

DYNAMICS OF CONTAMINANT MICROBES IN BIOETHANOL PRODUCTION FROM SUGARCANE

Authors: Ana Laura Ramos Romano¹, Natalia Coutouné¹, Artur Rego-Costa², Michael M. Desai^{2,3}, Marcelo F. Carazzolle⁴, Andreas K. Gombert^{1,#}

Affiliations:

¹School of Food Engineering, University of Campinas, Campinas, SP 13083-862, Brazil

²Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

³Department of Physics, Harvard University, Cambridge, MA 02138, USA

⁴Laboratory of Genomics and bioEnergy, Department of Genetics, Evolution, Microbiology and Immunology, Institute of Biology, University of Campinas, Campinas, SP, 13083-862, Brazil

#Corresponding author: gombert@unicamp.br

ABSTRACT

The dynamics and impact of microbial contaminants in industrial sugarcane bioethanol production in Brazil were investigated through a two-year metagenomic study across two biorefineries. Shotgun metagenomic sequencing revealed that temporal shifts in the contaminant microbiome dynamics within production seasons were more pronounced than inter-annual or inter-mill variations. While *Saccharomyces* spp. dominated, bacterial communities, primarily within the Firmicutes phylum and dominated by the genera *Lactobacillus*, *Limosilactobacillus*, and *Bacillus*, exhibited dynamic changes. Correlation analyses with industrial process parameters revealed a complex interplay: lower *Lactobacillus* levels in one mill were associated with increased ethanol yield, whereas higher levels in another mill correlated with reduced yeast viability and increased flocculation. The presence of *Limosilactobacillus* was linked to decreased yeast viability, whereas *Bacillus* showed potential for inhibiting both *Lactobacillus* and *Limosilactobacillus*. These findings highlight the nuanced and species-specific impacts of bacterial contaminants on bioethanol production, underscoring the need for strain-level

- 31 functional studies and targeted interventions to optimize fermentation efficiency
- 32 and stability in industrial settings.

33 INTRODUCTION

34 The global push for sustainable energy positions bioethanol as a key
 35 renewable fuel. Brazil is a leader in this market, ranking as the world's second-
 36 largest producer and top exporter (1)(2). Its sugarcane-based process is vastly
 37 more land- and energy-efficient than the corn-based method used by the largest
 38 producer, the United States (3)(4). Brazilian sugarcane bioethanol boasts an
 39 energy balance of 10.2:1 (renewable output to fossil input), while US corn
 40 bioethanol is significantly lower at roughly 1.4:1 (5). This production relies on
 41 the fermentation of sugarcane feedstocks by *Saccharomyces cerevisiae* in non-
 42 aseptic conditions, making it highly susceptible to microbial contamination and
 43 related process disruptions. These contaminants, primarily bacteria, may
 44 compete directly with *S. cerevisiae* for sugars, diverting resources that would
 45 otherwise be transformed into ethanol, reducing the final achievable ethanol
 46 yield. In an industry where profit margins are exceptionally tight, even minor
 47 gains in process efficiency can have a substantial economic impact.

48 The sugarcane bioethanol industry employs multiple strategies to limit
 49 microbial contamination in the fermentation environment. This includes
 50 prophylactic sulfuric acid treatment of recycled yeast cells (pH < 2.0 for 1 to 1.5
 51 h) to suppress bacterial growth (7), maintaining a high yeast inoculum fraction,
 52 and performing routine equipment cleaning to reduce biofilm formation.
 53 Although antibiotics such as penicillin, virginiamycin, streptomycin, tetracycline,
 54 and monensin have historically been used, rising concerns over increased
 55 antibiotic resistance and potential contamination of animal feed byproducts (8)
 56 have prompted the search for alternative antimicrobial approaches. Bacterial
 57 contamination levels can reach 10^7 – 10^8 cells/mL, leading to economic losses of
 58 up to 20,000 liters of bioethanol per day in a medium-sized biorefinery (9),
 59 primarily due to competition for nutrients and the production of inhibitory organic
 60 acids (10, 11). Moreover, species like *Limosilactobacillus fermentum* can cause
 61 yeast flocculation (12), disrupting yeast–substrate interactions, slowing
 62 fermentation kinetics, and reducing ethanol yield and productivity, while also

63 impairing downstream processes such as yeast separation before distillation
64 (13).

65 Despite the economic and operational significance of bacterial
66 contamination in bioethanol production, our understanding of the dynamic shifts
67 within these bacterial communities over time remains limited. Culture-
68 dependent methods have identified key genera like *Lactobacillus*, *Bacillus*,
69 *Staphylococcus*, and *Leuconostoc* (14-17), and more recent studies using 16S
70 rRNA gene sequencing consistently highlight *Lactobacillus* as the dominant
71 contaminant (10, 11, 18-21). However, these reports provided only a low-
72 resolution and rather static snapshot of microbial community composition,
73 because they relied on a small number of samples, making it difficult to establish
74 correlations between microbial composition dynamics and process performance
75 (22).

76 Shotgun metagenomic sequencing offers a high-resolution alternative to
77 observe microbial communities. Lino et al. (2024) applied this strategy to
78 samples from two sugarcane biorefineries, revealing a more diverse microbial
79 composition in the fermentation environment than previously described, and
80 proposing new hypotheses about the impact of some of the bacterial members
81 on the industrial process (23). However, such datasets are still scarce,
82 especially when sampling across different production cycles and biorefineries.
83 Further lack of temporal resolution prevents evaluation of the population
84 dynamics of these microbial communities.

85 To address this knowledge gap, in this study we use whole-metagenome
86 sequencing to characterize the microbiome present in the industrial
87 fermentation of sugarcane feedstocks. We evaluated the temporal changes in
88 this microbiome across two entire production years in two distinct biorefineries,
89 expanding upon the initial findings of Rego-Costa et al. (24) by specifically
90 analyzing DNA sequences not attributable to *S. cerevisiae*. This approach aims
91 at providing a deeper understanding of the bacterial dynamics within industrial
92 alcoholic fermentations in Brazilian sugarcane biorefineries, ultimately

93 contributing to the development of more effective strategies for contamination
94 control and process improvement.

95 **MATERIALS AND METHODS**

96 **Sample Collection**

97 Industrial fermentation samples were collected approximately weekly
98 over two consecutive sugarcane production seasons (2018 and 2019), spanning
99 from April to November of each year. Samples were obtained from two
100 independent sugarcane mills, designated as UIR and UCP (Fig. 1A,
101 corresponding to Site A and Site B, respectively, in (24)), both situated in the
102 region of Piracicaba, São Paulo State, Brazil. All samples were taken from the
103 fermented broth, or *wine*, in the industrial jargon, either directly from a fermentor
104 or from the holding tank used as an intermediary storage vessel downstream of
105 the fermentor and upstream of the continuous centrifuges. For each sampling
106 event, 15 mL Falcon tubes prefilled with 3 mL of sterile glycerol were filled with
107 12 mL of the sample. Immediately after collection, samples were thoroughly
108 homogenized by vortexing and then subjected to a two-stage freezing protocol:
109 initial storage at -20 °C for 1 to 3 months, followed by long-term storage at -
110 80 °C. Samples were later shipped in dry ice to Harvard University, where they
111 were kept at -80 °C until further processing. The collection and handling of these
112 samples were conducted in accordance with Brazilian federal regulations, and
113 the samples are duly registered in the SisGen database under the accession
114 numbers R40E57A, RB42674, R193AED, RAD5521, and AF14971.

115 **DNA Extraction and Sequencing**

116 A total of 70 samples collected from both mills and both production years
117 were selected for shotgun metagenomic sequencing. Information on sample
118 collection dates can be found in Rego-Costa et al. (2023) (24). Total genomic
119 DNA was extracted from each sample using a column-based protocol previously

described by Nguyen Ba et al. (2019) (25), which was assessed visually under a microscope to ensure efficient lysis of all cells in the samples. Sequencing libraries were prepared using a transposase-based kit (Illumina, San Diego, CA, USA) following the reduced-volume protocol described by Baym et al. (2015) (26). Library quality and size distribution were assessed using an Agilent TapeStation 4200 (Agilent Technologies, Santa Clara, CA, USA), and libraries were quantified using Thermo Fisher's QuBit dsDNA HS assay before sequencing. Sequencing was performed using two runs on an Illumina NextSeq 500 platform and one run on an Illumina MiSeq platform, generating an average of 9.1 million 300 bp paired-end reads per sample.

Read Preprocessing

Raw paired-end reads were trimmed using Trimmomatic version 0.39 (27), removing the reads shorter than 36 base pairs and trimming of low-quality bases from the 3' end of reads using a sliding window of 4 bases, with a minimum average quality score of 23, which removed ~20% of all the reads (Fig. 1B). Adapter sequences were removed using NGmerge version 0.2 (28) with the "-a" flag, employing stringent parameters to ensure complete adapter removal and accurate merging of overlapping paired-end reads, where applicable.

Taxonomic Profiling

Taxonomic classification of the processed reads was performed using Kraken 2 (v. 2.0.8-beta) (29), a *k*-mer-based classifier known for its speed and accuracy. Initially, all reads were classified at a high taxonomic level (kingdom) against the core_nt database available from the Kraken 2 repository (<https://benlangmead.github.io/aws-indexes/k2>) to delineate reads belonging to *Saccharomyces*, other fungi, bacteria, viruses, and archaea. Due to the overwhelming abundance of *Saccharomyces cerevisiae* reads, which could obscure the detection of less abundant microbial taxa, a targeted approach was

employed to remove these *Saccharomyces*-derived sequences. All processed reads were aligned against the reference genome of *Saccharomyces cerevisiae* S288C using the Burrows-Wheeler Aligner (BWA) version 0.7.15-r1140 (30). Reads that failed to align to the *S. cerevisiae* genome were retained for subsequent taxonomic classification of the remaining microbial community (Fig. 1B; Supp. Figure 1). These unmapped reads were then classified using Kraken 2 against its bacterial, fungal, viral, and archaeal databases. The raw output from Kraken 2 is included in Supplementary Material 1. The breakdown of the classification across all four site-years was visualized using Krona (Fig. 1C displays data for UCP 2018) (31). An interactive chart can be found in Supplementary Material 2.

Kraken 2 classifies reads down to the smallest taxonomic level that it can deal with statistical confidence. In our dataset, 67% of the reads could be classified to the coarsest taxonomic level, Domain, with 61% being classified down to Genus, followed by a dropoff to 55% at the level of Species (Fig. 1B). To balance classification depth and amount of classified reads, we retained all classifications down to the Genus level in downstream analyses. Given the inherent uncertainty in read classification, we employed a filter on classified genera to each site-year separately, only keeping high-confidence genera that are seen to have more than 100 reads in 4 or more samples in that site-year (Fig. 1D). We further removed any reads classified down to the genus *Homo*, which likely stem from contamination during any step of the process, as well as those classified down to *Cellulosimicrobium*, which are likely to stem from the enzymatic treatment used for yeast cell lysis in the genomic DNA preparation step. In all further analyses, we refer to the ratio between the number of reads classified down to a kept genus and the total number of reads fed to Kraken 2 (*Homo* and *Cellulosimicrobium* excluded) as either *Relative abundance* or *Fraction of Reads*.

Microbial community analysis

Microbial community structure was explored using multiple complementary approaches. Principal component analysis (PCA) was carried out over the relative abundance data for both years and sites together using the factoextra package (32) in R (v.4.3.1). This provided us with a way to visualize global patterns of variation in microbial composition and assess potential clustering by industrial site and production year.

The alpha diversity of bacterial communities was assessed using the Shannon diversity index (H'). Shannon's index was then computed as:

$$H' = - \sum_{i=1}^S p_i \ln(p_i)$$

where p_i represents the relative abundance of genus i among S genera detected in the sample (33). This index accounts for both richness and evenness of the microbial community, with higher values indicating greater diversity.

To investigate potential interactions within the microbial community, we inferred co-occurrence and co-exclusion networks from genus-level abundance data. Networks were constructed using the SpiecEasi package (34) with the Meinshausen–Bühlmann method and 99 regularization parameters, with model stability evaluated by pulsar subsampling (50 replicates). The resulting adjacency matrix of partial correlations was represented as an undirected weighted graph using igraph (35). Edges above the 80th percentile of absolute weight were retained and classified as positive (co-occurrence) or negative (co-exclusion) interactions. Final networks were visualized with ggraph (36) using a Fruchterman–Reingold layout, with node size scaled to degree and edge width proportional to correlation strength.

Industrial process metrics and regression analysis

Industrial process data were obtained from both sites and encompassed a range of fermentation and yeast performance variables (Supp. Table 1). These included yeast-related metrics (viability, budding, flocculation, and bacterial counts) and physicochemical parameters of the fermentation process (sugar concentration in must and wine, pH, fermentation yield, and volume of ethanol produced). The frequency and type of measurements differed between the two sugarcane mills: UIR performed detailed yeast viability and contamination monitoring, while UCP reported physicochemical data and weekly yield records. A detailed description of individual parameters can be found in Supplementary Table 1.

To test for associations between microbial community variation and industrial performance, we implemented linear regression models linking the abundance of the two main genera to occupy the process: *Lactobacillus* and *Limosilactobacillus* to each industrial variable. For each site (UIR and UCP), separate models were fitted with the general form:

$$Y_j = \beta_0 + \beta_1 \text{Lactobacillus} + \beta_2 \text{Limosilactobacillus} + \beta_3 \text{Year} + \epsilon$$

where Y_j represents each industrial metric j (e.g., yeast viability, pH, ethanol yield), and Year was included as a discrete covariate to account for interannual effects. Models were fitted using the `lm()` function in R, and p-values for regression coefficients were computed from t-tests to assess the statistical significance of each association.

Data and code availability

The raw sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA865262. All code developed and used for metagenomic processing and statistical analyses is publicly available in the GitHub repository: https://github.com/alrsilva5577/Contaminants_ethanol.

RESULTS

This study involved the collection of whole-population microbiological samples during the 2018 and 2019 fermentation seasons from two independent sugarcane biorefineries, referred to as UIR and UCP (Fig. 1A). These mills are located approximately 18 km apart in São Paulo state, Brazil. Operational differences — such as the yeast starter strains used at the beginning of the season and distinct contamination control strategies, among others — can influence the microbiological community structure of the fermentation process. Previous work by Rego-Costa et al. (24) described the dynamics of the yeast *S. cerevisiae* populations in these same samples, showing large population shifts, including the invasion of the process by new yeast strains likely originating from the environment. Here, we describe and analyze the temporal dynamics, microbial community structure, and correlation with industrial process metrics of the total microbial DNA excluding *S. cerevisiae*.

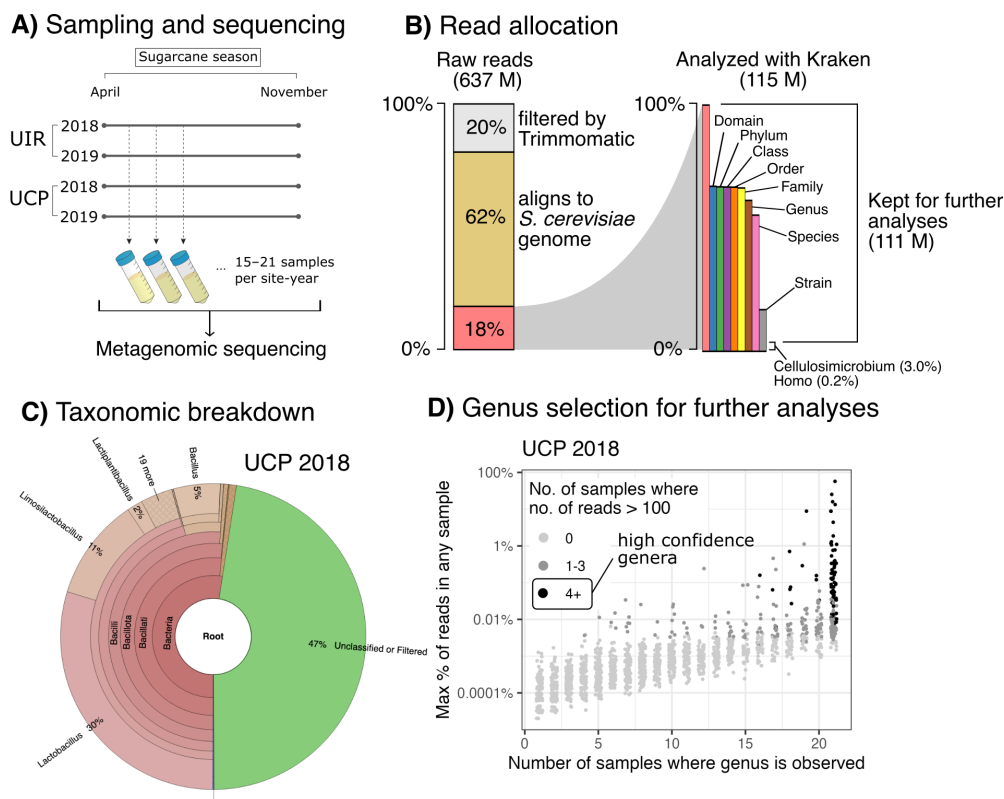


Figure 1. Metagenomic pipeline and selection of high-confidence genera for bioethanol fermentation contaminants. (A) Schematic of the time-series sampling strategy across two biorefineries (UIR and UCP) during the 2018 and 2019 sugarcane seasons, leading to shotgun metagenomic sequencing. (B) Read processing and allocation pipeline. Raw reads were filtered by quality (20%), and budding yeast-derived sequences were removed by alignment to the s288c reference genome (62%). The remaining high-quality reads (18%) were classified using Kraken2, with the fraction of classified reads decreasing at more granular taxonomic levels. See a breakdown of this classification in each site-year and across temporal samples in Supplementary Figure 1. (C) Illustrative taxonomic breakdown of classified reads in UCP 2018, showing dominance of bacterial taxa, notably *Lactobacillus* (30%) and *Limosilactobacillus* (11%). This data is available as an interactive plot in Supplementary Material 2. (D) Genus selection methodology (in UCP 2018) based on peak observed abundance (maximum fraction of reads in any sample) and prevalence (number of samples where genus is observed). High-confidence genera (black points) were retained for further analysis based on the criterion of ≥ 4 samples containing >100 reads. See the same figure for the other three site-years in Supplementary Figure 2.

Shotgun metagenomic sequencing of the 70 collected samples yielded a total of 639 million 300-bp paired-end reads, with an average of 9.1 million reads per sample (Fig. 1B). 20% of the reads were first removed due to low quality scores based on the established Trimmomatic filtering criteria. Approximately 62% of all reads aligned to the *Saccharomyces cerevisiae* S288C reference genome. The remaining 18% of high-quality, non-*Saccharomyces* reads were subjected to taxonomic classification with Kraken 2 in order to characterize the microbial contaminants present in the fermentation environment, including archaea, viruses, bacteria, and other fungi.

The fraction of reads classified to each taxonomic level was consistent across all site-years and time points (Fig. 1B; Supp. Fig. 1). Given the probabilistic nature of taxonomic assignment and the sharp drop-off in fraction of classified reads when comparing genus to species, we proceeded with genus-level classification data. That is, reads classified down to the genus or lower taxonomic level retain their genus classification, and reads classified down to a taxonomic level above genus remain labeled as Unclassified at the genus level. We further reassess the genus-level classification. We start by

removing the reads identified as likely contaminants: *Homo*, likely from human handling in both the industry and in the laboratory, and *Cellulosimicrobium*, likely related to the *Cellulosimicrobium*-derived zymolyase enzymatic treatment used in genomic DNA preparation. We then leveraged the longitudinal sampling to filter for high-confidence genera, retaining only those genera present with >100 reads in at least four samples within a given site-year (Fig. 1D; Supp. Fig. 2). This high-confidence set accounted for 94% of all reads classified to the genus level; reads from filtered-out genera were relabeled as Unclassified. We use this data for all further analyses, using read abundance (the fraction of all high-quality non-*Saccharomyces* reads classified to a given genus) as a proxy for population abundance that we do not attempt to correct for genome size.

The final filtered dataset comprised 98 unique bacterial genera and 5 bacteriophages (Supp. Fig. 3). Consistent with literature on the bioethanol industry (37), the communities were consistently dominated by the genera *Lactobacillus*, *Limosilactobacillus*, and *Bacillus*, which were the most prevalent genera across all datasets (Fig. 2; Supp. Fig. 3). The overall community structure was heavy-tailed and showed site-specific differences (Supp. Fig. 4). UCP 2018, for example, exhibited the highest richness with 81 different genera, 24 of which are unique to this site-year, while UIR 2019 was the least rich, with 36 total genera, and a single unique genus (Supp. Fig. 4A; Supp. Fig. 5). Beyond the dominant taxa, we identified a persistent community of low-abundance microbes, with 30 genera shared across all four site-years, underscoring their stable role in this fermentation environment (Supp. Fig. 4A).

Analysis of the temporal dynamics revealed the highly unstable nature of this microbial ecosystem (Fig. 2). The dominant genera, particularly *Lactobacillus* and *Limosilactobacillus*, exhibited large boom and crash cycles, with relative abundances fluctuating dramatically between successive time points (Fig. 2A). An exception was observed in the UIR 2019 time series, which showed a qualitatively different dynamic. Here, the community composition remained relatively stable throughout the season, a feature that also

corresponded to the lowest overall genus richness and Shannon diversity of all four datasets (Supp. Fig. 5). While *Lactobacillus* and *Limosilactobacillus* abundance fluctuations did not show a clear relationship, their cycles often coincide with fluctuations in multiple low-abundance genera, a pattern suggestive of system-wide perturbations and/or microbial interactions (Fig. 5B).

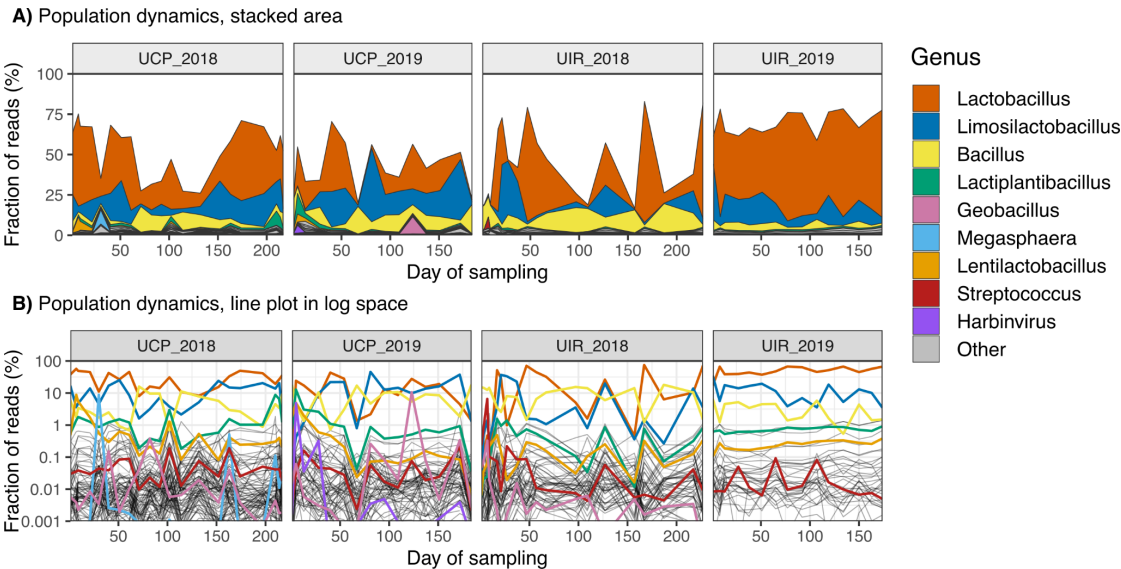


Figure 2. Temporal dynamics of the microbial (non-*Saccharomyces*) community in industrial bioethanol production. Relative abundance of identified bacterial genera tracked over the sampling period for all four site-years. (A) Stacked area plots showing community composition, dominated by *Lactobacillus* and *Limosilactobacillus*, followed by *Bacillus*. The top nine most abundant genera are distinctly colored, while the remainder are shown in grey. (B) Line plots of the same genus abundance data, displayed on a log-scale (0.001% to 100% fraction of reads). This visualization highlights the temporal fluctuations and persistence of low-abundance lineages across all four-time series.

We next sought to determine if the observed microbial dynamics were associated with industrial process metrics (Supp. Table 1; Supp. Fig. 6, 7). The specific metrics collected varied between the UIR and UCP facilities. As expected, many of these metrics correlate with one another (Fig. 3A,C), including trivial associations (e.g., between total reducing sugar, or TRS, and the Brix index in UCP) and non-trivial ones (e.g., yeast viability and sugarcane TRS in UIR).

To assess the impact of the dominant bacteria in the process, we focused on *Lactobacillus* and *Limosilactobacillus*. These two genera are the most abundant and, critically, their abundances do not correlate, as assessed by Pearson's correlation or by PCA (Fig. 3A,C; Supp. Fig. 8A), allowing their effects to be evaluated independently. We fit individual linear regression models for each industrial metric, using the abundances of both genera as explanatory variables and including the production year (2018/2019) as a covariate to account for annual variation.

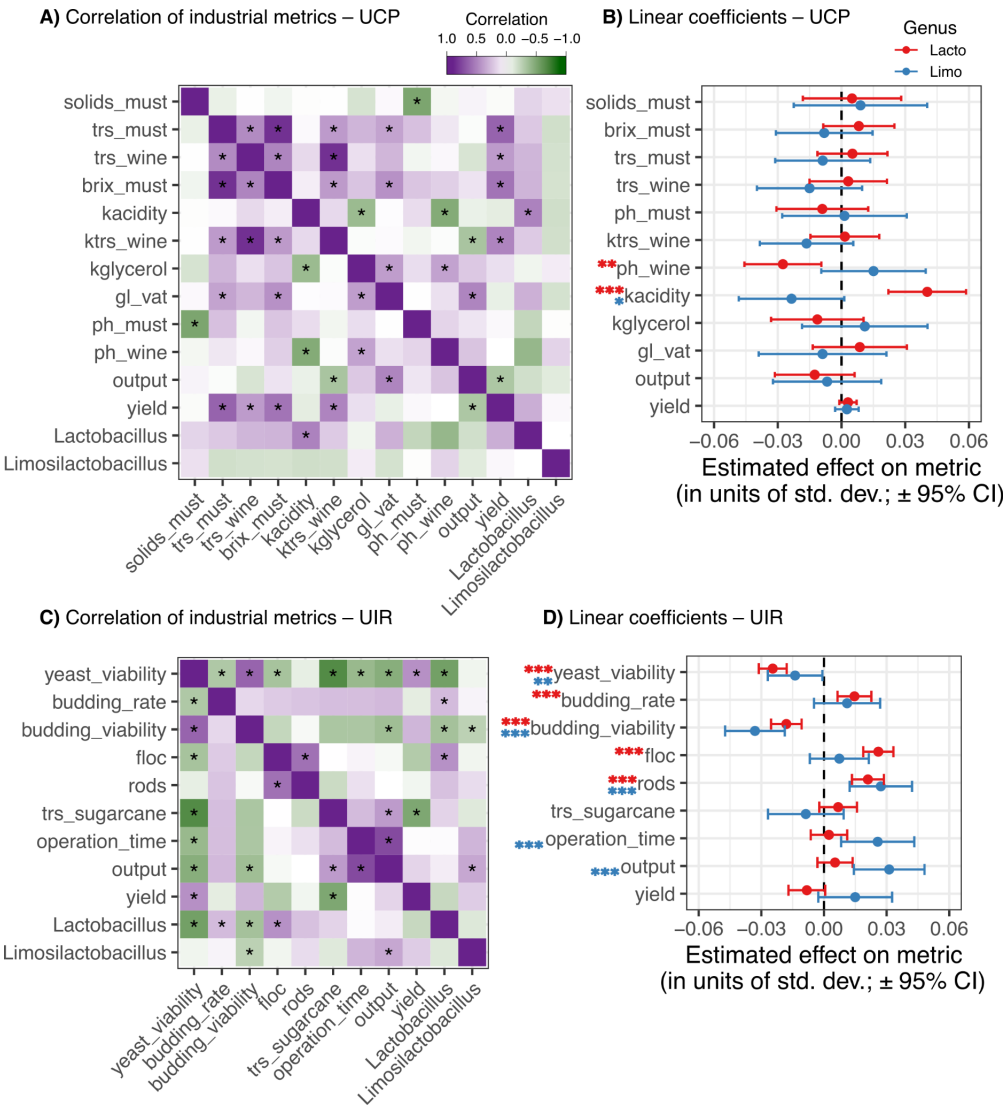


Figure 3. Statistical association of dominant bacterial genera with industrial process metrics. (A, C) Pearson's correlation matrices for UCP and UIR, respectively, showing correlations between genus abundance and metrics of the industrial process, kindly shared with us by the biorefineries. See Supplementary Tables 1, 2, and 3 for more information on these metrics and for the fit statistics of the regression models. Asterisk indicates a significant correlation under a t-test of the correlation coefficients: * : $p < 0.1$, ** : $p < 0.05$, *** : $p < 0.05/\text{no. of tests}$ as a Bonferroni-corrected significance threshold, where the number of tests is taken to be 12 for UCP and 9 for UIR. (B, D) Estimated linear coefficients and 95% confidence intervals for the effect of *Lactobacillus* and *Limosilactobacillus* abundance on industrial metrics at UCP and UIR, respectively, taking year as a covariate. Observed values and predictions from linear models are shown in Supplementary Figures 9 and 10.

The models revealed significant associations between bacterial abundance and process metrics at both facilities (Fig. 3B,D; Supp. Table 2,3). At UCP, *Lactobacillus* abundance showed a strong, statistically significant association with lower wine pH (i.e., higher acidity), a finding supported by both direct pH tests and titration methods (Fig. 3B). Conversely, *Limosilactobacillus* abundance was associated with a less acidic environment (a positive coefficient for pH). No other significant relationships were found at UCP between these two genera and other metrics, total ethanol output or production yield being relevant ones.

At UIR, the associations revealed a more complex interplay. Both *Lactobacillus* and *Limosilactobacillus* were significantly associated with decreased yeast cell and bud viability (Fig. 3D). *Lactobacillus* abundance was also correlated with a higher budding rate and increased flocculation, the latter referring to a state of cell aggregation associated with reduced fermentation efficiency. Most notably, *Limosilactobacillus* showed a significant positive association with key production metrics: Operation Time (the fraction of time the pipeline is operational) and Final Output (total ethanol volume). This suggests that while both dominant genera may negatively impact yeast health, *Limosilactobacillus* is linked to more favorable, large-scale industrial outcomes.

The co-occurrence network exhibited multiple interconnected taxa forming small clusters, mostly in pairs. Only a few nodes showed a higher degree, acting as

local hubs, suggesting taxa with broader association profiles within the community (Supplementary Figure 11). Several genera that emerged from the recent reclassification of the former genus *Lactobacillus* (38) (e.g., *Companilactobacillus*, *Latilactobacillus*, *Liquorilactobacillus*, *Loigolactobacillus*, and *Paucilactobacillus*) formed a subnetwork of positive associations. These genera share a common evolutionary origin and are adapted to carbohydrate-rich, fermentative environments, suggesting that their co-occurrence reflects shared ecological preferences rather than direct biological interactions. Most of the other co-expression clusters (such as *Corynebacterium*, *Polynucleobacter*, and *Rhizobium*) represent a functional module of plant-associated microorganisms, linked by an ecological niche in the rhizosphere and phyllosphere, that persist after sugarcane processing.

While all identified viral taxa were characterized as bacteriophages of the *Lactobacillus* genus, two of them exhibited positive correlations with non-host bacterial genera (*Colunavirus* and *Watanabevirus*), and only the genus *Harbinvirus* was positively associated with the genus *Lactiplantibacillus* within the co-expression networks. Notably, no co-exclusionary patterns were observed between these phages and any of the newly reclassified *Lactobacillus* genera. This lack of negative correlation could suggest that viral activity in this ecosystem may not be driven by acute lytic predation of the dominant lactic acid bacteria, but rather by stable lysogenic associations as previously described for lactobacilli phages (40,41,42,43).

This pattern of small sub-networks, rather than a single highly connected core, may suggest a wide diversity of origins for the bacteria found in the fermentation tank. Many of the genera described are commensals or parasites of insects and plants such as *Apilactobacillus*, *Bombilactobacillus*, *Blattabacterium*, *Candidatum Nardonella*, *Spiroplasma*, and *Xanthomonas* (38, 44, 45, 46, 47, 52); others are organisms associated with the rhizosphere or freshwater such as *Bacillus*, *Pantoea*, *Mycobacterium*, *Flavobacterium* and *Chlamydia* (48, 49, 50, 51, 52), and several other genera with a high number of

399 human or animals pathogens, like *Treponema*, *Haemophilus*, and *Yersinia* (53,
400 54, 55).

401 Interestingly, the co-exclusion subnetwork between *Bacillus* and
402 *Lactobacillus* indicates a strong antagonism. As lactic acid producers,
403 *Lactobacillus* species acidify the fermentation medium, creating unfavorable
404 conditions that could inhibit *Bacillus*' growth and induce sporulation, as *Bacillus*
405 prefers neutral to alkaline environments (48). This negative association could
406 be further driven by the production of bacteriocins by *Lactobacillus*, a widely
407 described phenomenon (56, 57, 58, 59, 60). Furthermore, the antimicrobial
408 repertoire of *Bacillus* spp. spans numerous chemical classes, ranging from
409 small non-ribosomal peptides to complex polyketides and lipopeptides, capable
410 of producing quorum-quenching and causing cell lysis in other bacteria (61).

411 DISCUSSION

412 Our study employed metagenomic sequencing to investigate the non-
413 *Saccharomyces* fraction of the microbial communities in the economically-
414 relevant fermentation process of bioethanol production from sugarcane. By
415 sampling these communities with high temporal resolution across biorefineries
416 and over different fermentation seasons, we were able to assess the consistent
417 and variable aspects of the microbial composition and dynamics of this industrial
418 process with higher confidence than has been possible so far.

419 We have found *Lactobacillus* and *Limosilactobacillus* to be the most
420 prevalent contaminants of this industrial alcoholic fermentation process, which
421 agrees with previous studies (22). While both genera produce lactic acid, they
422 differ in their metabolic impact on the process. *Lactobacillus* is strictly
423 homofermentative (38), producing virtually exclusively lactic acid as the product
424 of glucose fermentation, while *Limosilactobacillus* is heterofermentative,
425 producing ethanol and acetic acid, in addition to lactic acid. In UCP we have
426 found that *Lactobacillus* abundance is associated with a more acidic
427 fermentation environment, while the opposite is true for *Limosilactobacillus*. This
428 contrasts with the observations of Lino et al. (2024) (37), who found
429 *Limosilactobacillus* to be associated with higher acidity. Lino et al. (2024)
430 sampled the process over the course of a fermentation cycle, while we sampled
431 mostly the fermented broth at the end of a cycle over the course of whole
432 fermentation seasons. Differences in temporal resolution may account for these
433 contrasting observations. We also note that both studies are limited by the use
434 of correlation-based analyses, which hampers our ability to make causal claims.

435 Biorefineries in our study differed in their antimicrobial strategy during the
436 study period. In the 2018 production year, UIR initiated operations with a low
437 dose of monensin, subsequently transitioning at some unknown timepoint to the
438 natural antimicrobial hops and the chemical chlorine dioxide for the remainder
439 of the year. In 2019, UIR switched to using exclusively hops and chlorine
440 dioxide, with no antibiotic application. UCP utilized a combination of three

antibiotics – monensin, virginiamycin, and oxytetracycline – in varying proportions, along with chlorine dioxide. It is notable, therefore, that out of all four sampled site-years, we find that UIR 2019, which only had hops and chlorine dioxide in its antimicrobial strategy, had a significantly less diverse and more stable microbial composition through time. Even if UIR 2019 did not show increased ethanol output or yield in 2019, relative to 2018 (Supp. Fig. 6), such a stability in the microbial community may be desirable for the industrial process, if it leads to fewer disruptions and/or to a decreased use of antimicrobials, which always add costs and a negative environmental footprint. Limited disclosure of industrial data prevents us from being able to assess other microbiological features that could be informative for this question, such as yeast flocculation and viability.

A large fraction of reads remains unclassified by Kraken 2 (Fig. 2). This is not at all surprising, given the software’s conservative reliance on exact match of k-mers to its database. Therefore, sequence variants (either real or from sequencing error) and low-complexity regions will both lead to unclassified reads. There is a non-negligible chance that the unclassified fraction of reads hides some taxon that is not present in the Kraken 2 database, despite its extensive size. However, given previous knowledge of the microbial composition of the bioethanol fermentation environment, it is unlikely that any such taxon would be more prevalent than *Lactobacillus* or *Limosilactobacillus*.

Our findings underscore the dual-edged role of bacterial contaminants in bioethanol fermentation. We present evidence suggesting that certain *Lactobacillus* species, particularly *L. amylovorus*, *L. helveticus*, and *L. crispatus*, may exert beneficial effects, potentially enhancing process yield and yeast budding rate at lower abundances. However, exceeding a critical threshold, these same *Lactobacillus* populations are associated with detrimental effects on yeast viability and the occurrence of flocculation, a poorly understood phenomenon in this context that warrants further investigation.

Conversely, our data indicate that *Limosilactobacillus* species (*L. fermentum*, *L. panis*, and *L. mucosae*) are likely detrimental to fermentation performance. Their association with reduced yeast viability is consistent with their potential to produce inhibitory metabolites such as acetic acid, due to their heterofermentative metabolism.

In conclusion, our metagenomic analysis of industrial sugarcane bioethanol production across two production years and two distinct biorefineries in Brazil reveals nuanced insights into the dynamics and potential impact of microbial contaminants. Contrary to initial expectations, the temporal shifts in the contaminant microbiome within a single harvest appear to be more pronounced than the differences observed between different harvests or between the two mills studied. This highlights the dynamic nature of these microbial communities and the influence of within-season operational variations. A deeper understanding of the complex dynamics between yeast and bacterial communities in industrial bioethanol production holds great potential to optimize process efficiency and reduce the economic losses associated with contamination, opening new ways for novel bioprocess engineering strategies.

Current methods for studying microbial contamination in industrial bioethanol production, such as cultivation-dependent techniques, are limited because they require isolation for identification and cannot accurately quantify diverse bacterial taxa. Similarly, existing cultivation-independent methods often rely on metabarcoding, which only uses sequences of specific DNA regions. Our study introduces a novel approach using whole-metagenome sequencing and analysis, which offers several advantages: it provides access to taxonomic information for a wider range of microorganisms, including viruses and fungi, offers accurate abundance estimation, and achieves high taxonomic resolution down to the genus and species level. This comprehensive strategy allowed us to visualize the distribution of identified microorganisms in UIR and UCP during 2018 and 2019, as shown in Figure 1, specifically focusing on the percentage of *Saccharomyces* species, other fungi, bacteria, and viruses, along with a

magnified view of the most abundant bacterial species relative to all identified bacteria.

The spectrum of bacterial genera identified in this study is largely consistent with findings from both historical culture-dependent studies (Gallo, 1898 (13); Rodini, 1985 (15); Neto, 1990 (17)) and more recent cultivation-independent investigations (Bonatelli et al., 2017 (11); Costa et al., 2015 (21); Queiroz et al., 2020 (21); Lino et al., 2021 (23)). Notably, the dominant genera reported in these prior studies, including *Lactobacillus* and *Bacillus*, are also the most prevalent in our metagenomic dataset. While earlier culture-dependent studies identified species such as *Lactobacillus fermentum* (now *Limosilactobacillus fermentum*), *L. helveticus*, *L. plantarum* (now *Lactiplantibacillus plantarum*), *Bacillus subtilis*, *B. coagulans*, and *B. stearothermophilus*, more recent metagenomic studies, including Lino et al. (2021) (23), have provided species-level resolution, identifying *L. amylovorus*, *L. fermentum*, *L. helveticus*, *L. buchneri* (*Lentilactobacillus buchneri*), *L. plantarum*, and *Bacillus cereus* as key players.

Our metagenomic approach allowed for taxonomic classification down to the species level, although the genus level was the primary focus of our analyses, due to the lower confidence and higher proportion of single-species genera. The original Kraken 2 results for read classification beyond genera are included in Supplementary Material 1. Within the *Lactobacillus* genus, *L. amylovorus* was the most abundant species, with *L. helveticus* and *L. crispatus* present at lower levels in some samples. The *Limosilactobacillus* genus was primarily represented by *L. fermentum*, along with *L. panis* and *L. mucosae*. *Bacillus* was dominated by *Bacillus cereus*, with minor contributions from *Bacillus wiedmannii*. Other identified genera were largely represented by a single classified species: *Lentilactobacillus buchneri*, *Lactiplantibacillus plantarum*, *Streptococcus infantarius*, *Streptomyces* sp. ICC4, *Megasphaera elsdenii*, *Prevotella* sp. WR041, and *Geobacillus stearothermophilus*. A recent study on bioethanol production identified key bacterial influences on

fermentation performance. It found that an increased yeast-to-bacteria ratio correlates with higher production performance. Specifically, *Lactobacillus amylovorus* and *Weissella* species were associated with higher production performance, with *L. amylovorus* potentially having beneficial effects on the ethanol yield. Conversely, *Limosilactobacillus fermentum* was identified as a detrimental bacterium that can reduce ethanol yield by up to 5%. Interestingly, different strains of *L. fermentum* were found to have varying impacts (37).

CONCLUSION

Intriguingly, the presence of *Bacillus* species (*B. cereus* and *B. wiedmannii*) appears to be negatively correlated with the abundance of both *Lactobacillus* and *Limosilactobacillus*. This suggests a potential for *Bacillus* to act as a natural modulator of the fermentation microbiome, potentially preventing the overgrowth of *Lactobacillus* and the establishment of undesirable *Limosilactobacillus* populations. However, the specific mechanisms underlying this interaction within the complex industrial fermentation environment require further investigation. Interestingly, *B. cereus* was not identified as being part of the taxa that make up 80% of the biodiversity in the samples analysed by Lino et al (2021) (23).

The ability of our shotgun metagenomic approach to provide species-level resolution within key genera like *Lactobacillus* represents a significant advancement over studies that rely solely on genus-level identification. The impact observed even within the *Lactobacillus* genus highlights the importance of future research focusing on strain-specific functional characterization to fully understand their roles in bioethanol production, as had already been shown by Lino et al (2024) (37) and Rich et al (2018) (39). The insights gained from this study pave the way for exploring targeted interventions to manipulate the fermentation microbiome. For instance, further investigation into the antagonistic relationship between *Bacillus* and lactic acid bacteria could lead to the development of novel biocontrol strategies to enhance process stability and efficiency.

To validate the correlations observed in this industrial setting and to understand the underlying mechanisms of interaction between yeast and bacterial contaminants, controlled laboratory experiments and fermentation simulations will be required. Using defined co-cultures with varying proportions of the identified key bacterial species (a.k.a. synthetic microbial communities) would provide a more controlled environment to elucidate the causal relationships between specific microbial populations and fermentation performance, including yield, yeast health, and flocculation dynamics. Furthermore, detailed metabolomic analyses could identify specific microbial metabolites responsible for the observed effects.

Despite the different scopes, this study and that of Rego-Costa et al. (2023) (24) converge on similar ecological patterns. Both show that UCP displays higher microbial diversity, whereas UIR tends to harbor more homogeneous communities, particularly in 2019. Such convergence, emerging from independent analyses of yeast and bacterial populations, suggests that structural factors of the industrial setting (e.g., starter strain choice, recycling practices, or antimicrobial usage) exert a consistent influence on the broader fermentative microbiome. Also, this comparison indicates that, whereas yeast populations undergo rapid and recurrent strain turnover, bacterial contaminants follow more predictable trajectories and appear less affected by yeast dynamics. This distinction may have practical implications: controlling invasive yeast lineages and managing bacterial contaminants may require distinct, yet complementary, strategies to ensure stability and efficiency in industrial ethanol production.

ACKNOWLEDGMENTS

We thank the various members of the bioethanol mills UCP and UIR who provided extensive support and expertise along the way, especially Thaís Grano and Luciano Zamberlan. We thank the Bauer Core facility at Harvard for

assistance with sequencing, as well as the FAS Division of Science Research Computing Group, which manages the computer cluster used for the bioinformatic analysis of the sequencing data.

FUNDING

AKG acknowledges a scholarship provided by the São Paulo Research Foundation (FAPESP, São Paulo, Brazil, grant # 2018/04962-5) and the Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq, Brasília, Brazil, grant # 306190/2022-2). ALRR acknowledges a M.Sc. scholarship from CNPq (grant # 130582/2021-2).

SUPPLEMENTARY MATERIAL

This manuscript is accompanied by Supplementary Tables and Figures (3 tables and 11 figures), which can be found in this same document, and by Supplementary Material 1 and Supplementary Material 2, supplied as separate files.

REFERENCES

1. RFA: Renewable Fuels Association. Annual Ethanol Production - U.S. and World Ethanol Production. *Ethanolrfa*. [Online] 2022. <https://ethanolrfa.org/markets-and-statistics/annual-ethanol-production>. Accessed February, 2024.
2. USDA: United States Department of Agriculture – Foreign Agricultural Service. Biofuels Annual. *APPS FAS*. [Online] USDA. https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Biofuels%20Annual_Sao%20Paulo%20ATO_Brazil_BR2022-0047.pdf. Accessed February, 2024.
3. Macedo, I. C., Seabra, J. E., & Silva, J. E. (2008). Green house gases emissions in the production and use of ethanol from sugarcane in Brazil: the 2005/2006 averages and a prediction for 2020. *Biomass and bioenergy*, 32(7), 582-595.
4. de Barros, M. M. e Szklo, A. Petroleum refining flexibility and cost to address the risk of ethanol supply disruptions: the case of Brazil. *Renew Energy*. 2015, Vol. 77.
5. Goldemberg J. Ethanol for a sustainable energy future. *Science*. 2007 Feb 9;315(5813):808-10. doi: 10.1126/science.1137013. Erratum in: *Science*. 2007 Sep 7;317(5843):1325. PMID: 17289989.
6. Calderón, L A L. *Saccharomyces cerevisiae* transcriptional reprogramming due to bacterial contamination during industrial scale bioethanol production. *Micro Cell Fact*. 2015, Vol. 14:13.
7. Antonini, S R C. Conventional and nonconventional strategies for controlling bacterial contamination in fuel ethanol fermentations. *World Journal of Microbiology and Biotechnology*. 2018, Vol. 34, 80.
8. Muthaiyan, A, Limayem, A e Ricke, S C. Antimicrobial strategies for limiting bacterial contaminants in fuel bioethanol fermentations. *Progress in Energy and Combustion Science*. 2011, Vol. 37, 351-370.
9. Basso, T O, et al. Homo and heterofermentative lactobacilli differently affect sugarcane-based fuel ethanol fermentation. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*. 15, 2014, Vols. 169-177.
10. Firmino, F C, et al. Characterization of microbial communities in ethanol biorefineries. *Journal of Industrial Microbiology & Biotechnology*. 2020, Vol. 47, 183-195.
11. Bonatelli, M L, et al. Characterization of the contaminant bacterial communities in sugarcane first-generation industrial ethanol production. *FEMS Microbiology Letters*. 2017, Vol. 364.
12. Carvalho-Netto, O V, et al. *Saccharomyces cerevisiae* transcriptional reprogramming due to bacterial contamination during industrial-scale bioethanol production. *Microbial Cell Factories*. 2015, Vol. 14, 13.
13. Basso, L. C., Basso, T. O., & Rocha, S. N. (2011). Ethanol production in Brazil: the industrial process and its impact on yeast fermentation. *Biofuel production-recent developments and prospects*, 1530(3), 85-100.
14. Gallo, C R. Determinação da microbiota bacteriana de mosto e de dornas de fermentação alcoólica. s.l. : Doutorado - Faculdade de Engenharia de Alimentos/ UNICAMP, 1989.
15. Rodini, M A T. Isolamento, caracterização e identificação de bactérias contaminantes de dornas de fermentação nas destilarias de etanol . *Mestrado em Microbiologia Agrícola*. s.l. : Escola Superior de Agricultura Luiz de Queiroz/ USP, 1985.
16. Rosales, S Y R. Contaminantes bacterianos da fermentação etanólica: isolamento em meios diferenciais, identificação e avaliação de desinfetantes. *Doutorado em*

654 *Microbiologia Aplicada*. s.l. : Faculdade Estadual Paulista “Júlio de Mesquita Filho”,
655 1989.

656 17. Neto, P O. Influência da contaminação por bactérias lácticas na fermentação alcoólica
657 pelo processo de batelada alimentada. *Dissertação de Mestrado*. s.l. : Faculdade de
658 Engenharia de Alimentos/ UNICAMP, 1990.

659 18. Li, Q, Heist, E P e Moe, L A. Bacterial community structure and dynamics during corn-
660 based bioethanol fermentation. *Microbial Ecology*. 2016, Vol. 71, 409-421.

661 19. Rich, J O, et al. Microbial contamination of commercial corn-based fuel ethanol
662 fermentations. *Bioresource Technology Reports*. 2020, Vol. 11.

663 20. Queiroz, L L, et al. Dynamics of microbial contaminants is driven by selection during
664 ethanol production. *Brazilian Journal of Microbiology*. 2020, Vol. 51, 303-3012.

665 21. Costa, O Y A, et al. Microbial diversity in sugarcane ethanol production in a Brazilian
666 distillery using a culture-independent method. *Journal of Industrial Microbiology &*
667 *Biotechnology*. 2015, Vol. 42, 73-84.

668 22. Basso TO, Venturini AM, Ceccato-Antonini SR, Gombert AK. Microbial ecology applied to
669 fuel ethanol production from sugarcane. *FEMS Microbiol Ecol* 2025;101:fiaf100.
670 <https://doi.org/10.1093/femsec/fiaf100>

671 23. Lino SO, F., Bajic, D., Vila, J. C. C., Sánchez, A., & Sommer, M. O. A. 2021, Complex
672 yeast–bacteria interactions affect the yield of industrial ethanol fermentation.
673 *Nature communications*, 12(1), 1498.

674 24. Rego-Costa, A, et al. Yeast population dynamics in Brazilian bioethanol production. *G3*
675 *Genes| Genomes| Genetics*. 13, 2023, Vol. 7.

676 25. N, Nguyen Ba A, et al. High-resolution lineage tracking reveals travelling wave of
677 adaptation in laboratory yeast. *Nature*. 575, 2019, Vols. 494-499.

678 26. Baym, M, et al. Inexpensive multiplexed library preparation for megabase-sized genomes.
679 *PLoSOne*. e0128036, 2015.

680 27. Bolger, A. M., Lohse, M. e Usadel, B. Trimmomatic: A flexible read trimming tool for illumina
681 ngs data. 2014, *Bioinformatics*, Vol. 30, pp. 2114–2120.

682 28. Gaspar, John M. NGmerge: merging paired-end reads via novel empirically-derived
683 models of sequencing errors. 2018, *BMC Bioinformatics*, Vol. 19, p. n536.

684 29. Wood, D. E., Lu, J. e Langmead, B. *Improved metagenomic analysis with kraken 2*. 2019,
685 *Genome Biology*, Vol. 20, p. 257.

686 30. Li, H. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem.
687 2013, *asXiv*, Vol. 1303.3997v2.

688 31. Ondov, B.D., Bergman, N.H. & Phillippy, A.M. 2011, Interactive metagenomic visualization
689 in a Web browser. *BMC Bioinformatics* 12, 385. [https://doi.org/10.1186/1471-2105-](https://doi.org/10.1186/1471-2105-12-385)
690 [12-385](https://doi.org/10.1186/1471-2105-12-385)

691 32. A Kassambara, F Mund, 2017. CRAN repository: [https://cran.r-](https://cran.r-project.org/web/packages/factoextra/index.html)
692 [project.org/web/packages/factoextra/index.html](https://cran.r-project.org/web/packages/factoextra/index.html). URL:
693 <http://www.sthda.com/english/rpkgs/factoextra>.

694 33. Shannon, Claude E., and Warren Weaver. 1949. The Mathematical Theory of
695 Communication. *Urbana: University of Illinois Press*.

696 34. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA (2015) Sparse and
697 Compositionally Robust Inference of Microbial Ecological Networks. *PLoS*
698 *Comput Biol* 11(5): e1004226. <https://doi.org/10.1371/journal.pcbi.1004226>

699 35. Csardi, G., & Nepusz, T. 2006. The igraph software package for complex network
700 research. *InterJournal, Complex Systems*, 1695.

36. Pedersen T. 2025. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 2.2.2.9000. <https://ggraph.data-imaginist.com>
37. Lino, F S O, et al. Strain dynamics of contaminating bacteria modulate the yield of ethanol biorefineries. *Nature Communications*. 15, 2024, Vol. 5323.
38. Zheng, J, et al. A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. *International Journal of Systematic and Evolutionary Microbiology*. 70, 2020, Vols. 2782-2858.
39. Rich JO, Bischoff KM, Leathers TD, Anderson AM, Liu S, Skory CD. Resolving bacterial contamination of fuel ethanol fermentations with beneficial bacteria - An alternative to antibiotic treatment. *Bioresour Technol*. 2018 Jan;247:357-362. doi: 10.1016/j.biortech.2017.09.067. Epub 2017 Sep 14. PMID: 28954248.
40. Capra, M. L., Mercanti, D. J., Reinheimer, J. A., & Quiberoni, A. L. (2010). Characterisation of three temperate phages released from the same *Lactobacillus paracasei* commercial strain. *International Journal of Dairy Technology*, 63(3), 396–405. Portico. <https://doi.org/10.1111/j.1471-0307.2010.00600.x>
41. Séchaud L, Callegari M L, Rousseau M, Muller M C and Accolas J P (1989) Relationship between temperate bacteriophage 0241 and virulent bacteriophage 832-B1 of *Lactobacillus helveticus*. *Netherlands Milk and Dairy Journal* 43 261–277.
42. Davidson B E, Powell I B and Hillier A J (1990) Temperate bacteriophages and lysogeny in lactic acid bacteria. *FEMS Microbiology Reviews* 7 79–90.
43. Suarez V, Zago M, Quiberoni A, Carminati D, Giraffa G and Reinheimer J A (2008) Lysogeny in *Lactobacillus delbrueckii* strains and characterization of two new temperate prolate-headed bacteriophages. *Journal of Applied Microbiology* 105 1402–1411.
44. Vicente CSL, Mondal SI, Akter A, Ozawa S, Kikuchi T, Hasegawa K (2018) Genome analysis of new *Blattabacterium* spp., obligatory endosymbionts of *Periplaneta fuliginosa* and *P. japonica*. *PLoS ONE* 13(7): e0200512. <https://doi.org/10.1371/journal.pone.0200512>
45. Lefèvre, C., Charles, H., Vallier, A., Delobel, B., Farrell, B., & Heddi, A. (2004). Endosymbiont phylogenesis in the dryophthoridae weevils: evidence for bacterial replacement. *Molecular biology and evolution*, 21(6), 965–973. <https://doi.org/10.1093/molbev/msh063>.
46. Razin, S. (2006). The genus *Mycoplasma* and related genera (class Mollicutes). In *The prokaryotes* (pp. 836-904). Springer, New York, NY.
47. Ryan, R., Vorhölter, F.J., Potnis, N. et al. Pathogenomics of *Xanthomonas*: understanding bacterium–plant interactions. *Nat Rev Microbiol* 9, 344–355 (2011). <https://doi.org/10.1038/nrmicro2558>
48. Xu, X., & Kovács, Á. T. (2024). How to identify and quantify the members of the *Bacillus* genus?. *Environmental Microbiology*, 26(2), e16593.
49. Tambong, J. T. (2019). Taxogenomics and Systematics of the Genus *Pantoea*. *Frontiers in Microbiology*, 10, 487075. <https://doi.org/10.3389/fmicb.2019.02463>
50. Meehan, Conor J., et al. "Reconstituting the genus *Mycobacterium*." *International journal of systematic and evolutionary microbiology* 71.9 (2021): 004922.
51. Seo, Hyojun, et al. "The plant-associated *Flavobacterium*: a hidden helper for improving plant health." *The plant pathology journal* 40.3 (2024): 251.
52. Collingro, Astrid, Stephan Köstlbacher, and Matthias Horn. "Chlamydiae in the environment." *Trends in Microbiology* 28.11 (2020): 877-888.

53. Norris, S. J., Paster, B. J., Moter, A., & Göbel, U. B. (2006). The Genus *Treponema*. In *The Prokaryotes* (pp. 211–234). Springer New York. https://doi.org/10.1007/0-387-30747-8_8.
54. Nørskov-Lauritsen N. 2014. Classification, Identification, and Clinical Significance of *Haemophilus* and *Aggregatibacter* Species with Host Specificity for Humans. *Clin Microbiol Rev* 27: <https://doi.org/10.1128/cmr.00103-13>.
55. Seabaugh, J. A., & Anderson, D. M. (2024). Pathogenicity and virulence of *Yersinia*. *Virulence*, 15(1). <https://doi.org/10.1080/21505594.2024.2316439>
56. Du, Hechao, et al. "Purification, characterization, and mode of action of plantaricin GZ1-27, a novel bacteriocin against *Bacillus cereus*." *Journal of Agricultural and Food Chemistry* 66.18 (2018): 4716-4724.
57. Xu, Jue, et al. "Detection and development of lactic acid bacteria bacteriocins-A hint on the screening of bacteriocins of lactic acid bacteria." *LWT* (2025): 117627.
58. Aouadhi, Chedia, Manel Turki, and Abderrazak Maaroufi. "Screening and characterization of lactic acid bacteria with antibacterial effect against heat-resistant spore outgrowth of *Bacillus sporothermodurans*." *International Dairy Journal* 158 (2024): 106047.
59. Ren, Yanxin et al. "Bacteriocin production and inhibition of *Bacillus subtilis* by *Lactobacillus paracasei* HD1.7 in an indirect coculture system." *Preparative biochemistry & biotechnology* vol. 52,7 (2022): 783-788. doi:10.1080/10826068.2021.1995412
60. Bogovič-Matijašić, B., et al. "Isolation and characterization of two bacteriocins of *Lactobacillus acidophilus* LF221." *Applied Microbiology and Biotechnology* 49.5 (1998): 606-612.
61. Abriouel, H., Franz, C. M. A. P., Omar, N. B., & Gálvez, A. (2011). Diversity and applications of *Bacillus* bacteriocins. *FEMS Microbiology Reviews*, 35(1), 201–232. <https://doi.org/10.1111/j.1574-6976.2010.00244.x>.

SUPPLEMENTARY TABLES AND FIGURES

Supplementary Table 1. Description of the industrial metrics kindly shared by both UCP and UIR. These metrics primarily serve the internal purposes of these companies, and therefore will have been calculated using undisclosed methods, which may vary from site to site. Therefore, some of the units are unclear or unknown, but the respective metrics were kept in the analysis as they are still interpretable for our purposes. Must is the industry term for the sugary liquid to which yeast is added for fermentation, and wine is the alcoholic broth right at the end of a fermentation cycle.

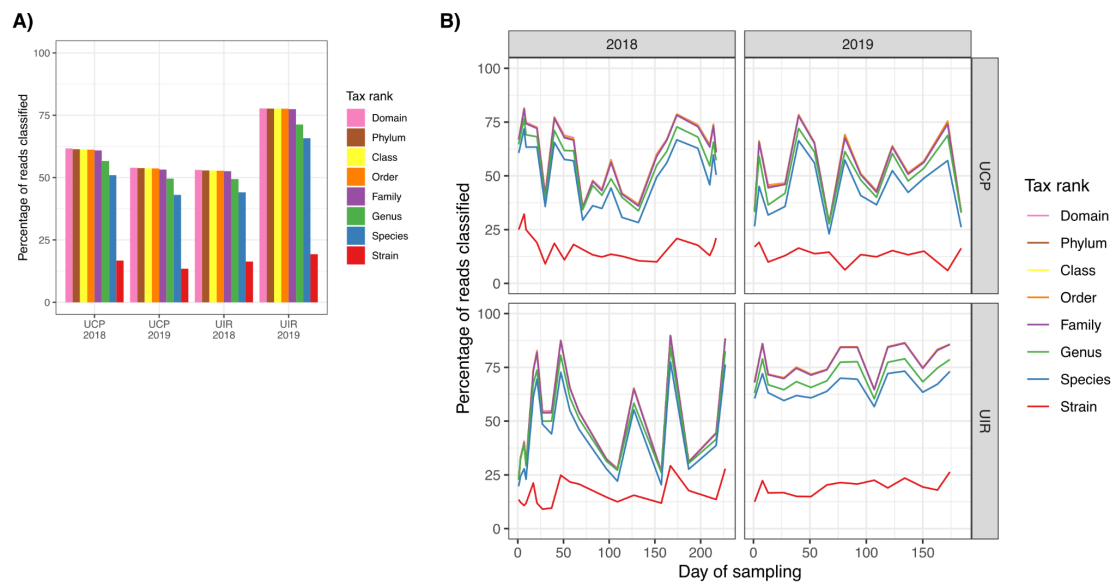
Mill	Metric	Description [unit]
UCP	solids_must	Fraction of sugarcane solids in must [%]
	trs_must	Total reducing sugars in must [%]
	trs_wine	Total reducing sugars in wine [%]
	brix_must	Brix level of must [degrees Brix]
	ktrs_wine	Total reducing sugars in wine, corrected by alcoholic content [unk.]
	kacidity	Acidity of wine, measured by titration [unk.]
	kglycerol	Glycerol content of wine, corrected by unknown factor [%]
	gl_vat	Ethanol content in fermentation vat [degrees Gay-Lussac]
	ph_must	pH of must
	ph_wine	pH of wine
	output	Total weekly ethanol production [liters]
	yield	Net efficiency of production of ethanol from sugar [%]
UIR	yeast_viability	Fraction of viable yeasts in fermentation [%]
	budding_rate	Fraction of budding yeast cells in fermentation [%]
	budding_viability	Fraction of viable buds [%]
	floc	Fraction of flocculated cells [%]
	rods	Amount of rod bacteria [rods/ml]
	operation_time	Fraction of time pipeline was operational [%]
	yield	Net efficiency of production of ethanol from sugar [%]
	trs_sugarcane	Total reducing sugars in sugarcane [%]
	output	Total weekly ethanol production [liters]

Supplementary Table 2. Linear regression of UCP industry metrics on *Lactobacillus* and *Limosilactobacillus* abundances. Significance of effect estimates: * : $p < 0.1$, ** : $p < 0.05$, *** : $p < 0.05/12$ as a Bonferroni-corrected significance threshold.

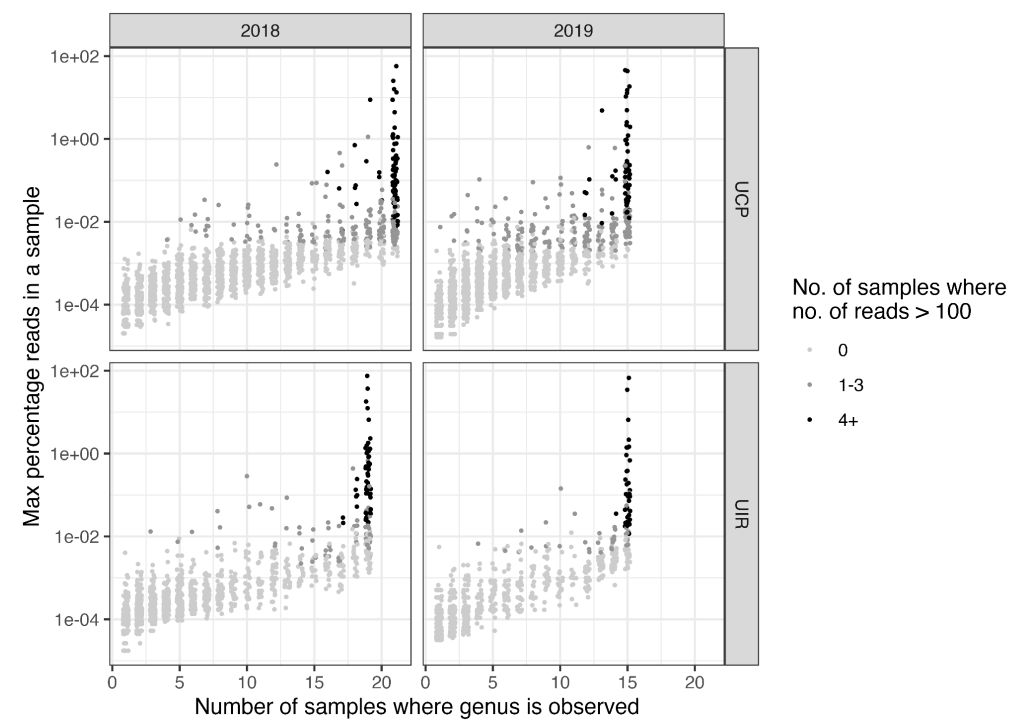
Response	Lactobacillus			Limosilactobacillus			R^2	Adj. R^2
	β (std. err.)	$t_{df=32}$	p	β (std. err.)	$t_{df=32}$	p		
solids.must	5×10^{-3} (1.2×10^{-2})	4.3×10^{-1}	0.672	8.9×10^{-3} (1.6×10^{-2})	5.7×10^{-1}	0.573	0.05	-0.03
trs.must	5.1×10^{-3} (8.2×10^{-3})	6.2×10^{-1}	0.537	-8.8×10^{-3} (1.1×10^{-2})	-7.9×10^{-1}	0.435	0.06	-0.02
trs.wine	3.2×10^{-3} (9.1×10^{-3})	3.5×10^{-1}	0.73	-1.5×10^{-2} (1.2×10^{-2})	-1.2	0.233	0.29	0.22
brix.must	8.2×10^{-3} (8.4×10^{-3})	9.7×10^{-1}	0.338	-8.1×10^{-3} (1.1×10^{-2})	-7.2×10^{-1}	0.48	0.10	0.01
kacidity	4×10^{-2} (9.1×10^{-3})	4.4	< 0.001***	-2.4×10^{-2} (1.2×10^{-2})	-1.9	0.067*	0.41	0.35
ktrs	1.6×10^{-3} (8.1×10^{-3})	2×10^{-1}	0.846	-1.6×10^{-2} (1.1×10^{-2})	-1.5	0.144	0.21	0.14
kglycerol	-1.1×10^{-2} (1.1×10^{-2})	-1	0.303	1.1×10^{-2} (1.5×10^{-2})	7.4×10^{-1}	0.463	0.06	-0.03
gl.vat	8.6×10^{-3} (1.1×10^{-2})	7.8×10^{-1}	0.443	-8.9×10^{-3} (1.5×10^{-2})	-5.9×10^{-1}	0.557	0.22	0.15
ph.must	-9×10^{-3} (1.1×10^{-2})	-8.4×10^{-1}	0.409	1.4×10^{-3} (1.5×10^{-2})	9.5×10^{-2}	0.925	0.12	0.03
ph.wine	-2.8×10^{-2} (9.1×10^{-3})	-3	0.005**	1.5×10^{-2} (1.2×10^{-2})	1.2	0.229	0.28	0.21
output	-1.3×10^{-2} (9.4×10^{-3})	-1.3	0.188	-6.8×10^{-3} (1.3×10^{-2})	-5.3×10^{-1}	0.599	0.07	-0.02
yield	3×10^{-3} (2×10^{-3})	1.5	0.146	2.5×10^{-3} (2.8×10^{-3})	9.1×10^{-1}	0.367	0.18	0.10

Supplementary Table 3. Linear regression of UIR industry metrics on *Lactobacillus* and *Limosilactobacillus* abundances. Significance of effect estimates: * : $p < 0.1$, ** : $p < 0.05$, *** : $p < 0.05/9$ as a Bonferroni-corrected significance threshold.

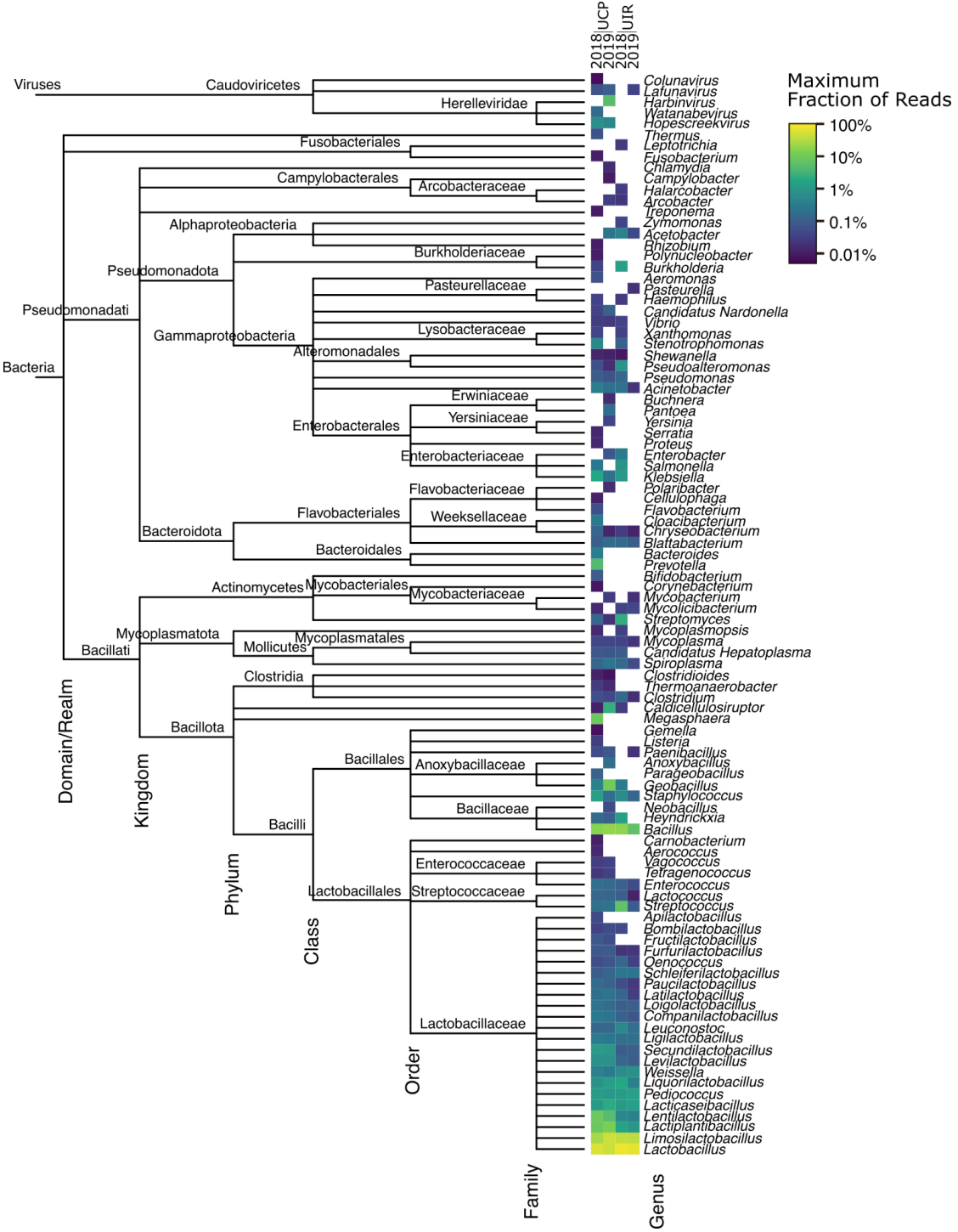
Response	Lactobacillus			Limosilactobacillus			R^2	Adj. R^2
	β (std. err.)	$t_{df=32}$	p	β (std. err.)	$t_{df=32}$	p		
yeast.viability	-2.5×10^{-2} (3.3×10^{-3})	-7.3	< 0.001***	-1.4×10^{-2} (6.5×10^{-3})	-2.1	0.036**	0.39	0.38
budding.rate	1.5×10^{-2} (4.1×10^{-3})	3.6	< 0.001***	1.1×10^{-2} (7.9×10^{-3})	1.4	0.165	0.11	0.09
budding.viability	-1.8×10^{-2} (3.7×10^{-3})	-4.9	< 0.001***	-3.3×10^{-2} (7.1×10^{-3})	-4.6	< 0.001***	0.27	0.26
floc	2.6×10^{-2} (3.6×10^{-3})	7.2	< 0.001***	7.3×10^{-3} (7.1×10^{-3})	1	0.3	0.29	0.27
rods	2.1×10^{-2} (3.8×10^{-3})	5.5	< 0.001***	2.7×10^{-2} (7.5×10^{-3})	3.6	< 0.001***	0.20	0.19
operation.time	2.4×10^{-3} (4.4×10^{-3})	5.5×10^{-1}	0.587	2.6×10^{-2} (8.8×10^{-3})	2.9	0.004***	0.11	0.09
yield	-8.2×10^{-3} (4.4×10^{-3})	-1.9	0.066*	1.5×10^{-2} (8.8×10^{-3})	1.7	0.095*	0.10	0.07
atr	6.8×10^{-3} (4.5×10^{-3})	1.5	0.137	-8.7×10^{-3} (9.1×10^{-3})	-9.6×10^{-1}	0.34	0.05	0.02
production	5.3×10^{-3} (4.2×10^{-3})	1.3	0.214	3.1×10^{-2} (8.5×10^{-3})	3.7	< 0.001***	0.17	0.15



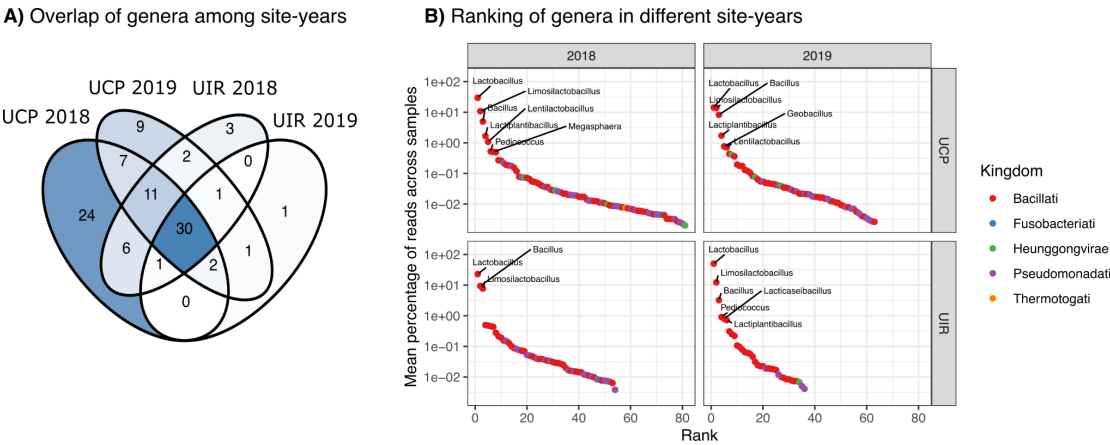
Supplementary Figure 1. Percentage of reads classified down to different taxonomic levels by Kraken by (A) site-year, and (B) across samples.



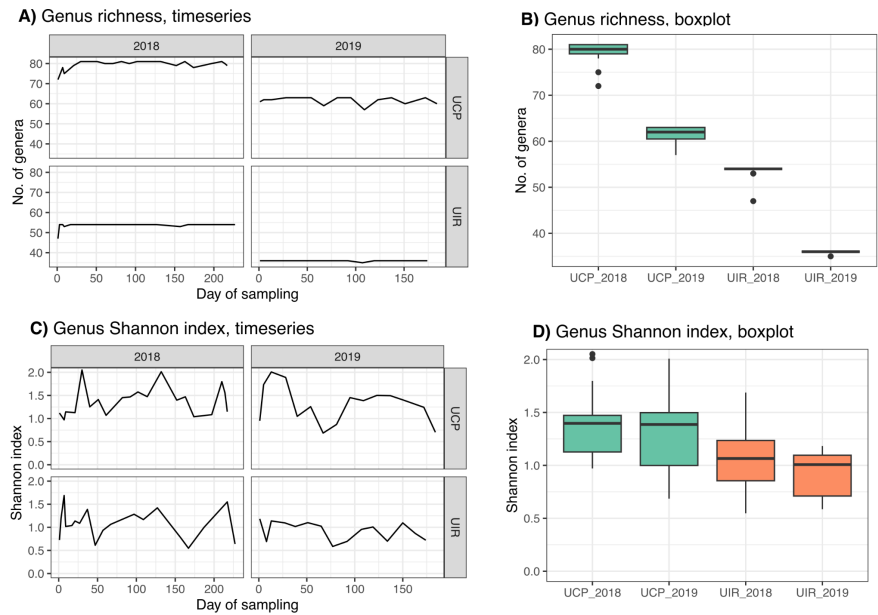
Supplementary Figure 2. Selection of high-confidence genera across all site-years based on minimum read count and number of samples. Each point represents a genus found at a specific site-year.



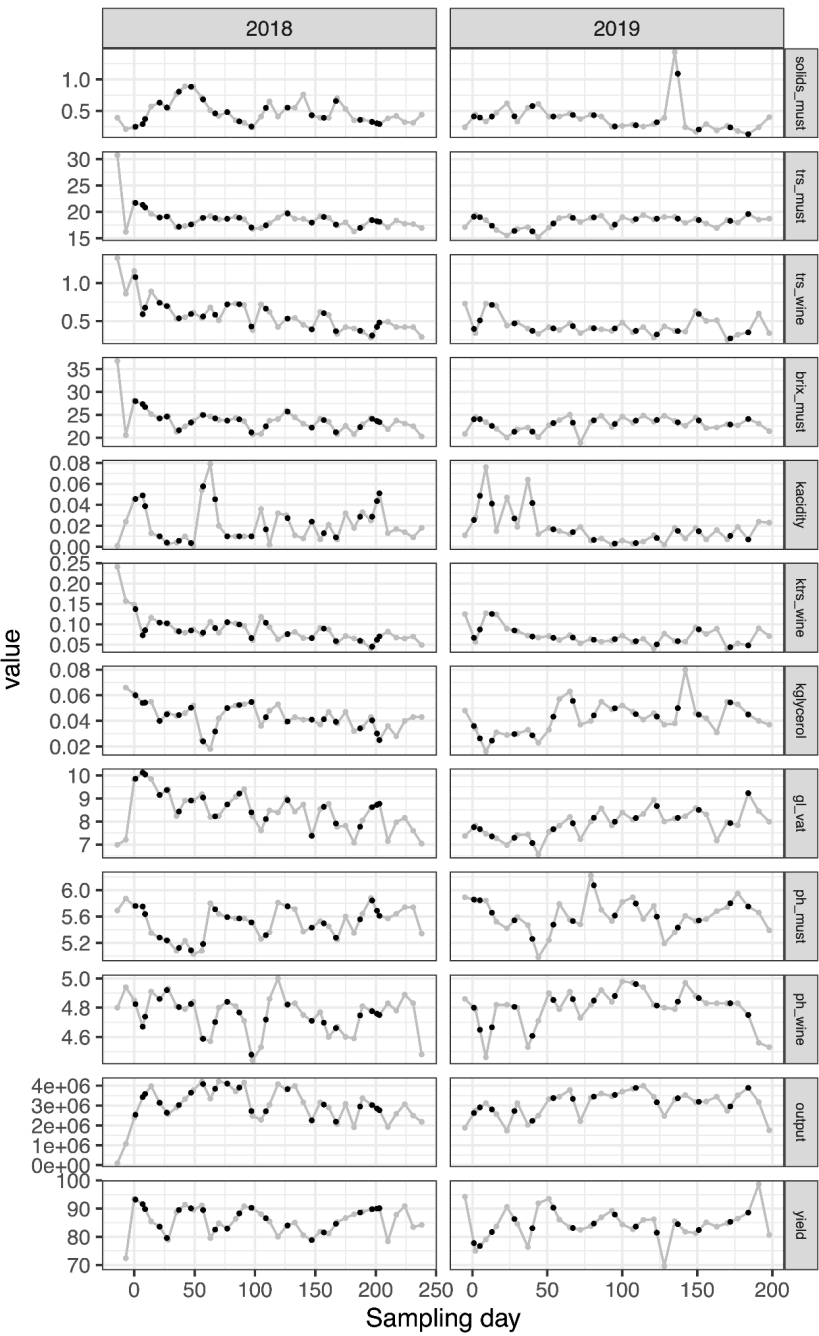
Supplementary Figure 3. Cladogram of all high-confidence genera selected in the study across all site-years. The maximum abundance of each genus is indicated with a heat map. White tiles in the heatmap indicate genera not observed in that site-year.



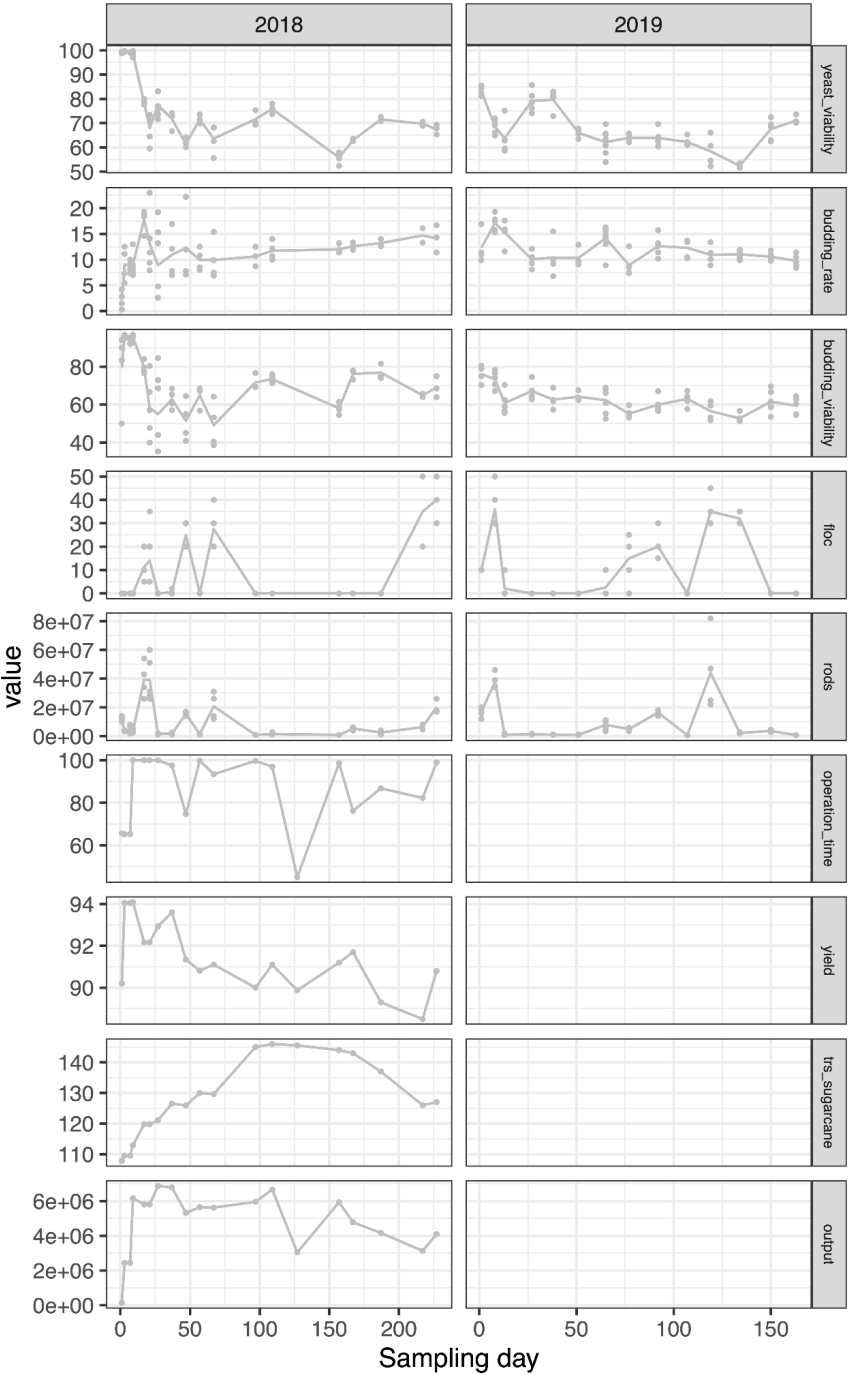
Supplementary Figure 4. Selected high-confidence genera. (A) Venn diagram of genera among the 4 different site-years. Shading is proportional to the number of genera in each cell. (B) Mean abundance of high-confidence genera selected in each site-year. Each point is a genus, ranked by mean fraction of reads classified to it across all samples in that site-year. Points are colored based on Kingdom classification. Some high abundance genera are labeled.



Supplementary Figure 5. (A, B) Genus richness and (C, D) Shannon index throughout the fermentation season and in each site-year.

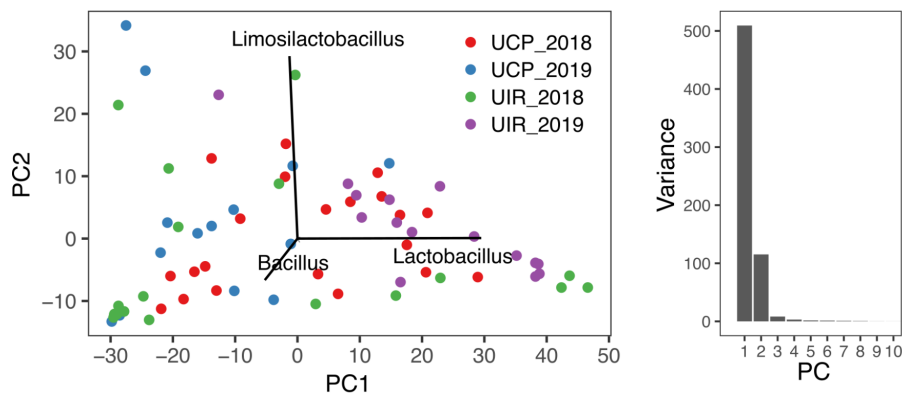


Supplementary Figure 6. Industrial metrics shared by UCP from the 2018 and 2019 fermentation seasons. Shared data is shown in grey. Linearly-interpolated data points are shown in black and represent estimates of metrics on the days of metagenomic sampling. See Supplementary Table 1 for a definition of the metrics.

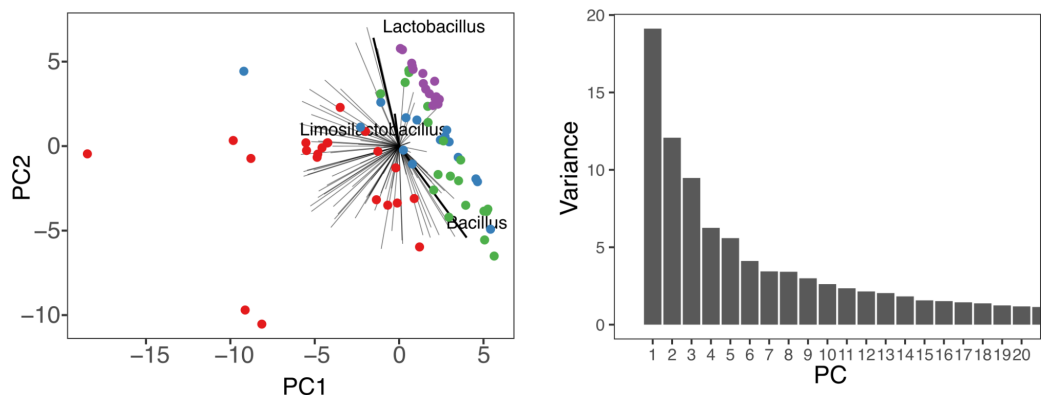


Supplementary Figure 7. Industrial metrics shared by UIR from the 2018 and 2019 fermentation seasons. Line represents the average value at each day of sampling. See Supplementary Table 1 for a definition of the metrics.

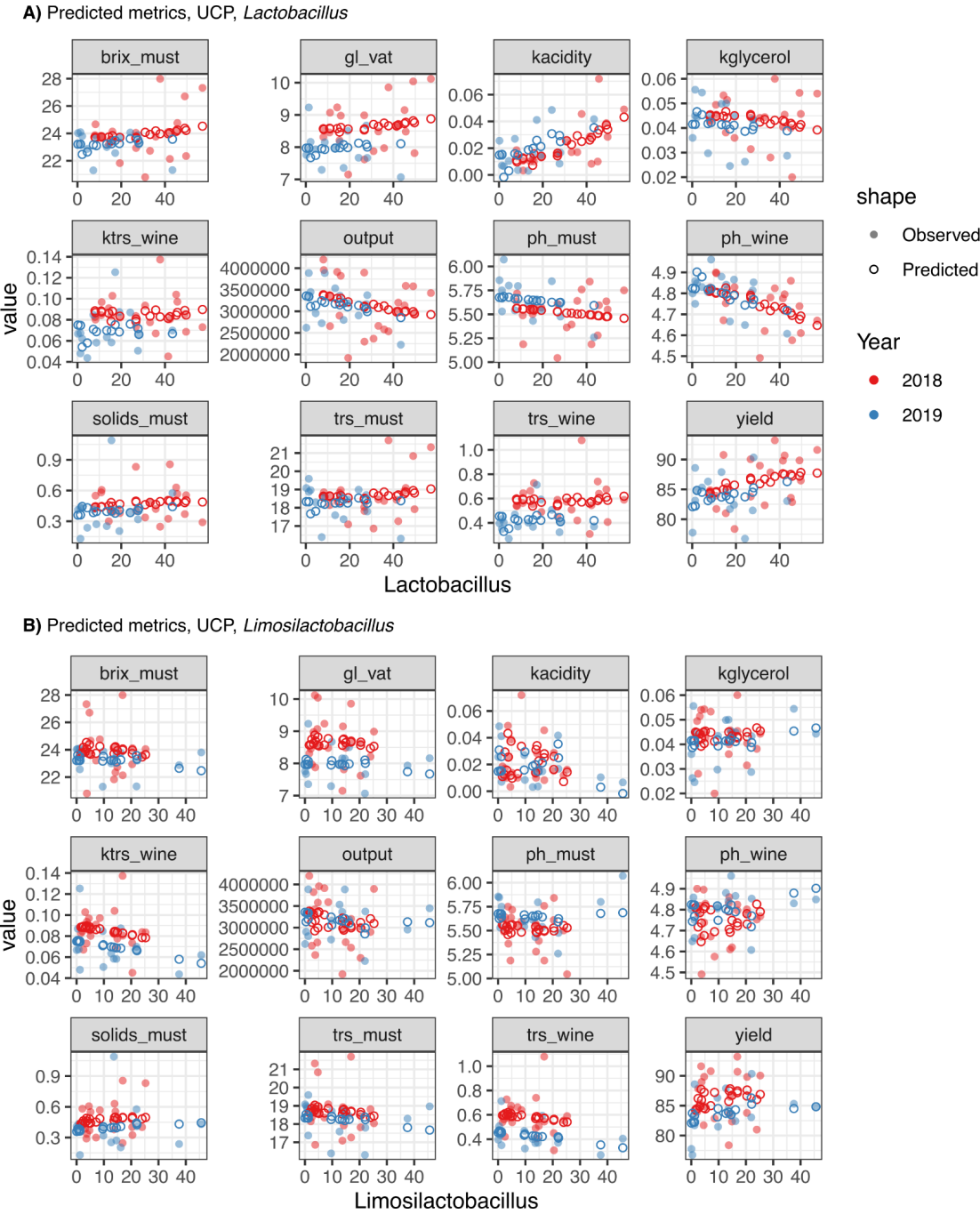
A) PCA, unscaled read fraction data



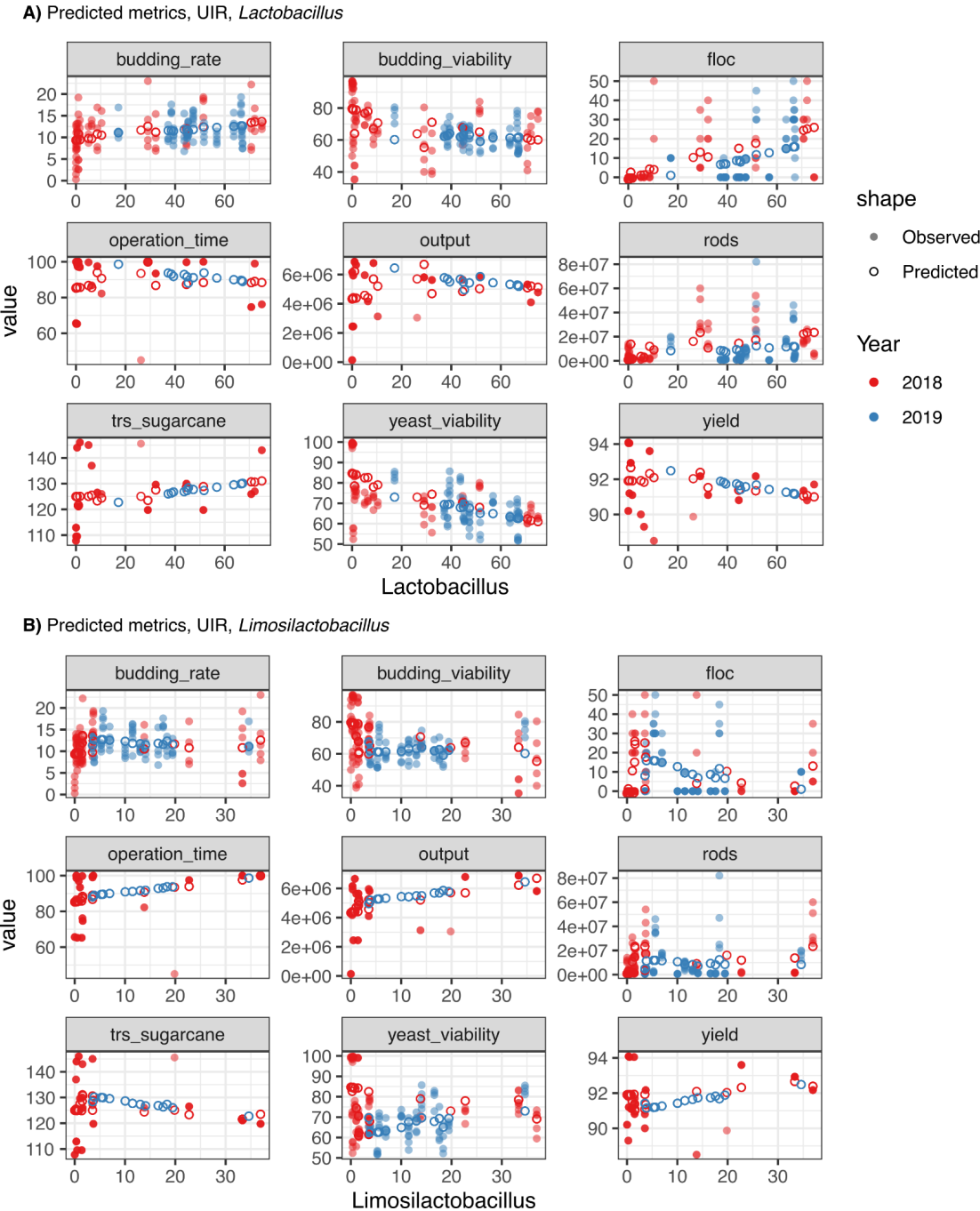
B) PCA, scaled read fraction data (not significant)



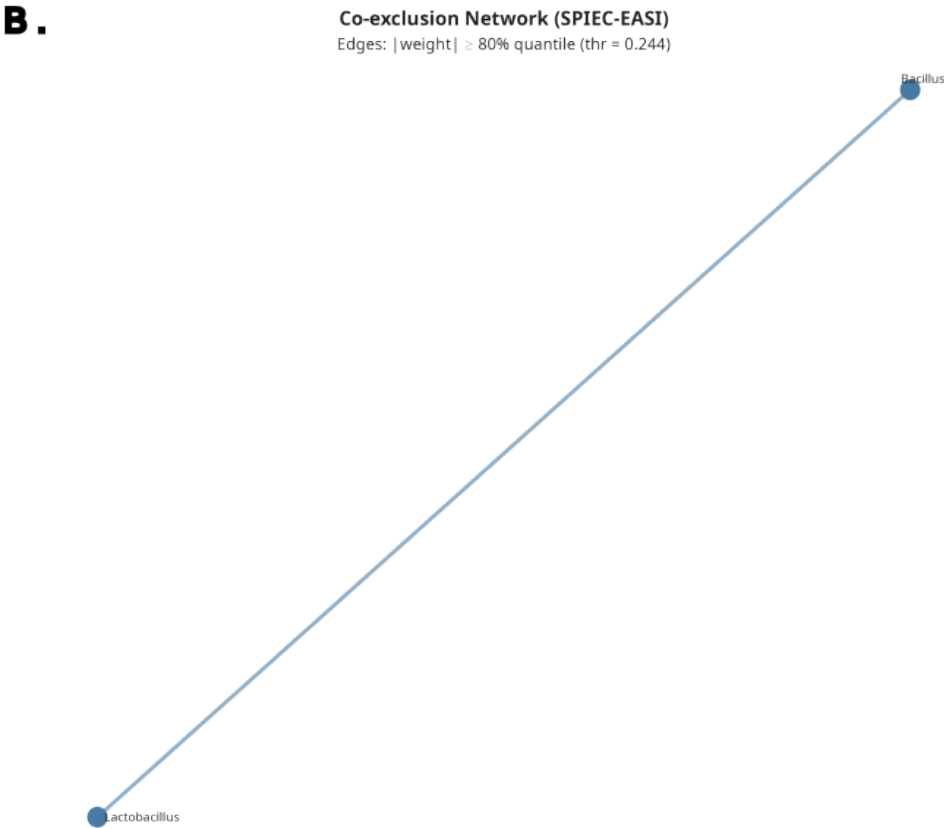
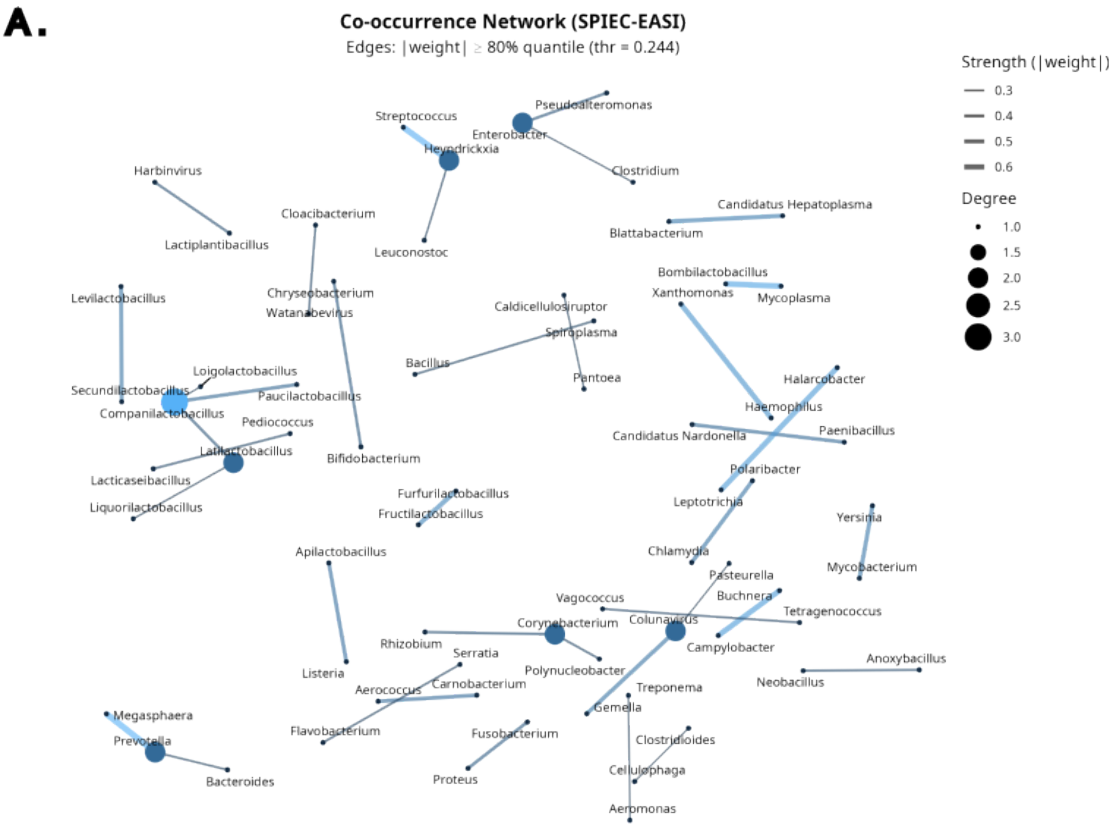
Supplementary Figure 8. Principal component analysis (PCA) of microbial community composition, using either (A) unscaled or (B) scaled genus compositional data. Samples are ordinated along the first two principal components (PC1 and PC2), colored by site and year (UCP_2018, UCP_2019, UIR_2018, UIR_2019). The vectors indicate the PCA loadings, interpreted as the taxa whose abundance contributes to each principal component. In the unscaled PCA, *Lactobacillus* and *Limosilactobacillus* each define a PC, after which a few with community variation, including *Bacillus*. The separation of samples along PC1 highlights temporal and site-specific differences in microbial community structure, with UIR_2019 samples clustering distinctly from other groups.



Supplementary Figure 9. Industrial metrics values against (A) *Lactobacillus* and (B) *Limosilactobacillus* abundance in UCP. Observed values are shown in light shade full circles, while the corresponding predicted values from the linear regression models are shown in empty circles.



Supplementary Figure 10. Industrial metrics values against (A) *Lactobacillus* and (B) *Limosilactobacillus* abundance in UIR. Observed values are shown in light shade full circles, while the corresponding predicted values from the linear regression models are shown in empty circles.



Supplementary Figure 11. Network analysis of microbial interactions constructed using the SPIEC-EASI. (A) Co-occurrence Network. The topology reveals a decentralized community structure characterized by discrete functional subnetworks rather than a single highly connected core. A prominent subnetwork comprises the newly reclassified *Lactobacillaceae* (e.g., *Companilactobacillus*, *Latilactobacillus*), reflecting shared ecological preferences. Other clusters include soil module (e.g., *Corynebacterium*, *Polynucleobacter*, *Rhizobium*), and diverse insect/plant commensals. (B) Co-exclusion Network. This subnetwork highlights the *ecological* antagonism between *Bacillus* and *Lactobacillus*.