Deciphering the Microbial Complexity of Chinese Traditional Sourdough through Integrated Learning

Yan Liu[1], Jinyu Wang[1], Shiwei Zhou[2], Feijiang Li[3] , Honghong Cheng[4], Zheqing Zhu[3], Jiale Wang[1], Aowen Liu[1], Ting Lu[1], Yujuan Yu[1], Senmiao Tian[1], Min Zhang[5], Faizan Ahmed Sadiq (Co-Corresponding Author)[*], and Guohua Zhang(Co-Corresponding Author)[*]

1 School of Life Science, Shanxi University, Taiyuan 030006, China.

2 Lesaffre Management (Shanghai) Co., Ltd, Shanghai 200030, China.

3 Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China.

4 School of Information, Shanxi University of Finance and Economics, Taiyuan 030006, China.

5 Research Institute of Applied Biology, Shanxi University, Taiyuan 030006, China.

* Correspondence:

Advanced Therapies Group, School of Dentistry, Cardiff University, Cardiff CF14 4XY, United Kingdom SadiqF@catdiff.ac.uk (F.S.).

School of Life Science, Shanxi University, Taiyuan 030006, China; zhanggh@sxu.edu.cn (G.Z.); Tel.: +86-15513091052.

Faizan Ahmed Sadiq and Guohua Zhang, these authors contributed equally as co-corresponding authors

Abstract:

Sourdough fermentation relies on a dynamic microbial consortium of fungi and bacteria, shaped by interspecies interactions and environmental conditions. We analysed the microbiota and physicochemical characteristics of 115 sourdough samples – including both 26 dry samples and 89 wet samples – collected across ten Chinese provinces, and used integrated learning approaches to correlate microbial composition with physicochemical properties, revealing associations that underpin fermentation outcomes and regional variation in community structure. In dry samples, *Pediococcus pentosaceus*（39.46%）and *Levilactobacillus brevis*（11.02%）dominated, whereas wet samples were dominated by *Fructilactobacillus sanfranciscensis* (51.27%). *Saccharomyces cerevisia*e was recognised as a dominant and core yeast species (detected in over 50% of all samples and ranking among the top 5 in any groups) in the sourdough microbiota. *Saccharomycopsis fibuligera, Lactiplantibacillus plantarum* and *Leuconostoc mesenteroides* were core bacterial species. Ensemble learning and abundance ranking identified rare taxa such as *Dermacoccus nishinomiyaensis*, *Kodamaea ohmeri* and *Buckleyzyma phyllomatis* that, despite their low abundance in the sourdough microbiota, showed significant correlations with physicochemical indices, although their functional roles remain uncharacterised. Microbial community structure correlated strongly with physicochemical conditions, including moisture content, pH and total acidity, underscoring the importance of considering these parameters when designing stable synthetic consortia for sourdough fermentation. This study provides foundational insights into the structure and function of sourdough microbiomes across diverse regions and sample types. By integrating microbiota data with physicochemical characteristics, we demonstrate the potential of machine learning to uncover key microbial-environment relationships. These findings support the rational design of stable, tailored starter cultures for improved sourdough fermentation.

Keywords: Sourdough; Fungi; Bacteria; Microbial diversity; Ensemble learning; Random forest

Introduction

Sourdough fermentation is among the oldest forms of human-guided microbial biotechnology. Traditional sourdough, a naturally fermented blend of flour and water, has supported breadmaking for millennia through the activity of complex communities of lactic acid bacteria (LAB) and yeasts (Chavan et al., 2011; Landis et al., 2021). Its global significance is evident in ancient records, including early evidence from China. By the Yuan Dynasty (1271 to 1368), sourdough techniques in China closely resembled modern practices, underscoring its historical and cultural importance in food science (Su, 2009).

Sourdough fermentation involves a diverse microbial consortium, primarily comprising LAB, yeasts, and moulds, that interact to form a complex, synergistic system (Calabrese et al., 2022; Coda et al., 2014; De Vuyst et al., 2016; GäNzle & Ripari, 2016; Gnäzle & Gobbetti, 2013; Gobbetti et al., 2016; Siepmann et al., 2018; Van Kerrebroeck et al., 2017; De Vuyst et al., 2014). Extensive research, both in China and internationally, has confirmed this microbial composition, with LAB typically outnumbering yeasts by a ratio of approximately 100:1 (Gobbetti, 1998; Minervini et al., 2012).

To date, more than 70 LAB species, predominantly heterofermentative, have been isolated from sourdough ecosystems (Fu et al., 2022; Van Kerrebroeck et al., 2017). Among these, *Fructilactobacillus sanfranciscensis* (formerly known as *Lactobacillus sanfranciscensis*) is consistently dominant in sourdough but rarely found in other environments (Liu et al., 2018). *F. sanfranciscensis* is commonly found on cereal grains and can enter the sourdough ecosystem via flour. Intriguingly, recent studies have also detected this species in the frass (fecal matter) of grain-feeding insects such as *Tribolium* spp., suggesting insects may act as a natural reservoir and vector into the sourdough process (Boiocchi et al., 2017). Other prevalent LAB in sourdough include *Lactiplantibacillus plantarum*, *Levilactobacillus brevis*, *Lactobacillus paralimentarius*, *Lactobacillus fermentum*, and *Pediococcus pentosaceus* (Liu et al., 2018; Minervini et al., 2015). In addition, some homofermentative lactic acid bacteria, such as

Lactobacillus acidophilus, *Pediococcus acidilactici*, and *Lactobacillus delbrueckii*, have also been found in the sourdough microbiota (Corsetti & Settanni, 2007). These bacteria convert hexose into lactic acid via the glycolysis pathway. Compared to LAB, the sourdough ecosystem harbors a less diverse range of yeast species. *Saccharomyces cerevisiae*, *Pichia anomala*, *Maudiozyma humilis*, *Torulaspora delbrueckii*, *Wickerhamomyces anomalus*, *Kazachstania exigua*, and *Pichia kudriavzevii* are among the most common yeasts found in sourdough. Although *S. cerevisiae* predominates, it is not unique to sourdough, its abundance reflects the use of commercial starters that diversify the microbial community. The widespread use of commercial yeast starters has enriched the microbial composition of sourdough (Gino et al., 2010; Minervini et al., 2015). In contrast, *M. humilis* and *K. exigua* are considered primary sourdough yeasts, frequently associated with traditional fermentation practices (De vuyst et al., 2016; Liu et al., 2018; Mielecki et al., 2024).

Globally, according to the applied processing technology, sourdough can be classified into four types. Type 0 sourdough refers to sponge or pre-dough processes, in which the flour–water mixture undergoes limited fermentation, allowing naturally occurring LAB from flour and baker's yeast contaminants to proliferate. Type I, the most prevalent, uses a portion of fermented dough as a starter for subsequent batches (Chavan et al., 2011). Types II and III comprise liquid and dried powdered industrial forms, respectively (Chavan et al., 2011; De Vuyst et al., 2023). In China, sourdough production diverges into two categories: Jiaozi, derived from wheat, corn, or rice flour supplemented with microbial starters (e.g., Daqu, Xiaoqu, or Hongqu), undergoes multiple fermentations and drying, yielding a low-moisture, shelf-stable product; and Laomian, akin to Type I, employs a similar iterative dough-based process(Yan et al., 2019).

These microbial interactions generate a repertoire of metabolites that enhance the quality of fermented products, contributing to improved flavour, aroma, colour, and nutritional value (Ma et al., 2024; Han et al., 2016). In recent years, the global rise of green and healthy lifestyles has contributed to the growing popularity of sourdough-fermented products as a dietary trend. Multi-strain fermentation in sourdough

outperforms commercial yeast fermentation, improving shelf life, flavour, vitamin and mineral content, glycemic control, and texture (Alfonzo et al., 2013; Hammes et al., 2005; Katina et al., 2005; Wang et al., 2020; Zhang et al., 2015). Microbial fermentation produces functional compounds like peptides, organic acids, and prebiotics, which can support digestion, gut health, and lower the glycaemic index (Clarke & Arendt, 2005; Gäenzle, 2014). Scaling up sourdough fermentation may help unlock these health and industrial benefits.

Although extensive research has shown that the diversity of sourdough microbiota directly influences product quality and aroma profiles (Landis et al. 2021; Liu et al., 2018; Zhang et al., 2019), the link between microbial composition and key physicochemical properties – such as moisture content and pH – across different sourdough types remains poorly understood. This knowledge gap hinders the development of standardised, safe synthetic microbial communities for controlled fermentation. Naturally fermented sourdough, whether homemade or workshop-processed (Liu et al., 2018), faces variability and safety challenges due to its open system, complicating industrial standardisation. Sequencing generates high-dimensional microbial data that challenges traditional statistical methods. Machine learning (ML), particularly ensemble approaches like random forest (Fernandez et al., 2014), excels at handling such complexity by reducing overfitting and improving accuracy. Random forests classify samples and highlight key microbial taxa, offering robust and interpretable results (Medina et al., 2022). These models not only classify microbial communities but also correlate key taxa with physicochemical traits such as pH, moisture, and acidity. In this study, ML helped link microbiota composition to fermentation characteristics, enabling the identification of core species and informing the design of stable synthetic communities for controlled sourdough production.

Materials and methods

Sample Collection and Processing

A total of 115 sourdough samples were collected from ten provinces in China. The geographic distribution of the samples is presented in Fig. S1, and detailed sample

information is provided in Supplementary Table. 1. Following collection, approximately 2 g of each sourdough sample was aseptically collected under a laminar flow hood. Samples were thoroughly homogenized using sterile tools to ensure uniformity. The collected aliquots were immediately transferred into sterile, DNA-free tubes and temporarily kept on ice. All samples were subsequently stored at -80°C until DNA extraction and high-throughput sequencing. The remainder of each sample was stored at 4°C for physical and chemical analysis, with sequencing samples processed within one week.

Determination of physical and chemical parameters of sourdough

Moisture content was determined by the direct drying method at 101–105°C according to the national standard GB 5009.3-2016, using constant weight as the endpoint. Moisture content (X) was calculated using the formula:

$$X=\frac{\mathrm{m}_1\text{-}m_2}{m_1\text{-}m_3}\times 100$$

X = moisture content (g/100 g),

m_1 = mass of the bottle and sample before drying (g),

m_2 = mass of the bottle and sample after drying (g),

m_3 = mass of the empty bottle (g).

Samples were categorized as “dry” or “wet” based on their observable physical state and preparation tradition, which correspond to low and high moisture content regimes, respectively. The measured moisture contents for each group are provided in Supplementary Table 2 and 3.

For pH measurement, 10.0 g of sourdough was suspended in carbon dioxide-free water, stirred magnetically for 10 min, then diluted to 100 mL with the same solvent. pH was measured in triplicate using a calibrated pH meter.

Total titratable acidity (TTA) was determined by titration with 0.1 mol/L NaOH to an endpoint of pH 8.2 and calculated as the lactic acid mass fraction (g/kg), with a blank correction applied (Liu et al., 2016; Wu et al., 2012).

$$X=\frac{c\times\ (V_1-V_2)\ \times 90}{m\times 1000}\times 1000$$

X — Total titratable acidity, as lactic acid (g/L)

c — Concentration of NaOH standard solution (mol/L)

V_1 — Volume of NaOH consumed by the test sample (mL)

V_2 — Volume of NaOH consumed by the blank (mL)

m — Volume of the test sample taken (mL)

90 — Molar mass of lactic acid (g/mol)

1000 — Conversion factor

Isolation and Identification of LAB and Yeast

Take 1 g of sourdough sample, add 9 mL of sterile water in a clean bench, mix thoroughly, and perform a tenfold serial dilution. Take 0.1 mL of the dilution from the three dilution gradients of 10^{-4}, 10^{-5}, and 10^{-6}, and evenly spread it onto MRS solid medium (containing 50 mg/L Cycloheximide) and YPD solid medium (containing 50 mg/L Chloramphenicol). The plates were incubated anaerobically at 37°C for 24 h and 30°C for 48 h, respectively (Xing et al., 2020).

Characteristic colonies were randomly selected from MRS solid medium and subjected to multiple rounds of purification until single colonies were obtained. Gram staining and the catalase test were performed to screen for suspected lactic acid bacteria strains. After enrichment in MRS liquid medium for 24 h, the cultures were divided into two portions. One portion was supplemented with 30% glycerol and stored at -80°C, while the other was kept at 4°C for subsequent bacterial identification.

Characteristic colonies from the YPD solid medium were randomly selected and purified through 2–3 rounds of subculturing until only the target strain were observed under the microscope. Cell morphology was examined using 0.1% Lu's alkaline methylene blue staining solution. The purified yeast isolates were enriched in YPD liquid medium for 24 h before being divided into two portions. One portion was preserved at -80°C with 30% glycerol, while the other was stored at 4°C for further strain identification.

DNA was extracted from the isolated strains using a DNA extraction kit. The

extracted DNA was sent to Sangon Biotech (Shanghai) Co., Ltd. for PCR amplifications and sequencing, bacterial identification was performed using primers 27F (5'-AGAGTTTGATCMTGGCTCAG-3') and 1492R (5'-GGTTACCTTGTTACGACTT-3') to target the nearly full-length 16S rRNA gene. Fungal identification was carried out with primers ITS1 (5'-TCCGTAGGTGAACCTGCGG-3') and ITS4 (5'-TCCTCCGCTTATTGATATGC-3') to amplify the ITS1-5.8S-ITS2 region. All PCR amplifications were performed under standardized conditions followed by Sanger sequencing. And the obtained sequences were analyzed using BLAST homology searches against the NCBI database. Strains with a sequence similarity of 97% or higher were identified at the species level.

Microbial diversity sequencing

Genomic DNA was extracted from sourdough samples using a commercial kit for subsequent microbial community sequencing (Shanghai Meiji Biopharmaceutical Technology Co., Ltd.), and its quality was verified by 1% agarose gel electrophoresis. For bacteria, the V3–V4 hypervariable region of the 16S rRNA gene was amplified using specific primers with barcodes: Forward primer (338F): 5'-ACTCCTACGGGAGGCAGCAG-3', Reverse primer (806R): 5'-GGACTACHVGGGTWTCTAAT-3'. For fungi in the sample, the Forward primer ITS1F (5'-CTTGGTCATTTAGAGGAAGTAA-3') and the Reverse primer ITS2R (5'-GCTGCGTTCTTCATCGATGC-3') were selected to perform PCR amplification of the ITS region. PCR products were verified using 2% agarose gel electrophoresis, and amplicons were gel-purified. A paired-end (PE) sequencing library was then constructed, followed by sequencing on the Illumina MiSeq PE300 platform. All sequencing was conducted at Shanghai Meiji Biopharmaceutical Technology Co., Ltd.

Raw FASTQ files were de-multiplexed using an in-house perl script, and then quality-filtered by fastp version 0.19.6 (Chen et al., 2018) and merged by FLASH version 1.2.7 (Mago & Salzberg, 2011) with the following criteria: (1) the reads were truncated at any site receiving an average quality score of < 20 over a 50 bp sliding window, and the truncated reads shorter than 50 bp were discarded, reads containing

ambiguous characters were also discarded; (2) only overlapping sequences longer than 10 bp were assembled according to their overlapped sequence. The maximum mismatch ratio of overlap region is 0.2. Reads that could not be assembled were discarded; (3) Samples were distinguished according to the barcode and primers, and the sequence direction was adjusted, exact barcode matching, 2 nucleotide mismatch in primer matching. Then the optimized sequences were clustered into operational taxonomic units (OTUs) using UPARSE 7.1 (Edgar, 2013; Stackebrandt & Goebel, 1994) with 97% sequence similarity level. The most abundant sequence for each OTU was selected as a representative sequence. The OTU table was manually filtered, i.e., chloroplast sequences in all samples were removed. To minimize the effects of sequencing depth on alpha and beta diversity measure, the number of 16S rRNA gene sequences from each sample were rarefied to 20,000, which still yielded an average Good's coverage of 99.09% , respectively.

OTU species taxonomy annotation was performed using RDP classifier (Wang et al., 2007) (http://rdp.cme.msu.edu/, version 2.11). Bacterial annotation was compared with the Silva 16S rRNA gene database (v138), and fungal annotation was compared with the UNITE fungal ITS database (v8.0). The confidence threshold was 70%, and the community composition of each sample was statistically analyzed at different species classification levels.

Abundance sorting

Species importance within the sourdough microbial community was determined based on their relative abundance across samples. Abundance values for each species, annotated during high-throughput sequencing, were calculated and ranked from highest to lowest. The SUM method was used to identify key species, with the top 50 and top 15 representing the most abundant species across all samples.

Ensemble learning

To analyse high-dimensional sequencing data, we applied Uniform Manifold Approximation and Projection (UMAP), a widely used nonlinear dimensionality

reduction algorithm (Han et al., 2001; Hastie et al., 2008). UMAP preserves both local and global structures by estimating distances in high-dimensional space and mapping them onto a low-dimensional manifold while maintaining relative relationships between data points (Mcinnes et al., 2018). This approach enables effective separation of microbial communities and captures nonlinear patterns, improving the resolution of microbial diversity analysis. The characteristics of the UMAP algorithm determine that it can retain both local and global structures, learn nonlinear patterns from data, and better separate samples.

The Random Forest algorithm was applied to perform regression analysis, where discrete strain data served as the independent variable, and physical and chemical index values were used as the dependent variable for function fitting. The algorithm outputs feature weights, ranking strains by importance based on their contribution to the model. In the regression analysis using a random forest model, the OTU data matrix and the physicochemical indicators from the microbial diversity sequencing of the sourdough samples were used as input data. The model was implemented with the default parameter n_estimators=100, meaning 100 decision trees were constructed. The output, feature_importances_, represented the contribution of each species to the model. The species with the highest contributions were identified, with the top 50 and top 15 ranked as the most important species. The model's performance was evaluated using the mean square error (MSE). Additionally, matplotlib and UMAP were applied for result visualization.

As shown in Fig. 1, all sample data were split into training and test sets. Features and samples were randomly selected to generate multiple sub-training sets, and the model underwent multiple rounds of training to develop corresponding base models. Test set data were then input into the base models to evaluate predictive performance, and the final output was derived by aggregating the optimal prediction results (Wilhelm et al., 2022).

Data analysis

Excel was used to perform data statistics on physical and chemical indicators and

calculate the mean and standard deviation. Bioinformatic analysis of the sourdough microbiota was carried out using the Majorbio Cloud platform (https://cloud.majorbio.com) (Han et al., 2024). Based on the OTUs information, rarefaction curves and alpha diversity indices including observed OTUs, Simpson, Shannon, sobs index and Good's coverage were calculated with Mothur v1.30.1 (Schloss et al., 2009).The similarity among the microbial communities in different samples was determined by principal coordinate analysis (PCoA) based on Bray-curtis dissimilarity using Vegan v2.5-3 package. The PERMANOVA test was used to assess the percentage of variation explained by the treatment along with its statistical significance using Vegan v2.5-3 package. The linear discriminant analysis (LDA) effect size (LEfSe) (Segata et al., 2011) (http://huttenhower.sph.harvard.edu/LEfSe) was performed to identify the significantly abundant taxa (genera to species) of bacteria and fungi among the different groups (LDA score > 2, $P < 0.05$). A correlation between two nodes was considered to be statistically robust if the spearman's correlation coefficient over 0.6 or less than -0.6, and the P-value less than 0.01. Random forest ranked the species by importance and presented them in a table. The distribution of important species was reduced to two-dimensional space through UMAP and visualized using Python.

Results

Bacterial diversity of sourdough

To elucidate the microbial disparities between sourdough types, we first assessed the bacterial community diversity. We first assessed alpha diversity differences between dry and wet sourdough samples using inter-group statistical tests. Both the wet and dry samples in our study align with the fundamental characteristics of Type I sourdough, as they are maintained through traditional backslopping processes. Specifically, the dry samples represent a unique Chinese traditional variant where the fermented dough undergoes a final drying step, resulting in a low-moisture, shelf-stable product. Crucially, these dry starters can be reactivated by inoculation into fresh flour and water,

fermenting for a period before potentially being dried again - thereby maintaining the essential "backslopping" characteristic of Type I sourdough. The box plots (Fig. S2) illustrate the distribution and dispersion of microbial diversity, with outliers represented as discrete points. The coverage index (Fig. S2D and S2H) did not differ significantly between groups and approached 1 for most samples, indicating sufficient sequencing depth. Bacterial diversity was lower in wet samples than in dry samples (Fig. S2A and S2B), and bacterial communities in dry samples were more tightly clustered, while those in wet samples were more dispersed (Fig. S2C). Although species richness did not differ significantly between groups, community diversity was significantly higher in dry samples, suggesting greater species evenness.

Analysis of Bacterial community species composition reveals that in dry samples, *P. pentosaceus*, *L. brevis*, *Acetobacter tropicalis*, and *Lp. plantarum* comprised a substantial portion of the bacterial community, with relative abundances of 39.46%, 11.02%, 10.20%, and 9.45%, respectively (Fig. 2A). Sourdough fermentation involves a diverse consortium of LAB and yeasts, with acetic acid bacteria such as *A. tropicalis* occasionally present, though their role remains debated (Sun et al., 2016). *F. sanfranciscensis* was also detected in the dry sample group, but the proportion was relatively small, accounting for only 0.14%. In wet samples, *F. sanfranciscensis* (51.27%), *Weissella cibaria* (9.25%), *P. pentosaceus* (6.52%), and *Leuconostoc mesenteroides* (5.24%) emerged as the dominant taxa (Fig. 2C). *F. sanfranciscensis* was detected in the wet sample group with a Prevalence of 78.65%, and its relative abundance was consistently high across individual wet samples, as indicated by the red portion of the columns in Fig. 2C. In this study, we operationally define dominant species as those constituting >10.00% of the microbial community under specific conditions (e.g., grouping) and exhibited a prevalence of >70% (bacteria or fungi); core species are operationally defined as taxa detected in more than 50% of all sourdough samples (regardless of wet or dry state) and ranking among the top 5 most abundant taxa in both dry and wet sample groups of sourdough microbiota. Therefore, the dominant microbial species in the dry sample group were *P. pentosaceus* and *L. brevis*, whereas only *F. sanfranciscensis* was identified as the dominant species in the

wet sample group; *Lp. plantarum* and *L. mesenteroides* were core bacterial species in the sourdough microbial community.

Through traditional cultivation methods, we successfully isolated 275 LAB strains (Supplementary Table 4), among which *P. pentosaceus* (29 strains), *Lp. plantarum* (42 strains), *F. sanfranciscensis* (50 strains), and *L. brevis* (22 strains) were identified as frequently occurring species. Their high isolation frequency showed strong consistency with the high-throughput sequencing results. These cultivation findings independently validated the reliability of the core microbiota identified by sequencing, while overcoming potential limitations of molecular methods in detecting viable strains. Moreover, the cultured strains not only verified the microbial composition revealed by sequencing but also provided valuable biological resources for subsequent functional studies, such as strain characterization and synthetic community construction, thereby enhancing both the robustness and practical applicability of this study.

The bacterial composition of sourdough was analysed using a Venn diagram, illustrating the proportion of shared and unique species between the two sample groups. Fig. 3A shows that the dry sample group contains 19 unique bacterial species, while the wet sample group harbours 152 unique species, with 191 species common to both groups.

Fig. 3B highlights that *F. sanfranciscensis* and *P. pentosaceus* are the most prevalent species in both dry and wet samples. Fig. 3C and 4D further reveal that *Lactocoocus garvieae* (29.63%), a bacterium associated with intestinal homeostasis, is the most abundant species exclusively present in the wet sample group. In contrast, the dry sample group contains *Acinetobacter soil*, soil-associated species with a high relative abundance.

Species-level differences between dry and wet sourdough samples were assessed using the Wilcoxon rank-sum test (Mann–Whitney U), two-tailed, with FDR correction. Among bacteria, *F. sanfranciscensis* was significantly more abundant in wet samples (47.62% vs 0.12%), while *P. pentosaceus* (32.72% vs 5.83%) and *Lp. plantarum* (8.04% vs 4.30%) were enriched in dry samples (Fig. 4A, $p < 0.001$ for all).

Fungal diversity of sourdough

Following the characterization of bacterial communities, we next profiled the fungal microbiota, which revealed distribution patterns distinct from those of bacteria. Fungal species diversity was significantly higher in wet sourdough samples than in dry ones, as indicated by differences in Simpson and Shannon indices between groups (Fig. S2E, F). Species richness, measured by the Sobs index, was also greater in wet samples, reflecting a notable difference in the number of observed fungal species (Fig. S2G). Inter-group difference tests identified few outliers, which did not influence the overall analysis. While species richness did not differ significantly between groups ($p > 0.05$), community diversity was significantly higher in wet samples, whereas dry samples exhibited greater evenness.

In fungal communities, as shown in Fig. 2B and Fig. 2D, *S. cerevisiae* was detected in 74.87% of dry samples and 55.05% of wet samples. Its relative abundance in dry samples ranged from 28.33% to 99.72%. Notably, *Saccharomycopsis fibuligera* was dominant in SXWY and DWH-1, accounting for 99.93% and 64.86% of the fungal community, respectively. The coverage and relative abundance of *W. anomalus* in dry samples were second only to *S. cerevisiae*. Fig. 2D also indicates that *S. cerevisiae* remained the most prevalent fungal species in wet samples, with *M. humilis* showing high coverage and, in some cases, being the most abundant taxon. *S. cerevisiae* was identified as the dominant fungal species in both dry and wet sample groups, and served as a core microbial member across all sourdough samples. Additionally, *S. fibuligera* was also characterized as a core fungal species.

Traditional cultivation yielded 160 yeast isolates, among which *S. cerevisiae* (60 strains), *M. humilis* (35 strains), and *W. anomalus* (37 strains) were identified as frequently isolated species. These cultivation results showed strong consistency with the fungal community structure revealed by ITS sequencing analysis. Moreover, the obtained yeast strains provide valuable biological resources for further investigation into their functional roles during sourdough fermentation.

Among fungi, *S. cerevisiae* (59.49%) dominate the sourdough microbiota. A comparison of Fig. 3F and 3G reveals that the most abundant unique fungal species in

the wet samples is *Cystofilobasidium macerans* (12.66%). In the dry sample group, *Rosellinia australiensis* (66.96%) emerge as noteworthy unique species.

For fungi, *M. humilis, Aspergillus cibarius*, *Epicoccum nigrum*, and *Cladosporium delicatulum* showed higher abundance in wet samples ($p < 0.001$), whereas *S. cerevisiae*, *W. anomalus*, *Fusarium concentricum*, and *Issatchenkia orientalis* were more prevalent in dry samples ($p < 0.05$; Fig. 4B).

Community heatmap profiling of the top 50 taxa revealed clear distributional differences across groups. Bacterial taxa such as *Lactobacillus zymae*, *Clostridium beijerinckii*, *L. garvieae*, *F. sanfranciscensis*, and *P. pentosaceus* differed significantly in abundance between sample types (Fig. 4C). In fungal communities, both groups harboured high levels of *S. cerevisiae*, *M. humilis*, *Hanseniaspora lachancei*, *C. delicatulum*, and *Candida sake*. *Torulospora sp.*, *Filobasidium sp.*, *M. humilis* and *Cladosporium rugosum* were significantly more abundant in wet samples (Fig. 4D), supporting statistical trends identified by the Wilcoxon test.

Microbial Community Clustering

Having confirmed differences in species abundance, we next employed cluster analysis and PCoA based on beta-diversity distance matrices to assess the overall structural similarity and divergence between dry and wet microbial communities. QIIME 2 (version 2020.2.0) was used to compute the distance matrix and generate a hierarchical clustering tree.

Fig. S3A and S3B show the results of hierarchical clustering using the weighted UniFrac algorithm, which incorporates phylogenetic relationships to calculate distances between samples. The analysis revealed that dry samples primarily clustered into three main groups, although some samples were partially mixed with wet samples. This suggests that, while the overall bacterial community structure differs between dry and wet samples, there is still some degree of compositional similarity. Variability within groups may reflect differences in species richness and evenness.

To further explore the compositional relationships among samples, PCoA was applied. As shown in Fig. S3C (bacterial community) and S3D (fungal community), dry

and wet samples generally formed two distinct but partially overlapping clusters, indicating that while they are compositionally different, complete separation is not achieved.

In the bacterial PCoA plot (Fig. S3C), sample points are more tightly clustered, suggesting greater similarity among bacterial communities. In contrast, the fungal PCoA plot (Fig. S3D) shows a more dispersed distribution, reflecting higher variability among fungal communities. These patterns support the idea that the core microbiota remains relatively stable between dry and wet samples, despite differences in their physical states.

Species-level contributions to similarity and dissimilarity were also examined. Among bacteria, *F. sanfranciscensis*, *P. pentosaceus*, and *Lp. plantarum* contributed most to the shared community structure between sample types. In the fungal community, *S. cerevisiae* and *W. anomalus* played similar roles. Conversely, *A. soli* and *L. garvieae* were major contributors to bacterial differences, while *M. humilis*, *Cystofilobasidiun macerans* and *Rhodotorula australiensis* drove fungal distinctions between wet and dry samples.

Microbiota-Physicochemistry Relationships

To link the observed community structures to environmental variables, we examined the associations between microbial composition and key physicochemical parameters using CCA and Spearman correlation analysis. CCA (Canonical Correspondence Analysis) was used to reveal the relationship between sourdough environmental factors and microbial community composition. Fig. 5A visually illustrates that pH is negatively correlated with TTA and moisture, while TTA and moisture content exhibit a positive correlation. The red arrows in the plot are of similar lengths, suggesting no significant difference in the extent to which these three physicochemical factors influence the bacterial community composition.

Spearman rank correlation analysis was performed between environmental factors and dominant bacterial species. The correlation matrix was visualised as a heatmap to illustrate both the strength and significance of correlations. Fig. 5C highlights that *F.*

sanfranciscensis （ρ = -0.56434, $p < 0.001$）, *L. brevis* （ρ = 0.50975, $p < 0.001$）, and *P. pentosaceus* （ρ = 0.40121, $p < 0.001$） were significantly correlated with pH. Additionally, *F. sanfranciscensis* （ρ = 0.39667, $p < 0.001$）*L. brevis* （ρ = -0.50526, $p < 0.001$）was strongly positively correlated with moisture, whereas *L. brevis* （ρ = -0.50526, $p < 0.001$）, *P. pentosaceus* （ρ = -0.40429, $p < 0.001$）and *Lp. plantarum* （ρ = -0.30529, $p < 0.001$） showed strongly negative correlation with moisture. In contrast, *L. mesenteroides* （ρ = -0.42419, $p < 0.001$） and *Lactobacillus sakei* （ρ = -0.33388, $p < 0.001$）were significantly associated with TTA. Fig. 5B further confirms that moisture content and TTA are negatively correlated with pH, with moisture having the strongest impact on fungal species composition. The length of the arrow representing moisture suggests that sourdough moisture content has the greatest influence, while TTA has the least effect. Additionally, dry samples exhibited a positive correlation with pH but a negative correlation with moisture content. The percentages along the horizontal and vertical axes indicate the variance explained by the two dimensions; however, both values are relatively low, suggesting that the model has a limited ability to explain the observed relationships.

In terms of fungal community composition, Fig. 5D illustrates that *Penicillium cinnamopurpureum* （ρ = 0.54652, $p < 0.001$）, *Cladosporium aggregatocicatricatum* （ρ = 0.53078, $p < 0.001$）, and *Sporidiobolus metaroseus* （ρ = 0.55722, $p < 0.001$） were significantly positively correlated with moisture ($p < 0.01$), whereas *S. cerevisiae* （ρ = -0.2484, $p < 0.01$） showed a significant negative correlation with moisture ($p < 0.01$). Additionally, *F. sp.* （ρ = -0.53728, $p < 0.001$）, *C. delicatulum* （ρ = -0.51899, $p < 0.001$）, *Alternaria infectoria* （ρ = -0.49545, $p < 0.001$）, *S. metaroseus* （ρ = -0.48221, $p < 0.001$） were significantly negatively correlated with pH, while *Thermoascus aurantiacus* （ρ = 0.30412, $p < 0.001$） was positively correlated with

TTA ($p < 0.001$). The heat map analysis of the correlation between environmental factors and microbial species compensated for the limitation of CCA analysis, which did not provide P-values for assessing the relationship between environmental factors and microbial species in sourdough samples. This approach provided clearer quantitative support for understanding the impact and contribution of environmental factors on microbial community composition.

Identification of important bacterial species

Moving beyond relative abundance, we employed a Random Forest ensemble learning approach to identify bacterial taxa with high predictive importance for physicochemical traits. Comparing Appendix Table. 8 and Table. 14, both identification methods detected a large number of bacterial taxa that could not be assigned to the species level (see Figs 6 and 7). Among the important bacterial species identified using the ensemble learning random forest method, 24% were classified at the species level, whereas only 4% were identified using the sum sorting method. However, there was a high degree of overlap at the genus level, with key taxa concentrated in *Leuconostoc*, *Pediococcus*, *Gluconobacter*, *Lactobacillus*, and *Acetobacter*.

A comparative analysis of Table. 14 and Table. 9 reveals that, despite technological constraints, for evolutionarily closely related species, the V3-V4 variable region of 16S rDNA alone may not provide sufficient resolution for accurate differentiation. Some species exhibit over 97% similarity in their 16S rRNA sequences while remaining distinct. Even when the commonly used species-level identification threshold of 97% similarity is met, misclassification or unclassified cases may still occur, some important bacterial species were identified at the genus level. Notably, OTU668, a non-cultured *Malikia* species, ranked among the top 15 important bacterial taxa in both the feature importance score ranking from the random forest analysis (where moisture content and pH were used as feature labels) and the SUM ranking method based on the sum of species abundance. Similarly, OTU143, a member of the genus *Acetobacter*, appeared in the top 15 rankings in both the SUM method and in the pH-based classification, it ranked 9th in the random forest analysis and 5th in the SUM abundance ranking,

suggesting that this bacterium made a substantial contribution to the pH of the sourdough samples based on its ranking. Additionally, its total abundance across all samples was high, reinforcing its significance in the microbial community.

In addition to the dominant genera, another notable species identified as important across all methods was *Dermacoccus nishinomiyaensis*. The prevalence of this species in sourdough samples was low, at only 1.74%, and its percent was minimal, accounting for just 0.0006%. However, random forest analysis prioritizes the contribution of species to physicochemical indicators rather than merely their abundance. This method, which employs multiple decision trees for regression, assesses feature importance by evaluating the influence of variables (species) on the predicted target (physicochemical indicators). Even at low abundance, a species may still receive a high importance score if it significantly impacts a specific physicochemical parameter. Consequently, *D. nishinomiyaensis was* identified as an important species under all three feature labels (sourdough physicochemical indicators). This species, originally isolated from sawdust, may have been introduced through raw materials, and suggests that it may be a low-abundance yet high-impact functional species, playing a crucial metabolic role in the sourdough fermentation process.

Identification of Important Fungal Species

This integrated analytical framework was likewise applied to the fungal community to pinpoint keystone species instrumental in predicting sourdough physicochemical properties. Compared to bacteria, a greater proportion of fungi could be classified at the species level, with most of these taxa being yeasts. A comparison of Table. 11 and Table. 15 (see Figs. S4 and S5) shows that both the SUM ranking and the random forest analysis using moisture as the feature label consistently identified *Kodamaea ohmeri*、*Schizothecium*、*Curvularia hominis*、*Lecanicillium antillanum*、*Rhizopus microspores* and *Buckleyzyma phyllomatis* as the top 15 important fungal.

Additionally, comparing Table. 12 and Table. 15, both SUM sorting and random

forest with pH as feature label identified *Penicillium citrinum*, *B. phyllomatis*, and *C. hominis*, and they were all important fungi ranked in the top 15. Moreover, *B. phyllomatis*, *P. citrinum*, *Candida argentea*, *L. antillanum*, and *C. hominis* were identified as important fungi by SUM sorting and random forest identification method with TTA as feature label. In addition, *B. phyllomatis* and *C. hominis* were consistently identified as top 15 keystone species across all analytical methods. Both are plant-derived fungi that likely entered the sourdough microbial community through raw materials, though their potential functional roles remain unclear.

Discussion

Sourdough has been used as a natural starter culture for leavening bread across the world for thousands of years. The microbiota of sourdough plays a critical role in determining the physicochemical and aromatic qualities of the resulting products (Clark et al., 2024; Ripari et al., 2016). Therefore, understanding how microorganisms respond to fermentation conditions is key to predicting their effects on product functionality and quality. The hydration state of the starter – whether wet (high hydration) or dry (low hydration) – also has a notable impact on the quality of the final product, influencing characteristics such as flavour, texture, and shelf life (Albagli et al., 2021). Although numerous studies have characterized the microbial composition of sourdough (GäNzle et al., 2016; Landis et al., 2021; De Vuyst et al., 2014), the effect of hydration on microbial diversity has received comparatively less attention. Specifically, the distinctions in microbial communities between wet and dry sourdough cultures remain underexplored. This study revealed significant differences in microbial diversity between dry and wet sourdough samples and, using machine learning approaches, identified correlations between microbial diversity and the chemical characteristics of sourdough, including hydration level, pH and TTA.

Bacterial diversity and evenness within individual samples were higher in dry sourdoughs, indicating more balanced microbial communities. In contrast, bacterial diversity in wet samples was relatively low, with communities dominated by fewer taxa.

Conversely, fungal species richness and diversity were significantly greater in wet samples than in dry ones. The even distribution of bacteria in dry sourdoughs may contribute to more stable fermentation and enhanced shelf life, as a broad range of species can buffer against rapid spoilage. Albagli et al. (2021) also reported that dried sourdough starters improve bread shelf life (Clark et al., 2024). Gobbetti et al. showed that freeze-dried sourdough starters preserve over 90% of antimicrobial compounds, such as phenyllactic acid, supporting LAB survival and extending bread shelf life (Gobbetti et al., 2016). In contrast, the higher fungal diversity observed in wet sourdoughs could be linked with enhanced flavour complexity, potentially driven by increased production of volatile aroma compounds and enzymatic activities (Birch et al., 2013). In dry sourdoughs, the abundance of *P. pentosaceus*, *L. brevis*, *A. tropicalis*, and *Lp. plantarum* suggests a selective advantage for these species under low-moisture conditions (De Vuyst et al., 2014; Liu et al., 2016; Marco Gobbetti et al., 2016; Zhang et al., 2019). *Lp. plantarum* strain from sourdough have been reported to show high tolerance to acid, osmotic and oxidative stresses (Parente et al., 2010; Zotta et al., 2009). We attribute the contrasting diversity trends between bacteria and fungi in dry versus wet sourdough to distinct microbial ecological strategies in response to moisture. The low-moisture conditions in dry samples likely acted as an environmental filter, selecting for desiccation-tolerant bacterial taxa such as *P. pentosaceus* and *L. brevis*, which together formed a community with higher diversity and evenness. Conversely, higher moisture in wet samples likely expanded ecological niches for fungi, promoting growth and coexistence of a wider range of taxa, including *M. humilis*, through enhanced nutrient availability and niche specialization. This divergence underscores that hydration acts as an environmental filter for bacteria but a resource driver for fungi, explaining their apparent opposite responses.

In contrast, *F. sanfranciscensis* dominated the wet sourdough samples, comprising 51.27% of the relative abundance with a prevalence of 78.65%, indicating strong adaptation to high-moisture environments. This pattern suggests that hydration level is a key ecological determinant in sourdough microbiomes, potentially influencing species distribution through differential tolerance to osmotic stress, acid accumulation,

or nutrient diffusion dynamics (Gänzle & Zheng, 2019; De Vuyst et al., 2014; De Vuyst et al., 2017). Fungal community composition also reflected hydration-driven selection pressures. *S. cerevisiae* remained the dominant species across both groups, *M. humilis* was more prevalent in wet sourdoughs, and *S. fibuligera* was enriched in some dry samples, potentially indicating differing metabolic strategies or moisture preferences between these yeasts (Santos et al., 2017). These species may contribute distinct enzymatic activities that affect flavour development and dough structure under variable hydration conditions.

Venn diagram analysis further supported the distinction in community structures. This suggests that while both environments share foundational microbial members, the wet matrix supports a broader ecological niche space, possibly due to increased nutrient solubilisation and microbial mobility (Gänzle et al., 2019; Landis et al., 2019; De Vuyst et al., 2014; De Vuyst et al., 2017). These insights are particularly relevant for industrial fermentations where hydration can be strategically manipulated to modulate microbial function and, ultimately, product characteristics such as flavour, texture, and shelf life. The effect of moisture levels on microbial dynamics has been studied on soil microbiome (Brangarí et al., 2021; Butcher et al., 2020; Evans et al., 2022; Zhang et al., 2024) and gut microbiota (Moghaddam et al., 2024).

Our analysis revealed distinct clustering patterns between dry and wet sourdough samples in hierarchical clustering and PCoA based on weighted UniFrac distances. However, the substantial overlap in bacterial communities indicates a stable core microbiota shared between both sample types. The observed differences primarily stemmed from variations in species richness and community evenness rather than complete taxonomic replacement. These results suggest that environmental factors shape community structure mainly by modulating relative abundances of existing taxa, not through wholesale microbial turnover.

High-dimensional data exploration using random forest, UMAP, and abundance-ranking techniques identified key functional taxa associated with sourdough traits. Notably, specific bacterial operational taxonomic units (e.g., OTU668 An unculturable species of the genus *Malikia* and OTU143 *Acetobacter sp.*) showed strong predictive

importance for humidity and pH, despite taxonomic ambiguity due to limitations in 16S rDNA resolution. Similarly, current research on *D. nishinomiyaensis* is limited. Although C. Joron et al. identified it as an etiological agent of persistent paediatric catheter-related bacteraemia, its functional role within the sourdough microbiota remains unexplored (Joron et al., 2019). This microorganism may represent a low-abundance but potentially high-impact fungal species, highlighting the ecological importance of rare taxa in fermentation ecosystems (Liang et al., 2020). Compared with bacteria, fungal species were more reliably classified at the species level using ITS sequencing. Core taxa such as *K. ohmeri*, *C. hominis*, *L. antillanum*, and *B. phyllomatis* were consistently identified by both machine learning and ordination analyses. Although recognized as important taxa by these methods, such fungi have been largely overlooked in previous studies of sourdough microbial diversity, and their functional roles remain poorly understood. Notably, *K. ohmeri* and *L. antillanum* may present potential pathogenic risks: *K. ohmeri* is an emerging human yeast pathogen (Zhou et al., 2021), while *L. antillanum* exhibits entomopathogenic activity (Zhou et al., 2018). This is the first study to provide proof of concept for how machine learning can be employed in the selection of starter cultures based on their association with physicochemical characteristics. This technique has been proposed as a potential method to develop microbiome therapeutics (Abavisani et al., 2024; Mccoubrey et al., 2021).

In this study, high-throughput sequencing combined with ensemble learning revealed the composition of dominant and core microorganisms in traditional sourdough, as well as their responses to physicochemical parameters such as moisture, pH, and TTA. These insights provide important guidance for the rational design of novel starter cultures. Core taxa such as *F. sanfranciscensis* and *S. cerevisiae* can serve as foundational strains to ensure fermentation stability, while low-abundance but functionally relevant species identified through machine learning may contribute to flavor complexity and system robustness. Therefore, this study offers both ecological and empirical evidence for developing stable and customizable starter cultures to improve sourdough fermentation.

The limited taxonomic resolution for bacterial communities reflects the inherent constraints of 16S rDNA sequencing and impedes comprehensive interpretation of bacterial contributions to fermentation dynamics. To overcome this, future studies should adopt full-length 16S rRNA gene sequencing, which enables higher resolution by capturing variation across all hypervariable regions. Moreover, accounting for intragenomic heterogeneity – the presence of multiple, slightly different copies of the 16S gene within a single bacterial genome – can further refine taxonomic resolution and strain-level discrimination (Johnson et al., 2019). The presence of unclassified or potentially pathogenic bacteria highlights a critical gap in microbial safety assessment. Furthermore, our current understanding of microbial interactions and the biochemical pathways underlying flavour development remains incomplete.

Future work employing metagenomic or metatranscriptomic approaches may provide the necessary resolution to resolve species-level identities and functional potentials of sourdough-associated microbes. Integrating these datasets with systems biology and AI-driven modeling could enable the design of synthetic microbial consortia optimized for functional stability, improved fermentation performance, and enhanced food safety.

Declarations

Ethics approval and consent to participate

Non-applicable.

Consent for publication

Non-applicable.

Availability of data and material

The dataset supporting the conclusions of this article is available in the Sequence Read Archive repository under accession PRJNA1291283.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Number 32172179).

Authors' contributions

Conceptualization, G.Z.; methodology, Y.L., F.L., H.C. and Z.Z.; validation, Y.L., F.L. and G.Z. ; formal analysis, J.W. and Y.L.; data curation, Y.L., J.W., A.L. and T.L.; writing—original draft preparation, Y.L., F.S., Y.Y. and S.T.; writing—review and editing, Y.L., F.S. and G.Z.; supervision, G.Z., S.Z. and M.Z.; investigation, S.Z.; project administration, G.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We thank Kemeng Chai and Jianpeng Zhang for their assistance in determining the physicochemical properties of sourdough samples, Xi Chen for providing valuable insights into data visualization, as well as Lingjun Ma, Juan Hao, Haowei Xu, Dongge Zhao, Xiao Qu, and 56 other fellow students for their help in sample collection.

References

Abavisani, M., Khoshrou, A., Foroushan, S. K., Ebadpour, N., & Sahebkar, A. (2024). Deciphering the gut microbiome: The revolution of artificial intelligence in microbiota analysis and intervention. *Current Research in Biotechnology, 7,* 100211. https://doi.org/10.1016/j.crbiot.2024.100211

Albagli, G., Schwartz, I. D. M., Amaral, P., Ferreira, T., & Finotelli, P. (2021). How dried sourdough starter can enable and spread the use of Sourdough bread. *Lwt - Food Science and Technology, 149*, 111888. https://doi.org/10.1016/j.lwt.2021.111888

Alfonzo, A., Ventimiglia, G., Corona, O., Gerlando, R. D., Gaglio, R., Francesca, N., Moschetti, G.,& Settanni, L. (2013). Diversity and technological potential of lactic acid bacteria of wheat flours. *Food Microbiology, 36*(2), 343-354. https://doi.org/10.1016/j.fm.2013.07.003

Birch, A. N., Petersen, M. A., Arneborg, N., & Hansen, A. S. (2013). Influence of commercial baker's yeasts on bread aroma profiles. *Food Research International, 52*(1), 160-166.

https://doi.org/10.1016/j.foodres.2013.03.011

Boiocchi, F., Porcellato, D., Limonta, L., Picozzi, C., Vigentini, I., Locatelli, D. P., & Foschino, R. (2017). Insect frass in stored cereal products as a potential source of *Lactobacillus sanfranciscensis* for sourdough ecosystem. *Journal of applied microbiology, 123*(4)*,* 944–955. https://doi.org/10.1111/jam.13546

Brangarí, A. C., Manzoni, S., & Rousk, J. (2021). The mechanisms underpinning microbial resilience to drying and rewetting – A model analysis. *Soil Biology and Biochemistry, 162*, 108400. https://doi.org/10.1016/j.soilbio.2021.108400

Butcher, K. R., Nasto, M. K., Norton, J. M., & Stark, J. M. (2020). Physical mechanisms for soil moisture effects on microbial carbon-use efficiency in a sandy loam soil in the western United States. *Soil Biology & Biochemistry, 150,* 107969. https://doi.org/10.1016/j.soilbio.2020.107969

Calabrese, F. M., Ameur, H., Nikoloudaki, O., Celano, G., Vacca, M., Junior, W. J., Manzari, C., Vertè, F., Di Cagno, R., Pesole, G., De Angelis, M., & Gobbetti, M. (2022). Metabolic framework of spontaneous and synthetic sourdough metacommunities to reveal microbial players responsible for resilience and performance. *Microbiome, 10*(1), 148. https://doi.org/10.1186/s40168-022-01301-3

Chavan, R. S., & Chavan, S. R. (2011). Sourdough Technology—A Traditional Way for Wholesome Foods: A Review. *Comprehensive Reviews in Food Science & Food Safety, 10*(3), 169-182. https://doi.org/10.1111/j.1541-4337.2011.00148.x

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics, 34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Clark, C. S., Ohstrom, A., Rolon, M. L., Smith, M., Wolfe, B. E., Wee, J., & Buiten, C. B. V. Sourdough starter culture microbiomes influence physical and chemical properties of wheat bread. *Journal of Food Science, 89*(3), 1414-1427. https://doi.org/10.1111/1750-3841.16957

Clarke, C. I., & Arendt, E. K. (2005). A Review of the Application of Sourdough Technology to Wheat Breads. *Advances in Food & Nutrition Research, 49*, 137-161. https://doi.org/10.1016/s1043-4526(05)49004-x

Coda, R., Cagno, R. D., Gobbetti, M., & Rizzello, C. G. (2014). Sourdough lactic acid bacteria: Exploration of non-wheat cereal-based fermentation. *Food Microbiology, 37*(2), 51-58. https://doi.org/10.1016/j.fm.2013.06.018

Corsetti, A., & Settanni, L. (2007). Lactobacilli in sourdough fermentation. *Food Research International, 40*(5), 539-558. https://doi.org/10.1016/j.foodres.2006.11.001

De Vuyst, L., Comasio, A., & Kerrebroeck, S. V. (2021). Sourdough production: fermentation strategies, microbial ecology, and use of non-flour ingredients. *Critical Reviews in Food Science and Nutrition*, 63(15), 2447–2479.https://doi.org/10.1080/10408398.2021.1976100

De Vuyst, L., Harth, H., Van Kerrebroeck, S., & Leroy, F. (2016). Yeast diversity of sourdoughs and associated metabolic properties and functionalities. *International Journal of Food Microbiology*, *239,* 26-34. https://doi.org/10.1016/j.ijfoodmicro.2016.07.018

DongMin, S. (2009). Probe into the origin of the steamed bun and its historical development. *Journal of Henan University of Technology(Social Science Edition), 2*(5),1673-1751. https://doi.org/10.16433/j.cnki.cn41-1379.2009.02.001

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods, 10*, 996-998. https://doi.org/10.1038/nmeth.2604

Evans, S. E., Allison, S. D., & Hawkes, C. V. (2022). Microbes, memory and moisture: Predicting microbial moisture responses and their impact on carbon cycling. *Functional Ecology*. *36*(6), 1430-1441. https://doi.org/10.1111/1365-2435.14034

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research, 15(1)*, 3133-3181. https://dl.acm.org/doi/epdf/10.5555/2627435.2697065

Fu, L., Nowak, A., Zhao, H., & Zhang, B. (2022). Relationship between Microbial Composition of Sourdough and Texture, Volatile Compounds of Chinese Steamed Bread. *Foods, 11*(13), 1908. https://doi.org/10.3390/foods11131908

GäNzle, M., & Ripari, V. (2016). Composition and function of sourdough microbiota: From ecological theory to bread quality. *International Journal of Food Microbiology*, *239,* 19-25. https://doi.org/10.1016/j.ijfoodmicro.2016.05.004

Gaenzle, M. G. (2014). Enzymatic and bacterial conversions during sourdough fermentation. *Food Microbiology, 37*, 2-10. https://doi.org/10.1016/j.fm.2013.04.007

Gino, V., Luc, D. V., Roel, V. D. M., Geert, H., Peter, V., & Heide-Marie, D. (2010). Yeast species composition differs between artisan bakery and spontaneous laboratory sourdoughs. *Fems Yeast Research, 10*(4), 471-481. https://doi.org/10.1111/j.1567-1364.2010.00621.x

Gnäzle, M., & Gobbetti, M. (2013). Handbook on Sourdough Biotechnology *(2013 ed.).* Springer, (Chapter 7 and 8).

Gänzle, M. G., & Zheng, J. (2019). Lifestyles of sourdough lactobacilli - Do they matter for microbial ecology and bread quality? *International Journal of Food Microbiology, 302,* 15-23. https://doi.org/10.1016/j.ijfoodmicro.2018.08.019

Gobbetti, M. (1998). The sourdough microflora: Interactions of lactic acid bacteria and yeasts. *Trends in Food Science & Technology, 9*(7), 267-274. https://doi.org/10.1016/S0924-2244(98)00053-3

Gobbetti, M., Minervini, F., Pontonio, E., Di Cagno, R., & De Angelis, M. (2016). Drivers for the establishment and composition of the sourdough lactic acid bacteria biota. *International Journal of Food Microbiology*, *239*, 3-18. https://doi.org/10.1016/j.ijfoodmicro.2016.05.022

Hammes, W. P., Brandt, M. J., Francis, K. L., Rosenheim, J., Seitter, M. F. H., & Vogelmann, S. A. (2005). Microbial ecology of cereal fermentations. *Trends in Food Science & Technology, 16*(1-3), 4-11. https://doi.org/10.1016/j.tifs.2004.02.010

Han, C., Shi, C., Liu, L., Han, J., Yang, Q., Wang, Y., Li, X., Fu, W., Gao, H., Huang, H., Zhang, X., Yu, K.(2024). Majorbio Cloud 2024: Update single-cell and multiomics workflows. *iMeta, 3*(4), e217. https://doi.org/10.1002/imt2.217

Han, J., & Kamber, M. (2011). *Data Mining Concept and Techniques*: Data Mining Concept and Techniques*(2th ed.).* Burlington (Chapter 2).

Han, J., Wang, Y., & Liu, C. (2016). Research on Different Fermentation Methods and Characteristics of Steamed Buns. *Grain Science and Technology and Economy, 41*(2), 60-62. https://doi.org/10.16465/j.gste.cn431252ts.20160218

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction *(2th ed.).* Springer (Chapter 10).

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstork, G. W. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis.

Nature Communications, 10, 5029. https://doi.org/10.1038/s41467-019-13036-1

Joron, Roméo, Flèche-Matéos, L., Rames, Samad, E., & Hamdad. (2019). *Dermacoccus nishinomiyaensis* as a cause of persistent paediatric catheter-related bacteraemia. *Clinical microbiology and infection, 25*(8), 1054-1055. https://doi.org/10.1016/j.cmi.2019.02.023

Katina, K., Arendt, E., Liukkonen, K. H., Autio, K., Flander, L., & Poutanen, K. (2005). Potential of sourdough for healthier cereal products. *Trends in Food Science & Technology, 16*(1-3), 104-112. https://doi.org/10.1016/j.tifs.2004.03.008

Landis, E. A., Oliverio, A. M., McKenney, E. A., Nichols, L. M., Kfoury, N., Biango-Daniels, M., Shell, L. K., Madden, A. A., Shapiro, L., Sakunala, S., Drake, K., Robbat, A., Booker, M., Dunn, R. R., Fierer, N., & Wolfe, B. E. (2021). The diversity and function of sourdough starter microbiomes. *elife, 10,* e61644. https://doi.org/10.7554/eLife.61644

Liu, T., Li, Y., Chen, J., Sadiq F. A., Zhang, G., Li, Y., & He, G.(2016). Prevalence and diversity of lactic acid bacteria in Chinese traditional sourdough revealed by culture dependent and pyrosequencing approaches. *Lwt Food Science & Technology*. *68,* 91-97. https://doi.org/10.1016/j.lwt.2015.12.025

Liu, T., Li, Y., Sadiq, F. A., Yang, H., & He, G. (2018). Predominant yeasts in Chinese traditional sourdough and their influence on aroma formation in Chinese steamed bread. *Food Chemistry, 242*, 404-411. https://doi.org/10.1016/j.foodchem.2017.09.081

Liang, Y., Xiao, X., Nuccio, E. E., Yuan, M., Zhang, N., Xue, K., Cohan, F. M., Zhou, J., & Sun, B. (2020). Differentiation strategies of soil rare and abundant microbial taxa in response to changing climatic regimes. *Environmental microbiology, 22*(4)*,* 1327–1340*.* https://doi.org/10.1111/1462-2920.14945

Mago, T., & Salzberg, S. L. (2011). FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies. *Bioinformatics, 27*(21), 2957-2963. https://doi.org/10.1093/bioinformatics/btr507

Ma, S., Wang, X., Qian, X., Sun, B., & Li, L. (2024). Review on flavor substances and nutritional characteristics of sourdough steamed bread. *Journal of Henan University of Technology (natural science edition), 45*(4), 134-142. https:/doi.org/10.16433/j.1673-2383.202403220001

Mccoubrey, L. E., Elbadawi, M., Orlu, M., Gaisford, S., & Basit, A. W. (2021). Harnessing machine learning for development of microbiome therapeutics. *Gut Microbes, 13*(1), 1-20. https://doi.org/10.1080/19490976.2021.1872323

Mcinnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *The Journal of Open Source Software, 3*(29), 861. https://doi.org/10.48550/arXiv.1802.03426

Medina, R. H., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., & Rasmussen, S. (2022). Machine learning and deep learning applications in microbiome research. *ISME Communications, 2,* 98. https://doi.org/10.1038/s43705-022-00182-9

Mielecki, D., Detman, A., Aleksandrzak-Piekarczyk, T., Widomska, M., Chojnacka, A., Stachurska-Skrodzka, A., Walczak, P., Grzesiuk, E., & Sikora, A. (2024).Unlocking the genome of the non-sourdough Kazachstania humilis MAW1: insights into inhibitory factors and phenotypic properties. *Microbial Cell Factories, 23*(1), 111. https://doi.org/10.1186/s12934-024-02380-7

Minervini, F., Di Cagno, R., Lattanzi, A., De Angelis, M., Antonielli, L., Cardinali, G., Cappelle, S., & Gobbetti, M. (2012). Lactic acid bacterium and yeast microbiotas of 19 sourdoughs used for traditional/typical italian breads: interactions between ingredients and microbial species

diversity. *Applied & Environmental Microbiology, 78*(4), 1251-1264. https://doi.org/10.1128/aem.07721-11

Minervini, F., Lattanzi, A., Angelis, M. D., Celano, G., & Gobbetti, M. (2015). House microbiotas as sources of lactic acid bacteria and yeasts in traditional Italian sourdoughs. *Food Microbiology, 52*, 66-76. https://doi.org/10.1016/j.fm.2015.06.009

Moghaddam, H. S., Abkar, L., & Fowler, S. J. (2024). Making waves: From tap to gut- exploring the impact of drinking water on gut microbiota. *Water Research, 267,* 122503. https://doi.org/10.1016/j.watres.2024.122503

Parente, E., Ciocia, F., Ricciardi, A., Zotta, T., Felis, G. E., & Torriani, S. (2010). Diversity of stress tolerance in *Lactobacillus plantarum*, *Lactobacillus pentosus* and *Lactobacillus paraplantarum*: A multivariate screening study. *International Journal of Food Microbiology, 144*(2), 270-279. https://doi.org/10.1016/j.ijfoodmicro.2010.10.005

Ripari, V., Cecchi, T., & Berardi, E. (2016). Microbiological characterisation and volatiles profile of model, ex-novo, and traditional Italian white wheat sourdoughs. *Food Chemistry, 205*, 297-307. https://doi.org/10.1016/j.foodchem.2016.02.150

Santos, M. C., Golt, C., Joerger, R. D., Mechor, G. D., Gerson, B. M., & Kung, L. (2017). Identification of the major yeasts isolated from high moisture corn and corn silages in the United States using genetic and biochemical methods. *Journal of Dairy Science, 100*(2), 1151-1160. https://doi.org/10.3168/jds.2016-11450

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F.(2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied & Environmental Microbiology, 75*(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology, 12*(6), R60. https://doi.org/10.1186/gb-2011-12-6-r60

Siepmann, F. B., Valery, R.,Waszczynskyj, N., & Spier, M. R. (2018). Overview of Sourdough Technology: from Production to Marketing. *Food and bioprocess technology, 11*, 242-270. https://doi.org/10.1007/s11947-017-1968-2

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology , 44*(4), 846-849. https://doi.org/10.1099/00207713-44-4-846

Sun, W., Qu, L., Chao, T., & Li, L. (2016). Identification of acetic acid bacteria in fermenting agent of Chinese steamed bread and optimization of fermentation conditions. *Science and Technology of Food Industry, 22,* 190-194. https://dx.doi.org/10.13386/j.issn1002-0306.2016.22.029

Van Kerrebroeck, S., Maes, D., & De Vuyst, L. (2017). Sourdoughs as a function of their species diversity and process conditions, a meta-analysis. *Trends in Food Science & Technology*, 68, 152-159. https://doi.org/10.1016/j.tifs.2017.08.016

Vuyst, L. D., Kerrebroeck, S. V., Harth, H., Huys, G., Daniel, H. M., & Weckx, S. (2014). Microbial ecology of sourdough fermentations: Diverse or uniform? *Food Microbiology, 37*, 11-29. https://doi.org/10.1016/j.fm.2013.06.002

Vuyst, L. D., Kerrebroeck, S. V., & Leroy, F. (2017). Microbial Ecology and Process Technology of Sourdough Fermentation. *Advances in applied microbiology, 100*, 49-160. https://doi.org/10.1016/bs.aambs.2017.02.003

Wu, C.; Liu, R.; Huang, W.; Duarte, P. R., Wang, F., & Yao, Y. (2012). Effect of sourdough fermentation on the quality of Chinese Northern-style steamed breads. *Journal of Cereal Science, 56(2),* 127–133. https://doi.org/10.1016/j.jcs.2012.03.007.

Wang, Q., Garrity, G., Tiedje, J., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*. *73(16),* 5261–5267. https://doi.org/10.1128/aem.00062-07

Wang, X., Zhu, X., Bi, Y., Zhao, R., & Yuan, W. (2020). Dynamics of microbial community and changes of metabolites during production of type I sourdough steamed bread made by retarded sponge-dough method. *Food Chemistry, 330*(5), 127316. https://doi.org/10.1016/j.foodchem.2020.127316

Wilhelm, R. C., Van, E. H. M., & Buckley, D. H. (2022). Predicting measures of soil health using the microbiome and supervised machine learning. *Soil Biology & Biochemistry, 164*, 108472. https://doi.org/10.1016/j.soilbio.2021.108472

Xing, X., Ma, J., Fu, Z., Zhao, Y., & Suo, B. (2020). Diversity of bacterial communities in traditional sourdough derived from three terrain conditions (mountain, plain and basin) in Henan Province, China. *Food Research International*, *133*, 109139. https://doi.org/10.1016/j.foodres.2020.109139

Yan, B., Sadiq, F. A., Cai, Y., Fan, D., Chen, W., Zhang, H., & Zhao, J. (2019). Microbial diversity in traditional type I sourdough and jiaozi and its influence on volatiles in Chinese steamed bread. *LWT-Food Science & Technology, 101*, 764-773. https://doi.org/10.1016/j.lwt.2018.12.004

Zhou, Y. M., Zhi, J. R., Ye, M., Zhang, Z. Y., Yue, W. B., & Zou, X. (2018). *Lecanicilliumcauligalbarum* sp. nov. (Cordycipitaceae, Hypocreales), a novel fungus isolated from a stemborer in the Yao Ren National Forest Mountain Park, Guizhou. *Mycokeys, 43*, 59-74. https://doi.org/10.3897/mycokeys.43.30203

Zhang, E., Wong, S. Y., Czechowski, P., Terauds, A., Ray, A. E., Benaud, N., Chelliah, D. S., Wilkins, D., Montgomery, K., & Ferrari, B. C. (2024). Effects of increasing soil moisture on Antarctic desert microbial ecosystems. *Conservation biology : the journal of the Society for Conservation Biology*, *38*(4), e14268. https://doi.org/10.1111/cobi.14268

Zhang, G., Sadiq, F. A., Zhu, L., Liu, T., Yang, H., Wang, X., & He, G. (2015). Investigation of Microbial Communities of Chinese Sourdoughs Using Culture-Dependent and DGGE Approaches. *Journal of Food Science, 80*(11), M2535-M2542. https://doi.org/10.1111/1750-3841.13093

Zhang, G., Tu, J., Sadiq, F. A., Zhang, W., & Wang, W. (2019). Prevalence, Genetic Diversity, and Technological Functions of the Lactobacillus sanfranciscensis in Sourdough: A Review. *Comprehensive Reviews in Food Science and Food Safety, 18*(4), 1209–1226. https://doi.org/10.1111/1541-4337.12459

Zhang, G., Zhang, W., Sadiq, F. A., Sakandar, H. A., & Guoqing, H. (2019). Microbiota succession and metabolite changes during the traditional sourdough fermentation of Chinese steamed bread. *Cyta Journal of Food, 17*(1), 172-179. https://dx.doi.org/10.1080/19476337.2019.1569166

Zhou, M., Li, Y., Kudinha, T., Xu, Y., & Liu, Z. *Kodamaea ohmeri* as an Emerging Human Pathogen: A Review and Update. *Frontiers in microbiology, 12*, 736582.

https://doi.org/10.3389/fmicb.2021.736582
Zotta, T., Parente, E., & Ricciardi, A. (2009). Viability staining and detection of metabolic activity of sourdough lactic acid bacteria under stress conditions. *World Journal of Microbiology & Biotechnology, 25*(6), 1119-1124. https://doi.org/10.1007/s11274-009-9972-y

Figure legend

Fig.1 Schematic diagram of ensemble learning working
The dataset is partitioned into training and test sets. For each base model, random subsampling of both data instances and features (physicochemical indicators) from the training set generates a sub-training set. After training and validation against the test set, base models make individual predictions. Final predictions are aggregated from all base models to yield highly robust integrated results.

Fig. 2 Species composition of sourdough
A Bacterial composition in dry sourdough samples was dominated by *P. pentosaceus* and *Lactobacillus brevis*; **B** Fungal composition featured *S. cerevisiae* and *W. anomalus* as predominant taxa; **C** In wet samples, *F. sanfranciscensis* and *W. cibaria* prevailed over *P. pentosaceus* and *Lp. plantarum* in bacterial composition; **D** Fungal communities remained dominated by *S. cerevisiae*, with *M. humilis* emerging as another major fungal component in wet samples.

Fig. 3 Venn diagram of sourdough sample species
A Venn diagram of bacterial composition in dry (D) and wet (S) sample groups, with numbers indicating bacterial species counts; **B** Proportional distribution of 91 shared bacterial species. *F. sanfranciscensis* (40.23%) and *P. pentosaceus* (14.10%) were the most prevalent shared species; **C** *A. soil* (19.44%) represented the most abundant unique taxon in dry samples; **D** *Lactococcus garvieae* (29.63%) dominated unique species in wet samples; **E** Venn diagram of fungal composition in dry and wet groups; **F** *C. macerans* (12.66%) constituted the largest proportion among shared fungal species; **G** *R. australiensis* (66.96%) was the predominant unique species in wet samples; **H** *S. cerevisiae* (59.49%) accounted for the highest proportion of unique species in dry samples.

Fig. 4 Analysis of species differences in sourdough
A Significance analysis of intergroup differences at species level. *F. sanfranciscensis*, *P. pentosaceus* and *Lp. Plantarum* are significantly different in wet and dry samples; **B** *M. humilis, A. cibarius, E. nigrum, and C. delicatulum* are highly significant differences between the dry and wet sample groups; **C** community heatmap analysis on species level. The difference of species abundance between groups was demonstrated by the color change of color blocks; the abundances of *F. sanfranciscensis*, *P. pentosaceus* and *Lp. Plantarum* were significantly different between the two groups; **D** Both dry and wet

sample groups contained *S. cerevisiae* and *M. humilis* at high abundances with significantly difference.

Fig. 5 Correlation analysis of sourdough microbial communities

A Canonical correspondence analysis (CCA) of sourdough samples against physicochemical parameters. Red points: dry samples; blue points: wet samples. Arrows represent environmental factors, where arrow length indicates the magnitude of explanatory power (i.e., influence on species distribution). Angles between arrows denote correlations (acute: positive; obtuse: negative; right angle: uncorrelated). The distance from the origin to the projection point of a sample onto a quantitative environmental factor arrow reflects the relative influence of that factor on community composition. **B** Red points: wet samples; blue points: dry samples. **C&D** X- and Y-axes represent environmental factors and species, respectively. R-values are color-coded; *P ≤ 0.05, **P ≤ 0.01, ***P ≤ 0.001.

Fig. 6 Identification of top 50 bacteria using physical and chemical indicators

A Top 50 important bacteria identified using moisture as the feature. Left and middle panels show species identified by abundance ranking and random forest methods respectively, with fixed spatial positions; Right panel displays value distributions of both methods. **B** Top 50 important bacteria using pH as feature. **C** Top 50 important bacteria using TTA as feature

Fig. 7 Identification of top 15 bacteria using physical and chemical indicators

A Top 15 important bacteria identified using moisture as the feature. Left and middle panels display species identified through abundance ranking and random forest methods respectively, with fixed spatial positions; Right panel shows value distributions of both methods. **B** Top 15 important bacteria using pH as feature. **C** Top 15 important bacteria using TTA as feature.

Supplementary Information for "Deciphering the Microbial Complexity of Chinese Traditional Sourdough through Integrated Learning"

Deciphering the Microbial Complexity of Chinese Traditional Sourdough through Integrated Learning

Yan Liu[1], Jinyu Wang[1], Shiwei Zhou[2], Feijiang Li[3] , Honghong Cheng[4], Zheqing Zhu[2], Jiale Wang[1], Aowen Liu[1], Ting Lu[1], Yujuan Yu[1], Senmiao Tian[1], Min Zhang[5], Faizan Ahmed Sadiq(Co-Corresponding Author)[6*], and Guohua Zhang(Co-Corresponding Author)[1*]

1 School of Life Science, Shanxi University, Taiyuan 030006, China.

2 Lesaffre Management (Shanghai) Co., Ltd, Shanghai 200030, China.

3 Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China.

4 School of Information, Shanxi University of Finance and Economics, Taiyuan 030006, China.

5 Research Institute of Applied Biology, Shanxi University, Taiyuan 030006, China.

6 Advanced Therapies Group, School of Dentistry, Cardiff University, Cardiff CF14 4XY, United Kingdom.

* Correspondence: Advanced Therapies Group, School of Dentistry, Cardiff University, Cardiff CF14 4XY, United Kingdom SadiqF@catdiff.ac.uk(F.S.); School of Life Science, Shanxi University, Taiyuan 030006，China; zhanggh@sxu.edu.cn(G.Z.); Tel.: +86-15513091052.

Faizan Ahmed Sadiq and Guohua Zhang, these authors contributed equally as co-corresponding authors

Fig. S1:

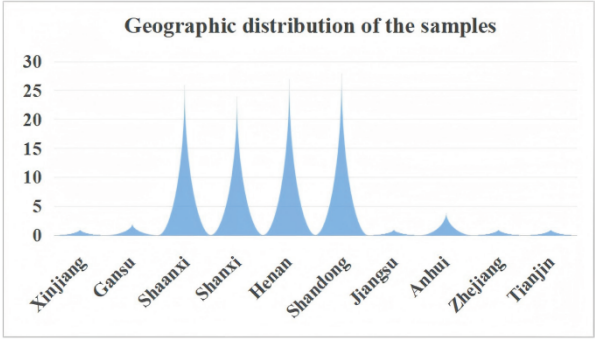

Fig. S1 Geographical distribution of sourdough samples

Numbers indicate sample counts collected from each province in China: Xinjiang Uygur Autonomous Region (1), Gansu (2), Shaanxi (26), Shanxi (24), Henan (27), Shandong (28), Anhui (4), Jiangsu (1), Zhejiang (1).

Fig. S2:

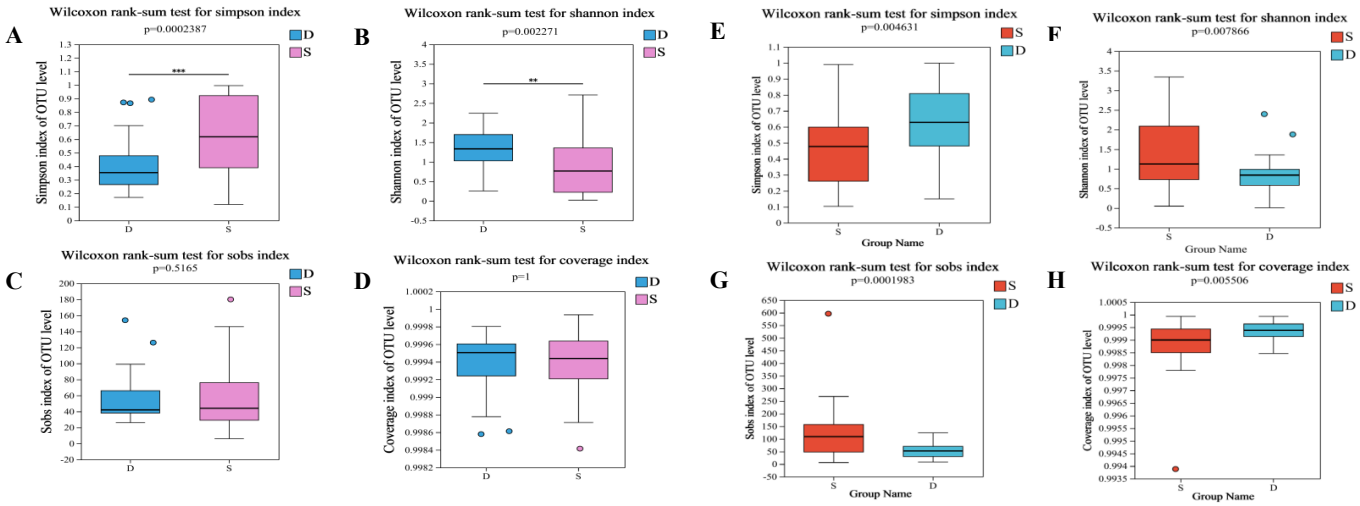

Fig. S2 Alpha diversity analysis of sourdough microbial community

A Simpson's diversity index (bacteria); **B** Shannon-Wiener index (bacteria); **C** Observed species richness (bacteria); **D** Good's coverage estimator (bacteria); **E** Simpson's diversity index (fungi); **F** Shannon-Wiener index (fungi); **G** Observed species richness (fungi); **H** Good's coverage estimator (fungi). Boxes represent interquartile ranges (IQR), horizontal lines indicate medians, whiskers show 1.5×IQR. Asterisks denote significance levels ($p < 0.05$; *$p < 0.01$; **$p < 0.001$; ns: not significant) determined by Mann-Whitney U tests.

Fig. S3:

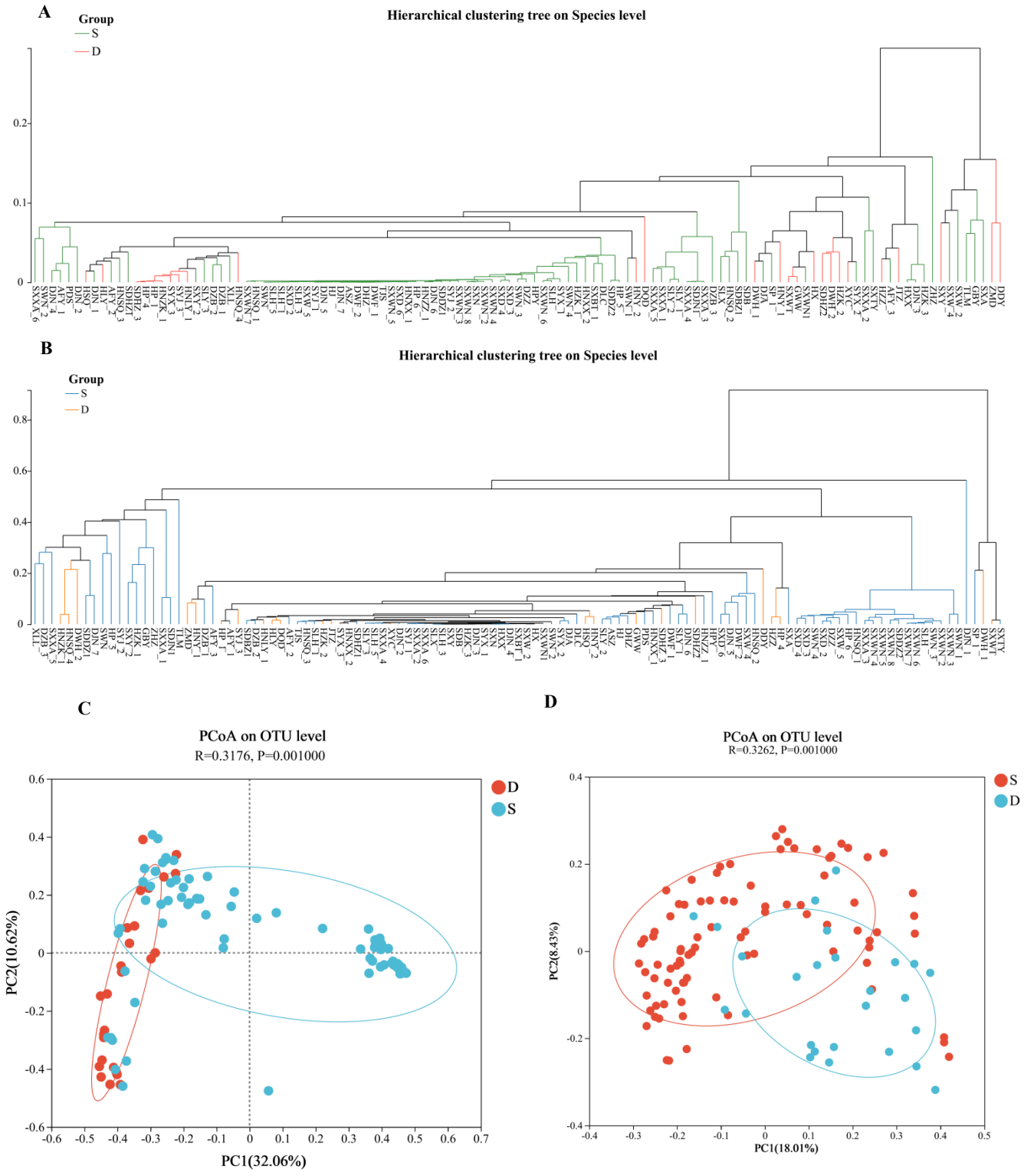

Fig. S3 Cluster analysis of sourdough samples

A Bacterial community dissimilarity (wet samples: S; dry samples: D); **B** Fungal community dissimilarity; **C** Inter-group bacterial divergence quantified by *weighted UniFrac* distances; **D** Inter-group fungal divergence.

Axes: PCo1 and PCo2 represent primary coordinates, with percentage values indicating

the proportion of variance explained. Scale ticks denote relative distance units. Points: Spatial proximity reflects similarity in species composition (e.g., clustered dry samples in Panel A indicate homogeneous bacterial communities). Statistical significance of group separation was confirmed by PERMANOVA ($p < 0.001$).

Fig. S4

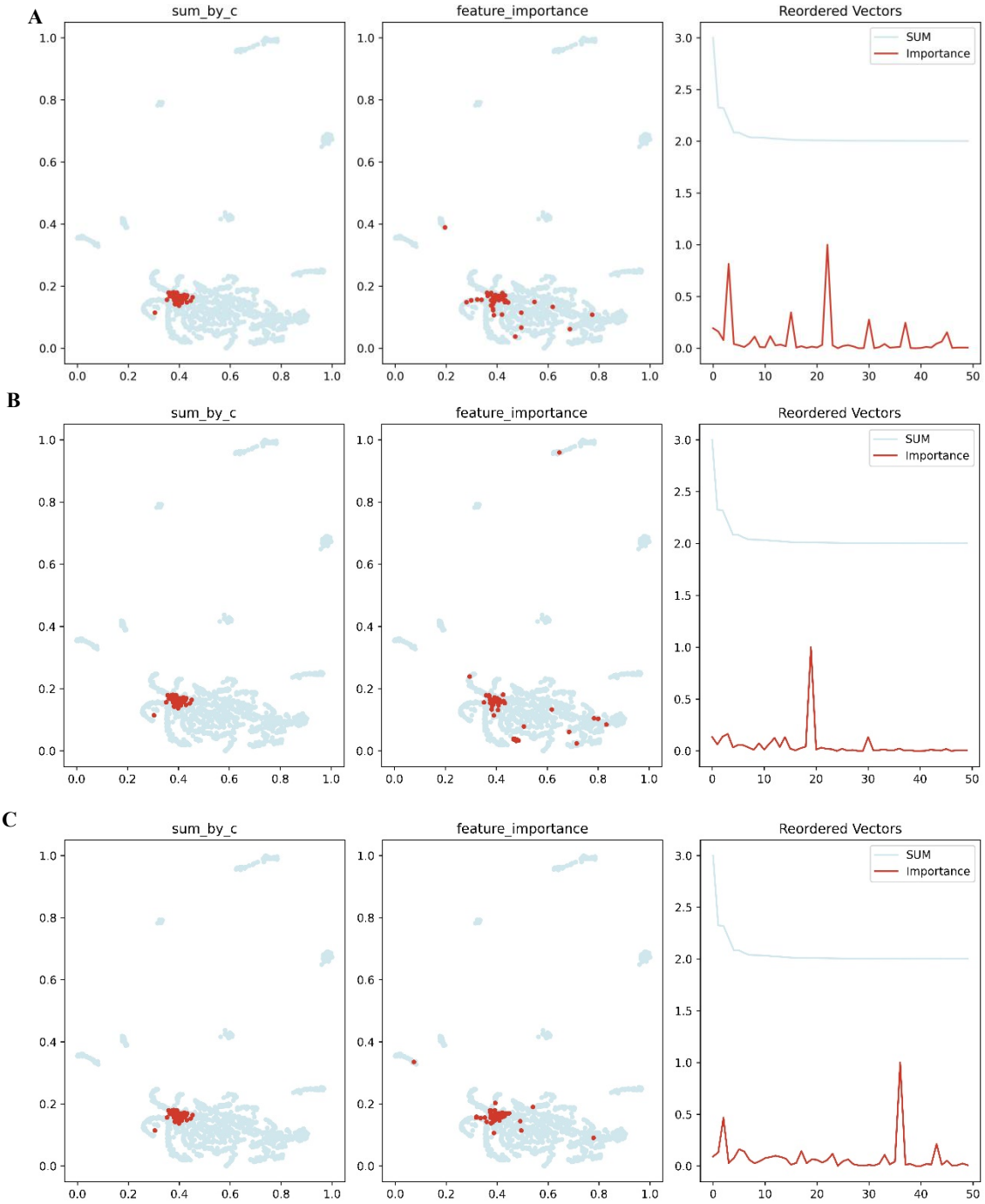

Fig. S4 Identification of top 50 fungi using physical and chemical indicators

A Top 50 important Fungi identified using moisture as the feature. Left and middle panels show species identified by abundance ranking and random forest methods respectively, with fixed spatial positions; Right panel displays value distributions of both methods. **B** Top 50 important Fungi using pH as feature. **C** Top 50 important Fungi using TTA as feature.

Fig. S5:

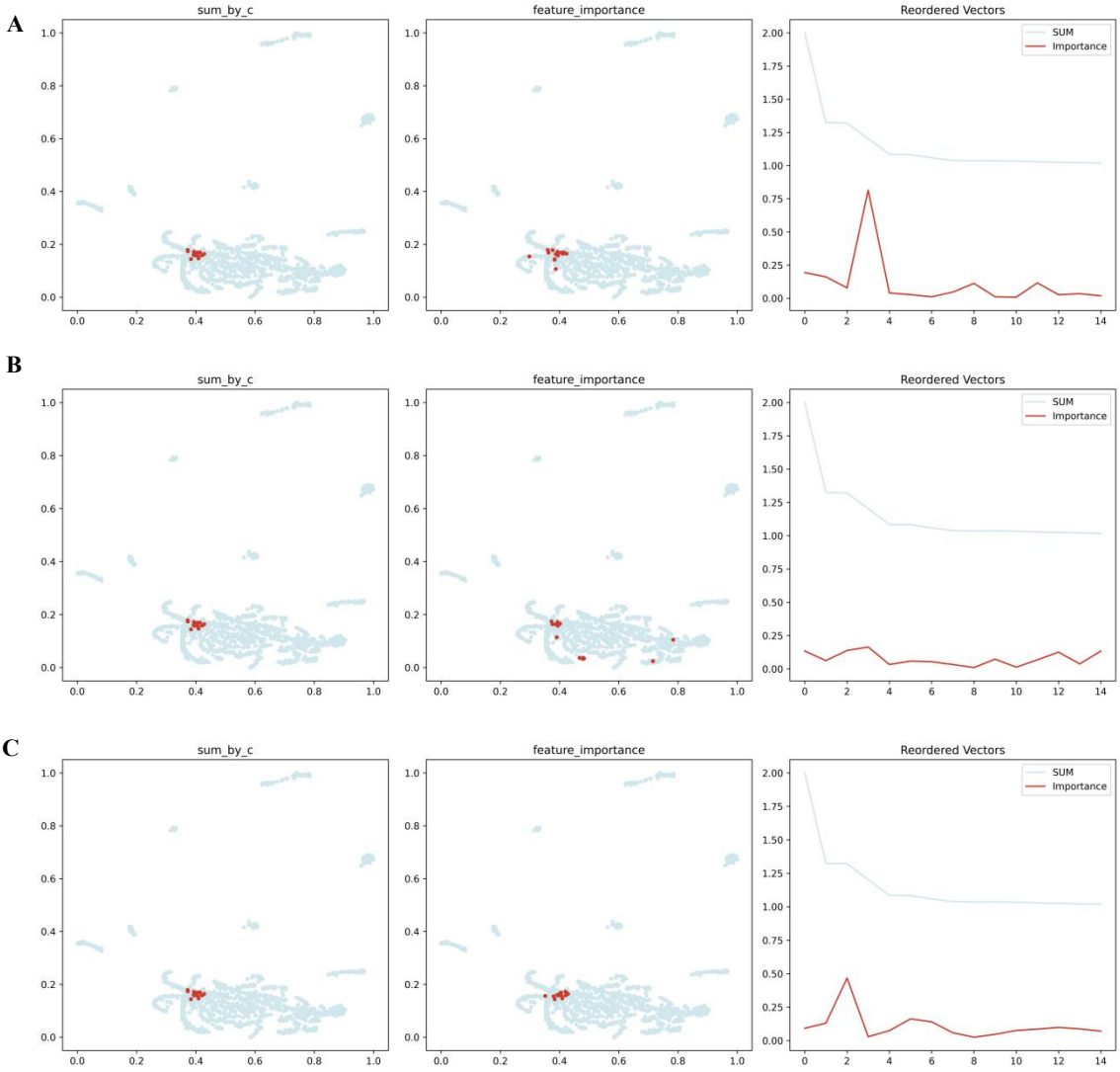

Fig. S5 12 Identification of top 15 fungi using physical and chemical indicators

A Top 15 important Fungi identified using moisture as the feature. Left and middle panels show species identified by abundance ranking and random forest methods respectively, with fixed spatial positions; Right panel displays value distributions of both methods. **B** Top 15 important Fungi using pH as feature. **C** Top 15 important Fungi using TTA as feature

Table. 1 Sample information list

Serial number	Sample name	Sample source	Sample state
1	HNY-1	Nanyang (Henan)	solid
2	HNY-2	Nanyang (Henan)	solid
3	HZK-2	Zhoukou (Henan)	solid
4	HLY	Luoyang (Henan)	solid
5	HSQ	Shangqiu (Henan)	solid
6	ZMD	Zhuamadian (Henan)	solid
7	DWH-2	Weihai (Shandong)	solid
8	DWH-1	Weihai (Shandong)	solid
9	DDY	Dongying (Shandong)	solid
10	HZZ	Zhengzhou (Henan)	solid
11	GWW	Wuwei (Gansu)	solid
12	DQD	Qingdao (Shandong)	solid
13	JTZ	Taizhou (Jiangsu)	solid
14	XYC	Yuncheng (Shanxi)	solid
15	HNLY-1	Luoyang (Henan)	solid
16	HNSQ-4	Shangqiu (Henan)	solid
17	HNZK-1	Zhoukou (Henan)	solid
18	HP-1	Puyang (Henan)	solid
19	HP-4	Puyang (Henan)	solid
20	HX	Xinxiang (Henan)	solid
21	SDHZ-2	Heze (Shandong)	solid
22	SDHZ-3	Heze (Shandong)	solid
23	SXWT1	Xinzhou (Shanxi)	solid
24	SYJ-3	Yuncheng (Shanxi)	solid
25	SYX-3	Yuncheng (Shanxi)	solid
26	SXY	Yulin (Shaanxi)	solid
27	DJN-1	Jining (Shandong)	Semi-solid
28	DJN-2	Jining (Shandong)	Semi-solid
29	DJA	Jinan (Shandong)	Semi-solid
30	HZK-3	Zhoukou (Henan)	Semi-solid

（continuation of table 1）

Serial number	Sample name	Sample source	Sample state

31	DZB-2	Zibo (Shandong)	Semi-solid
32	DJN-3	Jining (Shandong)	Semi-solid
33	TJS	Tianjin	Semi-solid
34	HPY	Puyang (Henan)	Semi-solid
35	DZB-3	Zibo (Shandong)	Semi-solid
36	DZB-1	Zibo (Shandong)	Semi-solid
37	DJN-6	Jining (Shandong)	Semi-solid
38	DJN-7	Jining (Shandong)	Semi-solid
39	XLL	Lvliang (Shanxi)	Semi-solid
40	DWF-2	Weifang (Shandong)	Semi-solid
41	DHZ	Heze (Shandong)	Semi-solid
42	DZZ	Zaozhuang (Shandong)	Semi-solid
43	DJN-5	Jining (Shandong)	Semi-solid
44	DJN-4	Jining (Shandong)	Semi-solid
45	AFY-2	Fuyang (Anhui)	Semi-solid
46	AFY-1	Fuyang (Anhui)	Semi-solid
47	DWF-1	Weifang (Shandong)	Semi-solid
48	HZK-1	Zhoukou (Henan)	Semi-solid
49	ASZ	Suzhou (Anhui)	Semi-solid
50	GBY	Baiyin (Gansu)	Semi-solid
51	PDS	Pingdingshan (Henan)	Semi-solid
52	AFY-3	Fuyang (Anhui)	Semi-solid
53	SWN	Weinan (Shaanxi)	Semi-solid
54	DLC	Liaocheng (Shandong)	Semi-solid
55	TLM	Tarim (Xinjiang)	Semi-solid
56	HXX	xinxiang (Henan)	Semi-solid
57	SXA	Xi 'an (Shaanxi)	Semi-solid
58	ZHZ	Hangzhou (Zhejiang)	Semi-solid
59	SWN-1	Weinan (Shaanxi)	Semi-solid
60	SWN-2	Weinan (Shaanxi)	Semi-solid
61	SWN-3	Weinan (Shaanxi)	Semi-solid
62	SWN-4	Weinan (Shaanxi)	Semi-solid

（continuation of table 1）

Serial number	Sample name	Sample source	Sample state

63	HJ	Jiaozuo (Henan)	Semi-solid
64	HNSQ-1	Shangqiu (Henan)	Semi-solid
65	HNSQ-2	Shangqiu (Henan)	Semi-solid
66	HNSQ-3	Shangqiu (Henan)	Semi-solid
67	HNXX-1	Xinxiang (Henan)	Semi-solid
68	HNXX-2	Xinxiang (Henan)	Semi-solid
69	HNZZ-1	Zhengzhou (Henan)	Semi-solid
70	HP-5	Puyang (Henan)	Semi-solid
71	HP-6	Puyang (Henan)	Semi-solid
72	SDB	Binzhou (Shandong)	Semi-solid
73	SDBZ1	Binzhou (Shandong)	Semi-solid
74	SDDZ1	Dezhou (Shandong)	Semi-solid
75	SDDZ2	Dezhou (Shandong)	Semi-solid
76	SDHZ1	Heze (Shandong)	Semi-solid
77	SDJN1	Jining (Shandong)	Semi-solid
78	SLH	Linfen (Shanxi)	Semi-solid
79	SLH-1	Linfen (Shanxi)	Semi-solid
80	SLH-3	Linfen (Shanxi)	Semi-solid
81	SLH-5	Linfen (Shanxi)	Semi-solid
82	SLX	Linfen (Shanxi)	Semi-solid
83	SLY-1	Linfen (Shanxi)	Semi-solid
84	SLY-2	Linfen (Shanxi)	Semi-solid
85	SLY-3	Linfen (Shanxi)	Semi-solid
86	SP-1	Jinzhong (Shanxi)	Semi-solid
87	SXD-2	Xinzhou (Shanxi)	Semi-solid
88	SXD-3	Xinzhou (Shanxi)	Semi-solid
89	SXD-4	Xinzhou (Shanxi)	Semi-solid
90	SXD-6	Xinzhou (Shanxi)	Semi-solid
91	SXTY	Taiyuan (Shanxi)	Semi-solid
92	SXY-2	Xinzhou (Shanxi)	Semi-solid
93	SYJ-1	Yuncheng (Shanxi)	Semi-solid
94	SYJ-2	Yuncheng (Shanxi)	Semi-solid

（continuation of table 1）

Serial number	Sample name	Sample source	Sample state

95	SYX-1	Yuncheng (Shanxi)	Semi-solid
96	SYX-2	Yuncheng (Shanxi)	Semi-solid
97	SXBT-1	Yan 'an (Shaanxi)	Semi-solid
98	SXN	Yan 'an (Shaanxi)	Semi-solid
99	SXW-2	Yan 'an (Shaanxi)	Semi-solid
100	SXW-4	Yan 'an (Shaanxi)	Semi-solid
101	SXW-5	Yan 'an (Shaanxi)	Semi-solid
102	SXWN-1	Weinan (Shaanxi)	Semi-solid
103	SXWN-2	Weinan (Shaanxi)	Semi-solid
104	SXWN-3	Weinan (Shaanxi)	Semi-solid
105	SXWN-4	Weinan (Shaanxi)	Semi-solid
106	SXWN-5	Weinan (Shaanxi)	Semi-solid
107	SXWN-6	Weinan (Shaanxi)	Semi-solid
108	SXWN-7	Weinan (Shaanxi)	Semi-solid
109	SXWN-8	Weinan (Shaanxi)	Semi-solid
110	SXXA-1	Xi 'an (Shaanxi)	Semi-solid
111	SXXA-2	Xi 'an (Shaanxi)	Semi-solid
112	SXXA-3	Xi 'an (Shaanxi)	Semi-solid
113	SXXA-4	Xi 'an (Shaanxi)	Semi-solid
114	SXXA-5	Xi 'an (Shaanxi)	Semi-solid
115	SXXA-6	Xi 'an (Shaanxi)	Semi-solid

Note: Semi-solid is wet sample, solid is dry sample. The name of the province is in parentheses.

Physicochemical index：

Table 2　Physicochemical indexes of dry samples

Dry Sample	Sample name	Moisture content	pH	TTA
1	HNY-1	13.98±0.08	5.00±0.01	6.88±0.02
2	HNY-2	15.80±0.03	4.12±0.02	8.19±0.08
3	HZK-2	11.10±0.23	3.99±0.03	12.40±0.18
4	HLY	11.54±0.02	5.09±0.02	12.60±0.10
5	HSQ	13.50±0.27	5.80±0.03	14.56±0.02
6	ZMD	12.74±0.00	4.82±0.00	6.95±0.04
7	DWH-2	11.38±0.81	5.07±0.02	11.40±0.52
8	DWH-1	13.74±0.18	5.36±0.02	15.88±0.02
9	DDY	16.88±0.35	5.53±0.02	12.29±0.51
10	HZZ	14.85±2.24	4.67±0.02	5.99±0.50
11	GWW	7.75±2.44	4.51±0.04	9.35±0.22
12	DQD	12.50±0.86	4.21±0.00	17.70±0.11
13	JTZ	13.99±0.07	9.69±0.06	/
14	XYC	12.94±0.14	5.52±0.1	5.73±0.33
15	HNLY-1	14.7±0.26	4.97±0.04	14.07±0.05
16	HNSQ-4	11.8±0.20	5.30±0.00	10.49±0.03
17	HNZK-1	11.8±0.20	5.30±0.00	10.49±0.03
18	HP-1	13.3±0.10	5.89±0.03	14.00±0.23
19	HP-4	14.6±0.10	5.70±0.02	17.94±0.63
20	HX	10.5±0.17	5.00±0.03	13.12±0.30
21	SDHZ-2	14.3±0.10	5.95±0.03	6.48±0.04
22	SDHZ-3	10.1±0.17	4.80±0.00	12.69±0.27
23	SXWT1	14.3±0.10	5.83±0.09	14.68±0.46
24	SYJ-3	9.7±0.17	5.55±0.16	9.93±0.26
25	SYX-3	8.8±0.10	5.59±0.07	13.00±0.38
26	SXY	11.5±0.17	4.40±0.09	10.05±0.55

Table 3 Physicochemical indices of wet samples

Wet Sample	Sample Name	Moisture content	pH	TTA

1	DJN-1	39.33±0.43	4.12±0.01	12.35±0.16
2	DJN-2	37.97±0.26	6.16±0.03	3.26±0.16
3	DJA	41.96±0.82	5.39±0.02	4.47±0.14
4	HZK-3	43.42±0.11	4.88±0.00	5.72±0.11
5	DZB-2	46.13±0.27	3.82±0.00	13.2±0.17
6	DJN-3	45.25±0.27	5.27±0.04	2.72±0.26
7	TJS	44.46±0.01	3.92±0.00	8.30±0.06
8	HPY	41.89±0.22	3.71±0.02	12.02±0.38
9	DZB-3	37.08±0.38	4.34±0.04	5.74±0.11
10	DZB-1	45.38±0.23	4.16±0.01	12.12±0.19
11	DJN-6	49.76±0.29	3.68±0.46	8.03±0.06
12	DJN-7	43.14±0.39	4.22±0.02	8.97±0.18
13	XLL	43.81±0.26	4.56±0.15	6.72±0.42
14	DWF-2	39.43±0.09	3.94±0.02	10.07±0.25
15	DHZ	43.11±0.19	4.15±0.01	6.53±0.06
16	DZZ	41.10±0.17	4.31±0.01	8.36±0.57
17	DJN-5	42.21±0.00	3.89±0.01	12.72±0.08
18	DJN-4	43.32±0.20	6.07±0.02	2.60±0.16
19	AFY-2	43.08±0.65	4.37±0.03	3.06±2.25
20	AFY-1	38.29±0.27	4.51±0.01	5.81±0.13
21	DWF-1	40.61±0.59	4.08±0.01	13.71±0.18
22	HZK-1	46.84±0.38	3.97±0.03	11.3±0.07
23	ASZ	43.68±0.18	4.00±0.02	8.76±0.11
24	GBY	35.81±0.38	4.08±0.02	6.42±0.49
25	PDS	45.15±0.96	4.22±0.02	7.61±0.04
26	AFY-3	31.19±0.24	4.53±0.05	9.51±0.75
27	SWN	42.96±0.49	4.47±0.23	9.36±0.04
28	DLC	40.29±0.03	3.78±0.01	13.34±0.9
29	TLM	42.69±0.20	4.18±0.02	8.88±0.15
30	HXX	49.45±0.10	5.51±0.03	3.57±0.56

Continuation of table 3:

Wet Sample	Sample Name	Moisture content	pH	TTA

31	SXA	49.98±0.21	3.9±0.07	12.48±0.08
32	ZHZ	41.59±0.35	3.91±0.02	13.32±0.39
33	SWN-1	48.52±0.19	4.62±0.00	10.02±0.24
34	SWN-2	18.83±0.25	4.16±0.02	14.52±0.49
35	SWN-3	40.98±0.07	4.21±0.01	10.88±0.02
36	SWN-4	47.95±0.04	4.20±0.04	12.87±0.46
37	HJ	19.5±0.36	3.79±0.03	18.05±1.65
38	HNSQ-1	19.5±0.35	4.06±0.02	11.61±0.04
39	HNSQ-2	39.8±0.44	4.51±0.01	8.50±0.06
40	HNSQ-3	37.9±0.69	4.40±0.00	8.90±0.38
41	HNXX-1	35.7±0.46	3.90±0.06	14.11±0.09
42	HNXX-2	40.6±0.10	4.65±0.01	6.05±0.07
43	HNZZ-1	37.8±0.70	3.85±0.01	13.24±0.12
44	HP-5	36.4±0.40	4.21±0.05	8.47±0.05
45	HP-6	43.6±0.62	3.72±0.00	13.05±0.45
46	SDB	20.0±0.26	4.78±0.04	14.87±0.34
47	SDBZ1	22.1±0.20	5.35±0.03	12.00±0.02
48	SDDZ1	36.2±0.69	4.08±0.00	10.33±0.02
49	SDDZ2	36.8±0.26	3.88±0.01	12.31±0.19
50	SDHZ1	19.2±0.35	4.13±0.02	12.89±0.25
51	SDJN1	36.4±0.70	4.40±0.01	9.79±0.17
52	SLH	24.9±0.35	4.09±0.02	19.85±0.08
53	SLH-1	30.3±0.46	3.70±0.01	19.67±0.17
54	SLH-3	26.4±0.44	3.89±0.01	18.97±0.33
55	SLH-5	37.6±0.72	3.74±0.02	19.37±0.08
56	SLX	28.1±0.50	5.31±0.01	4.79±0.03
57	SLY-1	26.4±0.36	4.13±0.01	10.79±0.13
58	SLY-2	25.0±0.20	4.06±0.01	11.20±0.10
59	SLY-3	28.9±0.52	3.68±0.01	10.21±0.03
60	SP-1	35.9±0.66	3.73±0.01	17.00±0.33

Continuation of table 3:

Wet Sample	Sample Name	Moisture content	pH	TTA

61	SXD-2	48.4±0.44	3.81±0.05	12.99±0.39
62	SXD-3	42.9±0.70	3.86±0.02	19.49±0.03
63	SXD-4	44.2±0.40	3.81±0.03	16.97±1.00
64	SXD-6	46.0±0.35	3.76±0.02	17.25±0.18
65	SXTY	18.3±0.30	3.71±0.03	17.66±0.38
66	SXY-2	38.5±0.44	3.94±0.02	12.19±0.07
67	SYJ-1	33.5±0.52	4.00±0.02	13.04±0.50
68	SYJ-2	30.1±0.46	3.72±0.01	17.99±0.84
69	SYX-1	36.0±0.69	3.94±0.02	15.55±0.03
70	SYX-2	16.5±0.26	6.50±0.04	8.91±0.31
71	SXBT-1	19.5±0.36	6.33±0.08	1.44±0.16
72	SXN	20.4±0.10	4.03±0.01	11.29±0.06
73	SXW-2	17.4±0.26	4.13±0.01	9.21±0.08
74	SXW-4	35.7±0.61	4.81±0.01	4.78±0.07
75	SXW-5	21.4±0.17	3.75±0.01	19.60±0.04
76	SXWN-1	19.8±0.36	5.15±0.03	11.77±0.05
77	SXWN-2	24.9±0.44	5.27±0.00	6.61±0.03
78	SXWN-3	39.0±0.36	5.18±0.03	5.72±0.30
79	SXWN-4	36.3±0.26	3.79±0.07	12.27±0.03
80	SXWN-5	24.8±0.17	4.09±0.03	11.58±0.96
81	SXWN-6	33.9±0.61	4.36±0.02	9.31±0.07
82	SXWN-7	27.2±0.26	4.05±0.06	13.72±0.07
83	SXWN-8	21.1±0.35	3.90±0.00	11.26±0.02
84	SXXA-1	34.5±0.61	4.08±0.00	8.90±0.02
85	SXXA-2	36.2±0.44	3.99±0.01	10.78±0.04
86	SXXA-3	29.2±0.56	4.75±0.02	7.52±0.20
87	SXXA-4	31.7±0.40	4.00±0.01	10.49±0.03
88	SXXA-5	30.6±0.56	6.30±0.02	3.56±0.04
89	SXXA-6	37.9±0.20	6.92±0.09	4.31±0.93

Table. 4 Lactic acid bacteria isolated from sourdough

No.	Strains name	Sample distribution

		Henan	Shandong	Shanxi	Shaanxi	Frequency of strain identification
1	*Enterococcus faecium*	-	2	-	-	2
2	*Weissella cibaria*	10	8	11	13	33
3	*Weissella confusa*	2	4	-	-	6
4	*Leuconostoc fallax*	-	1	-	-	1
5	*Leuconostoc citreum*	8	3	9	9	29
6	*Pediococcus acidilactici*	-	-	3	-	3
7	*Pediococcus pentosaceus*	7	6	11	5	29
8	*Levilactobacillus brevis*	5	3	9	5	22
9	*Latilactobacillus sakei*	-	-	-	3	3
10	*Leuconostoc mesenteroides*	-	-	-	5	5
11	*Latilactobacillus curvatus*	7	3	-	2	12
12	*Bacillus amyloliquefaciens*	-	-	1	1	2
13	*Furfurilactobacillus rossiae*	-	-	-	2	2
14	*Lacticaseibacillus paracasei*)	-	1	3	1	5
15	*Lactiplantibacillus plantarum*	8	5	11	9	42
16	*Companilactobacillus crustorum*	5	5	10	9	29
17	*Fructilactobacillus sanfranciscensis*	12	4	16	18	50
	Total	64	45	84	82	275

Table. 5 Yeast isolated from sourdough

No.	Strains name	Sample distribution				Frequency of strain identification
		Henan	Shandong	Shanxi	Shaanxi	
1	*Pichia kudriavzevii*	-	-	4	1	5
2	*Saccharomyces cerevisiae*	15	8	21	16	60
3	*Maudiozyma humilis*	5	2	12	16	35
4	*Pichia membranifaciens*	2	-	-	-	2
5	*Torulaspora delbrueckii*	3	2	-	-	5
6	*Saccharomycopsis fibuligera*	-	-	11	-	11
7	*Meyerozyma guilliermondii*	-	-	3	-	3
8	*Hanseniaspora opuntiae*	-	2	-	-	2
9	*Wickerhamomyces anomalus*	10	4	15	8	37
	Total	35	18	66	41	160

Alpha Diversity Index inter-group difference test:

Table 6 Index difference test data table (bacteria) :

Estimators	D-Mean	D-Sd	S-Mean	S-Sd	P_value(D-S)	P_adjust(D-S)
sobs	57.346	32.172	55.472	35.535	0.5165	1
shannon	1.2915	0.5263	0.86796	0.65291	0.002271	0.006814
simpson	0.40596	0.21113	0.62388	0.27614	0.0002387	0.001432
ace	92.206	52.887	93.041	54.408	0.891	1
chao	79.48	40.023	81.262	50.217	0.7943	1
coverage	0.9994	0.00033127	0.99942	0.00030026	1	1

Table 7 Index difference test data table (fungi) :

Estimators	D-Mean	D-Sd	S-Mean	S-Sd	P_value(D-S)	P_adjust(D-S)
sobs	54.538	32.505	115.71	85.53	0.0001983	0.00119
shannon	0.83599	0.51341	1.3529	0.85292	0.007866	0.007866
simpson	0.61944	0.21801	0.47612	0.25248	0.004631	0.006608
ace	101.01	61.986	161.96	104.44	0.002895	0.005789
chao	82.493	52.859	148.05	99.622	0.0005816	0.001745
coverage	0.99935	0.00040432	0.99896	0.00077049	0.005506	0.006608

Note: “S” stands for wet sample and “D” for dry sample. 1 is listed as the diversity index type and the rest as the difference test statistics for the corresponding grouping category. Sd is standard deviation; Pvalue is the false positive probability value, and Qvalue is the FDR value.

Characteristic strains identified by random forest algorithm and SUM method:

Table 8 Top 50 important bacteria identified with moisture as characteristic label

OTU ID	**Strain Name**
OTU496	Unclassified Bacteria
OTU389	Rhodobacteraceae
OTU3	*Leuconostoc*
OTU159	*Pediococcus*
OTU668	*Malikia*
OTU587	*Dermacoccus_nishinomiyaensis*
OTU612	*Prevotella*
OTU426	*Lolium_perenne*
OTU399	*Leuconostoc_fallax*
OTU750	*Carnobacterium_maltaromaticum*
OTU161	unclassified Bacteria
OTU34	Muribaculaceae

OTU510	Leuconostocaceae
OTU400	*Gluconobacter*
OTU347	UCG-010

Continue Table 8:

OTU ID	**OTU ID**
OTU143	*Acetobacter*
OTU176	Lactobacillaceae
OTU2	*Wolbachia*
OTU266	*Serratia_marcescens*
OTU686	Frankiales
OTU516	Lactobacillales
OTU165	Uncultured bacterium 67-14
OTU344	Lachnospiraceae_UCG-006
OTU306	*Triticum_aestivum*
OTU726	Chloroplast
OTU518	*Pseudomonas_rhizosphaerae*
OTU622	*Nocardioides*
OTU179	Gammaproteobacteria
OTU158	*Pediococcus*
OTU766	*Lactobacillus_reuteri*
OTU204	*Thermoactinomyces_vulgaris*
OTU429	Mitochondria
OTU783	*Rheinheimera*
OTU231	Mitochondria
OTU39	*Triticum_aestivum*
OTU441	*Triticum_aestivum*
OTU149	Blfdi19
OTU324	Uncultured bacterium 0319-7H2
OTU677	*Azospira*
OTU299	*Lactobacillus_delbrueckii_subsp._bulgaricus*
OTU309	Saccharimonadales
OTU507	Gammaproteobacteria
OTU520	*_Leuconostoc*
OTU414	*Lactobacillus*
OTU545	Enterobacterales
OTU584	*Pseudomonas_psychrophila*
OTU703	*Subsaxibacter*
OTU600	*Rothia_kristinae*
OTU126	*Jeotgalicoccus*
OTU780	Lactobacillales

Table 9 Top 50 important bacteria identified with pH as characteristic label

OTU ID	Strain Name
OTU3	*Leuconostoc*
OTU173	*Lactobacillus*
OTU668	*Malikia*
OTU283	*Alistipes*
OTU159	*Pediococcus*
OTU496	unclassified Bacteria
OTU587	*Dermacoccus_nishinomiyaensis*
OTU570	*Sphingomonas*
OTU143	*Acetobacter*
OTU34	Muribaculaceae
OTU727	*Halanaerobium_kushneri*
OTU473	*Lactobacillus_paracasei*
OTU776	*Pseudomonas_geniculata*
OTU766	*Lactobacillus_reuteri*
OTU2	*Wolbachia*
OTU161	unclassified Bacteria
OTU639	Sva0485
OTU622	*Nocardioides*
OTU168	*Gluconobacter*
OTU250	*Lactobacillus*
OTU400	*Gluconobacter*
OTU149	Blfdi19
OTU735	*Azospirillum*
OTU507	Gammaproteobacteria
OTU772	*Eubacterium_rectale*
OTU229	*Lactobacillus*
OTU389	Rhodobacteraceae
OTU344	Lachnospiraceae
OTU126	*Jeotgalicoccus*
OTU484	*Lactobacillus*
OTU158	*Pediococcus*
OTU347	UCG-010
OTU518	*Pseudomonas_rhizosphaerae*
OTU231	Mitochondria
OTU17	Lactobacillales
OTU59	*Hordeum_vulgare_subsp._vulgare*
OTU785	Bacilli
OTU193	*Serratia*

OTU426	*Lolium_perenne*
OTU487	unclassified Bacteria
OTU703	*Subsaxibacter*
OTU254	*Pseudomonas*

Continue Table 9:

OTU ID	**Strain Name**
OTU12	Rubrobacter
OTU608	*Prevotella_melaninogenica*
OTU75	Chloroplast
OTU118	*Triticum_aestivum*
OTU266	*Serratia_marcescens*
OTU812	Proteobacteria
OTU835	Bacilli
OTU117	unclassified Bacteria

Table 10 Top 50 important bacteria identified with TTA as the characteristic label

OTU ID	**Strain Name**
OTU776	*Pseudomonas_geniculata*
OTU3	*Leuconostoc*
OTU426	*Lolium_perenne*
OTU338	unclassified Bacteria
OTU587	*Dermacoccus_nishinomiyaensis*
OTU780	Lactobacillales
OTU400	*Gluconobacter*
OTU622	*Nocardioides*
OTU159	*Pediococcus*
OTU34	Muribaculaceae
OTU17	Lactobacillales
OTU584	*Pseudomonas_psychrophila*
OTU2	*Wolbachia*
OTU202	*Saccharopolyspora_rectivirgula*
OTU250	*Lactobacillus*
OTU518	*Pseudomonas_rhizosphaerae*
OTU143	*Acetobacter*
OTU231	Mitochondria
OTU685	Bacteroidetes
OTU507	Gammaproteobacteria
OTU739	*Helicobacter_typhlonius*
OTU376	*Triticum_aestivum*
OTU520	*Leuconostoc*
OTU496	unclassified Bacteria
OTU212	*Blastococcus_aggregatus*
OTU283	*Alistipes*
OTU545	Enterobacterales

OTU165	Uncultured bacterium
OTU808	*Lactobacillus_graminis*
OTU836	Lactobacillales
OTU519	*Algoriella*

Continue table 10:

OTU ID	**Strain Name**
OTU677	*Azospira_sp.*
OTU766	s__Lactobacillus_reuteri
OTU204	*Thermoactinomyces_vulgaris*
OTU600	*Rothia_kristinae*
OTU661	Rhizobiales
OTU741	unclassified Bacteria
OTU174	Enterobacterales
OTU703	*Subsaxibacter*
OTU512	Enterobacteriaceae
OTU302	*Lactobacillus*
OTU414	*_Lactobacillus*
OTU158	*Pediococcus*
OTU863	*Weissella*
OTU616	*Blautia*
OTU668	*Malikia*
OTU429	Mitochondria
OTU527	*Chryseobacterium*
OTU408	*Lactobacillus*
OTU211	*Sarcina*

Table 11 Top 50 important fungi identified with moisture as characteristic label

OTU ID	**Strain Name**
OTU965	*Kodamaea_ohmeri*
OTU217	Agaricomycetes
OTU1113	unclassified Fungi
OTU565	*Rozellomycota_sp*
OTU179	*Schizothecium_sp*
OTU819	*Curvularia_hominis*
OTU1455	unclassified Fungi
OTU30	*Rhizopus_microsporus*
OTU472	Hypocreales
OTU1086	*Candida_intermedia*
OTU152	*Lecanicillium_antillanum*
OTU473	Didymellaceae
OTU1237	*Acremonium_curvulum*
OTU477	*Buckleyzyma_phyllomatis*
OTU1351	*Alternaria_infectoria*
OTU337	*Chrysosporium_carmichaelii*

OTU624	Hypocreales
OTU361	*Aspergillus_fumigatus*
OTU1179	*Septoriella_phragmitis*
OTU1254	Pleosporales

Continue table 11:

OTU ID	**Strain Name**
OTU469	Bipolaris
OTU509	*Zygopleurage_zygospora*
OTU912	*Candida_argentea*
OTU1261	*Aspergillus_penicillioides*
OTU500	*Bionectria_pityrodes*
OTU66	*Apiotrichum*
OTU974	Ascomycota
OTU1405	*Cystofilobasidium_macerans*
OTU535	*Sistotrema_sp*
OTU1320	*Aspergillus*
OTU802	*Clitocybe_trulliformis*
OTU1484	*Neocosmospora_ramosa*
OTU355	*Penicillium_citrinum*
OTU242	*Byssochlamys_sp*
OTU333	*Phialosimplex_salinarum*
OTU1522	*Candida_sake*
OTU1688	*Pichia_cactophila*
OTU258	Ascomycota
OTU1080	*Issatchenkia_orientalis*
OTU56	unclassified Fungi
OTU1617	Ascomycota
OTU35	*Penicillium_sumatraense*
OTU1345	*Aspergillus_minisclerotigenes*
OTU1066	Ascomycota
OTU342	*Trichoderma_asperellum*
OTU580	Basidiomycota
OTU1625	*Emericellopsis*
OTU858	Ascomycota
OTU1480	*Coniochaeta_ligniaria*
OTU1445	*Millerozyma_farinosa*

Table 12 Top 50 important fungi identified with pH as characteristic label

OTU ID	**Strain Name**
OTU241	*Torula_sp*
OTU380	*Candida_saitoana*
OTU217	Agaricomycetes
OTU625	*Scutellinia*
OTU477	*Buckleyzyma_phyllomatis*

OTU819	*Curvularia_hominis*
OTU353	*Apiotrichum_brassicae*
OTU580	Basidiomycota
OTU529	*Mortierella_humilis*

Continue table 12:

OTU ID	**Strain Name**
OTU355	*Penicillium_citrinum*
OTU982	*Ramophialophora*
OTU214	*Talaromyces_radicus*
OTU453	*Pyrenophora_graminea*
OTU1164	unclassified Fungi
OTU1558	unclassified Fungi
OTU1423	*Alternaria_macrospora*
OTU576	unclassified Fungi
OTU152	*Lecanicillium_antillanum*
OTU461	Ascomycota
OTU1455	unclassified Fungi
OTU441	Wickerhamomyces_onychis
OTU1118	unclassified Fungi
OTU802	*Clitocybe_trulliformis*
OTU663	*Rozellomycota_sp*
OTU1163	unclassified Fungi
OTU361	Aspergillus_fumigatus
OTU243	unclassified Fungi
OTU1066	Ascomycota
OTU909	*Pichia_mandshurica*
OTU1480	*Coniochaeta_ligniaria*
OTU707	*Aspergillus_aculeatus*
OTU66	*Apiotrichum*
OTU1245	unclassified Fungi
OTU1587	*Paraconiothyrium_brasiliense*
OTU500	*Bionectria_pityrodes*
OTU912	*Candida_argentea*
OTU624	Hypocreales
OTU1405	*Cystofilobasidium_macerans*
OTU1022	Dipodascaceae
OTU719	*Clavispora_lusitaniae*
OTU1345	*Aspergillus_minisclerotigenes*
OTU1681	*Chalastospora_gossypii*
OTU1336	*Candida_tropicalis*
OTU967	*Dipodascaceae_sp*
OTU938	*Bipolaris*
OTU1113	unclassified Fungi
OTU979	*Trichoderma_simmonsii*

OTU1029	*Ceratobasidium_sp*
OTU1080	*Issatchenkia_orientalis*
OTU965	*Kodamaea_ohmeri*

Table 13 Top 50 important fungi identified with TTA as characteristic label

OTU ID	**Strain Name**
OTU932	Xylariales
OTU477	*Buckleyzyma_phyllomatis*
OTU1179	*Septoriella_phragmitis*
OTU802	*Clitocybe_trulliformis*
OTU1345	*Aspergillus_minisclerotigenes*
OTU663	*Rozellomycota_sp*
OTU1455	unclassified Fungi
OTU1320	*Aspergillus*
OTU509	*Zygopleurage_zygospora*
OTU355	*Penicillium_citrinum*
OTU819	*Curvularia_hominis*
OTU179	*Schizothecium_sp*
OTU152	*Lecanicillium_antillanum*
OTU449	*Alternaria*
OTU912	*Candida_argentea*
OTU580	Basidiomycota
OTU1454	Basidiomycota
OTU241	*Torula_sp*
OTU535	*Sistotrema_sp*
OTU1578	*Dothiorella_viticola*
OTU624	Hypocreales
OTU965	*Kodamaea_ohmeri*
OTU177	*Trichomeriaceae_sp*
OTU1326	*Saccharomycopsis_fibuligera*
OTU30	*Rhizopus_microsporus*
OTU243	Unclassified Fungi
OTU1617	Ascomycota
OTU1583	*Filobasidium_wieringae*
OTU576	Unclassified Fungi
OTU1080	*Issatchenkia_orientalis*
OTU781	*Rozellomycota_sp*
OTU1380	*Gibberella_intricans*
OTU1237	*Acremonium_curvulum*
OTU1086	*Candida_intermedia*
OTU1530	*Candida_blankii*
OTU337	*Chrysosporium_carmichaelii*
OTU1405	*Cystofilobasidium_macerans*
OTU1221	*Rozellomycota_sp*

OTU1685	*Rozellomycota_sp*
OTU217	*Agaricomycetes*
OTU701	Ascomycota
OTU909	*Pichia_mandshurica*

Continue table 13:

OTU ID	**Strain Name**
OTU259	*Rosellinia_australiensis*
OTU1190	*Hypoxylon_sp*
OTU473	Didymellaceae
OTU500	*Bionectria_pityrodes*
OTU1028	_Ascomycota
OTU476	*Trichomerium_sp*
OTU1285	*Microascus_brevicaulis*
OTU704	Archaeorhizomycetes

Table. 14 Top50 important bacteria identified by SUM method

OTU ID	**Strain Name**
OTU3	*Leuconostoc*
OTU34	Muribaculaceae
OTU159	*Pediococcus*
OTU400	*Gluconobacter*
OTU143	*_Acetobacter*
OTU622	*Nocardioides*
OTU250	*Lactobacillus*
OTU587	*Dermacoccus_nishinomiyaensis*
OTU496	unclassified Bacteria
OTU668	*Malikia*
OTU507	Gammaproteobacteria
OTU518	Pseudomonas_rhizosphaerae
OTU229	*Lactobacillus*
OTU158	*Pediococcus*
OTU502	*Clostridium*
OTU117	unclassified Bacteria
OTU17	Lactobacillales
OTU169	*Lactobacillus*
OTU355	*Sphingomonas*
OTU686	Frankiales
OTU834	*Lactobacillus*

OTU453	*Levilactobacillus_ brevis*
OTU241	Sphingobacterium
OTU2	Wolbachia

Continue table 14:

OTU ID	Strain Name
OTU776	*Pseudomonas_geniculata*
OTU113	*Triticum_aestivum*
OTU761	Lactobacillales
OTU520	*Leuconostoc*
OTU676	Uncultured bacterium
OTU344	Lachnospiraceae

Table. 15 Top50 important fungi identified by SUM method

OTU ID	Strain Name
OTU819	*Curvularia_hominis*
OTU1455	unclassified Fungi
OTU477	*Buckleyzyma_phyllomatis*
OTU217	Agaricomycetes
OTU912	*Candida_argentea*
OTU802	*Clitocybe_trulliformis*
OTU663	*Rozellomycota_sp*
OTU624	Hypocreales
OTU473	Didymellaceae
OTU576	unclassified Fungi
OTU449	*Alternaria*
OTU152	*Lecanicillium_antillanum*
OTU355	*Penicillium_citrinum*
OTU66	*Apiotrichum*
OTU580	Basidiomycota
OTU1113	unclassified Fungi
OTU701	Ascomycota
OTU1345	*Aspergillus_minisclerotigenes*
OTU909	*Pichia_mandshurica*
OTU241	*Torula_sp*
OTU1578	*Dothiorella_viticola*
OTU1405	*Cystofilobasidium_macerans*
OTU965	*Kodamaea_ohmeri*
OTU1320	*Aspergillus*
OTU1562	unclassified Fungi
OTU1080	s__Issatchenkia_orientalis
OTU535	*Sistotrema_sp*
OTU1339	*Aspergillus_penicillioides*
OTU957	*Apiospora_montagnei*
OTU1196	*Leptoxyphium*