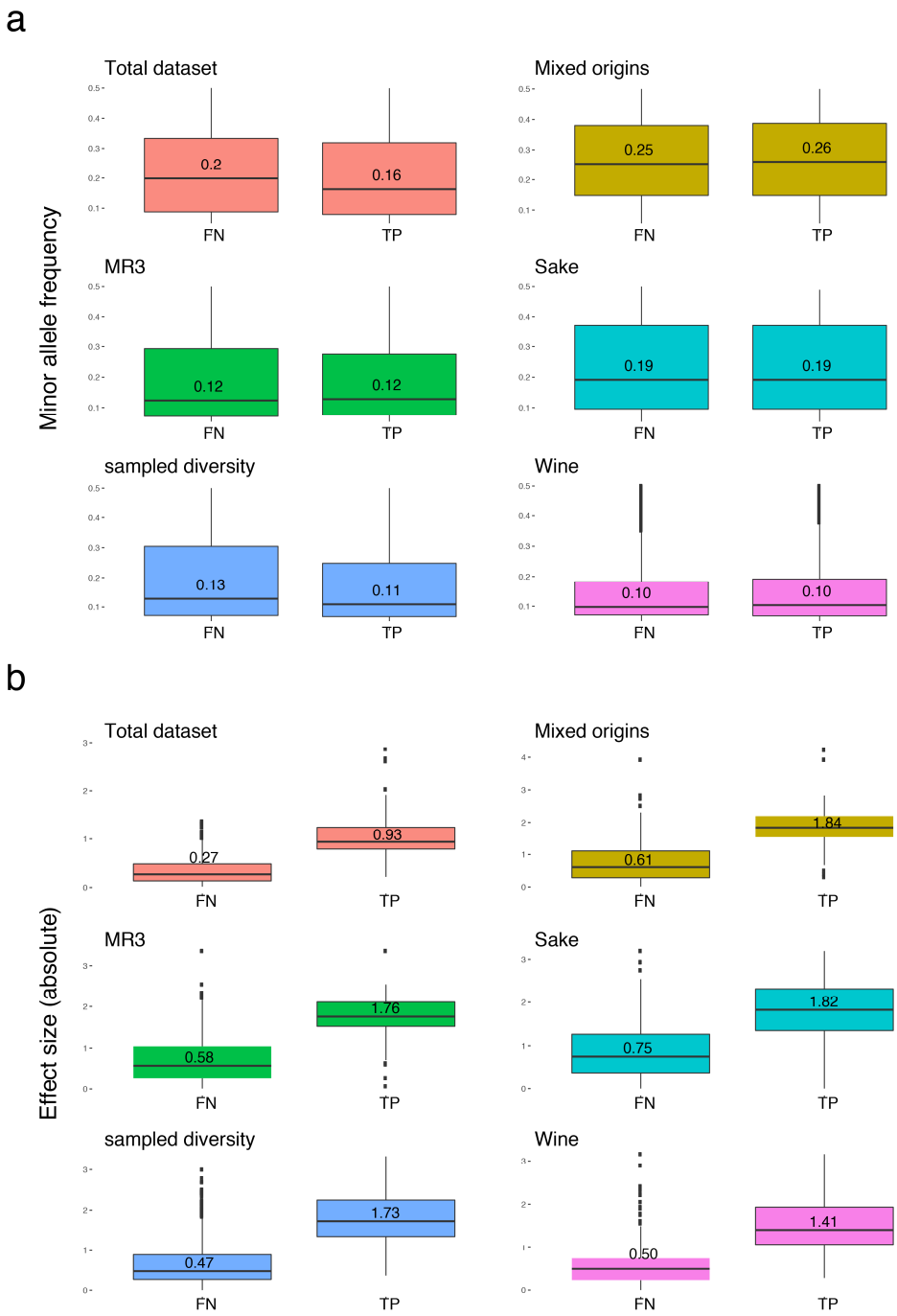


Figure 8: Evaluation of parameters influencing the power of detection. **a.** Minor allele frequency doesn't seem to influence the power of detection (**a.**) whereas variants with low effect size are harder to detect (**b.**)

Also, the p-values of the association tests attributed to the causal SNPs are positively correlated with the effect size for all datasets, indicating that SNPs with high effect sizes are definitively more likely to have high scores of association (Figure 9). Taken together, these results illustrate how the missing heritability can be hidden behind a large number of variants with small effects and that datasets with higher sample size are more likely to detect the causal variants.

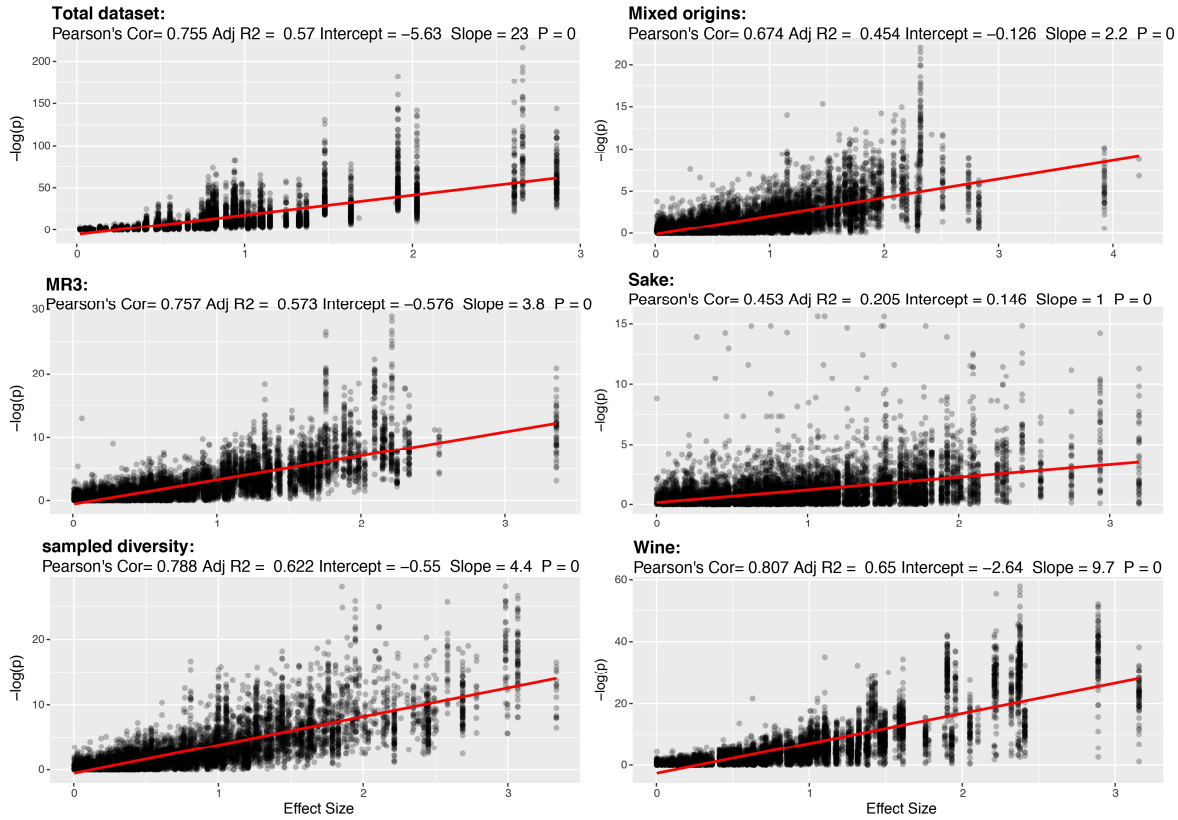


Figure 9: Association scores in function of effect size. Pearson's correlation is indicates above, as well as the coefficient of determination, the slope and the y-intercept value of the linear correlation

Dataset composition impacts the false positive rate

Surprisingly, the median of the FPR for each dataset is lower than when we simulated a Mendelian trait, and the order of the dataset sorted by variance is almost the same (Figure 4 and 10). The median values are low and range from 2.75e-05 for the mosaic dataset to 3.74e-04 for the 1,011 strains, but again, the variance is probably a better indicator to sort the datasets (Figure 10). As for the Mendelian traits, the sampled diversity and total subsets do not present association test with high values of FPR, indicating that the chance that these datasets will detect spurious associations is lower. By contrast, the sake subpopulation cannot give satisfying results, as the number of false positives it yields is too high. A total of 17 association tests using the sake subpopulation show a FPR of more than 10%.

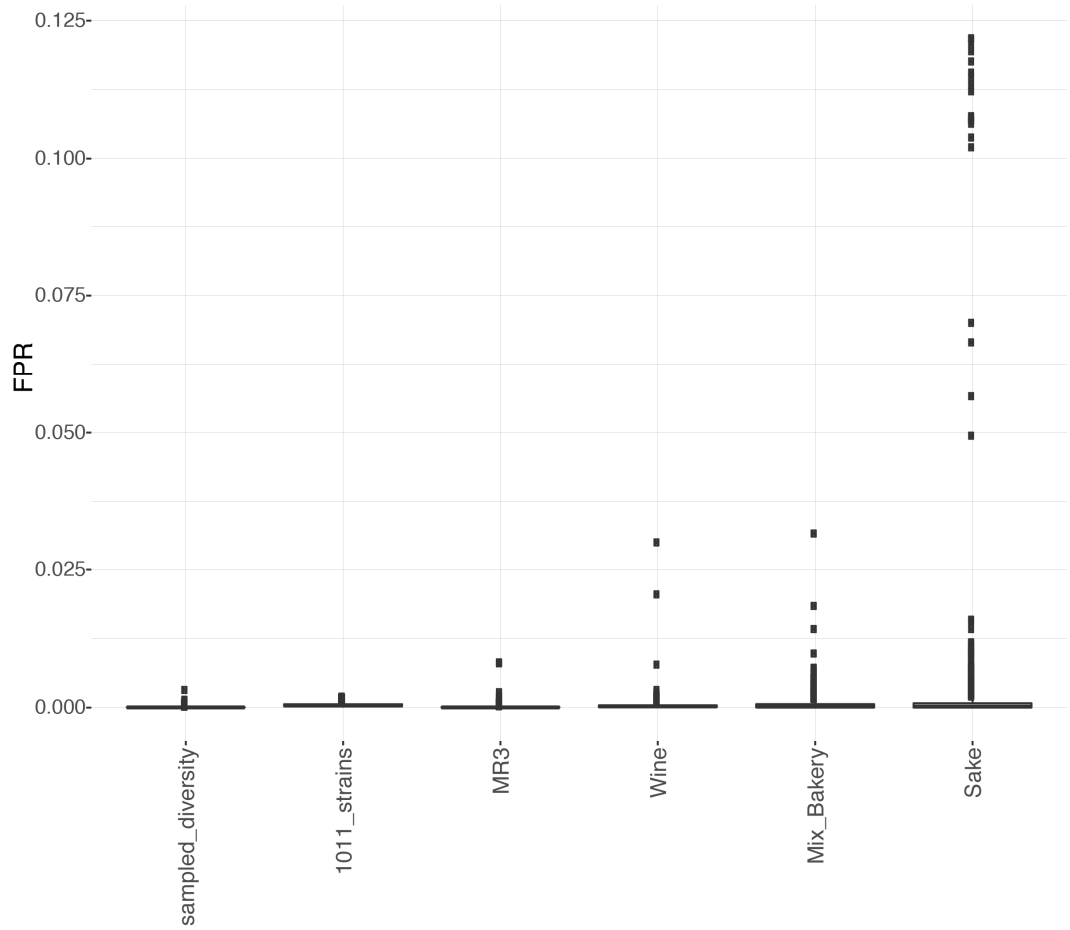


Figure 10: False positive rate by dataset for the detection of a complex trait.

By looking at the distribution of the genomic inflation factor λ of the 1,000 runs in each dataset, we can observe that the distributions that are showing values closest to 1 are the sampled diversity, the mixed dataset and the 1,011 strains. Generally, all datasets have λ values close or inferior than 1, except the sake cluster, for which the distribution is very wide, supporting once more the fact that we expect a lot of false positives using this dataset. By contrast, the distribution of the GIF for the sampled diversity dataset is very narrow and centered around 1, indicating that this dataset is expected to show few false positives (Figure 11a). Also, it seems that there is an increase of the FPR together with the TPR, but this tendency is less marked when the sample size is bigger (Figure 11b). This latter result illustrates the intuitive fact that causal SNPs might be mapped by chance if there are many SNPs that have been significantly identified, but is also consistent with the previous result showing that a high sample size will reduce the false positive rate. These results show that the sample size is not the only parameter that is influencing GWAS outcome since the sampled diversity dataset shows very good results with only 133 individuals.

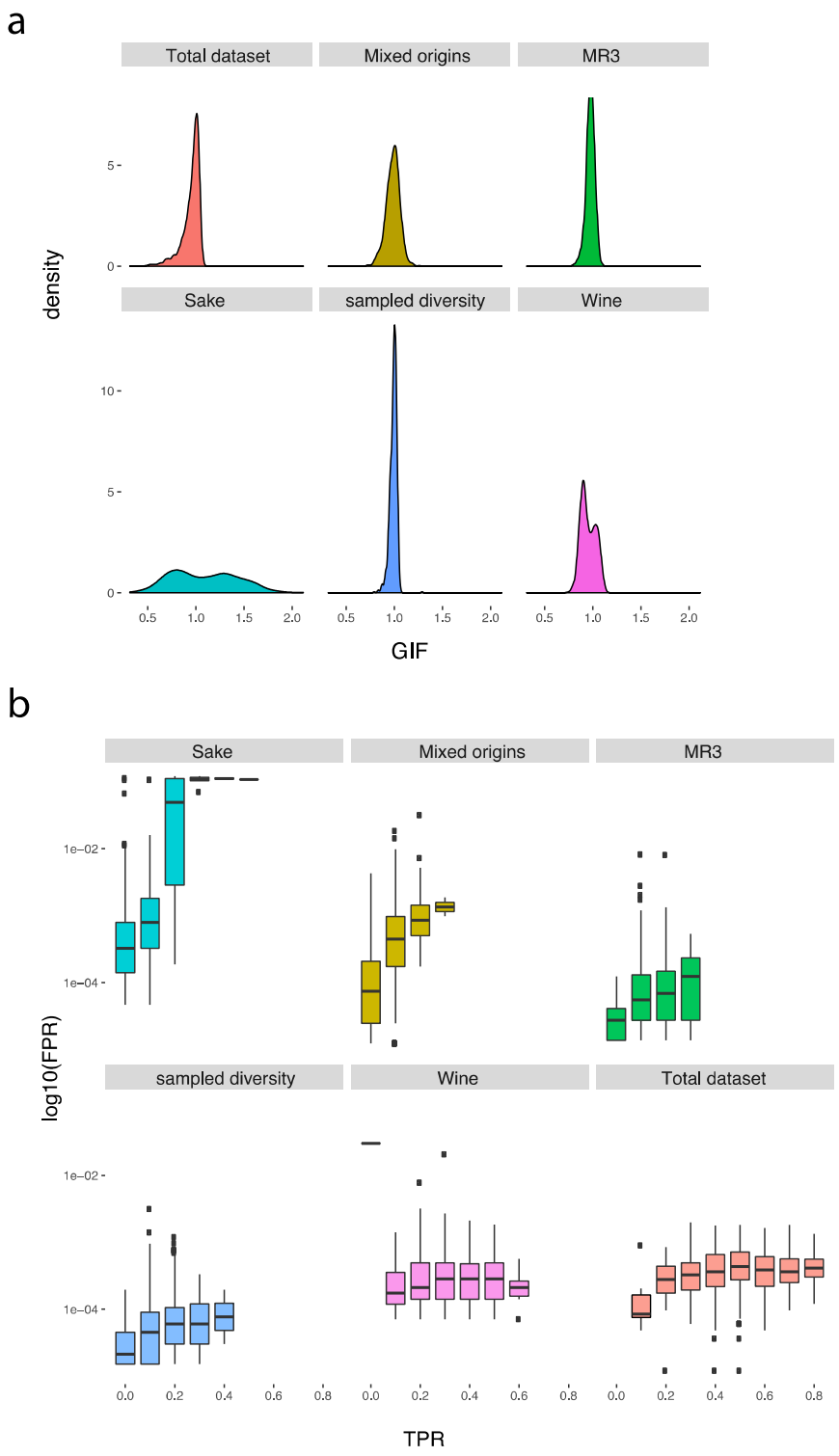


Figure 11: GIF distribution (**a**) and TPR by FPR (**b**) for each dataset.

Association mapping using a growth phenotype

After the evaluation of all the datasets using simulated phenotypes, we wanted to test association with real phenotypic values. To that end, we used the measures of growth for all our isolates on copper sulfate 10 mM normalized by the growth on the standard complete medium (YPD) at 30°C (see Material and Methods). These measures are proxy for the fitness of each isolates on a stress induced by copper sulphate.
We decided to perform association tests on this phenotype because it shows a very high genome-wide heritability for the complete dataset (83.78%), which is in the same order of magnitude as the heritability used to simulate the phenotypes in this study. In addition, the major genetic origin of the resistance to high concentration of copper sulfate in yeast is well known via linkage mapping analysis. A major gene, the gene *CUP1-2*, was mapped in different studies. It is the major copper-activated metallothionine in *S. cerevisiae*. The protein Cup1p binds and sequesters cupruous copper(I), Cu+, allowing the cell to control the copper ion homeostasis. This ion, while essential for yeast survival, is also an environmental heavy metal toxicant at high concentrations, which is used, for example, to kill downy mildew in vineyards[19], where *S. cerevisiae* is often found. Tandem duplications of the *CUP1* gene are common in budding yeasts, and the copper ion tolerance is correlated with the number of copies of this gene[11]. As the copy number variation of the *CUP1-2* gene is known to be of major importance for this trait, it is expected that association with well-suited datasets should easily detect this variant. For this purpose, we added the CNVs of the 1,011 genomes to our matrices.

Association has been performed with the same method used for the simulations with a dataset-specific threshold determined by 100 runs of permutations. The copy number variant of *CUP1-2* has not been detected in three datasets: the sake, mix bakery and the mosaic subpopulations, which are the 3 datasets that had the lowest true positive rates in our simulations. The mosaic and the sake subpopulations also show associations with other variants that are not known to be involved in the resistance to copper and probably falsepositives (Table 2). In the other datasets, the *CUP1-2* copy number variant has been significantly associated with the phenotype, with no other significant association. It was expected to be found in the wine subpopulation because the acquisition of resistance to copper sulphate is certainly reflecting convergent evolution due to human selection for industrial processes[20], thus emphasizing the fact that dataset composition is also depending on the phenotype of interest. The score of association is an indicator of the ease with which a dataset is able to detect the causal variant. In this case, the total dataset shows the lowest P-value (4.86 e-44), followed by the wine cluster (6.03e-18) and the sampled diversity dataset (1.04e-14). These results are consistent with our simulations, as the total dataset and the sampled diversity dataset were the ones that showed the best results.

Dataset	Hit localization	P-value	Threshold
Total dataset	*CUP1-2*	4.86e-44	5.20e-07
Wine	*CUP1-2*	6.03e-18	4.64e-07
Sampled diversity	*CUP1-2*	1.04e-14	8.45e-08
Sake	*ARP10*	2.77e-09	1.91e-06
Sake	*tRNA-Ile*	2.77e-09	1.91e-06
Sake	*tRNA-Ile*	2.77e-09	1.91e-06
MR3	*APL2*	3.62e-08	2.02e-07
MR3	*APL2*	3.62e-08	2.02e-07
Sake	*SPO71*	5.61e-08	1.91e-06
Sake	*SPO71*	5.61e-08	1.91e-06
Sake	*ERG5-SOC2*	5.61e-08	1.91e-06
Sake	*MYO5*	5.61e-08	1.91e-06
Sake	*MATALPHA1*	1.08e-07	1.91e-06
Sake	*IRC6*	2.53e-07	1.91e-06

Table 2: GWAS using growth in presence of copper sulphate. The variants shown are the ones that reached significance (association tests P-value superior than threshold determined by permutations).

As previously stated, the detection of the *CUP1-2* copy number variation for the wine subpopulation is attributed to the evolutionary history of this population, and its link with human activity. This result underlines the fact that even if the results of the simulations were not as convincing as for the total dataset, some populations might be relevant with the trait of interest and might be considered to perform association.

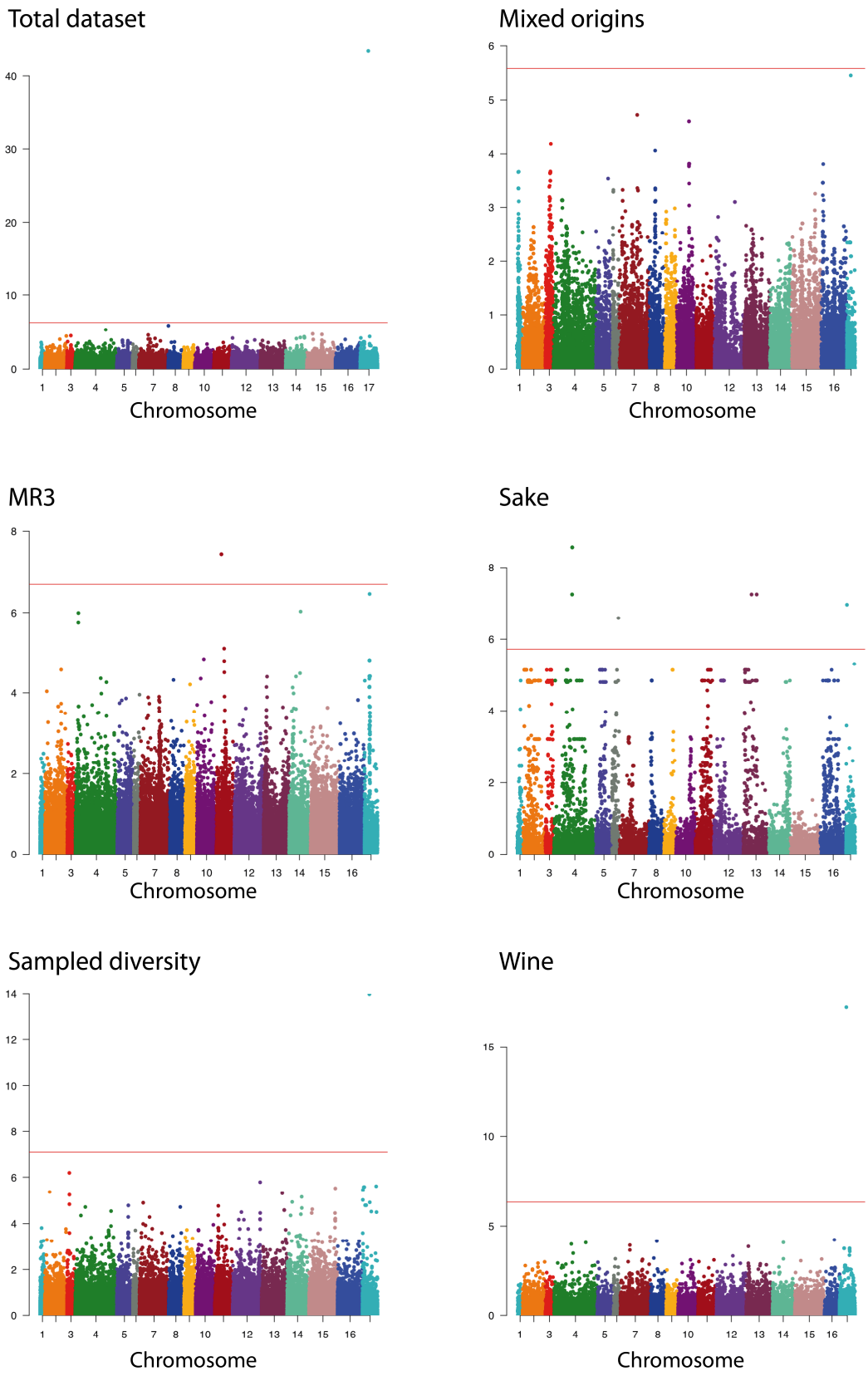


Figure 12: Manhattan plots of the genome-wide association with growth in presence of $CuSO_4$. The chromosomes are represented on the x-axis from 1 to 16 with an additional chromosome for the CNVs. The threshold in red is obtained with 100 permutations of the phenotypes.

Discussion

In this work, we performed extensive simulations of genome-wide association and we measured the capacity of linear mixed models to find significant association between causal genetic variants and simulated phenotypes. These association tests have been performed on five various *S. cerevisiae* subsets as well as the complete dataset of more than 1,000 genomes. While Mendelian traits are easily mapped independently from the dataset used, we can already see discrepancies in terms of type-1 error. Because the sake subpopulation is composed of very closely related isolates sharing a recent common ancestor, a high type-1 error rate is observed. The wine subpopulation is supposed to have somewhat the same problem to a lesser extent because of higher sample size and a less recent common ancestor. Concerning the mapping of complex traits, variation between datasets is already found in terms of detection. While the total dataset shows the best performances in terms of detection, the sampled diversity dataset shows also a low type-1 error rate, even though its sample size is reduced compared to the wine cluster or the total dataset, indicating that confounding factors can also be handled with well-suited datasets. Finally, we tried association mapping using a real phenotype, more precisely on the ability to grow in presence of copper sulphate and checked if we could identify the already well known CNV of the *CUP1-2* gene responsible for the variation in fitness for this condition. This CNV has been successfully identified in three datasets with no other associations while it has been missed in the three others. The detection of the causal variant was consistent with our simulations and showed us that some specific population might represent good dataset to perform associations with relevant phenotypes.

GWAS have been problematic in *S. cerevisiae* due to the highly stratified populations, which led to high type-1 error rates[9,10]. In this work, we showed that structure is well accounted for in most of the dataset by a LMM approach, as the type-1 error rates are low for most of the datasets. Our hypothesis is that sample size was the limiting factor in the aforementioned studies. Indeed, the type-1 error rates of our datasets with the lowest sample sizes were among the highest. But our results support the fact that other parameters are impacting GWAS outcomes, as the performance is not always in correlation with the sample size. For example, it seems that it is important to build datasets with individuals that do not share a recent common ancestor, as this relatedness will introduce a bias in allele frequencies that will also lead to spurious associations, as observed for the sake and the wine subpopulations. It is difficult to estimate the relative impact of these parameters individually, as each dataset will display its unique combination of parameters that will influence GWAS, and our simulations do not allow disentangling them.

The results of our associations vary widely from one cohort to another, allowing us to evaluate the limits of GWAS and to present what would make an ideal dataset for association studies. For the mapping of a Mendelian trait, the sampled diversity set is the best, with the lowest false positives. It is then followed by the total dataset with 1,011 strains. When the trait is complex, the whole population has the advantage to map more causal variants than any other. Also the false positive rate is close to the one we observe for the sampled diversity

dataset, which has, as for a Mendelian trait, the lowest value, but the total dataset has the advantage to retain more markers, which is important to find the causal markers when one wants to perform association using a real phenotype. It is now evident that the sake cluster is definitely poorly suited to perform GWAS. First, the number of sequences composing this dataset is too low and thus does not provide enough statistical power for trustworthy associations. Second, the strains composing this cluster are very close to each other, and are all sharing large parts of the genome du to the fact that the common ancestor is very recent compared to other datasets. For these reasons, the sake cluster then produces spurious associations of numerous markers, as the distinction between the ones that are purely due to ancestry and the ones really associated with the phenotype is impossible. We initially thought that the wine population, being a clear lineage and composed of a high number of strains (n=323), would constitute a good sample to perform GWAS as the population is not stratified. In fact, the results are already very noisy, even for the mapping of a Mendelian trait. The number of false positives by run is very variable, it has the 2nd and 3rd highest variance for the FPR for a Mendelian and complex trait, respectively, and therefore this dataset's results are to be taken with caution. This can be explained due to the influence of human selection for this cluster, generating a population bottleneck from which all wine strains are derived. This recent common ancestor is biasing allelic frequencies for this subpopulation the same way as previously described for the sake cluster. The results of association with the real phenotype showed us that the wine cluster was a well-suited subset to map the causal locus, due to the population history. This means that the best dataset to perform GWAS also depends on the phenotype that is tested, and that, depending on the need, one might need to consider building a specific dataset for a specific phenotype.

Conclusion and perspectives

This work sought to establish the feasibility and the limits of GWAS on the model organism *S. cerevisiae*, as well as the influence of the population used to perform association.
To go further, it would be interesting to test if higher sample sizes would allow to improve the detection of causal loci. Indeed, as there is a skew towards low frequency alleles, a dataset with higher sample size could be able to identify rarer genetic variants, and therefore increase the proportion of phenotypic variation explained by the associated variants. This improvement is possible due to the small genome of *S. cerevisiae* together with the always-diminishing cost of genome sequencing. Increasing the sample size will also allow building more subsets, while keeping a high sample size, and could be of great help when association is tested with a specific phenotype, which will make the phenotyping easier and still allow to identify statistically relevant associations.
Another solution would be to perform GWAS on an inbred population resulting from a diallel cross, *i.e* all the pairwise crosses between parental accessions for which the genomic sequence are known. This strategy offers several advantages that might be worth to consider for further studies. First, the genotypes of the hybrids can be constructed from the parental genotypes,

meaning we only need to sequence the parental strains. Second, we expect a good resolution, as the parental haplotypes will be shuffled. Third, this mating scheme will also modify the MAF, allowing rarer genetic variants to be detected with association mapping. Further important improvement would be to include other types of genetic variation in the genetic matrix. For example, it has been shown that CNVs can explain a large part of phenotypic variation in human disease (see [21] for a review) or other traits, like the rapid spreading of glyphosate resistance among the weed plant *Amaranthus palmeri*[22], or fitness traits in yeast[12]. Integrating aneuploidies could be an interesting improvement, as this type of variation results in imbalanced gene dosage with noticeable effect on several traits, for example Down's syndrome or tumour cells in human (see [23] for a review), *A. thaliana*'s seed fertility[24], or cell wall integrity for *S. cerevisiae*[25]. This could be particularly interesting for *S. cerevisiae*, as this type of genetic variation is more tolerated in yeast (or plants) than in animals, and therefore more common. The integration of these other genetic variants will give us a more precise view of genetic variation and is likely to reduce the missing heritability.

References

1. Pantalião, G. F. *et al.* Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. *Genetica* **144,** 651–64 (2016).
2. Francisco, M. *et al.* Genome wide association mapping in *Arabidopsis thaliana* identifies novel genes involved in linking allyl glucosinolate to altered biomass and defense. *Front. Plant Sci.* **7,** 1010 (2016).
3. Ivanov, D. K. *et al.* Longevity GWAS using the *drosophila* genetic reference panel. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **70,** 1470–78 (2015).
4. Lavinsky, J. *et al.* Genome-wide association study identifies *NOX3* as a critical gene for susceptibility to noise-induced hearing loss. *PLoS Genet.* **11,** 1–21 (2015).
5. Goffeau, a *et al.* Life with 6000 genes. *Science.* **274,** 546, 563– 7 (1996).
6. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387–91 (2002).
7. Fay, J. C. The molecular basis of phenotypic variation in yeast. *Curr. Opin. Genet. Dev.* **23,** 672–7 (2013).
8. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science.* **353,** 1381–96 (2016).
9. Connelly, C. F. & Akey, J. M. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191,** 1345–53 (2012).
10. Diao, L. & Chen, K. C. Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics* **192,** 1503–11 (2012).
11. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* **125,** 762–74 (2015).
12. Peter, J. *et al.* Yeast evolutionary history and natural variation revealed by 1,011 genomes. *In revision* (2017).
13. Ohnuki, S. *et al.* Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. *G3 (Bethesda)* **7**, 2807–20 (2017).
14. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–5 (2011).
15. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
16. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–44(2015).
17. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542,** 186–90 (2017).
18. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–86 (2014).
19. McBride, M., Tiller, K. & Merry, R. Copper in soils and plants. *Acad. Press. Sydney* (1981).
20. Warringer, J. *et al.* Trait variation in yeast is defined by population history. *PLoS Genet.* **7,** (2011).
21. Zhang, F., Gu, W., Hurles, M. & Lupski, J. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* 451–81 (2009).
22. Gaines, T. A. *et al.* Gene amplification confers glyphosate resistance in A*maranthus palmeri*. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1029–34 (2010).
23. Giam, M. & Rancati, G. Aneuploidy and chromosomal instability in cancer: a jackpot

to chaos. *Cell Div.* **10,** 3 (2015).
24. Henry, I. M., Dilkes, B. P., Miller, E. S., Burkart-Waco, D. & Comai, L. Phenotypic consequences of aneuploidy in *Arabidopsis thaliana*. *Genetics* **186,** 1231–45 (2010).
25. Dodgson, S. E. *et al.* Chromosome-specific and global effects of aneuploidy in *Saccharomyces cerevisiae*. *Genetics* **202,** 1395–409 (2016).

Chapter 3: Genome-wide association on a diallel hybrid panel

Introduction

The heritability of a trait describes how genetic components contribute to phenotypic variance. We distinguish broad-sense heritability (H^2), which corresponds to the total genetic variance of a trait, and narrow-sense heritability (h^2), which is the proportion explained by additive genetic components[1]. Understanding trait heritability represents an essential step towards the elucidation of the genotype-phenotype relationship and requires accurate estimations of the contribution of additive as well as non-additive effects.

Genome-wide screens for genetic variants associated with phenotypic variation increased our understanding of the genetic architecture of complex traits. Up to date, genome-wide association studies have identified more than 250,000 SNPs associated with human complex traits and common diseases[2]. These studies underline some trends for trait-associated genetic variation. First, most traits are governed by a large number of loci, with some of them explaining a small proportion of phenotypic variation. Second, associated loci fail to explain the totality of the genetic contribution to the trait, giving rise to the phenomenon called missing heritability. For example it has been shown recently that using genomic data from over 700,000 individuals, around 700 variants have been significantly identified as playing a role in the adult human height, but they only explain 20% of the heritability for this trait, which is estimated to be around 80%, depending on the estimates[3–5]. Several hypotheses have been formulated to explain the sources of missing heritability, such as (i) the difficulty to detect associations with small effect, (ii) the possibility that causal loci might be rare alleles, (iii) the fact that other type of genetic variants than SNPs, such as structural variants, might have a genetic contribution to a trait, (iv) the fact that GWAS are largely biased towards the identification of additive variants due to computational and statistical difficulties, or (v) the possible phenotypic impact of epigenetic modifications[6].

The investigation of rare and low-frequency variants (MAF between 0.5 and 1%) responsible for trait variability is of particular interest. Indeed, as evolutionary theory predicts, deleterious alleles are likely to be rare due to purifying selection. Moreover, loss-of-function variants, which prevent the generation of functional proteins, are especially rare[7,8]. Rare variants play an important role in human diseases (see [9] for a review). While it is possible to perform GWAS with rare variants, the sample size required to provide the necessary statistical power is very high, and thus very expensive to obtain.

A diallel cross is a type of mating system that involves all possible crosses among a group of parents. Plant or animal breeders have used it historically as a reliable method to obtain overall information on average performance of individual inbred lines. It is also useful to investigate the genetic causes underlying quantitative traits.

Here, we performed GWAS on a half-diallel mating panel using 34 parental strains, which were completely sequenced in the frame of the 1,002 yeast genomes project[10]. We constructed F1 genotypes *in silico* by combining known parental genotypes, resulting in a population of 595 individuals (see Material and Methods). This method constitutes a cheap and elegant way to build large genomic dataset from a small number of sequenced strains. Moreover, due to the design of a diallel mating scheme, all haplotypes are represented several times in the

matrix, which leads to a shift of minor allele frequency of alleles that were rare in the parental population to more frequent alleles in the F1 progeny. Also, we used different encodings to test whether we could identify additive genetic variation or deviation of progeny from parental mean. More precisely, we used a nonconventional SNP encoding, where both homozygous classes are represented the same way, as opposed to heterozygotes encoding, allowing us to test if a single mutation in heterozygous state is causal. We identified several significant associations in both the additive and the overdominant models. We also show an example of the identification of a rare allele, that fell under the minor allele frequency threshold of 5% in the 1,011 genomic sequences, reinforcing the potential power of such an approach for GWAS.

Results

Experimental design and model selection

One major benefit of performing GWAS on a diallel panel is the relatively low sequencing cost compared to classical studies with high number of genomes. Here, only parental genomes are required to generate *in silico* the genomes of every hybrid by merging the genotype of both parents (see Material and Methods). Based on 34 parental genomes, we could construct the population of 595 F1 hybrid genotypes matching one half matrix of the 561 hybrids resulting from pairwise crosses of the parents, plus the 34 homozygous diploids (Figure 1).

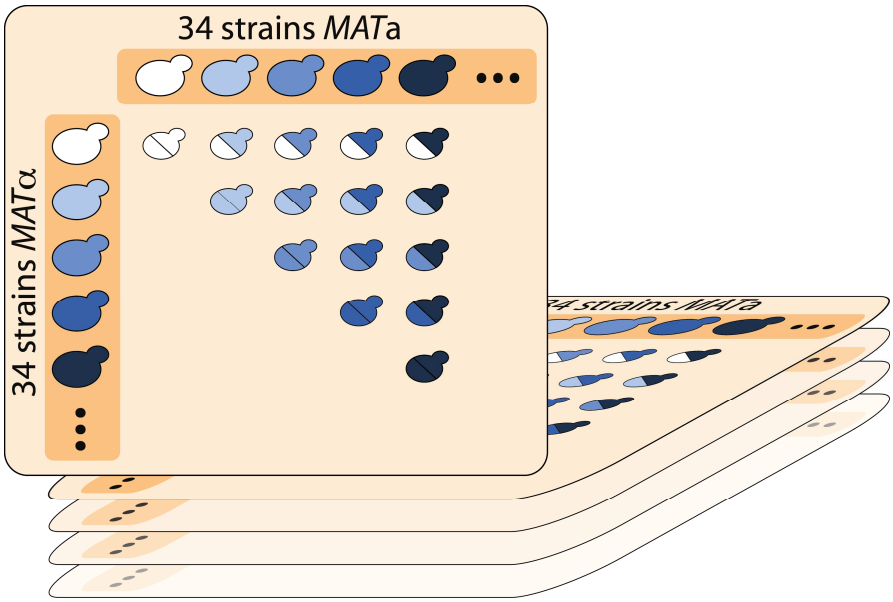


Figure 1: Diallel mating scheme. Half-matrix of all pairwise crosses from homozygous parental strains, with genotypes indicated by the colors of each individual.

Strain Name	Ecological origin	Geographical origin
DBVPG1564	Wine	Dolianova, Sardinia
EXF-5248	Water	Koper, Slovenia
DBVPG3591_1b	Nature	NA
CECT10109_1b	Nature	Spain
1560	Nature	Spain
UCD_09-448	Nature	California, USA
I14_1b	Wine	Italy
CLIB413_1b	Fermentation	China
CLIB382_1b	Beer	Ireland
2162	Soil	Hungary
2187	Soil	Hungary
EXF-7197	Tree	Montenegro
sample 40	Tree	Bordeaux, France
CLIB1071	Cider	Normandy, France
YJM627	NA	France
CLQCA_04-021	Insect	Los Rios, Ecuador
ZP_611	Tree	Canada, Vancouver,
YJM326_b	Human, clinical	California, USA
HN10	Nature	Hainan province, China
BJ20	Tree	Beijing, China
NPA03.1	Palm wine	Nigeria
HN15	Nature	Hainan province, China
HN16	Soil	Hainan province, China
CLQCA_20-184	Flower	Yasuni, Ecuador
CLQCA_20-246	Insect	Yasuni, Ecuador
YPS617	Tree	Pennsylvania, USA
ES4M07	Fruit	Nantou, Taiwan
YPS163	Soil	Pennsylvania, USA
YPS143	Wine	Pennsylvania, USA
NC_02_b	Tree	North Carolina, USA
Y12_1b	Palm wine	Ivory Coast
YJM421_b	Human, clinical	USA

Table 1: Ecological and geographical origins of the 34 parental strains.

The parental strains have been selected to be representative of the genetic diversity, with various geographical and ecological origins (Table 1). We built a matrix of genetic variants for this panel and filtered SNPs to only retain biallelic SNPs with no missing calls. Close to 150,000 SNPs were matching those criteria. However, due to the few number of parental genotypes, there is extensive long distance linkage disequilibrium (LD) that had to be removed. To that end, filtering had to be performed based on the basis of SNP pattern sharing across the population (see Material and Methods), drastically reducing the matrix to 31,632 SNPs.

In order to be able to map additive as well as non-additive variants impacting phenotypic variation, we performed GWA using two different models. First, we used a classical additive encoding for SNPs where linear relationship between trait and genotype is searched, *i.e.* every locus has a different encoding for each genotype (homozygous for minor allele encoded as 1-1, heterozygous as 1-2 and homozygous for major allele as 2-2). To account for non-additive inheritance, we also used an overdominant model, which only accounts for differences between heterozygous and homozygous positions, thus allowing to reveal overdominance and dominance effect. In this case, SNPs were encoded as 1-1 for both homozygous and 1-2 for heterozygous.

Overview of results of the association tests

We ran GWA for each of the two models described above using diploid fitness for 53 growth conditions as input phenotype. The selected conditions impact multiple cellular and molecular pathways such as transcription, translation, osmotic stress, oxidative stress, metal and drug resistance as well as carbon sources or temperature (Table 2).

Categories	Subcategories	Conditions
Comparer		SC
Cell Wall	**Membrane stability**	SC SDS 0.01%
		SC SDS 0.025%
		SC SDS 0.05%
		SC nystatin 5µg/ml (polyene)
		SC nystatin 10µg/ml (polyene)
		SC nystatin 25µg/ml (polyene)
	Ergosterol synthesis	SC fluconazole 1µg/ml (triazole)
		SC fluconazole 5µg/ml (triazole)
		SC fluconazole 10µg/ml (triazole
	Ergosterol synthesis + multiple targets	SC ketoconazole 10µg/ml (imidazole)
		SC ketoconazole 30µg/ml (imidazole
		SC ketoconazole 60µg/ml (imidazole
Low T°		SC 14°C
DNA metabolism	**Telomere dynamics**	SC sodium (meta)arsenite 1mM (SMA)
		SC sodium (meta)arsenite 2.5mM (SMA)
		SC sodium (meta)arsenite 5mM (SMA)
	DNA damage	SC 4-NQO 1µg/ml
		SC 4-NQO 2µg/ml
		SC 4-NQO 3µg/ml
	DNA synthesis	SC 5-FU 50µg/ml (pyrimidine analog)
		SC 5-FU 100µg/ml (pyrimidine analog)
		SC 5-FU 250µg/ml (pyrimidine analog)
General		SC CuSO4 0.1mM

cellular damage		SC CuSO4 0.5mM
		SC CuSO4 1mM
Metabolism	**Carbon sources utilization**	SC galactose 2%
		SC glycerol 2%
	Carbon starvation	SC glucose 0.01%
		SC galactose 0.01%
		SC glycerol 0.01%
	High carbon source tolerance	SC glucose 10%
		SC galactose 10%
		SC glycerol 10%
Osmotic stress		SC NaCl 0.5M
		SC NaCl 1M
Oxydative stress		SC methyl viologen 0.5mM
		SC methyl viologen 1mM
		SC methyl viologen 2.5mM
Protein stability		SC formamide 1%
		SC formamide 2%
		SC formamide 5%
Signal transduction pathways		SC caffeine 10 mM
		SC caffeine 20 mM
		SC caffeine 40 mM
Subcellular organization	**Vacuole function**	SC bafilomycin B1 1µg/ml
	Microtubules function	SC benomyl 50µg/ml
		SC benomyl 100µg/ml
Translation	**Ribosomes function**	SC cycloheximide 0.1µg/ml
		SC cycloheximide 0.25µg/ml
		SC cycloheximide 0.2µg/ml
Transcription	**GTP and UPT nucleotide pools**	SC 6-azauracil 50µg/ml
		SC 6-azauracil 100µg/ml
		SC 6-azauracil 200µg/ml

Table 2: Phenotyping conditions and the cellular and molecular pathway they affect

Taken together, GWA results revealed 2,293 significantly associated SNPs, with 1,523 (mean of 29 by condition) and 770 (mean of 15 by condition) for overdominant and additive model, respectively. We detected from 1 to 103 significant SNP by condition. Besides, 223 SNPs are significantly associated in both additive and overdominant model thus reinforcing their potential implication in phenotypic variation. Association scores also clearly depicts that loci associated with the overdominant model display significantly higher scores than the ones with additive model (Welch t-test, P-value = 1.8 $x10^{-3}$). However, highest association scores are with the additive model for conditions such as Methyl-viologen or caffeine, for example (Figure 2).

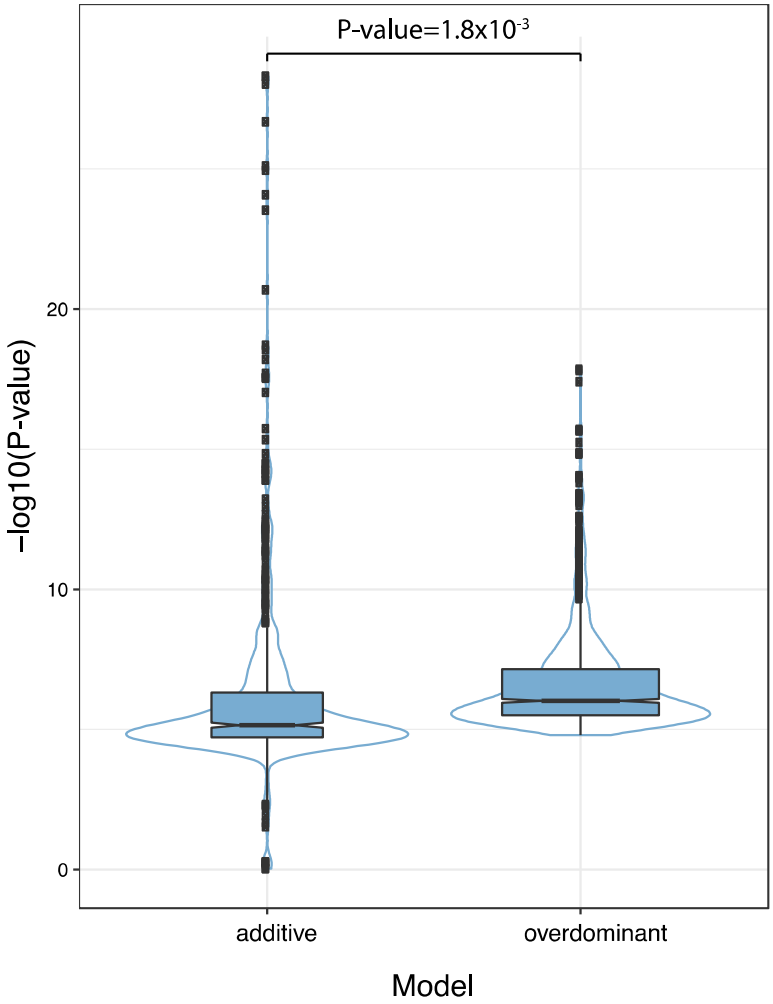


Figure 2: Distribution of P-values of association for the additive and the overdominant model

Identification of alleles with pleiotropic effects

Throughout all the genotype-phenotype associations found, 467 SNPs are significantly associated with more than one condition. As we often used up to 3 concentrations for the same molecule and/or condition, it is understandable that a SNP would be found for several concentrations. However, 127 SNPs are involved in more than 3 conditions with some up to

10 different conditions, regardless of the model used. This result highlights the presence of SNPs that may play a role in major yeast stress response pathways. For example, we detected 15 genotype-phenotype associations involving a SNP located in the coding sequence of *MIG3*, a transcriptional regulator involved response to toxic agents, suggesting the role of this gene in the fitness for these conditions.

Diallel design can be used to identify rare variants

One major motivation to use a diallel cross design to perform GWAS is that we can use the redundancy of the haplotypes which is intrinsic of the pairwise crosses to our advantage. Indeed, for a half diallel mating scheme based on 34 parents, each genotype will be represented at least 34 times in the progeny once for each hybrid deriving from this parent. This high level of haplotype shuffling and repetition gives the advantage of offering allele overrepresentation compared to the use of a population with the same number of independent individuals. Minor allele frequency (MAF) will be largely changed in a diallel compared to the species level because of the smaller number of parents involved in the population resulting from the diallel mating scheme. To exceed the classical MAF threshold of 5% for a SNP to be included in GWAS, only two out of the 34 parental genomes need to possess the minor allele. In our diallel panel, out of the total 31,632 SNPs retained (see Material and Methods), 3.5% (1,118) which had a MAF < 5% in the 1,011 *S. cerevisiae* genomes happen to surpass this threshold in the diallel, going up to a MAF of 32%[10]. Surprisingly, 12% (279) of the significantly associated SNPs also surpassed this threshold. It is obvious that the remapping of the MAF and thus the enrichment in rare variant discovery we observed here is entirely dependent of the parental sampling. However, a priori strain sampling based on a particular phenotype may be done to potentially enrich dataset in rarer variants that might play a role in the phenotype of interest.

A concrete example of the detection power of our design can be observed in the presence of 2% galactose as a carbon source. We detected 38 and 31 significantly associated SNPs with additive and overdominant model respectively, with 17 found in both models. Some of these SNPs are in genes related to galactose metabolism, diauxic shift, pentose phosphate pathway, TCA pathway and mitochondrial genes. Six of the associated SNP did not have a MAF superior 5% at the population level. Among them, we find the best hit, whose MAF is close to 9% in hybrids. It lies in the coding sequence of *GAL2*, a galactose permease on chromosome 12. Three distant strains are carrying this variant, namely CLIB1071, HN10 and UCD_09-448, suggesting that they have not been inherited from the same ancestor. Strikingly, this SNP alone explains 10.5% of the total phenotypic variance observed. This effect is confirmed when comparing fitness distributions for each genotype. Hybrids carrying at least one copy of the minor allele of this gene exhibit an improved fitness compared to strains homozygous for the major allele (Welch t-test, P-value = 2.9×10^{-16}) (Figure 3). However, no assumption regarding the overdominant or dominant character of the heterozygous combination of allele can be made. Indeed, given the small number of strains carrying this specific variant, no

statistical significance can be attained between phenotype of homozygous for this allele and the heterozygous.

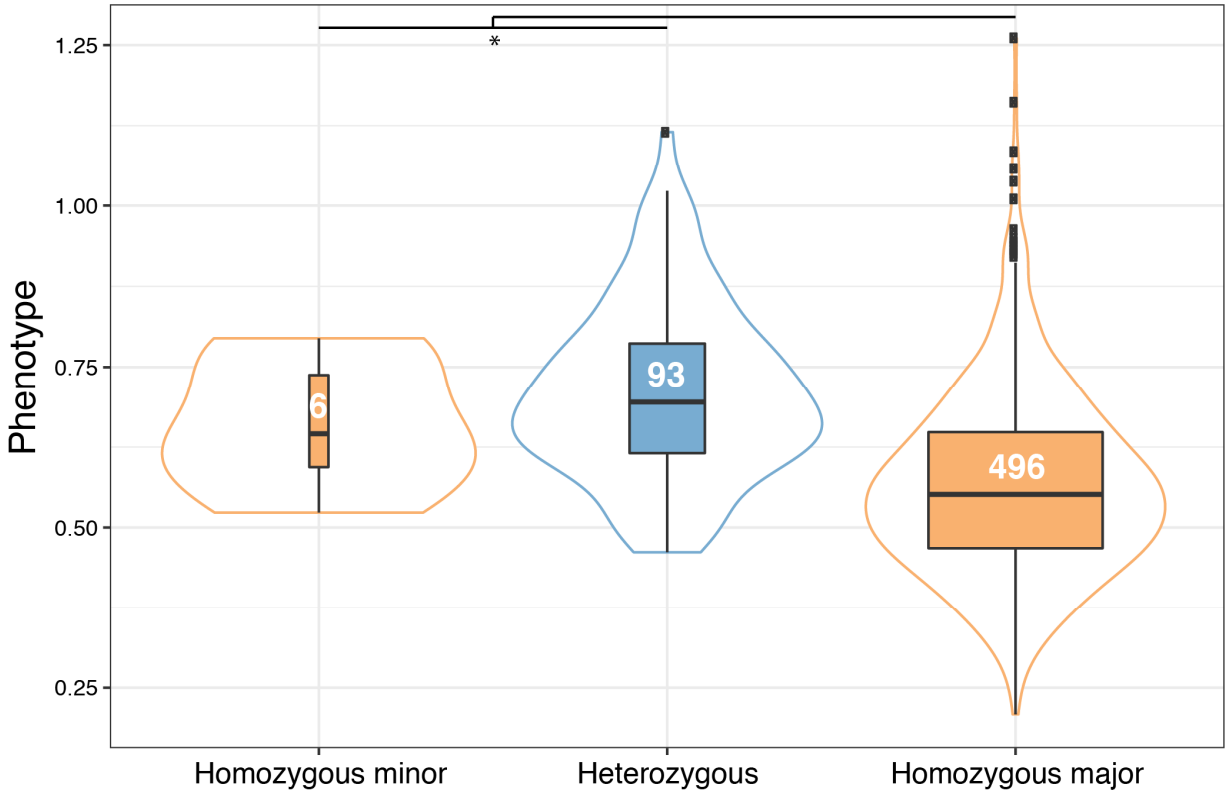


Figure 3: Distribution of phenotypic values by genotype.

The case illustrated by this genetic variant confirms the point that rare variant with important phenotypic effect can be efficiently detected by GWAS on a diallel panel, thus unveiling part of the missing heritability of traits alleged to be undetectable by this standard mapping strategy.

Conclusion

With the generation of a diallel mating scheme based on 34 parental strains with known genome, we could investigate the genotype-phenotype relationship of *S. cerevisiae* 595 individuals without the costly sequencing of those strains. We used an overdominant encoding in addition to the standard SNP encoding, allowing us to consider non-additive effects. Moreover, such a mating design can be used to identify rare genetic variants underlying phenotypic variation as it has been demonstrated here with the rare SNP located in the coding sequence of *GAL2* that bears a large effect. Taken together, these results demonstrate the potential power of using a diallel cross to perform GWAS, with the identification of 557 variants acting additively, 1,300 in an overdominant way as well as 223 found in both models. A careful selection of parental strains in accordance to the phenotype of

interest might therefore be an elegant way to go deeper into the genotype-phenotype relationship.

References

1. Falconer, D. S. & Mackay, T. F. C. *Introduction to quantitative genetics*. (Harlow, 1996).
2. Li, M. J. *et al.* GWASdb v2: An update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **44,** 869–76 (2016).
3. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* **542,** 186–90 (2017).
4. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–86 (2014).
5. Silventoinen, K. Determinants of variation in adult body height. *J. Biosoc. Sci.* **35,** 263–85 (2003).
6. Manolio, T. a *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (2009).
7. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
8. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science.* **335,** 823–28 (2012).
9. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18,** 77 (2017).
10. Peter, J. *et al.* Yeast evolutionary history and natural variation revealed by 1,011 genomes. *In revision* (2017).

Material & Methods

Sequencing, mapping and quality control

Saccharomyces cerevisiae sequenced isolates

The isolates included in this project were carefully selected to be representative of the *Saccharomyces cerevisiae* whole species. For this purpose, we maximized the ecological origins of the isolates: human-associated environments such as wine and sake fermentation, brewing, dairy products, as well as natural environments such as soil, insects, tree exudate and fruit are represented. Geographical origins are also highly diverse and distributed worldwide. In addition to the 918 isolates kindly provided from research laboratories and yeast collections, we have also included 93 strains sequenced in previous studies[1–3]. Finally, a total of 1,011 samples were analyzed throughout in this study. We sought to keep the studied isolates in their natural state before sequencing to provide a global picture of the ploidy and level of heterozygosity within the species. However, among the 918 selected isolates, 124 were non-natural haploid with the *HO* gene deleted and the 93 external isolates were genetically manipulated before sequencing.

Sequencing and quality filtering

Yeast cell cultures were grown overnight at 30°C in 15 mL of YPD medium to early stationary phase. Total genomic DNA was subsequently extracted using MasterPure™ Yeast DNA Purification Kit and Genomic Illumina HiSeq 2000 sequencing libraries were prepared for 918 strains with an insert size between 300 and 600 bp. 10 libraries were multiplexed per Illumina HiSeq2000 lane and subjected to paired-end sequencing, producing reads of 102 bases. An in-house quality control process was applied to the reads that passed the Illumina quality filters. Illumina sequencing adapters and primers sequences were removed from the reads and the low-quality nucleotides (Q<20) were discarded from both ends of the reads. Reads shorter than 30 nucleotides after trimming were removed. These trimming and removal steps were achieved using in-house-designed software based on the FastX package. The last step identifies and discards read pairs that mapped to the phage phiX genome (GenBank: NC_001422.1) using SOAP[4]. A total of 3.35 Tb of high-quality genomic sequence was generated with a median sequencing depth of 226X per isolate (ranging from 50X to 1,014X). For the publically available Illumina paired-end reads related to 93 strains (see *S. cerevisiae sequenced isolates* section), the median sequencing depth is 106X (from 20X to 570X).

Reads mapping and variant-calling

For each isolate, the reads were mapped to the *S. cerevisiae* S288C reference genome (version R64-1-1) with bwa (v0.7.4-r385)[5], using default parameters. Duplicated reads were marked with Picard-tools (v1.124) (http://broadinstitute.github.io/picard/) and local realignment around indels and variant calling were performed with GATK (v3.3-0)[6]. Default parameters were applied except for the realignment step (GATK IndelRealigner), where the following parameters were set: '--maxReadsForConsensuses 500 --maxReadsForRealignment 40000 --maxConsensuses 60 --maxPositionalMoveAllowed 400 --entropyThreshold 0.2'.
GATK Variant Annotator was run to add allele balance information in the vcf files.

SNPs filtering and matrix

For each sample, variants were first called with GATK HaplotypeCaller (see *Reads mapping and variant calling* section). At this stage, isolates with less than 5% of heterozygous sites were considered as homozygous. The raw files were then post-processed to deal with highly confident variants to be included in our complete SNPs matrix, based on both coverage and allele balance information:
(i) A minimal coverage depth of 50X was required for a SNP to be retained for the 918 isolates that were sequenced in this study, while it was lowered to 10X for the other 93 previously sequenced
(ii) For the haploid and homozygous isolates (<5% of heterozygous sites), the fraction of heterozygous SNPs detected were considered as false positives and therefore filtered out
(iii) For heterozygous isolates, heterozygous sites were filtered according to their allele balance ratio (ABHet). The thresholds for allele balance ratios were determined thanks to the allelic frequency distribution all over the heterozygous samples at each level of ploidy (from 2n to 5n). A heterozygous site was rejected while its ratio did not fit the expected range according to the number of copy of the considered chromosome (or region, in the case of segmental duplication).
The joint calling method of GATK was run with the cleaned vcf files to create a complete genotyping matrix (gvcf format). This SNPs matrix included 1,625,809 segregating sites accounting for a total of 58,912,916 high-quality SNPs across our 1,011 isolates (Chapter 1 Supplementary table 1).

Pangenome determination

de novo genome assembly

We used Abyss software (version 1.5.2)[7], with the option '-k 64', to assemble the entire genome collection. Pre-assembly filtering step was performed with condetri v2.2 perl script with parameters: '-cutfirst=6 -rmN -sc=33'. The resulting assemblies had a median N50 of 136 kb and a median number of contigs of 3,259. The median length of the genome is 12.1 Mb and the median GC content is 38.06 (Chapter 1 Supplementary table 17).

Detection of non-reference material

We set up a custom pipeline to identify non-reference genome material. Each genome was aligned to the reference sequence (S288C, version R64-1-1_20110203) using blastn with the following settings: '-gapopen 5 -gapextend 5 -penalty -5 -reward 1 -evalue 10 -word_size 11 -no_greedy'.
The CDH and CFH strains were excluded from the identification of non-reference genome material due to the presence of *Staphylococcus epidermis* contamination. The sequences aligned with an identity greater than 95% were divided in three categories to be further processed. If the aligned sequence belonged to contigs shorter than 100 bp or if the aligned sequence was up to 200 bp and belonged to a contig whose length was shorter than the length of the alignment plus 75 bp, the contig was discarded. If the aligned sequence was in the range 200 – 1,500 bp only the aligned sequence was discarded. If the aligned sequence was longer than 1,500 bp, it was divided into segments of 250 bp. Each sub sequence was aligned again to the reference and discarded if found with an identity over 95% on an alignment length of at least 187 bp (75% of the subsequence length). After this step the relative position of the retained sequence has been evaluated. If two or more of them belonged to the same contig and were separated each other by less than 100 bp, the sequence from the starting of the first one to the end of the final one was kept as a whole. In the subsequent step, all the kept sequences from the 1,011 genomes were sorted for length in decreasing order. The set of sequence was then aligned against itself (with the same criteria of the first step) to eliminate repeated elements. When two subsequences were found to have an identity over 95%, the one belonging to the shorter sequence was eliminated. The process led to 12,325 sequences for 9.3 total Mb.

Annotation of non-reference material

To annotate ORFs in dispensable regions, we set up an integrative yeast gene annotation pipeline by combining different existing annotation approaches, which give rise to an evidence-leveraged final annotation. We ran the three individual components: RATT, YGAP, and MAKER, for gene annotation independently and subsequently integrated their results by EVM. Proteomes of the *sensu stricto* species *S. cerevisiae, S. paradoxus, S. mikatae, S. bayanus, S. kudriavzevii* and *S. eubayanus* were retrieved and used in our annotation pipeline to provide protein alignment support for annotated gene models.

Pangenome definition

We compiled the pangenome by adding the 2,245 non-reference ORFs annotated here to the 6,713 genomic reference ORFs listed in the set "orf_genomic_all" from the SGD database (updated 2015-01-13, http://downloads.yeastgenome.org/sequence/S288C_reference/orf_dna/orf_genomic_all.fasta.gz). The 3 RDN genes were also added (RDN18-1, RDN25-1, RDN37-1) from the set "rna_genomic" available from the SGD database (http://downloads.yeastgenome.org/sequence/S288C_reference/rna/rna_genomic.fasta.gz).
We applied to this set of ORFs a graph-based pipeline to remove duplicate and closely related sequences. This step also removed overlapping ORFs present in the "orf_genomic_all" SGD dataset. A disconnected graph was created in which each node is an ORF and each edge means an identity over 95% upon at least 75% of the sequence of the smaller ORF in the couple. Each connected subgraph represents a single ORF family. For each of these families a representative has been chosen. The connectivity has been computed for each node. The first choice for the representative was the most central, non-dubious reference ORF, if any of them were present. The second choice was the most central reference ORF, if only non-reference ORFs were present in the family, the most central of these was taken. This led to a catalogue of 7,797 non-redundant ORFs, which represent the *S. cerevisiae* pangenome (Chapter 1 Supplementary table 5). Among the reference ORFs, the number of similar ORFs collapsed for each final ORFs have a wide range (up to 67, which is the cluster of Ty1 elements), although usually they not exceed 2. Other large clusters are the Y' elements one (59 ORFs) and the Ty2 (27 ORFs). Out of the 6,713 reference ORFs, 5,681 were not redundant while 1,032 were collapsed in 402 unique ORFs, the 89% of these (N=357) are duplicated ORFs.

Pangenome copy number variants

To assess the copy number of each ORF of the pangenome, we mapped the reads from each strain to the pangenomic ORFs with bwa using default parameter and option –U 0. The result was then filtered using samtools view with options –bSq 20 –F 260. The median coverage for

each ORFs was taken as coverage for the ORF in the specific isolate. The ratio between the values of individual ORFs and the values of genome coverage on the reference of the isolate (as the median of the median coverage for each nuclear chromosome) was considered as copy number for haploid genome.
The mapping was also used as confirmation step for the presence of the ORFs in each strain, leading to the identification of 4940 ORFs present in the 1011 strains of the collection, representing the core genome and 2857 ORFs, which are present in different subset of the population. Fifty-one ORFs were removed since they were present in single strains with low coverage (~10% of the genome wide coverage of strain) and were likely contaminations from *Escherichia coli* and *Clavispora lusitaniae*. Eighty-nine other ORFs not having a sufficient coverage were kept in the pangenome but they were not used for subsequent analyses due to poor mapping. Three of the core ORFs are present but not annotated in the S288C reference and were annotated by our annotation pipeline as 584-snap_masked-1700-AIE_1, 610-snap_masked-2999-BGP_1 and 611-snap_masked-3001-BGP_1.
To evaluate the difference between domesticated clades and wild clades, we normalized the data by calculating the clade CN median for each ORF to avoid sample bias. The distributions of medians in the domesticated and wild clades were then compared using the Wilcoxon test (R function wilcox.test) (Chapter 1 Supplementary fig. 8, Supplementary table 19-20).

Inference of pangenome origin

We constructed a local CDSs database for 57 representative species that deeply probe both closely related *Saccharomyces sensu stricto* species as well as a highly diverged yeast species (Chapter 1 Supplementary table 18). In addition, we added the CDSs of 12 representative *S. cerevisiae* and *S. paradoxus* strains with complete genome sequenced by long reads PacBio. For each annotated dispensable ORF, we first performed BLASTN search (-evalue 1E-6) against this local CDSs database to find its best hit. ORFs without hits in our local yeast CDSs database were further BLAST (-evalue 1E-6) against NCBI non-redundant database. Based on the sequence identity and query coverage of those top hits, we classified those dispensable genes into different categories.

dN/dS

For all isolates, sequences of the protein-coding genes were inferred from the filtered SNPs into the reference sequence with GATK FastaAlternateReferenceMaker. For each gene, the coding sequences were aligned and the ratio of nonsynonymous to synonymous polymorphisms (dN/dS) was computed with yn00 program of the PAML software. Median values were used for comparison

Whole species nucleotidic diversity characterization

Genomic and genetic distances

The 1,544,489 biallelic segregating sites were used to construct a neighbor-joining tree (Chapter 1, Fig. 1), using the R packages ape and SNPrelate[8]. The gvcf matrix was first converted into a gds file and individual dissimilarities were estimated for each pair of individuals with the *snpgdsDiss* function. The bionj algorithm was then run on the obtained distance matrix.
The genomic content distance has been calculated as number of ORF differences in the pangenome presence/absence profile (i.e. the number of ORFs present in only one strain for each pairwise strain comparison).

Genetic diversity

As an estimate of the scaled mutation rate, we computed π, the average pairwise nucleotide diversity θ_w, the proportion of segregating sites and Tajima's D, which stands for the difference between π and θ_w. Variscan 2.0[9] was run (runmode=12, 10-kb non-overlapping windows) on multiple alignments of the concatenated chromosomes representative of the isolates.

Ploidy, aneuploidies and segmental duplications

The natural ploidy of the 794 natural isolates (see *S. cerevisiae sequenced isolates* section), as well as their aneuploidy and segmental duplication content were investigated by combining 3 complementary approaches:
(1) Measurement of the cell DNA content using high-throughput flow cytometry: DNA content was analysed using a propidium iodide (PI) staining assay. Cells were first pulled out from glycerol stocks in liquid YPD in 96 well plates (30°C, overnight). 5 µL of the culture were transferred into 195 µL of fresh YPD and incubated 8 hours at 30°C. Then, 3 µL were taken and resuspended in 100 µL of cold 70% ethanol. Cells were fixed overnight at 4°C, washed twice with PBS, resuspended in 100 µL of staining solution (15 µM PI [Sigma-Aldrich], 100 µg/mL RNase A [Sigma-Aldrich], 0,1% v/v Triton-X, in PBS) and finally incubated 3 hours at 37°C in the dark. 10,000 cells for each sample were analysed on a FACS-Calibur flow cytometer using the HTS module for processing 96 well plates. Cells were excited at 488 nM and fluorescence was collected with FL2-A filter. The distributions of both FL2-A and FSC-H values have been processed to find the two main density peaks, which correspond to the two cell populations in G1 and G2. The peaks were detected using the

densityClust R package after removing the cells reaching the FACS saturation (either FLS-A or FSC-H values equal to 1000). We categorized the values of FLS-A, which correlate with the DNA quantity, to estimate the ploidy according to the following scheme: strains with G1 cells values between 39 and 181 and G2 values between 148 and 255 were labelled as haploid; strains with G1 cells values between 145 and 265 and G2 values between 295 and 500 were labelled as diploid; strains with G1 cells values between 245 and 355 and G2 values between 500 and 700 were labelled as triploid; strains with G1 cells values between 295 and 500 and G2 values between 700 and 905 were labelled as tetraploid; strains with G1 cells values between 395 and 605 and G2 values over 905 were labelled as over then 4N; strains with other combinations of values have been manually evaluated.
(2) Study of sequencing coverage: systematic analysis of the coverage depth along the genome was performed with 1 kb non-overlapping sliding windows, which allowed for the survey of chromosomal copy number variations as well as segmental duplications. The ratio between the coverage of the aneuploid chromosomes and the rest of the genome was also used to validate the ploidy of isolates.
(3) Investigation of the allele balance ratio associated to heterozygous SNPs, as heterozygous sites, should fit an expected range of ratio according to the number of copy of the considered chromosome (see *SNPs filtering and matrix* section). The precise locations of segmental duplications were manually investigated in the vcf files (Chapter 1Supplementary table 16).

SNP Annotation

In chapter 1, SnpEff (v4.1)[10] was used to annotate and predict the effect of the variants. Non-synonymous SNPs predicted as deleterious by SIFT (v5.2.2)[11] as well as nonsense mutations were considered as deleterious for protein function. Insertions and deletions were considered to cause frame-shifts when their sum is a single gene produced a number not divisible by 3. A sequence representative of each isolate was constructed by inferring these filtered SNPs into the reference sequence with GATK FastaAlternateReferenceMaker.

Model-based ancestry

Model-based ancestry estimation was performed on the biallelic SNPs using ADMIXTURE v.1.23 in unsupervised mode[12].

PCA

Principal components analysis on the biallelic SNPs was performed using EIGENSOFT v6.0.1. The “-w” argument was used to calculate the principal components using only a subset

of the samples, with the remaining samples then being projected onto the resulting components.

Discriminant analysis of principal components (DAPC)

The matrix of presence/absence of ORFs in the population has been analyzed using the DAPC algorithm implemented in the R package adegenet 2.0.1[13]. DAPC describes clusters maximizing the between-cluster variance while minimizing the within-cluster variance. The number of components retained for the PCA calculation was 150, accounting for >88% of total variance. For the subsequent DAPC calculation, the alpha-score indicates 25 as the optimal number of DPC to be retained. Clustering was performed using the K-means with different number of groups (N = 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50).

Linkage disequilibrium

The PLINK[14] package was used to compute r^2, the correlation coefficient between pairs of loci which stands as a measure of association for LD. All pairs of polymorphic sites were investigated through map and ped files generated with vcftools[15], excluding SNPs with a minor allele frequency (MAF) lower than 5%. We averaged r^2 based on the SNP distance (100 bp intervals) over 25 kb regions and calculated the half-length of r^2 which is the distance at which LD decays to half of its maximum value.

F_{ST} calculation

The Weir and Cockerham weighted F_{ST} index[16] was computed along the genome in 2-kb non-overlapping sliding windows with vcftools using all biallelic SNPs for each clade *vs* its complementary part of the population. Genes located in the 0.25% right tail of the F_{ST} empirical distribution were considered as showing significantly high values and investigated for functional enrichment with GO::TermFinder. P-values < 0.05 were considered as significant.

Loss of heterozygosity

Heterozygous isolates were investigated for loss of heterozygosity (LOH) regions with an in-house R script. Regions over 50 kb with less than 10 heterozygous sites per 50 kb were considered to be under LOH (Chapter 1 Supplementary table 8).

Saccharomyces sensu stricto rooted tree

To construct the tree, we used 22 *S. cerevisiae* isolates representative of the species genetic diversity that were sequenced with Oxford Nanopore technology. We annotated these 22 assemblies with the pipeline described above. The annotated protein-coding genes were pooled together with the *S. cerevisiae* reference genome (SGD R64-1-1) and another 18 yeast strains for orthology identification. Those 18 other yeast strains included 7 *S. cerevisiae* strains, and 5 *S. paradoxus* strains, and 6 out-group strains from other *sensu stricto* yeast species as described in our previous work. The orthology identification was carried out by Proteinortho (v5.15)[17] with synteny information considered (the PoFF feature of Proteinortho). This leads to the delineation of 2,018 1-to-1 orthologous groups across all the 41 sampled genomes. For each orthologous group, the protein sequences across the 41 strains were aligned with MUSCLE (v3.8.1551)[18], the resulting protein alignment was further used to guide the corresponding CDS alignment using PAL2NAL (v14)[19]. A concatenated multi-gene matrix was built for the CDS alignment of these 2,018 orthologous groups, which was further partitioned based on codon positions (e.g. 1st, 2nd, and 3rd codon positions). We used RAxML (v8.2.6)[20] to build the maximum likelihood (ML) tree based on the GTRGAMMA model with 100 rapid bootstraps. Alternatively, we also performed phylogenetic analysis using the consensus tree approach, in which we built individual gene trees for each of the 2,018 orthologous groups using the same method described for the concatenated tree. These individual gene trees were further summarized by ASTRAL (v4.7.12)[21] to infer the "species tree". Both the concatenated tree and the consensus tree were visualized in FigTree (v1.4.2) (http://tree.bio.ed.ac.uk/software/figtree/).

Genotype-Phenotype Relationship

Phenotyping

Quantitative high-throughput phenotyping was performed using end point colony growth on solid media. Strains were pregrown in flat bottom 96-well microplates containing liquid YPD medium. The replicating ROTOR HDA© benchtop robot (Singer instruments) was used to mix and pin strains onto a solid YPD matrix plate at a density of 384 wells. The matrix plates were incubated overnight at 30°C to allow sufficient growth and replicated on 36 media conditions including YPD 30°C as pinning and growth control (Chapter 1 Supplementary table 2). Each isolate was present in quadruplicates on the corresponding matrix (interplate replicates) and at two different positions (intraplate replicates). The plates were incubated at 30°C for 40 hours and were scanned at a resolution of 600 dpi and 16-bit grayscale. Quantification of the colony size from plate images was performed using the software package Gitter. Each value was normalized using growth ratio between stress media and standard YPD medium 30°C. Pairwise Pearson's correlations of fitness between replicates were calculated for each condition.

Phenotype simulation

In chapter 2, for each dataset, we simulated 1000 Mendelian traits (governed by only 1 causal SNP) and 1000 complex traits (governed by 10 SNPs). Causal SNPs were randomly chosen in the SNP matrix and a phenotype was generated accordingly using GCTA[22]. The heritability of all the simulated traits was chosen to be of 0.8 for each dataset. The command executed for the phenotype simulations was the following:

```
gcta --bfile $snp --maf $maf --simu-qt --simu-causal-loci $snplist --simu-hsq 0.8 --simu-rep 1 --out $output
```

With the simulation of a complex trait, we used GCTA's default effect size assignation method, which consists on generating them from a standard normal distribution among the 10 causal SNPs.

Phenotyping strategy

In chapter 3, High-Throughput Phenotyping and Growth Quantification Quantitative phenotyping was performed using endpoint colony growth on solid media. Strains were pregrown in liquid YPD medium and pinned onto a solid SC matrix plate to a 1536 density format using a replicating robot ROTOR™ (Singer Instruments). Two replicates of each parental strain were present on every plate and six replicates were present for each hybrid.

The matrix plates were incubated overnight to allow sufficient growth, which were then replicated in 53 media conditions, plus SC as a pinning control. The plates were incubated for 24 hr at 30 °C (except for 14 °C phenotyping) and were scanned at the 14-, 24-, and 38-hr time points with a resolution of 600 dpi at 16-bit grayscale. Quantification of the colony size was performed using the R package Gitter[23], and the fitness of each strain on the corresponding condition was measured by calculating the normalized growth ratio between stress media and SC. 24 hr time point fitness values have been retained for all subsequent analysis.
As each hybrid is present in six replicates, the value considered for its phenotype is the median of all its replicates, thus smoothing the effects of pinning defect or contamination.

Genome-wide association studies

Mixed-model association analysis was performed using FaST-LMM version 2.07. We used the normalized phenotypes by replacing the observed value by the corresponding quantile from a standard normal distribution, as FaST-LMM expects normally distributed phenotypes. In this step, we used the markers showing a MAF >5%. We also filtered missing genotypes with an arbitrary threshold set to exclude all variants present in less than 1,000 individuals for the total matrix, and we excluded all sites with missing calls for the subset matrices.The command used for association was the following:
'fastlmmc -bfile $snp -bfileSim $snp -pheno $pheno -out $assoc_file –verboseOutput'
The mixed model adds a polygenic term to the standard linear regression designed to circumvent the effects of relatedness and population stratification[24]. To quantify the extent of the bulk inflation and the excess false positive rate, we computed the genomic inflation factor λ for each condition (Chapter 1 Supplementary fig. 27). This factor is defined as the ratio between the median of the empirically observed distribution of the test statistic on the expected median. For example, the λ for a standard allelic test for association is based on the median (0.456) of the 1-d.f. $\chi 2$ distribution. Under a null model of no association and unlinked variants, the expectation is for the λ to be 1. A λ superior to 1 indicates inflated p-values of association, possibly due to a confounding factor not accounted for.

GWAS significance threshold

We estimated a trait-specific P-value threshold for each condition by permuting phenotypic values between individuals 100 times. The significance threshold was the 5% quantile (the 5th lowest p-value from the permutations). With that method, variants passing this threshold have a 5% family-wise error rate (Chapter 1 Supplementary fig. 28).

Genome-wide heritability computation

The estimations of genome-wide heritabilities were completed by dividing the genetic variance of the null model by the total variance on the null model (genetic variance and residual variance) computed by FaST-LMM (Chapter 1 Supplementary fig. 6). The values reported are those based on the quantile-normalized phenotypes. To compute the variance explained by our significantly associated markers, we included them in the covariance matrix with the '-bfileSim' option and did the same calculation again.

Simulations evaluation

In chapter 2, in order to evaluate and compare the power of our datasets to recover causal SNP(s), we built a table of confusion for each run of simulation, by measuring the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), transposed into rates with the following formulas:

- True positive rate (TPR) = TP / (TP + FN)
- True negative rate (TNR) = TN / (TN + FP)
- False positive rate (FPR) = FP / (FP + TN)
- False negative rate (FNR) = FN / (FN + TP)

Diallel Design

Stable haploid generation

In chapter 3, to easily obtain a large amount of crosses between a maximum of strains and avoid manual crossing of each hybrid, we inserted in each strains a resistance marker specific for each mating type. After mating, double selection will only allow hybrids carrying the two resistance markers to grow.
Due to *S cerevisiae* particular life and reproductive cycle, haploid strains needed to be stabilized as they easily switch their mating type because of *HO* endonuclease's ability to cut *MAT* cassettes that is then repaired with the opposite *MAT* cassette thus changing cell's mating type. This mating type switching allows for homothallism in *S. cerevisiae*.
To impede mating type switching, deletion of *HO* was performed in each strain. Δ*HO* strains were generated by homologous recombination with insertion of the resistance cassettes *KanMX* and *ClonNAT* in *MAT*a and *MAT*□, respectively.
71 diploid strains annotated as homozygous after sequencing and spanning all the genetic diversity of *S. cerevisiae* have been selected and *HO* successfully replaced by corresponding resistance cassette in both mating type.

Hybrids generation in diallel cross

Strains were pregrown in liquid YPD medium. Mating was performed with ROTOR™ (Singer Instruments) by pinning and mixing *MAT*a over *MAT*□ parental strains on YPD agar medium. 71 *MAT*a *HO::*Δ*KanMX* and corresponding 71 *MAT*□ *HO::*Δ*ClonNAT* strains were mated in a pairwise manner on YPD for 24 hours at 30°C. Plated were then replicated on YPD supplemented with G418 (200 µg/ml) and Nourseothricin (100 µg/ml) for double selection of hybrids. After 48 hours, plates were replicated again on the same media to eliminate potential residuals of non-hybrids cells.

Verification of strain homozygosity

In chapter 3, the 71 selected strains were annotated as homozygous after sequencing. However, phenotyping of the parental strains was performed to confirm the homozygosity between both mating type parents. A linear regression model was fitted against fitness from *MAT*a and *MAT*α parents of the same line. Only 55 lines displayed a linear regression coefficient (Pearson's r) higher than 0.8 (Figure1). They were considered homozygous and thus retained for the analysis.

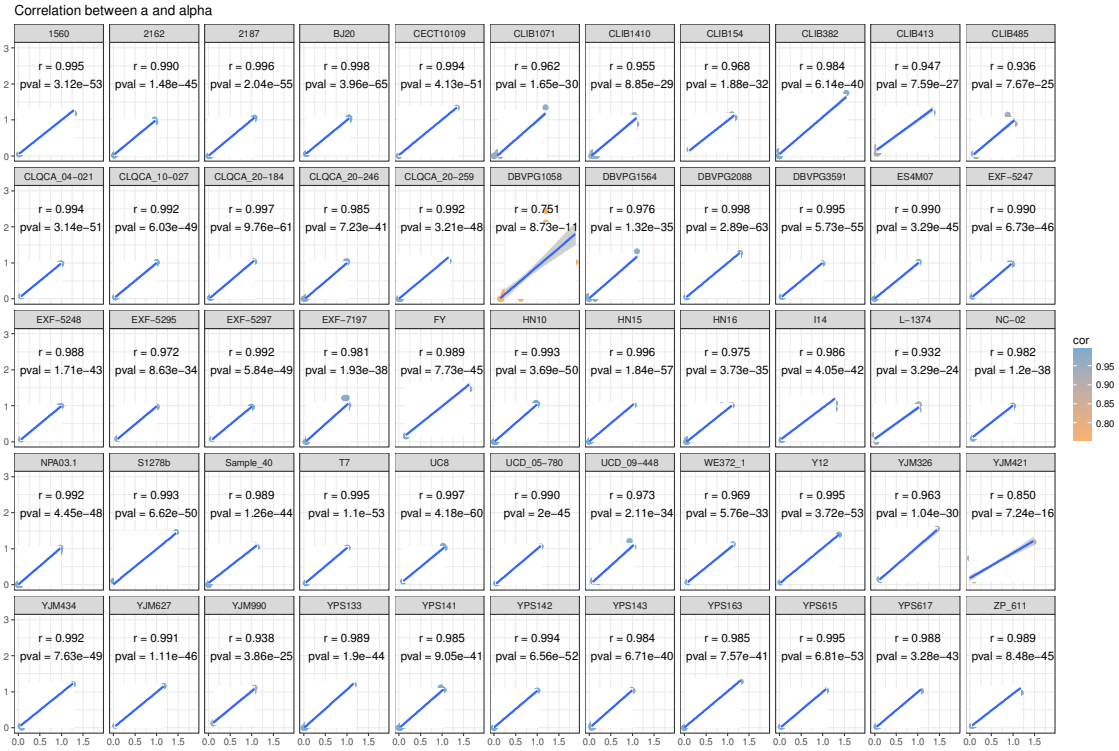


Figure 1: Phenotypic correlation between *MATa* and *MATα* parents

Generation of F1 Genotypes and SNP filtering

In silico F1 genotypes were constructed from 34 parental genomes sequences[25] that have been filtered to only biallelic polymorphic sites, resulting on a matrix containing 295,346 polymorphic sites encoded using the "recode12" function of PLINK[14]. Those genotypes correspond to the half-matrix of pairwise crosses, including the diagonal, i.e. the 34 homozygous parental genotypes (Figure 2). For each cross, we combined the genotypes from both parents to generate the diploid progeny. As a result, heterozygous sites correspond to sites for which the two parents had different allelic versions.

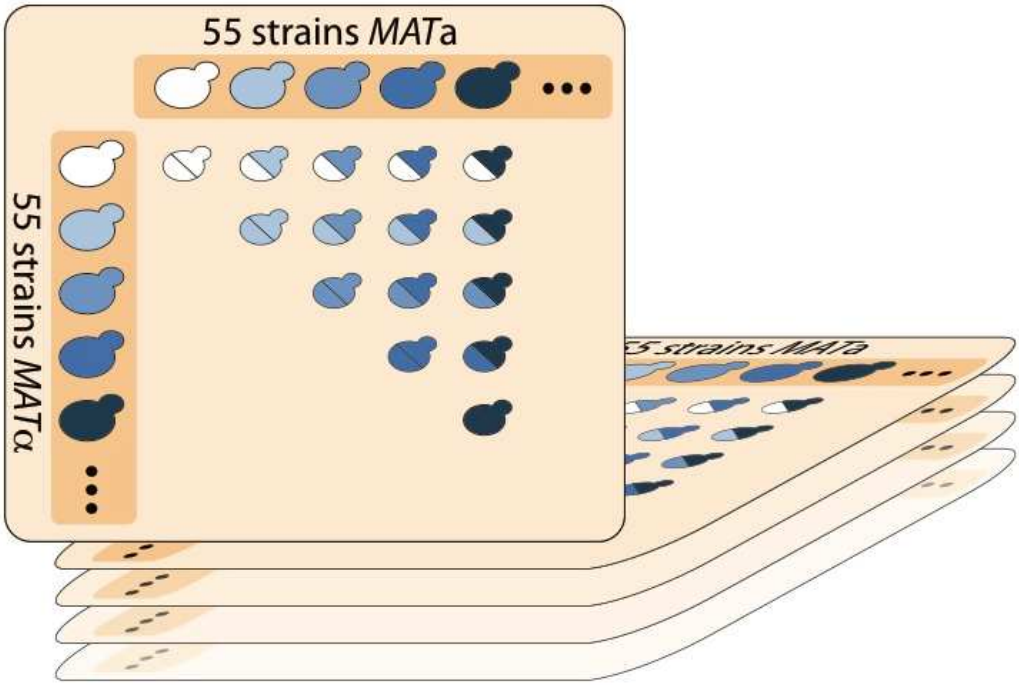


Figure 2: Overview of the diallel mating scheme

Long-range linkage disequilibrium filtering

Due to the small number of parental genotypes, there is an extensive long-distance LD between sites (Figure 3a). In order to evaluate this, we created categories for how often a SNP pattern across all individuals was observed within our dataset (pattern occurrence). These categories ranged from 1 to 7,230 meaning that if a SNP displays a pattern across the population that is present 1,000 times, 999 other SNPs will have the exact same pattern. In order to reduce the noise due to the impossibility to distinguish SNPs sharing the same pattern across our population when performing GWAS, we removed all sites whose pattern was shared more than twice across the population, resulting in a final dataset containing 31,632 polymorphic sites with a MAF superior than 5% (Figure 3b).

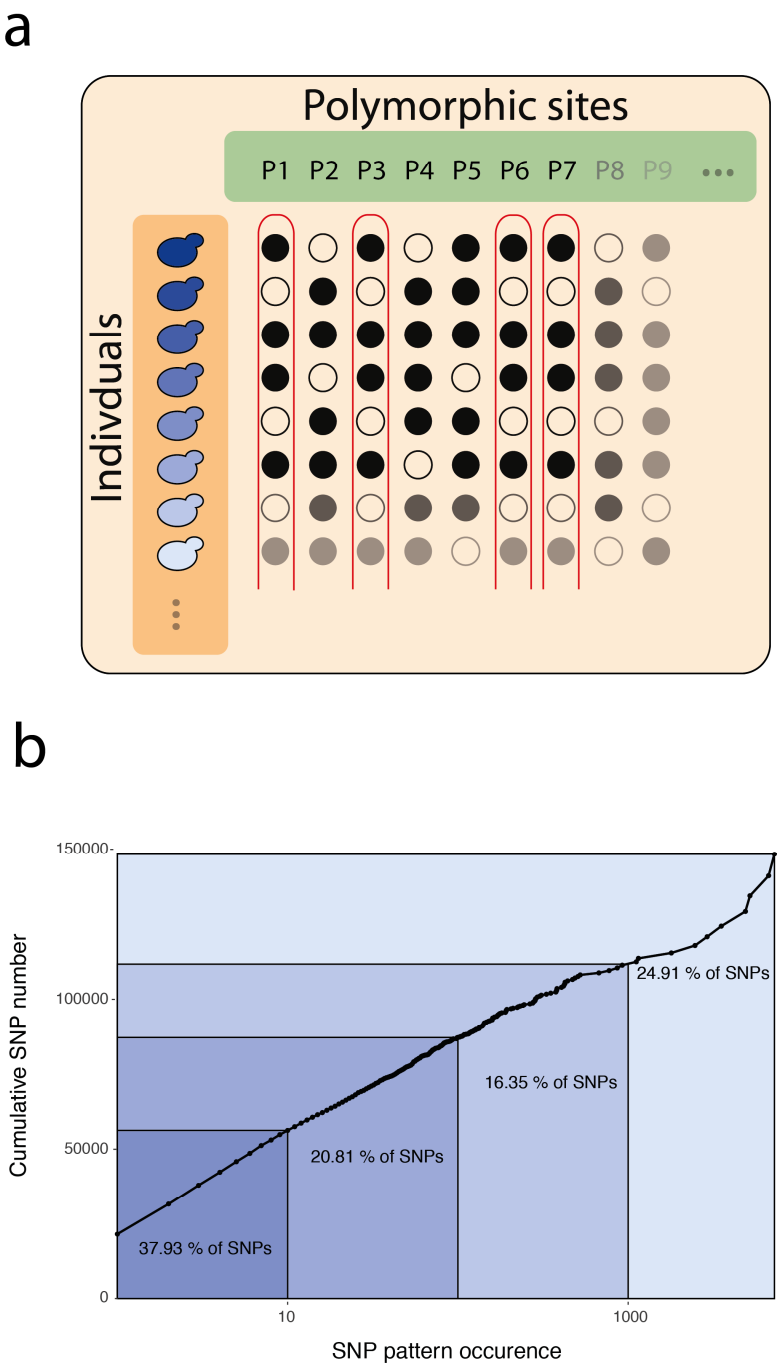


Figure 3: Long range linkage disequilibrium filtering. **a.** Schematic view of patterns of polymorphic sites across the population of 595 individuals. The allelic version of the SNP is indicated by its color (black or white). The polymorphic sites in red boxes share the same pattern. **b.** Cumulative distribution of SNP pattern occurrence. SNP pattern occurrence was calculated as the number of positions that share a genotypic pattern.

Matrix encoding

We performed GWA analyses with different encodings. In the additive model, the genotypes of the F1 progeny were simply the concatenation of the genotypes from the parents. As

homozygous parental alleles were encoded as 1 or 2, the possible alleles for each site in the F1 genotype were “11” and “22” for homozygous sites and “12” for heterozygous sites. We also used an overdominant genotype encoding, where both the homozygous minor and homozygous major alleles are encoded as “11” and heterozygous genotype is encoded as “22”.

References

1. Skelly, D. a. *et al.* Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23,** 1496–504 (2013).
2. Bergström, A. *et al.* A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31,** 872–88 (2014).
3. Strope, P. K. *et al.* The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res.* **125,** 762–74 (2015).
4. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1,** 18 (2012).
5. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).
6. Auwera, G. A. Van Der *et al.* From FastQ data to high confidence varant calls: the genonme analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* **11,** (2014).
7. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–23 (2009).
8. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28,** 3326–28 (2012).
9. Vilella, A. J., Blanco-Garcia, A., Hutter, S. & Rozas, J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics* **21,** 2791–3 (2005).
10. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6,** 80–92 (2012).
11. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
12. Alexander, D. H. & Novembre, J. Fast model-based estimation of ancestry in unrelated individuals. 1655–1664 (2009).
13. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24,** 1403–5 (2008).
14. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**,559–75 (2007).
15. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–8 (2011).
16. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38,** 1358–70 (1984).
17. Lechner, M. *et al.* Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12,** 124 (2011).
18. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).
19. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34,** W609–12 (2006).
20. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–3 (2014).
21. Mirarab, S. *et al.* ASTRAL: genome-scale coalescent-based species tree estimation.

Bioinformatics **30,** 541–8 (2014).
22. Yang, J., Lee, H., Goddard, M. & Visscher, P. GCTA : a tool for genome-wide complex trait analysis. 1–19 (2011).
23. Wagih, O. & Parts, L. gitter: a robust and accurate method for quantification of colony sizes from plate images. *G3 (Bethesda).* **4,** 547–552 (2014).
24. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38,** 203–8 (2006).
25. Peter, J. *et al.* Yeast Evolutionary history and natural variation revealed by 1,011 genomes. *In revision* (2017).

Conclusion & perspectives

Species-wide exploration of the genetic diversity within the *S. cerevisiae* model organism

The emergence and the accumulation of genetic differences within a species constitute the basis for phenotypic variation upon which natural selection will act. By using genomic sequences of large number of *Saccharomyces cerevisiae* isolates, we sought to describe in the most exhaustive way the extent of intraspecific genetic variation and investigate its phenotypic consequences.

To have an overview of the genetic diversity within our model species, we analyzed genomic sequences of 1,011 individuals that have been isolated in a wide range of ecological niches all around the world. This allowed us to have a precise view of the evolutionary history of *S. cerevisiae* from its origin in China to its several domestication events resulting from its long association with human for the production of fermented beverages or bread. Indeed, several populations have been identified with each one having its unique evolutionary history. We explored genetic variants such as single nucleotide polymorphisms, copy number variation, ploidy and aneuploidy levels and observed that the genome evolution of wild isolates is mostly driven by the accumulation of SNPs, while domesticated isolates gather most of the genome content, ploidy and aneuploidy variation. In accordance to what has been observed in other species such as *A. thaliana* or human, there is a strong bias towards low frequency SNPs[1,2]. We also described extensive loss-of-heterozygosity, which seems to be an important factor, allowing inter-individual variation in this mostly asexual species.

High-throughput phenotypic characterization of the sequenced strains has provided the opportunity to perform Genome-Wide Association Studies (GWAS) with an unprecedented power using this model species, demonstrating the feasibility of this approach, despite the limitations brought by the high levels of population stratification. Genetic variants detected by GWAS were mostly CNVs, which have a larger phenotypic impact than SNPs, suggesting the important role of these variants in reducing the missing heritability.

Genotype-phenotype relationship evolution in yeast

The genomic and dataset obtained constitutes a precious resource to explore the genotype-phenotype relationship. Indeed, previous attempts at performing GWA in *S. cerevisiae* suffered from a lack of power due to the limited sample size of the datasets they used[3–6]. In order to assess the feasibility of GWAS and estimate the limits of this approach using this species, we developed a phenotype simulation framework and evaluated the capacity to identify the causal variants that were used to simulate the traits for five subsets as well as the total dataset containing 1,011 genomes. Results showed that the difficulties associated with

confounding factors such as population structure or relatedness can be alleviated if the size of the population used is sufficient, and if the individuals do not have a too recent common ancestor.

Finally, to transcend GWAS limitations and further reduce the part of missing heritability, we used a diallel mating scheme and tested genotype-phenotype association. More precisely, we generated a half-diallel panel by generating all possible pairwise crosses between a population of 34 parental strains that are spread all over the genetic diversity of *S. cerevisiae*. The resulting population therefore contains 561 hybrids and 34 homozygous diploids corresponding to the parental strains. Results demonstrated the power of such an approach with the association of 2,293 significantly associated SNPs with 53 traits. We used several encodings also be able to detect variants acting non-additively. We were also able to map a variant located in the coding sequence of the *GAL2* gene that is a rare variant in the total population of 1,011 isolates. Due to the haplotype shuffling inherent of the diallel design, its minor allele frequency shifted and could be kept to perform GWAS. This variant alone explains 10.5% of phenotypic variation, suggesting that taking into account rare variants may be an efficient way to reduce the missing heritability.

What's next?

With the rapid advances in population genomics due to the advent of sequencing technologies, methods like GWAS and linkage mapping have been developed very quickly. Yet, the elucidation of the genotype-phenotype relationship is far being achieved. For example, the prediction of phenotypes based on the genotype is a goal that seemed reachable, but the numerous GWAS performed up to date demonstrated that associated variants only explain a fraction of the phenotypic variation[7]. The several possible hypotheses for the missing heritability have to be investigated and techniques have to be refined in order to improve our knowledge on the genotype-phenotype relationship.

One possibility that might reduce a part of the missing heritability is to perform GWA using all kind of genetic variants, and in particular structural variants. This is now conceivable with the advent of long-read sequencing such as Oxford Nanopore MinION, that offers the possibility of sequencing long DNA fragments in a very efficient way. A recent study already reported the genome sequences of 22 *S. cerevisiae* isolates and showed that genome assemblies generated by such a technology allow to detect structural variants such as inversions, transposable elements or translocations were successfully mapped, whereas they are generally missed using short-read sequencing strategies[8]. As the interest towards long-read sequencing grows, it is likely that whole genome data generated by these technologies will grow at a fast pace, allowing the precise characterization of structural variants at a species-wide scale. Integrating all variants identified to perform GWAS is likely to bring new insights into their role in phenotypic variation.

Another way to investigate further the genotype-phenotype relationship within *S. cerevisiae* is to increase the number of traits measured for each strain, enabling a higher scale exploration of the phenotypic landscape. One way to achieve this goal could be to obtain gene expression measures for all genes of the 1,011 *S. cerevisiae* isolates under a standard condition or even several conditions using a RNAseq strategy. With such an approach, we will obtain around 6,000 phenotypes for each condition (*i.e.* the expression level of each gene), which would constitute a goldmine for the exploration of the phenotypic diversity within *S. cerevisiae*. Several interesting aspects of phenotypic variation can be studied on a large panel of individuals. For example, it could be interesting to see if we can observe differences in gene expression between domesticated and wild isolates. Indeed, with human-related subpopulations being more prone to Copy Number Variation (CNV) of phenotypically relevant genes, we could characterize the relationship between gene dosage and gene products as well as identify genes whose change in expression allowed adaptation of individuals to colonize new environments.

Genetic variants can influence complex traits in many ways, including the modulation of gene expression. The conjugation of whole genome sequencing data with gene expression profiling by RNA-Seq could be an elegant and high-throughput way to obtain a global picture of how individuals react to their environment, and what are the genetic determinants responsible for differential gene expression and regulation (Figure 1).

Though these Transcription-Wide Association Studies (TWAS) are promising, they require both genomic sequences and expression data, which is limited by the cost of such experiments in human[9]. The budding yeast *Saccharomyces cerevisiae* could again constitute a model organism presenting the advantages of a reduced number of genes in the genome and the possibility to perform high-throughput phenotyping on clonal colonies, allowing the multiplication of phenotypic experiments (*i.e.* different conditions) on genetically similar individuals.

Overall, the wealth of data to be produced by this experiment will undoubtedly offer a clearer vision of the phenotypic outcomes of genetic variants within yeast natural populations, and will certainly be a precious resource to elucidate the genetic architecture of complex traits in general.

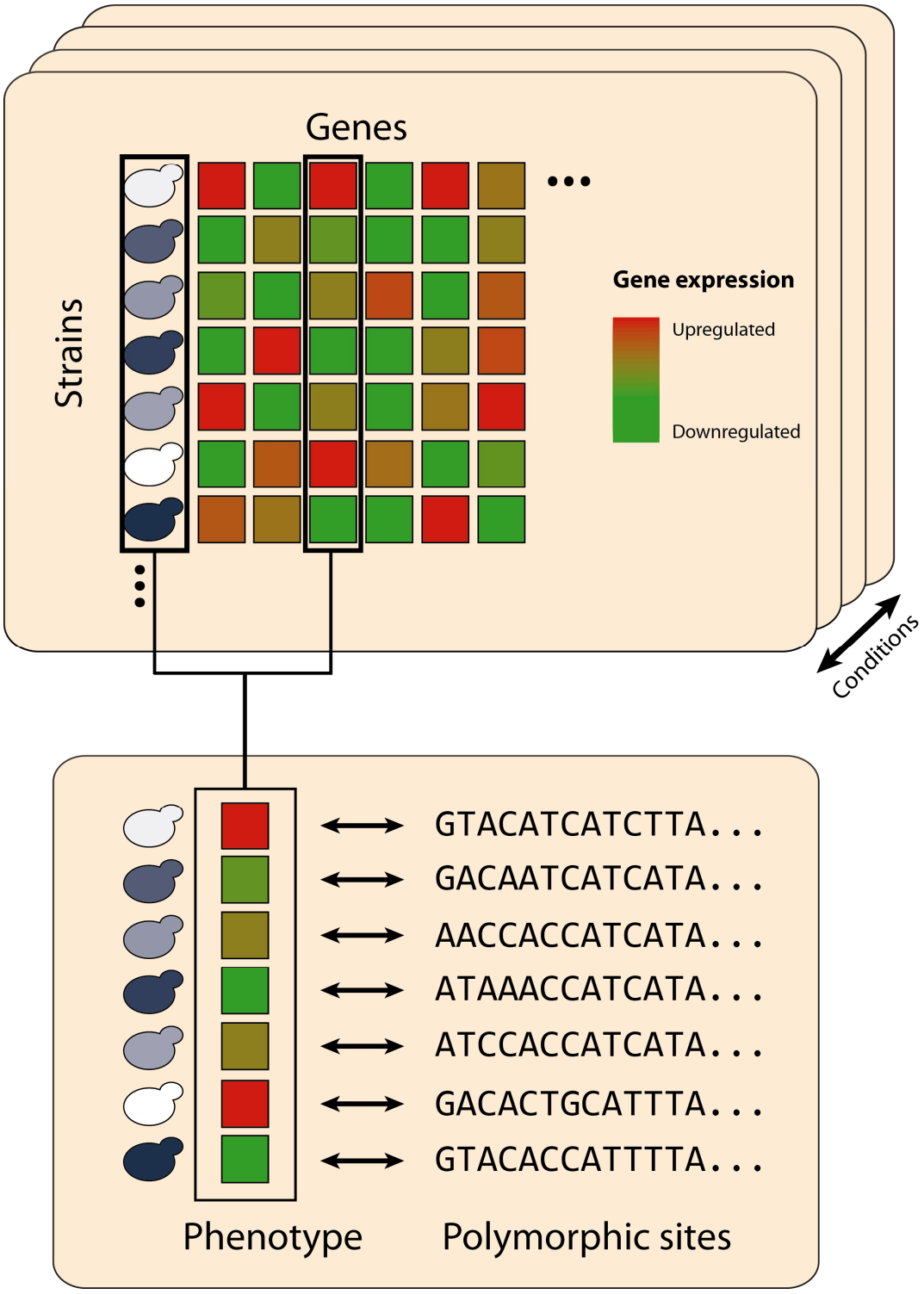


Figure 1: Overview of Transcriptome-Wide Association Studies.

References

1. 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–73 (2010).
2. Cao, J. *et al.* Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43,** 956–63 (2011).
3. Strope, P. K. *et al.* 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, 762–74 (2015).
4. Diao, L. & Chen, K. C. Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. *Genetics* **192,** 1503–1511 (2012).
5. Connelly, C. F. & Akey, J. M. On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* **191,** 1345–53 (2012).
6. Muller, L. A. H., Lucas, J. E., Georgianna, D. R. & McCusker, J. H. Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol. Ecol.* **20,** 4085–97 (2011).
7. Manolio, T. a *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–53 (2009).
8. Istace, B. *et al. de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience,* **6**, 1–13 (2017)
9. Gusev, A. *et al.* Whole population, genome-wide mapping of hidden relatedness. *Genome Res.* **19,** 318–26 (2009).

Appendices

List of publications

Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A, Freel K, Llored A, Cruaud C, Labadie, K, Aury J-M, Istace B, Lebrigand K, Barbry P, Engelen S, Lemainque A, Wincker P, Liti G and Schacherer J. (2017). Yeast evolutionary history and natural variation revealed by 1,011 genomes. **In review**

Brion C, Legrand S, **Peter J**, Caradec C, Pflieger D, Hou J, Friedrich A, Llorente B and Schacherer J. (2016). Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. ***PloS Genetics.*** 13(8):e1006917.

Hou J, Sigwalt A, Fournier T, Pflieger D, **Peter J**, De Montigny J, Dunham M-J and Schacherer J. (2016). The hidden complexity of Mendelian traits across yeast natural populations. ***Cell Reports*** 16(4):1106-14.

Peter J and Schacherer J. (2015). Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. ***Yeast***. 3:73-81.

List of communications

EMBO Conference Series: From Functional Genomics to Systems Biology
Heidelberg, Germany (2016)

New insights into the genotype-phenotype relationship by genome-wide association analysis in yeast. (Oral)

NGS Symposium
Illkirch, France (2016)

The 1002 yeast genomes project: Exploring the genotype-phenotype relationship of *Saccharomyces cerevisiae*. (Oral)

EMBO Conference Series: Exploring the genomic complexity and diversity of eukaryotes
Sant Feliu de Guíxols, Spain (2016)

The 1002 yeast genomes project: A framework for genome-wide association studies. (Oral)

Séminaire de Microbiologie de Strasbourg
Strasbourg, France (2015)

The 1002 yeast genomes project: A framework for genome-wide association studies. (Oral)

European Conference on computational biology
Strasbourg, France (2014)

The 1002 yeast genomes project: A framework for genome-wide association studies. (Poster)

EMBO Practical Course: Genotype mapping of complex traits
Hinxton, UK (2014)

The 1002 yeast genomes project: A framework for genome-wide association studies. (Poster)

Séminaire de Microbiologie de Strasbourg
Strasbourg, France (2014)

Meiotic recombination in *Lachancea kluyveri*. (Poster)

Levures, Modèle, Outils
Bordeaux, France (2014)

Meiotic recombination in *Lachancea kluyveri*. (Oral)

Résumé de la thèse

Dans toute espèce, la diversité phénotypique repose principalement sur la variabilité génétique observée entre individus. Un des buts fondamentaux en biologie consiste en la compréhension des forces générant et maintenant la diversité génétique au sein des individus, des populations et des espèces. De manière plus précise, la génomique des populations vise à obtenir une meilleure connaissance des relations qu'entretiennent le génotype et le phénotype au sein d'individus d'une même espèce. En effet, la capacité de détecter les variants génétiques responsables de la diversité phénotypique constituerait une étape cruciale vers l'élucidation de la relation génotype-phénotype, voire même la prédiction de phénotypes basés sur le génotype. Une telle connaissance pourrait être utilisée dans de nombreux domaines tels que la médecine ou l'industrie, par exemple. Une fois que le génome d'un organisme de référence a été séquencé, assemblé et annoté, la suite logique consiste au séquençage des génomes d'un grand nombre d'individus appartenant à la même espèce. En raison de l'émergence de technologies de séquençage, l'obtention de séquences génomiques pour de nombreux individus n'est aujourd'hui plus une étape limitante. C'est pourquoi de nombreux projets de reséquençage massifs ont été initiés, permettant de rassembler un nombre suffisant de génomes pour effectuer des études à haut-débit. Nous pouvons par exemple citer le projet 1000 génomes ou plus récemment le projet UK10K pour l'humain, ou encore le projet 1001 génomes pour *Arabidopsis thaliana*. Le principal but de ces projets est de rassembler un nombre suffisant de séquences génomiques pour pouvoir établir un catalogue des variants génétiques au sein d'une espèce. Les motivations pour initier de tels projets sont multiples et complémentaires. Premièrement, ces projets ont été conçus pour permettre une meilleure compréhension de la démographie et de l'histoire évolutive des populations. Deuxièmement, la génération de jeux de données aussi conséquents permet de comprendre les processus générant et maintenant la diversité génétique. Troisièmement, ces données rendent possible la distinction des effets agissant sur la totalité du génome, tels que la dérive génétique ou les migration, des effets agissant sur des loci individuels, tels que la sélection, les mutations ou la recombinaison. Pour finir, une des raisons majeures motivant les projets de reséquençage massifs est d'avoir une meilleure compréhension de la relation génotype-phénotype, et de manière plus précise, de pouvoir générer des jeux de données rendant possible l'identification de variants génétiques responsables de la diversité phénotypique. Une des manières d'accéder à cette identification est de conduire des études

pangénomiques d'association génotype-phénotype. Cette technique permet la d'associer les génotypes avec les phénotypes en tirant profit de la recombinaison des populations naturelles pour détecter des associations significatives entre des variants situés sur l'ensemble du génome et un phénotype d'intérêt.

Bien qu'étant un organisme modèle en génétique, les levures, et plus particulièrement *Saccharomyces cerevisiae*, est un modèle sous-représenté pour les études pangénomiques d'association génotype-phenotype. C'est pourtant un organisme idéal pour mener à bien une étude de génomique des populations à bien des égards. Pour commencer, le génome de *S. cerevisiae* est petit et compact, ce qui présente l'avantage de réduire les coûts du séquençage. De plus, cet organisme peut être isolé dans une grande gamme d'origines écologiques et géographiques, ce qui a pour effet de maximiser la diversité génétique et phénotypique. Enfin, les levures forment des colonies clonales, permettant de répliquer les mesures de phénotypes et donc garantir la reproductibilité des données. En dépit de ces avantages, le nombre de génomes disponibles chez *S. cerevisiae* est faible par rapport aux autres espèces modèles précédemment citées.

C'est dans ce cadre que s'inscrivent mes travaux de thèse qui se sont articulés autour de l'analyse de plus de 1000 génomes séquencés dans le cadre du projet 1002 génomes de levures (http://1002genomes.u-strasbg.fr/). Dans un premier temps, je me suis focalisé sur la génomique des populations au sein de l'espèce *S. cerevisiae* (figure 1a), ainsi qu'à la relation génotype-phénotype, en utilisant comme caractères observables les vitesses de croissances des colonies soumises à différentes conditions de stress. Dans un second temps, je me suis intéressé aux paramètres influençant la performance des études pangénomiques d'association génotype-phénotype (figure 1b). Pour ce faire, j'ai simulé des phénotypes à partir des génotypes pour ensuite tester la capacité de retrouver les variants impliqués et ainsi de pouvoir comparer plusieurs jeux de données entre eux et d'en apprécier les limites vis-à-vis des études d'association chez la levure. Dans un troisième temps, je me suis intéressé aux moyens de contourner les limites des études pangénomiques d'association génotype-phénotype chez *Saccharomyces cerevisiae*, et plus particulièrement de pouvoir détecter la variation phénotypique entraînée par des variants rares. Pour ce faire, j'ai réalisé des études d'association sur une population d'hybrides issue d'un croisement diallèle (figure 1c).

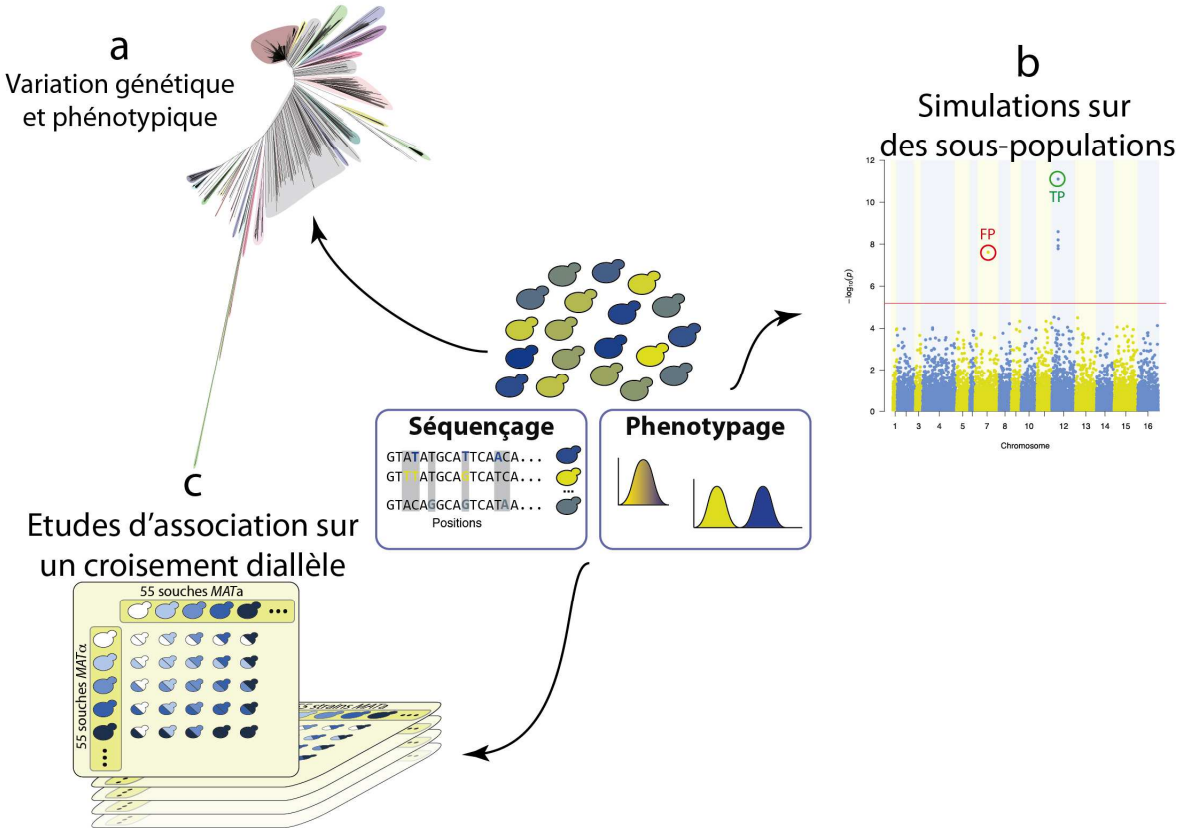


Figure 1 : Aperçu du projet de thèse. **a.** Première partie consistant en la description de la variation génétique ainsi que phénotypique au sein de l'espèce *Saccharomyces cerevisiae* basée sur l'étude de 1011 isolats naturels. **b.** Deuxième partie portant sur la simulation de phénotypes sur différentes sous-populations du jeu de données total pour mesurer la capacité d'identifier les variants causaux à l'aide d'études pangénomiques d'association génotype-phénotypes. **c.** Études d'associations sur une population issue de croisement diallèle.

L'étude de la variation génétique au sein de l'espèce *Saccharomyces cerevisiae* nous a permis d'avoir une vision globale de la diversité au sein de cette espèce. Celle-ci est élevée, avec un maximum de divergence nucléotidique de 1,1% entre les souches les plus éloignées. Les souches naturelles asiatiques sont celles qui montrent une très grande variabilité par rapport aux autres sous-populations. Cette observation, entre autres, suggère une origine asiatique de l'espèce *Saccharomyces cerevisiae*, suivie d'événements de domestication pour la production de pain ou de boissons fermentées ainsi de la colonisation de nouveaux habitats naturels à travers le monde (figure 2). L'existence de populations sauvages de *Saccharomyces cereivisiae* ainsi que d'autres populations domestiquées nous donne une opportunité unique d'étudier les histoires évolutives de chaque sous-population.

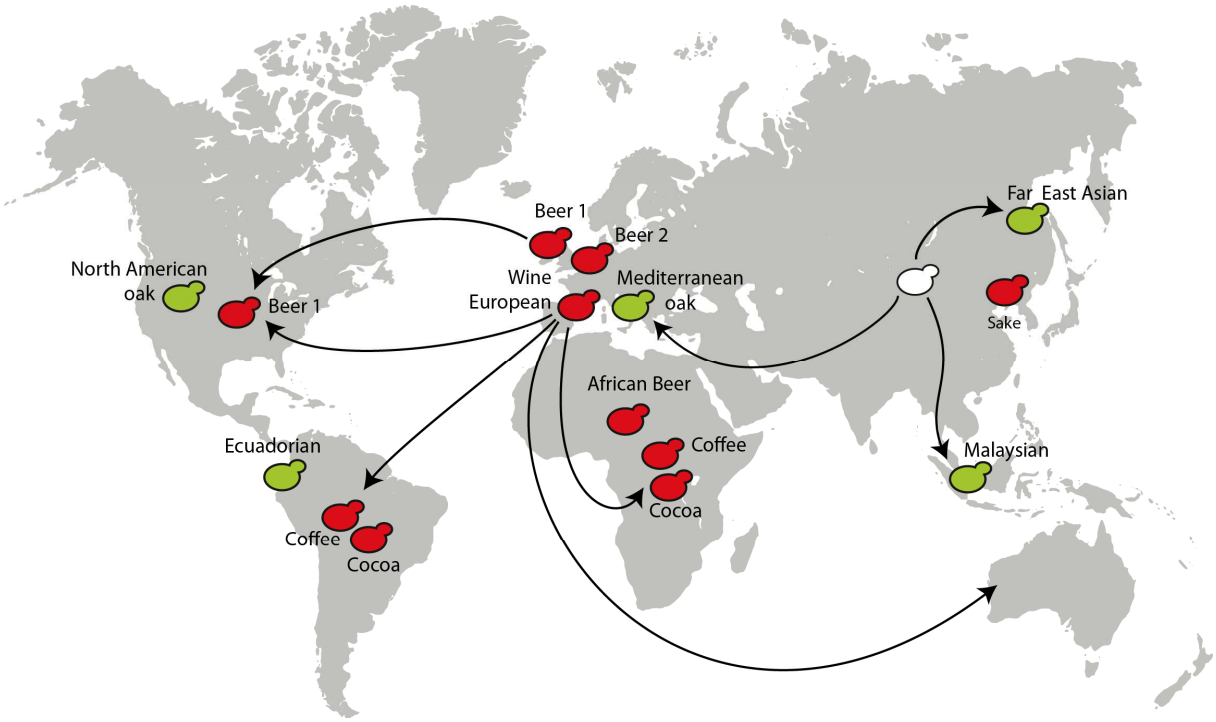


Figure 2 : Différentes sous-populations de *Saccharomyces cerevisiae* à travers le monde. L'origine asiatique supposée de l'espèce est indiquée en blanc. Les populations domestiquées sont en rouge tandis que les populations naturelles sont en vert.

Afin d'avoir une vision globale de l'architecture des traits chez la levure *S. cerevisiae*, j'ai profité de notre jeu de données avec un nombre de génomes jusque là sans précédent pour cette espèce dans le but étudier la relation génotype-phénotype. Pour ce faire, j'ai utilisé une matrice de variants génétiques qui comprend 82 869 mutations ponctuelles ainsi que 925 variants du nombre de copies déterminés à partir de la couverture des phases ouvertes de lecture du pangénome. Parallèlement, des mesures de la vitesse de croissance sur 35 conditions ayant un impact sur divers processus physiologiques et cellulaires (sources de carbone, stabilité de la membrane et des protéines, transduction du signal, biosynthèse du stérol, transcription, traduction, ainsi que les stress osmotiques ou oxydant) ont été effectuées pour 971 isolats naturels séquencés. Des indicateurs de fitness ont été déterminés en normalisant la taille de chaque colonie sur les différentes conditions de stress par la taille de la colonie sur milieu complet. J'ai utilisé ces indicateurs pour effectuer des tests d'association en utilisant un modèle linéaire mixte implémenté par FaST-LMM. De cette manière, j'ai pu mettre en évidence des associations significatives pour 35 variants (22 variants du nombre de copies, 13 mutations ponctuelles) avec 14 conditions, comme par exemple l'implication du variant du nombre de copie du gène *CUP1-2* dans la résistance au cuivre, ou celle des gènes *ARR* pour la résistance aux composés arsénites. Les variants associés expliquent plus de 25%

de la variance phénotypique pour 5 traits testés, avec une variance expliquée par les variants du nombre de copies généralement plus élevée, ainsi qu'une interprétation biologique plus aisée que pour les mutations ponctuelles. Ces résultats soulignent l'importance de prendre en compte tous les types de variants génétiques lorsque l'on réalise des études d'association pour réduire la part d'héritabilité manquante.

Cette première étude m'a permis de mettre en évidence la puissance des études pangénomiques d'association génotype-phénotype chez *S. cerevisiae* pour la détermination de variants génétiques jouant un rôle dans la variation phénotypique. Cependant, seulement la matrice de variants génétiques contenant l'ensemble des individus a été testée. Afin d'évaluer les paramètres influençant la performance des études d'associations, j'ai utilisé 6 populations avec un nombre d'individus et de variants différents (figure 3). Parmi ces populations, 4 forment une lignée génétique pure, une est composée de 133 individus balayant l'ensemble de la diversité de *S. cerevisiae* et la dernière contient l'ensemble des 1011 souches du projet 1002 génomes de levure.

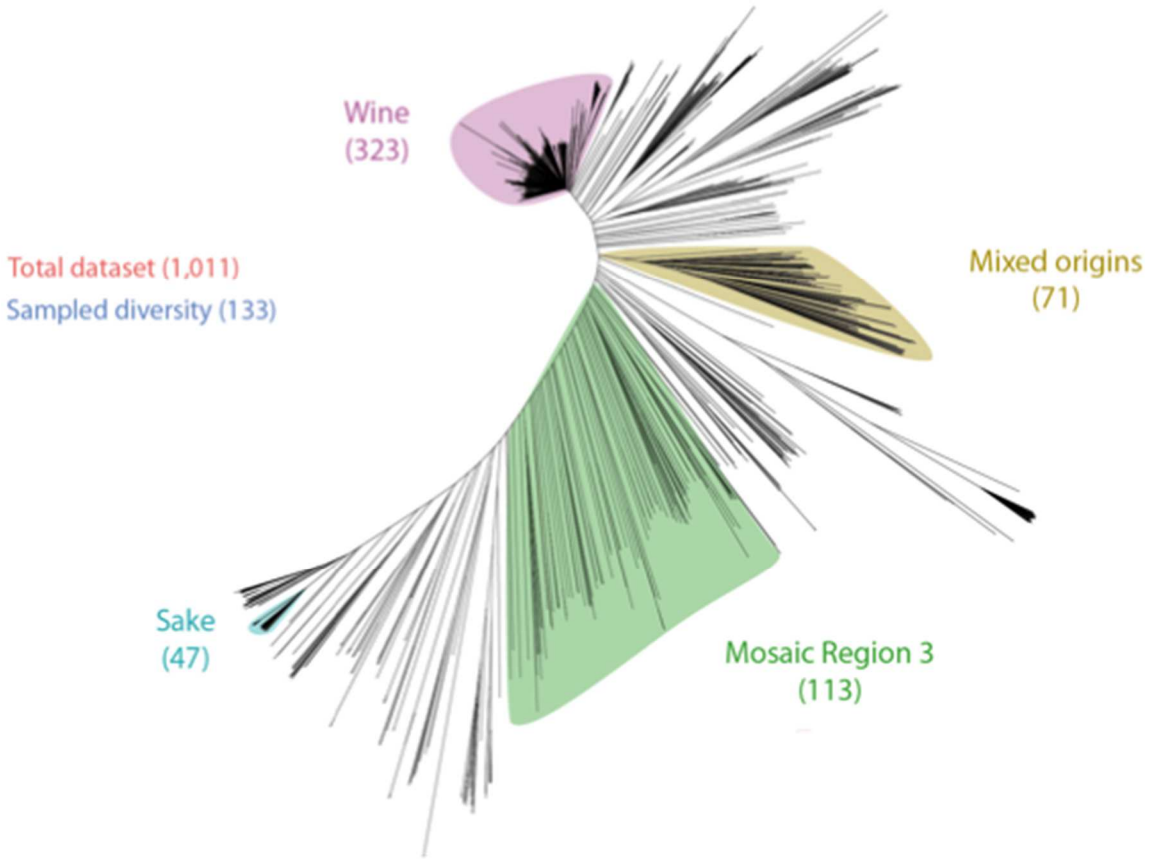


Figure 2 : Jeux de données utilisés pour réaliser les simulations d'études pangénomiques d'association génotype-phénotype. Le nombre d'individus est indiqué entre parenthèses.

Parallèlement, j'ai réalisé un programme permettant de tester ces jeux de données pour permettre leur comparaison. Plus concrètement, j'ai simulé 1000 phénotypes pour chaque jeu de données en choisissant le ou les variants causaux, puis j'ai effectué des études pangénomiques d'association génotype-phénotype pour chaque phénotype. Un seuil de significativité a été calculé en permutant aléatoirement les phénotypes 100 fois, puis en prenant la 5^{e} p-valeur la plus basse, obtenant ainsi un taux d'erreur de première espèce de 5% sur les 1000 séries de simulation. Cette procédure a été utilisée pour tester un phénotype Mendélien (un seul variant causal) ainsi qu'un trait complexe (ici déterminé par 10 variants causaux). L'analyse des résultats m'a permis d'évaluer pour chaque série de simulation, les risque de première et de deuxième espèce, et ainsi de comparer les jeux de données entre eux. Concernant la détection du variant causal d'un trait Mendélien, l'ensemble des séries ont détecté le variant causal à l'exception de 7 séries pour le jeu de données composé de souches utilisées pour la production de saké. Le taux de faux positif est quant à lui variable entre les jeux de données et permet de voir que le nombre d'individus composant le jeu de données n'est pas le seul facteur influençant les résultats des études d'association. En effet, un degré de parenté élevé entre les souches composant un jeu de données comme pour les souches de vin générera plus de faux positifs que le jeu de données balayant l'ensemble de la diversité de l'espèce, malgré un effectif plus élevé. Lorsque l'architecture du trait est complexe, les différences entre les jeux de données apparaissent déjà sur la capacité de détection. En effet, le jeu de données contenant l'ensemble des 1011 individus semble être le plus performant, avec une moyenne de 5 variants détectés, soulignant ici l'importance de la taille du jeu de données pour la détection de multiples variants. Si l'on se penche sur les faux positifs, il semble que la diversité génétique au sein d'un jeu de donnée soit cruciale pour limiter les détections erronées.

La différence entre l'héritabilité génomique des traits et la variance expliquée par les variants significativement associés avec les phénotypes mesurés dans la première partie nous montrent qu'il existe une part non-négligeable d'héritabilité manquante. Un des facteurs pouvant expliquer ce phénomène est la présence de variants rares ayant un grand effet phénotypique. L'influence de ces variants ne serait donc pas détectée avec des études d'association génotype-phénotype, vu que nous filtrons systématiquement les variants avec une fréquence de l'allèle mineur inférieure à 5%. De plus, nous avons noté que presque 93% des sites polymorphiques ont une fréquence de l'allèle mineur inférieure à 5%. Il est donc crucial de trouver des moyens de détecter et d'évaluer l'impact des variants rares. Pour ce faire, nous

avons effectué des études d'association sur une population de 595 hybrides issue d'un croisement par paires entre 34 souches parentales diploïdes homozygotes. Un des avantages d'un croisement diallèle est que cette approche ne nécessite le séquençage que des souches parentales. Les séquences des hybrides sont déduites des parents, ce qui nous permet de générer un jeu de données conséquent (595 individus) à partir de seulement 34 génomes. De plus, un tel schéma de croisement conduit à un mélange et une répétition des haplotypes, ce qui entraîne un changement des fréquences alléliques par rapport à la population totale. De cette manière, nous avons pu inclure 1118 sites polymorphiques qui ne l'étaient pas dans les précédentes études d'association. Les études d'association menées sur 53 conditions nous ont permis d'identifier 2293 mutations associées de manière significative. Parmis celles-ci, 12% sont rares au niveau de la population. Par exemple, nous avons détecté un site polymorphique significativement associé avec la croissance en utilisant du galactose comme source de carbone qui se situe dans la séquence codante du gène *GAL2*, une perméase du galactose (figure 4). Cette mutation explique à elle seule 10,5% de variation phénotypique. De plus la fréquence de l'allèle mineur pour de ce site est de 2% au sein de la population totale des 1011 individus. Cet exemple illustre bien l'importance du rôle joué par les variants rares dans la variation phénotypique, et donc l'importance de pouvoir les inclure dans nos analyses pour réduire la part d'héritabilité manquante.

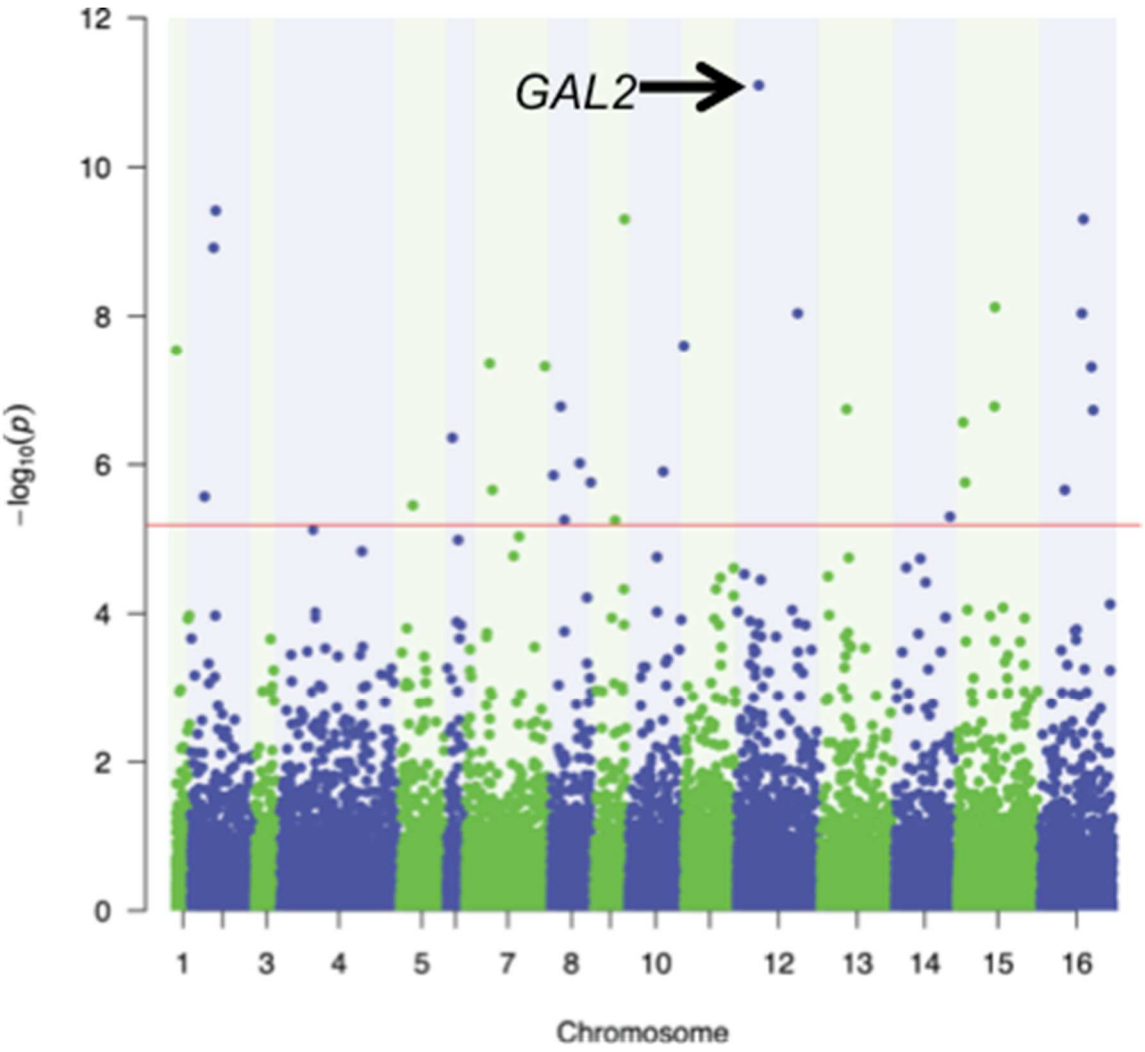


Figure 4 : Résultat de l'association entre les sites polymorphiques et la croissance en présence de galactose comme source de carbone. L'axe des abscisses indique la localisation des sites polymorphiques le long du génome. L'axe des ordonnées correspond au score d'association. La plus forte association a été détectée pour un marqueur présent dans la séquence codante du gène *GAL2*.

L'ensemble de ces travaux nous a permis d'avoir une meilleure compréhension de la variation génétique ainsi que phénotypique au sein d'isolats naturels de *S. cerevisiae*. Il s'agit de la première tentative d'effectuer des études pangénomiques d'association génotype-phénotype en utilisant un aussi grand nombre de génomes de cet organisme, réduisant l'impact des effets confondants sur la détection de variants génétiques responsables de la variation phénotypique. Ces résultats majeurs ouvrent la voie vers une exploration en profondeur de la relation génotype-phénotype chez la levure *S. cerevisiae*.

Jackson PETER

Dissection de la relation génotype-phénotype par des études d'association chez *Saccharomyces cerevisiae*

Résumé

Un objectif central en biologie est de comprendre la relation entre le génotype et le phénotype. Afin de disséquer les bases génétiques de la diversité phénotypique, il est nécessaire de disposer d'une collection de données génomiques d'un grand nombre d'individus d'une même espèce. Dans ce but, mes travaux de thèse se basent sur l'étude des séquences génomiques ainsi que des données phénotypiques de 1011 isolats naturels de la levure *Saccharomyces cerevisiae*. Dans un premier temps, je me suis intéressé à la description de la variation génétique et phénotypique pour dresser un portrait précis de l'histoire évolutive de cette espèce. Les données de phénotypage nous ont permis de réaliser des études pangénomiques d'association génotype-phénotype avec une puissance jusque là inégalée chez *Saccharomyces cerevisiae*. Je me suis par la suite penché sur l'évaluation des paramètres influençant le pouvoir de détection d'une telle approche, d'en apprécier limites pour tenter de les contourner.

Mots clés : génomique des populations, études d'association, levure

Summary

Elucidating the genetic origin of phenotypic diversity among individuals within the same species is essential to understand evolution. Using whole genome sequences of 1,011 Saccharomyces *cerevisiae* isolates, my work sought to describe intraspecific genetic variation and investigate of its phenotypic consequences. Doing so, I obtained a precise view of the evolutionary history of *S. cerevisiae*. Phenotypic characterization provided the opportunity to perform genotype-phenotype genome-wide association studies with unprecedented power. I then focused on the evaluation of the parameters influencing genome-wide association studies, the appreciation of the limits of such an approach, and ways to circumvent them.

Keywords: population genomics, genome-wide association studies, yeast