G3, 2023, 13(7), jkad104
https://doi.org/10.1093/g3journal/jkad104
Advance Access Publication Date: 2 June 2023
Investigation

Yeast population dynamics in Brazilian bioethanol production

Artur Rego-Costa,[1,†] I-Ting Huang,[1,†] Michael M. Desai,[1,2,3,4] Andreas K. Gombert [5,*]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Physics, Harvard University, Cambridge, MA 02138, USA
[3]NSF-Simons Center for Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge, MA 02138, USA
[4]Quantitative Biology Initiative, Harvard University, Cambridge, MA 02138, USA
[5]School of Food Engineering, University of Campinas, Campinas, SP 13083-862, Brazil

*Corresponding author: School of Food Engineering, University of Campinas, Rua Monteiro Lobato 80, Campinas, SP 13083-862, Brazil. Email: gombert@unicamp.br
[†]These authors contributed equally to this work.

Abstract

The large-scale and nonaseptic fermentation of sugarcane feedstocks into fuel ethanol in biorefineries represents a unique ecological niche, in which the yeast *Saccharomyces cerevisiae* is the predominant organism. Several factors, such as sugarcane variety, process design, and operating and weather conditions, make each of the ~400 industrial units currently operating in Brazil a unique ecosystem. Here, we track yeast population dynamics in 2 different biorefineries through 2 production seasons (April to November of 2018 and 2019), using a novel statistical framework on a combination of metagenomic and clonal sequencing data. We find that variation from season to season in 1 biorefinery is small compared to the differences between the 2 units. In 1 biorefinery, all lineages present during the entire production period derive from 1 of the starter strains, while in the other, invading lineages took over the population and displaced the starter strain. However, despite the presence of invading lineages and the nonaseptic nature of the process, all yeast clones we isolated are phylogenetically related to other previously sequenced bioethanol yeast strains, indicating a common origin from this industrial niche. Despite the substantial changes observed in yeast populations through time in each biorefinery, key process indicators remained quite stable through both production seasons, suggesting that the process is robust to the details of these population dynamics.

Keywords: eukaryote metagenomics, eco-evolutionary dynamics, bioethanol, industrial fermentation, yeast

Article summary

Microbial ecology and evolution is critical to many industrial processes, from the production of cheese to biofuel. Here, we provide the first high-resolution analysis of microbial evolution in 1 such process: fermentation of sugarcane into fuel ethanol in large-scale Brazilian biorefineries. We find that fuel production is robust despite complex eco-evolutionary dynamics of the baker's yeast populations that drive this process, which is characterized by enormous genetic diversity and substantial fluctuations in strain composition, including invasions by foreign strains.

Introduction

Fuel ethanol is used throughout the world to power light vehicles, either on its own or, more commonly, mixed with gasoline for increased octane rating (Johnson *et al.* 2015). Brazil is the second largest ethanol producer in the world, surpassed only by the United States, and accounts for roughly 30% (or 31.66 billion liters predicted for 2022) of the world's fuel ethanol production (Barros 2022). While American ethanol is mostly corn-based and requires enzymatic hydrolysis of starch prior to fermentation by the yeast *Saccharomyces cerevisiae*, most of Brazil's ethanol is produced from sucrose-, glucose-, and fructose-rich sugarcane products which can be directly fermented.

The Brazilian process is also unique in that it maintains a very large population of yeast in nonaseptic conditions throughout the 8-month-long sugarcane harvesting season (Amorim *et al.* 2011; Della-Bianca *et al.* 2013; Bermejo *et al.* 2021; Fig. 1a). The yeast cells are recycled at every ~12-h fed-batch fermentation-holding-centrifugation-treatment cycle, allowing for large inocula and short turnaround times. Acid wash and antimicrobials serve to control the ever-present bacterial contamination, which competes against yeast for carbon but also affects fermentation in ways that are not completely understood (Lino *et al.* 2021; Senne de Oliveira Lino *et al.* 2021). These practices are key to the high efficiency of the sugarcane–ethanol industrial process and drastically lower greenhouse gas emissions in comparison to corn-based ethanol (Crago *et al.* 2010; Pereira *et al.* 2019). However, inconsistencies in fermentation performance associated with cell recycling remain a costly challenge and point to microbiological routes for process improvement (Amorim *et al.* 2011; Rich *et al.* 2018; Senne de Oliveira Lino *et al.* 2021).

Yeast strains differ in their suitability for industrial-scale fermentation. Traditionally, the readily available baker's yeast was used to kick-start the fermentation season, but due to its

Received: December 23, 2022. **Accepted:** April 24, 2023

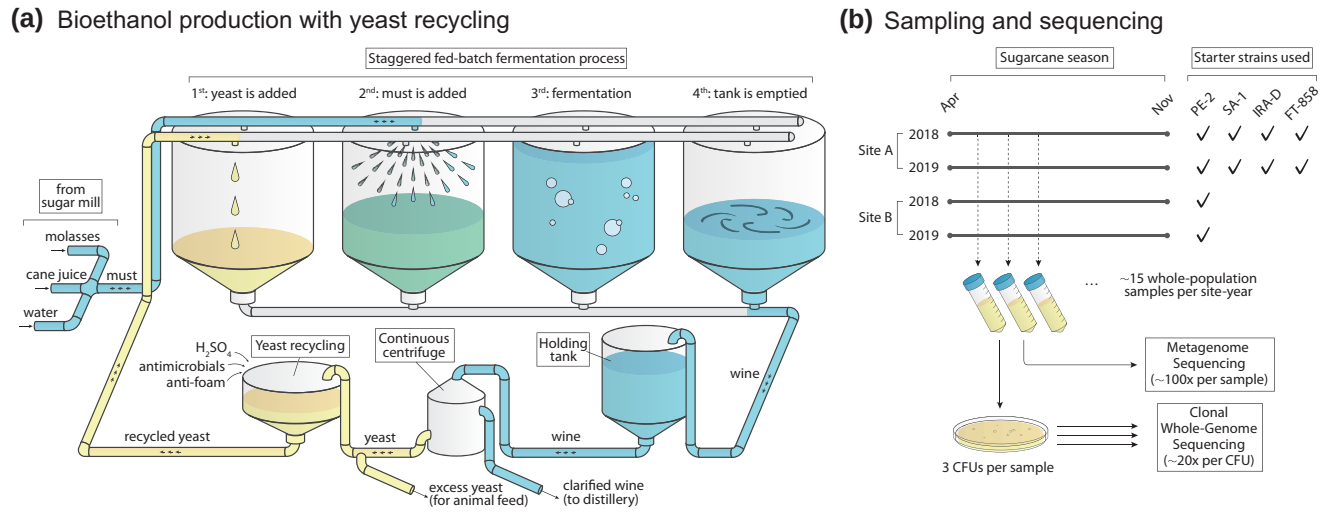

Fig. 1. Schematics of the fermentation process and sequencing strategy. a) A large population (~10^{17} individuals) of the yeast *S. cerevisiae* is maintained over the course of an 8-month-long fermentation season. Yeast ferments must, a mix of molasses, sugarcane juice, and water, to produce ethanol in a fed-batch process that takes ~8 h and runs in a staggered parallel fashion across several fermentors (8–16 in any 1 plant, each with an ~500,000 ℓ capacity). The fermented broth (wine) from different fermentors is loaded into a single holding tank, which continuously feeds a centrifuge for separation of the yeast from the liquid fraction. Holding tanks are larger than fermentors themselves and allow for mixing between batches. The yeast cells are then treated with chemicals to control for bacterial growth and are later reused in the process. The yeast population grows by ~10% every 12 h, leading to approximately 66 generations over the course of an ~8 months fermentation season. The season is started with selected industrial strains which are commercialized by yeast suppliers. b) We collected whole-population samples of the yeast used for fermentation through 2 seasons (2018 and 2019) in 2 plants (Site A and Site B) located ~18 km apart in the state of São Paulo, Brazil. The 2 plants are owned by different companies and use different sets of starter strains in their process. We employed a combination of whole-population metagenome sequencing and clonal whole-genome sequencing to observe the temporal dynamics of genetic diversity in each site-year. See Supplementary Tables 1–3 for a complete list of collected samples and isolates.

susceptibility to invasion by foreign *S. cerevisiae* lineages, production has largely shifted toward specialized starter strains. A major strain selection program conducted between 1993 and 2005 solidified the potential for these invading strains themselves to serve as a source of new industrially relevant variants (Basso *et al.* 2008). Strains isolated from this program, namely, PE-2, CAT-1, SA-1, BG-1, and VR-1, and their derivatives, as well as JP-1 (isolated from a similar effort; da Silva Filho *et al.* 2005), are the basis for the bulk of today's ethanol production and have successfully helped maintain the overall high yield of the industry. Still, invasion by foreign strains remains common, as fermentation conditions across the ~400 bioethanol plants operating around the country span a range of industrial practices, environmental conditions, sugarcane varieties, and other factors, in addition to the yet-little-explored possibility of evolutionary change over the course of a fermentation season.

To identify and track these yeast population dynamics in industry, chromosomal karyotyping became popular in the 1990s and is still commonly used for process monitoring (Basso 1993; da Silva Filho *et al.* 2005; Basso *et al.* 2008). More recently, PCR-based methods have helped in decreasing the cost of strain surveillance (da Silva-Filho *et al.* 2005; Antonangelo *et al.* 2013; Carvalho-Netto *et al.* 2013; Reis *et al.* 2017). However, these methods cannot readily differentiate closely related strains, which may differ by few mutations anywhere along the whole genome. Moreover, these methods estimate lineage frequencies based on fraction of picked isolates from agar plate streaks, which leaves room for biased assessments of strain dominance if strains differ in culturability.

Whole-genome metagenomic shotgun sequencing is a potential culture-independent alternative method for strain differentiation (Anyansi *et al.* 2020). Temporal metagenomic data sets have been used to assess microbial community dynamics with subspecies resolution, largely in the context of human gut microbiomes (Schloissnig *et al.* 2013; Franzosa *et al.* 2015; Luo *et al.* 2015; Scholz *et al.* 2016; Costea *et al.* 2017; Truong *et al.* 2017; Smillie *et al.* 2018: 20; Garud *et al.* 2019; Zhao *et al.* 2019; Roodgar *et al.* 2021). However, inference of the underlying strain movements from metagenomic frequency trajectories remains challenging and methods are mostly limited to low-diversity and prokaryotic populations. Nonhaploidy complicates this inference even further, as the diploid or polyploid genotype of individual variants (which itself may vary among individuals in a population) must also be accounted for.

Here, we present a novel framework for inferring the population dynamics of highly diverse, nonhaploid, asexual microbial populations from a combination of clonal sequences and temporal metagenomic data. We employ this method to investigate the dynamics of yeast genetic diversity across 2 fermentation seasons, in 2 independently run bioethanol plants in Brazil. More specifically, we ask whether starter strains tend to persist and dominate through an entire production season, and, if not, what strains they are replaced with. We also investigate the differences between seasons and production facilities, the origin of invading strains, and the effects they have on the process. Our focus here is on the yeast dynamics, but our sequencing data also contain information on other microbial species, which remains to be analyzed in future work.

Methods

Sample collection

We collected whole-population microbiological samples from 2 independent industrial units, which we refer to as *Site A* and *Site B*, through 2 fermentation seasons, 2018 and 2019, which ran from April/May through November/December (Fig. 1). Sampling started on the first day of the fermentation season for Site A 2018 and ~14 days into the season for the other site-years (see

sampling dates in Supplementary Table 1). The 2 sites are owned by different companies and are located 18 km apart in the region of Piracicaba, São Paulo, Brazil. Site A used a mix of 4 strains to start both the 2018 and 2019 fermentation periods—namely, strains PE-2, SA-1, FT-858, and IRA-D. While the first 3 are common commercially available industrial strains, IRA-D is an in-house strain isolated from Site A in a previous fermentation season. In contrast, Site B informed us that they have used PE-2 as their sole starter strain in both fermentation seasons, although we would later find evidence suggestive of a second starter strain being used, possibly unknowingly, in 2019 (see Results below).

Samples (~10 ml) were collected daily (2018) or weekly (2019), after fermentation was completed, directly from fermentors or holding tanks, into presterilized 15-ml tubes containing 3-ml glycerol. After mixing by vortexing, samples were stored at −20°C for a period of between 1 and 3 months before being transferred to a −80°C ultrafreezer. Finally, samples were shipped from Brazil to the United States in dry ice, where they were stored at −80°C. Starter strains PE-2, FT-858, and SA-1 were shipped as active dry yeast (ADY), whereas strain IRA-D was shipped as colonies on agar slants, without dry ice. The collection and shipping of samples has been registered at the Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado (SisGen, Brazilian federal government) under numbers R40E57A, RB42674, R193AED, and RAD5521 (for the shippings) and AF14971 (for the sampling). A full list of samples with associated collection dates can be found in Supplementary Table 1. Picked clonal isolates are made available upon request.

DNA extraction and sequencing

We selected 15–20 samples from each site-year for whole-genome metagenomic and clonal sequencing. For metagenomic sequencing, samples were completely thawed and vortexed, after which 1 ml was aliquoted and centrifuged to remove the supernatant. Whole DNA extraction was carried out using an in-house protocol (Nguyen Ba *et al.* 2019). Sequencing library preparation was done using the transposase-based protocol (Baym *et al.* 2015).

For clonal isolate sequencing, the same 15–20 thawed and homogenized samples were used for plating onto Yeast Extract-Peptone-Dextrose (YPD)-agar (Supplementary Table 2). Plates were incubated at 30°C for 24–48 h. From each plate, 2 or 3 CFUs were picked and grown in 5-ml liquid YPD overnight at 30°C, after which DNA extraction and library preparation proceeded as for metagenomic sequencing. Starter strains were inoculated in liquid YPD, left to grow overnight at 30°C, plated, and prepared in the same manner (Supplementary Table 3).

Sequencing was carried out in 2 Illumina NextSeq and 1 Illumina MiSeq runs, following a 300-bp paired-end workflow. Mean coverage after mapping to the reference strain S288c genome and haplotype inference (see section below) was 87× for metagenomic samples and 26× for clonal isolates. FASTQ files with all sequencing reads produced for this study were deposited in the NCBI SRA database (see Data availability).

Variant calling bioinformatic pipeline

We called variant sites (SNPs only) in relation to the *S. cerevisiae* S288c reference genome (yeastgenome.org, release R64) in all our metagenomic and clonal isolate data. The full pipelines with specific tools and settings used can be found in the GitHub repository (see Data availability). In summary, all sequencing reads were first trimmed of sequencing adapters using NGmerge (Gaspar 2018) and then aligned to the reference genome using BWA (Li and Durbin 2009). Variant calling was done with the haplotype inference tools in the Broad Institute's GATK (van der Auwera and O'Connor 2020). In essence, these tools assemble local haplotypes from aligned reads, calculate the posterior probability of each read coming from each of the assembled haplotypes, and finally infer variant sites jointly across a group of samples for added power to call true low-frequency variants: intuitively, an observed variant is less likely to be a sequencing error if it is observed in more than 1 sample. Given different probabilistic prior models of allele frequency for clonal and nonclonal data, variant calling of isolate clonal data is done with HaplotypeCaller jointly across all isolates, while that of the metagenomic data is done using Mutect2 jointly across all timepoints within each site-year, in line with GATK guidelines (van der Auwera and O'Connor 2020). Alternate and reference allele counts (AD field in the VCF files) outputted by the variant calling tools are estimates based on inferred haplotype membership of aligned reads (instead of being simple observations from aligned reads). These are the numbers that we use for all later analyses. For convenience, when referring to a variant site, we will often refer to alternate allele counts as simply *counts*, and the sum of alternate and reference allele counts as simply *depth*. In all further sections, *allele frequency* at a variant site is defined as the number that ranges from 0 to 1 given by counts divided by depth. For the sake of simplifying, we exclude from analyses the small number of variant sites for which we observe more than 1 alternate allele.

Isolate ploidy

Isolate ploidy was assessed based on visual examination of the distribution of allele frequencies in clonal isolate data over the whole genome (upper right corner of each panel in Supplementary File 1): diploid strains have a multimodal distribution peaked at values 0, 0.5, and 1, while triploid strains, at 0, 1/3, 2/3, and 1. Example allele frequency distributions from a diploid and a triploid strain are shown in Supplementary Fig. 8.

Phylogenetic analyses

We infer 2 phylogenetic trees in this study, both using whole-genome SNP data. *Tree 1* was run with the SNPhylo pipeline (Lee *et al.* 2014) using default parameters. The tree is inferred based on a total of 27,229 SNPs across all clonal isolates from all site-years, including isolates from the 4 starter strains (Newick format tree in Supplementary File 2). *Tree 2* includes the same clonal isolates, plus all isolates from the 1,011 yeast genomes project (YGP) (Peter *et al.* 2018; Supplementary Fig. 10; Newick format tree in Supplementary File 3). For this tree, SNPs were first filtered and aligned using SNPhylo with a missing rate of 0.001, and a maximum likelihood tree was constructed from 42,012 SNP markers using RAxML (Stamatakis 2014) with 1,000 bootstrap replicates, employing the general time reversible nucleotide substitution model with the GAMMA model of rate heterogeneity. For the purposes of downstream analyses and presentation, Tree 1 was re-rooted in a node analogue to that from which the Bioethanol subtree of Tree 2 branches from the remainder of the tree.

Inference of population dynamics

We assume the reproduction during fermentation is exclusively asexual. Therefore, the population is composed of some large but discrete number of clonal strains of asexually dividing individuals which may have 3 origins: (i) preexisting diversity in starting inoculum, (ii) invading strains during the course of the fermentation season, and (iii) new strains founded by de novo mutational events during fermentation.

Clonal strains share phylogenetic history and therefore alleles. Assuming no recombination and no de novo mutation reversal, we assume that these lineages organize themselves into a hierarchical tree-like structure which defines clades, herein referred to as *lineages*, each with a particular set of synapomorphic alleles: i.e. alleles that are shared by all clonal strains within that lineage but no strain outside of it. In effect, the inference pipeline should be able to handle some amount of departure from these assumptions due to past history of recombination, mutation reversals, and noise, but we expect this pattern to compose the bulk of the observed data.

Our goal was to use the metagenomic data to infer the frequencies through time of as many lineages as possible in order to characterize the population dynamics over the course of the fermentation season in each site-year. Our inference consists of (i) identifying lineages and their synapomorphic alleles based on a maximum likelihood phylogeny inferred from our sequenced clones and (ii) looking for each lineage's set of synapomorphic alleles among the metagenomic sequencing data to infer lineage frequencies using a maximum likelihood framework. The rationale for this approach is that the metagenomic data samples genetic diversity among chromosomes in the population in an unbiased way, while the clonal genome sequencing informs us of how to group alleles that segregate together in the same lineages. We note that these synapomorphies may be carried at different number of copies in the genome of each strain (i.e. as 1 or 2 copies in diploid strains and 1, 2, or 3 copies in triploid strains), convoluting the relationship between lineage frequency in the population and mutation frequency in the metagenome. Therefore, we restrict ourselves to synapomorphies that are carried at the same number of copies across all members of a given lineage, probabilistically inferring genotype of individual picked clones (Supplementary Fig. 1). We do not assume any particular dynamical model of evolution and instead infer lineage frequencies at each timepoint independently. A crucial feature of this inference is that genetic diversity that is not sampled among sequenced clones does not bias the frequency estimates of other lineages.

A detailed description of the inference pipeline is described in the Supplementary material, together with a validation analysis using subsampled clonal data (Supplementary Figs. 2–5). The code developed for this inference is available in the GitHub repository (see Data availability).

Results

We carried out temporal whole-population metagenome sequencing of the *S. cerevisiae* populations used to ferment sugarcane feedstock into bioethanol over 2 fermentation seasons (2018 and 2019), at 2 independently owned biorefineries (Site A and Site B) in the state of São Paulo, Brazil (Fig. 1). We also whole-genome sequenced ∼35 isolated clonal strains from each site-year. Metagenomic and clonal sequencing reads were aligned to the reference genome of strain s288c and used to call and count genomic variants in the data. See Methods for details.

High genetic diversity among industrial isolates

We began by investigating genetic diversity in the studied populations. Using our variant calling pipelines (see Methods), we find a total of 145,066 SNPs among all 134 fermentation and 11 starter strain isolates. A total of 14,200 (9.8%) of these mutations are singletons, while 15,749 (10.5%) are seen in all sequenced clones (see Supplementary Fig. 7 for the full distribution). We also find a similar number of SNPs (150,265) in the whole-population metagenome data across all 4 site-years, with an overlap of 126,845 between the clonal and the metagenomic data sets. This suggests that the clonal genotyping data covers a substantial fraction of the genetic diversity of these populations, especially given that the metagenomic data (i) samples from the whole population and (ii) represents a sequencing effort of 6154× over all timepoints, which is larger than that of clonal genotyping (4,341× over all isolates). The 168,486 SNPs uncovered in the whole data set are widely distributed along the genome, hitting 6,370 out of all 6,579 genes in the annotated S288c genome. A total of 129,697 of these SNPs have been previously observed in the 1,011 YGP, which itself uncovered 1,544,489 SNPs (Peter *et al.* 2018).

S. cerevisiae may exist at different ploidies, and so we examined allele frequencies in the clonal isolate data to infer isolate ploidy (see Methods for details). We found that 64 of our isolates are triploid, while the remaining 70 are diploid (Fig. 2a). All isolates of starter strains FT-858 and IRA-D are triploid, while those of PE-2 and SA-1 are diploid (as described in Basso *et al.* 2008; Argueso *et al.* 2009; Nagamatsu *et al.* 2019). An examination of allele frequencies and sequencing depth along the genome revealed that a small number of isolates carry structural variations, such as gain or loss of whole chromosomes or sections of chromosomes (Supplementary File 1). Given the small number of affected isolates and in each case a minor fraction of the genome being affected, we keep these isolates in all further analyses.

We then used the called SNP data to infer a maximum likelihood phylogenetic tree between all sequenced isolates (Fig. 2a). As expected, we find that several of the isolated clones are closely related to the starter strains used to initiate the industrial process. We note that PE-2 isolates form 2 major clades, which are both represented in starter and fermentation isolates from both sites and years. We also find several other groups of closely related isolates, mostly triploid, that diverge from the starter strains by thousands of SNPs. These groups are all composed of isolates from Site B, whereas all Site A isolates fall close to the known starter strains.

Lineage inference

We turned to the whole-population metagenomic data to investigate the yeast population dynamics through the fermentation season (Fig. 2b). We are interested in understanding how starter strains change in frequency through the fermentation, as well as identifying events of selection of de novo mutations or invasion by foreign strains. Examining the raw metagenomic allele frequencies through time, we observe periods when large cohorts of mutations move together, indicative of competition between divergent strains, as well as periods of stability when allele frequencies remain mostly constant. Correlation between allele frequency trajectories is indicative of co-segregation and has been used as the signal for inference of population dynamics in previous studies (Luo *et al.* 2015; Smillie *et al.* 2018). However, this type of inference is complicated by several factors. First, our populations are highly genetically diverse, and mutations are shared between different strains in complex patterns. These patterns are presumably created by earlier, potentially sexual population dynamics that led to the creation of these strains in other unknown environments in which they evolved. This means that individual metagenomic mutation trajectories can depend on the frequency changes of potentially multiple different strains that carry that mutation. This is complicated by the fact that these different strains may carry a given mutation at different genotypes (i.e. as homozygous or heterozygous diploids or in 1–3 copies in triploids). Finally, it is not immediately clear how to

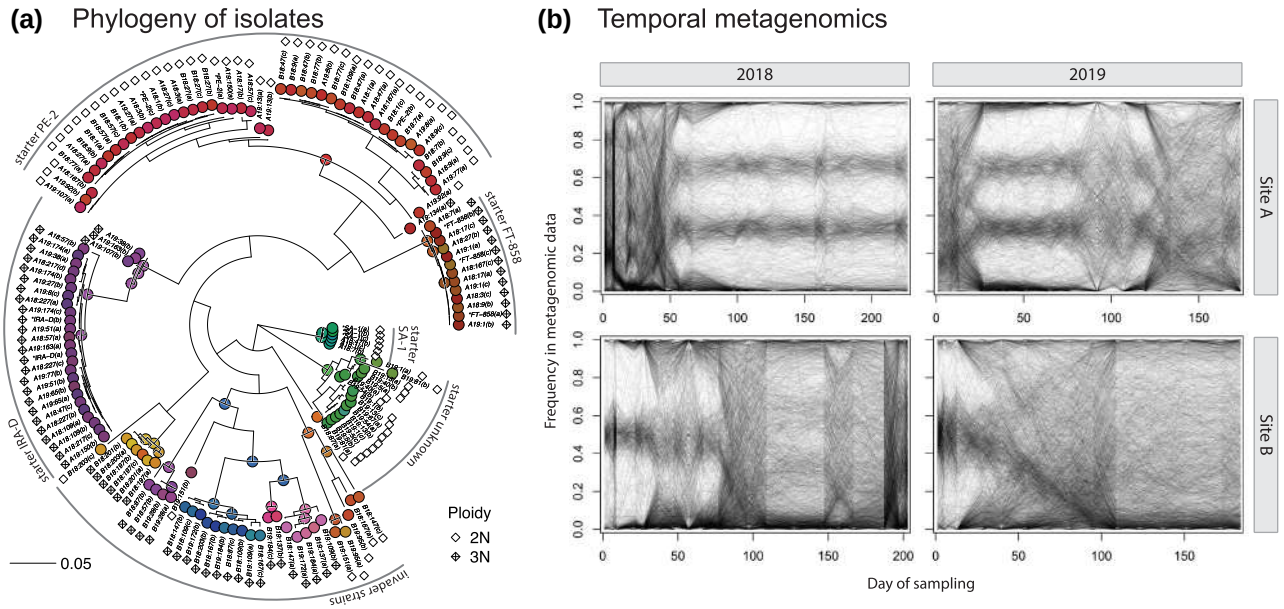

Fig. 2. Yeast populations in bioethanol fermentors are genetically diverse and dynamic. a) Phylogenetic tree of isolated clonal strains from all site-years, as well as known starter strains used. Most isolates are closely related to the known starter strains, but several are not. The tree was inferred with a maximum likelihood model using the data of 27,229 SNPs. Ploidy of each isolate, assessed as described in the Methods, is indicated by diamonds. Nodes and tips are colored as in Figs. 4 and 5. The tree is rooted in the same place as the independently inferred tree in Fig. 6. Isolates are grouped as in Figs. 4–6. Isolates are named as <*site*><*year*>: < *timepoint* > (<*letter identifier*>), while starter strain isolates are marked with an asterisk. The associated Newick tree can be found in Supplementary File 2. The allele frequency data used for ploidy assessment can be visualized in Supplementary File 1. Selected examples of a diploid and triploid strain can be seen in Supplementary Fig. 8. b) Frequency of alternate allele (in relation to the reference genome of strain s288c) through time for an arbitrary subset of 2,000 mutations (out of ~100,000) per site-year. Overall, mutation trajectories indicate alternation between periods of stasis, when 1 major strain dominates, and periods of transition, when many mutations change in frequency in a correlated way indicative of strain dynamics. Noise in mutation trajectories comes from random sampling (approximately binomial), as well as sequencing and mapping errors, which is not homogeneous across mutations.

polarize mutations for lineage frequency inference (i.e. which one should be considered the reference vs alternative allele), which leads to an overall pattern of mirrored mutation trajectories in the raw metagenomic data (Fig. 2b).

Here, we developed and employed a novel framework for jointly inferring the frequencies of nested asexual lineages of descent through time from whole-population metagenomic data (Fig. 3; see Supplementary Methods and Supplementary material for details). This approach takes advantage of our clonal sequencing data to phase an informative subset of all mutations into cohorts that segregate together in the population, completely ignoring the metagenomic data for this purpose. While we are limited to the genetic diversity that is sampled by picked isolates, by following this approach, we overcome the challenges described above, as well as have higher power to identify small lineages, whose metagenomic trajectories may be indistinguishable from sequencing noise in correlation-based grouping methods (Luo *et al.* 2015; Smillie *et al.* 2018). In doing so, our pipeline automates an approach similar to that of Zhao *et al.* (2019), while handling high genetic diversity and ploidy variation in the population.

Among the 4 site-years, we infer the frequencies of a total of 197 lineages, spanning a wide range of lineage sizes, with a median maximum lineage frequency of 6.7% (see Supplementary Fig. 9 for the full distribution). The inferred results pass basic soundness checks: the timepoints at which different isolates were picked largely correspond to times when their associated inferred lineage frequencies are high, and lineage frequency trajectories are smooth, even though timepoints are inferred independently from each other.

Stable dynamics dominated by in-house strain in Site A

In Site A, we only observe lineages closely related to the known starter strains (Fig. 4). In particular, we find that IRA-D, a triploid strain, dominates the process in both years. Curiously, IRA-D is an in-house strain which was found to invade the process in a previous fermentation season, and since then, it has been included in the starter strain mix. While these observations suggest that IRA-D is the best adapted to these fermentation conditions among all 4 starter strains, we observe that it does not completely displace PE-2 in 2019, which continues at a low frequency in the process even in later timepoints. Coexistence for such a long timescale is suggestive of some ecological process, such as niche partitioning, or negative frequency dependence. However, it is unclear why the same dynamics are not seen in 2018, when PE-2 seems to be completely outcompeted. Either the population itself is genetically different between the years (although isolates from both seasons are closely related), or differences in agricultural and industrial practices, or weather patterns, may have affected fermentation conditions.

Foreign lineages systematically invade Site B

In Site B, we observe a very different picture, where several large lineages are distantly related to the starter strain PE-2 (Fig. 5). While PE-2 dominates at the start of 2018, it is a minor fraction at the start of 2019, when the process is instead dominated by a different lineage (labeled "starter unknown" in Figs. 2a and 5), suggesting a different starter strain mix for that year.

In both years, the population gets substituted by a cohort of much fitter strains halfway into the season (labeled invader

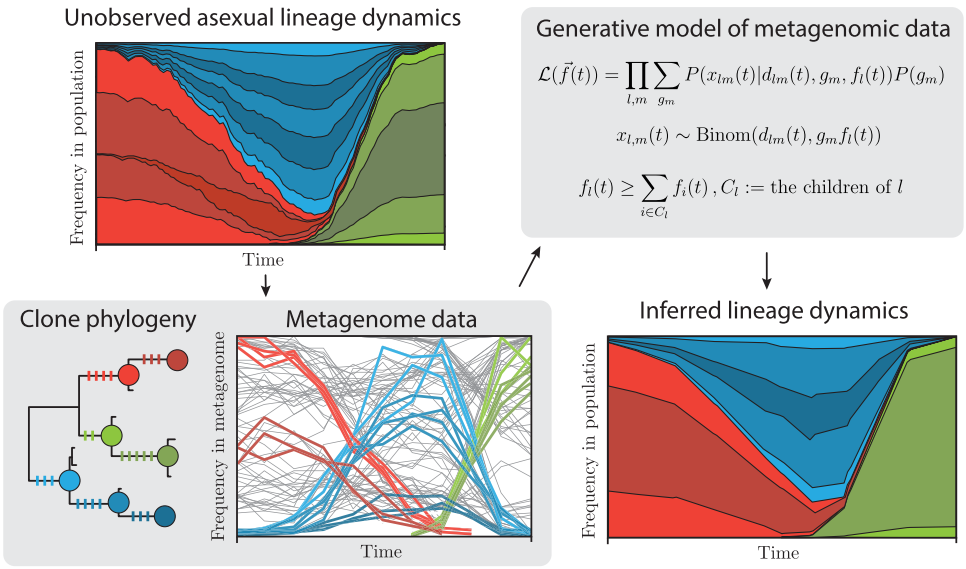

Fig. 3. Schematics of lineage inference procedure. We use temporal metagenomics and clonal isolate whole-genome sequencing to infer the unobserved frequencies of asexual lineages in the original population over the course of a fermentation season. (Upper left) Starter, invading, and newly mutated lineages change in frequency through time due to selective and random factors. (Lower left) A phylogeny of clonal isolates is used to select the sets of clade-defining variants (colored bars on tree branches) that we will later search in the metagenomic data and use for lineage inference. (Upper right) At each timepoint t, we jointly infer the frequencies $\vec{f}$ of all asexual lineages by optimizing a likelihood model of $\vec{f}$ given the metagenomic allele counts x_{lm} of variant m, which is a clade-defining variant for lineage l, the read depth d_{lm}, and the variant's genotype g_m (which takes values 0, 0.5, or 1 for diploid and 0, 1/3, 2/3, or 1 for triploid lineages). The frequencies of all lineages are jointly inferred and constrained such that the summed frequencies of sister lineages do not exceed that of the respective parent lineage. (Lower right) Undersampling of genetic diversity by isolates will cause whole lineages to be left out, but that should not bias the frequency estimation of included lineages.

strains in Figs. 2a and 5). Most of these strains are triploid, except for a small group present in both years (Figs. 2a and 5). While their genetic distance to other starter strains and minute presence in early timepoints suggest that they invade the fermentation process, we cannot rule out that they were already present in the starter inoculum or have their origin in the industrial equipment itself, where they might find a reservoir from one production season to the next. The fact that closely related isolates are seen in both 2018 and 2019 is indicative of some systematic source of contamination. Surprisingly, despite the large degree of genetic diversity and the ploidy variation within this cohort, these different invading strains stably coexist in the timescale of the fermentation season. Here again, an ecological explanation is suggested.

Finally, we observe a second substitution event in the final timepoints of Site B's 2018 season. The inference suggests that this set of strains were already present since early in the season, remaining at low frequency until they suddenly displace all other strains. This event does not seem to be driven by selection for a de novo mutation, since the expanding lineage retains significant diversity within itself, and instead may be caused by a sudden change in fermentation conditions.

Origin of invading yeast strains

We further investigate the origin of Site B's invader strains. While we cannot assess industrial procedures directly, we can examine the phylogenetic relationship of these strains to other known isolates. For that purpose, the 1,011 YGP represents the largest and broadest whole-genome sampling of *S. cerevisiae* genetic diversity (Peter *et al.* 2018). Most importantly, it includes 37 isolates related to the Brazilian bioethanol industry. Here, we compare all our picked isolates to the YGP collection by inferring a combined phylogeny of both studies (Fig. 6; see Methods for details). The inferred unrooted tree largely replicates the structure of previous inferred trees of broad yeast diversity (West *et al.* 2014; Gallone *et al.* 2016; Peter *et al.* 2018; Jacobus, Stephens, *et al.* 2021b).

First, we find that all Brazilian bioethanol isolates from both studies form a monophyletic group and are closely related to a large group of European wine strains, in agreement with previous studies (Fig. 6a; Peter *et al.* 2018; Jacobus, Stephens, *et al.* 2021b). As shown in Fig. 6b, we note that among the 37 isolates classified in the Brazilian bioethanol group in the 1,011 YGP, 3 were isolated from cachaça distilleries (a traditional sugarcane-based spirit), while 2 were from the sugarcane plant or from sugarcane juice (although further detail is missing), while the remainder were isolated from different bioethanol plants. Among these isolates from the bioethanol industry, several are closely related to PE-2, SA-1, and, most notably, to the "unknown starter" strain in Site B's 2019 season. Finally, Site B's "invader strains" do not seem to be represented in the 1,011 YGP, but their close association with other bioethanol isolates points to an industrial origin (e.g. shared equipment, supplies, or sugarcane), as opposed to invasion by wild strains brought to the industrial environment by vectors such as insects or birds from foreign niches.

Stability of macroscopic fermentation parameters despite strain dynamics

Yeast strains vary in their suitability for the industrial process due to, among other factors, their ability to produce and withstand high ethanol concentrations, their propensity to generate foam or cell aggregates in large industrial settings, or their tendency to be outcompeted by poorer performing strains (in terms of the final ethanol yield on sugars) (Basso *et al.* 2008). Thus, invasion by unknown strains may harm the fermentation process and the profitability of the industry, due to decreased ethanol

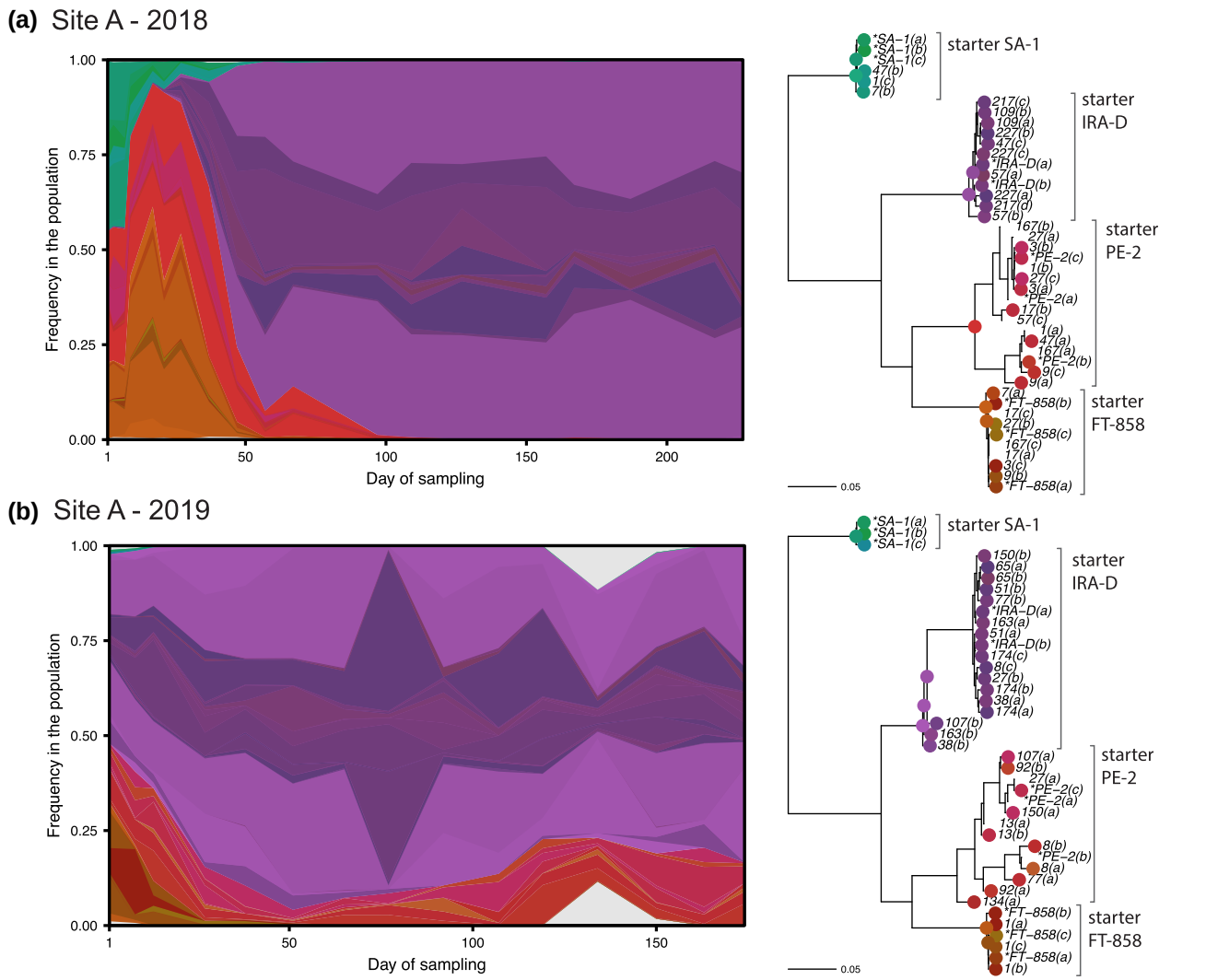

Fig. 4. In Site A the in-house starter strain IRA-D consistently dominates over other starter strains. On the left, inferred strain dynamics in Site A over the 2 fermentation seasons. White space corresponds to noninferred genetic diversity in the population. On the right, subtrees of the tree in Fig. 2a including only the isolates from each respective site-year. Circles on nodes and tips indicate inferred lineages and their respective colors.

production and/or to higher costs involved with the use of chemicals, such as sulfuric acid, antimicrobials, antifoaming agents, and dispersants. In the case of Site B's 2018 and 2019 seasons, we have not found a connection between general industrial metrics and inferred events of population substitution (Supplementary Fig. 11). Nonetheless, it may still be possible that this stability was accomplished by the employment of commonly used but costly corrective measures, such as those outlined above.

Discussion

In this study, we described the population dynamics of the yeast used for bioethanol production via fermentation in sugarcane-based biorefineries through the course of 2 fermentation seasons (2018 and 2019) in 2 independently run industrial plants. The method we developed for this purpose allowed for an unprecedented description of how the starter strains used in the process change in frequency through time and how the fermentation environment may be invaded by foreign strains. We observe that these large populations (estimated to be ~10^{17} individuals) harbor a vast amount of genetic diversity, recovering ~8% of alleles previously found in a *S. cerevisiae*-wide survey (Peter *et al.* 2018), plus novel ones. This diversity is observed not only in invading strains but also within the starter strains themselves, whose same subtypes are sampled across years and sites (most notably the 2 major groups within PE-2; Fig. 2a). This may be due to how propagation companies, which sell large initial inocula to bioethanol producers, keep and propagate their own stocks: companies may not start from single colonies every year, and de novo mutations may accumulate during propagation. Similar observations of strain genotypic (and phenotypic) heterogeneity have also been made in the baking, wine, and beer industries (Rácz *et al.* 2021).

Such large populations must harbor many de novo mutations. At an approximate rate of 5×10^{-10} mutations/bp/generation (Lang and Murray 2008) and at least 66 generations during 1 fermentation season, a total of 8×10^{16} or more mutations should occur in a diploid population of this size. In fact, at this rate, any given SNP in the yeast genome should independently occur ~3×10^{7} times per generation. We cannot know how many of these mutations would be adaptive in the industrial environment, but decades of microbial experimental evolution, including in yeast populations, show that adaptation in large asexual

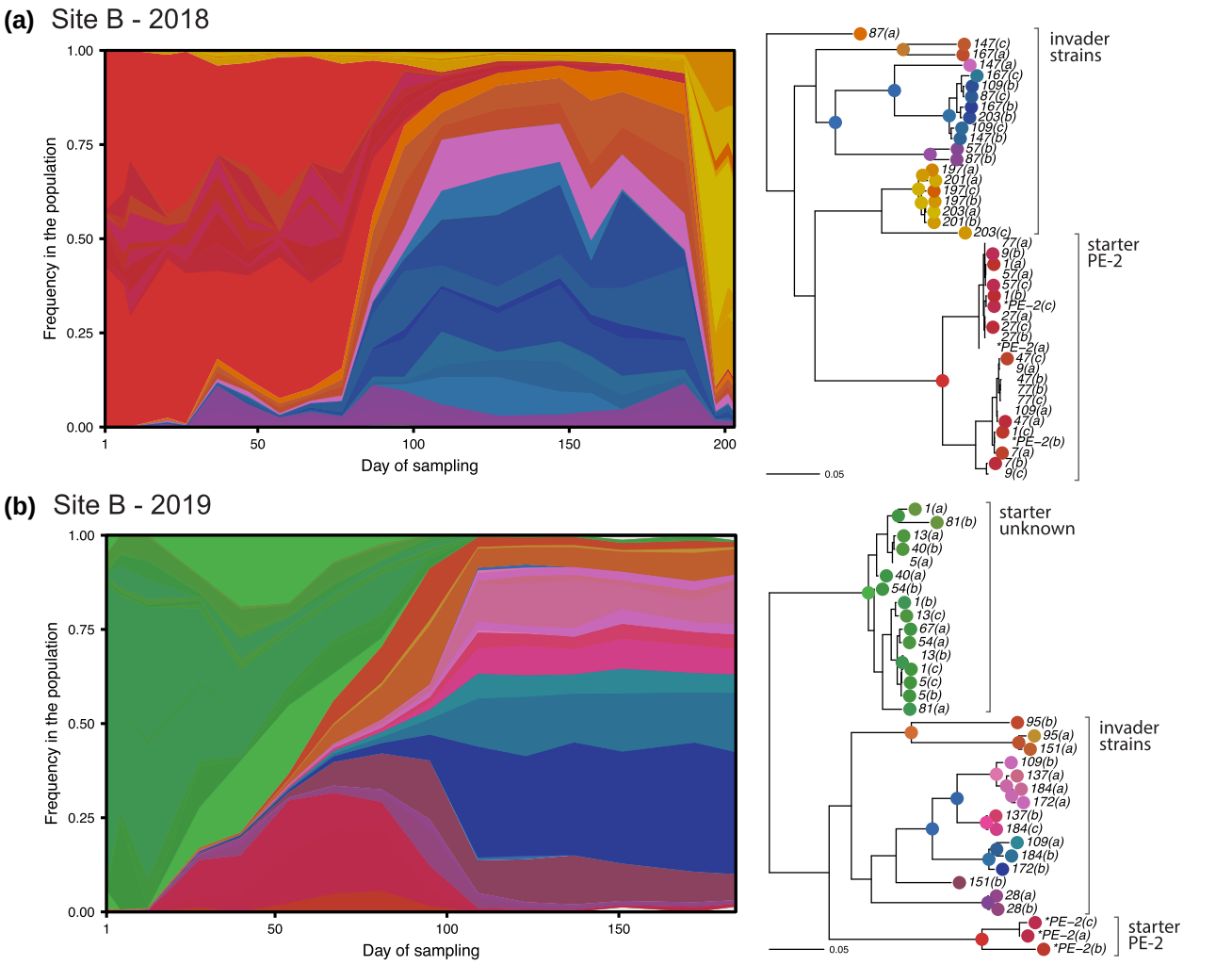

Fig. 5. In Site B, a group of diverse invading strains systematically takes over the process. Despite the genetic diversity among invader strains, they seem to coexist, except for the second substitution event in 2018, which involves a different set of invading strains. In the 2019 fermentation season, the process starts with a large amount of an unexpected unknown strain. See Fig. 4 for a description of the diagrams.

populations is not mutation limited (Barrick and Lenski 2009; Levy *et al.* 2015; Maddamsetti *et al.* 2015; Good *et al.* 2017; Nguyen Ba *et al.* 2019; Johnson *et al.* 2021). Yet, we do not find clear signs of selection for de novo mutations in our results, which would be observed as either an inferred lineage that increases in frequency much faster than its closely related counterparts or inferred lineages being deflected by some unobserved rising lineage. A likely explanation is that the timescale of a fermentation season (in number of generations) is too short for selected lineages, carrying de novo adaptive mutations of a typical fitness effect, to increase in frequency enough to be sampled by our sparse isolate picking strategy. All in all, what this suggests is that as long as starter inocula are not produced from the previous year's final population or that the equipment itself is not contaminated with large amounts of previous populations, evolution on a single-strain background is likely not a consequential factor in the timescale of a fermentation season due strictly to the large population sizes and dynamics of selection.

Ecological dynamics may explain the observed long periods of coexistence between distantly related lineages in both sites, such as in PE-2's permanence in Site A 2019, or the stable relative frequencies of invader strains in Site B 2019. While it is possible that these observations simply reflect small differences in fitness in the fermentation environment, the large phylogenetic distance between strains argues against this hypothesis. Large genetic differences may lead to diversity in resource usage (niche partitioning) and/or in how strains benefit or not from each other's presence (frequency dependence). Such ecological dynamics are by no means rare in microbiological communities in the wild (Faust and Raes 2012; Mitri and Richard Foster 2013) and have been unintentionally evolved in laboratory *E. coli* and *S. cerevisiae* populations (Frenkel *et al.* 2015; Good *et al.* 2017). Strain interactions could open up avenues for designed strain mixes that take advantage of synergistic interactions in terms of fermentation output and management. We also should not discount the potential bacterial contribution to these dynamics, as bacteria have been shown to interact both positively and negatively with yeast during fermentation (Rich *et al.* 2018; Senne de Oliveira Lino *et al.* 2021). The analyses carried out for the current study do not include bacterial data, but such microbial consortia compose an interesting avenue for future work.

The fact that results have varied more between industrial plants than between years suggests that systematic differences in industrial practices and/or starter strain mix largely explain differences in population dynamics. Additionally, observed fluctuations in strain frequencies through time (e.g. the strain

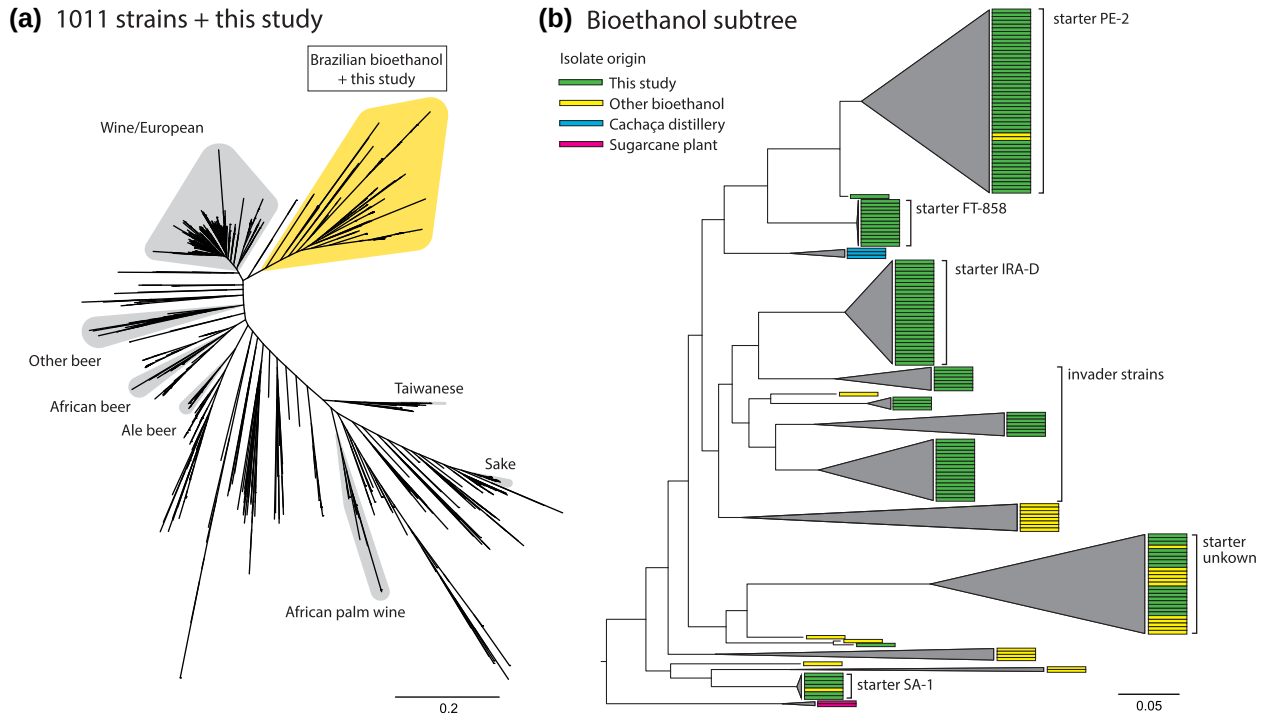

Fig. 6. Starter and invader isolates all cluster together within a larger group of Brazilian bioethanol strains. a) A SNP-based maximum likelihood phylogeny combining isolates from the current study and from the 1,011 Yeast Genomes Project (Peter *et al.* 2018). Other groups of domesticated strains are highlighted for reference. This tree was inferred based on 42,012 SNPs. b) Subtree of bioethanol-related isolates. Isolates from the current study are closely associated with isolates from the bioethanol industry and cachaça distilleries (a sugarcane-based spirit). Individual isolate origins are indicated with colored rectangles. Branches are collapsed to aid visualization. A full phylogeny can be seen in Supplementary Fig. 10, and its associated Newick tree can be found in Supplementary File 3.

responsible for the second substitution event in Site B 2018) indicate that fluctuations in fermentation conditions may make certain strains more or less fit to the industrial environment. This is not unexpected, as (i) fermentors are only partially protected from external temperature fluctuations, (ii) incoming sugarcane varieties change through the year and result in different must compositions, (iii) the ratio of sugarcane juice and molasses in the must is adjusted daily depending on current sugar and ethanol prices, (iv) clean-in-place (CIP) practices are carried out on a regular or as-needed basis, and (v) recycling practice may be adjusted depending on levels of bacterial contamination, among other factors. Further collaborations with companies, including access to a detailed record of industrial practices and strain tracking as done in this study, may shed further light into the causes behind fermentation fluctuations. These records should especially contain information on the usage of chemicals (e.g. sulfuric acid, antimicrobials, antifoaming agent, and dispersant, among others), which remediate fermentation output, but add to production cost and greenhouse gas emissions.

Our observation that the in-house strain IRA-D dominates the process throughout the 2 observed seasons in Site A underscores the potential of in loco isolation of industrial strains. Invading strains have been documented to cause harm, but they also served as the source for most if not all of the currently used strains in the industry (Basso *et al.* 2008; Lopes *et al.* 2015; Jacobus, Gross, *et al.* 2021a). Previous studies had shown that these known bioethanol strains are phylogenetically related and harbor genomic signals of domestication, some of which are shared with wine strains and others that are specific to bioethanol strains (Jacobus, Stephens, *et al.* 2021b). These strains also cluster very far apart from known natural *S. cerevisiae* isolates from other Brazilian biomes, further suggesting a nonnatural origin (Barbosa *et al.* 2016, 2018). Our results show that currently invading strains in Site B are closely related to these known domesticated bioethanol strains. On top of that, we note that the dominant strains across all sites and years are largely triploid, suggesting a systematic advantage of higher ploidy in this industrial environment (Supplementary Fig. 6). Taken all together, we hypothesize that the same patterns hold in most strain invasion events in bioethanol plants that follow a process similar to Sites A and B (Fig. 1a). The observed large genetic diversity among invading strains should be further explored as a potential resource for future strain isolation. Strain tracking as carried out in the current study is thus not only a useful process-monitoring tool but also a productive assistive strategy for the selection of novel and locally adapted industrial strains. For this purpose, industrial plants should have protocols in place for the isolation of invading strains, record keeping of associated fermentation metrics, and subsequent testing in blocked off portions of the industrial pipeline and scaled-down systems that mimic the industrial process (Raghavendran *et al.* 2017).

Our study used metagenomics and a newly developed framework to extract individual lineages to illuminate the yeast population dynamics in industrial sugarcane-based bioethanol production, with the goal of finding routes toward more consistent fermentation performance. The resolution obtained with this approach surpasses by far previously described and utilized methods, such as chromosomal karyotyping and PCR-based methods. Our approach also requires less clonal picking effort than these methods, as corroborated by inference on rarefied clonal data

(see Supplementary material). We observed that over 2 sampled production periods in 2 independent industrial units, the yeast population dynamics varied more dramatically between units than between years. In one site, we observed dominance and persistence of an in-house strain in both years, whereas in the other site, foreign strains invaded the process and displaced the starter strain used to initiate the production period. The several individual clones sequenced, including invading strains, are phylogenetically grouped with other known bioethanol strains, producing strong evidence that the invading strains originate from the sugarcane environment itself, and not from natural niches. The data presented, as well as the statistical framework developed, represent useful material for future investigations on sugarcane biorefineries (as well as other microbial communities of mixed ploidy). This, in turn, might lead us to a deeper understanding of the yeast and other microbial ecology in this peculiar environment, opening the way for process improvements, decreased consumption of costly chemicals, and increased ethanol yields. A potential new paradigm of industrial practice includes the design of synergistic yeast strain mixes and the inoculation of beneficial (or probiotic) bacteria in the process.

Data availability

Clonal isolates are available upon request. The Supplementary material contains a detailed description of the lineage inference pipeline, as well as all Supplementary Figures. Supplementary File 1 shows the allele frequency and coverage along the genome for all clonal isolates. Supplementary Files 2 and 3 contain the Newick format data for trees in Figs. 2a and 6a. Supplementary Tables 1–4 have information on sampled fermentation timepoints, clonal isolates, and Site B fermentation metrics. Raw sequencing reads for clonal and metagenomic samples have been deposited in the NCBI BioProject database under accession number PRJNA865262. Code for the variant calling pipeline, lineage inference, and figure generation, as well as parsed called variant data for clonal and metagenomic samples, can be found in the GitHub repository (https://github.com/arturrc/bioethanol_inference). Supplementary Material available at figshare: https://doi.org/10.25387/g3.22640890.

Acknowledgments

We thank the various members of Site A and Site B who provided extensive support and expertise along the way, specially Thaís Granço and Luciano Zamberlan. We thank the Bauer Core facility at Harvard for assistance with sequencing.

Funding

M.M.D. acknowledges support from grant PHY-1914916 from the National Science Foundation (NSF). A.K.G. acknowledges support from the Harvard Lemann Brazil Research Fund and from grant 2018/04962-5 from the São Paulo Research Foundation (FAPESP). The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Conflicts of interest

The author(s) declare no conflict of interest.

Author contributions

M.M.D. and A.K.G. designed the project; A.K.G. sequenced samples; A.R.-C and I.H. developed inference methods; A.R.-C., I.H., and A.K.G. analyzed the data; and A.R.-C., I.H., M.M.D., and A.K.G. wrote the paper.

Literature cited

Amorim HV, Lopes ML, de Castro Oliveira JV, Buckeridge MS, Goldman GH. Scientific challenges of bioethanol production in Brazil. Appl Microbiol Biotechnol. 2011;91(5):1267–1275. doi:10.1007/s00253-011-3437-6.

Antonangelo ATBF, Alonso DP, Ribolla PEM, Colombi D. Microsatellite marker-based assessment of the biodiversity of native bioethanol yeast strains: microsatellite assessment of native bioethanol yeast strains. Yeast 2013;30(8):307–317. doi:10.1002/yea.2964.

Anyansi C, Straub TJ, Manson AL, Earl AM, Abeel T. Computational methods for strain-level microbial detection in colony and metagenome sequencing data. Front Microbiol. 2020;11:1925. doi:10.3389/fmicb.2020.01925.

Argueso JL, Carazzolle MF, Mieczkowski PA, Duarte FM, Netto OVC, Missawa SK, Galzerani F, Costa GGL, Vidal RO, Noronha MF, *et al.* Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. Genome Res. 2009;19(12):2258–2270. doi:10.1101/gr.091777.109.

Barbosa R, Almeida P, Safar SVB, Santos RO, Morais PB, Nielly-Thibault L, Leducq J-B, Landry CR, Gonçalves P, Rosa CA, *et al.* Evidence of natural hybridization in Brazilian wild lineages of *Saccharomyces cerevisiae*. Genome Biol Evol. 2016;8(2):317–329. doi:10.1093/gbe/evv263.

Barbosa R, Pontes A, Santos RO, Montandon GG, de Ponzzes-Gomes CM, Morais PB, Gonçalves P, Rosa CA, Sampaio JP. Multiple rounds of artificial selection promote microbe secondary domestication—the case of cachaça yeasts. Genome Biol Evol. 2018;10(8):1939–1955. doi:10.1093/gbe/evy132.

Barrick JE, Lenski RE. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. Cold Spring Harb Symp Quant Biol. 2009;74(0):119–129. doi:10.1101/sqb.2009.74.018.

Barros S. Biofuels Annual. Brazil: United States Department of Agriculture; 2022.

Basso LC. Dominância das Leveduras Contaminantes Sobre as Linhagens Industriais Avaliada Pela Técnica da Cariotipagem. Águas de São Pedro, SP, Brazil: V Congresso Nacional da STAB; 1993.

Basso LC, de Amorim HV, de Oliveira AJ, Lopes ML. Yeast selection for fuel ethanol production in Brazil. FEMS Yeast Res. 2008;8(7):1155–1163. doi:10.1111/j.1567-1364.2008.00428.x.

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony RK. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One 2015;10(5):e0128036. doi:10.1371/journal.pone.0128036.

Bermejo PM, Badino A, Zamberlan L, Raghavendran V, Basso TO, Gombert AK. Ethanol yield calculations in biorefineries. FEMS Yeast Res. 2021;21(8):foab065. doi:10.1093/femsyr/foab065.

Carvalho-Netto OV, Carazzolle MF, Rodrigues A, Bragança WO, Costa GGL, Argueso JL, Pereira GAG. A simple and effective set of PCR-based molecular markers for the monitoring of the *Saccharomyces cerevisiae* cell population during bioethanol fermentation. J Biotechnol. 2013;168(4):701–709. doi:10.1016/j.jbiotec.2013.08.025.

Costea PI, Coelho LP, Sunagawa S, Munch R, Huerta-Cepas J, Forslund K, Hildebrand F, Kushugulova A, Zeller G, Bork P. Subspecies in the global human gut microbiome. Mol Syst Biol. 2017;13(12):960. doi:10.15252/msb.20177589.

Crago CL, Khanna M, Barton J, Giuliani E, Amaral W. Competitiveness of Brazilian sugarcane ethanol compared to US corn ethanol. Energy Policy 2010;38(11):7404–7415. doi:10.1016/j.enpol.2010.08.016.

da Silva-Filho EA, dos Santos SKB, do Resende AM, de Morais JOF, de Morais MA, Simões DA. Yeast population dynamics of industrial fuel-ethanol fermentation process assessed by PCR-fingerprinting. Antonie Van Leeuwenhoek. 2005;88(1):13–23. doi:10.1007/s10482-005-7283-3.

da Silva Filho EA, de Melo HF, Antunes DF, dos Santos SKB, do Resende AM, Simões DA, de Morais MA Jr. Isolation by genetic and physiological characteristics of a fuel-ethanol fermentative *Saccharomyces cerevisiae* strain with potential for genetic manipulation. J Ind Microbiol Biotechnol. 2005;32(10):481–486. doi:10.1007/s10295-005-0027-6.

Della-Bianca BE, Basso TO, Stambuk BU, Basso LC, Gombert AK. What do we know about the yeast strains from the Brazilian fuel ethanol industry? Appl Microbiol Biotechnol. 2013;97(3):979–991. doi:10.1007/s00253-012-4631-x.

Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–550. doi:10.1038/nrmicro2832.

Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci. 2015;112(22), E2930-E2938. doi:10.1073/pnas.1423854112.

Frenkel EM, McDonald MJ, Van Dyken JD, Kosheleva K, Lang GI, Desai MM. Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. Proc Natl Acad Sci. 2015;112(36):11306–11311. doi:10.1073/pnas.1506184112.

Gallone B, Steensels J, Prahl T, Soriaga L, Saels V, Herrera-Malaver B, Merlevede A, Roncoroni M, Voordeckers K, Miraglia L, *et al.* Domestication and divergence of *Saccharomyces cerevisiae* beer yeasts. Cell. 2016;166(6):1397–1410.e16. doi:10.1016/j.cell.2016.08.020.

Garud NR, Good BH, Hallatschek O, Pollard KS. Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. Gordo I, editor. PLoS Biol. 2019;17(1):e3000102. doi:10.1371/journal.pbio.3000102.

Gaspar JM. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. BMC Bioinformatics 2018;19(1):536. doi:10.1186/s12859-018-2579-2.

Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature 2017;551(7678):45–50. doi:10.1038/nature24287.

Jacobus AP, Gross J, Evans JH, Ceccato-Antonini SR, Gombert AK. *Saccharomyces cerevisiae* strains used industrially for bioethanol production. Essays Biochem. 2021a;65(2):147–161. doi:10.1042/EBC20200160.

Jacobus AP, Stephens TG, Youssef P, González-Pech R, Ciccotosto-Camp MM, Dougan KE, Chen Y, Basso LC, Frazzon J, Chan CX, *et al.* Comparative genomics supports that Brazilian bioethanol *Saccharomyces cerevisiae* comprise a unified group of domesticated strains related to cachaça spirit yeasts. Front Microbiol. 2021b;12:644089. doi:10.3389/fmicb.2021.644089.

Johnson *Cetal* 2015. High-Octane Mid-Level Ethanol Blend Market Assessment (NREL/TP-5400-63698). Golden, CO: National Renewable Energy Laboratory, U.S. Department of Energy. [accessed 2022 September 25]. https://doi.org/10.2172/1351596.

Johnson MS, Gopalakrishnan S, Goyal J, Dillingham ME, Bakerlee CW, Humphrey PT, Jagdish T, Jerison ER, Kosheleva K, Lawrence KR, *et al.* Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. eLife 2021;10:e63910. doi:10.7554/eLife.63910.

Lang GI, Murray AW. Estimating the per-base-pair mutation rate in the yeast *Saccharomyces cerevisiae*. Genetics 2008;178(1):67–82. doi:10.1534/genetics.107.071506.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 2014;15(1):162. doi:10.1186/1471-2164-15-162.

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature 2015;519(7542):181–186. doi:10.1038/nature14279.

Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Lino FSO, Misiakou M-A, Kang K, Li SS, da Costa BLV, Basso TO, Panagiotou G, Sommer MOA. 2021. Strain dynamics of specific contaminant bacteria modulate the performance of ethanol biorefineries. bioRxiv 430133. https://doi.org/10.1101/2021.02.07.430133. [accessed 2022 October 14]. http://biorxiv.org/lookup/doi/10.1101/2021.02.07.430133.

Lopes M, Paulillo SC, Cherubin R, Godoy A, Amorim Neto H, Amorim H. Tailored Yeast Strains for Ethanol Production: Process-Driven Selection. Piracicaba: Fermentec Sugar and Alcohol Technologies Ltd; 2015.

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. Constrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015;33(10):1045–1052. doi:10.1038/nbt.3319.

Maddamsetti R, Lenski RE, Barrick JE. Adaptation, clonal interference, and frequency-dependent interactions in a long-term evolution experiment with escherichia coli. Genetics 2015;200(2):619–631. doi:10.1534/genetics.115.176677.

Mitri S, Richard Foster K. The genotypic view of social interactions in microbial communities. Annu Rev Genet. 2013;47(1):247–273. doi:10.1146/annurev-genet-111212-133307.

Nagamatsu ST, Teixeira GS, de Mello FD, Tizei PA, Nakagawa BT, de Carvalho LM, Pereira GA, Carazzolle MF. Genome assembly of a highly aldehyde-resistant *Saccharomyces cerevisiae* SA1-derived industrial strain. Microbiol Resour Announc. 2019;8(13):e00071-19. doi:10.1128/MRA.00071-19.

Nguyen Ba AN, Cvijović I, Rojas Echenique JI, Lawrence KR, Rego-Costa A, Liu X, Levy SF, Desai MM. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature 2019;575(7783):494–499. doi:10.1038/s41586-019-1749-3.

Pereira LG, Cavalett O, Bonomi A, Zhang Y, Warner E, Chum HL. Comparison of biofuel life-cycle GHG emissions assessment tools: the case studies of ethanol produced from sugarcane, corn, and wheat. Renew Sustain Energy Rev. 2019;110:1–12. doi:10.1016/j.rser.2019.04.043.

Peter J, De Chiara M, Friedrich A, Yue JX, Pflieger D, Bergström A, Sigwalt A, Barre B, Freel K, Llored A, *et al.* Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature 2018;556(7701):339–344. doi:10.1038/s41586-018-0030-5.

Rácz HV, Mukhtar F, Imre A, Rádai Z, Gombert AK, Rátonyi T, Nagy J, Pócsi I, Pfliegler WP. How to characterize a strain? Clonal heterogeneity in industrial *Saccharomyces* influences both phenotypes and heterogeneity in phenotypes. Yeast 2021;38(8):453–470. doi:10.1002/yea.3562.

Raghavendran V, Basso TP, da Silva JB, Basso LC, Gombert AK. A simple scaled down system to mimic the industrial production of first generation fuel ethanol in Brazil. Antonie Van Leeuwenhoek. 2017;110(7):971–983. doi:10.1007/s10482-017-0868-9.

Reis VR, Antonangelo ATBF, Bassi APG, Colombi D, Ceccato-Antonini SR. Bioethanol strains of *Saccharomyces cerevisiae* characterised by microsatellite and stress resistance. Braz J Microbiol. 2017;48(2):268–274. doi:10.1016/j.bjm.2016.09.017.

Rich JO, Bischoff KM, Leathers TD, Anderson AM, Liu S, Skory CD. Resolving bacterial contamination of fuel ethanol fermentations with beneficial bacteria—an alternative to antibiotic treatment. Bioresour Technol. 2018;247:357–362. doi:10.1016/j.biortech.2017.09.067.

Roodgar M, Good BH, Garud NR, Martis S, Avula M, Zhou W, Lancaster SM, Lee H, Babveyh A, Nesamoney S, *et al.* Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. Genome Res. 2021;31(8):1433–1446. doi:10.1101/gr.265058.120.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, *et al.* Genomic variation landscape of the human gut microbiome. Nature 2013;493(7430):45–50. doi:10.1038/nature11711.

Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, Truong DT, Tett A, Morrow AL, Segata N. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435–438. doi:10.1038/nmeth.3802.

Senne de Oliveira Lino F, Bajic D, Vila JCC, Sánchez A, Sommer MOA. Complex yeast–bacteria interactions affect the yield of industrial ethanol fermentation. Nat Commun. 2021;12(1):1498. doi:10.1038/s41467-021-21844-7.

Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, *et al.* Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal Microbiota transplantation. Cell Host Microbe 2018;23(2):229–240.e5. doi:10.1016/j.chom.2018.01.003.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–1313. doi:10.1093/bioinformatics/btu033.

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27(4):626–638. doi:10.1101/gr.216242.116.

van der Auwera G, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. Sebastopol, CA: O'Reilly Media; 2020.

West C, James SA, Davey RP, Dicks J, Roberts IN. Ribosomal DNA sequence heterogeneity reflects intraspecies phylogenies and predicts genome structure in two contrasting yeast species. Syst Biol. 2014;63(4):543–554. doi:10.1093/sysbio/syu019.

Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, Xavier RJ, Alm EJ. Adaptive evolution within gut microbiomes of healthy people. Cell Host Microbe. 2019;25(5):656–667.e8. doi:10.1016/j.chom.2019.03.007.

Editor: R. Kulathinal