

Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates

Jackson Peter^{1,6}, Matteo De Chiara^{2,6}, Anne Friedrich¹, Jia-Xing Yue², David Pflieger¹, Anders Bergström², Anastasie Sigwalt¹, Benjamin Barre², Kelle Freel¹, Agnès Llored², Corinne Cruaud³, Karine Labadie³, Jean-Marc Aury³, Benjamin Istace³, Kevin Lebrigand⁴, Pascal Barbry⁴, Stefan Engelen³, Arnaud Lemainque³, Patrick Wincker^{3,5,7}, Gianni Liti^{2,7*} & Joseph Schacherer^{1,7*}

Large-scale population genomic surveys are essential to explore the phenotypic diversity of natural populations. Here we report the whole-genome sequencing and phenotyping of 1,011 *Saccharomyces cerevisiae* isolates, which together provide an accurate evolutionary picture of the genomic variants that shape the species-wide phenotypic landscape of this yeast. Genomic analyses support a single ‘out-of-China’ origin for this species, followed by several independent domestication events. Although domesticated isolates exhibit high variation in ploidy, aneuploidy and genome content, genome evolution in wild isolates is mainly driven by the accumulation of single nucleotide polymorphisms. A common feature is the extensive loss of heterozygosity, which represents an essential source of inter-individual variation in this mainly asexual species. Most of the single nucleotide polymorphisms, including experimentally identified functional polymorphisms, are present at very low frequencies. The largest numbers of variants identified by genome-wide association are copy-number changes, which have a greater phenotypic effect than do single nucleotide polymorphisms. This resource will guide future population genomics and genotype-phenotype studies in this classic model system.

The budding yeast *S. cerevisiae* is a powerful model system for understanding eukaryotic biology at the cellular, molecular and genomic levels^{1,2}. *S. cerevisiae* has recently emerged as a model in population genomics^{3–5}, because it can be found worldwide in a broad array of human-associated (for example, wine, sake, beer and other fermented beverages) and wild (for example, plant, soil and insect) biotopes. Recent years have seen a spike in the number of published *S. cerevisiae* genome sequences, which together have revealed a high level of genetic diversity and a complex population structure in this yeast^{6–12}. However, the number of available sequenced genomes from natural isolates remains limited and stands in contrast to the wealth of data on *Arabidopsis thaliana*¹³ and humans^{14,15}; this small sample size for yeast genomes has not fully captured the global evolutionary processes relevant to the species. Here we apply deep coverage genome sequencing to more than 1,000 natural *S. cerevisiae* isolates and explore their phenotypic landscape. To our knowledge, our large-scale genome analysis provides the first comprehensive view of genome evolution at different levels (for example, accounting for differences among ploidy, aneuploidy, genetic variants, hybridization and introgressions), which is challenging to obtain at this scale and accuracy for other model organisms. This sequencing effort substantiates previous hypotheses but also reveals novel aspects of *S. cerevisiae* evolutionary history. Natural *S. cerevisiae* isolates have previously proven to be a powerful tool for investigating the genotype-phenotype relationship via linkage mapping¹⁶. Although genome-wide association studies (GWAS) have led to the identification of common alleles with strong effects in other organisms, the small sample size of yeast genomes has so far prevented similar attempts for yeast. Our dataset enables GWAS and exhaustively captures genetic variants (single nucleotide polymorphisms (SNPs), copy-number variants (CNVs) and genome content), providing insights into the genetic architecture of traits and the source of missing heritability.

Species-wide genetic and phenotypic diversity

We assembled a collection of 1,011 *S. cerevisiae* isolates that maximized the breadth of their ecological and geographical origins (Supplementary Fig. 1a, b and Supplementary Table 1). We deeply sequenced 918 isolates using the Illumina paired-end strategy with a 232-fold mean coverage, and we also included 93 strains that had previously been sequenced^{6–8} (Supplementary Fig. 1c). 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. Most of these SNPs are present at very low frequencies, with 31.3% of the polymorphic positions being singletons and 93% with a minor allele frequency (MAF) < 0.1 (Supplementary Fig. 2). This bias might in part be driven by the sampling scheme. Deleterious mutations as predicted by SIFT¹⁷ show the strongest bias toward rare alleles, consistent with the notion that selection prevents such SNPs from spreading in the population (Supplementary Fig. 2). In addition, we detected 125,701 small-scale insertions and deletions (indels) (up to 50 bp) with the majority exhibiting a low frequency (Supplementary Fig. 3a). Most indels are short (42.6% are 1 bp in length) and those present in coding regions are strongly biased to lengths that are multiples of three, which reflects the influence of purifying selection (Supplementary Fig. 3b, c). We also characterized CNVs by measuring the coverage ratio of each individual pangenomic open reading frame (ORF) (see below) normalized to the genome of each respective strain (Methods). CNVs are heavily enriched in subtelomeric regions, whereas internal chromosomal regions are largely copy-number stable (Supplementary Fig. 4a). Nearly all ORFs have at least one strain with a CNV. Most CNVs associated with individual ORFs are rare in the population (Supplementary Fig. 4b). Variants with high copy numbers affect only a small fraction of ORFs (Supplementary Fig. 4c) and extreme cases (> 20 copies) include

¹Université de Strasbourg, CNRS, GMGM UMR 7156, Strasbourg, France. ²Université Côte d’Azur, CNRS, INSERM, IRCAN, Nice, France. ³Commissariat à l’Energie Atomique (CEA), Genoscope, Institut de Biologie François-Jacob, Evry, France. ⁴Université Côte d’Azur, CNRS, IPMC, Sophia Antipolis, Valbonne, France. ⁵CNRS UMR 8030, Université d’Evry Val d’Essonne, Evry, France.

⁶These authors contributed equally: Jackson Peter, Matteo De Chiara. ⁷These authors jointly supervised this work: Patrick Wincker, Gianni Liti, Joseph Schacherer. *e-mail: gianni.liti@unice.fr; schacherer@unistra.fr

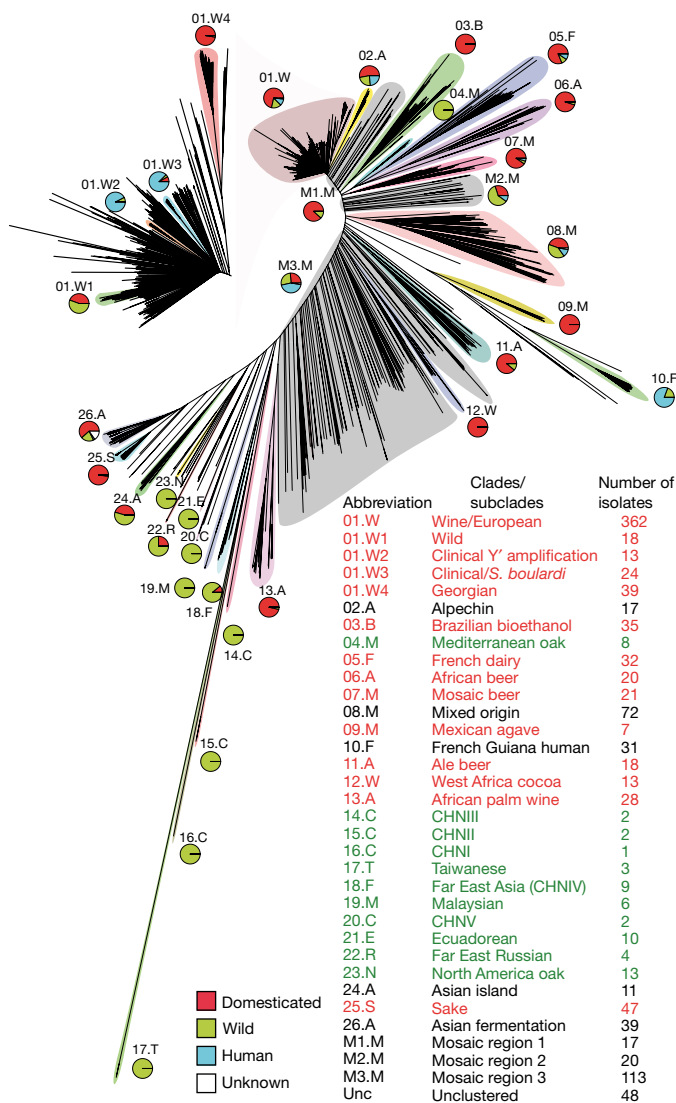


Fig. 1 | Neighbour-joining tree built using the biallelic SNPs. We identified 26 clades (numbered clockwise from 1 to 26) and three mosaic groups (M1–M3). The pie charts represent the ecological origins of the clade: domesticated (red), wild (green) and human (cyan). The colour of the clade name indicates its assignment: domesticated (red) and wild (green). The top left inset represents a magnification of the wine/European clade with four major subclades highlighted.

2 μ plasmid ORFs, mitochondrial genome, ribosomal DNA and repetitive elements such as Ty and Y'.

In parallel, 971 strains were phenotyped in different conditions that affect various physiological and cellular responses (Methods and Supplementary Table 2). In total, we analysed 34,956 phenotypic measurements that covered 36 traits, providing a comprehensive analysis of their inheritance patterns. Most of the traits vary continuously across the population in a manner consistent with their genetic complexity (Supplementary Fig. 5). However, some traits—such as resistance to copper sulfate (CuSO₄) or anisomycin¹⁸—follow a bimodal distribution model, and therefore a Mendelian inheritance pattern. Estimates of the narrow-sense heritability, h^2 , from genome-wide SNPs genotyped¹⁹ show a substantial amount of variance explained across all the traits, with a mean of 0.69, which suggests the feasibility of performing GWAS (Supplementary Fig. 6).

Population structure supports out-of-China origin

The phylogenetic tree of the 1,011 strains shows well-defined clades, loose clusters and isolated branches (Fig. 1). Most of the strains (813 in total) fall into 26 clades, and another 150 strains belong to

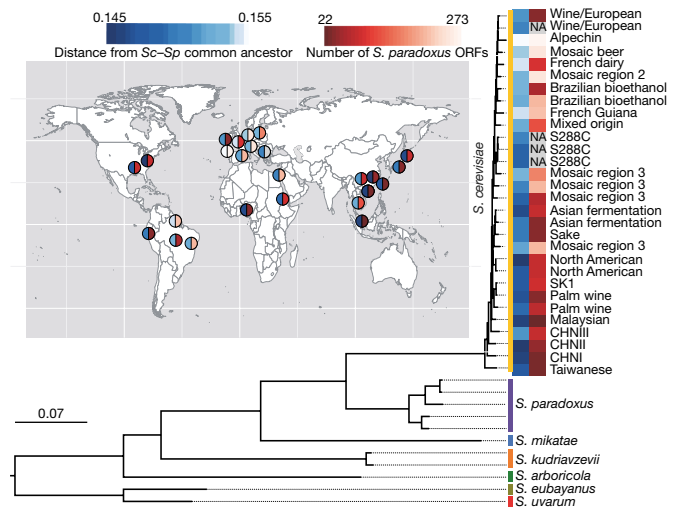


Fig. 2 | Chinese origin of *S. cerevisiae*. Maximum-likelihood rooted tree of the *Saccharomyces* complex, based on the alignment of 2,018 concatenated conserved genes. Heat maps display the distance from the last common ancestor of *S. cerevisiae* (*Sc*)–*S. paradoxus* (*Sp*) (white–blue), and the number of introgressed *S. paradoxus* ORFs (white–red). The map shows the geographical origins of the strains.

three groups of poorly related strains (Supplementary Information note 1). Our data revealed a complex pattern of genetic differentiation with distinct lineages that correlate with geography, environmental niche and the degree of human association, as has previously been reported^{8,9,11,12,20}. Domesticated and wild clades largely fall into two well-delineated sides of the tree, and are separated by a large group of mosaic strains. The main exceptions are the wild Mediterranean oak strains, which group with the domesticated clades, and the sake strains, which group with wild clades. However, the Mediterranean oak lineage groups together with the other wild lineages on the basis of ORF-content strain clustering (Supplementary Fig. 7).

We used ADMIXTURE²¹ to investigate ancestry in the genomes of individual strains. Mosaic strains are characterized by admixture from two or more lineages derived by outbreeding^{3,4} and frequently manifest as isolated branches in the phylogenetic tree. We identified three groups of mosaic strains that are mostly associated with human-related environments (Fig. 1). Population structure analysis revealed different sources of ancestry and degrees of mosaicism, consistent with multiple hybridization events (Supplementary Fig. 8). These findings underscore the role of human-driven admixture in shaping the population structure of *S. cerevisiae*.

The recent discovery of highly diverged wild Chinese lineages suggests that East Asia may represent the geographic origin of *S. cerevisiae*²². The Taiwanese wild lineage represents the most divergent population that has yet been described (average of 1.1% sequence divergence to non-Taiwanese strains). This lineage also contains an extremely divergent 2 μ plasmid that shares only 80% of identity with known plasmid variants (Supplementary Fig. 9 and Supplementary Information note 2). We used a subset of highly contiguous de novo assemblies that sample the main *S. cerevisiae* lineages and closely related *Saccharomyces* species^{23,24} to generate a rooted phylogenetic tree (Fig. 2). The outgroup species branched off near the Taiwanese and Chinese lineages, which strongly supports a Chinese origin for *S. cerevisiae*. This scenario is also consistent with the isolation of closely related *Saccharomyces* species such as *S. mikatae* and *S. arboricola*^{25,26}, which are restricted to East Asia, and the broad genetic diversity of the Japanese *S. kudriavzevii* populations²⁷. Together, these observations suggest an Asian origin for 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 genetic structure of Chinese strains. We performed a principal component analysis on SNPs that

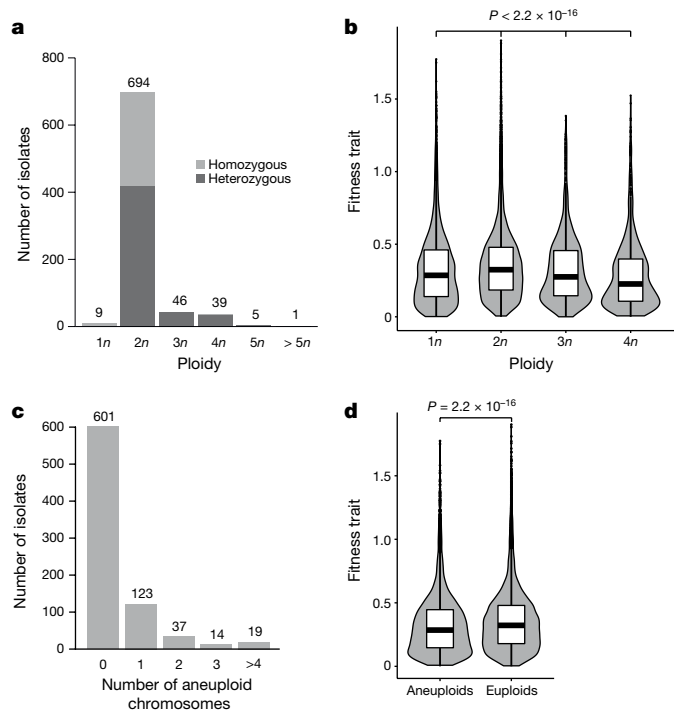


Fig. 3 | Ploidy and aneuploidy natural variation. **a**, Distribution of ploidy and fraction of heterozygous isolates. **b**, Violin plot of growth fitness trait by ploidy. Diploid isolates are globally fitter than individuals with other ploidy levels. Number of trait values for $1n$ isolates = 4,585; for $2n$ isolates = 26,249; for $3n$ isolates = 1,610; and for $4n$ isolates = 1,330. **c**, Distribution of aneuploid chromosomes per individual. **d**, Violin plot of growth fitness trait of aneuploid ($n = 6,510$) and euploid ($n = 20,719$) isolates shows a significant difference in fitness trait between the two categories. All P values were calculated using a two-sided Mann–Whitney–Wilcoxon test. Centre lines, median; boxes, interquartile range (IQR); whiskers, $1.5 \times$ IQR. Data points beyond the whiskers are outliers.

included only the CHN I–V (Chinese isolates) and Taiwanese strains and then projected the rest of the strains onto the principal component space defined by these highly diverged lineages (Supplementary Fig. 10). Principal component 1 defines the separation of the Taiwanese strains from all other strains, consistent with the deep genetic differentiation of this lineage. Principal components 2–4 then define differentiation between different Chinese lineages. Notably, the non-Chinese strains all project onto the same part of the space, which implies that they are essentially all equally related to the different Chinese lineages. In other words, it appears that different non-Chinese strains do not derive from different Chinese lineages but instead share a single out-of-China origin.

Ploidy and aneuploidy variation by ecological origin

Variation in ploidy and aneuploidy is not uncommon in yeast, and this genomic plasticity is often described as a strategy for rapid adaptation to environmental changes^{28–30}. As 217 strains were genetically manipulated and no longer in their natural ploidy states, we assigned a relative ploidy state to the other 794 isolates (Methods), and found 9 haploids, 694 diploids and 91 isolates with higher ploidy (Fig. 3a). Our results reveal that most (about 87%) of the natural isolates are in a diploid state. Polyploid isolates (3–5n) are not frequent (approximately 11.5%) and enriched in specific subpopulations such as the beer, mixed-origin and African palm wine clades, which strongly suggests that some human-related environments have had an effect on the ploidy level (Supplementary Fig. 11). By testing the effect of ploidy on the growth fitness trait across the species, a general advantage was observed for diploid isolates compared to other types of ploidy (Fig. 3b). This result supports a general mitotic growth advantage for diploidy in yeast^{30,31} that holds across every condition tested, and shows no major environment-specific effects (Supplementary Fig. 12).

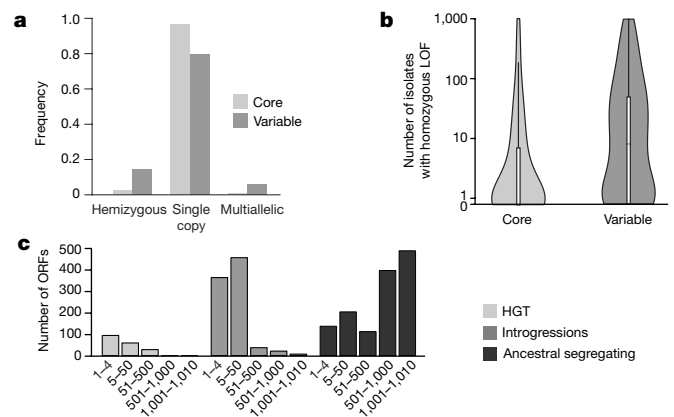


Fig. 4 | The *S. cerevisiae* pangenome. **a**, Copy number distribution for core and variable ORFs. Variable ORFs have a greater frequency of both hemizygous and multiallelic genes. **b**, Logarithmic-scale distribution of isolates carrying loss-of-function mutations for core ($n = 4,931$) and variable ORFs ($n = 1,111$). The core genome is characterized by far fewer loss-of-function mutations compared to variable ORFs (P value = 6.45×10^{-78} , two-sided Mann–Whitney–Wilcoxon test). Centre lines, median; boxes, IQR; whiskers, $1.5 \times$ IQR. Data points beyond the whiskers are outliers. **c**, Different types of variable ORFs have marked differences in distribution.

By combining coverage analysis and allele frequency distributions, we determined the copy number of each chromosome together with instances of segmental duplications (Fig. 3c, Supplementary Fig. 13a and Supplementary Table 1). We identified a total of 342 cases of aneuploidy that affected 193 isolates (19.1% of the 1,011 strains). The most frequently observed cases of aneuploidy involve chromosomes I, III and IX and are only weakly correlated with chromosome size (Supplementary Fig. 13b). There is a strong enrichment of aneuploid strains in the sake, ale beer and mixed-origin clades (Supplementary Fig. 14). Aneuploidy is therefore not uncommon but its relationship with fitness is paradoxical. Indeed, cases of aneuploidy arise under a variety of selective regimes but lead to a decrease of global cellular fitness^{32–34}. We found a general mitotic growth advantage in euploid versus aneuploid strains, consistent with a fitness cost for chromosomal aneuploidy (Fig. 3d).

A portrait of the *S. cerevisiae* pangenome

The 1,011 genomes provided an opportunity to determine the yeast pangenome³⁵ using de novo genome assemblies and detection of non-reference material (Methods). Containing a total of 7,796 ORFs, the pangenome (Supplementary Table 3) is composed of 4,940 core ORFs and 2,856 ORFs that are variable within the population. Most core ORFs are present as a single copy per haploid genome, whereas variable ORFs show a higher frequency of both hemizygous and multiallelic copy numbers (Fig. 4a). Analysis of the 6,081 non-redundant ORFs present in the well-annotated S288C reference genome (4,937 core and 1,144 variable ORFs) highlighted different trends. First, the distribution of variable ORFs is biased towards subtelomeric regions (Supplementary Fig. 15a), known as hotspots of gene content variation^{7,24,36}, and shows a strong functional enrichment for cell–cell interactions, secondary metabolisms and stress responses (Supplementary Table 4). Second, the core genome is characterized by lower levels of loss-of-function mutations and has a different ratio of substitution rates (ratio of nonsynonymous to synonymous polymorphisms, d_N/d_S), which reflects differences in selective constraints (Fig. 4b and Supplementary Fig. 15b). Out of the 1,072 essential genes defined in the S288C background, 123 do not belong to the core genome, although the absence of 71 of these is complemented by closely related orthologues (Supplementary Table 5).

To trace the origins of the variable ORFs, we implemented a phylogenetic approach by inspecting the evolution of each individual

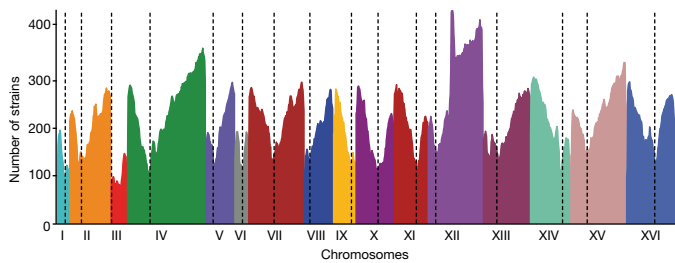


Fig. 5 | Landscape of loss-of-heterozygosity events. Density of the regions under LOH across the 16 nuclear chromosomes (I–XVI) within our population. Each colour corresponds to a chromosome, and centromere locations are represented by dotted lines.

ORF (Methods and Supplementary Fig. 16). We defined 1,380 ancestral segregating ORFs with sequence-similarity levels consistent with genome-wide species relatedness. We identified 913 introgressed ORFs, with a clear majority ($n = 885$) being unambiguously traced to a *Saccharomyces paradoxus* origin. All *S. cerevisiae* isolates carry at least one *S. paradoxus* ORF (median of 26), indicating ubiquitous gene flow between these two closely related species³⁷. We also detected 14 ORFs in the Taiwanese lineage that were introgressed from an undocumented species (Supplementary Fig. 17). The amount of introgressed content is highly variable, with an enrichment in four human-associated subpopulations in which two species might coexist and which might represent interspecific hybrid zones (Supplementary Fig. 18 and Supplementary Information note 3). By contrast, introgressions are rare in the highly diverged lineages, consistent with secondary contacts with *S. paradoxus* occurring mainly after the out-of-China dispersal (Fig. 2). There is a notable match between the geographic origins of the *S. cerevisiae* clades and the ancestry of introgressed ORFs from *S. paradoxus* (Supplementary Fig. 19 and Supplementary Information note 3).

Finally, 183 ORFs are likely to be the result of horizontal gene transfer (HGT) events from highly divergent yeast species. Introgressed ORFs tend to replace *S. cerevisiae* orthologues, which suggests that they are integrated by homologous recombination; by contrast, HGT segments localize mainly at subtelomeres. We traced the donor for 85 HGT ORFs and found an enrichment for *Torulaspora* and *Zygosaccharomyces* species. These species coexist with *S. cerevisiae* in fermentative environments, which might promote HGT events or favour their retention. We identified 6 large (38–165 kb) HGT events with most isolates retaining only small segments in complex patterns, consistent with multiple independent rearrangements leading to partial deletions of the large ancestral HGT (Supplementary Information note 3, Supplementary Figs. 20–22 and Supplementary Table 6). Together, these introgression and HGT events with distinct population frequency distributions (Fig. 4c) correspond to important evolutionary processes that have shaped the *S. cerevisiae* species genome.

Patterns of extensive loss-of-heterozygosity

S. cerevisiae is highly inbred and characterized by rare sexual cycles^{38–40}. The frequency of outcrossing has considerable effects on genome variation, and particularly on patterns of heterozygosity. Among the 794 natural isolates, 505 isolates (about 63%) are heterozygous (Fig. 3a), with the proportion varying across subpopulations and with marked differences between domesticated and wild clades (Supplementary Fig. 23), as has previously been observed⁴¹. Heterozygous sites were spread across the genomes, but we also detected large regions of loss-of-heterozygosity (LOH) and generated an accurate genome-wide LOH map (Fig. 5, Methods and Supplementary Fig. 24a). LOH events range from 2 to 56 regions per strain and represent up to 80% of the sake isolate genomes (Supplementary Fig. 24b). Although LOH levels are variable across subpopulations (Supplementary Fig. 25), we observed an overall high LOH level with 25 regions covering approximately 50% of the genome on average (Supplementary Fig. 24b, c and Supplementary

Table 7). However, LOH events are not evenly distributed along the genome (Fig. 5) and centromere-proximal regions exhibit low levels of recombination initiation, consistent with them experiencing few LOH events.

By masking the LOH regions we precisely determined the level of heterozygosity, which ranges from 0.63 to 6.56 heterozygous sites per kb (Supplementary Table 8). The distribution of the level of heterozygosity in the population is bimodal, reflecting the variability observed across subpopulations (Supplementary Fig. 26). These observed patterns are likely to be related to variation in the rate of outcrossing rate, as LOH and heterozygosity correlate with one another (Supplementary Fig. 27). The prevalence of LOH events agrees with the low outcrossing rate³⁸ and most LOH is thought to be the result of mitotic recombination⁴² or cells returning to mitotic growth after entering the meiotic phase^{43,44}. Overall, our data support the idea that *S. cerevisiae* undergoes clonal expansion followed by LOH-mediated diversification, enabling the expression of recessive alleles and generating novel allele combinations with potential effects on phenotypic diversity.

Genetic diversity and evolution by subpopulation

The comparison of genome content variation and levels of SNPs in domesticated and wild clades (Supplementary Fig. 28) shows higher SNP density (median 0.55% versus 0.41%) and lower genome content variation (median 115 ORFs that are not shared, versus 161 shared ORFs) in wild versus domesticated clades, respectively (Supplementary Fig. 28 and Supplementary Tables 9, 10). These findings suggest a shift in evolutionary mechanisms during the domestication process. The wild clades share similar genome content, and their evolution is mainly driven by the accumulation of SNPs. The specific artificial environments colonized by domesticated clades probably promote rapid ORF expansion and/or loss, leading to considerable variation in genome content and CNVs. Some domesticated clades also exhibit high copy numbers of Ty1 and Ty2 transposable elements (Supplementary Fig. 29).

We investigated evolutionary patterns across multiple *S. cerevisiae* subpopulations using 19 clades that contained at least 10 isolates. By determining the ploidy and measuring the genome-wide levels of genetic diversity (π and θ_w), the MAF spectrum and Tajima's D values, our results revealed distinct evolutionary histories across subpopulations (Supplementary Table 11). We also estimated the timings of the *S. cerevisiae* out-of-China and domestication events (Supplementary Information note 4).

Analyses of human-related *S. cerevisiae* subpopulations provide strong evidence for various independent and lineage-specific domestication events. Although *S. cerevisiae* beer isolates are polyphyletic, they are characterized by higher ploidy ($\geq 3n$) and a higher number of aneuploidy events than other domesticated lineages. In addition, the genetic diversity of beer isolates is very high ($\pi = 2.8 \times 10^{-3}$ on average) and the number of heterozygous sites they possess is elevated (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, therefore, marked by genome-level modification and high nucleotide variation.

By contrast, wine and sake isolates are primarily diploids, monophyletic and have limited genetic diversity (π values of 1×10^{-3} and 0.8×10^{-3} , respectively). The wine cluster is characterized by a strong bias towards low-frequency polymorphisms, with more than 95% of the polymorphic sites with $MAF < 0.1$. In addition, wine isolates harbour low heterozygosity and extensive LOH, which suggests a low outcrossing rate. All of these observations indicate that wine isolates experienced a population expansion after a domestication bottleneck. The effect of the domestication event on the sake genomes is very similar. Indeed, these genomes have low levels of genetic diversity, low heterozygosity and extensive LOH regions. However, the sake subpopulation does not exhibit a bias towards low-frequency alleles (Tajima's D value of 0.0481), reflecting their more recent origin (Supplementary Information note 1)⁴⁵.

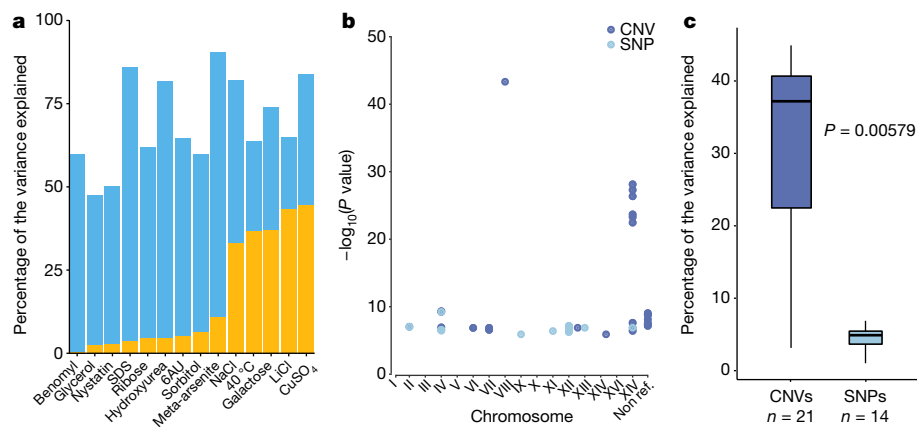


Fig. 6 | Genotype–phenotype relationship in *S. cerevisiae*. **a**, Narrow-sense heritability (blue) and phenotypic variance explained (yellow) for phenotypes with associated variants. **b**, Association scores of the detected genetic variants across the 16 chromosomes and the non-reference ORFs. **c**, Variance explained by CNVs and SNPs associated with traits.

New insights into the genotype–phenotype relationship

Natural *S. cerevisiae* isolates have been a powerful tool for linkage mapping, revealing an large number of quantitative trait loci, over 100 quantitative trait genes and 50 quantitative trait nucleotides¹⁶. We investigated the allele frequency associated with these polymorphisms, which underlie quantitative trait variation (Supplementary Table 12). Among 36 well-characterized quantitative trait nucleotides, 24 were found with at a frequency of lower than 5%, and were therefore undetectable using a GWAS strategy. This strong bias towards rare alleles for these polymorphisms is consistent with the overall MAF spectrum. Consequently, this result highlights that a fraction of the missing heritability of complex traits might be explained by rare variants in species with an allele frequency distribution similar to that of yeast.

The high genetic diversity ($\pi = 3 \times 10^{-3}$) and low linkage disequilibrium ($LD_{1/2} = 500$ bp) (Supplementary Fig. 30) among *S. cerevisiae* isolates indicates that this species could represent a powerful resource for GWAS. We built a matrix of genetic variants that included comprehensive sets of SNPs, CNVs and non-reference variable ORFs. Our matrix contains a total of 82,869 SNPs and 925 CNVs, which represents a dense map with, on average, one marker every 143 bp. We estimated genome-wide phenotypic heritability¹⁹ and revealed that traits are not stratified by subpopulations (Fig. 6a, Supplementary Figs. 31, 32 and Supplementary Tables 13, 14).

We performed a mixed-model association¹⁹ and detected 35 genetic variants associated with 14 conditions (22 CNVs versus 13 SNPs), with an enrichment and high association scores for CNVs (Fig. 6b, Supplementary Fig. 33 and Supplementary Table 15). In addition, four of the variants we detected are linked to non-reference ORFs. For five traits, the phenotypic variation explained is greater than 25% (Fig. 6a). CNVs explained larger proportions of trait variance compared to SNPs (a median of 36.8% and 4.49% of the variance explained, respectively; Fig. 6c). As an example, we found the *CUP1* gene strongly associated with resistance to copper sulfate (P value = 4.85×10^{-44}) (Supplementary Fig. 33 and Supplementary Table 15). Amplification of this locus strongly contributes to the resistance to high concentrations of copper and cadmium⁴⁶ with copy number variation alone explaining 44.5% of phenotypic variation. Our GWAS analyses, which included an exhaustive catalogue of genome content and CNVs, highlighted the overall importance of these genetic variants for phenotypic diversity. The high number of associated CNVs is consistent with the notion that these variants can contribute to a large amount of phenotypic variation⁴⁷.

Conclusion

Our whole-genome sequencing of 1,011 isolates, combined with our phenotyping efforts, provides a detailed view of *S. cerevisiae* variation.

Association scores and variance explained are higher for CNVs compared to SNPs (P value = 0.00579, two-sided Mann–Whitney–Wilcoxon test). Centre lines, median; boxes, IQR; whiskers, $1.5 \times$ IQR. Data points beyond the whiskers are outliers.

This resource has revealed previously undescribed evolutionary history as well as the driving forces of genome evolution, and has provided insights into the genotype–phenotype relationship. Our study lays the foundation for GWAS in *S. cerevisiae* and provides a population genomic resource at a scale that matches those of other model organisms^{13,14}. The difference between the estimated genome-wide heritability and explained phenotypic variance gives an overview of the extent of missing heritability^{48,49}. Many SNPs are present at low frequencies, which echoes observations previously made in human GWAS¹⁵ and raises the question of whether rare SNPs have an important role in modulating the phenotypic landscape. Furthermore, the comprehensive characterization of the species pangenome can further improve the genotype-to-phenotype map. The availability of end-to-end genome assemblies will enable the organization of such a dataset with genome graphs⁵⁰. This collection of genetic and phenotypic variants will therefore enable novel functional approaches in a powerful model system.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at <https://doi.org/10.1038/s41586-018-0030-5>.

Received: 4 May 2017; Accepted: 9 January 2018;

Published online: 11 April 2018

- Botstein, D. & Fink, G. R. Yeast: an experimental organism for 21st century biology. *Genetics* **189**, 695–704 (2011).
- Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546–567 (1996).
- Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**, 342–345 (2009).
- Liti, G. et al. Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
- Peter, J. & Schacherer, J. Population genomics of yeasts: towards a comprehensive view across a broad evolutionary scale. *Yeast* **33**, 73–81 (2016).
- Skelly, D. A. et al. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23**, 1496–1504 (2013).
- Bergström, A. et al. A high-definition view of functional genetic variation from natural yeast genomes. *Mol. Biol. Evol.* **31**, 872–888 (2014).
- 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**, 762–774 (2015).
- Almeida, P. et al. A population genomics insight into the Mediterranean origins of wine yeast domestication. *Mol. Ecol.* **24**, 5412–5427 (2015).
- Zhu, Y. O., Sherlock, G. & Petrov, D. A. Whole genome analysis of 132 clinical *Saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3 (Bethesda)* **6**, 2421–2434 (2016).
- Gallone, B. et al. Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. *Cell* **166**, 1397–1410.e16 (2016).

12. Gonçalves, M. et al. Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. *Curr. Biol.* **26**, 2750–2761 (2016).
13. The 1001 Genomes Consortium. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* **166**, 481–491 (2016).
14. The 1001 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
15. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
16. Fay, J. C. The molecular basis of phenotypic variation in yeast. *Curr. Opin. Genet. Dev.* **23**, 672–677 (2013).
17. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols* **4**, 1073–1081 (2009).
18. Hou, J. et al. The hidden complexity of Mendelian traits across natural yeast populations. *Cell Reports* **16**, 1106–1114 (2016).
19. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
20. Ludlow, C. L. et al. Independent origins of yeast associated with coffee and cacao fermentation. *Curr. Biol.* **26**, 965–971 (2016).
21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
22. 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**, 5404–5417 (2012).
23. Borneman, A. R. & Pretorius, I. S. Genomic insights into the *Saccharomyces sensu stricto* complex. *Genetics* **199**, 281–291 (2015).
24. Yue, J. X. et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. *Nat. Genet.* **49**, 913–924 (2017).
25. Boynton, P. J. & Greig, D. The ecology and evolution of non-domesticated *Saccharomyces* species. *Yeast* **31**, 449–462 (2014).
26. Liti, G. et al. High quality *de novo* sequencing and assembly of the *Saccharomyces arboricolus* genome. *BMC Genomics* **14**, 69 (2013).
27. Hittinger, C. T. et al. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**, 54–58 (2010).
28. Hose, J. et al. Dosage compensation can buffer copy-number variation in wild yeast. *eLife* **4**, e05462 (2015).
29. Sunshine, A. B. et al. The fitness consequences of aneuploidy are driven by condition-dependent gene effects. *PLoS Biol.* **13**, e1002155 (2015).
30. Gerstein, A. C., Chun, H. J., Grant, A. & Otto, S. P. Genomic convergence toward diploidy in *Saccharomyces cerevisiae*. *PLoS Genet.* **2**, e145 (2006).
31. Zörgö, E. et al. Ancient evolutionary trade-offs between yeast ploidy states. *PLoS Genet.* **9**, e1003388 (2013).
32. Torres, E. M. et al. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317**, 916–924 (2007).
33. Yona, A. H. et al. Chromosomal duplication is a transient evolutionary solution to stress. *Proc. Natl Acad. Sci. USA* **109**, 21010–21015 (2012).
34. Tan, Z. et al. Aneuploidy underlies a multicellular phenotypic switch. *Proc. Natl Acad. Sci. USA* **110**, 12367–12372 (2013).
35. Tettelin, H. et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl Acad. Sci. USA* **102**, 13950–13955 (2005).
36. 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).
37. Dunn, B., Richter, C., Kvitek, D. J., Pugh, T. & Sherlock, G. Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res.* **22**, 908–924 (2012).
38. Ruderfer, D. M., Pratt, S. C., Seidel, H. S. & Kruglyak, L. Population genomic analysis of outcrossing and recombination in yeast. *Nat. Genet.* **38**, 1077–1081 (2006).
39. 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. USA* **105**, 4957–4962 (2008).
40. 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**, 184–192 (2015).
41. Magwene, P. M. et al. Outcrossing, mitotic recombination, and life-history trade-offs shape genome evolution in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **108**, 1987–1992 (2011).
42. Llorente, B., Smith, C. E. & Symington, L. S. Break-induced replication: what is it and what is it for? *Cell Cycle* **7**, 859–864 (2008).
43. Laureau, R. et al. Extensive recombination of a yeast diploid hybrid through meiotic reversion. *PLoS Genet.* **12**, e1005781 (2016).
44. Brion, C. et al. Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. *PLoS Genet.* **13**, e1006917 (2017).
45. Ohnuki, S. et al. Phenotypic diagnosis of lineage and differentiation during sake yeast breeding. *G3 (Bethesda)* **7**, 2807–2820 (2017).
46. Fogel, S. & Welch, J. W. Tandem gene amplification mediates copper resistance in yeast. *Proc. Natl Acad. Sci. USA* **79**, 5342–5346 (1982).
47. Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
48. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
49. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
50. Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res.* **27**, 665–676 (2017).

Acknowledgements We thank our colleagues (listed in Supplementary Table 1) for their gifts of yeast strains, and particularly the DBVPG and NCYC yeast collections for donating large batches of strains and assistance with identification. We thank B. Dujon and D. Speed for discussions. The sequencing effort was funded by France Génomique (ANR-10-INBS-09-08). This work was also supported by the National Institutes of Health (NIH) grant R01 GM101091-01 to J.S. and by ATIP-Avenir (CNRS/INSERM), Fondation ARC (SFI20111203947), the French National Research Agency (ANR-13-BSV6-0006-01, ANR-16-CE12-0019, ANR-11-LABX-0028-01), Cancéropôle PACA (AAP emergence) and DuPont Young Professor Award to G.L. J.P. was supported by Ministère de l'Enseignement Supérieur, J.-X.Y. by Fondation ARC (PDF20150602803) and B.B. by La Ligue Contre le Cancer (GB-MA-CD-11287). We also acknowledge the IRCAN Flow Cytometry Facility CytoMed. J.S. is a Fellow of the University of Strasbourg Institute for Advanced Study (USIAS) and a member of the Institut Universitaire de France.

Reviewer information Nature thanks D. Gresham, G. Sherlock and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.F. and A.LI. prepared the DNA samples. A.S. and B.B. performed the phenotyping. J.P., M.D.C., A.F., J.-X.Y., D.P., A.B. and K.Le. analysed the data. C.C., K.La., J.-M.A., B.I., S.E., A.Le. and P.W. sequenced the genomes and produced the data. P.B., P.W., G.L. and J.S. contributed tools and materials. G.L. and J.S. conceived and coordinated the project. G.L. and J.S. wrote the paper with the input of J.P., M.D.C. and A.F.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0030-5>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to G.L. or J.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

***S. cerevisiae* sequenced isolates.** The isolates included in this project were carefully selected to be representative of the *S. cerevisiae* whole species. All the isolate details, including ecological and geographical origins, providers and references, are listed in Supplementary Table 1. We maximized the isolate ecological origins by including both human-associated environments such as wine and sake fermentation, brewing and dairy products, as well as natural environments such as soil, insects, tree exudate and fruit. Geographical origins are also highly diverse and have a worldwide distribution (Supplementary Table 1). In addition to the 918 isolates provided by research laboratories and yeast collections, we included 93 strains sequenced in previous studies^{6–8}, to give a total of 1,011 samples analysed in this study. We sought to keep the isolates in their natural state before sequencing to provide a global picture of the ploidy and level of heterozygosity. However, among the 918 selected isolates, 124 were non-natural haploid with the *HO* gene deleted and the 93 external isolates were genetically manipulated and made homozygous 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. Ten 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 software designed in-house, based on the FastX package. The last step identifies and discards read pairs that mapped to the phage phiX genome (GenBank code: NC_001422.1) using SOAP⁵¹. A total of 3.35 Tb of high-quality genomic sequence was generated with a mean sequencing depth of $232 \times$ per isolate (ranging from $50 \times$ to $1,014 \times$, Supplementary Fig. 1c). For the publically available Illumina paired-end reads related to 93 strains (see '*S. cerevisiae* sequenced isolates'), the mean sequencing depth is $169 \times$ (from $20 \times$ to $570 \times$, Supplementary Fig. 1c).

Reads mapping and variant calling. For each isolate, the reads were mapped to the *S. cerevisiae* S288C reference genome (version R64-1-1) with Burrows–Wheeler Aligner (BWA v.0.7.4-r385)⁵², using default parameters. Duplicated reads were marked with Picard-tools (v.1.124) (<http://picard.sourceforge.net>) and local realignment around indels and variant calling were performed with GATK (v.3.3-0)⁵³. Default parameters were applied except for the realignment step (GATK IndelRealigner), in which 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.

Ploidy, types of aneuploidy and segmental duplications. The natural ploidy of the 794 natural isolates (see '*S. cerevisiae* sequenced isolates'), as well as their aneuploidy and segmental duplication content were investigated by combining three complementary approaches:

First, 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). Five microlitres of the culture were transferred into 195 µl of fresh YPD and incubated for 8 h 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, 100 µg/ml RNase A, 0.1% v/v Triton-X, in PBS) and finally incubated for 3 h at 37 °C in the dark. Ten thousand 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 a 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 1,000). 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 cell values between 39 and 181 and G2 values between 148 and 255 were labelled as haploid; strains with G1 cell values between 145 and 265 and G2 values between 295 and 500 were labelled as diploid; strains with G1 cell values between 245 and 355 and G2 values between 500 and 700 were labelled as triploid; strains with G1 cell values between 295 and 500 and G2 values between 700 and 905 were labelled as tetraploid; strains with G1 cell values between 395 and 605 and G2 values over 905 were labelled as over $4n$; strains with other combinations of values have been manually evaluated.

Second, study of sequencing coverage: systematic analysis of the coverage depth along the genome was performed with 1-kb non-overlapping sliding windows, which enabled 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.

Third, investigation of the allele balance ratio associated with heterozygous SNPs, as heterozygous sites should fit an expected range of ratios according to the copy number of the chromosome being considered (see 'SNPs filtering and matrix').

The precise locations of segmental duplications were manually investigated in the .vcf files (Supplementary Table 16).

SNP filtering and matrix. For each sample, variants were first called with GATK HaplotypeCaller (see 'Reads mapping and variant calling'). At this stage, isolates with less than 5% of heterozygous sites (average percentage of heterozygous sites detected in a sample of 104 haploid and/or homozygous diploid isolates) 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:

First, a minimal coverage depth of $50 \times$ was required for a SNP to be retained for the 918 isolates that were sequenced in this study; this coverage depth was lowered to $10 \times$ for the other 93 previously sequenced isolates.

Second, for the haploid and homozygous isolates ($< 5\%$ of heterozygous sites), the fraction of heterozygous SNPs detected was considered as representing false positives and was therefore filtered out.

Third, for heterozygous isolates, heterozygous sites were filtered according to their allele balance ratio (ABHet). The thresholds for allele balance ratios were determined according to the allelic frequency distribution all over the heterozygous samples at each level of ploidy (from $2n$ to $5n$). A heterozygous site was rejected when its ratio did not fit the expected range according to the copy number of the chromosome being considered (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, see 'Data availability'). This matrix of SNPs included 1,625,809 segregating sites across our 1,011 isolates (Supplementary Table 1).

SnEff (v.4.1)⁵⁴ was used to annotate and predict the effect of the variants. Non-synonymous SNPs predicted as deleterious by SIFT (v.5.2.2)¹⁷ as well as non-sense mutations were considered as deleterious for protein function. Insertions and deletions were considered to cause frame shifts when their sum produced a number not divisible by three in a single gene.

A sequence representative of each isolate was constructed by inserting these filtered SNPs into the reference sequence with GATK FastaAlternateReferenceMaker (see 'Data availability').

Pangenome. De novo genome assembly. We used Abyss software (v.1.5.2)⁵⁵ with the option '-k 64' to produce the de novo assemblies (see 'Data availability'). The pre-assembly filtering step was performed with condetri v.2.2 to remove the 6 bases closest to the 5' end and to discard low-quality 3'-end bases of reads. 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 (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) 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 owing to the presence of *Staphylococcus epidermidis* 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 with a length that 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 alignment identity of 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 as the first step) to eliminate repeated elements. When two sub-sequences were found to have an alignment identity of 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 gave rise to an evidence-leveraged final annotation²⁴. We independently ran the three individual components (RATT⁵⁶, YGAP⁵⁷ and MAKER⁵⁸) for gene annotation, and subsequently integrated their results using EVM⁵⁹. Proteomes of the *Saccharomyces* species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. arboricolus*, *S. uvarum* and *S. eubayanus*) were retrieved and used in our annotation pipeline to provide protein alignment support for annotated gene models^{60,61}.

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 *Saccharomyces* Genome Database (SGD) database (updated 13 January 2015, https://downloads.yeastgenome.org/sequence/S288C_reference/orf_dna/). Three RDN genes were also added (*RDN18-1*, *RDN25-1* and *RDN37-1*) from the set 'rna_genomic' available from the SGD database (https://downloads.yeastgenome.org/sequence/S288C_reference/rna/). We applied a graph-based pipeline to this set of ORFs, 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 represents an alignment identity of over 95% on 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 7,796 non-redundant ORFs (see 'Data availability'), which represent the *S. cerevisiae* pangenome (Supplementary Table 3). Among the reference ORFs, the number of similar ORFs collapsed for each final ORFs has a wide range (up to 67, which is the cluster of Ty1 elements), although usually this number does not exceed 2. Other large clusters are the Y' elements one (59 ORFs) and the Ty2 elements (27 ORFs). Out of the 6,713 reference ORFs, 5,679 were not redundant and 1,032 were collapsed into 402 unique ORFs; 89% of these unique ORFs ($n = 357$) are duplicated ORFs.

Pangenome CNVs. 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 parameters and the option `-U 0`. The result was then filtered using samtools view with the options `-bSq 20 -F 260`. The median coverage for each ORF 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 the copy number for the haploid genome (see 'Data availability').

The mapping was also used as a confirmation step for the presence of the ORFs in each strain, leading to the identification of 4,940 ORFs present in the 1,011 strains of the collection, representing the core genome plus 2,856 ORFs present in different subsets of the population. Fifty-two ORFs were removed because they were present in single strains with low coverage (~10% of the genome wide coverage) and were likely to be contaminations from *Escherichia coli* and *Clavispora lusitaniae*. Eighty-nine other ORFs that did not have sufficient coverage were kept in the pangenome, but were not used for subsequent analyses owing 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 copy-number median for each ORF to avoid sample bias. The distributions of medians in the domesticated and wild clades were then compared using the Mann–Whitney–Wilcoxon test (R function `wilcox.test`) (Supplementary Fig. 28 and Supplementary Tables 18, 19).

Inference of pangenome origin. We constructed a local ORFs database for 57 representative species that deeply probed both closely related *Saccharomyces* species as well as a highly divergent yeast species⁶² (Supplementary Table 20). In addition, we added the ORFs of 12 representative *S. cerevisiae* and *S. paradoxus* strains with complete genome sequenced by long reads (PacBio)²⁴. For each annotated variable ORF, we first performed a BLASTN search ('-eval 1E-6') against this local ORFs database to find its best hit. ORFs without hits in our local yeast ORFs database underwent a further round of BLAST searching ('-eval 1E-6') against the NCBI non-redundant database (<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>). Based on the sequence identity and query coverage of the top hits, we classified the variable genes into different categories.

d_N/d_S . For all isolates, sequences of the protein-coding genes were inferred from the filtered SNPs and inserted into the reference sequence with GATK FastaAlternateReferenceMaker. For each gene, the coding sequences were aligned and the ratio of nonsynonymous to synonymous polymorphisms (d_N/d_S) was

computed with the yn00 program in PAML software⁶³. Median values were used for comparison.

Genomic diversity characterization. *Genomic and genetic distances.* The 1,544,489 biallelic segregating sites were used to construct a neighbour-joining tree (Fig. 1), using the R packages `ape` and `SNPrelate`. 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 distance matrix that was obtained.

The genomic content distance (see 'Data availability') has been calculated as the number of ORF differences in the pangenome presence/absence profile (that is, the number of ORFs present in only one strain for each pairwise strain comparison) (see 'Data availability').

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 value, which represents the difference between π and θ_w .

Variscan 2.0⁶⁴ was run (`runmode = 12`, 10-kb non-overlapping windows) on multiple alignments of the concatenated chromosomes that were representative of the isolates.

Model-based ancestry. Model-based ancestry estimation was performed on the biallelic SNPs using ADMIXTURE v.1.23²¹ in unsupervised mode.

Principal component analysis. Principal components analysis on the biallelic SNPs was performed using EIGENSOFT v.6.0.1. 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. The matrix of presence/absence of ORFs in the population has been analysed using the discriminant analysis of principal components (DAPC) algorithm implemented in the R package `adegenet` 2.0.1. DAPC describes clusters by maximizing the between-cluster variance while minimizing the within-cluster variance. The number of components retained for the principal component analysis calculation was 150, accounting for > 88% of total variance. For the subsequent DAPC calculation, the alpha-score indicates 25 as the optimal number of discriminant principal components 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` package⁶⁵ was used to compute r^2 , the correlation coefficient between pairs of loci that stands as a measure of association for linkage disequilibrium. All pairs of polymorphic sites were investigated through `.map` and `.ped` files generated with `vcftools`⁶⁶, excluding SNPs with a 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 linkage disequilibrium decays to half of its maximum value.

Loss of heterozygosity. Heterozygous isolates were investigated for LOH regions with an R script generated in-house (see 'Data availability'). Regions over 50 kb with less than 10 heterozygous sites per 50 kb were considered to be under LOH (Supplementary Table 7).

***Saccharomyces* rooted tree.** To construct the tree, we used 22 *S. cerevisiae* isolates representative of the species genetic diversity that were sequenced with Oxford Nanopore technology⁶⁷. 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. These 18 other yeast strains included 7 *S. cerevisiae* strains, 5 *S. paradoxus* strains and 6 out-group strains from other *Saccharomyces* yeast species as previously described²⁴. The orthology identification was carried out using Proteinortho (v.5.15)⁶⁸ and synteny information was 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 (v.3.8.1551)⁶⁹, and the resulting protein alignment was further used to guide the corresponding CDS alignment using PAL2NAL (v.14)⁷⁰. 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 (for example, 1st, 2nd and 3rd codon positions). We used RAXML (v.8.2.6) to build the maximum likelihood tree based on the GTRGAMMA model with 100 rapid bootstraps. As an alternative, 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 (v.4.7.12)⁷¹ to produce the 'species tree'. Both the concatenated tree and the consensus tree were visualized in FigTree (v.1.4.2) (<http://tree.bio.ed.ac.uk/software/figtree/>).

Phenotyping. Quantitative high-throughput phenotyping was performed using end-point colony growth on solid medium. 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 medium conditions, including YPD 30 °C as the pinning and growth control condition (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 h 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 Gitter⁷². Each value was normalized using the growth ratio between stress media and standard YPD medium at 30 °C (see ‘Data availability’). Pairwise Pearson’s correlations of fitness trait values between replicates were calculated for each condition.

Genome-wide association studies. Mixed-model association analysis was performed using FaST-LMM v.2.07¹⁹. We used the normalized phenotypes by replacing the observed value with 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 as ‘fs’: an arbitrary threshold has been set to exclude all variants present in less than 1,000 individuals for the total matrix.

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. To quantify the extent of the bulk inflation and the excess false positive rate, we computed the genomic inflation factor, λ , for each condition (Supplementary Fig. 34). 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-degree-of-freedom χ^2 distribution. Under a null model of no association and unlinked variants, the expectation is for the λ to be 1. A λ greater than 1 indicates inflated P values of association, possibly owing to a confounding factor that has not been accounted for.

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). Using this method, variants passing this threshold have a 5% family-wise error rate (Supplementary Fig. 33).

The estimations of genome-wide heritabilities were completed by dividing the genetic variance of the null model by the total variance of the null model (genetic variance and residual variance), computed using FaST-LMM (Supplementary Fig. 6). The values reported here are 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 performed the same calculation again (Fig. 6).

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. The 1002 Yeast Genome website (<http://1002genomes.u-strasbg.fr/files/>) provides access to ‘-scripts.tar.gz’, which contains the Perl and R scripts used to (1) extract all the non-reference material from a large set of assemblies; (2) collapse ORFs that are more similar than a specific threshold; and (3) detect regions of LOH.

Data availability. All the strains are available on request except for 11 isolates, which cannot be distributed (see Supplementary Table 1).

The Illumina reads are available in the Sequence Read Archive under accession number ERP014555.

The 1002 Yeast Genome website (<http://1002genomes.u-strasbg.fr/files/>) provides access to:

-1011Matrix.gvcf.gz: all SNPs and indels called at the population level (.gvcf format).

-1011GWASMatrix.tar.gz: the matrix used for GWAS, which contains all biallelic positions known for 1,000 isolates or more with MAF > 5% as well as CNVs (encoded 0/1/2 for absence/0.5–1 copy/multiple copies) (.bed, .bim and .fam formats).

-1011DistanceMatrixBasedOnSNPs.tab.gz: for each pair of strains, the value is the percentage, based on SNPs, of non-identical bases. Heterozygous differences were half-weighted compared to the homozygous differences.

-1011DistanceMatrixBasedOnORFs.tab.gz: for each couple of strains, the value is the number of ORFs that are present in only one out of the two isolates.

-1011Assemblies.tar.gz: de novo assemblies of the 1,011 isolates (.fasta format).
-allReferenceGenesWithSNPsAndIndelsInferred.tar.gz: sequences of the genes found in the reference genome in which SNPs and indels have been automatically inferred for each isolate.

-allORFs_pangenome.fasta.gz: sequences of the 7,796 pangenomic ORFs (.fasta format).

-genesMatrix_PresenceAbsence.tab.gz: pattern of presence and/or absence of pangenomic ORFs for each isolate, in which the presence of an ORF is marked as 1 and its absence is marked as 0.

-genesMatrix_CopyNumber.tab.gz: estimated copy number for each pangenomic ORF, per isolate. Values are given for the haploid genome, so that non-integer values can be found (different copy number on homologous chromosomes).

-genesMatrix_Frameshift.tab.gz: for each isolate, the presence or absence (indicated by 1 or 0, respectively) of homozygous frameshift is reported in each gene, based on the number of bases affected by indels.

-phenoMatrix_35ConditionsNormalizedByYPD.tab.gz: growth ratio between 35 stress conditions and standard YPD medium at 30 °C, for 971 isolates.

All other data are available from the corresponding authors upon reasonable request.

- Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1**, 18 (2012).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- 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 (Austin)* **6**, 80–92 (2012).
- Simpson, J. T. et al. ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
- Otto, T. D., Dillon, G. P., Degraeve, W. S. & Berriman, M. RATT: rapid annotation transfer tool. *Nucleic Acids Res* **39**, e57 (2011).
- Proux-Wéra, E., Armisen, D., Byrne, K. P. & Wolfe, K. H. A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach. *BMC Bioinformatics* **13**, 237 (2012).
- Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
- Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
- Scannell, D. R. et al. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces sensu stricto* genus. *G3 (Bethesda)* **1**, 11–25 (2011).
- Baker, E. et al. The genome sequence of *Saccharomyces eubayanus* and the domestication of lager-brewing yeasts. *Mol. Biol. Evol.* **32**, 2818–2831 (2015).
- Hittinger, C. T. et al. Genomics and the making of yeast biodiversity. *Curr. Opin. Genet. Dev* **35**, 100–109 (2015).
- Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
- Hutter, S., Vilella, A. J. & Rozas, J. Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinformatics* **7**, 409 (2006).
- Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
- Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
- 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).
- Lechner, M. et al. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
- Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
- Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609–W612 (2006).
- Mirarab, S. et al. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–i548 (2014).
- Wagh, O. & Parts, L. Gitter: a robust and accurate method for quantification of colony sizes from plate images. *G3 (Bethesda)* **4**, 547–552 (2014).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were used to predetermine sample sizes.

2. Data exclusions

Describe any data exclusions.

- Ploidy and level of aneuploidy: only strains in their natural state were considered for these analysis. We consequently excluded strains that were genetically manipulated.
- Pangenome: The CDH and CFH strains were excluded from this analysis due to the presence of Staphylococcus epidermis contamination,
- SNPs with a minor allele frequency (MAF) lower than 5% were excluded from the linkage disequilibrium analysis, as well as from the genome-wide association studies,
- SNPs with missing genotypes in more than 11 individuals were excluded from the linkage disequilibrium analysis, as well as from the genome-wide association studies,
- Core and variable ORFs: essential genes were excluded from the dN/dS comparison (Fig S14)
- CNVs between wild and domesticated clades: only euploid diploid isolates were considered for this analysis (Fig S27). We consequently excluded aneuploid strains as well as haploid and polyploid strains.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Experimental replication was not attempted in our study.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples were not randomized for the experiments.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was not used during data collection.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The <u>exact sample size</u> (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement indicating how many times each experiment was replicated |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as an adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clearly defined error bars |

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

Abyss (v1.5.2), ADMIXTURE (v.1.23), ASTRAL (v4.7.12), BLAST package (v2.2.28+), bwa (v0.7.4-r385), condetri v2.2, EIGENSOFT (v6.0.1), FastX toolkit (v0.0.13.2), FaST-LMM (v2.07), FigTree (v1.4.2), FlowJo, GATK (v3.3-0), Maker (v2.31.8), MUSCLE (v3.8.1551), PAL2NAL (v14), PAML (v4.8), Picard-tools (v1.124), Plink (v2.0), Proteinortho (v5.15), RATT, RAxML (v8.2.6), SIFT (v5.2.2), SnpEff (v4.1), SOAP (v2.21), Variscan (v2.0), vcftools (v0.1.12), YGAP.
R packages: adegenet(v2.01), ape(v3.4), densityClust(v0.2.1), flowCore(v1.32.2), flowViz(v1.30.1), gitter(v1.1.1), SNPrelate(v1.0.1)
The custom codes can be accessed via the following links:
<http://1002genomes.u-strasbg.fr/files/>
<http://www.genoscope.cns.fr/fastxtend/>

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All the strains are available on request except for 11 isolates, which cannot be distributed.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

n/a

b. Describe the method of cell line authentication used.

n/a

c. Report whether the cell lines were tested for mycoplasma contamination.

n/a

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

n/a

► Animals and human research participantsPolicy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

n/a

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

n/a

Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

▶ Data presentation

For all flow cytometry data, confirm that:

- 1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- 2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- 3. All plots are contour plots with outliers or pseudocolor plots.
- 4. A numerical value for number of cells or percentage (with statistics) is provided.

▶ Methodological details

- | | |
|--|--|
| 5. Describe the sample preparation. | 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. |
| 6. Identify the instrument used for data collection. | FACS-Calibur flow cytometer using the HTS module for processing 96 well plates |
| 7. Describe the software used to collect and analyze the flow cytometry data. | FlowJo and R packages FlowViz, flowCore, and densityClust |
| 8. Describe the abundance of the relevant cell populations within post-sort fractions. | no sorting applied |
| 9. Describe the gating strategy used. | no gate used |

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.