TITLE: Fungal and bacterial community composition and structure in fermented 'hairy' tofu (Mao tofu)

Authors: Gian Maria Niccolò Benucci[1†], Xinxin Wang[1,2†], Li Zhang[3], Gregory Bonito[1*], Fuqiang Yu[3*]

Affiliations:

[1]Plant Soil and Microbial Sciences, Michigan State University, 1066 Bogue St. 48824 MI, USA

[2]Department of Plant Protection, Shenyang Agricultural University, Shenyang 110866, China

[3]The Germplasm Bank of Wild Species, Yunnan Key Laboratory for Fungal Diversity and Green Development, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, China

*** Co-Correspondence**: Gregory Bonito - bonito@msu.edu

Fuqiang Yu - fqyu@mail.kib.ac.cn

† These authors contributed equally.

Abstract

The process of fermenting tofu extends thousands of years. Despite a resurgent interest in microbial communities and fermented foods, little knowledge exists concerning microbial diversity of communities of fermented ‘hairy’ tofu known in China as Mao tofu. We used high-throughput metagenomic sequencing of the ITS, LSU and 16S rDNA marker genes to disentangle the Mao tofu fungal and bacterial community composition and diversity across the four most important markets in the Yunnan region of China. We show that hairy tofu in this region consists of around 170 fungal and 365 bacterial taxa. Significant differences in community structure were found between markets and niches. Machine learning random forest models were able to accurately classify both market and niche of sample origin. An over-abundance of yeast taxa were detected, and *Geotrichum* were the most abundant fungal taxa, followed by *Torulaspora*, *Trichosporon,* and *Pichia*. *Mucor* (Mucormycota) was also abundant in the LSU data and especially in the outside niche (rind), which consists of the visible ‘hairy’ mycelium. The majority of the bacterial OTUs belonged to Proteobacteria, Firmicutes, and Bacteroidota, with *Acinetobacter*, *Lactobacillus*, *Sphingobacterium* and *Flavobacterium* the most abundant members. Of interest, putative fungal pathogens of plants (e.g. *Cercospora*, *Diaporthe, Fusarium*) and animal (e.g. *Metarhizium*, *Entomomortierella*, *Pyxidiophora*, *Candida*, *Clavispora*), as well as bacterial (e.g. *Legionella*) pathogens, were detected. Non-target eukaryotic taxa detected in by LSU amplicon sequencing included soybean (*Glycine max*), Protozoa, Metazoa (e.g. Nematoda and Platyhelminthes), Rhizaria and Chromista, providing evidence of additional biocomplexity and diversity in the tofu microbiome.

Keywords

Soybean, lactic-acid bacteria, Tremellomycetes, *Geotrichum*, UPARSE, MiSeq amplicon sequencing

1. Introduction

Soybean is an important crop with wide applications in food, livestock and biofuels streams given its high level of protein, fiber, vitamins, minerals, and lipids (Jayachandran & Xu, 2019). Many traditions foods in Asia are produced by fermenting soy, including soy sauces, soy cheese, soy yogurt, stinky tofu, and Mao tofu. Through fermentation, carbohydrates, proteins, and lipids from soybean are broken down by microbial enzymes (Jang et al., 2008), which significantly increases concentration of beneficial compounds including isoflavones, antioxidant capacity, B vitamins, and gamma-Aminobutyric acid levels (Xu, Cai, & Xu, 2017). Fermented soybean products have been shown to have anti-diabetic (Kim et al., 2008), antioxidant (Yoon & Park, 2014), anticancer (Pisani, Parkin, Bray, & Ferlay, 1999; Zhu et al., 2006), anti-inflammatory (Lee et al., 2013), anti-hyperlipidemic (Ren, Chen, Li, Mcgowan, & Lin, 2017) properties. Fermented tofu has also been shown to stimulate blood pressure (Pisani et al., 1999; Tsai, Lin, Pan, & Chen, 2006) immunity (Lee et al., 2013), neural activity (Kang et al., 2016) and provide other health benefits.

Stinky tofu and Mao tofu are popular fermented soy products with origins in China. Through fermentation, stinky tofu and Mao tofu develop a pungent odor and flavor (Gu et al., 2018). Topically, stinky tofu has a smooth surface with a color that varies from golden to black, while Mao tofu is typically hairy in appearance, owing to the growth of fluffy white fungal mycelia. The fermentation principles of stinky tofu and Mao tofu are very different, and the manufacturing procedures can also vary from region to region. Stinky tofu is made by soaking soybean curds into fermented stinky brine for a few hours to several days whereas Mao tofu is made by exposing the curds to open-air

without the explicit addition of any microbiota (Zhao & Zheng, 2009).

Previous studies on the production of stinky tofu have shown lactic acid bacteria (e.g. *Lactobacillus*, *Leuconostoc*, *Streptococcus* etc.) and other species of bacteria play a dominant role in the fermentation of tofu (Chao, Tomii, Sasamoto, et al., 2008; Chao, Tomii, Watanabe, & Tsai, 2008; Liang, Deng, & Lin, 2013; Sun, Zhang, Wang, Wang, & Xie, 2010; Yu, Hu, & Li, 2012). Due to its characteristic moldy and fluffy appearance, fungi are assumed to be particularly important in the fermentation of Mao tofu. It has been hypothesized that the main causative fermenters of Mao tofu are *Mucor spp.*, which are responsible for the fluffy appearance (Zhao & Zheng, 2009). The fermentation of Mao tofu is influenced both by starting materials and environmental factors, including temperature, humidity, fermentation duration, and processing conditions. It is currently unknown which other microbiota contribute to the fermentation of Mao tofu and, and there is a general lack of knowledge on the composition and structure of the microbial communities associated with this food.

To address this, we studied 72 Mao tofu samples across four markets in Yunnan, China. We investigated the fungal and bacterial communities of Mao tofu through high-throughput metagenomic sequencing of the internal transcribed spacer (ITS), the large subunit (LSU) of the nuclear ribosomal DNA, and the 16S ribosomal RNA gene. We aimed to determine: i) the microbial composition and core taxa in fermented-tofu ii) whether microbial communities are structured by geographical patterns by assessing the level of similarity within and between different markets, and; iii) how microbial communities in external and internal niches differ.

2. Material and Methods

2.1 Sampling

We sampled Mao tofu from four markets of Kunming City (Ciba) and Jianshui City (Longjin, Wanyao, and New District) in Yunnan, China (Fig. S1). From each market three individual pieces of Mao tofu were sampled from three independent vendors, and from each piece of tofu we sampled both inside (internal) and outside (external) microbial niches. A total of 72 samples, 36 for each niche were studied.

2.2 *DNA extraction, amplification and NGS library preparation*

Tofu samples were carried back to the lab in sterile plastic bags and ~1g of was collected from the outside of each piece and placed in a cetyl-trimethylammonium bromide (CTAB) 4x DNA buffer. Samples were then carefully split open, and ~1g of internal tofu was sampled with a sterile forcep and placed in a different tube containing CTAB. DNA was extracted from samples through chloroform extraction, and then precipitated and washed with ethanol as previously described (Benucci et al., 2019). Fungal internal transcribed spacer (ITS) region of the ribosomal RNA (rRNA) was amplified with the ITS1F-ITS4 primer set (Gardes & Bruns, 1993; White et al., 1990), and eukaryotic large subunit (LSU) region of the ribosomal RNA (rRNA) was amplified with the primers LROR-LR3 (Hopple & Vilgalys, 1994). Prokaryotic V4 region of the 16S rRNA was amplified with the primers 515F-806R (Caporaso et al., 2010). Amplicons libraries were prepared as described in previous studies (Benucci et al., 2019; Longley et al., 2019; Noel, Chang, & Chilvers, 2020). Amplicon libraries were sequenced on a MiSeq Illumina platform with v3 300 PE chemistry.

2.3 Bioinformatics

Raw data quality of ITS, LSU, and 16S rDNA read data was assessed by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). For each dataset, sequences were then demultiplexed in QIIME according to barcode indices (G. J. Caporaso et al., 2010). Subsequently, Illumina adapters and primers were trimmed off the reads with Cutadapt (Martin, 2011). Reads were then filtered according to maximum expected errors = 0.5 for ITS, 1.0 for LSU, and 0.1 for 16S to account for different error rates obtained during each sequencing run. Conserved regions upstream (SSU) and downstream (5.8S) of ITS1 were removed as described by Benucci et al. (2020). Sequences were trimmed to 200 nucleotides length for ITS and 250 bp for LSU and 16S (R. Edgar, 2016; Edgar & Flyvbjerg, 2015). Sequences were then de-replicated, singleton sequences were removed and remaining sequences were used to create operational taxonomic units (OTUs) at 97% similarity threshold with the UPARSE algorithm (Edgar, 2013).

Taxonomic assignments of OTUs were performed with the RDP Naïve Bayesian Classifier (Wang et al., 2007) against the 16S and LSU representative sequences releases for the 16S and LSU data, respectively, and with CONSTAX (Gdanetz et al., 2017) against the UNITE (V.04.02.2020) fungal rDNA reference database (Abarenkov et al., 2020) for ITS. Ambiguous taxonomy assignments were manually checked with the BLAST algorithm against the NCBI GenBank database (Clark et al., 2016) (https://www.ncbi.nlm.nih.gov/).

2.4 Statistical analyses

OTU tables with 16S, LSU, and ITS rDNA amplicon metadata, taxonomy, and reference sequences files were imported in the R statistical environment (R Core Team, 2020) and combined in *phyloseq* (McMurdie & Holmes, 2014) as objects for the subsequent analysis. Datasets were then filtered, removing all sequences belonging to mitochondria, chloroplast, non-target organisms, and potential contaminants as detected by the decontam package (Davis et al., 2018). Control samples were then removed from the datasets. Plots showing contaminant OTU frequency and histogram of sample libraries distribution are available in Fig. S2. Rarefaction curves were calculated in *vegan* (Oksanen et al., 2019) using the function “rarecurve”. Observed OTU richness and Shannon diversity index were calculated in *vegan* with the “specnumber” and “diversity” functions (Oksanen et al., 2019). Shannon index [

$H = \sum_{i=1}^{k} p_i \log(p_i)$] was then transformed into the Shannon equitability index (

$EH = 1 - \frac{H}{\log(k)}$), with *k* denoting the number of species (i.e. OTUs) and p_i the proportional abundance of species *i*. This normalizes the Shannon index to a value between 0 and 1 with higher values indicating greater evenness.

We adopted Random Forest (RF) models to identify which OTUs across ITS, LSU and 16S datasets differentiated markets and microbial niches. This was accomplished with the “randomforest” function in the *randomForest* R package (Liaw & Wiener, 2002). Random forest models were optimized by testing different numbers of trees to reach the lowest and stable out-of-bag (OOB) error estimate possible, and the

best mtry value (number of features randomly sampled from the entire pool for each tree at each split) with the "tuneRF" function in *randomForest* R package. We then generated a matrix of 1 - proximity matrix and used it to build a multidimensional scaling (MDS) ordination (analog to PCoA) using the “cmdscale” function in the *stats* package to graphically show the prediction obtained with the RF model. The importance of features to differentiate sample groups was assessed by calculating the mean decrease accuracy of the model when a particular OTU is removed from the community. Significance of RF models was assessed using 999 permutations (RF models were repeated 999 times) with the “rf.significance” function in the *rfUtilities* R package (Murphy, Evans, & Storfer, 2010). Additionally, a taxon-group association analysis was used to assess the degree of correlation and significance of each OTU for the target group in relation to other groups and the overlap with the most important features for classification obtained with the RF models. An association analysis was performed with the function *multipatt* with the ”r.g” parameter in the “indicspecies” R package (De Cáceres, Legendre, & Moretti, 2010).

For β-diversity we studied two components: i) community structure, defined as the difference in multivariate space between samples and sample groups, and ii) community dispersion, defined as multivariate variance within each sample group. To visualize these components, we first standardize the data by rescaling each OTU count to 0-1. In this way each OTU is independent from the others; all OTUs have the same scale and different slope, which removes differences in sequencing depth caused by differing library sizes between taxa (Weiss et al., 2017). Second, we performed principal coordinate analysis (PCoA) on Bray-Curtis distance matrix with the function “ordinate” in

phyloseq (McMurdie & Holmes, 2014). A permutational multivariate analysis of variance (PERMANOVA) was used to test differences among a priori defined sample groups (Anderson, 2001) with the functions “adonis” in the *vegan* R package. To identify which groups were significantly different to others (calculate pairwise post-hoc comparisons between factor levels) we used the “calc_pairwise_permanovas” function in the *mctoolsr* R package (Leff, 2017). To assess the amount of multivariate dispersions (Anderson, Ellingsen, & McArdle, 2006) around centroids we used the “betadisper” function in the *vegan* (Oksanen et al., 2019).

To investigate the concordance between the fungal and bacterial communities, and to test whether ITS and LSU showed similar structure, a Procrustes analysis (Gower, 1975) was carried out, combined with a randomization test (Jackson, 1995) with the “protest” function in *vegan* (Oksanen et al., 2019). Concordance represents similarity in β-diversity between two communities across samples or sample groups and can indicate co-occurrence or similar responses of both communities to environmental factors.

Heatmap trees were generated to report taxon relative abundance in different tofu niches with the function “plot_heat” in the *metacoder* (Foster, Sharpton, & Grünwald, 2017) R package. For the LSU heat tree, 11 non-fungal OTUs were retained because of their biological and sanitary importance. All graphs were plotted using *ggplot2* (Wickham, 2016) and *ggpubr* (Kassambara, 2020) R packages.

Overall, significant differences were tested with non-parametric Kruskal-Wallis, pairwise Wilcoxon, and ANOVA-like permutation tests and p-value corrected for multiple comparisons (Bonferroni method) unless specified differently.

3. Results

3.1 Generating OTUs from MiSeq data

After the demultiplexing step we obtained a total of 6,518,019, 16,408,779, and 3,168,370 sequence reads in the ITS, LSU, and 16S rDNA marker datasets, respectively. In total, the 6 negative controls included in the MiSeq run had 0.22, 0.02, and 0.007% of the demultiplexed reads in the ITS, LSU, and 16S datasets, respectively. After removing non-target organisms, unclassified and contaminants OTUs (Fig. S2), we obtained a total of 169, 167 and 365 OTUs for the ITS, LSU, and 16S datasets, distributed in 72 total respective samples. The ITS otu_table had 3,362,770 total counts with an average of 46,705.1 (± 23,018.4 standard deviation) sequence reads per sample, the LSU otu_table had 8,596,493 counts and 119,395.7 ± 75,525.0, and the 16S otu_table had 2,459,152 counts and 34,154.9 ± 17,513.9 sequence reads per sample. We report cumulative read number for each marker divided by market (Fig. S3). Non-target Eukaryota organisms from the LSU dataset were represented by 16 OTUs in total (46,825 counts and 793.6 ±1678.3 sequence reads).

3.2 Alpha diversity

Rarefaction curves (Fig. S4) showed that most of the samples were exhaustively sampled as OTU richness plateaued, but there were a few exceptions. In general, we detected significant differences in OTU richness across markets and niches (Fig. 1). Across the different markets, the inside niche generally had significantly lower richness in respect to those on the outside niche with the exception of Wanyao (and also Longjin

for 16S). The Ciba market had the lowest fungal (Fig. 1 A, B) and bacterial (Fig. 1 C) richness overall, while Mao tofu from Longjin and New District markets had significantly higher richness across all three DNA markers.

The Shannon equitability index (EH) showed no significant differences between markets in the ITS inside niche, but New District market had higher evenness than Ciba on the outside niche (Fig. 2 A). In contrast, the LSU (Fig. 2 B) and 16S (Fig. 2 C) data showed no significant differences between markets in the outside niche, but Ciba and Lonjin samples showed higher evenness than other markets. No significant differences were present when comparing inside and outside niches within different markets (Fig. 2 A,B) with the exception of the New District market (Fig. 2 C) in the 16S dataset, where the inside niche has significantly higher evenness.

3.3 Beta diversity

The principal coordinate analysis (PCoA) ordination graphs showed that samples from different markets cluster separately, primarily, according to the market of origin, and secondarily the inside or outside tofu niche (Fig 3). The first axis, with highest fraction of explained variance (10.5 -11.4%), was clearly driven by differences in community structure between markets in the ITS data (Fig. 3 A) and niche in the LSU and 16S data (Fig. 3 B, C), although the pattern was less robust in the LSU data. Generally, samples from the Ciba market clustered closely together and separately from those of other markets.

The results of PERMANOVA showed that market, niche, and the interaction between the previous two, are significant ($P \leq 0.01$ after Bonferroni correction) in all three DNA marker datasets (Table 1). Market had the highest R^2 in the ITS data

(19.3%) followed by 16S (17.0%) and LSU (15.8%). The niche factor had an R^2 about 5 to 5.6% across markets, and the interaction factor varied from 7.9 to 10.2%, with the lowest value in the ITS dataset (Table 1). Group dispersion around centroids is another important layer of β-diversity since it helps to understand differences between samples within the same group. We found that samples from the Ciba market had higher and significantly different spread with respect to other markets in the ITS and 16S datasets (Fig. 3 D, F), while non-significant differences were found in the LSU dataset (Fig. 3 E). Again, samples from the inside niche were significantly more dispersed (less similar) than those of the outside one in the ITS and LSU datasets (Fig. 3 D, E) but not in the 16S dataset (Fig. 3 F). Pairwise PERMANOVAs were not significant for Market in any of the DNA markers, while inside and outside niches were significantly different in the LSU (R^2 = 0.199, p-value = 0.042 after Bonferroni correction) and 16S (R^2 = 0.192, p-value = 0.014 after Bonferroni correction) datasets.

From the Procrustes analysis we found statistically significant (p= 0.0001, permutations=9999) concordance (i.e. similarity in multivariate β-diversity or community structure) between the ITS and LSU (Fig. 4A), ITS and 16S (Fig. 4B), and LSU and 16S (Fig. 4C) ordinations. As expected, ITS and LSU ordinations had the highest correlation and the lowest m^2 suggesting a good concordance of both markers on capturing the fungal community structure of Mao-tofu. Although lower, ITS-16S and LSU-16S Procrustes rotations showed high concordance between the two ordinations suggesting co-occurrence or interdependence of these sets or organisms. Procrustes residual error plots (Fig. 4D, E, F, Fig. S5) allowed identification of individual samples or sample groups that had the highest concordance. For example, in the ITS-LSU Procrustes

rotation the New District samples had significantly lower mean residual (i.e. better fit) than those of Ciba, while Ciba was the market with the lowest mean residual in the 16S dataset. Interestingly, in the ITS-16S Procrustes rotation, the samples of the inside niche showed significantly lower residual values than those of the outside niche (Fig 4F).

3.4 Random forest models and indicator taxa

We were able to build very accurate Random Forest (RF) models to classify samples to a market or a tofu niche based on the high-throughput DNA marker data. A graphical visualization of model prediction accuracy was generated using a MDS ordination of the 1-proximity matrix and is reported in Fig. 5. The out-of-bag (OOB) error rate estimate, that represents the amount of misclassification performed by the modes, was generally, and lower for market than niche. The best model was obtained for the 16S dataset for tofu prediction of the market of origin where the 9.61% of the predictions were correct (Fig. 5 C). Models obtained for the LSU dataset (Fig.5 B) were not as good, but still 87.5% of the predictions were accurate. Models for the ITS data performed well with 95.83 and 91.67% model accuracy for market and niche (Fig. 5A).

To identify how much the RF model accuracy decreases if we drop a variable front he model we plotted the Mean Decrease Accuracy of the top 20 OTUs for each RF model (Fig. 6). Additionally, we included the *r.g* correlation value (with 0.05 after fdr correction) of each of the 20 OTUs to a sample group. Interestingly, the most important OTUs for classification in the ITS (Fig. 6 A) and LSU (Fig.6 B) RF models are quite different, and correlated to different markets and niches. In particular, OTUs classified

as *Lachancea* sp. (FOTU_12), *Candida tropicalis* (FOTU_14), *Diutina catenulata* (FOTU_7) in the ITS dataset, while *Trichosporon* sp. (EOTU_4), *Mucor* sp. (EOTU_9), and *Torulaspora* sp. (EOTU_3) in the LSU dataset, were the most important to classify different markets. Similarly, *Candida* sp. (FOTU_124), *Geotrichum candidum* (FOTU_163), and *Clavispora lusitaniae* (FOTU_43) for the ITS dataset and *Saturnispora* sp. (EOTU_5), *Debaryomyces* sp. (EOTU_12), and *Candida* sp. (EOTU_2) for the LSU dataset had the highest Mean Decrease Accuracy. Additionally, *Lactobacillus mucosae* (BOTU_33), *Acinetobacter* sp. (BOTU_16), and *Weissella* sp. (BOTU_5) for the markets, and *Acetobacter* sp. (BOTU_15), *Enterococcus* sp. (BOTU_12), and *Corynebacterium* sp. (BOTU_125) for niches were the top OTUs in the 16S dataset (Fig. 6 C). Most of the highly important OTUs for classification were correlated to the Longjin or Ciba market, in the ITS dataset, and to Wanyao, Longjin and New District in the LSU dataset. Similarly, most of the top OTUs to classify different niches were correlated to outside, in the ITS data, and to the inside niche in the LSU dataset. Most of the highly important OTUs to classify different niches in the 16S dataset were also correlated to the outside niche.

3.5 Microbial diversity and composition

Fungal and bacterial taxonomic diversity as well as core taxa (defined here as taxa shared across > 90% of the samples) of Mao tofu was visualized in the heatmap trees (Fig. 7) with an emphasis on the inside niche (i.e. colored nodes).

The ITS dataset showed a dominance of Ascomycota (78.9% relative abundance) with respect to Basidiomycota (16.7%) and Mucoromycota (3.4%).

Geotrichum was the most abundant genus (17.2%) followed by *Trichosporon* (8.0%), *Pichia* (7.7%), *Clavispora* (6.3%), *Dipodascus* (5.7%), and *Candida* (5.6%). These genera were also core taxa in the overall dataset and present in >80% (n=28-31) of the samples in the inside niche (Fig. 7 A) together with *Apiotichum* that showed high frequency (79.2%) but low abundance (1.4%). *Geotrichum candidum* (16.5) and *Clavispora lusitaniae* (5.4%), *Lachancea* sp. (2.3%) and *Diutina catenulata* (1.4%) were present with high frequency in Wanyao and New District but not in Ciba and Lonjin accounting for their power for classifying these markets through the RF models (see Fig. 5 A). *Mucor*, which was observed and expected to be present in this kind of tofu, was in low abundance (1.1%) and frequency (26.3% of the total samples and 30.6% of the inside samples) in the ITS dataset. Other interesting taxa included plant and insect pathogens, such as *Fusarium* (0.6%), *Epicoccum* (0.3%), *Alternaria*, (1.6%) *Metarhizium* (0.3%), *Entomortierella* (0.7%), all with low frequencies (1.4 - 15.3%) across samples.

The LSU dataset (Fig. 7 B) was also composed mostly of Ascomycota (60.1%) and Basidiomycota (14.6%) but with a larger amount of Mucoromycota (7.7%). Mucor was abundant (5.6%) and frequent (present in the 72.2% of the total samples, 58.3% in the inside samples, and 86.1% in the outside samples) with respect to the ITS data. Other core genera were *Geotrichum* (10.7%), *Torulaspora* (8.8%), *Trichosporon* (5.5%), *Pichia* (4.5%), and *Candida* (4.2%), all with frequencies above 80% in the whole dataset and part of the top taxa for market classification in the RF models (See Fig. 6 B). Several plant pathogen and insect associated fungi were detected through this marker, such as *Cladosporium* (0.4%), *Wallemia* (0.4%), *Acremonium* (0.2%), *Fusarium* (0.1%)

and *Cercospora sojina* (0.1%), an actual soybean pathogen, all with low frequencies (1.4 - 4.2% overall) and mostly in the outside niche. One OTU in *Entomortierella* (1.1%) and one in the Laboulbeniomycetes (0.1%) were also present in the outside niche and New District market. Many non-fungal taxa of interest were detected in the LSU data. For example, soybean [*Glycine max* (0.2%)] and other Plantae (1.5%), Protozoa (0.7%) with 9.7% frequency, Metazoa (8.4%) with 44% frequency with the Nematoda (0.5%) with 11.1% frequency and Platyhelminthes (6.3%) with 29% frequency (44.4 in the outside niche) phyla represented, Chromista (1.5%) with 15.3% frequency and Rhizaria (2.7%) with 56.9% frequency (86.1% in the inside niche), all interesting form human health point of view.

Proteobacteria (34.7&), Firmicutes (33.5%), Bacteroidota (15.4) and Actinobacteria (8.5%) were the most abundant and core taxa (Fig. 6 C). Core genera in the inside niche were also core genre overall, *Lactobacillus* (10.7%) present in all samples, *Leuconostoc* (3.2%) with 98.6% frequency; *Dysgonomonas* 3.1%) 69% frequency overall and 25 samples in the inside niche; *Acinetobacter* (2.9%) with 93.1% frequency; *Sphingobacterium* (2.4%) with 87% frequency overall and present in 28 samples in the inside niche; and *Flavobacterium* (2.2%) with 61.1% frequency and in 23 samples in the inside niche. OTUs within *Acinetobacter*, *Lactobacillus mucosae* (0.1%) with 29.1% frequency; *Weissella* (1.4%) with 98.6% frequency; *Enterococcus* (0.9%) with 94.4% frequency; and *Corynebacterium* (0.9%) with 69.4% frequency overall were among those with the highest Mean Decrease Accuracy in the RF models for market and niche classification, respectively. One OTU belonging to *Legionella* (0.01%) was detected in the outside niche of the New District market.

4 Discussion

Microbes play key roles in determining the general quality attributes of any fermented food. To date, several studies about composition and dynamics of microbial communities associated with Chinese traditional fermented foods are available in the literature (He et al., 2017). However, there is no information on the composition and diversity of Mao tofu, an important fermented food, deeply rooted Chinese traditional cuisine, consumed as an appetizer for hundreds of years (He et al., 2017; Yan et al., 2020). Through high-throughput metagenomic sequencing of the internal transcribed spacer (ITS), the large subunit (LSU) of the nuclear rDNA, and the 16S rDNA gene, we investigated fungal and bacterial communities both inside and outside of Mao tofu samples taken across 4 geographically distant markets.

Mao tofu fungal communities were dominated by Ascomycota (varying from 60.1-78.9%), Basidiomycota (14.6-16.7%) and Mucoromycota (3.4-7.7%). Mucoromycota were better represented in the LSU dataset compared to the ITS. This was not unexpected, since ITS can bias against basel fungal lineages which can have longer ITS sequence lengths (Reynolds et al., 2019), and given that the ITS1F primer has multiple central mismatches to nearly all taxa in the subphylum Mucoromycotina (Tedersoo & Lindahl, 2016). *Mucor* was visually apparent and particularly abundant in our tofu samples from the LSU dataset (especially in the outside niche where most of the mycelium develops). As shown previously, this fungus impacts the quality of the final product through proteolytic processes and release of nutrients, impacting the texture and flavor of Mao tofu (Zhang & Zhao, 2010; Zhao & Zheng, 2009). At Class

level, Tremellomycetes and Saccharomycetes included most of the core members. For example, *Geotrichum candidum* was the most abundant OTU (10.7-16.5%) and was shared across almost all samples in the ITS dataset. This yeast is present in soil and plant material but has also been identified as one of the main components of the microflora of soft cheeses such as Camembert and semi-fresh goat's and ewe's milk cheese (Boutrou & Guéguen, 2005; Morel et al., 2015). Comparative genomic studies of yeasts in the Saccharomycetales have shown that *G. candidum* has retained in its genome a set of cellulases that can be used to break down cellulose in the environment. This may explain the ecological ability of *G. candidum* to grow on cellulose-rich plant-derived material like fermented soybean curds (Tamang et al., 2016; Zhao & Zheng, 2009). Other taxa, such as *Clavispora*, *Candida*, *Dipodascus*, *Pichia, Torulaspora, Diutina,* and *Trichosporon,* were also abundant and frequent both inside and outside Mao tofu. Most of these taxa are also components of the cheese microbiome (Büchl & Seiler, 2011), but also other kinds of tofu, such as stinky tofu (Gu et al., 2018), or other fermented soybean products, such as tempeh (Dimidi et al., 2019). Interestingly, some of these non-*Saccharomyces* taxa have been proposed as potential mixed starters for their beneficial activities for the production of various fermented foods and beverages. For example, they increase acidity and improve primary and secondary aroma of wines (Combina et al., 2005; Padilla, Gil, & Manzanares, 2016; van Breda, Jolly, & van Wyk, 2013) or influence foam stability and flavor in beer (Michel et al., 2016). Some others are considered contaminants of dairy products (Delavenne et al., 2011; O'Brien et al., 2018).

Mao tofu bacterial communities were dominated by Proteobacteria, Firmicutes,

and Bacteroidetes (83.6 % total). Within these groups, lactic acid bacteria such as *Lactobacillus* and *Leuconostoc* were present in about all samples regardless of the niche. These bacteria are generally common in fermented foods and already reported for soybean-based fermented products (Fei et al., 2018; Tamang et al., 2016). Although *Lactobacillus*, *Enterococcus* and *Bifidobacterium* are considered probiotic microorganisms due to their antimicrobial and antioxidant properties, radical scavenging and peptide production activities (Tamang et al., 2016), there is still debate on their actual effects on gastrointestinal health and disease in humans (Dimidi et al., 2019).

It has been shown that DNA-based next generation sequencing techniques are sensitive to DNA traces and allow detection of extracellular DNA from dead microorganisms that persist in soil for weeks to years (Carini et al., 2016). We detected potential fungal pathogens of soybean (e.g. *Epicoccum*, *Cercospora*, *Acremonium, Fusarium*) and insects (e.g. *Entomortierella, Metarhizium*, Laboulbeniomycetes). For example, *Cercospora sojina*, the agent of the Frogeye leaf spot disease, often overwinters in soybean residue and seeds, while *Metarhizium anisopliae* is an entomopathogen that has a wide host range (Chen et al., 1999; Wrather et al., 2010; Zimmermann, 1993). These findings may reflect the soybean material used, crop management, and the environment during production or sale. Along with these findings, we detected potentially harmful bacteria (i.e. *Legionella*), flatworms (i.e. Platyhelminthes) and roundworms (i.e. Nematoda), mostly on the outside surface and most of which were not targeted and have not been reported from Mao tofu previously.

Regarding the microbial diversity of Mao tofu, we found that the inside niche had significantly lower richness than the externa tofu niche, and variation in microbial

communities was evident between different markets. Although having slightly different patterns in PCoA ordinations, the community structure component of β-diversity was primarily driven by the geographic location (i.e. the market) in all the selected DNA markers. The market component explains a higher variance (15.8 - 19.3%) in the data with respect to the niche (5.0 - 5.7%). However, these two factors are linked (one influences the other) as shown by the important variance explained by the market:niche interaction (7.9 - 10.2%), depriving our ability to assess their importance as main factors.

These site-specific variations can be explained by the existence of market-specific microbiomes, similar to that which has been shown for cheese-making plants (Bokulich & Mills, 2013), and general environmental variations due to geographic distance, as shows for bacterial communities in fermented meat products (Van Reckem et al., 2019). Yet, cheeses made in geographically distant parts of the world can have strikingly similar rind communities if similar environmental conditions are maintained (Wolfe et al., 2014).

In Mao tofu, we believe that the inside niche represents a more selective compartment, and stable environment for microbes to survive (e.g. in terms of oxygen, pH, competition) compared to the outside, which is subjected to environmental fluxes and random dispersal from other sources (e.g. market tables, air, humans). We found a lower number of reads in the samples from the inside niche compared to the outside, reflecting a lower amount of microbial DNA template, rather than biased sequencing results. This pattern was consistent across all three of the investigated molecular markers. Consistent community changes between bacteria and fungi, which implies

similar community-structuring factors, are supported by Procrustes analyses. In particular, samples from the internal niche had a significantly better fit in the Procrustes rotation, meaning that both bacteria and fungi were exposed to similar environmental conditions in the internal niche, regardless of the market, and compose the core microbiome of Mao tofu. This can be considered as additional evidence that the internal niche is a more confined environment, less subjected to random changes and contamination. Again, it is possible that the microbial communities of the inside niches are strictly linked to the production processes and microbiota living on tofu, tools and surfaces (Bokulich & Mills, 2013). In this way, microbes may be adapted, selected, and domesticated in this niche, similar to that which has happened in fermentation of cheese and wine (Almeida et al., 2015; Dumas et al., 2020).

It has been shown before that microbiome compositions can be used to predict the geographical origin of grapes (Mezzasalma et al., 2018), to distinguish soybean under organic, no-till, and conventional management (Longley et al., 2020), or even identification of human body niches and disease states (Statnikov et al., 2013). Our random forest models showed high accuracy in the classification of samples belonging to different markets (1.39 - 4.17% OOB error estimate) and niches (8.33 - 12.5%), demonstrating microbiomes have utility for determining provincial origin of fermented foods. Several taxa that showed the highest mean decrease in accuracy (top important OTUs for classification) were also group indicators (OTUs highly correlated with samples groups).

5. Conclusions

We found that Mao tofu diversity of both fungal and bacterial communities varied across geographical gradients and niches, with strong and significant interaction between the two factors. We showed that fungal and bacterial communities undergo similar environmental pressure especially in the inside niche. We found it noteworthy that several taxa abundant in Mao tofu overlap with those of other fermented food, cheese in particular. ITS, LSU and 16S microbial community profiles and machine learning models were used to accurately predict the market of origin, and whether samples were from the inside or the outside niche. Finally, our data demonstrate the presence of diverse non-target eukaryotes, further illuminating the complex microbiology of fermented foods. Similar to the cheese microbiome, we suspect most of fungi and bacteria comprising Mao tofu can be isolated. Culture studies, coupled with RNA and genome sequencing may help to disentangle the community ecology of fermented tofu as well as inform on microbial functions and interactions.

References

Abarenkov, K., Zirk, A., Piirmann, T., Pöhönen, R., Ivanov, F., Nilsson, R. H., & Kõljalg, U. (2020). *UNITE Community. UNITE general FASTA release for eukaryotes. Version 04.02.2020* [Data set]. doi: 10.15156/BIO/786370

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., … Sampaio, J. P. (2015). A population genomics insight into the Mediterranean origins of wine yeast domestication. *Molecular Ecology*, *24*(21), 5412–5427. doi: 10.1111/mec.13341

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Anderson, M. J., Ellingsen, K. E., & McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters*, *9*(6), 683–693. doi: 10.1111/j.1461-0248.2006.00926.x

Benucci, G. M. N., Longley, R., Zhang, P., Zhao, Q., Bonito, G., & Yu, F. (2019). Microbial communities associated with the black morel cultivated in greenhouses. *PeerJ*, *7*, e7744. doi: 10.7717/peerj.7744

Benucci, G. M. N., Rennick, B., & Bonito, G. (2020). Patient propagules: Do soil archives preserve the legacy of fungal and prokaryotic communities? *PloS One*, *15*(8), e0237368. doi: 10.1371/journal.pone.0237368

Bokulich, N. A., & Mills, D. A. (2013). Facility-Specific “House” Microbiome Drives Microbial Landscapes of Artisan Cheesemaking Plants. *Applied and Environmental Microbiology*, *79*(17), 5214–5223. doi: 10.1128/aem.00934-13

Boutrou, R., & Guéguen, M. (2005). Interests in Geotrichum candidum for cheese technology. *International Journal of Food Microbiology*, *102*(1), 1–20. doi: 10.1016/j.ijfoodmicro.2004.12.028

Büchl, N. R., & Seiler, H. (2011). *YEASTS AND MOLDS | Yeasts in Milk and Dairy Products* (pp. 744–753). Elsevier Ltd. doi: 10.1016/b978-0-12-374407-4.00498-2

Caporaso, G. J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. doi: 10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., … Knight, R. (2010). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, *108*(Supplement_1), 4516–4522. doi: 10.1073/pnas.1000080107

Carini, P., Marsden, P. J., Leff, J. W., Morgan, E. E., Strickland, M. S., & Fierer, N. (2016). Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature Microbiology*, *2*, 16242. doi: 10.1038/nmicrobiol.2016.242

Chao, S.-H., Tomii, Y., Sasamoto, M., Fujimoto, J., Tsai, Y.-C., & Watanabe, K. (2008). Lactobacillus capillatus sp. nov., a motile bacterium isolated from stinky tofu brine. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, *58*(11), 2555–2559. doi: 10.1099/ijs.0.65834-0

Chao, S.-H., Tomii, Y., Watanabe, K., & Tsai, Y.-C. (2008). Diversity of lactic acid bacteria in fermented brines used to make stinky tofu. *International Journal of Food Microbiology*, *123*(1-2), 134–141. doi: 10.1016/j.ijfoodmicro.2007.12.010

Chen, W., Gray, L. E., Kurle, J. E., & Grau, C. R. (1999). Specific detection ofPhialophora gregataandPlectosporium tabacinumin infected soybean plants using

polymerase chain reaction. *Molecular Ecology*, Vol. 8, pp. 871–877. doi: 10.1046/j.1365-294x.1999.00645.x
Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. doi: 10.1093/nar/gkv1276
Combina, M., Elía, A., Mercado, L., Catania, C., Ganga, A., & Martinez, C. (2005). Dynamics of indigenous yeast populations during spontaneous fermentation of wines from Mendoza, Argentina. *International Journal of Food Microbiology*, *99*(3), 237–243. doi: 10.1016/j.ijfoodmicro.2004.08.017
Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. doi: 10.1186/s40168-018-0605-2
De Cáceres, M., Legendre, P., & Moretti, M. (2010). Improving indicator species analysis by combining groups of sites. *Oikos* , *119*(10), 1674–1684. doi: 10.1111/j.1600-0706.2010.18334.x
Delavenne, E., Mounier, J., Asmani, K., Jany, J.-L., Barbier, G., & Le Blay, G. (2011). Fungal diversity in cow, goat and ewe milk. *International Journal of Food Microbiology*, *151*(2), 247–251. doi: 10.1016/j.ijfoodmicro.2011.08.029
Dimidi, E., Cox, S. R., Rossi, M., & Whelan, K. (2019). Fermented Foods: Definitions and Characteristics, Impact on the Gut Microbiota and Effects on Gastrointestinal Health and Disease. *Nutrients*, *11*(8). doi: 10.3390/nu11081806
Dumas, E., Feurtey, A., Rodríguez de la Vega, R. C., Le Prieur, S., Snirc, A., Coton, M., … Giraud, T. (2020). Independent domestication events in the blue-cheese fungus Penicillium roqueforti. *Molecular Ecology*, *29*(14), 2639–2660. doi: 10.1111/mec.15359
Edgar, R. (2016). *UCHIME2: improved chimera prediction for amplicon sequencing*. doi: 10.1101/074252
Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. doi: 10.1038/nmeth.2604
Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* , *31*(21), 3476–3482. doi: 10.1093/bioinformatics/btv401
Fei, Y., Li, L., Chen, L., Zheng, Y., & Yu, B. (2018). High-throughput sequencing and culture-based approaches to analyze microbial diversity associated with chemical changes in naturally fermented tofu whey, a traditional Chinese tofu-coagulant. *Food Microbiology*, *76*, 69–77. doi: 10.1016/j.fm.2018.04.004
Foster, Z. S. L., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS Computational Biology*, *13*(2), e1005404. doi: 10.1371/journal.pcbi.1005404
Gardes, M., & Bruns, T. D. (1993). ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. *Molecular Ecology*, *2*(2), 113–118. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8180733
Gdanetz, K., Benucci, G. M. N., Vande Pol, N., & Bonito, G. (2017). CONSTAX: a tool for improved taxonomic resolution of environmental fungal ITS sequences. *BMC Bioinformatics*, *18*(1), 538. doi: 10.1186/s12859-017-1952-x

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51. doi: 10.1007/bf02291478

Gu, J., Liu, T., Sadiq, F. A., Yang, H., Yuan, L., Zhang, G., & He, G. (2018). Biogenic amines content and assessment of bacterial and fungal diversity in stinky tofu – A traditional fermented soy curd. *LWT*, *88*, 26–34. doi: 10.1016/j.lwt.2017.08.085

He, G.-Q., Liu, T.-J., Sadiq, F. A., Gu, J.-S., & Zhang, G.-H. (2017). Insights into the microbial diversity and community dynamics of Chinese traditional fermented foods from using high-throughput sequencing approaches. *Journal of Zhejiang University. Science. B*, *18*(4), 289–302. doi: 10.1631/jzus.B1600148

Hopple, J. S., & Vilgalys, R. (1994). Phylogenetic Relationships among Coprinoid Taxa and Allies Based on Data from Restriction Site Mapping of Nuclear rDNA. *Mycologia*, Vol. 86, p. 96. doi: 10.2307/3760723

Jackson, D. A. (1995). PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Écoscience*, *2*(3), 297–303. doi: 10.1080/11956860.1995.11682297

Jang, C. H., Park, C. S., Lim, J. K., Kim, J. H., Kwon, D. Y., & Kim, Y. S. (2008). Metabolism of isoflavone derivatives during manufacturing of traditional Meju and Doenjang. *Food Science and Biotechnology*.

Jayachandran, M., & Xu, B. (2019). An insight into the health benefits of fermented soy products. *Food Chemistry*, *271*, 362–371. doi: 10.1016/j.foodchem.2018.07.158

Kang, S. J., Seo, J. Y., Cho, K. M., Lee, C. K., Kim, J. H., & Kim, J.-S. (2016). Antioxidant and Neuroprotective Effects of Doenjang Prepared with Rhizopus, Pichia, and Bacillus. *Preventive Nutrition and Food Science*, *21*(3), 221–226. doi: 10.3746/pnf.2016.21.3.221

Kassambara, A. (2020). ggpubr: “ggplot2” Based Publication Ready Plots. R package version 0.4.0. Retrieved from https://CRAN.R-project.org/package=ggpubr

Kim, D.-J., Jeong, Y.-J., Kwon, J.-H., Moon, K.-D., Kim, H.-J., Jeon, S.-M., … Choi, M.-S. (2008). Beneficial effect of chungkukjang on regulating blood glucose and pancreatic beta-cell functions in C75BL/KsJ-db/db mice. *Journal of Medicinal Food*, *11*(2), 215–223. doi: 10.1089/jmf.2007.560

Lee, S.-M., Kim, Y., Choi, H. J., Choi, J., Yi, Y., & Yoon, S. (2013). Soy milk suppresses cholesterol-induced inflammatory gene expression and improves the fatty acid profile in the skin of SD rats. *Biochemical and Biophysical Research Communications*, *430*(1), 202–207. doi: 10.1016/j.bbrc.2012.10.074

Leff, W. J. (2017). mctoolsr: Microbial Community Data Analysis Tools. R package version 0.1.1.2. doi: 10.1556/168.2019.20.3.3

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database issue), D19–D21. doi: 10.1093/nar/gkq1019

Liang, H., Deng, L., & Lin, H. (2013). Distribution, functions and applications of lactic acid bacteria in traditional fermented soybean foods. *Food Science*.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from https://www.researchgate.net/profile/Andy_Liaw/publication/228451484_Classification_and_Regression_by_RandomForest/links/53fb24cc0cf20a45497047ab/Classification-and-Regression-by-RandomForest.pdf

Longley, R., Benucci, G. M. N., Mills, G., & Bonito, G. (2019). Fungal and bacterial community dynamics in substrates during the cultivation of morels (Morchella rufobrunnea) indoors. *FEMS Microbiology Letters*, *366*(17). doi: 10.1093/femsle/fnz215

Longley, R., Noel, Z. A., Benucci, G. M. N., Chilvers, M. I., Trail, F., & Bonito, G. (2020). Crop Management Impacts the Soybean (Glycine max) Microbiome. *Frontiers in Microbiology*, *11*. doi: 10.3389/fmicb.2020.01116

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, *17*(1), 10. doi: 10.14806/ej.17.1.200

McMurdie, P. J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, *10*(4), e1003531. doi: 10.1371/journal.pcbi.1003531

Mezzasalma, V., Sandionigi, A., Guzzetti, L., Galimberti, A., Grando, M. S., Tardaguila, J., & Labra, M. (2018). Geographical and Cultivar Features Differentiate Grape Microbiota in Northern Italy and Spain Vineyards. *Frontiers in Microbiology*, *9*, 946. doi: 10.3389/fmicb.2018.00946

Michel, M., Kopecká, J., Meier-Dörnberg, T., Zarnkow, M., Jacob, F., & Hutzler, M. (2016). Screening for new brewing yeasts in the non-Saccharomyces sector with Torulaspora delbrueckii as model. *Yeast* , *33*(4), 129–144. doi: 10.1002/yea.3146

Morel, G., Sterck, L., Swennen, D., Marcet-Houben, M., Onesime, D., Levasseur, A., … Casaregola, S. (2015). Differential gene retention as an evolutionary mechanism to generate biodiversity and adaptation in yeasts. *Scientific Reports*, *5*, 11571. doi: 10.1038/srep11571

Murphy, M. A., Evans, J. S., & Storfer, A. (2010). Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology*, *91*(1), 252–261. doi: 10.1890/08-0879.1

Noel, Z. A., Chang, H.-X., & Chilvers, M. I. (2020). Variation in soybean rhizosphere oomycete communities from Michigan fields with contrasting disease pressures. *Applied Soil Ecology*, Vol. 150, p. 103435. doi: 10.1016/j.apsoil.2019.103435

O'Brien, C. E., McCarthy, C. G. P., Walshe, A. E., Shaw, D. R., Sumski, D. A., Krassowski, T., … Butler, G. (2018). Genome analysis of the yeast Diutina catenulata, a member of the Debaryomycetaceae/Metschnikowiaceae (CTG-Ser) clade. *PloS One*, *13*(6), e0198957. doi: 10.1371/journal.pone.0198957

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2019). *vegan: Community Ecology Package, R package version 2.5-6*. Retrieved from https://CRAN.R-project.org/package=vegan

Padilla, B., Gil, J. V., & Manzanares, P. (2016). Past and Future of Non-Saccharomyces Yeasts: From Spoilage Microorganisms to Biotechnological Tools for Improving Wine Aroma Complexity. *Frontiers in Microbiology*, *7*, 411. doi: 10.3389/fmicb.2016.00411

Pisani, P., Parkin, D. M., Bray, F., & Ferlay, J. (1999). Estimates of the worldwide mortality from 25 cancers in 1990. *International Journal of Cancer. Journal International Du Cancer*, *83*(1), 18–29. doi: 3.0.co;2-m">10.1002/(sici)1097-0215(19990924)83:1<18::aid-ijc5>3.0.co;2-m

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from

https://www.R-project.org/.
Ren N. N., Chen H. J., Li Y., Mcgowan G. W., & Lin Y. G. (2017). [A clinical study on the effect of nattokinase on carotid artery atherosclerosis and hyperlipidaemia]. *Zhonghua yi xue za zhi*, *97*(26), 2038–2042. doi: 10.3760/cma.j.issn.0376-2491.2017.26.005
Reynolds, N. K., Benny, G. L., Ho, H.-M., Hou, Y.-H., Crous, P. W., & Smith, M. E. (2019). Phylogenetic and morphological analyses of the mycoparasitic genus Piptocephalis. *Mycologia*, *111*(1), 54–68. doi: 10.1080/00275514.2018.1538439
Statnikov, A., Henaff, M., Narendra, V., Konganti, K., Li, Z., Yang, L., … Alekseyenko, A. V. (2013). A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, *1*(1), 11. doi: 10.1186/2049-2618-1-11
Sun, G. P., Zhang, X. J., Wang, Y., Wang, D., & Xie, J. L. (2010). The investigation of bacteria diversity in stinky tofu brine. *Xiandai Shipin Keji = Modern Food Science and Technology*.
Tamang, J. P., Shin, D.-H., Jung, S.-J., & Chae, S.-W. (2016). Functional Properties of Microorganisms in Fermented Foods. *Frontiers in Microbiology*, *7: 578*. doi: 10.3389/fmicb.2016.00578
Tedersoo, L., & Lindahl, B. (2016). Fungal identification biases in microbiome projects. *Environmental Microbiology Reports*, *8*(5), 774–779. doi: 10.1111/1758-2229.12438
Tsai, J. S., Lin, Y. S., Pan, B. S., & Chen, T. J. (2006). Antihypertensive peptides and γ-aminobutyric acid from prozyme 6 facilitated lactic acid bacteria fermentation of soymilk. *Process Biochemistry*, *41*(6), 1282–1288. doi: 10.1016/j.procbio.2005.12.026
van Breda, V., Jolly, N., & van Wyk, J. (2013). Characterisation of commercial and natural Torulaspora delbrueckii wine yeast strains. *International Journal of Food Microbiology*, *163*(2-3), 80–88. doi: 10.1016/j.ijfoodmicro.2013.02.011
Van Reckem, E., Geeraerts, W., Charmpi, C., Van der Veken, D., De Vuyst, L., & Leroy, F. (2019). Exploring the Link Between the Geographical Origin of European Fermented Foods and the Diversity of Their Bacterial Communities: The Case of Fermented Meats. *Frontiers in Microbiology*, *10*, 2302. doi: 10.3389/fmicb.2019.02302
Wang, Q., Garrity, G. M., Tiedje, J. M., & 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. doi: 10.1128/AEM.00062-07
Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., … Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. doi: 10.1186/s40168-017-0237-y
White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). AMPLIFICATION AND DIRECT SEQUENCING OF FUNGAL RIBOSOMAL RNA GENES FOR PHYLOGENETICS. In *PCR Protocols* (pp. 315–322). doi: 10.1016/b978-0-12-372180-8.50042-1
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org
Wolfe, B. E., Button, J. E., Santarelli, M., & Dutton, R. J. (2014). Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. *Cell*, *158*(2), 422–433. doi: 10.1016/j.cell.2014.05.041

Wrather, A., Shannon, G., Balardin, R., Carregal, L., Escobar, R., Gupta, G. K., … Tenuta, A. (2010). Effect of Diseases on Soybean Yield in the Top Eight Producing Countries in 2006. *Plant Health Progress*, *11*(1), 29. doi: 10.1094/php-2010-0102-01-rs

Xu, L., Cai, W. X., & Xu, B. J. (2017). A Systematic Assesment on Vitamins (B2, B12) and GABA Profiles in Fermented Soy Products Marketed in China. *Journal of Food Processing and Preservation*, Vol. 41, p. e13126. doi: 10.1111/jfpp.13126

Yan, S., Liu, H., Zhang, J., & Tong, Q. (2020). Lactobacillus delbrueckii is the key functional microorganism of natural fermented tofu sour water involved in the traditional coagulation of Chinese Huizhou Mao-tofu. *LWT*, *131*, 109706. doi: 10.1016/j.lwt.2020.109706

Yoon, G.-A., & Park, S. (2014). Antioxidant action of soy isoflavones on oxidative stress and antioxidant enzyme activities in exercised rats. *Nutrition Research and Practice*, *8*(6), 618–624. doi: 10.4162/nrp.2014.8.6.618

Yu, Z., Hu, H., & Li, L. (2012). Analysis of microbial flora in the steep juice of Zhejiang shaoxing stinky tofu. *Science and Technology of Food Industry*.

Zhang, N., & Zhao, X.-H. (2010). Study of Mucor spp. in semi-hard cheese ripening. *Journal of Food Science and Technology*, *47*(6), 613–619. doi: 10.1007/s13197-010-0108-z

Zhao, X., & Zheng, X. (2009). A primary study on texture modification and proteolysis of mao-tofu during fermentation. *African Journal of Biotechnology*, *8*(10), 2294–2300. Retrieved from http://www.academicjournals.org/AJB

Zhu, Y., Wang, A., Liu, M. C., Zwart, A., Lee, R. Y., Gallagher, A., … Clarke, R. (2006). Estrogen receptor alpha positive breast tumors and breast cancer cell lines share similarities in their transcriptome data structures. *International Journal of Oncology*, *29*(6), 1581–1589. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17089000

Zimmermann, G. (1993). The entomopathogenic fungusMetarhizium anisopliaeand its potential as a biocontrol agent. *Pesticide Science*, *37*(4), 375–379. doi: 10.1002/ps.2780370410

Data Availability

Raw sequence reads have been deposited to the Sequence Read Archive (Leinonen, Sugawara, Shumway, & International Nucleotide Sequence Database Collaboration, 2011) with links to BioProject accession number PRJNA661071 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Supporting information

Fig. S1 - 5 (File S1) , R scripts (File S2), ITS, LSU, and 16S otu_table.txt files with taxonomic classifications (S3, S4 and S5 Files), metadata file (S6 File), ITS, LSU, and 16S OTU representative sequences (S7, S8, and S9 Files), are also provided as supporting information and available at https://github.com/Gian77/Scientific-Papers-R-Code.

Acknowledgments

We are grateful to Hong Ling for part of the sample collecting.

Author Contributions

Gian Maria Niccolò Benucci: Methodology, Software, Formal analysis, Validation, Data curation, Supervision, Writing- Original draft preparation, Writing- Reviewing and Editing; **XinXin Wang:** Methodology, Software, Data curation, Resources, Writing- Original draft preparation, Investigation, Writing- Reviewing and Editing; **Li Zhang:** Sampling, Microscopy, Photography, Wet lab; **Gregory Bonito:** Conceptualization, Methodology, Validation, Supervision, Project administration, Funding acquisition, Investigation, Writing- Original draft preparation, Writing- Reviewing and Editing **Fuqiang Yu:** Conceptualization, Supervision, Project administration, Funding acquisition, Reviewing and Editing

Funding

We acknowledge financial support of NSF DEB 1737898 and USDA MICL02416 to GB and CAS science and technology supported poverty alleviation project (KFJ-FP-201905) awarded to Yu Fuqiang. Xinxin Wang was supported by a China Scholarship Council.

Table Captions

Table 1. Results from permutational multivariate analysis of variance (PERMANOVA) of ITS, LSU and 16S DNA marker datasets for Market (Ciba, Longjin, Wanyao, New District) and Tofu Niche (inside and outside) as well as their interaction (Market:Niche) are shown. All p-values were adjusted using the Bonferroni method.

Factor	ITS				LSU				16S			
	Df	**F-value**	**R2**	**P-value**	**Df**	**F-value**	**R2**	**P-value**	**Df**	**F-value**	**R2**	**P-value**
Market	3	6.070	0.193	0.0003	3	4.891	0.158	0.0003	3	5.423	0.170	0.0003
Niche	1	4.708	0.050	0.0003	1	5.244	0.057	0.0003	1	5.420	0.057	0.0003
Market:Niche	3	2.502	0.079	0.0003	3	2.934	0.095	0.0003	3	3.260	0.102	0.0003
Residuals	64		0.678		62		0.690		64		0.670	
Total	71		1.000		69		1.000		71		1.000	

Figure Captions

Fig 1. Boxplot of samples distribution of observed species richness. Red diamonds represent the mean of the distribution. Letters, when present, represent pairwise Wilcoxon tests among groups after Kruskall-Wallis test ($p \leq 0.05$ after Bonferroni adjustment).

Fig 2. Boxplot of samples distribution of Shannon equitability (diversity) index. Red diamonds represent the mean of the distribution. Letters, when present, represent pairwise Wilcoxon tests among groups after Kruskall-Wallis test ($p \leq 0.05$ after Bonferroni adjustment).

Fig 3. Principal coordinate Analysis (PCoA) ordinations and sample distance form group centroids (dispersion) distributions. Model significance was tested using ANOVA-like permutation test (permutations= 9999) and p-values adjusted using the Bonferroni method. Letters, when present, represent pairwise permutation tests.

Fig 4. Procrustes plots for A) ITS and LSU rDNA OTU ordinations, B) ITS and 16S rDNA OTU ordinations, and C) LSU and 16S rDNA OTU ordinations. Each sample is represented by two points, connected by an arrow; the arrow starts at the target community and points toward the rotated community. m^2 represents the Procrustes Sum of Squares and r the Correlation in a symmetric Procrustes rotation ($r = \sqrt{(1 - m^2)}$). Tests were run using 9999 permutations. Distribution of Procrustes residual is shown in D, E and F. Letters, when present, represent pairwise Wilcoxon tests among groups after Kruskall-Wallis test ($p \leq 0.05$ after Bonferroni adjustment).

Fig 5. Metric Multidimensional Scaling (MDS) ordinations of Random Forest models 1-proximity matrix to visualize accuracy of market and niche sample classification in the A) ITS dataset, B) LSU dataset, and C) 16S dataset. Samples that cluster within the wrong groups may represent misclassifications.

Fig 6. Top 20 OTUs with the highest mean decrease accuracy obtained in the Random Forest models for the A) ITS, the B) LSU, and C) 16S rDNA marker datasets. Significant (with $p \leq 0.05$ after fdr correction) correlations (i.e. r.g. value) to samples of different markets or niches is reported within each bar, while colors specify the samples groups OTUs are correlated to.

Fig 7. Heatmap abundance trees to visualize tofu microbial composition according the A) ITS, the B) LSU, and C) 16S metagenomic marker datasets. The plot shows i) the number of “Inside” samples that have counts (i.e. sequence reads) for each taxon as the color of each taxon. Core taxa have darker colors while grey represents taxa that were absent in the “Inside” samples. ii) the number of OTUs assigned to each taxon in the overall dataset as node size.

Supplementary Files

Suppl. File 1 Supplementary figures 1-5.

Suppl. File 2 R scripts for the analyses.

Suppl. File 3 ITS otu table with taxonomy.

Supple File 4 LSU otu_table with taxonomy.

Supple File 5 16S otu table with taxonomy.

Suppl File 6 metadata file.

Suppl. File 7 ITS OTUs representative sequences.

Suppl. File 8 LSU OTUs representative sequences.

Suppl. File 9 16S OTUs representative sequences.