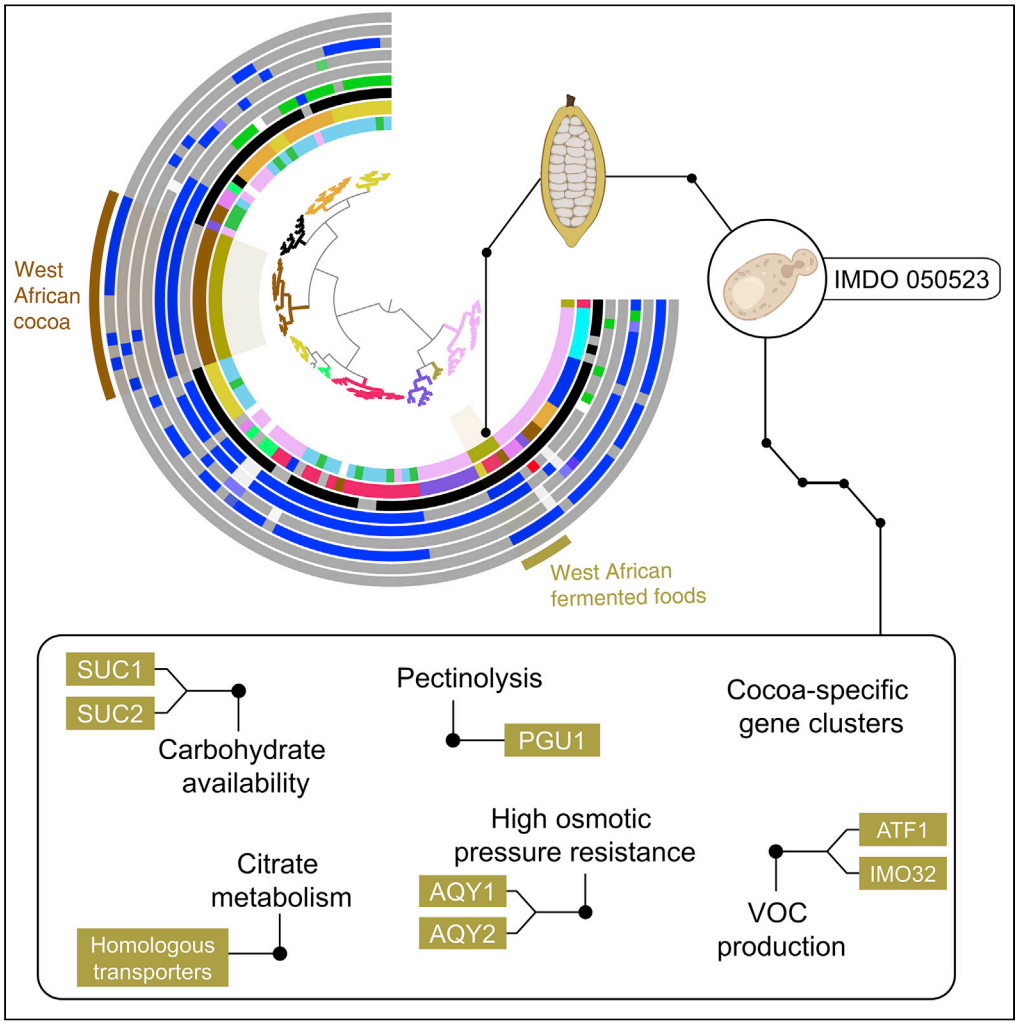
iScience

CellPress
OPEN ACCESS

Article

Phylogenomics of a *Saccharomyces cerevisiae* cocoa strain reveals adaptation to a West African fermented food population



Cristian Díaz-Muñoz, Marko Verce, Luc De Vuyst, Stefan Weckx

stefan.weckx@vub.be

Highlights

Cocoa strains clustered according to their geographical origin

The IMDO 050523 strain belonged to a West African fermented foods population

Cocoa strains contained unique genes within the *Saccharomyces cerevisiae* accessory genome

The IMDO 050523 strain showed an increased genetic fitness to cocoa fermentation

Díaz-Muñoz et al., iScience 25, 105309
November 18, 2022
https://doi.org/10.1016/j.isci.2022.105309

iScience

CellPress
OPEN ACCESS

Article

Phylogenomics of a *Saccharomyces cerevisiae* cocoa strain reveals adaptation to a West African fermented food population

Cristian Díaz-Muñoz,[1] Marko Verce,[1] Luc De Vuyst,[1] and Stefan Weckx[1,2,*]

SUMMARY

Various yeast strains have been proposed as candidate starter cultures for cocoa fermentation, especially strains of *Saccharomyces cerevisiae*. In the current study, the genome of the cocoa strain *S. cerevisiae* IMDO 050523 was unraveled based on a combination of long- and short-read sequencing. It consisted of 16 nuclear chromosomes and a mitochondrial chromosome, which were organized in 20 contigs, with only two small gaps. A phylogenomic analysis of this genome together with another 105 *S cerevisiae* genomes, among which 20 from cocoa strains showed a geographical distribution of the latter, including *S. cerevisiae* IMDO 050523. Its genome clustered together with that of a West African fermented food population, indicating a wider adaptation to West African food niches than cocoa. Furthermore, *S. cerevisiae* IMDO 050523 contained genetic signatures involved in sucrose hydrolysis, pectin degradation, osmotolerance, and conserved amino acid changes in key ester-producing enzymes that could point toward specific niche adaptations.

INTRODUCTION

Fermented foods are almost as old as human civilization (Hutkins, 2019), and so is cocoa (De Vuyst and Leroy, 2020; Ozturk and Young, 2017). However, the concept of what fermented foods and beverages are is still being redefined. The most recent definition refers to them as "foods made through desired microbial growth and enzymatic conversions of food components" (Marco et al., 2021). Moreover, fermented foods would not have been understood without the knowledge on yeasts, as these microorganisms were the first ones to be associated with food fermentation processes (Barnett, 2003; Hutkins, 2019). In particular, yeasts are responsible for alcoholic fermentation, thus capable of metabolizing the saccharides present in raw food materials to ethanol and carbon dioxide (Barnett, 2003). During fermentation, yeasts also contribute to an improvement of the organoleptic and nutritional properties of the final fermented foods and beverages (Hutkins, 2019).

The Ascomycete yeast species *Saccharomyces cerevisiae* is a main contributor to the production of many fermented foods, such as beer, wine, bread, and cocoa (Dequin and Casaregola, 2011; Lahue et al., 2020; Liti, 2015; Steensels and Verstrepen, 2014). The fact that *S. cerevisiae* is the most studied yeast species, mainly because of its model organism status in the early genomics era, contributes to the understanding of its involvement in food fermentation processes, including cocoa fermentation (De Vuyst and Leroy, 2020; Díaz-Muñoz and De Vuyst, 2022). Indeed, in the last decades, a number of genomic analyses have been performed to assess the different evolutionary traits and origins of domesticated and wild *S. cerevisiae* strains from different origins (Almeida et al., 2015; Borneman et al., 2016; Cromie et al., 2013; Duan et al., 2018; Fay et al., 2019; Fay and Benavides, 2005; Gallone et al., 2016; Gonçalves et al., 2016; Liti et al., 2009; Peter et al., 2018; Ramazzotti et al., 2019; Sicard and Legras, 2011; Strope et al., 2015). These studies have demonstrated that the ecological niches, from which *S. cerevisiae* strains have been isolated, play an important role in not only shifting their phenotype but also modifying their genomic architecture (Liti, 2015; Steensels and Verstrepen, 2014). Different kinds of genomic signatures, such as gene expansion, contraction and introgression, relative polymorphism frequency, or ploidy variation can explain the evolution and adaptation of *S. cerevisiae* to fermented food niches, as has been shown for beer, sake, and wine strains (Dequin and Casaregola, 2011; Duan et al., 2018; Fay et al., 2019; Fay and Benavides, 2005; Liti, 2015; Ramazzotti et al., 2019). In all those cases, the term domestication is used when

[1]Research Group of Industrial Microbiology and Food Biotechnology, Faculty of Sciences and Bioengineering Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium

[2]Lead contact

*Correspondence: stefan.weckx@vub.be

https://doi.org/10.1016/j.isci.2022.105309

these yeast strains optimally ferment the specific substrates of the raw food materials colonized (Liti et al., 2009; Liti, 2015). However, based on their phylogenomic analysis, some populations have been ascribed to be wild instead, as is the case for many *S. cerevisiae* strains isolated from oak tree barks in America and Japan, primeval forests in China, or rainforests in Malaysia (Duan et al., 2018; Fay et al., 2019; Fay and Benavides, 2005; Liti, 2015). The domestication traits associated with *S. cerevisiae* strains isolated from beer, wine, or bread productions have been well studied (Almeida et al., 2015; Borneman et al., 2016; Gallone et al., 2016; Gonçalves et al., 2016; Lahue et al., 2020). In contrast, strains belonging to the West African population have not been analyzed at the same level and specific genomic features from these groups have been explored only briefly (Cromie et al., 2013; Ezeronye and Legras, 2009; Fay and Benavides, 2005; Gonçalves et al., 2016; Han et al., 2021; Liti et al., 2009; Tapsoba et al., 2015). Many of these West African strains have been isolated from cocoa, although it is not always specified whether they have been isolated from fermenting cocoa pulp-bean mass, fresh cocoa pulp, or cocoa pods. To the best of the authors' knowledge, only one study has focused on the origin of cocoa isolates, albeit not as part of a whole-genome sequencing analysis but restricted to a single nucleotide polymorphism (SNP) investigation that was limited to genomic regions obtained with restriction site-associated DNA sequencing (RAD-seq; Ludlow et al., 2016).

Cocoa fermentation is a key step in the production chain of chocolate, as cocoa beans need to be cured by fermentation and drying before their roasting and further processing (De Vuyst and Leroy, 2020; Ozturk and Young, 2017; Saltini et al., 2013; Santander Muñoz et al., 2020; Schwan and Wheals, 2004). The participating yeasts perform an alcoholic fermentation in the carbohydrate- and citrate-rich cocoa pulp-bean mass and likely contribute to liquefaction of the pulp through the degradation of pectin, sucrose hydrolysis by means of an invertase, and possibly citrate conversion by means of a citrate lyase. However, it is not clear to what extent yeast metabolism is complementary to pulp pectinase and pulp invertase activities and to citrate conversion by lactic acid bacteria (De Vuyst and Leroy, 2020; Díaz-Muñoz and De Vuyst, 2022). Furthermore, yeasts are important contributors to the final cocoa flavor of the cured cocoa beans by the production of diverse volatile organic compounds (VOCs), such as higher alcohols, higher aldehydes, and esters, which diffuse from the pulp into the beans upon fermentation (De Vuyst and Leroy, 2020; Díaz-Muñoz et al., 2021; Díaz-Muñoz and De Vuyst, 2022; Ho et al., 2014; Maura et al., 2016; Meersman et al., 2016; Rodriguez-Campos et al., 2011; Schwan and Wheals, 2004). Albeit that a wide diversity of yeasts is encountered during spontaneous cocoa fermentation processes, *S. cerevisiae* is one of the yeast species most commonly found (Daniel et al., 2009; Ho et al., 2018; Meersman et al., 2013; Papalexandratou et al., 2013; Papalexandratou and De Vuyst, 2011). Also, it is widely used as part of starter culture mixtures to improve the quality of cured cocoa beans (Díaz-Muñoz et al., 2021; Ho et al., 2018; Lefeber et al., 2012; Meersman et al., 2016; Moreira et al., 2017). In particular, the Ghanaian *S. cerevisiae* IMDO 050523 strain has repeatedly been tested as member of functional starter culture mixtures and has proven to be a successful candidate starter culture strain for cocoa fermentation processes (Díaz-Muñoz et al., 2021; Lefeber et al., 2012). In general, starter culture-initiated cocoa fermentation processes inoculated with *S. cerevisiae* have demonstrated to proceed faster and, in many cases, the cured cocoa beans display a richer flavor (Díaz-Muñoz et al., 2021; Ho et al., 2018; Lefeber et al., 2012; Meersman et al., 2016; Moreira et al., 2017). However, whether yeast species, and specifically strains of *S. cerevisiae*, present in cocoa fermentations have adapted to this ecological niche is still an open question.

The present study aimed at a further understanding of the role of *S. cerevisiae* IMDO 050523 as candidate yeast starter culture strain for cocoa fermentation processes by determining its whole-genome sequence, making use of an optimized sequencing, assembling, and polishing strategy to obtain a high-quality yeast genome, and analyzing its genetic signatures of adaptation to function in a fermenting cocoa pulp-bean mass. Furthermore, its positioning in a phylogenomic tree was assessed, in comparison with 105 genomes of *S. cerevisiae* strains from different, mainly food-related, sources and geographical origins, thereby positioning tens of strains of a West-African origin.

RESULTS

Whole-genome sequencing, assembly, polishing, alignment, and annotation

Pure, high-molecular-mass DNA of the *S. cerevisiae* IMDO 050523 cocoa strain was subjected to both long-read and short-read sequencing to be able to obtain a high-quality, whole-genome sequence (Figure S1 and Table S1). An optimized genome assembly strategy to achieve a telomere-to-telomere representation of the chromosomes was followed. The final genome assembly included a manual curation to remove

contigs smaller than 800 bp, wrongly assigned regions corresponding to the template *lambda* DNA that was used during long-read sequencing, and the mitochondrial contigs generated with Canu (Koren et al., 2017), as well as to include the mitochondrial contig generated with NOVOPlasty (Dierckxsens et al., 2017). The *S. cerevisiae* IMDO 050523 genome consisted of 20 contigs, representing 16 nuclear chromosomes (chromosomes I to XVI) and a mitochondrial chromosome. A contig-per-chromosome was achieved for 14 of the 17 chromosomes. Chromosomes IX, XII, and XIII were resolved into two contigs each, leaving a gap of 2,061 and 1,991 bp in the case of chromosomes IX and XII, respectively, and showing an overlap region of 24,887 bp in the case of chromosome XIII. Haplotype phasing was performed with Purge Haplotigs (Roach et al., 2018) to obtain the final *S. cerevisiae* IMDO 050523 genome, which resulted in a final length of 12.15 Mbp, with 92.41% of complete BUSCO genes (Simão et al., 2015) and 5,265 unique genes predicted with GeneMark-ES (Ter-Hovhannisyan et al., 2008).

Analysis of ploidy and copy number variation

With an assembled, phased genome in hand, the whole-genome sequence of the *S. cerevisiae* IMDO 050523 cocoa strain was examined for ploidy estimation and degree of heterozygosity. The sequence coverage distribution over the genome, calculated using the short reads, showed a homogeneous pattern for each contig (270.05 ± 6.10), demonstrating the absence of aneuploidy in the genome of *S. cerevisiae* IMDO 050523 (Figure 1A). Nonetheless, the mitochondrial genome and the extremity of chromosome XII.b were sequenced at 20-fold and 15-fold higher sequence coverage, respectively. The latter genomic region corresponded to the location of the fungal rRNA genes, which suggested that the assembly underestimated the copy number of these genes, hence causing an increased sequence coverage. Furthermore, chromosome XII was split into two contigs at this genomic region, suggesting the impossibility of the assembler to join both extremities of the rRNA gene cluster.

The ploidy was estimated by calculating the alternative allele frequencies compared with the *S. cerevisiae* S288C reference genome, using the SNP data generated with the short reads. Overall, a biallelic distribution over the entire genome was demonstrated, suggesting that the *S. cerevisiae* IMDO 050523 genome was diploid (Figure 1B). Nevertheless, the heterozygosity levels were not homogeneous throughout the genome, because a loss of heterozygosity (LOH) of several genomic regions was found for many chromosomes. These LOH regions were especially present in the telomeric and subtelomeric regions of certain chromosomes (*e.g.,* chromosomes IV, X, and XVI), whereas other chromosomes presented large LOH regions that spanned most of the chromosome length (*e.g.,* chromosomes II, VIII, XII, and XIII).

Phylogenomic analysis

The number of genes in the 106 *S. cerevisiae* genomes examined varied from 5,164 to 6,933, embedded within a number of contigs per genome that ranged from 16 to 1,429, indicating an impact of the sequencing and assembly strategy (Figure 2). In the phylogenomic tree constructed, these genomes grouped into ten different clusters, based on the average nucleotide identity (ANI) values calculated, and, in general, were correlated with the source the concomitant strains were isolated from or with the geographical origin of these strains (Figure 3). Two Beer clusters could be distinguished, referred to as Beer 1 and Beer 2. Although the CFC strain, isolated from a traditional Belgian Trappist beer (Westmalle) clustered within the Beer 1 cluster, its genome displayed a lower ANI value compared to the other genomes of this cluster. Furthermore, a Bread cluster was found, containing genomes of strains from bakery origin and encompassing the genomes of two beer strains, all originating from Asia, Europe, and America. Genomes of strains isolated from wine productions also clustered together, thereby displaying small genetic distances among each other. Also here, genomes of strains from different geographical origin were included in this Wine cluster, which, together with their close phylogenomic relatedness, demonstrated signatures of domestication present in these strains isolated from wine productions. Furthermore, this Wine cluster contained a genome of a strain isolated from milk kefir (BFC) as well as a water kefir metagenome-assembled genome (MAG1), indicating commonalities between kefir and wine. However, the remaining strains isolated from kefirs (ARS and YJM1478) clustered with milk strains from China or were more distantly related, indicating a high heterogeneity of this fermented food ecosystem. A big Asian cluster was found, for which at least four different subclusters could be differentiated, three of which containing genomes of strains isolated from sake, rice wine/vinegar (including the Korean rice vinegar MAG2) and bread (traditional Mantou bread) productions, as well as a cocoa/Chinese forest subcluster. Regarding the latter subcluster, the genomes of two strains (ARH and ARI) isolated from cocoa in Indonesia were closely related

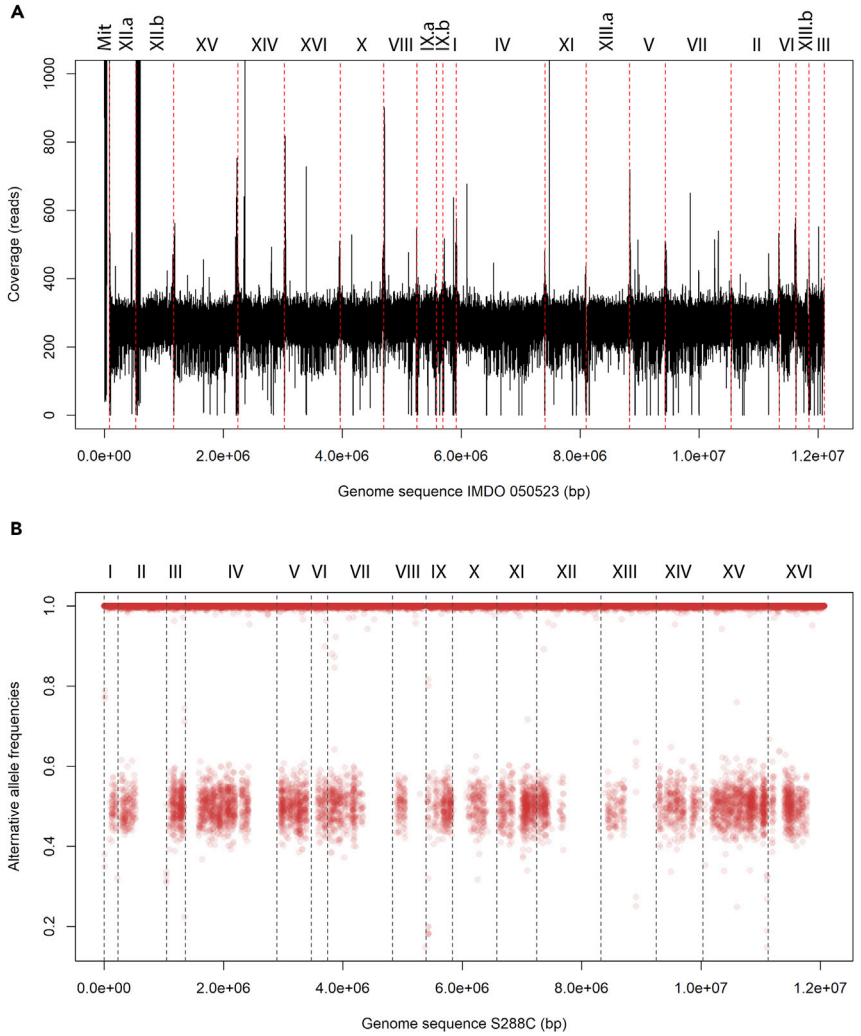


Figure 1. Read coverage and allele frequency distribution of the *Saccharomyces cerevisiae* IMDO 050523 genome

(A) Short read sequence coverage distribution for this genome, with dashed lines representing the end of the IMDO 050523 contigs; and (B) alternative allele frequency distribution of this genome compared with the *S. cerevisiae* S288C reference genome, with dashed lines representing the beginning of the S288C chromosomes.

to those of two other strains (YJM1400 and YJM1401) isolated from fruits from a relatively close geographical area (The Philippines). Most of the genomes of strains isolated from Chinese forests did cluster separately from those of the other Asian strains, reflecting a higher genetic diversity, which indicated a wild, human-independent origin of these strains (Duan et al., 2018). Closely related to this cluster, a cluster containing genomes of strains originating from West African fermented foods could be distinguished. This cluster included genomes of strains isolated from traditional palm wine and beer productions (YJM195, YJM1248, and YJM1439). The genome of the Ghanaian cocoa strain IMDO 050523, sequenced in the present study, also belonged to this West African fermented food cluster, and was hence separated from the cluster of genomes of West African cocoa strains. The genomes of strains contained within the West African cocoa cluster showed a closer relation to those of the Wine cluster than to those of the West African fermented food cluster. Finally, two genomes related to cocoa strains (including the Costa Rican cocoa MAG3) were found in the Mixed cluster, constituted by the genomes of two Mantou bread strains, one milk strain, two fruit strains, one Chinese forest strain, and the reference strain S288C. Again, the genome of the cocoa-related MAG3 (Costa Rica) and that of the cocoa strain ALG (Ecuador) were

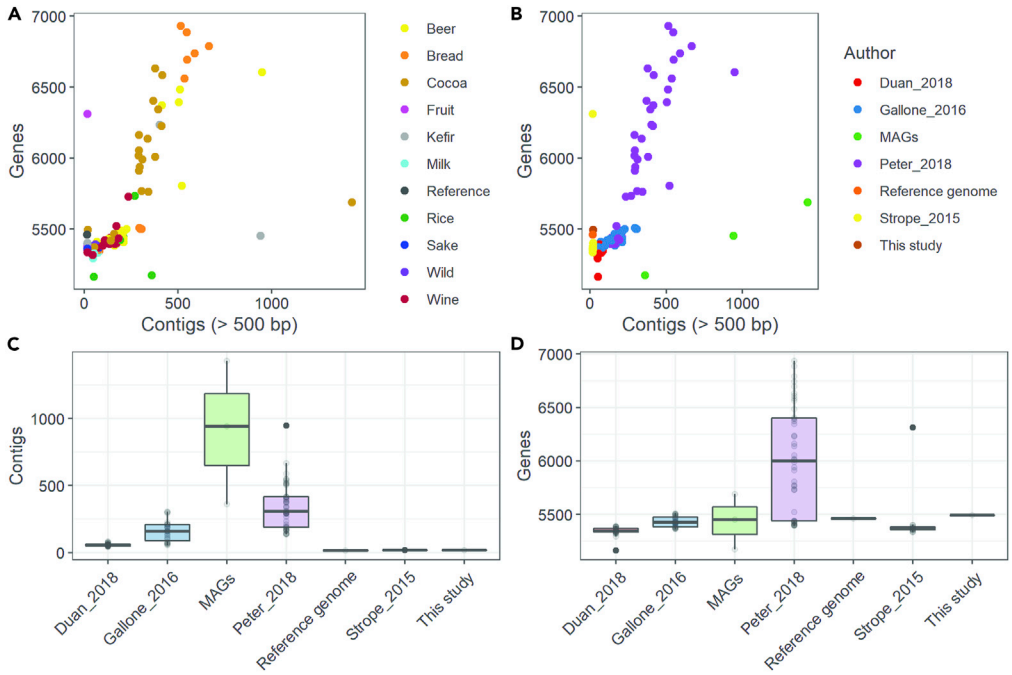


Figure 2. Gene and contig distribution across the 106 *Saccharomyces cerevisiae* genomes examined

The results are shown as a function of the isolation source of the strains (A) and the research teams involved (B, C, and D). The lower and upper hinges of the boxes correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend from the hinge to the largest value no further than 1.5 times the inter-quartile range (IQR) from the hinge. Semi-transparent dots represent the number of invertase-encoding genes each genome harbored, whereas filled dots are outlying points. MAG, metagenome-assembled genome.

closely related to the genome of a strain isolated from fruit (YJM1386) from a relatively close geographical area (Jamaica).

Furthermore, a pangenome was constructed that consisted of 5,193 gene clusters (GCs), encompassing 596,832 genes that were identified in the 106 genomes examined (Figure 4). A core genome could be distinguished, based on GC presence/absence, and consisted of 4,701 GCs (90.5% of the total GCs). The other 492 GCs (9.5%) established the accessory genome. The genome of the strain BDL (Ecuadorian beer), MAG1 (Belgian water kefir), and MAG3 (Costa Rican cocoa) grouped separately from those of the strains belonging to the same source or geographical location, which could be linked to their reduced genome completion percentage and increased number of singletons and contigs.

Population structure analysis

Because the strains isolated from cocoa ended up in different clusters of the phylogenomic analysis, a population structure analysis was performed to identify their ancestral populations and possible levels of admixture. Overall, the populations inferred were in line with the phylogenomic clusters described above and revealed a common origin for the Bread and Beer 1 strains as well as for the Wine, Beer 2, and Milk strains (Figure 5). Discriminant analysis of principal components (DAPC) identified seven distinct populations (Figure S2), among which the West African fermented food population showed the highest between-groups genetic diversity (Figure 5A). Alternatively, fastStructure identified six ancestral populations (likelihood = −0.567), as the strains belonging to the Mixed cluster identified with DAPC showed a high level of mosaicism that was linked to four of the populations inferred by fastStructure (Figure 5B). Furthermore, the Indonesian strains isolated from cocoa (ARI, ARH) and the strains isolated from fruits in The Philippines were admixed with the Chinese forest and Asian fermented food populations. Likewise, the Ecuadorian cocoa strain ALG showed evidence of admixture with the Chinese forest and Wine/Beer 2/Milk population. No evidence of admixture was found for the strains belonging to the West African fermented food population (including the IMDO 050523 strain), suggesting a separated evolutionary origin. Oppositely, the West African cocoa population contained strains with

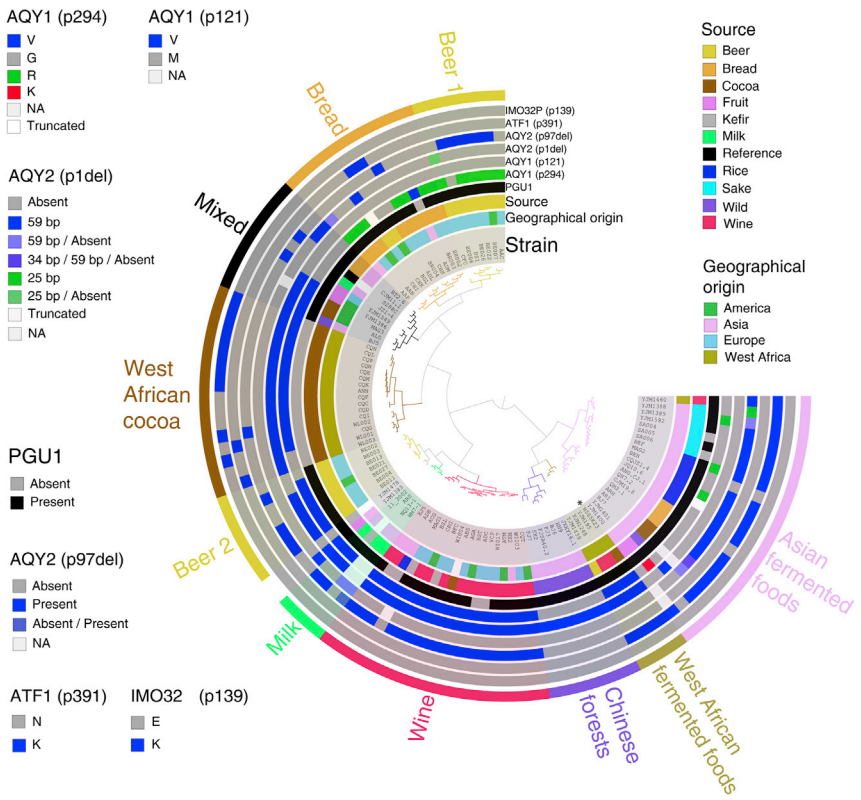


Figure 3. Unrooted phylogenomic tree of 106 *Saccharomyces cerevisiae* genomes that was generated based on the average nucleotide identity (ANI) values between their genomes

The isolation source and geographical origin of the strains are indicated with colors. The absence/presence of polygalacturonase (PGU1) and polymorphisms in the aquaporins (AQY1 and AQY2) and acetyltransferases (ATF1 and IMO32P) amino acid sequences are also colour-coded. Colored branches and bins represent phylogenomic clusters, identified by their geographical origin and/or isolation source. The *S. cerevisiae* IMDO 050523 cocoa strain, indicated as H5S5K23 accompanied with a black dot, belongs to the West African fermented food cluster, together with two wine strains and a beer strain from the same geographical origin. NA, not applicable.

signs of admixture with the Wine/Beer 2/Milk population (CQP, CQD, and CQC), the Chinese forest population (CQI and WL001), and with both populations (WL002, CQG, and WL003).

In silico functional analysis

Manual inspection of the GCs present in the accessory genome revealed 22 GCs that were more abundant in the genomes of the cocoa strains, often including the *S. cerevisiae* IMDO 050523 strain (Table S2). Using the information retrieved from their homology results in the HHPred (Zimmermann et al., 2018) and SUPERFAMILY 2.0 (Pandurangan et al., 2019) databases and the conserved domain database (CDD; Lu et al., 2020), these GCs could be classified according to their putative biological function (Figure S3). They were related to metabolism (*e.g.*, hydrolases and aminotransferases), transport (*e.g.*, transmembrane proteins and ABC transporters), structure (*e.g.*, flocculins), or regulation (*e.g.*, fungal transcription factors). Furthermore, four GCs (GC_00005075, GC_00005059, GC_00004937, and GC_00004865) were exclusive for the *S. cerevisiae* IMDO 050523 genome, although an overall low similarity for these GCs was found in the databases mentioned above.

Finally, metabolic traits of interest regarding the functional role of *S. cerevisiae* in cocoa fermentation processes were examined, in particular volatile organic compound (VOC) production pathways, pectinolysis, invertase activity, citrate metabolism, and high-osmolarity adaptation. Regarding VOC production, all 106 *S. cerevisiae* genomes possessed the necessary genes to produce the main VOCs that are commonly found in cocoa fermentation processes (Figure 6). Also, conserved amino acid changes within the *S. cerevisiae* phylogenomic clusters of key ester-producing enzymes could be found (Figure 3). A change from

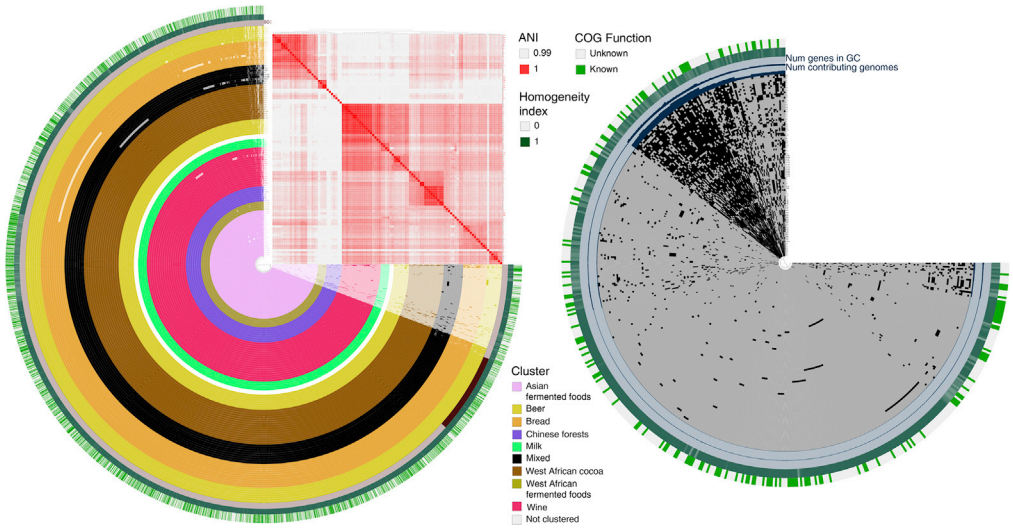


Figure 4. *Saccharomyces cerevisiae* pangenome (left) and accessory genome (right), based on the 106 genomes examined

The genomes constituting the pangenome are colored according to the phylogenomic cluster they belong to and ordered based on the average nucleotide identity (ANI) values. The order of the gene clusters (GCs; represented as radii) across the pangenome is based on their occurrence in the genomes. Bars in the inner 106 layers represent presence (full) or absence (empty) of GCs in each of the genomes. Single-copy gene clusters (SCGCs) are represented with a dark red bar in the last but two layers (pangenome), whereas the number of genes in the GCs and the contributing genomes to the GCs are represented with dark blue bars (accessory genome). The GCs that were more abundant in *S. cerevisiae* cocoa strains are detailed in Figure S3 and Table S2. The homogeneity index (an indication of the amino acid sequence similarity between all genomes) and the cluster of orthologous group (COG) functions are shown in the two outer layers for both the pangenome and the accessory genome.

asparagine to lysine in position 391 of the alcohol O-acetyltransferase (ATF1) was shown for strains belonging to the Asian fermented food and West African fermented food clusters. A change from glutamate to lysine in position 139 of the IMO32P enzyme, an alcohol acetyltransferase, was shown for several strains belonging to the West African cocoa cluster, but not in those of strains of West African fermented foods or in those of cocoa strains from other clusters. The esterase YMR210w of three strains from the Bread cluster contained a deletion of 35 amino acids, and a change from alanine to threonine in position 358 of this esterase was mainly shown for rice and sake strains from the Asian fermented food cluster.

With regard to the presence of genes encoding pectin-degrading enzymes, the amino acid sequences of 296 polygalacturonases, 245 pectate lyases, and 197 pectin esterases were used to search for homologous sequences in the genome of *S. cerevisiae* IMDO 050523. Among the cluster representative sequences, 30 significant alignments were found, of which 29 could be aligned to the same locus of chromosome X of the *S. cerevisiae* IMDO 050523 genome, containing a gene that was annotated as polygalacturonase *PGU1* (EC 3.2.1.15) (Table S3). This gene was present in only 27.3% of the cocoa strains, whereas it was present in 91.7% of the non-cocoa strains. Only one amino acid sequence, corresponding to a pectin esterase, was aligned to a different locus, also located on chromosome X. In particular, this amino acid sequence belonged to *Aspergillus terreus* and has been described as a putative feruloyl esterase (*FaeA*; EC 3.1.1.73), which can hydrolyze feruloyl-galactose ester bonds in pectin and helps pectinases to break down the plant cell wall. However, the corresponding region of the *S. cerevisiae* IMDO 050523 genome was annotated as a putative lipase (*LIH1*; EC 3.1.1.3), leaving the possibility open for a role in common lipid metabolism or pectin degradation. A comparative analysis of the *PGU1* gene revealed the presence of at least one intron (as predicted by AUGUSTUS; Stanke et al., 2006) in many of the genomes examined. An average of 436 $\pm$ 77 introns was predicted, accounting for 7.7% of the average total number of genes (5,644), which was in line with the expected values for the *S. cerevisiae* genome. However, canonical splicing sites and essential splicing motifs were not found in the predicted intron sequence of the *PGU1* gene of the IMDO 050523 genome.

A higher number of invertase genes was found in the genomes of strains belonging to the Beer 1, Beer 2, Bread, and West African fermented food clusters, but not in those of the West African cocoa cluster

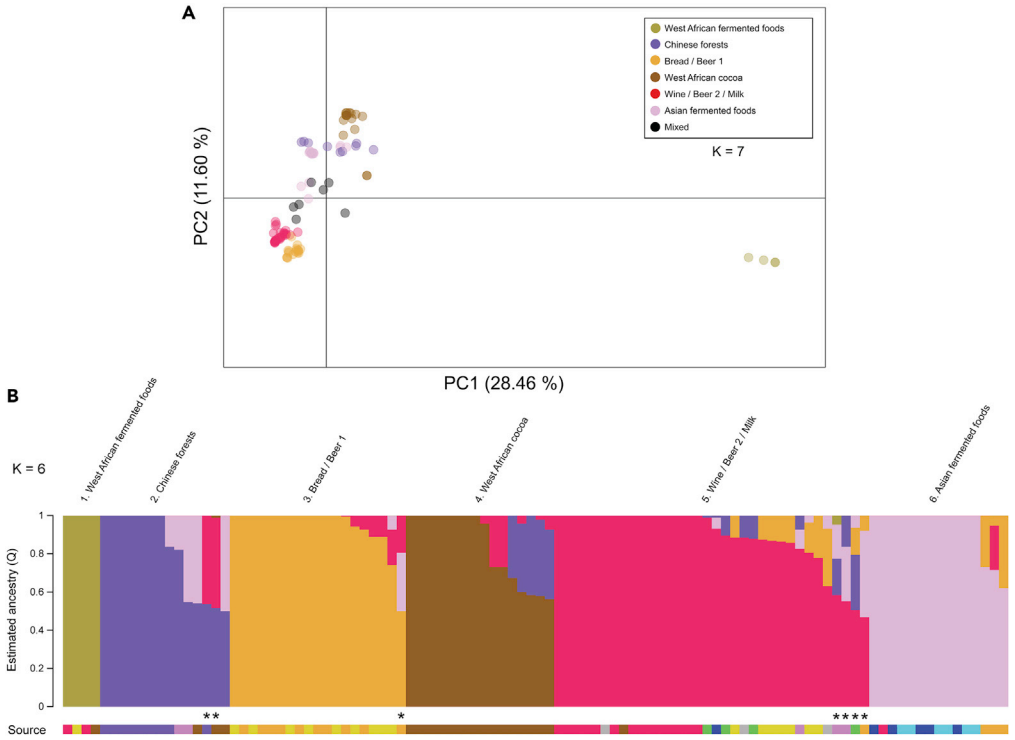


Figure 5. Population structure analysis based on 43,875 biallelic single nucleotide polymorphisms (SNPs) from 102 genomes of *S. cerevisiae* strains

(A) Scatterplot of a discriminant analysis of principal components (DAPC) assuming *K* = 7 populations and five retained PCs (Figure S2).

(B) Population structure plot for *K* = 6 assumed ancestral populations. Strains belonging to the Mixed cluster in the phylogenomic tree of Figure 3 are indicated with an asterisk below the corresponding bars in the structure plot. The color codes used for the source the strains have been isolated from are as presented in Figure 3. The names given to the populations are linked to the phylogenomic clusters represented in Figure 3.

(Figure 7). In particular, *S. cerevisiae* IMDO 050523 possessed three genes coding for invertases (annotated as *SUC1/SUC2*) located in three different genomic loci. The presence of other putative invertases was checked using the amino acid sequences of 245 invertases, 60 sucrases, 15 saccharases, and 346 alpha-glucosidases (Table S4). Among the cluster representative sequences, 55 significant alignments to loci different from those of the *SUC1/SUC2* genes were found (Table S4). However, most of these chromosomal positions were annotated as alpha-glucosidases, related to the degradation of maltose (*e.g., IMA1, IMA2, IMA5, MAL32,* and *MAL33*), glycoproteins (*e.g., ROT2* and *GTB1*), or glucans (*e.g., GDB1, EXG1, EXG2,* and *BGL2*). The only amino acid sequence annotated as invertase that showed a significant alignment to a position in chromosome XVI, was an ABC1 family protein with unknown function (*YPL109C*).

A gene screening as to the presence of putative enzymes related to the consumption or further degradation of extracellular citrate, besides those involved in the tricarboxylic acid (TCA) cycle, was performed by using the amino acid sequences of 431 citrate lyases and 356 citrate synthases. Only cluster representative amino acid sequences that did not align to the *CIT1*, *CIT2*, and *CIT3* genes encoding citrate synthases/citrate-oxaloacetate lyases of the TCA cycle, were considered (Table S5). As such, 42 significant alignments were found, of which 21 contained motifs similar to those of a wide variety of kinases (Table S5). The highest similarities were obtained with chromosomal regions annotated as *PDH1* (2-methylcitrate dehydratase) and *ACO2* (homocitrate dehydratase), two enzymes involved in the degradation of TCA citrate derivatives. Three alignments were found to loci of chromosomes X, XI, and XIV without any functional annotation. Nevertheless, these loci aligned with the human *AKT1* (serine/threonine kinase) gene, indicating that these non-annotated homologous regions may correspond to kinase motifs too instead of citrate lyases. Next to CTP1 (tricarboxylate transport protein), two homologous transporter proteins needed for the secretion of

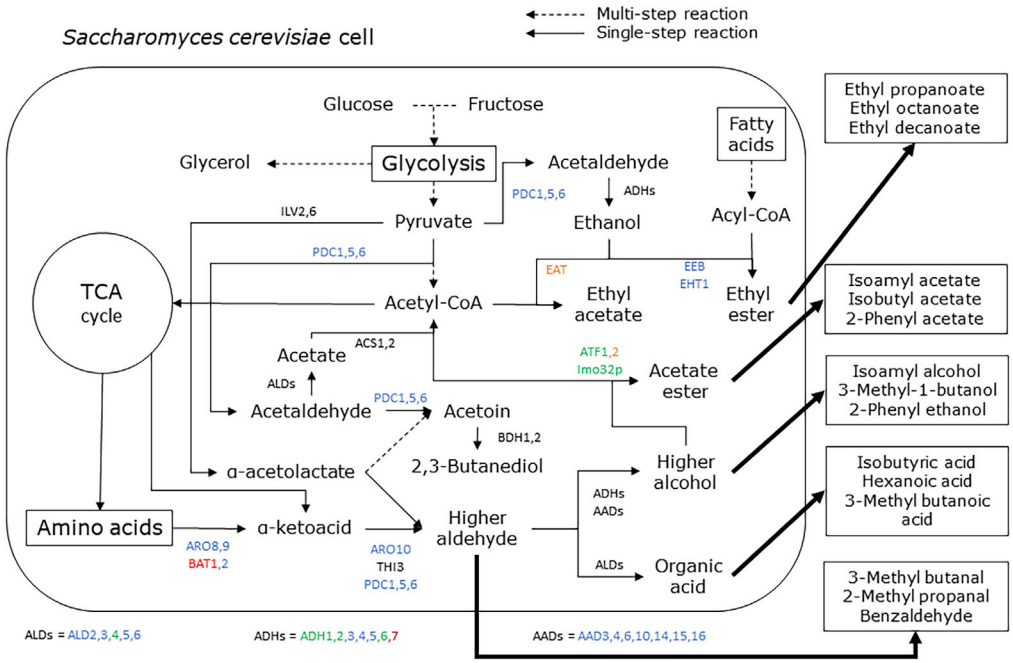


Figure 6. Metabolic reconstruction of volatile organic compound (VOC) production, based on genomic information comprised in the 106 *Saccharomyces cerevisiae* genomes examined

The genes needed for each step are indicated in color according to the number of genomes that harbor them (dark red if less than 103 genomes, red for 103, orange for 104, green for 105, and blue for 106 genomes).

citrate were found in the *S. cerevisiae* IMDO 050523 genome, namely the florfenicol exporter and the siderochrome-iron transporter (sequence similarities of 46–52% and sequence coverage of 76–97%). The florfenicol exporter (UniProtKB: GAA89875.1) aligned to regions of the *S. cerevisiae* IMDO 050523 genome and was annotated as *QDR1* and *QDR2* [multidrug resistance transporter (quinidine)] located on chromosome IX and *AQR1* (probable transporter of short-chain monocarboxylic acids) located on chromosome XIV. The siderochrome-iron transporter (UniProtKB: GAA92527.1) showed homology to two genes located on chromosomes V and VI that were also annotated in the *S. cerevisiae* IMDO 050523 genome as siderochrome-iron transporters (*SIT1*).

Different variants were found for the AQY1 protein (aquaporin) among the 106 *S. cerevisiae* genomes examined (Figure 3). First, a premature stop codon known to be present in the genomes of beer strains was found for several strains of the Beer 1 and Bread clusters, resulting in an arginine at position 294 of the amino acid sequence (Ale A882 insertion; Gonçalves et al., 2016). The A881 deletion (Will et al., 2010) was found in all strains of the West African cocoa cluster, with the exception of WL001 and WL002, and the West African fermented food cluster, leading to a protein of 327 amino acids and a valine at position 294. The cocoa strains isolated in Latin America (MAG3 and ALG) also possessed this polymorphism. On the contrary, one of the Indonesian cocoa strains (ARH) did not contain any of these *AQY1* polymorphisms, whereas the ARI strain lacked the whole gene. Furthermore, the fruit strain YJM1401 isolated in The Philippines possessed still another polymorphism affecting the amino acid position 294 (lysine), leading to a protein of 328 amino acids. Another commonly occurring variant of the protein occurred too, as also glycine was found at position 294, leading to a final protein of 305 amino acids (V121M polymorphism; Gonçalves et al., 2016). Of all cocoa strains, this variant was only found in the ARH strain. A polymorphism in the *AQY1* involving a thymine deletion in position 498 of the gene sequence (Barbosa et al., 2018) was not found in any of the genomes examined. Finally, previously reported deletions in the *AQY2* gene (aquaporin) of wine, beer, and Asian strains were found in the strains included in the current study. The first deletion (Asian G25 deletion; Will et al., 2010) occurred at the start of the amino acid sequence (Asian fermented food strains) and the second one (11-bp deletion; Will et al., 2010) occurred at position 97 (Wine, Beer 1, and Beer 2 strains). None of these variants were found in the cocoa strains examined.

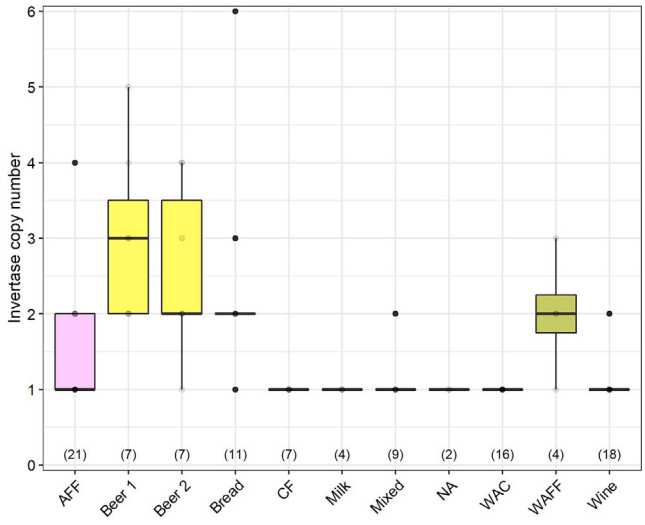


Figure 7. Number of invertases found in the 106 *Saccharomyces cerevisiae* genomes examined, grouped according to the phylogenomic cluster they belonged to

The lower and upper hinges of the boxes correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend from the hinge to the largest value no further than 1.5 times the inter-quartile range (IQR) from the hinge. Semi-transparent dots represent the number of invertase-encoding genes each genome harbored, whereas filled dots are outlying points. AFF, Asian fermented foods; CF, Chinese forests; NA, not clustered; WAC, West African cocoa; WAFF, West African fermented foods.

DISCUSSION

Thousands of *S. cerevisiae* genomes are currently available, harboring differences in quality, completeness, ploidy representation, and heterozygosity (Almeida et al., 2015; Duan et al., 2018; Gallone et al., 2016; Gonçalves et al., 2016; Libkind et al., 2021; Liti et al., 2009; Liti, 2015; Peter et al., 2018; Ramazzotti et al., 2019; Shen et al., 2018; Strope et al., 2015). Choosing appropriate sequence data processing tools to achieve a high-quality yeast genome is not always straightforward and the approach followed influences any downstream analysis. To elucidate the whole-genome sequence of the cocoa strain *S. cerevisiae* IMDO 050523, originally isolated from a Ghanaian cocoa heap fermentation process (Camu et al., 2007; Daniel et al., 2009), the present study made use of an optimized bioinformatics pipeline, combining Canu to assemble the long reads, Purge Haplotigs to perform haplotype phasing, and Racon and Medaka as polishing tools, to obtain a highly accurate consensus sequence. Moreover, Pilon was used for short read-based polishing, improving the genome assembly in terms of completeness, redundancy, indels, and misassemblies.

The *S. cerevisiae* IMDO 050523 genome sequenced in the present study consisted of 16 nuclear chromosomes and a mitochondrial chromosome, representing a length of 12.15 Mbp and harboring 5,265 genes. The differences found in the total number of genes in the 106 *S. cerevisiae* genomes examined during a phylogenomic analysis reflected a substantial influence of the sequencing and assembly strategy used. For example, some research teams sequenced haploid genomes (Strope et al., 2015), whereas others sequenced all genomes with their natural ploidy (Gallone et al., 2016; Peter et al., 2018), pointing to the importance of reporting haploid genomes for a correct interpretation of gene content and functional analysis. Haplotype phasing was successfully achieved in the present study and, hence, the genome assembly of the sequenced IMDO 050523 cocoa strain corresponded to its haploid genome. However, assessment of the natural ploidy of the sequenced strain revealed that it was a diploid genome, with a high percentage of LOH regions. The latter indicates a low level of outcrossing, comparable with the genomes of wine and sake strains reported before and hence representing a wide genetic diversity (Peter et al., 2018).

The phylogenomic analysis of the *S. cerevisiae* IMDO 050523 genome together with another 105 *S. cerevisiae* genomes, covering 101 strains of different origins (Latin America, Asia, Europe, and West Africa), one reference genome (laboratory strain *S. cerevisiae* S288C), and three MAGs (from Belgian water kefir, Korean rice vinegar, and Costa Rican cocoa) showed clustering patterns that were correlated with the

geographical origin or isolation source of the concomitant fermented food strains. The data presented supported those of previous studies that distinguished several genome clusters among *S. cerevisiae* strains according to their geographical origin (*e.g.,* West Africa, Asia, and North America) or isolation source (*e.g.,* wine, beer, and sake) (Almeida et al., 2015; Barbosa et al., 2016; Borneman et al., 2016; Cromie et al., 2013; Duan et al., 2018; Fay et al., 2019; Fay and Benavides, 2005; Gallone et al., 2016; Gonçalves et al., 2016; Liti et al., 2009; Peter et al., 2018; Pontes et al., 2020; Ramazzotti et al., 2019; Strope et al., 2015). A clustering based on the isolation source of the strains, regardless of the geographical origin, could be ascribed to domesticated populations, as the ones of wine, bread, beer, sake, and rice wine/vinegar strains. Many of these fermented foods and beverages are inoculated either through backslopping or after addition of a microbial starter culture (De Vuyst et al., 2021; De Vuyst and Leroy, 2020; Leroy and De Vuyst, 2004). Backslopping likely contributes to the adaptation of strains to a specific ecological niche of a fermented food matrix. However, cocoa fermentation is still a spontaneous process, meaning that the initial inoculation of the cocoa pulp-bean mass comes from microorganisms present in the close environment, *in casu* the cocoa pod surfaces, banana or plantain leaves, fermentation equipment (*e.g.,* baskets and boxes), insects, etc. (De Vuyst and Leroy, 2020; De Vuyst and Weckx, 2016; Maura et al., 2016). This feature could explain the geographical clustering of the genomes of the cocoa strains examined, including *S. cerevisiae* IMDO 050523, and their closeness to genomes of fruit strains or strains from traditional spontaneous food fermentation processes, which could be ascribed as wild (Gallone et al., 2016). Furthermore, the genomes of the cocoa strains belonged to at least three different populations, scattered by their geographical origin. In addition, it has been shown before that the genomes of cocoa and coffee strains are the result of admixtures from *S. cerevisiae* populations from a geographical proximity (Ludlow et al., 2016). Similarly, a relatedness has been shown for genomes of African and Southeast Asian strains (Cromie et al., 2013; Han et al., 2021) and West African and secondary Chinese forest strains (Duan et al., 2018). The wildness of the cocoa strains was also in line with the findings of GCs with structural functions (*e.g.,* flocculins), which were more abundant in the genomes of this group of strains. These genes are indeed typically contracted or lost in domesticated strains (Duan et al., 2018).

In contrast to the overall geographical clustering of the cocoa strain genomes, that of the Ghanaian *S. cerevisiae* IMDO 050523 cocoa strain belonged to a cluster of genomes of strains from West African fermented foods, whose phylogenomic position was located far from the genomes of 16 other West African cocoa strains. Furthermore, the presence of two differentiated West African populations was demonstrated, with no evidence of admixture or common evolutionary history. A differential clustering of cocoa strain genomes from those of other West African non-cocoa strains could already be seen in the phylogenies reported in previous studies (Cromie et al., 2013; Peter et al., 2018; Pontes et al., 2020; Tapsoba et al., 2015). A possible explanation for the differential clustering pattern of the IMDO 0505023 cocoa strain must be sought in the environment, fermentation practices, and specific source from which the cocoa strains examined were isolated (cocoa pulp-bean mass, fermentation baskets and boxes, surfaces of banana leaves, harvesting tools, etc.). The genetic relatedness between the West African cocoa population and the Wine/Beer 2/Milk population and the mosaicism found in some of these strains may indicate a common evolutionary relationship, which could shed light onto the African origin of vineyard strains, as suggested before (Barbosa et al., 2016; Ezeronye and Legras, 2009; Fay and Benavides, 2005; Han et al., 2021; Peter et al., 2018). The West African population described in several studies was barely associated with a specific fermentation process or it subdivided into different clades (Barbosa et al., 2016; Gallone et al., 2016; Gonçalves et al., 2016; Liti et al., 2009; Strope et al., 2015). Later, the West African population was separated in two clades, according to respective cocoa and palm wine fermentation processes (Peter et al., 2018). Indeed, the West African palm wine population has already been described as a domesticated population that originated from a specific population of *S. cerevisiae,* triggered by geographic and/or ecological isolation (Ezeronye and Legras, 2009). Finally, a much more source-specific clustering has been found, suggesting that the West African population could have been domesticated to the local, spontaneous fermentation processes (Han et al., 2021; Pontes et al., 2020).

The three MAGs included in the phylogenomic analysis clustered together with the genomes of strains isolated from the same or similar sources, indicating that retrieving MAGs from complex metagenomes could be a good strategy to represent "unique" genomes, albeit that their quality, expressed as completeness and redundancy, was in general lower compared to genomes obtained from isolates, due to the specific data analysis steps performed (Meziti et al., 2021).

The *S. cerevisiae* pangenome built up during the present study consisted of a large number of genes that were common to all genomes examined (core genome), whereas only a minority was made up of variable genes (accessory genome), as has been shown before for a pangenome based on 1,011 *S. cerevisiae* genomes (Peter et al., 2018; Richard, 2020). For the latter pangenome, consisting of 7,796 open reading frames (ORFs), 4,940 ORFs are part of the core genome, whereas 2,856 ORFs are variable (Peter et al., 2018). The proportion of variable ORFs is, however, lower in the S288C reference genome (1,144 ORFs of a total of 6,081 ORFs; Richard, 2020). The lower proportion of variable GCs obtained in the present study was mainly due to the fact that GCs may contain different copies of the same gene or genes encoding similar proteins. As it has been shown that a higher gene copy number variability is present in those variable ORFs, rather than in ORFs contained within the *S. cerevisiae* core genome (Peter et al., 2018), it was logic to find a lower proportion of variable GCs than ORFs. However, the gene clustering performed during the present study may grant a better representation of the functional differences between the genomes examined and the actual size of the *S. cerevisiae* core and accessory genomes obtained, as it did not consider differences in gene copy number variation.

The manual inspection of the accessory genome (variable GCs) led to unravel specific GCs that were more abundant in the genomes of the *S. cerevisiae* cocoa strains, such as flocculins, transmembrane proteins, ABC transporters, fungal transcription factors, and a range of enzymes related to yeast metabolism. These findings corroborated previous results that showed that, in general, the *S. cerevisiae* variable ORFs are enriched with genes related to cell wall and membrane components, cell-cell interactions, and secondary metabolism (Richard, 2020).

The present study showed that *S. cerevisiae* IMDO 050523 harbored genes that reflect specific adaptations to cocoa fermentation. Although the microbial species present in the last stages of cocoa fermentation differ from those at the initial ones, some *S. cerevisiae* strains can be tolerant to late fermentation stage conditions (high temperature and high ethanol and acetic acid concentrations) and thus resist the entire cocoa fermentation course. For instance, *S. cerevisiae* IMDO 050523 seems to be adapted to the carbohydrate-rich cocoa pulp at the start of the fermentation, which could explain its prevalence. Indeed, it is known that a loss of function of the aquaporin genes (*AQY1* and *AQY2*) provides an increased resistance to high-osmolarity environments (Will et al., 2010; Gonçalves et al., 2016; Pontes et al., 2020). This adaptative loss of function has been reported in wine strains (including a West African one) too, which provides *S. cerevisiae* strains with an increased fitness toward the carbohydrate-rich must environment (Gonçalves et al., 2016). For the *AQY1* gene, an adenine deletion (coding sequence position 881, protein position 294) is responsible for its inactivity in *S. cerevisiae* wine strains (Gonçalves et al., 2016). The inactivation of the *AQY1* gene shown for strains from the West African cocoa and West African fermented food clusters suggested adaptation of *S. cerevisiae* IMDO 050523 to carbohydrate-rich niches.

Investigation of the metabolic pathway for the production of VOCs by *S. cerevisiae* showed that there was almost no strain variability regarding the presence of key flavour-producing enzymes. Overall, these findings suggested that the production of VOCs in the cocoa pulp-bean mass will not primarily depend on the presence of different *S. cerevisiae* strains but rather on the biochemical composition of the cocoa pulp (providing flavor precursors) or gene expression levels. Still, genomic comparison between strains of species of different yeast genera naturally occurring during cocoa fermentation processes, such as *Hanseniaspora* and *Pichia*, should be performed, as previous studies have demonstrated differential VOC productions according to the yeast species inoculated (Díaz-Muñoz et al., 2021; Moreira et al., 2021). However, the single amino acid changes found in the acetyltransferases ATF1 and IMO32P, affecting strains from the West African fermented food cluster and West African cocoa cluster, respectively, may result in a different activity of those ester-synthesizing enzymes and, thus, may have an impact on the VOC production by *S. cerevisiae* cocoa strains. To prove this hypothesis, the activity of these enzymes should be tested throughout the cocoa fermentation course.

Furthermore, yeasts have been postulated to contribute to pectin degradation at the initial stage of cocoa fermentation processes (Schwan and Wheals, 2004; De Vuyst and Leroy, 2020; Díaz-Muñoz and De Vuyst, 2022). A thorough screening of more than 700 pectin-degrading enzymes showed the presence of the polygalacturonase gene, *PGU1*, as the only gene encoding a pectin-degrading enzyme in *S. cerevisiae* IMDO 050523. It has been shown before that a polygalacturonase gene is the main gene responsible for pectin

degradation by cocoa strains of *S. cerevisiae* (Meersman et al., 2017), although the current study showed that several cocoa strains lacked this gene, in turn pointing indirectly toward the role of the pulp pectinase. In contrast, this gene was present in most of the *S. cerevisiae* strains from other sources. The presence of at least one intron in the *PGU1* gene in many of the strains examined was surprising, as less than 10% of the genes in *S. cerevisiae* are thought to contain introns (7.7% in the dataset examined). Consequently, a molecular approach should be performed to clarify whether these introns play a role in the activity of the enzyme, as alternative splicing to respond to environmental conditions, although rare, can occur in *S. cerevisiae* (Juneau et al., 2009; Hossain et al., 2011; Pleiss et al., 2007; Skelly et al., 2009). Nevertheless, the lack of conserved splicing sites and motifs typically present in *S. cerevisiae* (Langford et al., 1984) may indicate an inaccuracy of the gene prediction tool rather than a true intron.

Also, an increased number of invertase-encoding genes found in *S. cerevisiae* IMDO 050523 could suggest an increased sucrose hydrolysis activity, as happens with an increased number of *MAL* genes for maltose degradation in bread, beer, or sake strains (Gallone et al., 2016; Lahue et al., 2020), and with several traits characterizing wine production processes (Jeffares et al., 2017). As yeast invertase activity is thought to propitiate a fast conversion of sucrose into fermentable saccharides, it can speed up cocoa fermentation when considerable amounts of sucrose remain upon opening of (unripe) cocoa pods at the start of the fermentation process (as a result of an incomplete hydrolysis of sucrose into glucose and fructose by cocoa pulp invertase activity). However, a metatranscriptomic analysis of a Costa Rican cocoa fermentation process has shown a low expression of *S. cerevisiae* invertase genes, which appeared only late into the fermentation process (Verce et al., 2021). In line with these previous findings, the copy numbers of the invertase genes found in the West African cocoa population were lower than in other populations and could indicate a reduced invertase activity.

A possible active role of yeasts in citrate assimilation or conversion during cocoa fermentation processes has been postulated before (Daniel et al., 2009; Jespersen et al., 2005; Schwan and Wheals, 2004; Thompson et al., 2007). However, *in vitro* experiments have shown that not many cocoa yeast isolates are able to assimilate citrate. Nevertheless, it is well known that citrate is a central intermediate of the *S. cerevisiae* oxidative metabolism through the TCA cycle. Indeed, this species can generate citrate through the mitochondrial (CIT1 and CIT3; EC 2.3.3.1) or peroxisomal (CIT2; EC 2.3.3.16) citrate synthases, to be further converted into isocitrate through aconitate hydratase (ACO1; EC 4.2.1.3), which is then in turn used in the TCA cycle and/or glyoxylate cycle. This oxidative pathway is downregulated in the presence of glucose, as this carbon source shifts the *S. cerevisiae* oxidative metabolism toward a high fermentative metabolism (Casal et al., 2008). Therefore, the contribution of *S. cerevisiae* IMDO 050523 to the assimilation of citrate during the initial stages of cocoa fermentation processes (in which citrate is typically intensively consumed by LAB in particular) seems improbable. However, the same pathway is upregulated in the absence of glucose, allowing *S. cerevisiae* to convert the ethanol produced into acetyl-CoA and to incorporate the latter into the TCA cycle, thereby producing excess citrate (Casal et al., 2008; Odoni et al., 2019). In the present study, other putative citrate lyases and synthases were screened as to their occurrence in the genome of *S. cerevisiae* IMDO 050523, with no evidence of existing enzymes to provide additional pathways to consume or further degrade citrate. Yet, two transporter proteins, a florfenicol exporter and a siderochrome-iron transporter, homologous to transporters involved in citrate secretion in *Aspergillus luchuensis*, were present in the genome of *S. cerevisiae* IMDO 050523. These transporters have been proposed to actively participate in the secretion of citrate by *S. cerevisiae* when heterologously expressed (Odoni et al., 2019). As a small increase of citrate concentrations at the end of cocoa fermentation processes is sometimes found, characterized by aerobic conditions and remaining appropriate carbon sources such as ethanol and acetate (Batista et al., 2015; Camu et al., 2007; Díaz-Muñoz et al., 2021; Ho et al., 2018; Moreira et al., 2017; Papalexandratou et al., 2011), it may be ascribed to the appearance of disadvantageous filamentous fungi as well as to the action of still active *S. cerevisiae* strains.

Thus, the thorough functional study of the *S. cerevisiae* pangenome built up during the present study led to unravel specific signatures of possible niche adaptations in cocoa strains. Finally, the fact that the Ghanaian *S. cerevisiae* IMDO 050523 cocoa strain fitted into the West African fermented food population and not into the West African cocoa population may be linked to its broader physiological capabilities. Yet, it still harbored one polygalacturonase and several invertase genes, likely enabling its degradation of cocoa pulp pectin and hydrolysis of cocoa pulp sucrose, respectively. Overall, these findings contributed to a

better understanding of the functional roles of *S. cerevisiae* in cocoa fermentation processes and at the same time could explain the success of strains of this species when used as part of functional starter cultures.

LIMITATIONS OF THE STUDY

Although one of the goals of this research was to position the industrially relevant cocoa starter culture strain *S. cerevisiae* IMDO 050523 within the main populations of *S. cerevisiae* reported before, the focus was on the comparative genomic analysis performed to understand the adaptations of the sequenced strain to cocoa fermentation. As a consequence, only strains for which the genome assemblies were publicly available were considered. A balance between selecting a sufficient number of strains to ensure a wide representation of other fermented foods than cocoa and keeping the strains isolated from cocoa as the main focus was attempted.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - DNA extraction, DNA sequencing and quality assessment, and whole-genome assembly, polishing and data processing
  - *Saccharomyces cerevisiae* pangenome, accessory genome, and phylogenomic tree inference
  - Population structure analysis
  - *In silico* functional analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105309.

ACKNOWLEDGMENTS

This work was supported by the University Council of the Vrije Universiteit Brussel (SRP7, IRP2, IOF2442, and IOF3017 projects), the Hercules Foundation (projects UABR09004 and UAB13002), and the Research Foundation Flanders (FWO-Vlaanderen, REVICO project S004617N). Part of this work relied on the resources and services provided by the Flemish Supercomputer Center (VSC), funded by FWO and the Flemish Government. Delphine Sicard and Hugo Devillers are acknowledged to host CD-M in their facilities to assist for ploidy and heterozygosity determinations.

AUTHOR CONTRIBUTIONS

C.D-M. and M.V. performed the DNA extraction, long-read sequencing, and genome assembly. CD-M. performed the genome annotation, phylogenomic analysis, and *in silico* functional analysis. L.D.V. and S.W. designed and supervised the work. L.D.V. was responsible for funding. CD-M. drafted the manuscript. L.D.V. and S.W. revised the manuscript. All authors read, revised, edited, and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 9, 2022
Revised: July 22, 2022
Accepted: October 3, 2022
Published: November 18, 2022

REFERENCES

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J.L., Serra, M., Dequin, S., Couloux, A., et al. (2015). A population genomics insight into the Mediterranean origins of wine yeast domestication. Mol. Ecol. *24*, 5412–5427.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Barbosa, R., Almeida, P., Safar, S.V.B., Santos, R.O., Morais, P.B., Nielly-Thibault, L., Leducq, J.-B., Landry, C.R., Gonçalves, P., Rosa, C.A., and Sampaio, J.P. (2016). Evidence of natural hybridization in Brazilian wild lineages of *Saccharomyces cerevisiae*. Genome Biol. Evol. *8*, 317–329.

Barbosa, R., Pontes, A., Santos, R.O., Montandon, G.G., De Ponzzes-Gomes, C.M., Morais, P.B., Gonçalves, P., Rosa, C.A., and Sampaio, J.P. (2018). Multiple rounds of artificial selection promote microbe secondary domestication - the case of cachaça yeasts. Genome Biol. Evol. *10*, 1939–1955.

Barnett, J.A. (2003). Beginnings of microbiology and biochemistry: the contribution of yeast research. Microbiology *149*, 557–567.

Batista, N.N., Ramos, C.L., Ribeiro, D.D., Pinheiro, A.C.M., and Schwan, R.F. (2015). Dynamic behavior of *Saccharomyces cerevisiae, Pichia kluyveri* and *Hanseniaspora uvarum* during spontaneous and inoculated cocoa fermentations and their effect on sensory characteristics of chocolate. LWT - Food Sci. Technol. *63*, 221–227.

Bigey, F., Segond, D., Friedrich, A., Guezenec, S., Bourgais, A., Huyghe, L., Agier, N., Nidelet, T., and Sicard, D. (2021). Evidence for two main domestication trajectories in *Saccharomyces cerevisiae* linked to distinct bread-making processes. Curr. Biol. *31*, 722–732.e5.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Borneman, A.R., Forgan, A.H., Kolouchova, R., Fraser, J.A., and Schmidt, S.A. (2016). Whole genome comparison reveals high levels of inbreeding and strain redundancy across the spectrum of commercial wine strains of *Saccharomyces cerevisiae*. G3 *6*, 957–971.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60.

Camu, N., De Winter, T., Verbrugghe, K., Cleenwerck, I., Vandamme, P., Takrama, J.S., Vancanneyt, M., and De Vuyst, L. (2007). Dynamics and biodiversity of populations of lactic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. Appl. Environ. Microbiol. *73*, 1809–1824.

Casal, M., Paiva, S., Queirós, O., and Soares-Silva, I. (2008). Transport of carboxylic acids in yeasts. FEMS Microbiol. Rev. *32*, 974–994.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7.

Cherry, J.M. (2015). The *Saccharomyces* genome database: a tool for discovery. Cold Spring Harb. Protoc. *2015*. pdb.top083840.

Cromie, G.A., Hyma, K.E., Ludlow, C.L., Garmendia-Torres, C., Gilbert, T.L., May, P., Huang, A.A., Dudley, A.M., and Fay, J.C. (2013). Genomic sequence diversity and population structure of *Saccharomyces cerevisiae* assessed by RAD-seq. G3 *3*, 2163–2171.

Daniel, H.-M., Vrancken, G., Takrama, J.F., Camu, N., De Vos, P., and De Vuyst, L. (2009). Yeast diversity of Ghanaian cocoa bean heap fermentations. FEMS Yeast Res. *9*, 774–783.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics *34*, 2666–2669.

Dequin, S., and Casaregola, S. (2011). The genomes of fermentative *Saccharomyces*. C. R. Biol. *334*, 687–693.

Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ *6*, e4320.

De Vuyst, L., Comasio, A., and Kerrebroeck, S.V. (2021). Sourdough production: fermentation strategies, microbial ecology, and use of non-flour ingredients. Crit. Rev. Food Sci. Nutr. 1–33. https://doi.org/10.1080/10408398.2021.1976100.

De Vuyst, L., and Leroy, F. (2020). Functional role of yeasts, lactic acid bacteria and acetic acid bacteria in cocoa fermentation processes. FEMS Microbiol. Rev. *44*, 432–453.

De Vuyst, L., and Weckx, S. (2016). The cocoa bean fermentation process: from ecosystem analysis to starter culture development. J. Appl. Microbiol. *121*, 5–17.

Díaz-Muñoz, C., and De Vuyst, L. (2022). Functional yeast starter cultures for cocoa fermentation. J. Appl. Microbiol. *133*, 39–66.

Díaz-Muñoz, C., Van de Voorde, D., Comasio, A., Verce, M., Hernandez, C.E., Weckx, S., and De Vuyst, L. (2021). Curing of cocoa beans: fine-scale monitoring of the starter cultures applied and metabolomics of the fermentation and drying steps. Front. Microbiol. *11*, 616875.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. Nucleic Acids Res. *45*, e18.

Duan, S.F., Han, P.J., Wang, Q.M., Liu, W.Q., Shi, J.Y., Li, K., Zhang, X.L., and Bai, F.Y. (2018). The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. Nat. Commun. *9*, 2690.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Emms, D.M., and Kelly, S. (2018). OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. *20*, 238.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ *3*, e1319.

Ezeronye, O.U., and Legras, J.L. (2009). Genetic analysis of *Saccharomyces cerevisiae* strains isolated from palm wine in eastern Nigeria. Comparison with other African strains. J. Appl. Microbiol. *106*, 1569–1578.

Fay, J.C., and Benavides, J.A. (2005). Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. PLoS Genet. *1*, 66–71.

Fay, J.C., Liu, P., Ong, G.T., Dunham, M.J., Cromie, G.A., Jeffery, E.W., Ludlow, C.L., and Dudley, A.M. (2019). A polyploid admixed origin of beer yeasts derived from European and Asian wine populations. PLoS Biol. *17*, e3000147.

Gallone, B., Steensels, J., Prahl, T., Soriaga, L., Saels, V., Herrera-Malaver, B., Merlevede, A., Roncoroni, M., Voordeckers, K., Miraglia, L., et al. (2016). Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell *166*, 1397–1410.e16.

Gonçalves, M., Pontes, A., Almeida, P., Barbosa, R., Serra, M., Libkind, D., Hutzler, M., Gonçalves, P., and Sampaio, J.P. (2016). Distinct domestication trajectories in top-fermenting beer yeasts and wine yeasts. Curr. Biol. *26*, 2750–2761.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics *29*, 1072–1075.

Han, D.Y., Han, P.J., Rumbold, K., Koricha, A.D., Duan, S.F., Song, L., Shi, J.-Y., Li, K., Wang, Q.-M., and Bai, F.Y. (2021). Adaptive gene content and allele distribution variations in the wild and domesticated populations of *Saccharomyces cerevisiae*. Front. Microbiol. *12*, 631250.

Holt, C., and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinf. *12*, 491.

Hossain, M.A., Rodriguez, C.M., and Johnson, T.L. (2011). Key features of the two-intron *Saccharomyces cerevisiae* gene *SUS1* contribute to its alternative splicing. Nucleic Acids Res. *39*, 8612–8627.

Ho, V.T.T., Fleet, G.H., and Zhao, J. (2018). Unravelling the contribution of lactic acid bacteria and acetic acid bacteria to cocoa fermentation using inoculated organisms. Int. J. Food Microbiol. *279*, 43–56.

Ho, V.T.T., Zhao, J., and Fleet, G. (2014). Yeasts are essential for cocoa bean fermentation. Int. J. Food Microbiol. *174*, 72–87.

Hutkins, R.W. (2019). Microbiology and Technology of Fermented Foods, 2nd edition (Wiley-Blackwell Publishing).

Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. *8*, 14061.

Jespersen, L., Nielsen, D.S., Hønholt, S., and Jakobsen, M. (2005). Occurrence and diversity of yeasts involved in fermentation of West African cocoa beans. FEMS Yeast Res. *5*, 441–453.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics *24*, 1403–1405.

Juneau, K., Nislow, C., and Davis, R.W. (2009). Alternative splicing of *PTC7* in *Saccharomyces cerevisiae* determines protein localization. Genetics *183*, 185–194.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. *27*, 722–736.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. Genome Biol. *5*, 12.

Lahue, C., Madden, A.A., Dunn, R.R., and Smukowski Heil, C. (2020). History and domestication of *Saccharomyces cerevisiae* in bread baking. Front. Genet. *11*, 584718.

Langford, C.J., Klinz, F.J., Donath, C., and Gallwitz, D. (1984). Point mutations identify the conserved, intron-contained TACTAAC box as an essential splicing signal sequence in yeast. Cell *36*, 645–653.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lefeber, T., Papalexandratou, Z., Gobert, W., Camu, N., and De Vuyst, L. (2012). On-farm implementation of a starter culture for improved cocoa bean fermentation and its influence on the flavour of chocolates produced thereof. Food Microbiol. *30*, 379–392.

Leroy, F., and De Vuyst, L. (2004). Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends Food Sci. Technol. *15*, 67–78.

Libkind, D., Peris, D., Cubillos, F.A., Steenwyk, J.L., Opulente, D.A., Langdon, Q.K., Rokas, A., and Hittinger, C.T. (2021). Into the wild: new yeast genomes from natural environments and new tools for their analysis. FEMS Yeast Res. *20*, foaa008.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). 1000 genome project data processing subgroup. The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liti, G. (2015). The natural history of model organisms: the fascinating and secret wild life of the budding yeast *S. cerevisiae*. Elife *2015*, e05835.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A., Koufopanou, V., et al. (2009). Population genomics of domestic and wild yeasts. Nature *458*, 337–341.

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Marchler, G.H., Song, J.S., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. *48*, D265–D268.

Ludlow, C.L., Cromie, G.A., Garmendia-Torres, C., Sirr, A., Hays, M., Field, C., Jeffery, E.W., Fay, J.C., and Dudley, A.M. (2016). Independent origins of yeast associated with coffee and cacao fermentation. Curr. Biol. *26*, 965–971.

Marco, M.L., Sanders, M.E., Gänzle, M., Arrieta, M.C., Cotter, P.D., De Vuyst, L., Hill, C., Holzapfel, W., Lebeer, S., Merenstein, D., et al. (2021). The international scientific association for probiotics and prebiotics (ISAPP) consensus statement on fermented foods. Nat. Rev. Gastroenterol. Hepatol. *18*, 196–208.

Maura, Y.F., Balzarini, T., Borges, P.C., Evrard, P., De Vuyst, L., and Daniel, H.-M. (2016). The environmental and intrinsic yeast diversity of Cuban cocoa bean heap fermentations. Int. J. Food Microbiol. *233*, 34–43.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Meersman, E., Steensels, J., Mathawan, M., Wittocx, P.-J., Saels, V., Struyf, N., Bernaert, H., Vrancken, G., and Verstrepen, K.J. (2013). Detailed analysis of the microbial population in Malaysian spontaneous cocoa pulp fermentations reveals a core and variable microbiota. PLoS One *8*, e81559.

Meersman, E., Steensels, J., Struyf, N., Paulus, T., Saels, V., Mathawan, M., Allegaert, L., Vrancken, G., and Verstrepen, K.J. (2016). Tuning chocolate flavor through development of thermotolerant *Saccharomyces cerevisiae* starter cultures with increased acetate ester production. Appl. Environ. Microbiol. *82*, 732–746.

Meersman, E., Struyf, N., Kyomugasho, C., Jamsazzadeh Kermani, Z., Santiago, J.S., Baert, E., Hemdane, S., Vrancken, G., Verstrepen, K.J., Courtin, C.M., et al. (2017). Characterization and degradation of pectic polysaccharides in cocoa pulp. J. Agric. Food Chem. *65*, 9726–9734.

Meziti, A., Rodriguez-R, L.M., Hatt, J.K., Peña-Gonzalez, A., Levy, K., and Konstantinidis, K.T. (2021). The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. Appl. Environ. Microbiol. *87*, 025933-20.

Moreira, I., Costa, J., Vilela, L., Lima, N., Santos, C., and Schwan, R. (2021). Influence of *S. cerevisiae* and *P. kluyveri* as starters on chocolate flavour. J. Sci. Food Agric. *101*, 4409–4419.

Moreira, I.M.V., Vilela, L.F., Miguel, M.C.P., Santos, C., Lima, N., and Schwan, R.F. (2017). Impact of a microbial cocktail used as a starter culture on cocoa fermentation and chocolate flavor. Molecules *22*, 766.

Odoni, D.I., Vazquez-Vilar, M., Van Gaal, M.P., Schonewille, T., Martins Dos Santos, V.A.P., Tamayo-Ramos, J.A., Suarez-Diez, M., and Schaap, P.J. (2019). *Aspergillus niger* citrate exporter revealed by comparison of two alternative citrate producing conditions. FEMS Microbiol. Lett. *366*, fnz071.

Ozturk, G., and Young, G.M. (2017). Food evolution: the impact of society and science on the fermentation of cocoa beans. Compr. Rev. Food Sci. Food Saf. *16*, 431–455.

Pandurangan, A.P., Stahlhacke, J., Oates, M.E., Smithers, B., and Gough, J. (2019). The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. Nucleic Acids Res. *47*, D490–D494.

Papalexandratou, Z., and De Vuyst, L. (2011). Assessment of the yeast species composition of cocoa bean fermentations in different cocoa-producing regions using denaturing gradient gel electrophoresis. FEMS Yeast Res. *11*, 564–574.

Papalexandratou, Z., Lefeber, T., Bahrim, B., Lee, O.S., Daniel, H.-M., and De Vuyst, L. (2013). *Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum*, and *Acetobacter pasteurianus* predominate during well-performed Malaysian cocoa bean box fermentations, underlining the importance of these microbial species for a successful cocoa bean fermentation process. Food Microbiol. *35*, 73–85.

Papalexandratou, Z., Vrancken, G., De Bruyne, K., Vandamme, P., and De Vuyst, L. (2011). Spontaneous organic cocoa bean box fermentations in Brazil are characterized by a restricted species diversity of lactic acid bacteria and acetic acid bacteria. Food Microbiol. *28*, 1326–1338.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics *26*, 419–420.

Pearson, W.R. (2013). An introduction to sequence similarity ("homology") searches. Curr. Protoc. Bioinformatics *42*. 3.1.1.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., et al. (2018). Genome evolution across 1, 011 *Saccharomyces cerevisiae* isolates. Nature *556*, 339–344.

Pleiss, J.A., Whitworth, G.B., Bergkessel, M., and Guthrie, C. (2007). Rapid, transcript-specific

changes in splicing in response to environmental stress. Mol. Cell *27*, 928–937.

Pontes, A., Hutzler, M., Brito, P.H., and Sampaio, J.P. (2020). Revisiting the taxonomic synonyms and populations of *Saccharomyces cerevisiae*—phylogeny, phenotypes, ecology and domestication. Microorganisms *8*, E903.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490.

Pritchard, L., Glover, R.H., Humphris, S., Elphinstone, J.G., and Toth, I.K. (2016). Genomics and taxonomy in diagnostics for food security, soft-rotting enterobacterial plant pathogens. Anal. Methods *8*, 12–24.

Raj, A., Stephens, M., and Pritchard, J.K. (2014). FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics *197*, 573–589.

Ramasamy, R.K., Ramasamy, S., Bindroo, B.B., and Naik, V.G. (2014). Structure PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus *3*, 431.

Ramazzotti, M., Stefanini, I., Di Paola, M., De Filippo, C., Rizzetto, L., Berná, L., Dapporto, L., Rivero, D., Tocci, N., Weil, T., et al. (2019). Population genomics reveals evolution and variation of *Saccharomyces cerevisiae* in the human and insects gut. Environ. Microbiol. *21*, 50–71.

R Core Team (2019). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Richard, G.F. (2020). Eukaryotic pangenomes. In The Pangenome: Diversity, Dynamics and Evolution of Genomes, H. Tettelin and D. Medini, eds. (Springer).

Roach, M.J., Schmidt, S.A., and Borneman, A.R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinf. *19*, 460.

Rodriguez-Campos, J., Escalona-Buendía, H., Orozco-Avila, I., Lugo-Cervantes, E., and Jaramillo-Flores, M.E. (2011). Dynamics of volatile and non-volatile compounds in cocoa (*Theobroma cacao* L.) during fermentation and drying processes using principal components analysis. Food Res. Int. *44*, 250–258.

Saltini, R., Akkerman, R., and Frosch, S. (2013). Optimizing chocolate production through traceability: a review of the influence of farming practices on cocoa bean quality. Food Control *29*, 167–187.

Santander Muñoz, M., Rodríguez Cortina, J., Vaillant, F.E., and Escobar Parra, S. (2020). An overview of the physical and biochemical transformation of cocoa seeds to beans and to chocolate: flavor formation. Crit. Rev. Food Sci. Nutr. *60*, 1593–1613.

Saubin, M., Devillers, H., Proust, L., Brier, C., Grondin, C., Pradal, M., Legras, J.L., and Neuvéglise, C. (2020). Investigation of genetic relationships between *Hanseniaspora* species found in grape musts revealed interspecific hybrids with dynamic genome structures. Front. Microbiol. *10*, 2960.

Schwan, R.F., and Wheals, A.E. (2004). The microbiology of cocoa fermentation and its role in chocolate quality. Crit. Rev. Food Sci. Nutr. *44*, 205–221.

Shen, X.X., Opulente, D.A., Kominek, J., Zhou, X., Steenwyk, J.L., Buh, K.V., Haase, M.A.B., Wisecaver, J.H., Wang, M., Doering, D.T., et al. (2018). Tempo and mode of genome evolution in the budding yeast subphylum. Cell *175*, 1533–1545.e20.

Sicard, D., and Legras, J.L. (2011). Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex. C. R. Biol. *334*, 229–236.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212.

Skelly, D.A., Ronald, J., Connelly, C.F., and Akey, J.M. (2009). Population genomics of intron splicing in 38 *Saccharomyces cerevisiae* genome sequences. Genome Biol. Evol. *1*, 466–478.

Stanke, M., Schöffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinf. *7*, 62.

Steensels, J., and Verstrepen, K.J. (2014). Taming wild yeast: potential of conventional and nonconventional yeasts in industrial fermentations. Annu. Rev. Microbiol. *68*, 61–80.

Strope, P.K., Skelly, D.A., Kozmin, S.G., Mahadevan, G., Stone, E.A., Magwene, P.M., Dietrich, F.S., and McCusker, J.H. (2015). The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. Genome Res. *25*, 762–774.

Tapsoba, F., Legras, J.L., Savadogo, A., Dequin, S., and Traore, A.S. (2015). Diversity of *Saccharomyces cerevisiae* strains isolated from *Borassus akeassii* palm wines from Burkina Faso in comparison to other African beverages. Int. J. Food Microbiol. *211*, 128–133.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. *18*, 1979–1990.

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.

Thompson, S.S., Miller, K.B., and Lopez, A.S. (2007). Cocoa and coffee. In Food Microbiology: Fundamentals and Frontiers, 2nd ed., M.P. Doyle, L.R. Beuchat, and T.J. Montville, eds. (ASM Press), pp. 837–849.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. *27*, 737–746.

Verce, M., De Vuyst, L., and Weckx, S. (2019). Shotgun metagenomics of a water kefir fermentation ecosystem reveals a novel *Oenococcus* species. Front. Microbiol. *10*, 479.

Verce, M., Schoonejans, J., Hernandez Aguirre, C., Molina-Bravo, R., De Vuyst, L., and Weckx, S. (2021). A combined metagenomics and metatranscriptomics approach to unravel Costa Rican cocoa box fermentation processes reveals yet unreported microbial species and functionalities. Front. Microbiol. *12*, 641185.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., and Earl, A.M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*, e112963.

Will, J.L., Kim, H.S., Clarke, J., Painter, J.C., Fay, J.C., and Gasch, A.P. (2010). Incipient balancing selection through adaptive loss of aquaporins in natural *Saccharomyces cerevisiae* populations. PLoS Genet. *6*, e1000893.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. *430*, 2237–2243.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Newly sequenced *Saccharomyces cerevisiae* IMDO 050523 cocoa strain	This paper	European Nucleotide Archive Project ID PRJEB46227 Assembly accession GCA_932563355.1 Long reads SRA accession ERR10089878 Short reads SRA accession ERR10089777
Saccharomyces cerevisiae MAG obtained from a water kefir metagenome (MAG1 – MAG_SC_WK15)	This paper	European Nucleotide Archive Project ID PRJEB55444 BioSample ID SAMEA110668653 Assembly accession GCA_946466885
Saccharomyces cerevisiae MAG obtained from a rice vinegar metagenome (MAG2 – MAG_SC_RV18)	This paper	European Nucleotide Archive Project ID PRJEB55444 BioSample ID SAMEA110691117 Assembly accession GCA_946466795
Saccharomyces cerevisiae MAG obtained from a cocoa pulp metagenome (MAG3 – MAG_SC_CP16)	This paper	European Nucleotide Archive Project ID PRJEB55444 BioSample ID SAMEA110668654 Assembly accession GCA_946462225
Experimental models: Microorganisms/strains		
S. cerevisiae strains	N/A	Table S6
Chemicals, peptides, and recombinant proteins		
Yeast extract	Oxoid	Cat# LP0021
Peptone	Oxoid	Cat# LP0037
Glucose	Merck	Cat# 1083375000
Nuclease-free water	VWR Chemicals	Cat# E476
Critical commercial assays		
Qiagen Genomic tip 20/G kit	Qiagen	Cat# 10223
Qubit™ dsDNA HS Assay kit	Thermo Fisher Scientific	Cat# Q32851
SQK-LSK109 Ligation Sequencing kit	Oxford Nanopore Technologies	Cat# SQK-LSK109
Software and algorithms		
Albacore v2.3.3	Oxford Nanopore Technologies	N/A
Guppy v2.3.7	Oxford Nanopore Technologies	https://community.nanoporetech.com/downloads
NanoPack	DeCoster et al., 2018	https://github.com/wdecoster/nanopack
FastQC v0.11.3	Babraham Institute	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Trimmomatic v0.36	Bolger et al., 2014	https://github.com/usadellab/Trimmomatic
Canu v1.8	Koren et al., 2017	https://github.com/marbl/canu
Purge Haplotigs	Roach et al., 2018	https://bitbucket.org/mroachawri/purge_haplotigs/src/master/
Racon v1.3.2	Vaser et al., 2017	https://github.com/isovic/racon
Medaka v0.7.1	Oxford Nanopore Technologies	https://github.com/nanoporetech/medaka
minimap2 v2.17	Li, 2018	https://lh3.github.io/minimap2/
Pilon v1.23	Walker et al., 2014	https://github.com/broadinstitute/pilon

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bowtie2 v2.3.5.1	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
NOVOPlasty v3.8.3	Dierckxsens et al., 2017	https://github.com/ndierckx/NOVOPlasty
QUAST	Gurevich et al., 2013	http://quast.sourceforge.net/quast
GeneMark-ES	Ter-Hovhannisyan et al., 2008	http://exon.gatech.edu/GeneMark/
BUSCO	Simão et al., 2015	https://busco.ezlab.org/
MUMmer v3.23	Kurtz et al., 2004	http://mummer.sourceforge.net/
SAMtools v1.4	Li et al., 2009	https://github.com/samtools/samtools
GATK	McKenna et al., 2010	https://gatk.broadinstitute.org/hc/en-us
MAKER2	Holt and Yandell, 2011	http://www.yandell-lab.org/software/maker.html
AUGUSTUS v3.3.3	Stanke et al., 2006	http://augustus.gobics.de/
blastp v2.10	Altschul et al., 1997	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
InterProScan v5.36	Zdobnov and Apweiler, 2001	https://www.ebi.ac.uk/interpro/about/interproscan/
anvi'o v6.2	Eren et al., 2015	https://anvio.org/
pyANI v0.2.10	Pritchard et al., 2016	https://pypi.org/project/pyani/
FastTree v2.1.11	Price et al., 2010	http://www.microbesonline.org/fasttree/
R	R Core Team, 2019	https://cran.r-project.org/
PLINK v1.90	Chang et al., 2015	https://www.cog-genomics.org/plink2/
fastSTRUCTURE v1.0	Raj et al., 2014	https://github.com/rajanil/fastStructure
Structure Plot v2.0	Ramasamy et al., 2014	http://omicsspeaks.com/strplot2/
adegenet v2.1.5	Jombart, 2008	http://adegenet.r-forge.r-project.org/
pegas v1.1	Paradis, 2010	https://cran.r-project.org/web/packages/pegas/
OrthoFinder v2.3.8	Emms and Kelly, 2018	http://www.stevekellylab.com/software/orthofinder
DIAMOND	Buchfink et al., 2015	https://github.com/bbuchfink/diamond
MUSCLE v3.8.31	Edgar, 2004	https://www.ebi.ac.uk/Tools/msa/muscle/
SeaView v5.0.4	Gouy et al., 2010	http://doua.prabi.fr/software/seaview
tblastn v2.10	Altschul et al., 1997	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download

RESOURCE AVAILABILITY

Lead contact

DNA sequence information is publicly available from the NCBI and EBI databases. Further information and requests for resources should be directed to and will be fulfilled by the lead contact: Stefan Weckx (stefan.weckx@vub.be).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The genome assembly and annotation data generated in the present study have been submitted to the European Nucleotide Archive of the European Bioinformatics Institute (ENA/EBI; https://www.ebi.ac.uk/ena/home) under accession numberPRJEB46227.

- This paper does not report original codes.
- Any additional information about the analysis in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The yeast strain *S. cerevisiae* IMDO 050523 (also referred to as *S. cerevisiae* H5S5K23), which was originally isolated from a spontaneous cocoa fermentation process carried out in Ghana (Camu et al., 2007; Daniel et al., 2009), was used throughout this study for sequencing and functional analysis of its genome.

For a phylogenomic analysis, the genome of *S. cerevisiae* IMDO 050523 was compared with the genomes of 101 *S. cerevisiae* strains, originating from (non)-fermented food products (Strope et al., 2015; Gallone et al., 2016; Duan et al., 2018; Peter et al., 2018; Table S6). These genomes were retrieved from the Genome database of the National Center for Biotechnology Information (NCBI, Bethesda, Maryland, USA). This selection of genomes was made to maximize the number of strains originating from cocoa (22 in total), while ensuring a wide representation of other fermented foods. Further, three MAGs corresponding with strains of *S. cerevisiae* from Belgian water kefir (MAG1; Verce et al., 2019), Korean rice vinegar (MAG2; L. Vermote, L. DeVuyst and S. Weckx, unpublished results), and Costa Rican cocoa (MAG3; Verce et al., 2021) fermentation processes were included in the present study. Finally, the reference genome of the laboratory baker's yeast strain *S. cerevisiae* S288C, retrieved from the *Saccharomyces* genome database (SGD; Cherry, 2015), was included.

METHOD DETAILS

DNA extraction, DNA sequencing and quality assessment, and whole-genome assembly, polishing and data processing

DNA extraction

Saccharomyces cerevisiae IMDO 050523 was grown in yeast extract-peptone-glucose medium that contained 10 g/L of yeast extract (Oxoid, Basingstoke, Hampshire, UK), 20 g/L of peptone (Oxoid), and 20 g/L of glucose (Avantor, Radnor, Pennsylvania, USA). High-molecular-mass DNA was extracted and purified using a Genomic tip 20/G kit following the manufacturer's instructions (Qiagen, Hilden, Germany). The genomic DNA was finally collected with a glass rod to avoid shearing and was dissolved in 250 µL of nuclease-free water (Avantor) by incubation in a water bath at 55°C for 2 h. The DNA integrity was checked by agarose gel electrophoresis, the purity was assessed with a NanoDrop ND-2000 spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), and the concentration was measured using a Qubit fluorometer (Thermo Fisher Scientific).

DNA sequencing

The genomic DNA was subjected to both long-read and short-read sequencing. Long-read sequencing was performed using the Oxford Nanopore Technologies' MinION sequencer with a r9.4.1 FLO-MIN106 flow cell (Oxford Nanopore Technologies, Oxford, UK). One µg of high-molecular-mass DNA was used as starting material for library preparation, using the SQK-LSK109 Ligation Sequencing kit (Oxford Nanopore Technologies), according to the manufacturer's instructions. Pipette tips were end-cut to avoid extra shearing of the DNA during this step. Short-read paired-end sequencing was performed using Illumina's NovaSeq platform (Illumina, San Diego, California, USA) in the university's core facility (BRIGHTcore, Jette, Belgium).

The MinION sequencing signal was simultaneously basecalled in MinKNOW2.2 using Albacore (v2.3.3; Oxford Nanopore Technologies). The FAST5 files were then classified in "fail" or "pass" reads according to their qscore. To improve the basecalling accuracy, the raw "pass" FAST5 reads were basecalled using Guppy v2.3.7 in high-accuracy GPU-accelerated mode, using "dna_r9.4.1_450bps_flipflop.cfg" as configuration file (Oxford Nanopore Technologies). NanoPack (DeCoster et al., 2018) was used as the set of tools for visualization and quality-processing of the basecalled reads.

Long reads were quality-checked using NanoPlot and NanoQC, concatenated in a single fastq file, and subsequently quality-filtered and trimmed with NanoFilt, using the following parameters: quality, 12; headcrop, 35; and tailcrop, 25. Short reads were quality-checked with FastQC (v0.11.3; http://www.

bioinformatics.babraham.ac.uk/projects/fastqc/). They were quality-filtered and trimmed using Trimmomatic (v0.36; Bolger et al., 2014) with the following parameters: headcrop, 10; leading, 30; trailing, 30; slidingwindow, 4:15; and minlen, 20.

Whole-genome assembly and polishing

Quality-filtered and trimmed long reads were further corrected and assembled into contigs using Canu (v1.8; Koren et al., 2017), setting the expected genome size to 12 Mbp and using default parameters for the Nanopore raw input data. This raw assembly was then haplotype-phased with Purge Haplotigs (Roach et al., 2018), and then subjected to an optimized polishing strategy. Unless stated otherwise, all polishing and aligning tools were used with their default settings. Briefly, four iterations of Racon (v1.3.2; Vaser et al., 2017) and one iteration of Medaka (v0.7.1; Oxford Nanopore Technologies) were performed to obtain a more accurate consensus sequence, using the long reads and minimap2 (v2.17; Li, 2018) as alignment tool. The short reads were then incorporated with Pilon (v1.23; Walker et al., 2014) to correct the assembly for local misassemblies, single base differences, and indels. In this case, Bowtie2 (v2.3.5.1; Langmead and Salzberg, 2012) was the alignment tool used.

Quality-filtered and trimmed short reads were used to assemble the mitochondrial genome, using NOVOPlasty (v3.8.3; Dierckxsens et al., 2017). This mitochondrial assembly was then polished using Pilon and incorporated into the final assembly.

Assembly evaluation and alignment to the reference genome

To assess the quality of the genome assembly, QUAST (Gurevich et al., 2013) was used with the genome of *S. cerevisiae* S288C as reference, GeneMark-ES (Ter-Hovhannisyan et al., 2008) for gene prediction, and the benchmarking universal single-copy orthologs methodology (BUSCO; Simão et al., 2015) to measure the genome completeness.

To correlate contigs and chromosomes, the genome assembly was aligned to the chromosomes of the *S. cerevisiae* S288C reference genome, using minimap2 and MUMmer (v3.23; Kurtz et al., 2004).

To determine the ploidy of the genome of *S. cerevisiae* IMDO 050523, the short reads were mapped to the genome assembly, using Bowtie2 and SAMtools (v1.4; Li et al., 2009), and the coverage depth along the genome was plotted.

Finally, the heterozygosity level was estimated by calculating the alternative allele frequency in comparison to the *S. cerevisiae* S288C reference genome. Briefly, short reads were mapped to the genome sequence of the *S. cerevisiae* S288C reference genome and variant calling was performed using GATK (McKenna et al., 2010), as described before (Saubin et al., 2020).

Genome annotation

The final polished genome assembly was annotated using the MAKER2 pipeline (Holt and Yandell, 2011), with AUGUSTUS (v3.3.3; Stanke et al., 2006) as gene prediction program. The RNA sequences and transposable elements from the *S. cerevisiae* S288C reference genome (SGD; Cherry, 2015) and *Saccharomycetales* protein sequences from Swiss-Prot were used as references for the functional annotation. Further, the MAKER2 output, containing protein information, was used as query for alignment searches using the blastp algorithm (v2.10; Altschul et al., 1997), setting the *Saccharomycetales* protein sequences from Swiss-Prot as database. Only the best match was considered, with a maximum e-value set to 10^{-6} to obtain high accuracy matches only. Finally, InterProScan (v5.36; Zdobnov and Apweiler, 2001) was used to add information on protein domains, families, and functional sites to complete the functional annotation of the *S. cerevisiae* IMDO 050523 genome.

***Saccharomyces cerevisiae* pangenome, accessory genome, and phylogenomic tree inference**

The 106 *S. cerevisiae* genomes examined were imported into anvi'o (v6.2; Eren et al., 2015). The contigs smaller than 500 bp were removed, and genes were predicted with AUGUSTUS, using the *S. cerevisiae* S288C reference genome. The AUGUSTUS output was reformatted to fit into the anvi'o pipeline, using the command "anvi-script-augustus-output-to-external-gene-calls-partial". The amino acid sequences generated with AUGUSTUS were extracted from the gene prediction files and functionally annotated using

blastp. The pangenome was generated as described before (Delmont and Eren, 2018). Based on the presence/absence of GCs, the accessory genome was manually binned and analysed, separately from the rest of the pangenome, to assess differences between the phylogenomic clusters. Finally, to assess the ANI among all genomes examined, the program pyANI (v0.2.10; Pritchard et al., 2016) was used with the command "anvi-compute-genome-similarity". A phylogenomic tree was generated based on the calculated ANI values and was built using FastTree (v2.1.11; Price et al., 2010).

Population structure analysis

The raw sequence reads were quality-filtered and trimmed, as detailed above. The trimmed reads were aligned to the S288C reference genome using Bowtie2, the alignment files processed using SAMtools, and the variant calling performed using GATK, as described before (Saubin et al., 2020). The variant call format (VCF) file consisted of 481,775 SNPs, which were further filtered using PLINK (v1.90; Chang et al., 2015), as described before (Bigey et al., 2021). The final 43,875 biallelic SNPs were subsequently used to infer the ancestral populations present in the dataset. To that end, two different approaches were followed. First, twenty iterations of fastStructure v1.0 (Raj et al., 2014) using the simple prior were performed, varying the number of ancestral populations used to describe the population structure from 1 to 20. The function ChooseK was then used to select the model complexity that maximized marginal likelihood. Finally, Structure Plot (Ramasamy et al., 2014) was used to generate the structure plots. Alternatively, the VCF file was processed using the R package adegenet (v2.1.5; Jombart, 2008) and pegas (v1.1; Paradis, 2010) to perform a discriminant analysis of principal components (DAPC). The function find.clusters was used to identify the number (*K*) of clusters based on the resulting Bayesian information criterion (BIC) obtained for increasing values of *K*. The dapc function was finally used to describe the relationship between the inferred clusters. The number of retained PCs was selected considering the a-scores obtained for each cluster.

In silico functional analysis

To perform a detailed analysis of the functional differences between strains from cocoa and non-cocoa origins, a manual inspection of the GCs present in the accessory genome of the *S. cerevisiae* pangenome built up was performed, as not all GCs could be assigned to a specific cluster of orthologous group (COG) functions. Thus, the anvi'o interactive interface was used to access the GCs information. The amino acid sequences that were more abundant in the genomes of cocoa strains compared to all other genomes examined were used as queries for sequence and structure homology searches in HHPred (Zimmermann et al., 2018), the SUPERFAMILY 2.0 database (Pandurangan et al., 2019), and the CDD (Lu et al., 2020).

The amino acid sequences obtained from the 106 *S. cerevisiae* genomes examined were also used for orthogroup (OG) inference, which was performed with OrthoFinder (v2.3.8; Emms and Kelly, 2018). DIAMOND (Buchfink et al., 2015) and MUSCLE (v3.8.31; Edgar, 2004) were used to search common sequences among the 106 genomes and to align them, respectively. For an *in silico* analysis, enzymes of interest regarding cocoa fermentation traits, in particular those involved in flavour formation, pectin degradation, sucrose metabolism (invertase activity), citrate consumption, and proteins involved in osmotolerance (aquaporins), were retrieved from UniProt (The UniProt Consortium, 2021), prioritizing the Swiss-Prot entry if available. Their amino acid sequences were aligned against all ortholog sequences produced with OrthoFinder using blastp. The presence, absence, and/or copy number variation were checked by means of a custom Python script. SeaView (v5.0.4; Gouy et al., 2010) was used to align the amino acid sequences and manually check differences across the OGs.

To perform a dedicated screening as to the presence of putative pectin-degrading enzymes in the genome of *S. cerevisiae* IMDO 050523, the Swiss-Prot database was used. Therefore, the keywords "polygalacturonase", "pectate lyase", and "pectin esterase" were used as query searches and the results were downloaded as a fasta file. The same strategy was followed to perform a dedicated screening as to the presence of putative invertases (keywords "invertase", "sucrase", "saccharase", and "alpha-glucosidase"), and citrate lyases/synthases (keywords "citrate lyase" and "citrate synthase"). The amino acid sequences of each fasta file were then aligned using MUSCLE (v3.8.31; Edgar, 2004) and the identity matrix generated was used to examine the clusters formed. An amino acid sequence representative for each cluster was picked to perform a sequence alignment to the whole-genome sequence of *S. cerevisiae*

IMDO 050523, using the algorithm tblastn (Altschul et al., 1997). Furthermore, citrate transporter proteins were screened for, using the same strategy as elaborated above.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical details can be found in the figure legends. The maximum e-value to consider an alignment as significant was set to 0.001, as has been suggested before (Pearson, 2013).