Comparison of metabarcoding taxonomic markers to describe fungal communities in fermented foods

Olivier Rué[1,2], Monika Coton[3], Eric Dugat-Bony[4], Kate Howell[5], Françoise Irlinger[4], Jean-Luc Legras[6], Valentin Loux[1,2], Elisa Michel[6], Jérôme Mounier[3], Cécile Neuvéglise[6], Delphine Sicard[6]*

[1] Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

[2] Université Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, 78350, Jouy-en-Josas, France

[3] Univ. Brest, INRAE, Laboratoire Universitaire de Biodiversité et Ecologie Microbienne, F-29280 Plouzané, France

[4] Université Paris Saclay, INRAE, AgroParisTech, UMR SayFood, 91120 Palaiseau, France

[5] School of Agriculture and Food, Faculty of Veterinary & Agricultural Sciences, The University of Melbourne, Parkville, Victoria, Australia

[6] SPO, Univ Montpellier, INRAE, Institut Agro, Montpellier, France

*Corresponding author

Correspondence: delphine.sicard@inrae.fr

Abstract

Next generation sequencing offers several ways to study microbial communities. For agri-food sciences, identifying species in diverse food ecosystems is key for both food sustainability and food security. The aim of this study was to compare metabarcoding pipelines and markers to determine fungal diversity in food ecosystems, from Illumina short reads. We built mock communities combining the most representative fungal species in fermented meat, cheese, wine and bread. Four barcodes (ITS1, ITS2, D1/D2 and RPB2) were tested for each mock and on real fermented products. We created a database, including all mock species sequences for each

barcode to compensate for the lack of curated data in available databases. Four bioinformatics tools (DADA2, QIIME, FROGS and a combination of DADA2 and FROGS) were compared. Our results clearly showed that the combined DADA2 and FROGS tool gave the most accurate results. Most mock community species were not identified by the RPB2 barcode due to unsuccessful barcode amplification. When comparing the three rDNA markers, ITS markers performed better than D1D2, as they are better represented in public databases and have better specificity to distinguish species. Between ITS1 and ITS2, differences in the best marker were observed according to the studied ecosystem. While ITS2 is best suited to characterize cheese, wine and fermented meat communities, ITS1 performs better for sourdough bread communities. Our results also emphasized the need for a dedicated database and enriched fungal-specific public databases with novel barcode sequences for 118 major species in food ecosystems.

Keywords: fermented foods, metabarcoding, fungi, mock communities, barcode marker comparison

Introduction

In the field of microbial ecology, amplicon-based metagenomic analysis (also named metabarcoding) is one of the most popular routes to describe microbial communities as it is a high throughput method with relatively low-cost nowadays. This approach relies on amplifying a phylogenetic biomarker from total community DNA purified from the samples to be characterized, followed by massive amplicon sequencing (Shokralla et al., 2012), usually with Illumina MiSeq technology (Caporaso et al., 2012). For bacteria, the SSU rRNA gene is commonly accepted as the most suitable biomarker for metabarcoding studies although different variable regions can be targeted (Zhang et al., 2018). For fungal communities, there is still no international consensus regarding the choice of the best phylogenetic biomarker for such an approach.

Ten years ago, the Fungal Barcoding Consortium recommended the use of the Internal Transcribed Spacer (ITS) region as the primary marker for fungal identifications due to superior species-level resolution compared to LSU and SSU rRNA genes, and higher amplification success compared to protein coding genes (Schoch et al., 2012). However, amplicon-based metagenetic analysis using MiSeq sequencing imposes more constraints than classical species identification. Most importantly, the typical amplicon size (~500bp) will not provide complete ITS region sequences or full length rRNA or protein coding genes. Consequently, the most popular barcodes currently used for fungal community analysis by metabarcoding are ITS1, ITS2 and the LSU D1/D2 domain (R. Henrik Nilsson et al., 2019).

Amplicon-based metagenetic analyses using these markers has been largely facilitated over the past few years by the availability of a broad range of high-quality reference sequences in public databases from different sequencing initiatives of collection strains (Vu et al., 2016).

However, the multicopy nature of the rDNA operon negatively affects fungal community compositions detected in complex samples by metabarcoding (Lavrinienko et al., 2021). In addition, the variable length of the entire ITS region across fungi represents an additional bias resulting in lower representation of species with longer amplicons in the datasets (Ihrmark et al., 2012). Nevertheless, using only ITS1 or ITS2, which produce shorter amplicons, was shown to represent the quantitative composition of the sample (Ihrmark et al., 2012). Several very promising single-copy marker genes were thus proposed to overcome these limitations, including the rpb2 gene, encoding for the second largest ribosomal polymerase II subunit (Větrovský et al., 2016). In addition to the above-mentioned specificity, the species-resolving power of rpb2 was found to be higher than rDNA genes and ITS regions. This marker was also found to be particularly suited to study basal fungal lineages. Yet, due to the lack of universality and lower specificity of RPB2 primers (when compared to others) as well as the lower numbers of rpb2 gene sequences in public databases, further applications to study fungal communities from food ecosystems may be limited.

Some authors evaluated the reliability of different markers (ITS1-2, D1/D2 LSU and SSU) to describe fungal diversity by amplicon-based metagenomic analysis (De Filippis et al., 2017). A mock community, composed of 19 strains representative of common fungal species, as well as environmental samples including soil, human saliva, human feces and grape must were used. Although all markers were able to correctly detect the different species in the mock community, the results suggested that there was an important quantification bias when using ITS1-2. This could be due to the high heterogeneity in marker length across fungal species. However, marker performance is likely to be highly influenced by fungal species composition of the sample (e.g., composed mainly of Basidiomycota versus Ascomycota) which, in turn, depends on the studied environment.

Numerous tools are available for fungal metabarcoding data analyses from Illumina sequencing technology (R. Henrik Nilsson et al., 2019) but many differences between pipelines exist (Anslan et al., 2018) which highlights the need to pay attention to the choice of the tool. Indeed, amplicon length is one crucial characteristic to consider. Using an *in silico* approach for fungal sequences, a recent showed that the length of the extracted ITS1 portions from UNITE ranged from 9 bp to 1181 bp, with an average length of 177 bp while the extracted ITS2 portions ranged from 14 bp to 730 bp, with an average of 182 bp (Yang et al., 2018). For

D1/D2 and RPB2 amplicons, lengths are often above 600 bp; in this case, common strategies that merge paired-end reads are not suitable because Illumina sequencing, the most used technology in metabarcoding analyses, provides paired-end reads of maximum 2 x 300 bp.

Among bioinformatics solutions for fungal communities with short reads, some are dedicated to ITS such as PIPITS (Gweon et al., 2015) or DAnIEL (Loos et al., 2021) while others are more generic (Bernard et al., 2021; Bolyen et al., 2019; Callahan et al., 2016; Edgar, 2010; Escudié et al., 2018; Özkurt et al., 2022). However, one downfall is that few of them can process short and long amplicons simultaneously. PIPITS and DAnIEL reject non-overlapping reads, so long amplicons will be excluded. QIIME2 and DADA2 require a choice to be made between merging reads or keeping only R1 reads. USEARCH recommends taking into account merged sequences and 5' R1 reads of non-overlapping paired-end sequences. FROGS deals with mergeable reads and creates artificial sequences from non-mergeable reads, therefore all sequenced information is kept throughout the pipeline. It is worth mentioning that FROGS has shown better results than QIIME2, DADA2 and USEARCH on simulated data (Bernard et al., 2021).

Another important aspect in metabarcoding data analysis is the way to build correct biological sequences from reads. It can be under the form of Operational Taxonomic Units (OTUs), Amplicon Sequence Variants (ASVs) or zero-radius OTUs (ZOTUs) depending on tools. Numerous studies have compared the results from OTU-based and ASV-based approaches (Callahan et al., 2017). Using mock communities, ASV-based methods had higher sensitivity and detected bacterial strains present, sometimes at the expense of specificity (Caruso et al., 2019). However, a different study concluded that for broadscale (e.g., all bacteria or all fungi) α and β diversity analyses, ASV or OTU methods often provided similar ecological results (Glassman and Martiny, 2018). From a practical point of view, an important advantage of ASV-based approaches is the consistent labels with intrinsic biological meaning identified independently from a reference database. Thus, ASVs independently inferred from different studies and different samples (for the same targeted region) can be compared.

In this study, we compared the performance of four phylogenetic markers (ITS1, ITS2, D1/D2 LSU and RPB2) for metabarcoding analysis of complex fungal communities in different fermented foods, after assessing which bioinformatic strategy was most suitable for analyzing such datasets. To perform these analyses, we compared seven strategies based on commonly used tools for metabarcoding data (QIIME2, DADA2, USEARCH, FROGS) and a combination of DADA2 and FROGS (named DADA2_FROGS) by analyzing four separate mock communities' representative of the fungal diversity found in meat sausage (fermented meat), cheese, grape must (wine) and sourdough (bread), for a total of 118 species, as well as 24 real fermented food samples.

Methods

Mock community sample preparation

Species diversity

For each fermented food type, representative species were selected based on an inventory of the most frequently described species in the literature. One strain (mainly available type-strains) was included for each selected species. The complete list of strains used for mock community design is available on Recherche Data Gouv platform (https://doi.org/10.57745/AZNJFE) and a summary is presented in Table 1.

Table 1 - Number of Species, Genus and Family in each mock community

	Bread	Cheese	Meat	Wine

Species	27	25	40	60
Genus	15	16	14	37
Family	4	11	8	8

Figure 1 shows the species distribution at the genus level in the different mock community samples.

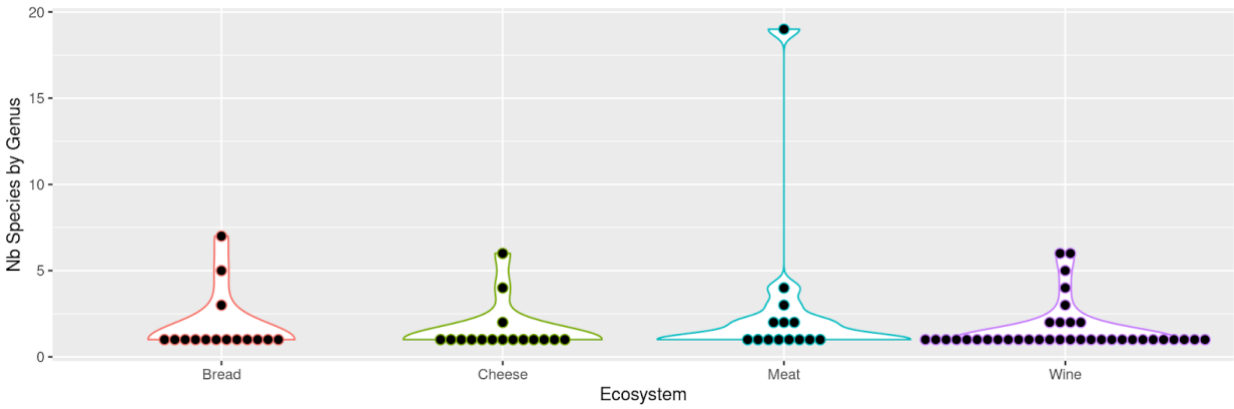

Figure 1 - Genus and species diversity in mock community samples. Each point corresponds to a genus. The number of species per genus is indicated on the vertical axis. For example, in meat, one genus is represented by 19 species in the mock community (upper point) and 8 genera are represented by a single species.

The sourdough bread mock community was composed of 27 strains, 26 belonged to Ascomycota phylum while one was a Basidiomycota. At the genus level, 7, 5 and three species belonged to *Kazachstania*, *Pichia* and *Candida* spp. respectively, while all others belonged to a distinct genus. The choice of these species was based on a review paper on sourdough yeasts (Von Gastrow et al., 2023).

The cheese mock community contained 25 strains including 19, 5 and one from the Ascomycota, Mucoromycota and Basidiomycota phylum, respectively. Six strains belonged to *Penicillium*, 4 to *Mucor* and 2 to *Clavispora* spp. while the others belonged to a distinct genus. This choice of fungal species was based on the reviews by Montel et al. 2014 and Irlinger et al., 2015.

Among the 40 strains composing the fermented meat mock community, 36 and 4 belonged to the Ascomycota and Basidiomycota phylum, respectively. Nineteen, 4, 3, 2, 2 and 2 strains were affiliated to *Penicillium*, *Yarrowia*, *Cladosporium*, *Aspergillus*, *Rhododotorula* and *Candida*, respectively. All others belonged to a distinct genus. This choice was based on litterature data on fermented meats (Coton et al., 2021; Franciosa et al., 2021; Berni, 2014 and Selgas and Garcia, 2014).

For the wine mock community, it contained 60 strains belonging to Ascomycota (45 strains) and Basidiomycota (15 strains). Six strains were affiliated to *Hanseniaspora*, 6 to *Pichia*, 5 to *Candida*, 4 to *Rhodotorula*, 3 to *Papiliotrema* and 2 to *Clavispora*, *Cystobasidium*, *Metschnikowia* and *Meyerozyma*. All others belonged to a distinct genus. The selection was made from a review of papers investigating grape and wine microbiota (Setati et al., 2012; Rossouw and Bauer, 2016; Jolly et al., 2003; Setati et al., 2015; Bokulich et al., 2013 and Garofalo et al., 2016).

Twenty-seven of the 118 strains were common to at least 2 different mock communities (Bread, Cheese, Meat, Wine). One, i.e. *Torulaspora delbrueckii*, was present in all mocks.

DNA extraction from single strains

Bread and wine

Each strain was grown overnight at 25°C in 15 mL of YEPD before centrifuging for 10 minutes at 1,500 × g. The cell pellet was resuspended in one mL of sterile water and transferred to a 2 mL tube. After a second centrifugation at 11,800 × g for 2 minutes, the pellet was resuspended in the yeast cell lysis solution from the MasterPure Yeast DNA extraction kit (Epicentre) and DNA was extracted according to the kit procedure.

Cheese

Each strain was grown overnight at 25°C under agitation at 200 rpm in 10 mL of YEGC. One mL of the culture was centrifuged at 10,000 × g for 10 minutes and the cell pellet was used for DNA extraction using the FastDNA SPIN Kit (MP Biomedicals).

Fermented meat

DNA was extracted from scraped colonies for yeasts or mycelial plugs for molds using the FastDNA SPIN Kit (MP Biomedicals) according to the manufacturer's instructions.

Mock community design

For each food environment (bread, wine, cheese, fermented meat), two different mock communities were prepared, a “DNA” mock community and a “PCR” mock community. For the DNA mock community, genomic DNA from each strain was quantified using the Qubit DNA Broad Range assay (ThermoFisher Scientific), diluted to the same concentration (10 ng/µL) and pooled. Then, the four markers were amplified in separate reactions from 20 ng of pooled DNA. For the PCR mock community, the four markers were individually amplified from each strain using 20 ng of genomic DNA as input, and PCR products were quantified using the Qubit DNA Broad Range assay (ThermoFisher Scientific). Then, 300 ng of PCR-product from each strain were pooled and diluted to a final concentration of 10 ng/µL before metabarcoding analysis. All mock community samples were prepared in triplicate.

Real samples preparation

DNA extraction was performed for each food environment (bread, wine, cheese, fermented meat) according to different protocols adapted for each matrix. DNA concentration was determined using a Qubit fluorometer (Life Sciences) according to the Broad Range DNA assay kit protocol.

Bread

Three types of sourdough coming from different French bakeries were analyzed. All sourdoughs were made of wheat flour. Sourdough 1 was sampled from an artisanal bakery in Azillanet (Occitanie Region), sourdough 2 from a local baker in Assas (Occitanie region) and sourdough 3 from a local baker in Amilly (Centre-Val de Loire region). For each sourdough, DNA extraction was performed from 200 mg of three independent samples using the MO BIO’s Powersoil DNA isolation kit procedure (Qiagen 12888-100) as described previously (von Gastrow et al., 2022).

202 *Cheese*
203 Three ready-to-consume cheeses, namely Saint-Nectaire, Livarot and Epoisses, were analyzed. For each
204 cheese type, three independent cheeses from the same producer were purchased on the same date. Rind was
205 gently separated from the core using sterile knives, and only the rind fraction was analyzed. Rind samples were
206 diluted 1:10 (w/v) in sterile distilled water and homogenized with an Ultra Turrax® (Labortechnik) at 8,000 rpm
207 for 1 min to obtain a homogeneous mixture. DNA extraction was performed on 0.5 mL using the bead beating-
208 based protocol detailed in a previous study (Dugat-Bony et al., 2015).
209

210 *Fermented meat*
211 DNA extractions were performed on casing samples obtained from French fermented sausages as
212 described previously (Coton et al., 2021). Briefly, 5 cm × 1 cm casing samples were mixed with 9 mL sterile
213 Tween (0.01% v/v) water followed by vigorous vortexing, before removing the casings. After centrifugation at
214 8,000 × g for 15 min, cell pellets were stored at −20°C until use. For DNA extractions, slightly thawed cell pellets
215 were resuspended in 500 µL yeast lysis solution, divided in two, and DNA extracted using the FastDNA spin kit
216 (MP Biomedicals) as described by the manufacturer. After extraction, DNA samples were purified using the
217 DNeasy Tissue Kit silica-based columns (Qiagen) according to the manufacturer's instructions.
218

219 *Wine*
220 Cells from 1 liter of a grape must were collected after centrifugation for 15 min at 10,000 × g. The pellet
221 was resuspended in 10 mL YPD supplemented with 20% glycerol and stored at -80°C. Three samples of
222 Sauvignon (1) and Viognier (2) grape must from the INRAE experimental wineries in Gruissan (France) were
223 chosen. For DNA extraction, 1 mL of this cell suspension was sampled and centrifuged, then cells were
224 resuspended in 1 mL of freshly prepared PBS supplemented with 1% Polyvinylpyrrolidone 25 to remove
225 polyphenolic compounds that could further inhibit target amplification. After a second centrifugation at 15,000
226 × g for 10 minutes, DNA was extracted with the DNeasy Plant kit (QIAGEN, Hilden, Germany) with some
227 modifications. The pellet was resuspended in 0.5 mL AP1 buffer and 4 µL of RNAse A solution and 300 µL of
228 0.3 mm glass beads were added. Cells were disrupted in a Precellys grinder (6,000 rpm, 320 seconds - 3 times).
229 After centrifugation for 5 minutes at 15,000 × g, the supernatant was used for the downstream DNA extraction
230 steps according to the manufacturer's instructions. The resulting DNA samples were then used for
231 metabarcoding.
232

233 **Library preparation and sequencing**
234 Target markers were amplified with the primers presented in Table 2. The ITS1, ITS2, D1/D2 and RPB2
235 regions were amplified with the primers F (CTTTCCCTACACGACGCTCTTCCG-forward primer sequence) and R
236 (GGAGTTCAGACGTGTGCTCTTCCG-reverse primer sequence) using 30 amplification cycles with an annealing
237 temperature of 48 or 55°C (Table 2), 0.5 U MTP Taq (Sigma-Aldrich), 1.25 µL each primer (20 µM), 1 µL dNTP
238 (10 µM each) in 50 µL final volume.
239 Single multiplexing was performed using an in-house 6 bp index, which was added to the reverse primer
240 during a second PCR with 12 cycles using forward primer
241 (AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC) and reverse primer
242 (CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGT). The resulting PCR products were
243 purified and loaded onto the Illumina MiSeq cartridge according to the manufacturer instructions, and paired-
244 end read sequencing was performed for 2 × 250 cycles. The quality of the run was checked internally using
245 PhiX as a control, and then each paired-end sequence was assigned to its sample with the help of the previously

integrated index. The sequencing data from this study are available in NCBI SRA repository under the Bioproject number PRJNA685292.

Table 2 - Description of the primers used to target ITS1, ITS2, D1/D2 and RPB2 regions.

Targeted region	Fwd name	Fwd sequence	Rv name	Rv sequence	Tm	Reference
D1/D2	NL1	5'-GCATATCAATAAGCGGAGGAAAAG-3'	NL4	5'-GGTCCGTGTTTCAAGACGG-3'	48°C	(O'Donnell, 1993)
RPB2	RPB2-6F	5'-TGGGGYATGGTNTGYCCYGC-3'	RPB2-7R	5'-GAYTGRTTRTGRTCRGGGAAVGG-3'	55°C	(Matheny, 2005)
ITS1	ITS1F	5'-CTTGGTCATTTAGAGGAAGTAA-3'	ITS2	5'-GCTGCGTTCTTCATCGATGC-3'	55°C	(Gardes and Bruns, 1993; White et al., 1990)
ITS2	ITS3	5'-GCATCGATGAAGAACGCAGC-3'	ITS4Kyo	5'-TCCTCCGCTTWTTGWTWTGC-3'	55°C	(Toju et al., 2012; White et al., 1990)

Bioinformatic analysis

Construction of the reference databank for mock community analysis

For each reference strain (118 in total), the sequence of the four phylogenetic markers (ITS1, ITS2, D1/D2, RPB2) was obtained from public databases or from unpublished sequences obtained in our labs. Only 3 RPB2 sequences were not available and missing due to PCR amplification failure (*Cryptococcus neoformans*, *Mucor lanceolatus* and *Rhodotorula glutinis*). The length distribution of the 469 sequences is represented in Figure 2. As expected, variations in length are visible, ranging from 70 to 835 bp. On average, ITS1 sequences are shortest, followed by ITS2, D1/D2 and RPB2 sequences.

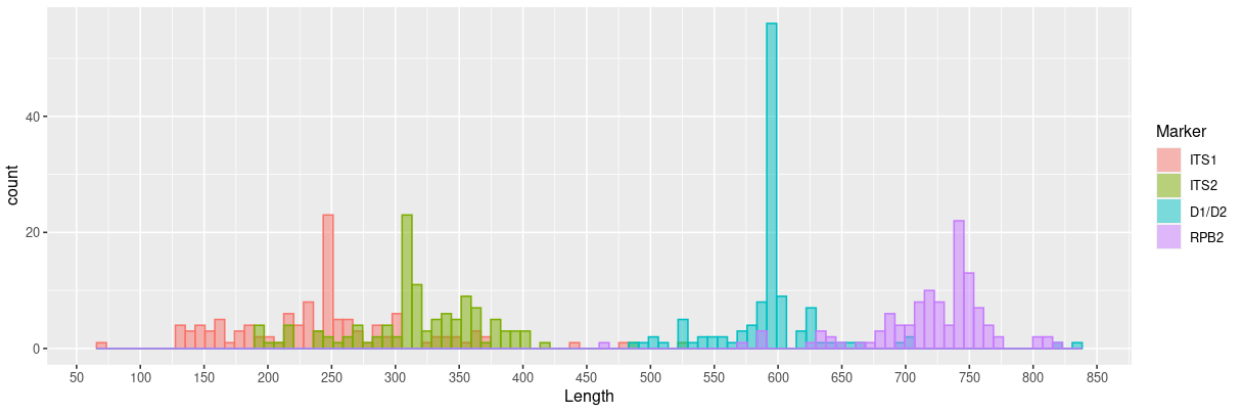

Figure 2 - Sequence length distribution of the four markers for the 118 strains present in mock samples inferred from the reference database..

From all sequences and for each marker, we built phylogenetic trees with FastTree (Price et al., 2010) and Phangorn R package (Schliep, 2011) after a multiple alignment of sequences with Mafft (Katoh et al., 2009).

Figure 3 shows the diversity of the 118 ITS1 sequences from a phylogenetic point of view. Phylogenetic trees for other markers are available on Recherche Data Gouv platform (https://doi.org/10.57745/AZNJFE).

Database construction was crucial for benchmarking and was necessary in our case because many sequences are absent in public databanks. Figure 4 shows the best hits of the 469 sequences after a blast against each representative databank (UNITE v9.0 for ITS, SILVA v138 for D1/D2 and nt release 2021-07-30 for RPB2). If the sequence is present in the databank, the corresponding point is at the top right of the graph. If no hit is found, the corresponding point is at the bottom left of the graph. In the middle of the graph are points corresponding to sequences for which the percentage of identity and coverage are less than one hundred percent. This result illustrates the incompleteness of public databases for food ecosystems, particularly for RPB2 sequences.

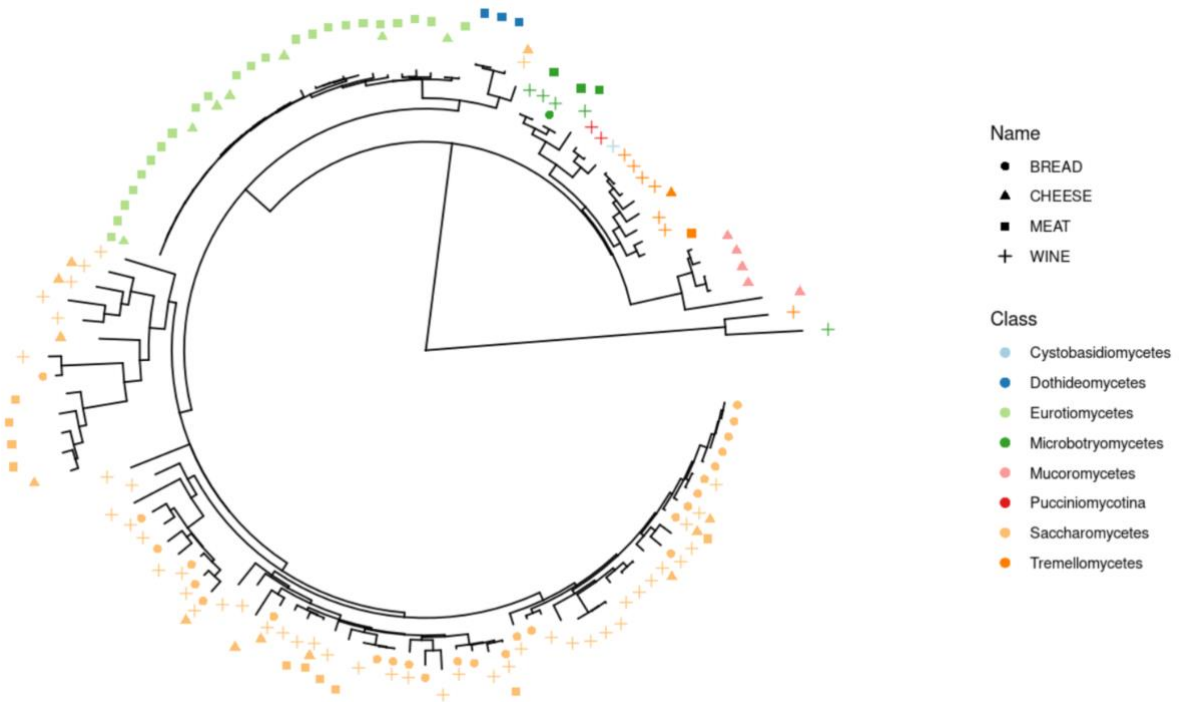

Figure 3 - Phylogenetic tree built from D1/D2 sequences of the 118 strains of the mock samples, colored by Class. Shapes indicate the presence in bread, cheese, fermented meat or wine mock samples.

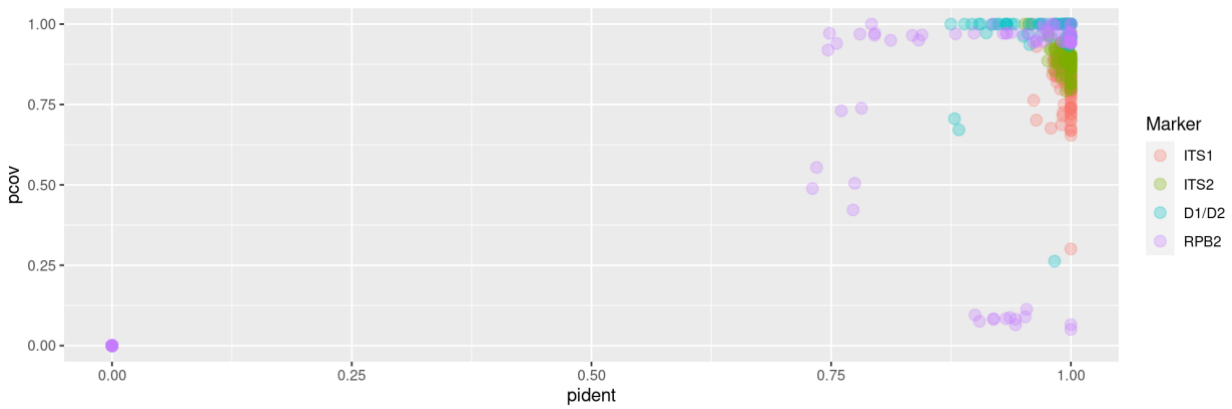

Figure 4 - Representation of blast results (identity and coverage percentage) of the 468 sequences against dedicated databases (UNITE for ITS, SILVA for D1/D2 and nt for RPB2). The percentage of coverage and percentage identity are shown on the Y and X axis, respectively.

Benchmark of metabarcoding approaches

Code and figures are available on Recherche Data Gouv platform (https://doi.org/10.57745/109NNP).

We used FROGS (Bernard et al., 2021), USEARCH (Edgar, 2010), QIIME2 (Bolyen et al., 2019) and DADA2 (Callahan et al., 2016) following their own guidelines, and a custom combination of DADA2 and FROGS that we named DADA2_FROGS.

One strength of FROGS is the ability to deal with overlapping and non-overlapping reads at the same time. This tool processes ITS1, ITS2, D1/D2 and RPB2 markers equally by a preprocess step to merge paired-end reads or create "artificial" sequences if they do not merge. In this case, R1 and R2 sequences are concatenated with a stretch of Ns in the middle.

DADA2 is a widely used tool for metabarcoding analyses. It infers exact amplicon sequence variants (ASVs) from amplicon data, resolving biological differences of even 1 or 2 nucleotides. By default, DADA2 does not deal with overlapping and non-overlapping reads at the same time. We used single-end and paired-end modes (DADA2-se and DADA2-pe, respectively). For DADA2-se, only R1 reads were used and only overlapping reads for DADA2-pe. In the same way, QIIME2 was used in single and paired-end modes (QIIME-pe and QIIME-se), following authors guidelines.

For USEARCH, we followed the instructions provided by the author on his website (https://www.drive5.com/usearch/manual/global_trimming_its.html) by taking into account merged sequences and 5' R1 reads of non-overlapping reads.

For the DADA2_FROGS strategy, the DADA2 recommendations were followed until obtaining the ASV table (cutadapt, filterAndTrim, dada, mergePairs and makeSequenceTable functions). At this step, we followed the FROGS guidelines after the clustering step: remove chimera, ITSx for ITS data and taxonomic affiliation. The aim was to benefit from the denoising algorithm that is, in theory, able to produce high-resolutive ASVs. As we wanted to keep merged and unmerged reads, we kept them by using the returnRejects parameter of the dada2 mergePairs function.

For each tool, we used our internal database available on Recherche Data Gouv platform (https://doi.org/10.57745/AZNJFE), consisting of our mock sequences, for taxonomic affiliation of the OTUs/ASVs/ZOTUs, as described above. No additional sequence was added to avoid unnecessary noise to analyze the different mock communities.

Different metrics were computed in order to compare the above-mentioned methods: (i) the divergence rate, computed as the Bray-Curtis distance between expected and observed abundance profiles at the species level; (ii) the number of false-negative taxa (FN) corresponding to the number of expected taxa that were not recovered by the method, (iii) the number of false positive taxa (FP) corresponding to the number of recovered taxa that were not expected, (iv) the number of true positive taxa (TP) corresponding to the number of recovered taxa that were expected. From these metrics we computed the precision (TP/(TP+FP)) and the recall rate (TP/(TP+FN)). Finally, as we knew the exact expected sequences, we computed the number of sequences perfectly identified (OTUs/ASVs/ZOTUs with nucleic sequence was strictly identical to the known reference sequence). For long sequences (i.e. > 500 bp), the middle was not sequenced and only the sequenced part was used for taxonomic affiliation. In this case, 100% identity between the reference and the OTU/ASV/zOTU resulted in a perfect identification.

Analysis of real samples

DADA2_FROGS, the bioinformatics approach with the best results from mock samples, was used to analyze real samples (Code and figures are available on Recherche Data Gouv platform, https://doi.org/10.57745/ENE09G). For the taxonomic affiliation of these samples (composition unknown), and for each marker, we added the 118 sequences from our mock communities to Unite (v. 9,0) (Rolf Henrik

Nilsson et al., 2019) for ITS data and SILVA (v. 138) (Quast et al., 2013) 28S rDNA sequences for D1/D2. For RPB2, we needed to build an in-house database because no dedicated one was publicly available to the best of our knowledge. We first extracted sequences from the "Fungi" division from NCBI nt databank (release 2021-07-30) (Sayers et al., 2022) using taxonkit (v. 0.6.0) (Shen and Ren, 2021) and then used cutadapt (Martin, 2011) with RPB2 primers to target sequences of interest. The databases were composed of 206,184 ITS sequences, 16,293 D1/D2 sequences and 13,055 RPB2 sequences.

For each ASV obtained, the taxonomic affiliation was manually checked and corrected when needed. More precisely, ASVs were blasted against different databases (e.g., NCBI, YEASTIP) to confirm or correct the affiliation, and we removed some ASVs (remaining chimera, contaminations). When taxonomic resolution at the species level was not possible (identical sequences between two or several species), we defined groups of species and labeled ASVs accordingly.

This manual curation step was performed for the most abundant ASVs (an abundance of at least 150 by marker and food ecosystem).

Results

In this study, we compare the efficiency of 4 barcodes (ITS1, ITS2, D1D2, RPB2) and seven bioinformatic workflows to detect the species in microbial community of 4 fermented products (bread, wine, cheese, fermented meat) using mocks and real samples. The phylogenetic diversity of fungal species analyzed is illustrated Figure 3.

Choice of the most accurate bioinformatic approach

Our benchmark of tools was only performed on mock community samples, and the results of the four markers were analyzed together (Figure 5).

Recall rate

The median recall rate (sensitivity), reflecting the capacity of tools to detect expected species, is between 0.75 and 0.8 for FROGS, DADA2_FROGS, USEARCH, QIIME-se and DADA2-se. It is lower (~0.5) for QIIME-pe and DADA2-pe, due to the fact that these methods always reject reads if they do not overlap (all D1/D2 and RPB2 sequences and some ITS1/2 were always missing).

Precision

Regarding precision, the four methods yield values of 0.95-0.97 (FROGS, DADA2_FROGS, USEARCH and QIIME-se). QIIME-pe is slightly less efficient (0.92), and both DADA2-se (0.84) and DADA2-pe (0.69) are worse.

Divergence rate

The divergence rate is computed as the Bray-Curtis distance between expected and observed abundance profiles at the species level. It therefore reflects the ability of the tool to detect species in the right proportions. The divergence rates obtained in this study are very high and, as expected, are lower for PCR mocks, as we can see in Figure 7. FROGS, DADA2_FROGS, USEARCH, QIIME-se and DADA2-se yield equivalent results for this indicator (~46-47% on average) while DADA2-pe and QIIME-pe show higher divergence rates (65-69%).

Reconstruction of sequences

DADA2_FROGS is able to reconstruct more sequences than the other methods. Indeed, 79.2% of expected sequences are found without errors. FROGS is very close with 75.9%, followed by DADA2-pe (75.1%) and QIIME-pe (74.9%). USEARCH (66%), DADA2-se (17.6%) and QIIME-se (1.2%).

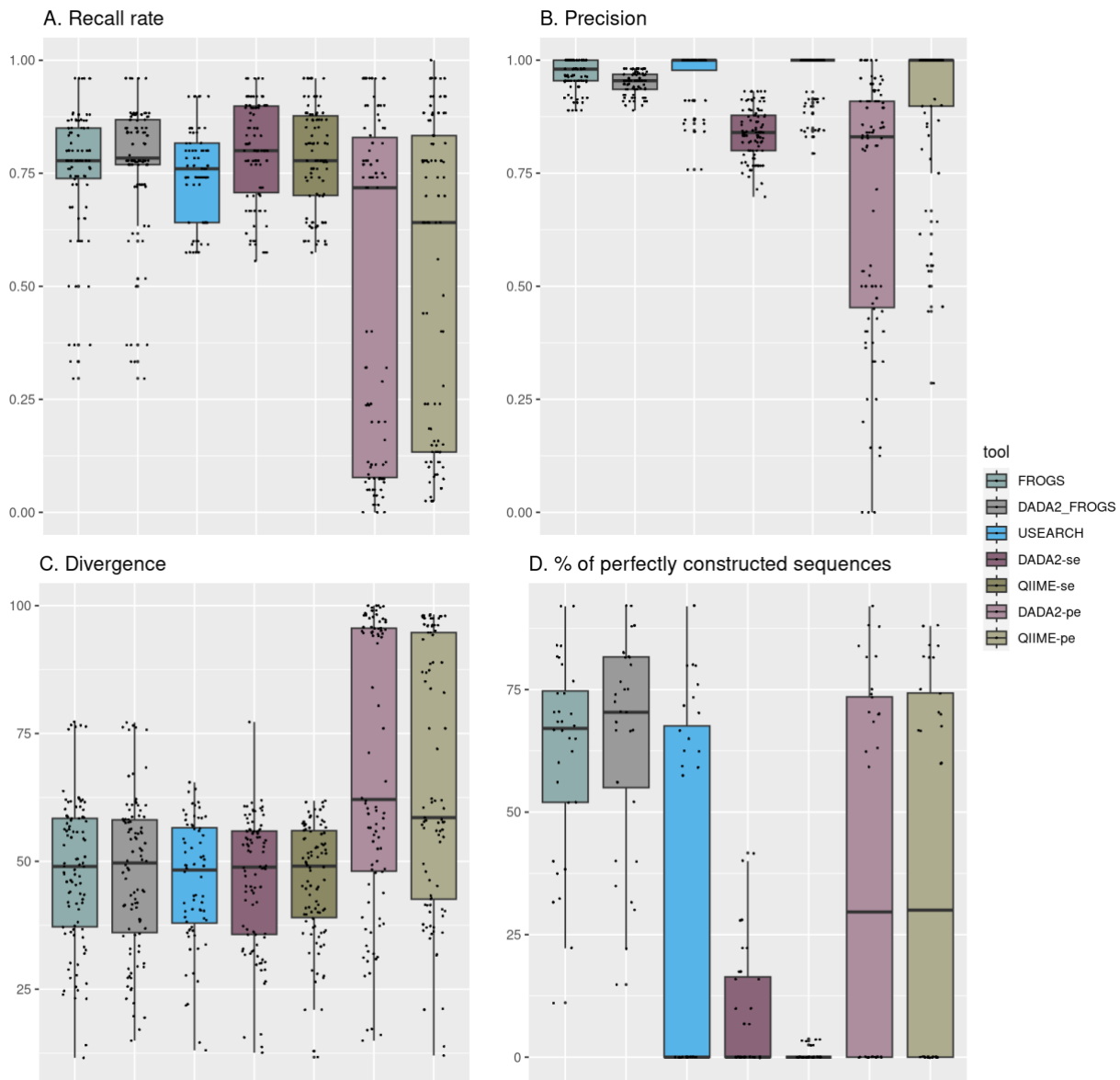

Figure 5 - Quality parameters obtained with the seven bioinformatics pipelines. A) Recall rate (TP/(TP+FN)) reflects the capacity of the tools to detect expected species. B) Precision (TP/(TP+FP)) shows the fraction of relevant species among the retrieved species. C) Divergence rate is the Bray-Curtis distance between expected and observed species abundance. D. Percentage of perfectly reconstructed sequences is the fraction of predicted sequences with 100% of identity with the expected ones.

Overall, the results obtained with the four indicators reveal that the DADA2_FROGS approach performs the best for analyzing ITS1, ITS2, D1/D2 and RPB2 mock samples. We thus selected this approach for all subsequent analysis. Nevertheless, it should be noted that the FROGS tool also performs well as it yields indicator values that are similar to DADA2_FROGS. The main difference is due to species harboring very similar sequences, such as those belonging to Penicillium spp.

Effect of amplicon length on the detected relative abundance

The ITS1 and ITS2 amplicon size is highly variable depending on the considered fungal species, as observed for those included in our mock communities (Figure 2). The effect of amplicon size on the relative abundance of the different species was evaluated using the PCR mock dataset (Figure 6) (code and figures are available on Recherche Data Gouv platform: https://doi.org/10.57745/APNOH8).

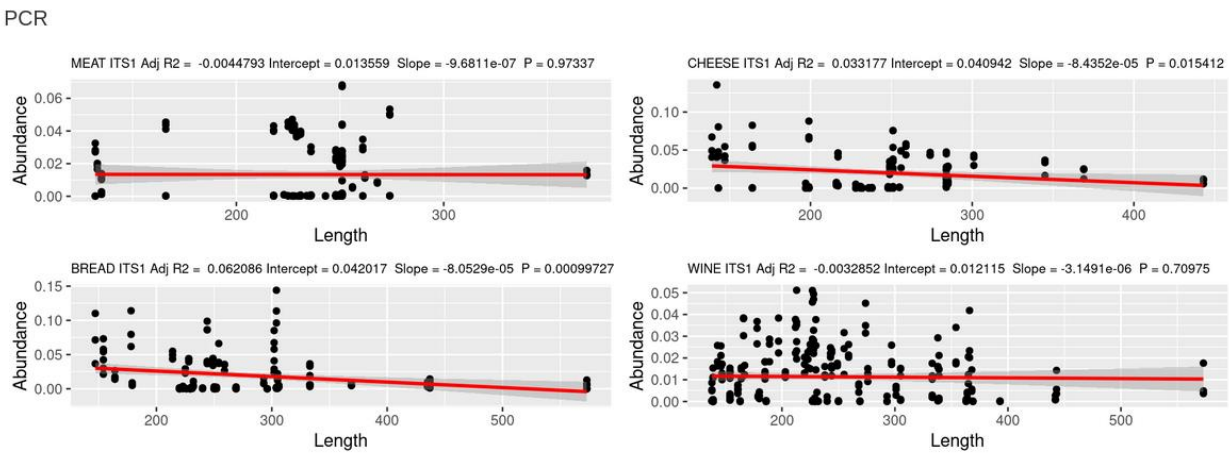

Figure 6 - Effect of ITS1 amplicon size on the relative abundance of the detected ASVs in the PCR mock datasets.

A significant negative relationship is observed between amplicon length and relative abundance in two out of four tested mock communities for ITS1 (cheese and bread, but not meat and wine) and ITS2 (meat and wine, but not cheese and bread). Furthermore, the determination coefficient (R2), which indicates the proportion of variation in the relative abundance data that is predictable from the amplicon length, is comprised between 0.013 and 0.085. This parameter therefore only has a limited impact on the observed proportions when using ITS1 and ITS2 as barcode markers.

Comparison of markers

To compare the capacity of each marker to correctly reflect the fungal diversity present in fermented food samples, we only focused on the results obtained with the DADA2_FROGS approach. Code and figures are available on Recherche Data Gouv platform (https://doi.org/10.57745/X6AXA6). Figure 7 shows the divergence rate obtained for each marker and mock community type (DNA or PCR mock communities).

Figure 7 - Divergence at species level for each tested barcode marker and mock community. DNA mock communities are colored in red and PCR mock communities are colored in blue.

Figure 8 shows the amount of absent, partially and perfectly reconstructed sequences for each ecosystem.

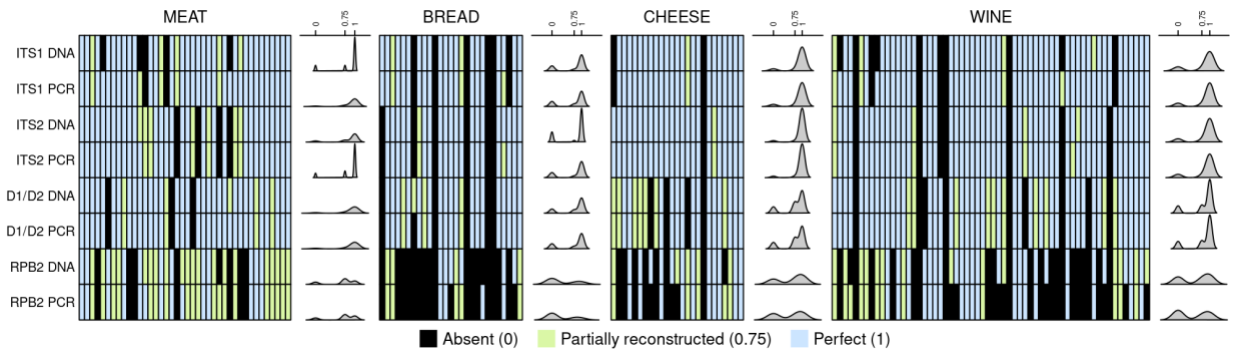

Figure 8 - Heatmap of expected species for each mock community and barcode marker using the DADA2_FROGS approach. Dark green represents perfectly reconstructed sequences (score of 1), light green partially reconstructed sequences (score of 0.75) and orange species that are missing (0). Density of results is indicated for each line, representing the ability of markers to be efficient.

Bread

Regarding the sourdough bread mock community, divergence varies according to the tested barcode marker. While RPB2 presents the highest divergence (75%), D1/D2 shows the lowest (34%-42%) while ITS1 and ITS2 show an intermediate level of divergence. Contrary to what was found for cheese and meat mock communities, divergence is not higher for the mock community DNA mixture than for the mock community PCR mixture, except for ITS1 where the PCR mock mixture presents 35% divergence while the DNA mock mixture ranges from 55% to 61%.

ITS1 and ITS2 perform equally well in terms of false negatives (n=6), false positives (n=1) and true positives (n=21). D1/D2 exhibits one less false negatives (n=5), one more true positives (n=22) and one more false positives than both ITS barcode markers. RPB2 is the worst performing marker with the highest number of false negatives (n=12-14), and the lowest number of true positives (n=8-10).

These results show that D1/D2, ITS1 and ITS2 are all relevant for the analysis of sourdough microbiota but it cannot be concluded which one is best.

Cheese

Regarding the cheese mock community, D1/D2, ITS1 and ITS2 show comparable performance in terms of divergence and are more accurate than RPB2. It is noteworthy that divergence is higher for DNA (between 45 and 70%) than PCR (between 15 to 50%) mock communities. ITS2 exhibits less false negatives (only one) and more true positives (24/25) than other markers. However, it generates two false positives with the DNA mock community and one with the PCR mock community samples while D1/D2 and RPB2 generates one false positive with both DNA and PCR samples. Altogether, these results indicate that, according to the four markers used in this study, ITS2 provides the best compromise to accurately profile cheese fungal communities.

Fermented meat

Regarding the fermented meat mock community, all tested markers show comparable performance in terms of divergence, ranging from 52 to 56% and 23 to 35% for the mock DNA and mock PCR mixture, respectively. As observed for the cheese mock community, divergence is 1.5-2 times higher for the mock DNA

mixture than for the mock PCR mixture, the lowest divergence being observed with ITS1 marker in the mock PCR mixture (23% divergence). D1/D2 exhibits the highest number of false negatives (n=8) as compared to other markers while for PCR mock community mixture, ITS1 and RPB2 exhibit 4 and 5 false negatives, respectively. Concerning the true positive metric, all markers perform equally well with between 32 and 33 true positives out of 40 expected species for both mock DNA and PCR mixtures with the exception of ITS1 marker in the mock PCR mixture which yields 36 true positives. Based on the above mentioned results, we conclude that the ITS1 barcode is slightly more accurate for profiling fermented meat fungal communities, although ITS2 and RPB2 also performed well.

Wine

RPB2 marker does not detect most species (22/60 not found). In contrast, D1/D2, ITS1 and ITS2 display similar results to describe the mock community although not completely; ITS2 is slightly better than the other markers. For the latter three markers, similar performance in terms of divergence is obtained, but better for the DNA mock community than the PCR mock community. However, at least 7 species out of 60 are not identified. ITS2 is also shown to be the most efficient.

Similar to the cheese ecosystem, the ITS2 barcode was the most accurate to explore wine mycobiota, followed by ITS1.

Analysis of real samples

We then compared the efficiency of the four barcodes to detect species in real samples in order to validate our mock results and take into account the fermented food matrix. Code and figures are available on Recherche Data Gouv platform (https://doi.org/10.57745/ENE09G).

Bread

Metabarcoding results from the wheat sourdough sample analyses showed that hits with identities above 80% were not detected using the RPB2 marker. Besides fungal DNA, the other three markers amplified plant DNA. The number of plant DNA hits was much higher for ITS2 and D1/D2 than ITS1. Moreover, ITS2 and D1/D2 markers amplified DNA from *Triticum* species (*Triticum aestivum*, *Triticum monococcum*, *Triticum durum*), and crop weeds, such as *Viciae* sp., *Gallium* sp. and *Calystegia* sp., often found in cereal fields. The ITSx tool automatically removed plant-derived ASVs in the final ITS1 and ITS2 ASV table whereas those from the D1D2 dataset had to be manually removed.

Regarding filamentous fungi, several genera were not detected with D1/D2 including *Aspergillus*, *Aureobasidium* and *Tilletia*. Regarding mycotoxin-producing wheat pathogens (e.g., *Fusarium* and *Penicillium* spp.) or species involved in negatively impacting grain and flour quality for bread making (e.g., rotten fish smell due to *Tilletia* sp.), results showed that the ITS1 barcode did not detect *Penicillium* spp. contrary to D1/D2 and ITS2 barcode markers. On the other hand, ITS1 allowed a better resolution to the species level within the *Fusarium* and *Tilletia* genus.

Regarding fermenting yeast, the well-known bakery yeast *Saccharomyces cerevisiae* was detected by all three markers. In contrast, *Wickerhamomyces anomalus*, frequently encountered among dominant yeast species in sourdoughs worldwide, was only detected by the ITS1 and ITS2 markers although not consistently. It was found in all sourdough samples using the ITS1 marker but only in two out of the three sourdough samples using the ITS2 marker.

Based on the comparison of the RPB2, ITS1, ITS2, D1/D2 markers on real sourdough samples, ITS1 is the best adapted marker to describe sourdough mycobiota as lower reads due to plant DNA were observed and the best detection of sourdough fungal species was obtained.

487 *Cheese*

488 At the genus level, D1/D2, ITS1 and ITS2 all detected *Geotrichum* and *Debaryomyces* as the major fungal
489 genera present on the surface of the three studied cheeses, contrary to RPB2, which placed *Debaryomyces* and
490 *Kluyveromyces* as the dominant taxa and exhibited high variations between biological replicates. So, we
491 decided to exclude RBP2 from the comparison. Regarding the three other markers, some important
492 discrepancies were observed. First, *Yarrowia* and *Mucor* species were only detected in real cheese samples
493 with D1/D2 and ITS2. These species are among the most prevalent fungi in cheese products. Secondly, within
494 *Geotrichum*, ITS1 detected ASVs affiliated with *G. candidum* but also with *Geotrichum* sp. while both D1/D2
495 and ITS2 only detected *G. candidum* species. Thirdly, within the *Mucor* genus, better species level resolution
496 was observed with ITS2 compared to D1/D2. Similarly, it was possible to correctly assign the species
497 *Kluyveromyces lactis* to the corresponding ASV when using ITS1 and ITS2, but the one obtained with D1/D2
498 could not be differentiated between *K. lactis* and *K. marxianus*. As these two species are frequently used as
499 ripening cultures in cheese production, it may be important to choose the ITS1 or ITS2 primers to discriminate
500 between them. Finally, *Candida*, which aggregates with species from different genera including species with
501 uncertain affiliations, was abundant in sample 1 (Saint-Nectaire cheese) based on both ITS regions but was
502 only slightly detected with D1/D2. Altogether, these results indicate that ITS2 performs best to describe cheese
503 fungal communities, followed by D1/D2.

504 *Fermented meat*

505 At the genus level, *Debaryomyces* and *Penicillium*, which are major genera on the casings of fermented
506 meat, were detected with all tested barcode markers. Interestingly, *Kurtzmaniella* sp. (*"Candida" zeylanoides*)
507 was only identified using D1D2 and ITS2 barcodes while *Yarrowia* and *Scopulariopsis* sp. were only identified
508 using D1D2 and ITS2, and, D1D2, ITS2 and RPB2, respectively. Noteworthy, both genera were found in a much
509 larger number of samples using the ITS2 barcode. At the species level, *D. hansenii* and *Penicillium nalgiovense*
510 were found, as expected, to be the most dominant taxa by all 4 barcode markers. Noteworthy, an unambiguous
511 assignation of ITS1 and ITS2 ASVs to *P. nalgiovense* was not possible as these ASVs shared 100% similarity with
512 other related *Penicillium* species and with *Penicillium melanoconidium* for ITS1 and ITS2 ASVs, respectively.
513 Among other major species in fermented meat, *P. salamii*, was only identified with ITS1, ITS2 and RPB2
514 barcodes while *P. nordicum*, a mycotoxin-producing fungus, was only found using RPB2. ITS2 was the only
515 barcode that accurately identified *Yarrowia* species (i.e., *Y. deformans* and *Y. lipolytica*). Overall, among the
516 different tested barcodes, ITS2 was the most efficient for identifying the majority of fungal meat species, the
517 only exception being for the very diversified *Penicillium* genus.

518 *Wine*

519 In comparison to the other studied ecosystems, species diversity in grape must samples was much more
520 complex. More than one hundred yeast and filamentous fungal species were detected including 33 that were
521 part of the wine mock.

522 When comparing the main species encountered, striking differences were observed between ITS1, ITS2
523 and D1D2 results. *Aspergillus* and *Aureobasidium* were detected in all samples, although with large differences
524 according to barcode. Nearly 50% was identified in the triplicates of sample 1 using ITS2 while only 25% with
525 ITS1 and about 40% with D1D2 (higher abundances of *Aspergillus* than *Aureobasidium* were also noted). An
526 opposite situation was observed for *Botrytis*. Indeed, this genus was much more abundant using ITS1 versus
527 ITS2 and D1D2. In agreement with mock analysis results, *Starmerella bacillaris*, a major component of grape
528 must microbiota, was detected at very low abundance in the different samples with ITS1, whereas this species
529 was detected in all samples with ITS2 and D1D2, and represented up to 40% in sample 2 according to ITS2 read
530 counts. In a similar manner, *Hanseniaspora uvarum* was poorly detected with ITS1, well detected with ITS2
531 and most abundant with D1D2. In contrast, *Metschnikowia* species were detected in a similar manner with

ITS1 and ITS2 (more than 40% abundance in samples 3.1 and 3.2) despite the short length and the polymorphism of the amplified sequences.

In conclusion, the ITS2 barcode provides the most comprehensive description of grape must mycobiota.

Discussion

The aim of this study was to compare bioinformatic tools and barcodes used to describe fungal communities in different fermented foods. Choosing the most robust barcode marker for accurate description of fungal communities is crucial. However, various challenges/biases have to be addressed or taken into account for their accurate characterization such as incompleteness of reference databases, low taxonomic depth and PCR amplification biases. In addition, dedicated pipelines also need to be evaluated. In the present study, we built mock fungal communities that gathered the most representative species of fermented meat, cheese, sourdough bread and grape must (wine). These mock communities were used to compare the performances of the main bioinformatic tools available to the scientific community (FROGS, USEARCH, QIIME and DADA2) as well as a combination of DADA2 and FROGS using reads obtained from four commonly used barcodes for fungal community assessment, i.e. ITS1, ITS2, D1D2 of the rDNA as well as RPB2. In addition, to compare these bioinformatic pipelines, we built an in-house database of barcode sequences as many sequences from major fungal species found in these fermented foods were missing in currently available databases. Finally, after selecting the best bioinformatic pipeline, we compared the performances of these four barcodes using real fermented food samples.

In the first part of this study, we compared several commonly used bioinformatic tools. By combining the denoising step of DADA2 followed by the FROGS pipeline, we defined a “universal” pipeline for all barcodes. It combines the advantages of FROGS (dealing with all amplicon lengths) and those of denoising approaches (best resolution and stable ASVs to compare datasets from different studies). Our pipeline avoids the pitfall of other tools in which targeting short barcodes rather than long ones is required (Brandon-Mong et al., 2015; Leray et al., 2013).

One of the main challenges in fermented foods is to characterize microbial communities at the species level including fungi. Food fungal communities are less diversified at the genus level but diversity within genera needs to be determined as it can be relatively high. Robust barcode markers are thus required to reach species level descriptions among genera. Protein-coding genes can be useful to reach this goal and among them, RPB2 is one of the most commonly used taxonomic marker. However, our results clearly highlighted the poor performance of this gene as a barcode, due to a lack of amplification of the barcode. This might be related to the choice of primers, which were originally designed for Basidiomycota (Matheny, 2005). To overcome this limitation, it would be worth designing new consensus primers suitable for Basidiomycota, Ascomycota and Mucoromycotina. Alternative protein-coding genes also need to be tested.

Besides protein-coding genes, rDNA barcodes provide a good global view of mycobiota. Previous studies compared ITS1 versus ITS2 (Bokulich and Mills, 2013) or ITS versus nuclear ribosomal large subunit (LSU) barcodes (Brown et al., 2014). All studies converged on the proposition of using ITS as the primary fungal barcode (Schoch et al., 2012). The LSU appeared to have superior species resolution in some taxonomic groups (Mota-Gutierrez et al., 2019), such as the early diverging lineages and ascomycete yeasts, but was otherwise slightly inferior to ITS (Schoch et al., 2012). ITS1 and ITS2 are, in general, more resolutive markers than D1D2, in particular for filamentous fungi (Mota-Gutierrez et al., 2019). ITS1 locus generally has the shortest mean amplicon lengths for all phyla, the smallest difference between Ascomycota and Basidiomycota amplicon lengths, and the highest species- and genus-level classification accuracy for short amplicon reads, arguing for the primacy of this locus, compared to ITS2 (Bokulich and Mills, 2013). However, in the present study, we found that none of the rDNA markers allowed us to unequivocally discriminate between all species in real fermented

foods. For example, this was emphasized for several mold species, especially species belonging to *Penicillium* or *Pichia spp.*

While D1D2 is among the reference sequences for fungal taxonomy, we found that it had less discriminating power to differentiate species as compared to ITS1 and ITS2 barcodes. This agrees with the previous analysis of 9,000 yeast strains, showing that 6 and 9.5% of the yeast species could not be distinguished by ITS and LSU, respectively (Vu et al., 2016). Indeed, LSU is more conserved than ITS.

The ITS1 and ITS2 markers performed better than D1D2 but their performance varied according to the tested fermented foods. Indeed, while ITS2 performed better for cheese, meat and wine, ITS1 seems better for bread. Concerning the latter, ITS2 primers amplified the ITS from wheat and several weeds which hampered its efficiency for cereal-based products. Besides the fermented product being studied, the choice of the ITS barcode also depends on the expected species diversity in the ecosystem and may be driven by fungal species of interest. Moreover, choice of ITS primers can also be adapted to targeted species. Indeed, some ITS barcoding primers may have mismatches with the sequence of fungal species of interest, such as *Yarrowia* species (Tedersoo and Lindahl, 2016). Finally, although ITS1 and ITS2 seem to be the best barcodes for distinguishing between species, their difference in size may hinder sequence alignment and therefore beta diversity estimates that take phylogenetic distances into account.

One of the limits addressed in the present study was the availability of a complete database adapted to the chosen fermented foods. We thus developed an in-house sequence database for all four major fermented foods in order to fill in this gap. This database significantly improved species level affiliations although manual curation was still required for some genera with complex taxonomy such as *Penicillium* spp. These results also illustrate the need to expand public databases with specific databases.

Metabarcoding is known to be a semi-quantitative method. It is considered to suffer from amplification biases caused by fragment length polymorphism. We did not find evidence for any correlation between amplicon size and read abundance. The divergence between the expected and observed frequency likely results from differences in copy number of rDNA genes (Sternes et al., 2017). Normalization of relative abundance by qPCR targeting a standard reference (Zemb et al., 2020) or by digital PCR (Floren et al., 2015; Zimmer-Faust et al., 2021) might also correct for DNA extraction bias.

In conclusion, although ITS2 appears as the most accurate barcode marker for fermented meat, cheese and wine samples and ITS1 for sourdough bread, no generic recommendation for all fermented food types can be made. This is mainly due to the fact that taxonomic resolution within some genera is not efficient which highlights the need to combine metabarcoding with culture-dependent analysis such as culturomic approaches. The availability of long-read technologies, like Oxford Nanopore Technologies or PacBio technology, provides the opportunity to sequence longer fragments of the fungal ribosomal operon, up to 6 Kb (18S-ITS1-5.8S-ITS2-28S) and to improve the taxonomy assignment of the communities up to species level (D'Andreano et al., 2021) but their current cost is still a brake to replace short reads technologies. Shotgun metagenomics sequencing is also an alternative or a complementary method to study food fermentations (Leech et al., 2020). It may provide a less biased vision of food microbiota than metabarcoding (Sternes et al., 2017), a more comprehensive insight into the microbial composition, and functional potential but at a much higher cost for low abundant species.

Acknowledgements

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources. We thank Julien Lebrat, Stéphane Guezenec, Anne-Sophie Sarthou for technical assistance and Claire Vincent for her

participation in building the reference databank. We also thank Mary and Loulou Bonneau, Stéphane Marrou, Anna and Maximilien for providing security for their sourdough.

Data, scripts, code, and supplementary information availability

Sequencing data are available in NCBI SRA repository under the Bioproject number PRJNA685292.
Scripts and code are available online on Gitlab (https://forgemia.inra.fr/migale/metabarfood).
Supplementary information is available online: https://doi.org/10.57745/AZNJFE, https://doi.org/10.57745/109NNP, https://doi.org/10.57745/ENE09G, https://doi.org/10.57745/X6AXA6 and https://doi.org/10.57745/APNOH8.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. D. Sicard, C. Neuvéglise, K. Howell and J.L. Legras are PCI recommenders.

Funding

This work was supported by the French “Microbial Ecosystems & Meta-omics” (MEM) metaprogram from INRAE. Migale is part of the Institut Français de Bioinformatique (ANR-11-INBS-0013).

References

Anslan, S., Nilsson, R.H., Wurzbacher, C., Baldrian, P., Tedersoo, L., Bahram, M., 2018. Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. MycoKeys 39, 29–40. https://doi.org/10.3897/mycokeys.39.28109

Bernard, M., Rué, O., Mariadassou, M., Pascal, G., 2021. FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics 22, bbab318. https://doi.org/10.1093/bib/bbab318

Berni, E., 2014. Molds, in: Handbook of Fermented Meat and Poultry. John Wiley & Sons, Ltd, pp. 147–153. https://doi.org/10.1002/9781118522653.ch17

Bokulich, N.A., Mills, D.A., 2013. Improved Selection of Internal Transcribed Spacer-Specific Primers Enables Quantitative, Ultra-High-Throughput Profiling of Fungal Communities. Applied and Environmental Microbiology 79, 2519–2526. https://doi.org/10.1128/AEM.03870-12

Bokulich, N.A., Ohta, M., Richardson, P.M., Mills, D.A., 2013. Monitoring Seasonal Changes in Winery-Resident Microbiota. PLoS One 8:e66437. https://doi.org/10.1371/journal.pone.0066437

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B.,

Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37, 852–857. https://doi.org/10.1038/s41587-019-0209-9

Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P.-S., Lim, P.-E., Wilson, J.-J., 2015. DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. Bulletin of Entomological Research 105, 717–727. https://doi.org/10.1017/S0007485315000681

Brown, S.P., Rigdon-Huss, A.R., Jumpponen, A., 2014. Analyses of ITS and LSU gene regions provide congruent results on fungal community responses. Fungal Ecology 9, 65–68. https://doi.org/10.1016/j.funeco.2014.02.002

Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11, 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G., Knight, R., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6, 1621–1624. https://doi.org/10.1038/ismej.2012.8

Caruso, V., Song, X., Asquith, M., Karstens, L., 2019. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. mSystems 4, e00163-18. https://doi.org/10.1128/mSystems.00163-18

Coton, M., Deniel, F., Mounier, J., Joubrel, R., Robieu, E., Pawtowski, A., Jeuge, S., Taminiau, B., Daube, G., Coton, E., Frémaux, B., 2021. Microbial Ecology of French Dry Fermented Sausages and Mycotoxin Risk Evaluation During Storage. Frontiers in Microbiology 12.

D'Andreano, S., Cuscó, A., Francino, O., 2021. Rapid and real-time identification of fungi up to species level with long amplicon nanopore sequencing from clinical samples. Biology Methods and Protocols 6, bpaa026. https://doi.org/10.1093/biomethods/bpaa026

De Filippis, F., Laiola, M., Blaiotta, G., Ercolini, D., 2017. Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity. Appl Environ Microbiol 83, e00905-17. https://doi.org/10.1128/AEM.00905-17

Dugat-Bony, E., Straub, C., Teissandier, A., Onésime, D., Loux, V., Monnet, C., Irlinger, F., Landaud, S., Leclercq-Perlat, M.-N., Bento, P., Fraud, S., Gibrat, J.-F., Aubert, J., Fer, F., Guédon, E., Pons, N., Kennedy, S., Beckerich, J.-M., Swennen, D., Bonnarme, P., 2015. Overview of a surface-ripened cheese community functioning by meta-omics analyses. PLoS ONE 10, e0124360. https://doi.org/10.1371/journal.pone.0124360

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., Pascal, G., 2018. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics 34, 1287–1294. https://doi.org/10.1093/bioinformatics/btx791

Floren, C., Wiedemann, I., Brenig, B., Schütz, E., Beck, J., 2015. Species identification and quantification in meat and meat products using droplet digital PCR (ddPCR). Food Chemistry 173, 1054–1058. https://doi.org/10.1016/j.foodchem.2014.10.138

Franciosa, I., Coton, M., Ferrocino, I., Corvaglia, M.R., Poirier, E., Jany, J.-L., Rantsiou, K., Cocolin, L., Mounier, J., 2021. Mycobiota dynamics and mycotoxin detection in PGI Salame Piemonte. Journal of Applied Microbiology 131, 2336–2350. https://doi.org/10.1111/jam.15114

Gardes, M., Bruns, T.D., 1993. ITS primers with enhanced specificity for basidiomycetes--application to the identification of mycorrhizae and rusts. Mol. Ecol. 2, 113–118.

Garofalo, C., Tristezza, M., Grieco, F., Spano, G., Capozzi, V., 2016. From grape berries to wine: population dynamics of cultivable yeasts associated to "Nero di Troia" autochthonous grape cultivar. World J Microbiol Biotechnol 32, 59. https://doi.org/10.1007/s11274-016-2017-4

Glassman, S.I., Martiny, J.B.H., 2018. Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. mSphere 3, e00148-18. https://doi.org/10.1128/mSphere.00148-18

Gweon, H.S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D.S., Griffiths, R.I., Schonrogge, K., 2015. PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods in Ecology and Evolution 6, 973–980. https://doi.org/10.1111/2041-210X.12399

Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D., 2012. New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. FEMS Microbiology Ecology 82, 666–677. https://doi.org/10.1111/j.1574-6941.2012.01437.x

Irlinger, F., Layec, S., Hélinck, S., Dugat-Bony, E., 2015. Cheese rind microbial communities: diversity, composition and origin. FEMS microbiology letters, 362(2), 1-11. https://doi.org/10.1093/femsle/fnu015

Jolly, N.P., Augustyn, O.P.H., Pretorius, I.S., 2003. The Occurrence of Non-Saccharomyces cerevisiae Yeast Species Over Three Vintages in Four Vineyards and Grape Musts From Four Production Regions of the Western Cape, South Africa. South African J. Enol. Vitic. 24:8–10. https://doi.org/10.21548/24-2-2640

Katoh, K., Asimenos, G., Toh, H., 2009. Multiple Alignment of DNA Sequences with MAFFT, in: Posada, D. (Ed.), Bioinformatics for DNA Sequence Analysis, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 39–64. https://doi.org/10.1007/978-1-59745-251-9_3

Lavrinienko, A., Jernfors, T., Koskimäki, J.J., Pirttilä, A.M., Watts, P.C., 2021. Does Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities? Trends in Microbiology 29, 19–27. https://doi.org/10.1016/j.tim.2020.05.019

Leech, J., Cabrera-Rubio, R., Walsh, A.M., Macori, G., Walsh, C.J., Barton, W., Finnegan, L., Crispie, F., O'Sullivan, O., Claesson, M.J., Cotter, P.D., 2020. Fermented-Food Metagenomics Reveals Substrate-Associated Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants. mSystems 5, e00522-20. https://doi.org/10.1128/mSystems.00522-20

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T., Machida, R.J., 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10, 34. https://doi.org/10.1186/1742-9994-10-34

Loos, D., Zhang, L., Beemelmanns, C., Kurzai, O., Panagiotou, G., 2021. DAnIEL: A User-Friendly Web Server for Fungal ITS Amplicon Sequencing Data. Frontiers in Microbiology 12.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12. https://doi.org/10.14806/ej.17.1.200

Matheny, P.B., 2005. Improving phylogenetic inference of mushrooms with RPB1 and RPB2 nucleotide sequences (Inocybe; Agaricales). Molecular Phylogenetics and Evolution 35, 1–20.https://doi.org/10.1016/j.ympev.2004.11.014

Montel, M. C., Buchin, S., Mallet, A., Delbes-Paus, C., Vuitton, D. A., Desmasures, N., Berthier, F., 2014. Traditional cheeses: Rich and diverse microbiota with associated benefits. International journal of food microbiology, 177, 136-154. http://dx.doi.org/10.1016/j.ijfoodmicro.2014.02.019

Mota-Gutierrez, J., Ferrocino, I., Rantsiou, K., Cocolin, L., 2019. Metataxonomic comparison between internal transcribed spacer and 26S ribosomal large subunit (LSU) rDNA gene. International Journal of Food Microbiology 290, 132–140. https://doi.org/10.1016/j.ijfoodmicro.2018.10.010

Nilsson, R. Henrik, Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., Tedersoo, L., 2019. Mycobiome diversity: high-throughput sequencing and identification of fungi. Nat Rev Microbiol 17, 95–109. https://doi.org/10.1038/s41579-018-0116-y

Nilsson, Rolf Henrik, Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L., Saar, I., Kõljalg, U., Abarenkov, K., 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research 47, D259–D264. https://doi.org/10.1093/nar/gky1022

O'Donnell, K., 1993. Fusarium and its near relatives, in: Reynolds, D., Taylor, J. (Eds.), . Presented at the The fungal holomorph: mitotic, meiotic and pleomorphic speciation in fungal systematics, pp. 225–233.

Özkurt, E., Fritscher, J., Soranzo, N., Ng, D.Y.K., Davey, R.P., Bahram, M., Hildebrand, F., 2022. LotuS2: an ultrafast and highly accurate tool for amplicon sequencing analysis. Microbiome 10, 176. https://doi.org/10.1186/s40168-022-01365-1

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE 5, e9490. https://doi.org/10.1371/journal.pone.0009490

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Rossouw, D., Bauer, F., 2016. Exploring the phenotypic space of non-Saccharomyces wine yeast biodiversity. Food Microbiol. 55:32–46. http://dx.doi.org/10.1016/j.fm.2015.11.017

Rué, Olivier, 2022a, "METABARFOOD - Description of the sequences used in mock communities", https://doi.org/10.57745/AZNJFE, Recherche Data Gouv, V1

Rué, Olivier, 2022b, "METABARFOOD - Results on mock communities, choice of the bioinformatics solution", https://doi.org/10.57745/109NNP, Recherche Data Gouv, V1

Rué, Olivier, 2022c, "METABARFOOD - Results on real communities", https://doi.org/10.57745/ENE09G, Recherche Data Gouv, V1

Rué, Olivier, 2022d, "METABARFOOD - Comparison of amplicons", https://doi.org/10.57745/X6AXA6, Recherche Data Gouv, V1

Rué, Olivier, 2022e, "METABARFOOD - Complementary analyses", https://doi.org/10.57745/APNOH8, Recherche Data Gouv, V1

Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., Sherry, S.T., 2022. Database resources of the national center for biotechnology information. Nucleic Acids Research 50, D20–D26. https://doi.org/10.1093/nar/gkab1112

Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593. https://doi.org/10.1093/bioinformatics/btq706

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Consortium, F.B., List, F.B.C.A., Bolchacova, E., Voigt, K., Crous, P.W., Miller, A.N., Wingfield, M.J., Aime, M.C., An, K.-D., Bai, F.-Y., Barreto, R.W., Begerow, D., Bergeron, M.-J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N., Burgaz, A.R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P., Cummings, C., Damm, U., Beer, Z.W. de, Hoog, G.S. de, Del-Prado, R., Dentinger, B., Diéguez-Uribeondo, J.,

Divakar, P.K., Douglas, B., Dueñas, M., Duong, T.A., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova, K., Furtado, M., García, M.A., Ge, Z.-W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M., Grube, M., Gryzenhout, M., Guo, L.-D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P., Heller, G., Herrera, C., Hirayama, K., Hirooka, Y., Ho, H.-M., Hoffmann, K., Hofstetter, V., Högnabba, F., Hollingsworth, P.M., Hong, S.-B., Hosaka, K., Houbraken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T., Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B.G., Kelly, L.J., Kirk, P.M., Knapp, D.G., Kõljalg, U., Kovács, G.M., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Liggenstoffer, A.S., Liimatainen, K., Lombard, L., Luangsa-ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S.N., Martin, M.P., May, T.W., McTaggart, A.R., Methven, A.S., Meyer, W., Moncalvo, J.-M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H., Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas, R., Quaedvlieg, W., Raja, H.A., Redecker, D., Rintoul, T.L., Ruibal, C., Sarmiento-Ramírez, J.M., Schmitt, I., Schüßler, A., Shearer, C., Sotome, K., Stefani, F.O.P., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S., Suh, S.-O., Sung, G.-H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W.A., Urbina, H., Vágvölgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.-M., Wang, Y., Weir, B.S., Weiß, M., White, M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.-C., Zhang, N., Zhuang, W.-Y., Schindel, D., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS 109, 6241–6246. https://doi.org/10.1073/pnas.1117018109

Selgas, M. d., García, M. l., 2014. Yeasts, in: Handbook of Fermented Meat and Poultry. John Wiley & Sons, Ltd, pp. 139–146. https://doi.org/10.1002/9781118522653.ch16

Setati, M.E., Jacobson, D., Andong, U.C., Bauer, F., 2012. The Vineyard Yeast Microbiome, a Mixed Model Microbial Map. PLoS One 7:e52609. https://doi.org/10.1371/annotation/b9d307d9-f5c1-4e0d-8945-c5a747b6f58e

Setati, M.E., Jacobson, D., Bauer, F., 2015. Sequence-based analysis of the Vitis vinifera L. cv cabernet sauvignon grape must mycobiome in three South African vineyards employing distinct agronomic systems. Front. Microbiol. 6:1–12. https://doi.org/10.3389/fmicb.2015.01358

Shen, W., Ren, H., 2021. TaxonKit: A practical and efficient NCBI taxonomy toolkit. Journal of Genetics and Genomics, Special issue on Microbiome 48, 844–850. https://doi.org/10.1016/j.jgg.2021.03.006

Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21, 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Sternes, P.R., Lee, D., Kutyna, D.R., Borneman, A.R., 2017. A combined meta-barcoding and shotgun metagenomic analysis of spontaneous wine fermentation. GigaScience 6, gix040. https://doi.org/10.1093/gigascience/gix040

Tedersoo, L., Lindahl, B., 2016. Fungal identification biases in microbiome projects. Environmental microbiology reports, 8(5), 774-779. https://doi.org/10.1111/1758-2229.12438

Toju, H., Tanabe, A.S., Yamamoto, S., Sato, H., 2012. High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLoS One 7. https://doi.org/10.1371/journal.pone.0040863

Větrovský, T., Kolařík, M., Žifčáková, L., Zelenka, T., Baldrian, P., 2016. The rpb2 gene represents a viable alternative molecular marker for the analysis of environmental fungal communities. Mol Ecol Resour 16, 388–401. https://doi.org/10.1111/1755-0998.12456

von Gastrow, L., Gianotti, A., Vernocchi, P., Serrazanetti, D.I., Sicard, D., 2023. Taxonomy, Biodiversity, and Physiology of Sourdough Yeasts. In: Gobbetti, M., Gänzle, M. (eds) Handbook on Sourdough Biotechnology. Springer, Cham. https://doi.org/10.1007/978-3-031-23084-4_7

von Gastrow, L., Michel, E., Legrand, J., Amelot, R., Segond, D., Guezenec, S., Rué, O., Chable, V., Goldringer, I., Dousset, X., Serpolay-Bessoni, E., Taupier-Letage, B., Vindras-Fouillet, C., Onno, B., Valence, F., Sicard, D.,

2022. Microbial community dispersal from wheat grains to sourdoughs: A contribution of participatory research. Molecular Ecology mec.16630. https://doi.org/10.1111/mec.16630

Vu, D., Groenewald, M., Szöke, S., Cardinali, G., Eberhardt, U., Stielow, B., de Vries, M., Verkleij, G.J.M., Crous, P.W., Boekhout, T., Robert, V., 2016. DNA barcoding analysis of more than 9 000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. Studies in Mycology 85, 91–105. https://doi.org/10.1016/j.simyco.2016.11.007

White, T., Bruns, T., Lee, S., Taylor, J., 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, in: Innis, M., Gelfand, D., Shinsky, J., White, T. (Eds.), PCR Protocols: A Guide to Methods and Applications. Academic Press, pp. 315–322.

Yang, R.-H., Su, J.-H., Shang, J.-J., Wu, Y.-Y., Li, Y., Bao, D.-P., Yao, Y.-J., 2018. Evaluation of the ribosomal DNA internal transcribed spacer (ITS), specifically ITS1 and ITS2, for the analysis of fungal diversity by deep sequencing. PLOS ONE 13, e0206428. https://doi.org/10.1371/journal.pone.0206428

Zemb, O., Achard, C.S., Hamelin, J., De Almeida, M.-L., Gabinaud, B., Cauquil, L., Verschuren, L.M.G., Godon, J.-J., 2020. Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard. MicrobiologyOpen 9, e977. https://doi.org/10.1002/mbo3.977

Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., Lu, Z., 2018. Evaluation of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. Science of The Total Environment 618, 1254–1267. https://doi.org/10.1016/j.scitotenv.2017.09.228

Zimmer-Faust, A.G., Steele, J.A., Xiong, X., Staley, C., Griffith, M., Sadowsky, M.J., Diaz, M., Griffith, J.F., 2021. A Combined Digital PCR and Next Generation DNA-Sequencing Based Approach for Tracking Nearshore Pollutant Dynamics Along the Southwest United States/Mexico Border. Frontiers in Microbiology 12.