Yeast population dynamics in Brazilian bioethanol production

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

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA 02138,

[2]Department of Physics, Harvard University, Cambridge, MA 02138,

[3]NSF-Simons Center for Mathematical and Statistical Analysis of Biology, Harvard University, Cambridge MA 02138,

[4]Quantitative Biology Initiative, Harvard University, Cambridge MA 02138,

[5]School of Food Engineering, University of Campinas, Rua Monteiro Lobato 80, 13083-862, Campinas, SP, Brazil.

*Both authors contributed equally to this work.

†Correspondence should be directed Andreas K. Gombert. Email: gombert@unicamp.br. Mailing address: School of Food Engineering, University of Campinas, Rua Monteiro Lobato 80, 13083-862, Campinas, SP, Brazil.

Running title: Yeast dynamics in bioethanol production

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

Abstract

The large scale and non-aseptic 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 two different biorefineries through two 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 one biorefinery is small compared to the differences between the two units. In one biorefinery, all lineages present during the entire production period derive from one 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 non-aseptic 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.

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 one 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 *S. 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 non-aseptic 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 kickstart the fermentation season, but due to its susceptibility to invasion by foreign *S. cerevisiae* lineages, production has largely shifted towards 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, 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 datasets 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. Non-haploidy 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, non-haploid, 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 two fermentation seasons, in two 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

93 investigate the differences between seasons and production facilities, the origin of invading strains, and the
94 effects they have on the process. Our focus here is on the yeast dynamics, but our sequencing data also contains
95 information on other microbial species, which remains to be analyzed in future work.

METHODS

Sample collection

We collected whole-population microbiological samples from two independent industrial units, which we refer to as *Site A* and *Site B*, through two 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 Table S1). The two 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 four strains to start both the 2018 and 2019 fermentation periods—namely strains PE-2, SA-1, FT-858, and IRA-D. While the first three 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 pre-sterilized 15 ml tubes containing 3 ml glycerol. After mixing by vortexing, samples were stored at –20°C for a period of between one and three months before being transferred to a –80°C ultrafreezer. Finally, samples were shipped from Brazil to the US 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 Table S1. Picked clonal isolates are made available upon request.

DNA extraction and sequencing

We selected 15 to 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 to 20 thawed and homogenized samples were used for plating onto Yeast Extract-Peptone-Dextrose(YPD)-agar (Table S2). 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 (Table S3).

Sequencing was carried out in two Illumina NextSeq and one 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 87x for metagenomic samples and 26x for clonal isolates. FASTQ files with all sequencing reads produced for this study were deposited in the NCBI SRA database (see Data and Code 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 and Code 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 one sample.

Given different probabilistic prior models of allele frequency for clonal and non-clonal 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 VCFs) 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 one 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 File S1): 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 Fig. S8 in Supplementary Information.

Phylogenetic analyses

We infer two 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 four starter strains (Newick format tree in File S2). *Tree 2* includes the same clonal isolates, plus all isolates from the 1011 Yeast Genomes Project (Peter et al. 2018; Fig. S10 in Supplementary Information; Newick format tree in File S3). 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 1000 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 rerooted 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 three origins: (1) preexisting diversity in starting inoculum; (2) invading strains during the course of the fermentation season; (3) 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 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 Information, together with a validation analysis using subsampled clonal data. The code developed for this inference is available in the GitHub repository (see Data and Code Availability).

RESULTS

We carried out temporal whole-population metagenome sequencing of the *S. cerevisiae* populations used to ferment sugarcane products into bioethanol over two fermentation seasons (2018 and 2019), at two 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. 14,200 (9.8%) of these mutations are singletons, while 15,749 (10.5%) are seen in all sequenced clones (see Fig. S7 in Supplementary Information for the full distribution). We also find a similar number of SNPs (150,265) in the whole-population metagenome data across all four site-years, with an overlap of 126,845 between the clonal and the metagenomic datasets. 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 6154x over all timepoints, which is larger than that of clonal genotyping (4,341x over all isolates). The 168,486 SNPs uncovered in the whole dataset are widely distributed along the genome, hitting 6,370 out of all 6,579 genes in the annotated S288c genome. 129,697 of these SNPs have been previously observed in the 1011 yeast genomes project, 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, and 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 (File S1). 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 two 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 the unknown other 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 one to three copies in triploids). Finally, it is not immediately clear how to polarize mutations for lineage frequency inference (i.e. which one should be considered

the references versus 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 Methods and Supplementary Information 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 and colleagues (2019), while handling high genetic diversity and ploidy variation in the population.

Among the four 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 Fig. S9 in Supplementary Information 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 four 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 Fig. 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 strains in Fig. 2A and 5). Most of these strains are triploid, except for a small group present in both years (Fig. 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 1011 Yeast Genomes Project (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. 2021).

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. 2021). As shown in Fig. 6B, we note that among the 37 isolates classified in the Brazilian bioethanol group in the 1011 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 1011 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 (Basso et al. 2008; in terms of the final ethanol yield on sugars). Thus, invasion by unknown strains may harm the fermentation process and the

317 profitability of the industry, due to decreased ethanol production and/or to higher costs involved with the use of
318 chemicals, such as sulfuric acid, antimicrobials, antifoaming agents and dispersants. In the case of Site B's 2018
319 and 2019 seasons, we have not found a connection between general industrial metrics and inferred events of
320 population substitution (Fig. S11 in Supplementary Information). Nonetheless, it may still be possible that this
321 stability was accomplished by the employment of commonly used but costly corrective measures, such as those
322 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 two fermentation seasons (2018 and 2019) in two 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 not only observed in invading strains, but also within the starter strains themselves, whose same subtypes are sampled across years and sites (most notably the two 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 one 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 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 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 two 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. 2021). Previous studies had shown that these known bioethanol strains are phylogenetically related and harbor genomic signals of domestication, some which are shared with wine strains and others that are specific to bioethanol strains (Jacobus, Stephens, et al. 2021). These strains also cluster very far apart known natural *S. cerevisiae* isolates from other Brazilian biomes, further suggesting a non-natural origin (Barbosa et al. 2016; Barbosa et al. 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 (Fig. S6 in Supplementary Information). Taken all together, we hypothesize that the same patterns hold in most strain invasion events in bioethanol plants that follow a process similar to Site 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 towards 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 Information). We observed that over two sampled production periods in two 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.

MATERIAL, DATA AND CODE AVAILABILITY

Clonal isolates are available upon request. The Supplementary Information contains a detailed description of the lineage inference pipeline, as well as all Supplementary Figures. File S1 shows the allele frequency and coverage along the genome for all clonal isolates. Files S2 and S3 contain the Newick format data for trees in Figs. 2A and 6A. Tables S1–S4 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).

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. A.R.-C., I.H., M.M.D., and A.K.G. wrote the paper.

ACKNOWLEDGMENTS

We thank the Bauer Core facility at Harvard for assistance with sequencing. M.M.D. acknowledges support from grant PHY-1914916 from the NSF. A.K.G. acknowledges support from the Harvard Lemann Brazil Research Fund, and from grant 2018/04962-5 from 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.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

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

439 Antonangelo ATBF, Alonso DP, Ribolla PEM, Colombi D. 2013. Microsatellite marker-based assessment of the
440 biodiversity of native bioethanol yeast strains: Microsatellite assessment of native bioethanol yeast strains. Yeast.
441 30(8):307–317. doi:10.1002/yea.2964.

442 Anyansi C, Straub TJ, Manson AL, Earl AM, Abeel T. 2020. Computational Methods for Strain-Level Microbial
443 Detection in Colony and Metagenome Sequencing Data. Front Microbiol. 11:1925. doi:10.3389/fmicb.2020.01925.

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

447 van der Auwera G, O'Connor BD. 2020. Genomics in the cloud: Using Docker, GATK, and WDL in Terra. 1st ed.
448 O'Reilly Media.

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

452 Barbosa R, Pontes A, Santos RO, Montandon GG, de Ponzzes-Gomes CM, Morais PB, Gonçalves P, Rosa CA,
453 Sampaio JP. 2018. Multiple Rounds of Artificial Selection Promote Microbe Secondary Domestication—The Case of
454 Cachaça Yeasts. Wolfe K, editor. Genome Biol Evol. 10(8):1939–1955. doi:10.1093/gbe/evy132.

455 Barrick JE, Lenski RE. 2009. Genome-wide Mutational Diversity in an Evolving Population of Escherichia coli. Cold
456 Spring Harb Symp Quant Biol. 74(0):119–129. doi:10.1101/sqb.2009.74.018.

457 Barros S. 2022. Biofuels Annual. Country: Brazil (BR2022-0047). United States Department of Agriculture.

458 Basso LC. 1993. Dominância das leveduras contaminantes sobre as linhagens industriais avaliada pela técnica da
459 cariotipagem. In: V Congresso Nacional da STAB.

460 Basso LC, de Amorim HV, de Oliveira AJ, Lopes ML. 2008. Yeast selection for fuel ethanol production in Brazil.
461 FEMS Yeast Res. 8(7):1155–1163.

462 Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony RK. 2015. Inexpensive multiplexed library
463 preparation for megabase-sized genomes. PLoS ONE. 10(5):1–15. doi:10.1371/journal.pone.0128036.

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

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

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

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

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

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

Franzosa EA, Huang K, Meadow JF, Gevers D, Lemon KP, Bohannan BJM, Huttenhower C. 2015. Identifying personal microbiomes using metagenomic codes. Proc Natl Acad Sci. 112(22). doi:10.1073/pnas.1423854112. [accessed 2022 Oct 19]. https://pnas.org/doi/full/10.1073/pnas.1423854112.

Frenkel EM, McDonald MJ, Van Dyken JD, Kosheleva K, Lang GI, Desai MM. 2015. Crowded growth leads to the spontaneous evolution of semistable coexistence in laboratory yeast populations. Proc Natl Acad Sci. 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. 2016. Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. Cell. 166(6):1397-1410.e16. doi:10.1016/j.cell.2016.08.020.

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

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

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

Jacobus AP, Gross J, Evans JH, Ceccato-Antonini SR, Gombert AK. 2021. *Saccharomyces cerevisiae* strains used industrially for bioethanol production. Mattanovich D, Ivan Nikel P, editors. Essays Biochem. 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. 2021. Comparative Genomics Supports That Brazilian Bioethanol *Saccharomyces cerevisiae* Comprise a Unified Group of Domesticated Strains Related to Cachaça Spirit Yeasts. Front Microbiol. 12:644089. doi:10.3389/fmicb.2021.644089.

Johnson C, Newes E, Brooker A, McCormick R, Peterson S, Leiby P, Martinez RU, Oladosu G, Brown ML. 2015. High-Octane Mid-Level Ethanol Blend Market Assessment (NREL/TP-5400-63698). National Renewable Energy Laboratory, U.S. Department of Energy. [accessed 2022 Sep 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. 2021. Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. eLife. 10:e63910. doi:10.7554/eLife.63910.

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

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

Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. 2015. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature. 519(7542):181–6. doi:10.1038/nature14279.

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

Lino FS de O, 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. doi:10.1101/2021.02.07.430133. [accessed 2022 Oct 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. 2015. Tailored yeast strains for ethanol production: process-driven selection. Piracicaba: Fermentec Sugar and Alcohol Technologies Ltd.

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

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

Mitri S, Richard Foster K. 2013. The Genotypic View of Social Interactions in Microbial Communities. Annu Rev Genet. 47(1):247–273. doi:10.1146/annurev-genet-111212-133307.

Nagamatsu ST, Teixeira GS, de Mello F da SB, Tizei PAG, Nakagawa BTG, de Carvalho LM, Pereira GAG, Carazzolle MF. 2019. Genome Assembly of a Highly Aldehyde-Resistant *Saccharomyces cerevisiae* SA1-Derived Industrial Strain. Cuomo C, editor. Microbiol Resour Announc. 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. 2019. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature. 575(7783):494–499. doi:10.1038/s41586-019-1749-3.

Pereira LG, Cavalett O, Bonomi A, Zhang Y, Warner E, Chum HL. 2019. Comparison of biofuel life-cycle GHG emissions assessment tools: The case studies of ethanol produced from sugarcane, corn, and wheat. Renew Sustain Energy Rev. 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. 2018. Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. Nature. 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. 2021. How to characterize a strain? Clonal heterogeneity in industrial *Saccharomyces* influences both phenotypes and heterogeneity in phenotypes. Yeast. 38(8):453–470. doi:10.1002/yea.3562.

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

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

Rich JO, Bischoff KM, Leathers TD, Anderson AM, Liu S, Skory CD. 2018. Resolving bacterial contamination of fuel ethanol fermentations with beneficial bacteria – An alternative to antibiotic treatment. Bioresour Technol. 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. 2021. Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. Genome Res. 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. 2013. Genomic variation landscape of the human gut microbiome. Nature. 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. 2016. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 13(5):435–438. doi:10.1038/nmeth.3802.

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

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

da Silva-Filho EA, Santos SKB dos, Resende A do M, de Morais JOF, de Morais MA, Simões DA. 2005. Yeast population dynamics of industrial fuel-ethanol fermentation process assessed by PCR-fingerprinting. Antonie Van Leeuwenhoek. 88(2):13–23. doi:10.1007/s10482-005-7283-3.

Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, Hohmann EL, Staley C, Khoruts A, Sadowsky MJ, et al. 2018. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host Microbe. 23(2):229-240.e5. doi:10.1016/j.chom.2018.01.003.

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

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

West C, James SA, Davey RP, Dicks J, Roberts IN. 2014. Ribosomal DNA Sequence Heterogeneity Reflects Intraspecies Phylogenies and Predicts Genome Structure in Two Contrasting Yeast Species. Syst Biol. 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. 2019. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host Microbe. 25(5):656-667.e8. doi:10.1016/j.chom.2019.03.007.

583 **Figure 1. Schematics of the fermentation process and sequencing strategy. (A)** A large population (~10^{17}
584 individuals) of the yeast *S. cerevisiae* is maintained over the course of an eight-month-long fermentation season.
585 Yeast ferments must, a mix of molasses, sugarcane juice and water, to produce ethanol in a fed-batch process
586 that takes ~8h and runs in a staggered parallel fashion across several fermentors (8–16 in any one plant, each with
587 a ~500,000 ℓ capacity). The fermented broth (wine) from different fermentors is loaded into a single holding tank,
588 which continuously feeds a centrifuge for separation of the yeast from the liquid fraction. Holding tanks are larger
589 than fermentors themselves and allow for mixing between batches. The yeast cells are then treated with
590 chemicals to control for bacterial growth and are later reused in the process. The yeast population grows by ~10%
591 every 12h, leading to approximately 66 generations over the course of an ~8 months fermentation season. The
592 season is started with selected industrial strains which are commercialized by yeast suppliers. **(B)** We collected
593 whole-population samples of the yeast used for fermentation through two seasons (2018 and 2019) in two plants
594 (Site A and Site B) located ~18 km apart in the state of São Paulo, Brazil. The two plants are owned by different
595 companies and use different sets of starter strains in their process. We employed a combination of whole-
596 population metagenome sequencing and clonal whole-genome sequencing to observe the temporal dynamics of
597 genetic diversity in each site-year. See Tables S1–3 for a complete list of collected samples and isolates.

598 **Figure 2. Yeast populations in bioethanol fermentors are genetically diverse and dynamic. (A)** Phylogenetic tree
599 of isolated clonal strains from all site-years, as well as known starter strains used. Most isolates are closely related
600 the known starter strains, but several are not. The tree was inferred with a maximum likelihood model using the
601 data of 27,229 SNPs. Ploidy of each isolate, assessed as described in the Methods, is indicated by diamonds.
602 Nodes and tips are colored as in Figs. 4 and 5. The tree is rooted in the same place as the independently inferred
603 tree in Fig. 6. Isolates are grouped as in Figs. 4–6. Isolates are named as *<site><year>:<timepoint>(<letter*
604 *identifier>)*, while starter strain isolates are marked with an asterisk. The associated Newick tree can be found in
605 File S2. The allele frequency data used for ploidy assessment can be visualized in File S1. Selected examples of a
606 diploid and triploid strain can be seen in Fig. S8 in Supplementary Information. **(B)** Frequency of alternate allele (in
607 relation to the reference genome of strain s288c) through time for an arbitrary subset of 2000 mutations (out of
608 ~100k) per site-year. Overall, mutation trajectories indicate alternation between periods of stasis, when one

609 major strain dominates, and periods of transition, when many mutations change in frequency in a correlated way
610 indicative of strain dynamics. Noise in mutation trajectories comes from random sampling (approximately
611 binomial), as well as sequencing and mapping errors, which is not homogeneous across mutations.

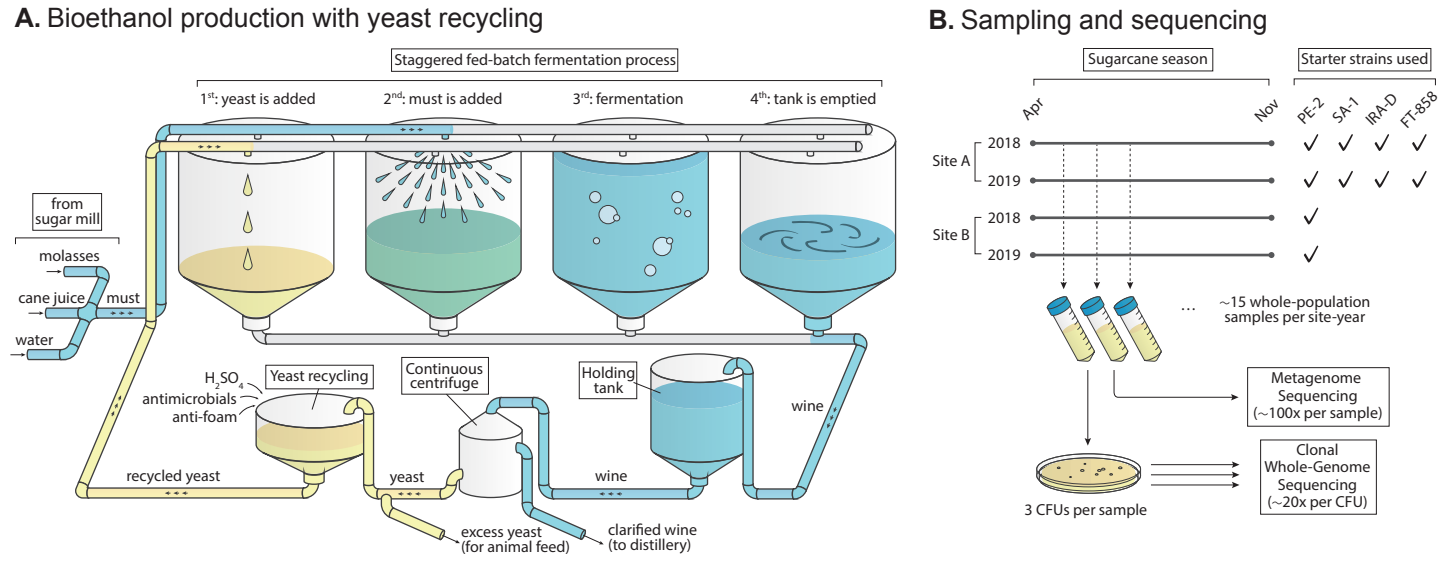
612 **Figure 3. Schematics of lineage inference procedure.** We use temporal metagenomics and clonal isolate whole-
613 genome sequencing to infer the unobserved frequencies of asexual lineages in the original population over the
614 course of a fermentation season. (Upper left) Starter, invading, and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

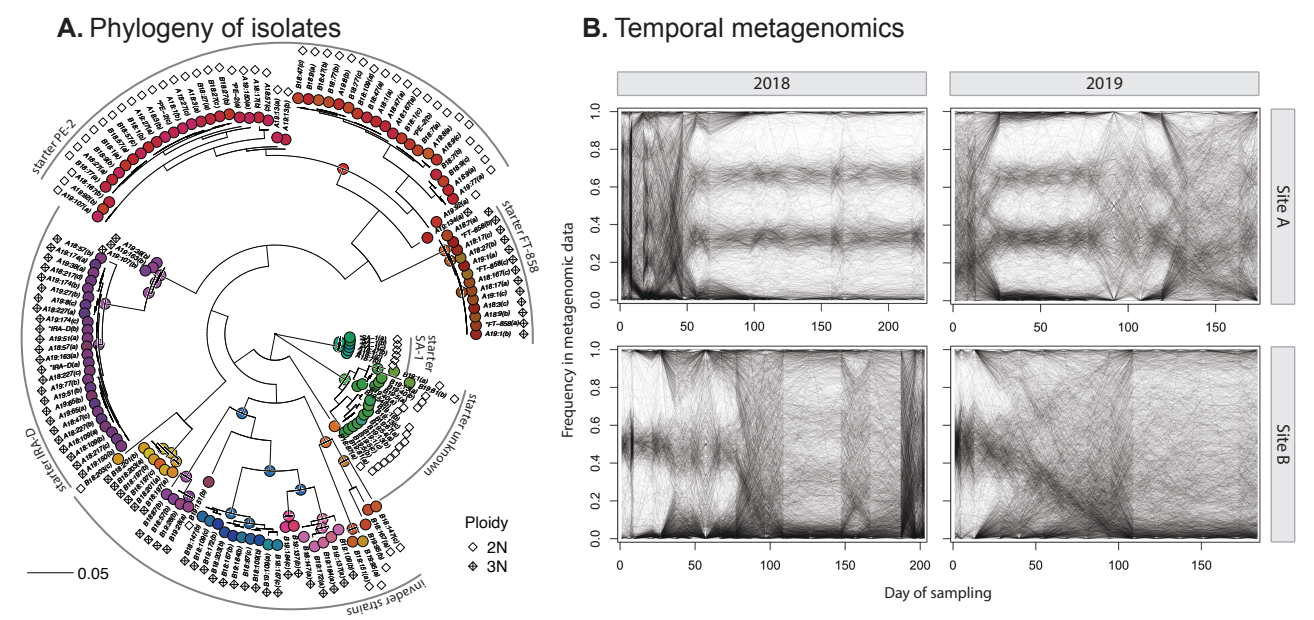
624 **Figure 4. In Site A the in-house starter strain IRA-D consistently dominates over other starter strains.** On the
625 left, inferred strain dynamics in Site A over the two fermentation seasons. White space corresponds to non-
626 inferred genetic diversity in the population. On the right, subtrees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

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

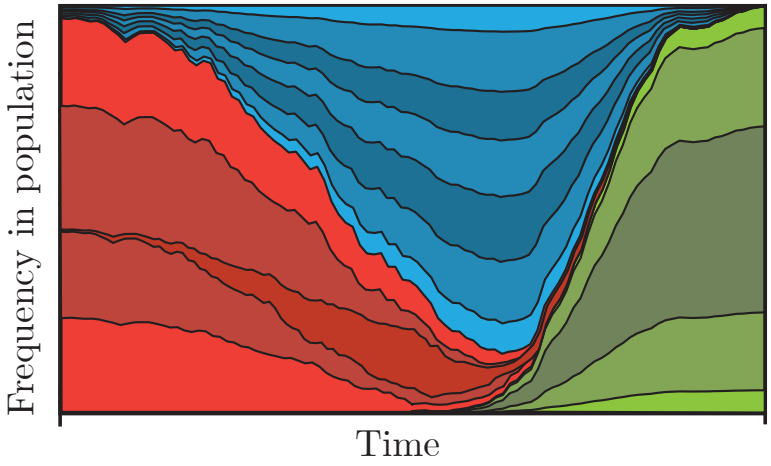
632 **Figure 6. Starter and invader isolates all cluster together within a larger group of Brazilian Bioethanol strains.**
633 **(A)** A SNP-based maximum likelihood phylogeny combining isolates from the current study and from the 1011

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 Fig. S10 in Supplementary Information, and its associated Newick tree can be found in File S3.


A. Bioethanol production with yeast recycling
Staggered fed-batch fermentation process
1st: yeast is added
2nd: must is added
3rd: fermentation
4th: tank is emptied
from sugar mill
molasses
cane juice
must
water
H2SO4
antimicrobials
anti-foam
Yeast recycling
Continuous centrifuge
Holding tank
wine
recycled yeast
yeast
wine
excess yeast (for animal feed)
clarified wine (to distillery)
B. Sampling and sequencing
Sugarcane season
Starter strains used
Apr
Nov
PE-2
SA-1
IRA-D
FT-858
Site A
2018
2019
Site B
2018
2019
~15 whole-population samples per site-year
Metagenome Sequencing (~100x per sample)
Clonal Whole-Genome Sequencing (~20x per CFU)
3 CFUs per sample


A. Phylogeny of isolates
starter PE-2
starter FT-858
starter SA-1
starter unknown
invader strains
starter IRA-D
Ploidy
2N
3N
0.05
B. Temporal metagenomics
2018
2019
Site A
Site B
Frequency in metagenomic data
Day of sampling

Unobserved asexual lineage dynamics



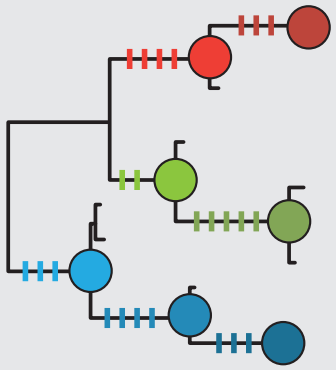
Generative model of metagenomic data

$$\mathcal{L}(\vec{f}(t)) = \prod_{l,m} \sum_{g_m} P(x_{lm}(t)|d_{lm}(t), g_m, f_l(t))P(g_m)$$

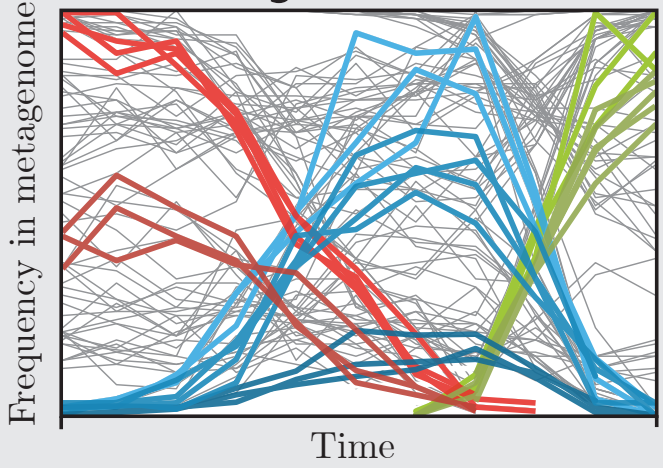
$$x_{l,m}(t) \sim \text{Binom}(d_{lm}(t), g_m f_l(t))$$

$$f_l(t) \geq \sum_{i \in C_l} f_i(t)\,, C_l := \text{the children of } l$$

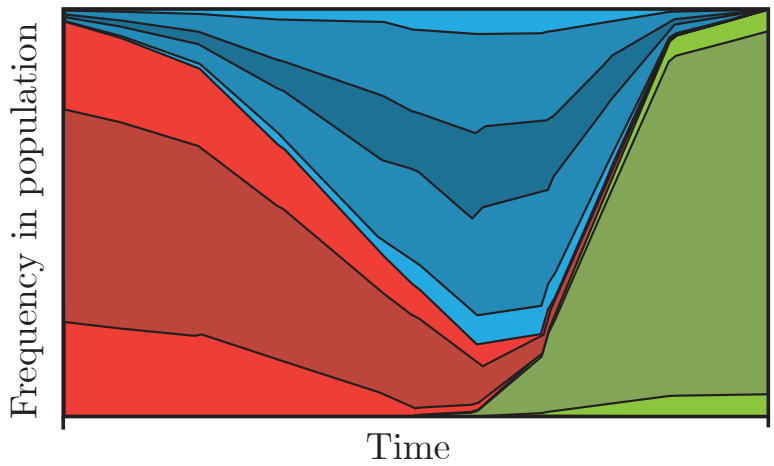
Clone phylogeny



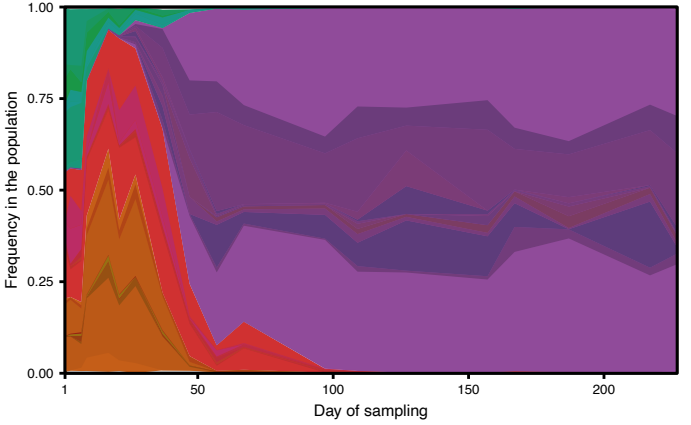
Metagenome data

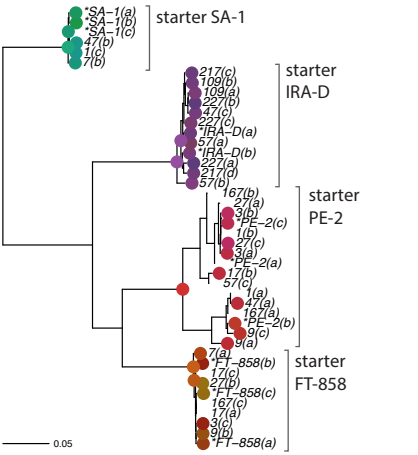


Inferred lineage dynamics

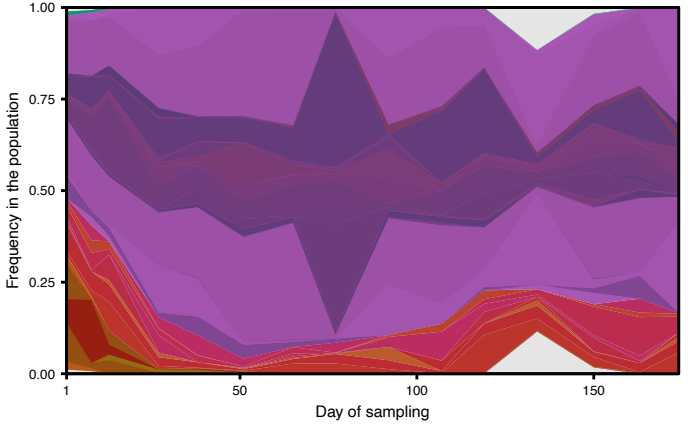


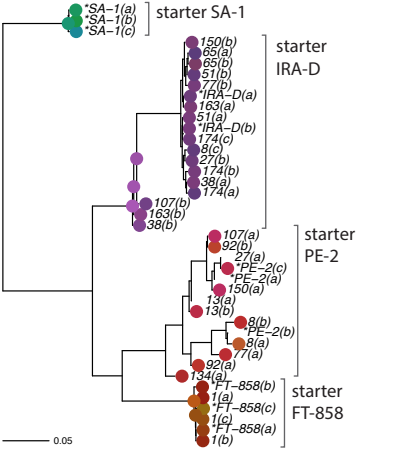
A. Site A - 2018

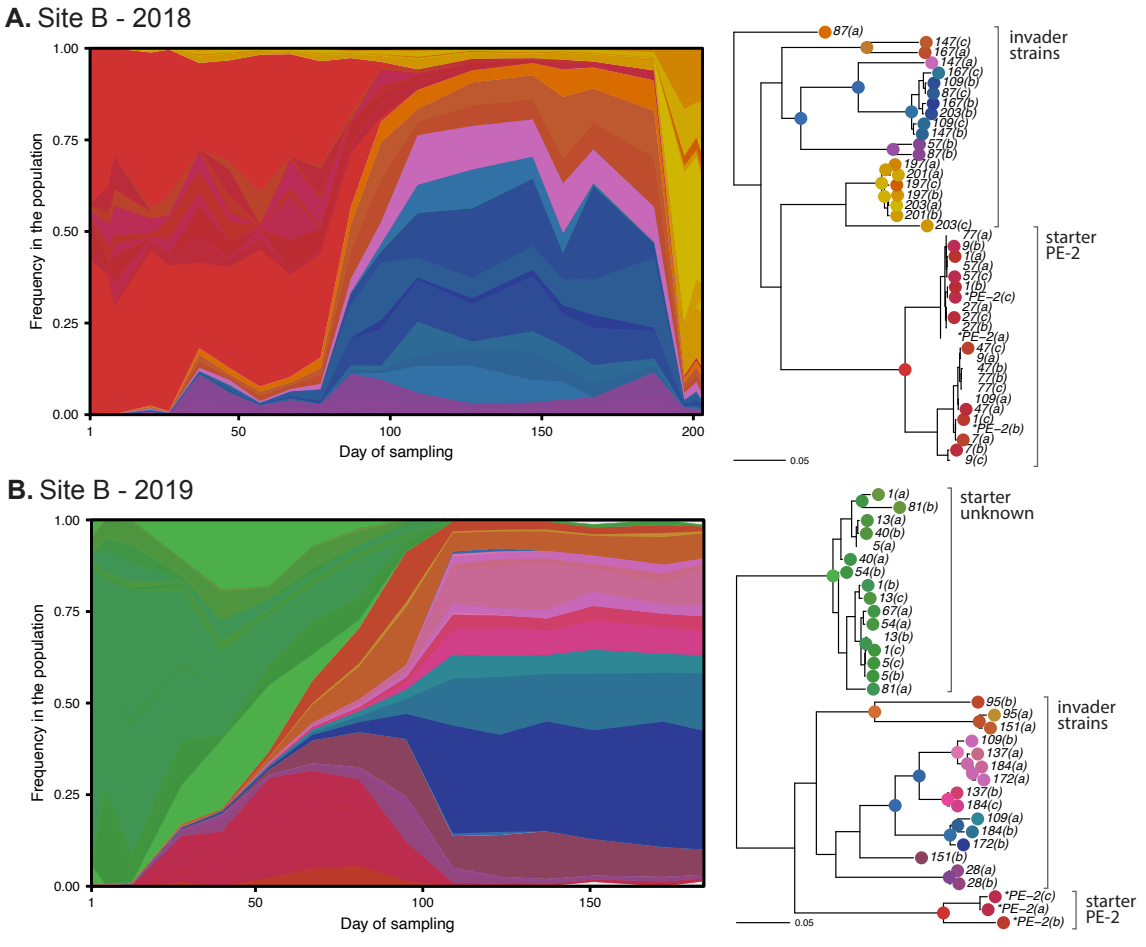


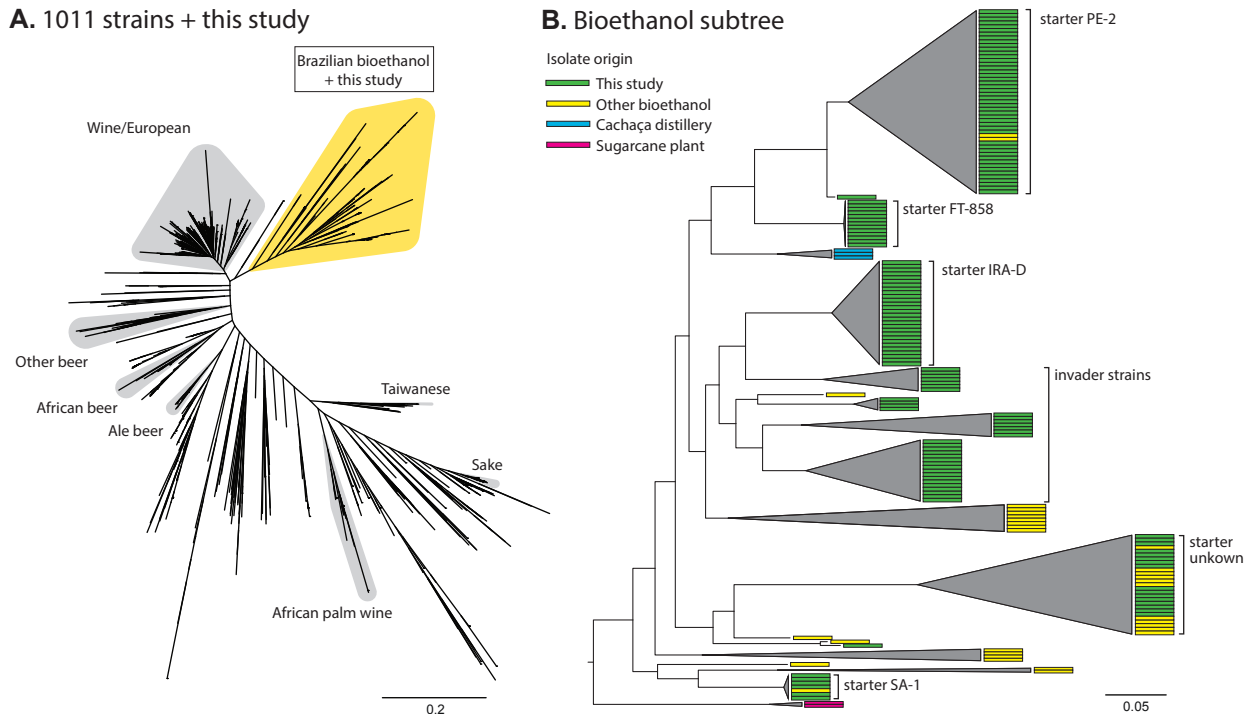


B. Site A - 2019






A. Site B - 2018
Frequency in the population
1.00
0.75
0.50
0.25
0.00
1
50
100
150
200
Day of sampling
87(a)
147(c)
167(a)
147(a)
167(c)
109(b)
87(c)
167(b)
203(b)
109(c)
147(b)
57(b)
87(b)
197(a)
201(a)
197(c)
197(b)
203(a)
201(b)
203(c)
77(a)
9(b)
1(a)
27(a)
57(c)
1(b)
*PE–2(c)
27(a)
27(c)
27(b)
*PE–2(a)
47(c)
9(a)
47(b)
77(b)
77(c)
109(a)
47(a)
1(c)
*PE–2(b)
7(a)
7(b)
9(c)
0.05
invader strains
starter PE-2
B. Site B - 2019
Frequency in the population
1.00
0.75
0.50
0.25
0.00
1
50
100
150
Day of sampling
1(a)
81(b)
13(a)
40(b)
5(a)
40(a)
54(b)
1(b)
13(c)
67(a)
54(a)
13(b)
1(c)
5(c)
5(b)
81(a)
95(b)
95(a)
151(a)
109(b)
137(a)
184(a)
172(a)
137(b)
184(c)
109(a)
184(b)
172(b)
151(b)
28(a)
28(b)
*PE–2(c)
*PE–2(a)
*PE–2(b)
0.05
starter unknown
invader strains
starter PE-2


A. 1011 strains + this study
Brazilian bioethanol + this study
Wine/European
Other beer
African beer
Ale beer
Taiwanese
Sake
African palm wine
0.2
B. Bioethanol subtree
Isolate origin
This study
Other bioethanol
Cachaça distillery
Sugarcane plant
starter PE-2
starter FT-858
starter IRA-D
invader strains
starter unkown
starter SA-1
0.05

Supplementary Information

INFERENCE OF POPULATION DYNAMICS

As described in the Methods, we assume that the population at each site-year is composed of a large but finite number of clonal strains which are related by some phylogenetic history in a tree-like manner. Clades in this tree represent lineages of descent from a common ancestor and is what we will be referring to as *lineages* throughout the text.

Our goal here is to (i) use the whole-genome clonal isolate data to identify as many as possible sets of lineage-defining synapomorphic alleles, and (ii) use the metagenomic frequencies of these synapomorphic alleles to infer their respective lineage frequencies in the population through the course of a fermentation season. By doing this, we ignore correlation between mutations in the metagenomic data as signal of coinheritance, something that has been previously done in literature [refs]. The advantage of following this route is higher power to identify low-frequency lineages, whose mutations' metagenomic trajectories would be too overpowered by noise to ever have a significant correlation signal (although our ability to identify these low-frequency lineages is still ultimately limited by the clonal isolate sampling).

In spirit, we follow a strategy similar to that of [Tami's paper], with the important difference that our populations are highly diverse and non-haploid. The consequence is that a large number of mutations will be unsuitable for inference, either because they are not monophyletically shared in the inferred phylogeny, or because their genotype (i.e. number of allele copies within an isolate's genome) varies among isolates that carry it, thus complicating the mathematical relationship between lineage frequency in the population and allele frequency in the metagenomic data.

Lineage assignment

We will define lineages as monophyletic clades in the phylogenetic tree inferred for all isolates from our experiment, which is in principle an unrooted tree (Fig. 2A). Since we observe (in a second inferred tree of all our isolates and those from the 1011 genomes project; see Methods for details) that all Brazilian bioethanol isolates cluster together, and within that cluster the SA-1 isolates are the most basal among our isolates (Fig. 6B), we root that first tree of isolates in the analogue node (as shown in Fig. 2A). From this rerooted tree, we define all lineages, (i) which include the very base of the tree with all isolates in the experiment, (ii) all internal nodes and their respective descendant isolates, and (iii) each tip with its associated isolate. Note that since this tree is inferred from isolates, it is most likely undersampling the genetic diversity of the population. Some lineages, especially the smaller ones, will most likely be missed (as illustrated in Fig. 3).

Finding lineage-specific alleles

For each one of the lineages defined above, we first would like to find a set of alleles that are unique to it. We cannot assess all individuals in the original population, and so instead we use the observed alternate allele counts and depth of coverage at variant sites in the clonal isolate data as a proxy. Therefore, for each lineage, we first flag all variant sites for which either (i) counts in all lineage members are larger than zero, while counts in non-lineage members are zero, or (ii) counts in all lineage members are less than the depth, while counts in all non-lineage members equal the depth. The second case covers variant sites for which the reference allele is the derived (synapomorphic) one in the phylogeny. For these mutations, in all analyses described below, *counts* will refer to the count of reference allele (instead of alternate allele).

If the lineage under consideration has a single isolate, then all flagged mutations are kept. Otherwise, we must select only those mutations for which we believe all isolates in the lineage to have the same genotype. For a diploid strain, the genotype of the mutation m in isolate i takes values $g_{mi} \in \{0, {}^1\!/_2, 1\}$, while for a triploid strain, $g_{mi} \in \{0, {}^1\!/_3, {}^2\!/_3, 1\}$. For this reason, we exclude from further analyses any lineages composed of a mix of diploid and triploid isolates. For each of the mutations flagged for a lineage we apply a statistical test of genotype heterogeneity, explained in more detail in the section below, where the null hypothesis is that all isolates in the lineage carry that mutation at the same genotype. We then use a procedure similar to Benjamini-Hochberg to select mutations for which we do not reject the null at a False Omission Rate of 0.05 (defined as false negatives/[false negatives + true negatives]).

We apply some filters before arriving at a final list of lineages and mutations for later frequency inference. First, we only keep those mutations that we also observe in the metagenomic dataset. Second, we limit the total number of mutations in a lineage to 500 to keep later steps computationally tractable. When this limit is imposed, mutations are chosen arbitrarily. Third, we filter mutations based on their observed depths in the metagenomic dataset, as they suggest underlying read mapping issues: we remove any mutations that have median depth in the metagenomic data lower than 10, or that has any metagenomic timepoint with depth equal to 0. Finally, we exclude any lineages for which we have selected 3 or less mutations, as we have observed that to result in noisy frequency inference.

Genotype heterogeneity test

As described in the section above, we would like to test whether a mutation is carried at the same genotype across all isolates from a lineage. For that we do a chi-squared test of goodness of fit to the model that all isolates have the same genotype.

Let a_{mi} and b_{mi} be the counts and depths of mutation m in isolate i. We first would like to define a generative model for the data so that we can compute the likelihood $P(a_{mi}|b_{mi}g_{mi})$. We choose a simple approach that assumes that a_{mi} is largely binomially distributed, except for a small probability of random errors, which can shift the count a_{mi} upwards or downwards. These errors may come from any of the preceding steps in data generation and analysis (e.g. sequencing and mapping errors), and they need to be accounted for the correct genotyping of homozygous sites that show a small (erroneous) count towards the opposite allele. We assume that the observed count a_{mi} is the result of a mixture of two populations of reads observed at site i: *true* and *error* reads. The b^T_{mi} true reads contribute with an alternate allele count $a^T_{mi} \sim \mathrm{Binom}\left(b^T_{mi}, g_{mi}\right)$, while the b^E_{mi} error reads contribute with an alternate allele count $a^E_{mi} \sim \mathrm{Binom}\left(b^E_{mi}, 0.5\right)$. We further assume that error reads are independent of each other and occur with equal probability p_{error}, such that $b^E_{mi} \sim \mathrm{Binom}(b_{mi}, p_{error})$. Since b^E_{mi} and a^E_{mi} are unobserved quantities, we marginalize over their possible values, and thus

$$P(a_{mi}|b_{mi}g_{mi}) = \sum_{b^E_{mi}=0}^{b_{mi}} \sum_{a^E_{mi}=0}^{\min(b^E_{mi}, a_{mi})} P\left(a^T_{mi} = a_{mi} - a^E_{mi} \middle| b^T_{mi} = b_{mi} - b^E_{mi}, g_{mi}\right) P\left(a^E_{mi} \middle| b^E_{mi}\right) P\left(b^E_{mi} \middle| b_{mi}\right),$$

where each probability above is calculated based on the probability mass function of the binomial distribution. Finally, we assume $p_{error} = 0.01$, which accomplishes our goal of a less stringent genotyping criterion at homozygous sites (Fig. S1).

If the null hypothesis that all isolates have the same genotype is true, then all inference could be done on the summed counts and depths $a_m = \Sigma_i a_{mi}$ and $b_m = \Sigma_i b_{mi}$, in which case the most likely genotype $\hat{g}_m$ for that mutation is

$$\hat{g}_m = \max_{g_m}[P(a_m|b_m, g_m)],$$

where $P(a_m|b_m, g_m)$ is calculated as described above.

We calculate the expected counts if the null is true as $\hat{a}_{mi} = \hat{g}_m b_{mi}$, with which we compute the test statistic

$$x^2 = \sum_i \frac{(a_{mi} - \hat{a}_{mi})^2}{\hat{a}_{mi}}.$$

If $\hat{a}_{mi} > 5$ for all i, we compute an exact p-value taking $x^2 \sim \chi^2_{\text{df=\#of isolates}-1}$ under the null assumption. Otherwise, we calculate an empirical p-value from 1,000 permutations of alternate and reference allele observations keeping the isolate depths constant.

Genotype posterior probability

In the later lineage frequency inference step, we would like to marginalize the likelihood of a mutation's metagenomic counts and depths by its genotype g_m, which effectively serves to downweight mutations for which we have less certainty about their genotype. For that we use an Expectation-Maximization procedure. We compute the posterior probability of the genotype g_m given the summed isolate clonal counts and depths a_m and b_m (see section above) as

$$P(g_m|a_m, b_m) = \frac{P(a_m|g_m, b_m)P(g_m)}{\sum_{g_m^*} P(a_m|g_m^*, b_m)P(g_m^*)}, \; a_m \sim \text{Binom}(b_m, g_m).$$

At first, we assume a uniform prior for $P(g_m)$, but having calculated the posteriors, we can update the priors as

$$P(g_m) = \sum_{m^*} P(g_{m^*} = g_m|a_{m^*}, b_{m^*}),$$

where m^* iterates over all mutations selected for a given lineage. We iterate over the two equations above until values converge enough, using a stop criterion on the change per iteration of the total likelihood of the data.

Joint inference of lineage frequencies in the metagenome

At this point, we have a list of lineages and their associated synapomorphic mutations. Note that, by definition, there is no overlap between the mutations used to identify any two lineages. We would like to use the metagenomic data for these mutations to infer the frequencies of the lineages during the fermentation season. For now, we will infer the frequency $f_l(t)$ of *chromosomes* of lineage l among all chromosomes in the population. This differs from the frequency $f_l^*(t)$ of *individuals* of lineage l among all individuals in the population because our populations are composed of a mix of diploid and triploid strains. We calculate this latter quantity in the section below.

We will do this inference independently for each timepoint, to avoid having to assume any particular model about how these lineages change in frequency through time. At each timepoint, we infer frequencies for all lineages jointly. If we allowed frequencies to vary freely, this would be equivalent to inferring each lineage's frequency independently. However, our lineages are hierarchically organized according to the inferred phylogenetic tree used to define them (as shown in Fig. 2A): we will use the term parent, child, and sibling lineages to point to the relationship between lineages in this hierarchy. In the most basal part of the tree, we will have one or more lineages that have no parent. Therefore, the frequencies $\vec{f}(t)$ of all lineages at a timepoint t are constrained by the set of inequalities

$$\sum_{l \in B} f_l(t) \leq 1, \text{ for the set of sibling basal lineages } B, \text{ and}$$

$$\sum_{l \in C_p} f_l(t) \leq f_p(t), \text{ for the set } C_p \text{ of children of a given lineage } p.$$

We assume that the error in metagenomic counts for different mutations are independent from each other, which is an assumption that only breaks in the case of mutations that are close enough in the genome that they may be covered by a same sequencing read. We therefore calculate the likelihood of a given model of lineage frequencies given the data as (suppressing t for convenience)

$$\mathcal{L}(\vec{f}|\text{data}) = \prod_l \prod_m \sum_{g_m} P(x_m|d_m, g_m, f_l) P(g_m|a_m, b_m),$$

where x_m and d_m are the counts and depths of mutation m in the metagenomic data, and we assume $x_m \sim \text{Binom}(d_m, g_m f_l)$.

We maximize the likelihood model above using a gradient descent method with a log-barrier that bounds solutions to the inequalities above as implemented in the function `constrOptim` in base R [ref]. To make this inference computationally tractable we do not infer the frequencies of all lineages at once, and instead follow an iterative procedure where at each step we infer the frequencies of a parent and all its children jointly starting from the most basal lineages:

(1) jointly fit frequencies of basal lineages $l \in B$, keeping $\sum_{l \in B} f_l(t) \leq 1$;
(2) randomly sort basal lineages; following this order jointly fit the frequency of basal lineage p and children lineages C_p, with inequalities

$$f_p \leq 1 - \sum_{p^* \in B | p^* \neq p} f_{p^*}, \text{ and}$$

$$\sum_{l \in C_p} f_l(t) \leq f_p(t);$$

(3) keep this new frequency f_p;
(4) for each fit grandparent lineage g, randomly sort its (also already fit) children C_g; following this order, fit jointly the frequencies of lineage $p \in C_g$ and its respective children $l \in C_p$, with inequalities

$$f_p \leq f_g - \sum_{p^* \in C_g | p^* \neq p} f_{p^*}, \text{ and}$$

$$\sum_{l \in C_p} f_l(t) \leq f_p(t);$$

(5) keep this new frequency f_p;
(6) repeat steps (4) and (5) until there are no more lineages to be fit.

We show inferred $\vec{f}(t)$ for all four site-years in Figs. S2A, S3A, S4A, and S5A.

Calculation of lineage frequency in the population

Having inferred the frequencies $\vec{f}(t)$ of all lineages in the metagenome, we proceed to calculating frequencies $\vec{f}^*(t)$ of all lineages in the population. These two quantities are related as (suppressing t for convenience)

$$f_l = \frac{p_l}{\bar{p}} f_l^*$$

where $p_l \in \{2,3\}$ is the ploidy of lineage l, and $\bar{p}$ is the mean ploidy in the population. Notice that if the whole population is composed of individuals of the same ploidy, then $f_l = f_l^*$.

We cannot directly assess the ploidy of all individuals in the original population, so instead we use inferred $\vec{f}(t)$ and respective lineage ploidies to estimate the mean ploidy in the population, but with two caveats. First, our isolate sampling may have missed ploidy heterogeneity within lineages. Second, our inference is not bound to infer frequencies that sum to 1 in the population, and thus may leave some portion of the population uninferred and of unknown ploidy. This is not a significant fraction in our study (see Figs. S2–S5), but it may be in other systems. We therefore make two assumptions: that (*i*) we are not missing ploidy heterogeneity in the inferred

portion of the population, and that (*ii*) any non-inferred portion of the population has the same mean ploidy as the inferred portion.

Let $F_2(t)$ and $F_3(t)$ be the total frequency of diploid and triploid strains in the metagenome as computed from inferred $\vec{f}(t)$. The frequencies $F_2^*(t)$ and $F_3^*(t)$ of diploid and triploid strains in the population are, thus, given by (suppressing t for convenience)

$$F_p^* = \frac{\frac{F_p}{p}}{\frac{F_2}{2} + \frac{F_3}{3}},$$

from which we compute the mean ploidy in the population as

$$\bar{p} = 2F_2^* + 3F_3^*.$$

We show computed $F_p(t)$ and $F_p^*(t)$ in Fig. S6, and inferred $\vec{f}^*(t)$ in Figures 4 and 5 of the main text. Effectively, they only slightly deviate from inferred $\vec{f}(t)$ (Figs. S2–S5).

VALIDATION ON RAREFIED CLONAL DATA

In this section, we assess the robustness of the inference procedure described above with respect to changes in the composition of picked clones in our dataset. To do this, we rarefy the data by selecting a simple random sample of 20, 10, or 5 among picked and starter clones for each of the four site-years. We then infer the lineages and their frequencies using only this subset of the clonal sequencing data while keeping the metagenomic sequencing dataset constant. To restrict this validation to the lineage inference procedure itself, we do not reinfer the clone phylogeny based on the rarefied clone dataset. A full account of phylogenetic uncertainty on the results of the inference requires substantial investigation and is beyond the scope of the current work.

Our analysis reveals that the rarefied clonal data largely preserves the large-scale lineage dynamics across all four site-years (Figs. S2–S5). This finding indicates that our inference method is generally robust to clonal undersampling. Reducing the number of picked clones reduces the number of inferred lineages in a size-dependent way. Larger lineages that dominate the dynamics are also more likely to be represented among picked clones, and their inferred frequencies are overall robust to undersampling. On the other hand, increasing the number of clones breaks large lineages down into smaller sublineages, allowing for the observation of finer-grain dynamics.

As anticipated from the inequality-constrained joint inference procedure, we note that the estimate of lineage frequencies becomes less constrained the less lineages there are in the inference. For example, the significant sweep observed in the last few timepoints in Site A – 2018 is not reflected in the estimate of sampled lineages in the rarefied dataset of 5 clones (Fig. S2D). Consequently, it remains desirable to sample sufficient clonal diversity in the population to more effectively constrain the inference. In practical terms, we suggest a similar rarefaction analysis to assess whether enough clones have been sampled in any particular study that uses this inference procedure.

Supplementary Figures

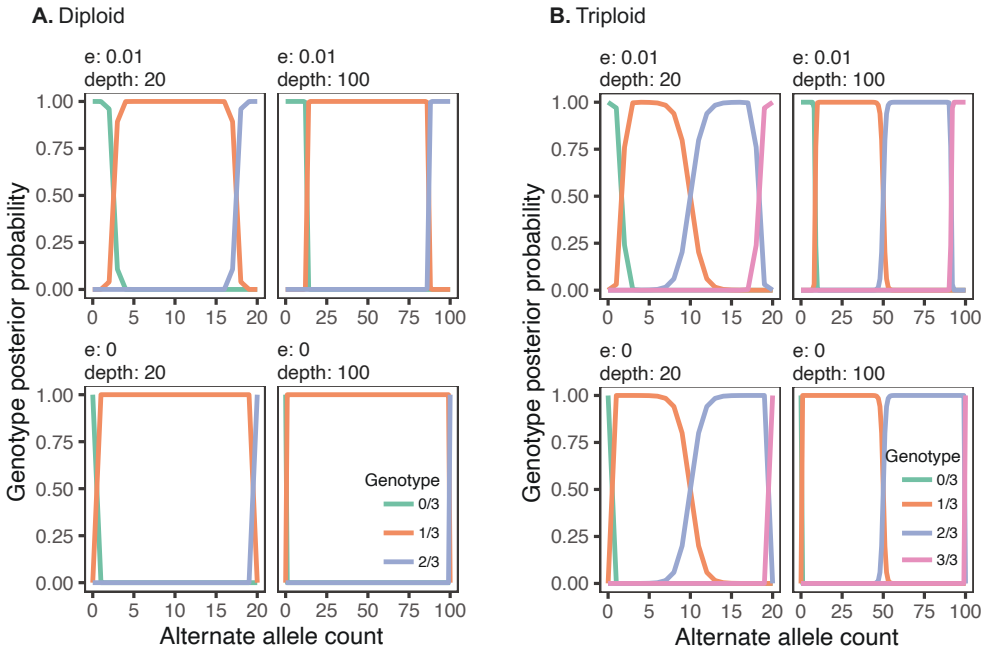


Figure S1. Probability of isolate data given genotype allowing for sequencing error. We show the computed probability of observing an alternate allele count value based on a given depth of coverage at that site, the probability of count errors p_{error} (*e* in the figure), and the isolate ploidy.

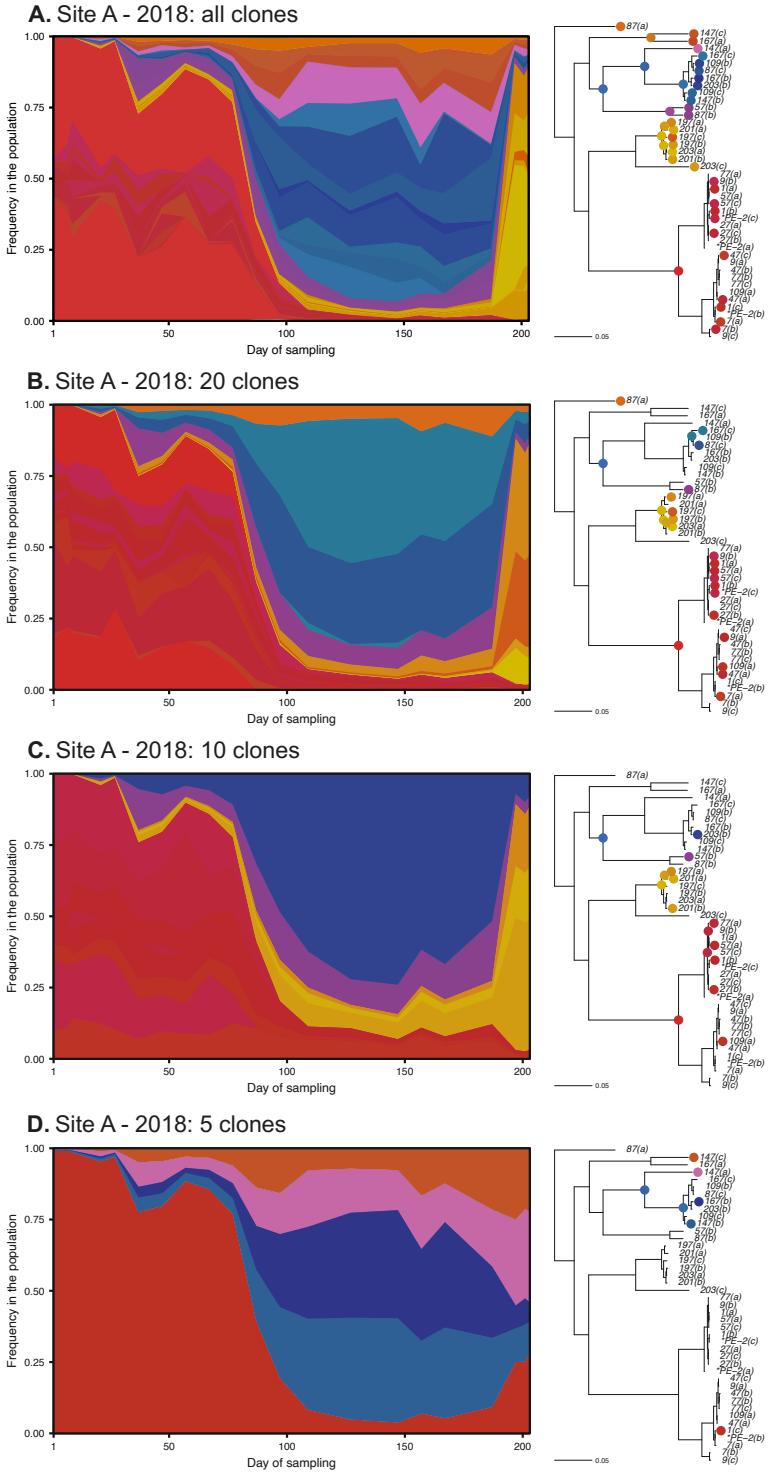


Figure S2. Inferred frequency of lineages in the metagenome for Site A – 2018. We show the inference results for **(A)** all picked clones, or a simple random sample of **(B)** 20, **(C)** 10, or **(D)** 5 of clones. Lineage frequencies $\vec{f}(t)$ are inferred with the procedure described in the sections above and are later used to compute the frequencies $\vec{f} * (t)$ of lineages in the population, as shown in Figs. 4 and 5. Lineages are color-labeled as in Fig. 4 and 5.

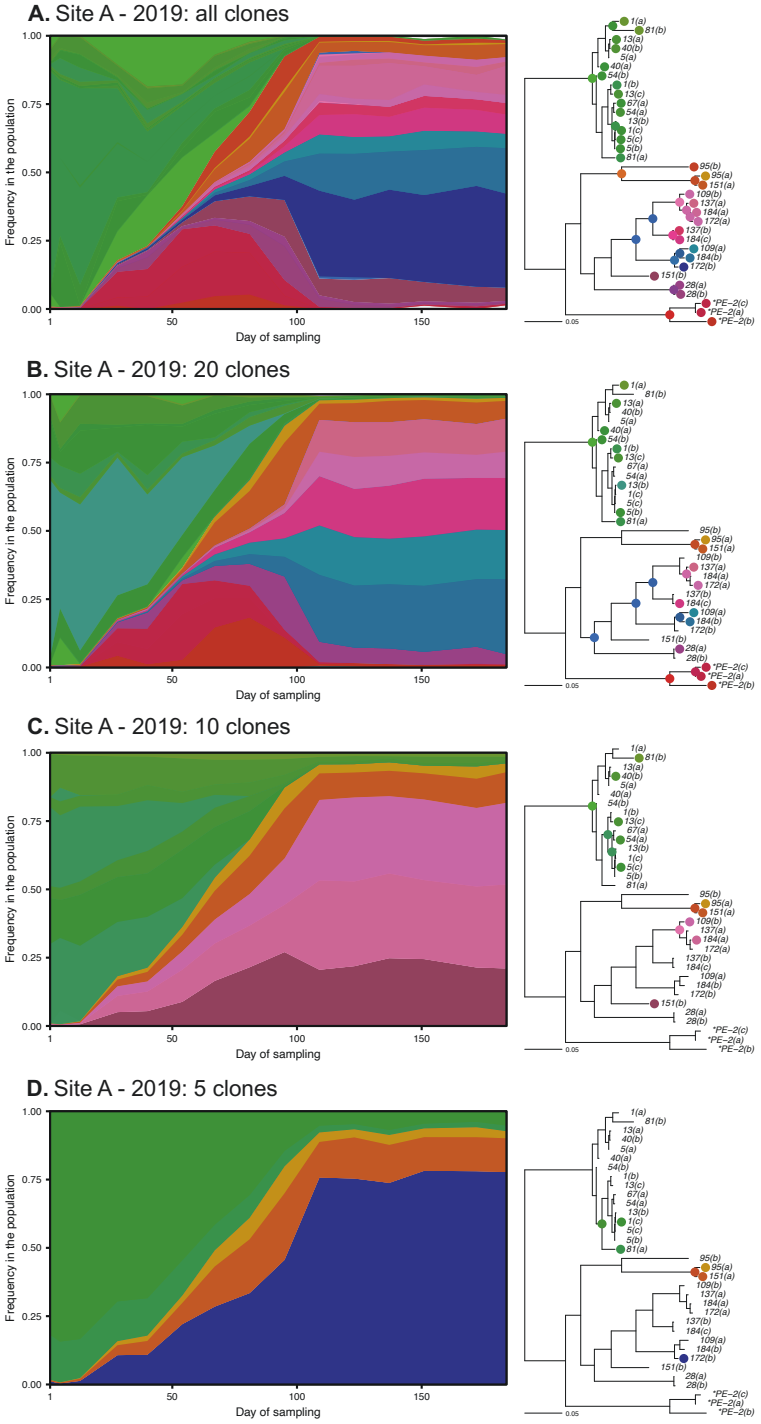


Figure S3. Inferred frequency of lineages in the metagenome for Site A – 2019. We show the inference results for **(A)** all picked clones, or a simple random sample of **(B)** 20, **(C)** 10, or **(D)** 5 of clones. Lineage frequencies $\vec{f}(t)$ are inferred with the procedure described in the sections above and are later used to compute the frequencies $\vec{f} * (t)$ of lineages in the population, as shown in Figs. 4 and 5. Lineages are color-labeled as in Fig. 4 and 5.

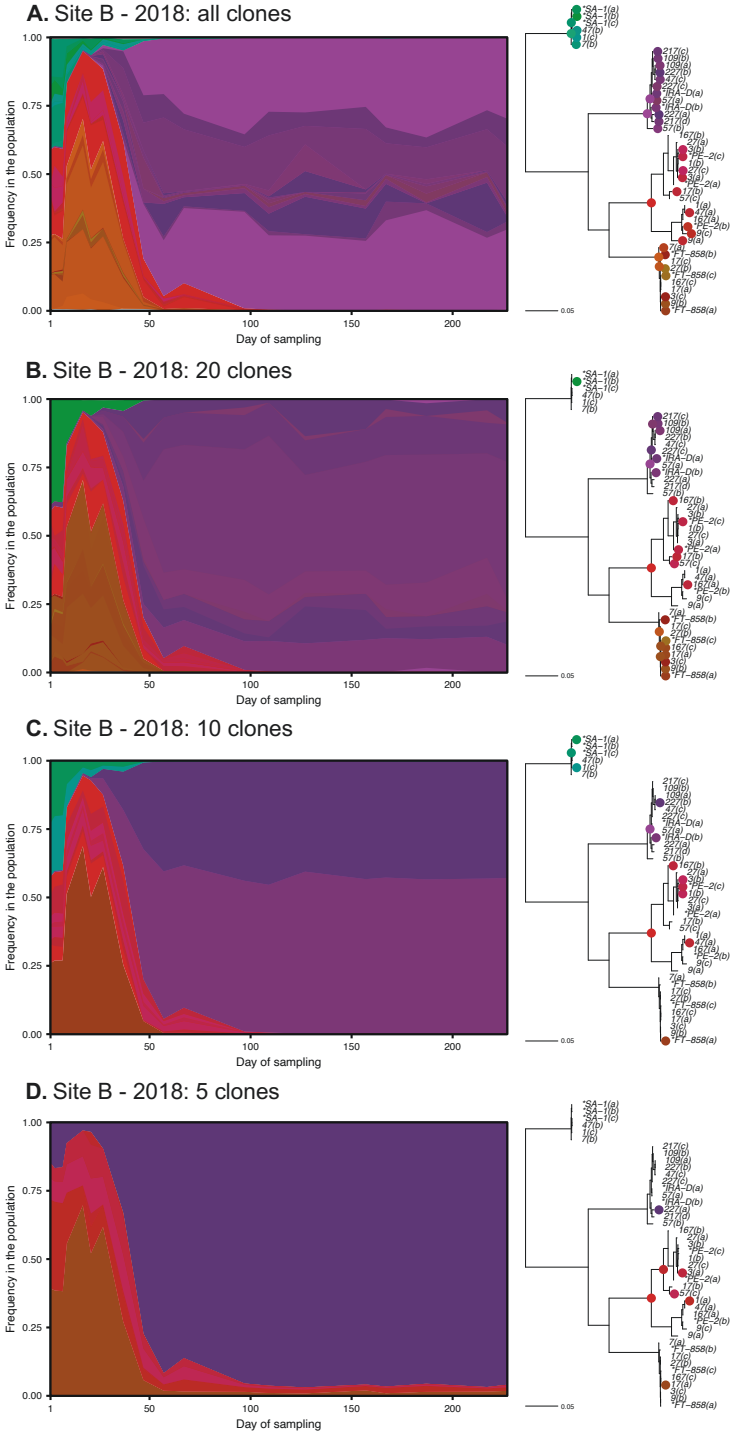


Figure S4. Inferred frequency of lineages in the metagenome for Site B – 2018. We show the inference results for **(A)** all picked clones, or a simple random sample of **(B)** 20, **(C)** 10, or **(D)** 5 of clones. Lineage frequencies $\vec{f}(t)$ are inferred with the procedure described in the sections above and are later used to compute the frequencies $\vec{f} * (t)$ of lineages in the population, as shown in Figs. 4 and 5. Lineages are color-labeled as in Fig. 4 and 5.

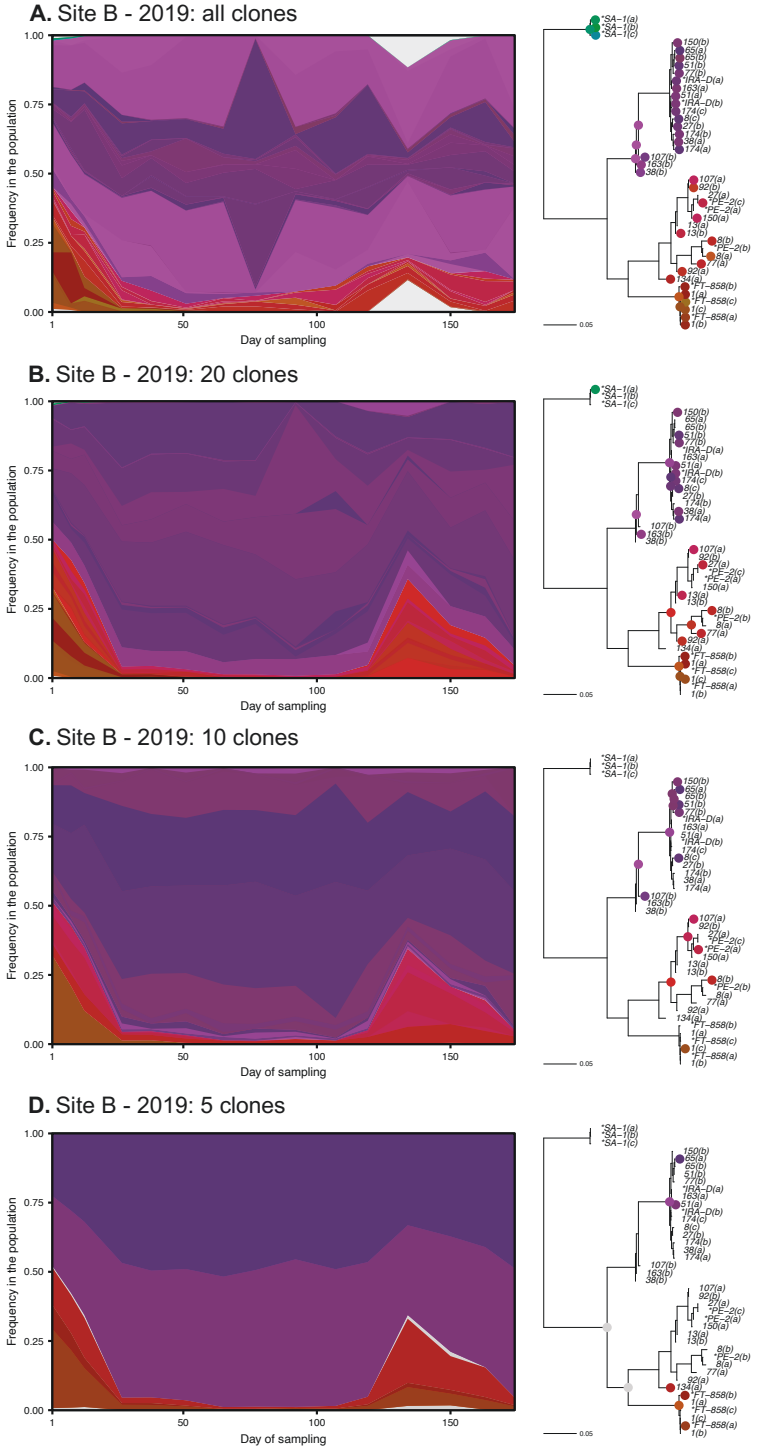


Figure S5. Inferred frequency of lineages in the metagenome for Site B – 2019. We show the inference results for **(A)** all picked clones, or a simple random sample of **(B)** 20, **(C)** 10, or **(D)** 5 of clones. Lineage frequencies $\vec{f}(t)$ are inferred with the procedure described in the sections above and are later used to compute the frequencies $\vec{f} * (t)$ of lineages in the population, as shown in Figs. 4 and 5. Lineages are color-labeled as in Fig. 4 and 5.

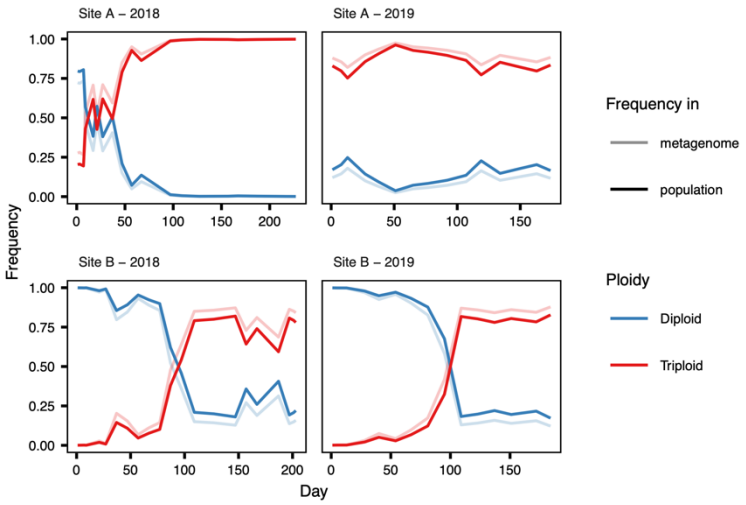


Figure S6. Inferred fraction of diploid and triploid strains along time based on inferred lineages' frequencies and ploidies. Estimated frequencies in both the metagenome (*i.e.* fraction of genetic material of the population that can be assigned to diploid or triploid individuals) and in the population (fraction of individuals) are shown. See Section "Calculation of lineage frequency in the population" of the Supp. Information for details.

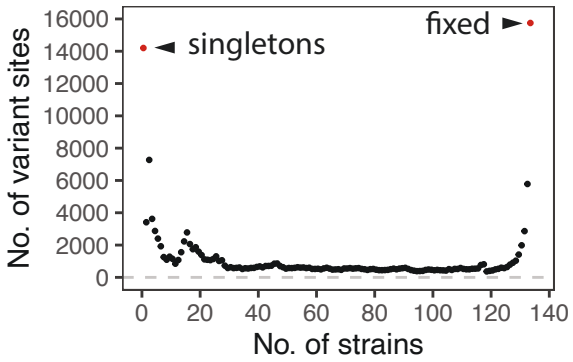


Figure S7. Histogram of number of isolates observed to carry a given alternate allele in the clonal sequencing data. Starter strains were excluded.

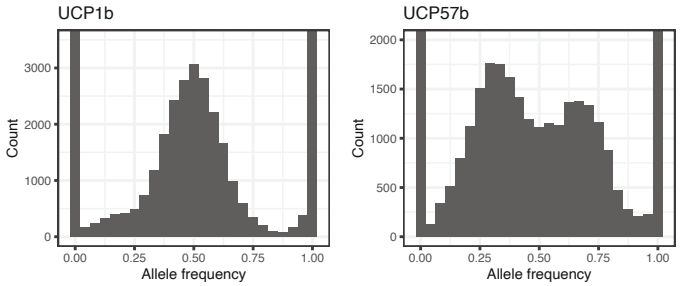


Figure S8. Representative examples of diploid and triploid whole-genome allele frequency distribution in the clonal sequencing data. The y-axes are cropped for better visualization.

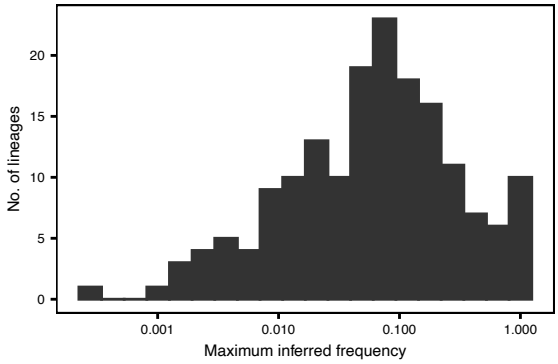


Figure S9. Distribution of maximum inferred frequency (over all timepoints) for all 197 inferred lineages across all site-years.

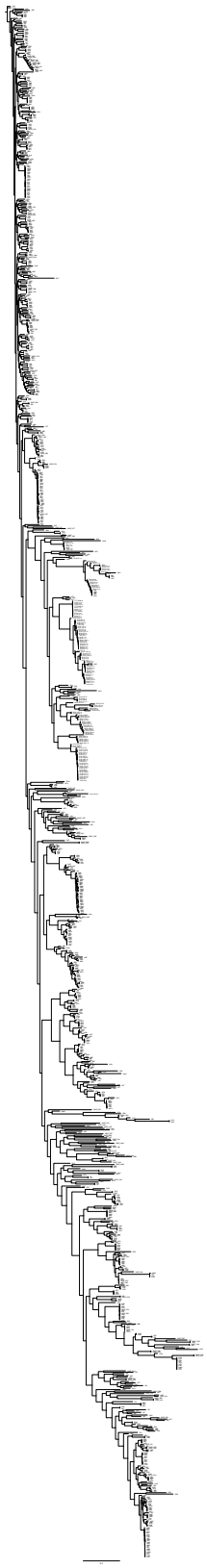


Figure S10. Midrooted labeled version of the tree in Fig. 6A. Clones from this study are labeled as in Table S2 and S3. Clones from the 1011 YGP are labeled as in Supp. Table 1 of Peter and colleagues (2018).

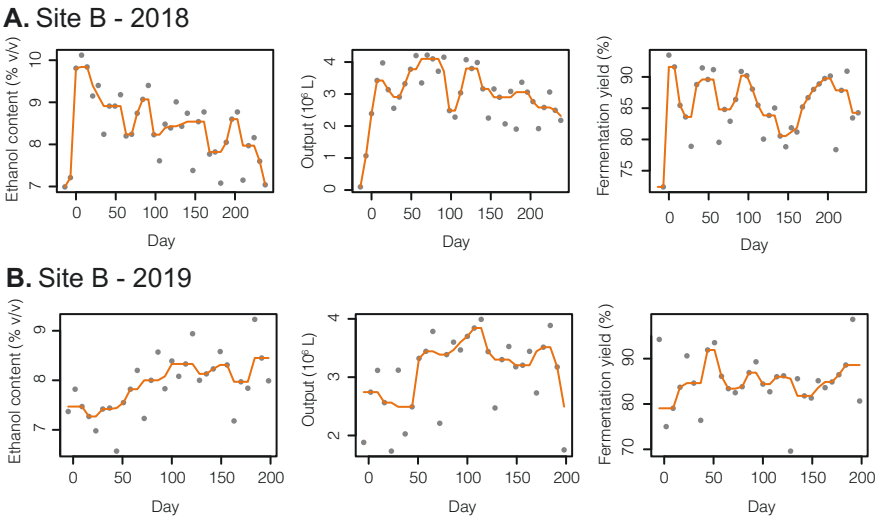


Figure S11. Fermentation metrics in Site B show no clear relationship with invasion by foreign strains. We show weekly data over the 2018 and 2019 fermentation seasons for (left) ethanol content of fermented wine, (middle) total bioethanol output, and (right) fermentation yield, as a measure of amount of ethanol produced out of a theoretical maximum. A running average is shown as an aid (orange line). The raw data can be found in Table S4.