



HAL
open science

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

Olivier Rué, Monika Coton, Eric Dugat-Bony, Kate Howell, Françoise Irlinger, Jean-Luc Legras, Valentin Loux, Elisa Michel, Jérôme Mounier, Cécile Neuvéglise, et al.

► To cite this version:

Olivier Rué, Monika Coton, Eric Dugat-Bony, Kate Howell, Françoise Irlinger, et al.. Comparison of metabarcoding taxonomic markers to describe fungal communities in fermented foods. Peer Community Journal, 2023, 3 (e97), pp.1-25. 10.24072/pcjournal.321 . hal-04237955v1

HAL Id: hal-04237955

<https://hal.inrae.fr/hal-04237955v1>

Submitted on 11 Oct 2023 (v1), last revised 25 Jan 2024 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

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

Olivier Rué^{1,2}, Monika Coton³, Eric Dugat-Bony⁴, Kate Howell⁵, Françoise Irlinger⁴, Jean-Luc Legras⁶, Valentin Loux^{1,2}, Elisa Michel⁶, Jérôme Mounier³, Cécile Neuvéglise⁶, 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

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

40

41

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

43

44

45

Introduction

47 In the field of microbial ecology, amplicon-based metagenomic analysis (also named metabarcoding) is one
48 of the most popular routes to describe microbial communities as it is a high throughput method with relatively
49 low-cost nowadays. This approach relies on amplifying a phylogenetic biomarker from total community DNA
50 purified from the samples to be characterized, followed by massive amplicon sequencing (Shokralla et al.,
51 2012), usually with Illumina MiSeq technology (Caporaso et al., 2012). For bacteria, the SSU rRNA gene is
52 commonly accepted as the most suitable biomarker for metabarcoding studies although different variable
53 regions can be targeted (Zhang et al., 2018). For fungal communities, there is still no international consensus
54 regarding the choice of the best phylogenetic biomarker for such an approach.

55 Ten years ago, the Fungal Barcoding Consortium recommended the use of the Internal Transcribed Spacer
56 (ITS) region as the primary marker for fungal identifications due to superior species-level resolution compared
57 to LSU and SSU rRNA genes, and higher amplification success compared to protein coding genes (Schoch et al.,
58 2012). However, amplicon-based metagenetic analysis using MiSeq sequencing imposes more constraints than
59 classical species identification. Most importantly, the typical amplicon size (~500bp) will not provide complete
60 ITS region sequences or full length rRNA or protein coding genes. Consequently, the most popular barcodes
61 currently used for fungal community analysis by metabarcoding are ITS1, ITS2 and the LSU D1/D2 domain (R.
62 Henrik Nilsson et al., 2019).

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

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

78 Some authors evaluated the reliability of different markers (ITS1-2, D1/D2 LSU and SSU) to describe fungal
79 diversity by amplicon-based metagenomic analysis (De Filippis et al., 2017). A mock community, composed of
80 19 strains representative of common fungal species, as well as environmental samples including soil, human
81 saliva, human feces and grape must were used. Although all markers were able to correctly detect the different
82 species in the mock community, the results suggested that there was an important quantification bias when
83 using ITS1-2. This could be due to the high heterogeneity in marker length across fungal species. However,
84 marker performance is likely to be highly influenced by fungal species composition of the sample (e.g.,
85 composed mainly of Basidiomycota versus Ascomycota) which, in turn, depends on the studied environment.

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

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

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

105 Another important aspect in metabarcoding data analysis is the way to build representative biological
106 sequences from reads. It can be under the form of Operational Taxonomic Units (OTUs), Amplicon Sequence
107 Variants (ASVs) or zero-radius OTUs (ZOTUs) depending on tools. Numerous studies have compared the results
108 from OTU-based and ASV-based approaches (Callahan et al., 2017). Using mock communities, ASV-based
109 methods had higher sensitivity and detected bacterial strains present, sometimes at the expense of specificity
110 (Caruso et al., 2019). However, a different study concluded that for broadscale (e.g., all bacteria or all fungi) α
111 and β diversity analyses, ASV or OTU methods often provided similar ecological results (Glassman and Martiny,
112 2018). From a practical point of view, an important advantage of ASV-based approaches is the consistent labels
113 with intrinsic biological meaning identified independently from a reference database. Thus, ASVs
114 independently inferred from different studies and different samples (for the same targeted region) can be
115 compared.

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

123 **Methods**

124 **Mock community sample preparation**

125 *Species diversity*

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

130

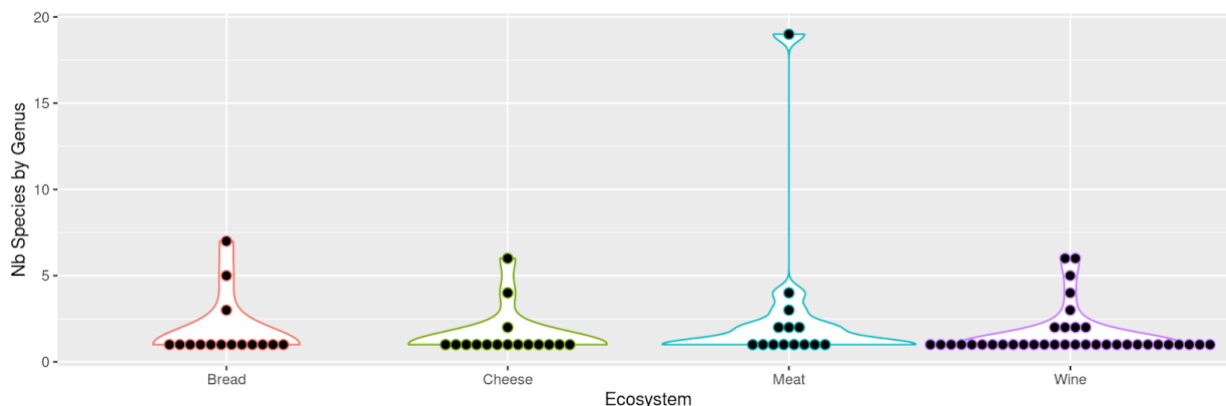
131 **Table 1** - Number of Species, Genus and Family in each mock community

	Bread	Cheese	Meat	Wine

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

132

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



134

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

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

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

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

151 For the wine mock community, it contained 60 strains belonging to Ascomycota (45 strains) and
 152 Basidiomycota (15 strains). Six strains were affiliated to *Hanseniaspora*, 6 to *Pichia*, 5 to *Candida*, 4 to
 153 *Rhodotorula*, 3 to *Papiliotrema* and 2 to *Clavispora*, *Cystobasidium*, *Metschnikowia* and *Meyerozyma*. All others
 154 belonged to a distinct genus. The selection was made from a review of papers investigating grape and wine
 155 microbiota (Setati et al., 2012; Rossouw and Bauer, 2016; Jolly et al., 2003; Setati et al., 2015; Bokulich et al.,
 156 2013 and Garofalo et al., 2016).

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

159

160 **DNA extraction from single strains**

161 *Bread and wine*

162 Each strain was grown overnight at 25°C in 15 mL of YEPD before centrifuging for 10 minutes at 1,500 × g.
163 The cell pellet was resuspended in one mL of sterile water and transferred to a 2 mL tube. After a second
164 centrifugation at 11,800 × g for 2 minutes, the pellet was resuspended in the yeast cell lysis solution from the
165 MasterPure Yeast DNA extraction kit (Epicentre) and DNA was extracted according to the kit procedure.
166

167 *Cheese*

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

171 *Fermented meat*

172 DNA was extracted from scraped colonies for yeasts or mycelial plugs for molds using the FastDNA SPIN Kit
173 (MP Biomedicals) according to the manufacturer's instructions.
174
175
176

177 **Mock community design**

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

187

188

189 **Real samples preparation**

190 DNA extraction was performed for each food environment (bread, wine, cheese, fermented meat)
191 according to different protocols adapted for each matrix. DNA concentration was determined using a Qubit
192 fluorometer (Life Sciences) according to the Broad Range DNA assay kit protocol.
193

194 *Bread*

195 Three types of sourdough coming from different French bakeries were analyzed. All sourdoughs were made
196 of wheat flour. Sourdough 1 was sampled from an artisanal bakery in Azillanet (Occitanie Region), sourdough
197 2 from a local baker in Assas (Occitanie region) and sourdough 3 from a local baker in Amilly (Centre-Val de
198 Loire region). For each sourdough, DNA extraction was performed from 200 mg of three independent samples
199 using the MO BIO's Powersoil DNA isolation kit procedure (Qiagen 12888-100) as described previously (von
200 Gastrow et al., 2022).
201

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 (CTTTCCTACACGACGCTCTCCG-forward primer sequence) and R
236 (GGAGTTCAGACGTGTGCTCTCCG-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 (AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGAC) 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

246 integrated index. The sequencing data from this study are available in NCBI SRA repository under the Bioproject
 247 number PRJNA685292.
 248

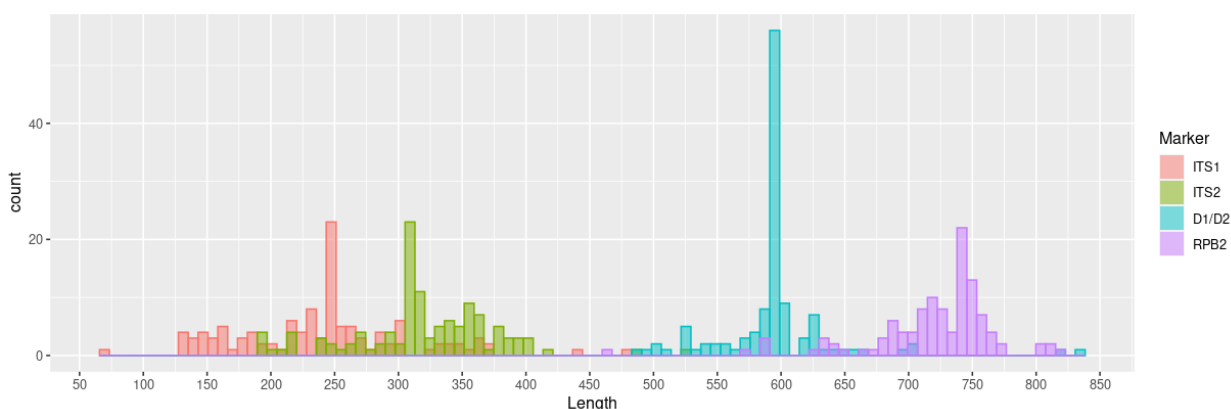
249 **Table 2** - Description of the primers used to target ITS1, ITS2, D1/D2 and RPB2 regions.

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'-GAYTGRTRTGRTRCGGGAAVGG-3'	55°C	(Matheny, 2005)
ITS1	ITS1F	5'-CTTGTCATTAGAGGAAGTAA-3'	ITS2	5'-GCTGCGTTCATCGATGC-3'	55°C	(Gardes and Bruns, 1993; White et al., 1990)
ITS2	ITS3	5'-GCATCGATGAAGAACGCAGC-3'	ITS4Kyo	5'-TCCTCCGCTTWTGWTWTGC-3'	55°C	(Toju et al., 2012; White et al., 1990)

250
 251 **Bioinformatic analysis**

252 *Construction of the reference databank for mock community analysis*

253 For each reference strain (118 in total), the sequence of the four phylogenetic markers (ITS1, ITS2, D1/D2,
 254 RPB2) was obtained from public databases or from unpublished sequences obtained in our labs. Only 3 RPB2
 255 sequences were not available and missing due to PCR amplification failure (*Cryptococcus neoformans*, *Mucor*
 256 *lanceolatus* and *Rhodotorula glutinis*). The length distribution of the 469 sequences is represented in Figure 2.
 257 As expected, variations in length are visible, ranging from 70 to 835 bp. On average, ITS1 sequences are
 258 shortest, followed by ITS2, D1/D2 and RPB2 sequences.
 259

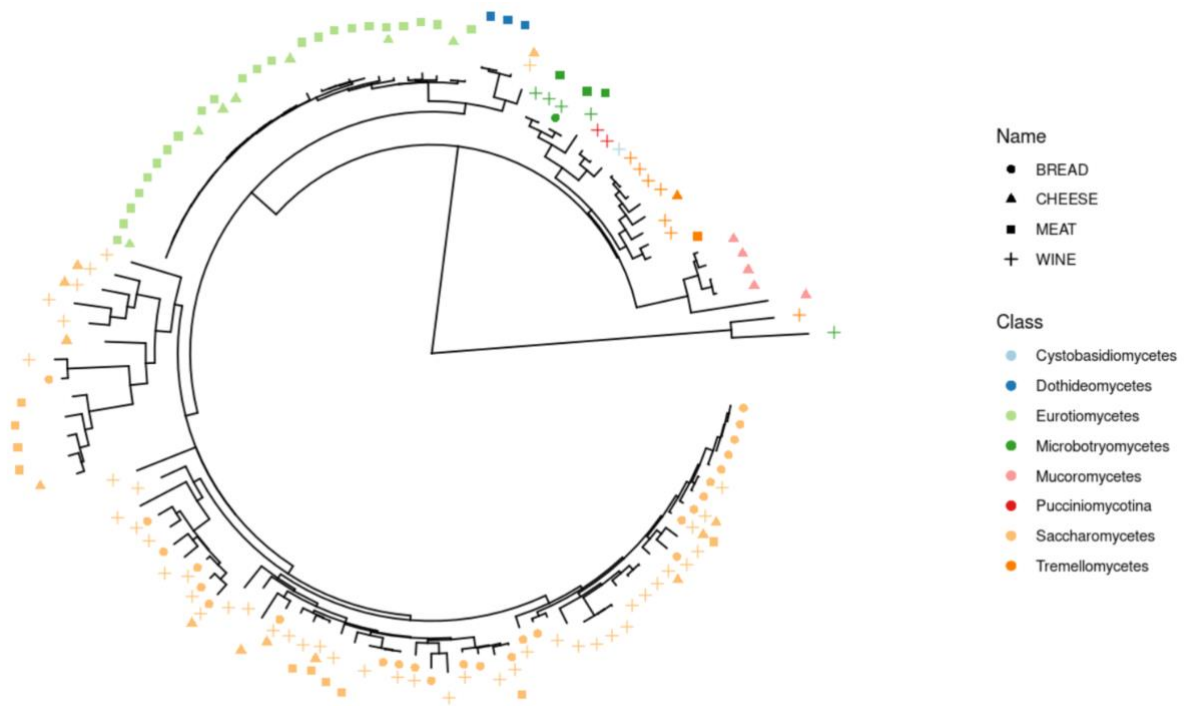


260
 261 **Figure 2** - Sequence length distribution of the four markers for the 118 strains present in mock samples inferred from the
 262 reference database..

263
 264 From all sequences and for each marker, we built phylogenetic trees with FastTree (Price et al., 2010) and
 265 Phangorn R package (Schliep, 2011) after a multiple alignment of sequences with Mafft (Katoh et al., 2009).

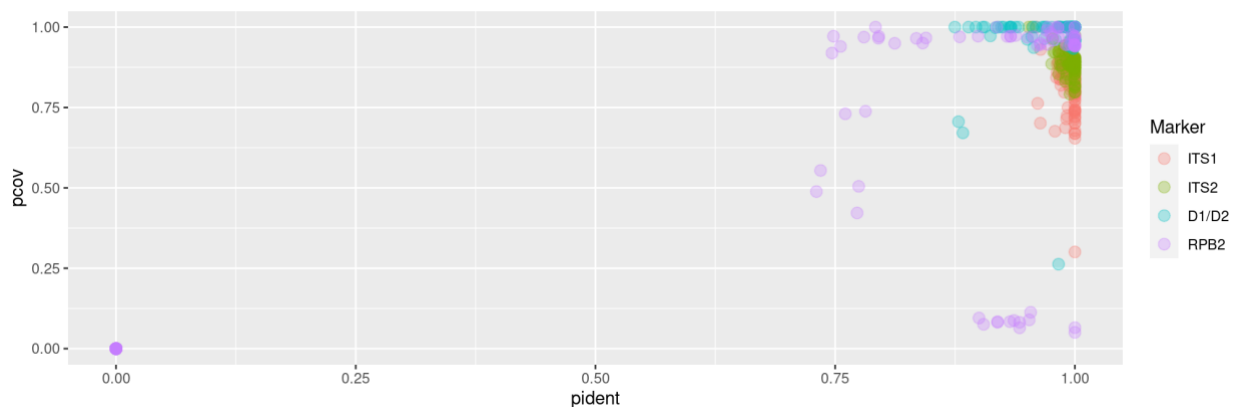
266 Figure 3 shows the diversity of the 118 ITS1 sequences from a phylogenetic point of view. Phylogenetic trees
267 for other markers are available on Recherche Data Gouv platform (<https://doi.org/10.57745/AZNJFE>).

268 Database construction was crucial for benchmarking and was necessary in our case because many
269 sequences are absent in public databanks. Figure 4 shows the best hits of the 469 sequences after a blast
270 against each representative databank (UNITE v9.0 for ITS, SILVA v138 for D1/D2 and nt release 2021-07-30 for
271 RPB2). If the sequence is present in the databank, the corresponding point is at the top right of the graph.
272 If no hit is found, the corresponding point is at the bottom left of the graph. In the middle of the graph are points
273 corresponding to sequences for which the percentage of identity and coverage are less than one hundred
274 percent. This result illustrates the incompleteness of public databases for food ecosystems, particularly for
275 RPB2 sequences.
276



277

278 **Figure 3** - Phylogenetic tree built from D1/D2 sequences of the 118 strains of the mock samples, colored by Class. Shapes
279 indicate the presence in bread, cheese, fermented meat or wine mock samples.
280



281

282 **Figure 4** - Representation of blast results (identity and coverage percentage) of the 468 sequences against dedicated databases
283 (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
284 axis, respectively.

285

286 *Benchmark of metabarcoding approaches*

287 Full codes and figures are available on Recherche Data Gouv platform (<https://doi.org/10.57745/109NNP>).

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

291 One strength of FROGS is the ability to deal with overlapping and non-overlapping reads at the same time.
292 This tool processes ITS1, ITS2, D1/D2 and RPB2 markers equally by a preprocess step (*preprocess.py*) to merge
293 paired-end reads or create “artificial” sequences if they do not merge. In this case, R1 and R2 sequences are
294 concatenated with a stretch of Ns in the middle. The subcommands *preprocess.py*, *clustering.py* with
295 parameters *--fastidious* and *--distance 1*, *remove_chimera.py*, *otu_filters.py* with parameter *--min-abundance*
296 *0.00005*, *itsx.py*, *affiliation_OTU.py* and *affiliation_filters.py* with parameters *--min-blast-coverage 0.9*, *--min-*
297 *blast-identity 0.9* and *--delete* were used.

298 DADA2 is a widely used tool for metabarcoding analyses. It infers exact amplicon sequence variants (ASVs)
299 from amplicon data, resolving biological differences of even 1 or 2 nucleotides. Cutadapt and then the
300 functions *filterAndTrim* (*maxN = 0*, *maxEE = 2*, *truncQ = 2*, *minLen = 50*, *rm.phix = TRUE*), *dada* and *mergePairs*
301 were used. By default, DADA2 does not deal with overlapping and non-overlapping reads at the same time.
302 We used single-end and paired-end modes (DADA2-se and DADA2-pe, respectively). For DADA2-se, only R1
303 reads were taken into account and only overlapping reads for DADA2-pe. Then, for both strategies,
304 *makeSequenceTable*, *removeBimeraDenovo* and *assignTaxonomy* functions were finally used.

305 In the same way, QIIME2 was used in single and paired-end modes (QIIME-pe and QIIME-se). The
306 commands used were *qiime cutadapt*, *qiime dereplicate-sequences* and then we performed an open-reference
307 clustering using the *qiime vsearch cluster-features-open-reference* command to build OTUs with the parameter
308 *--p-perc-identity 0.99*. Chimera were removed with *vsearch uchime-denovo* command and the taxonomic
309 affiliation was done with *qiime feature-classifier classify-sklearn*.

310 For USEARCH, we followed the instructions provided by the author on his website
311 (https://www.drive5.com/usearch/manual/global_trimming_its.html) by taking into account merged
312 sequences and 5' R1 reads of non-overlapping reads and used successively the parameters *-fastq_mergepairs*,
313 *-search_oligodb*, *-fastq_filter*, *-fastx_uniques*, *-unoise3* and *-otutab* to produce ZOTUs and *-sintax* for
314 taxonomic affiliation.

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

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

325 Different metrics were computed in order to compare the above-mentioned methods: (i) the divergence
326 rate, computed as the Bray-Curtis distance between expected and observed abundance profiles at the species
327 level; (ii) the number of false-negative taxa (FN) corresponding to the number of expected taxa that were not
328 recovered by the method, (iii) the number of false positive taxa (FP) corresponding to the number of recovered
329 taxa that were not expected, (iv) the number of true positive taxa (TP) corresponding to the number of
330 recovered taxa that were expected. From these metrics we computed the precision (TP/(TP+FP)) and the recall

331 rate (TP/(TP+FN)). Finally, as we knew the exact expected sequences, we computed the number of sequences
332 perfectly identified (OTUs/ASVs/ZOTUs with nucleic sequence was strictly identical to the known reference
333 sequence). For long sequences (i.e. > 500 bp), the middle was not sequenced and only the sequenced part was
334 used for taxonomic affiliation. In this case, 100% identity between the reference and the OTU/ASV/ZOTU
335 resulted in a perfect identification.

336 *Analysis of real samples*

337 DADA2_FROGS, the bioinformatics approach with the best results from mock samples, was used to analyze
338 real samples (Code and figures are available on Recherche Data Gouv platform,
339 <https://doi.org/10.57745/ENE09G>). For the taxonomic affiliation of these samples (composition unknown),
340 and for each marker, we added the 118 sequences from our mock communities to Unite (v. 9,0) (Rolf Henrik
341 Nilsson et al., 2019) for ITS data and SILVA (v. 138) (Quast et al., 2013) 28S rDNA sequences for D1/D2. For
342 RPB2, we needed to build an in-house database because no dedicated one was publicly available to the best
343 of our knowledge. We first extracted sequences from the “Fungi” division from NCBI nt databank (release
344 2021-07-30) (Sayers et al., 2022) using taxonkit (v. 0.6.0) (Shen and Ren, 2021) and then used cutadapt (Martin,
345 2011) with RPB2 primers to target sequences of interest. The databases were composed of 206,184 ITS
346 sequences, 16,293 D1/D2 sequences and 13,055 RPB2 sequences.

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

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

354 **Results**

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

359 **Choice of the most accurate bioinformatic approach**

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

362 **Recall rate**

363 The median recall rate (sensitivity), reflecting the capacity of tools to detect expected species, is between
364 0.75 and 0.8 for FROGS, DADA2_FROGS, USEARCH, QIIME-se and DADA2-se. It is lower (~0.5) for QIIME-pe and
365 DADA2-pe, due to the fact that these methods always reject reads if they do not overlap (all D1/D2 and RPB2
366 sequences and some ITS1/2 were always missing).

367 **Precision**

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

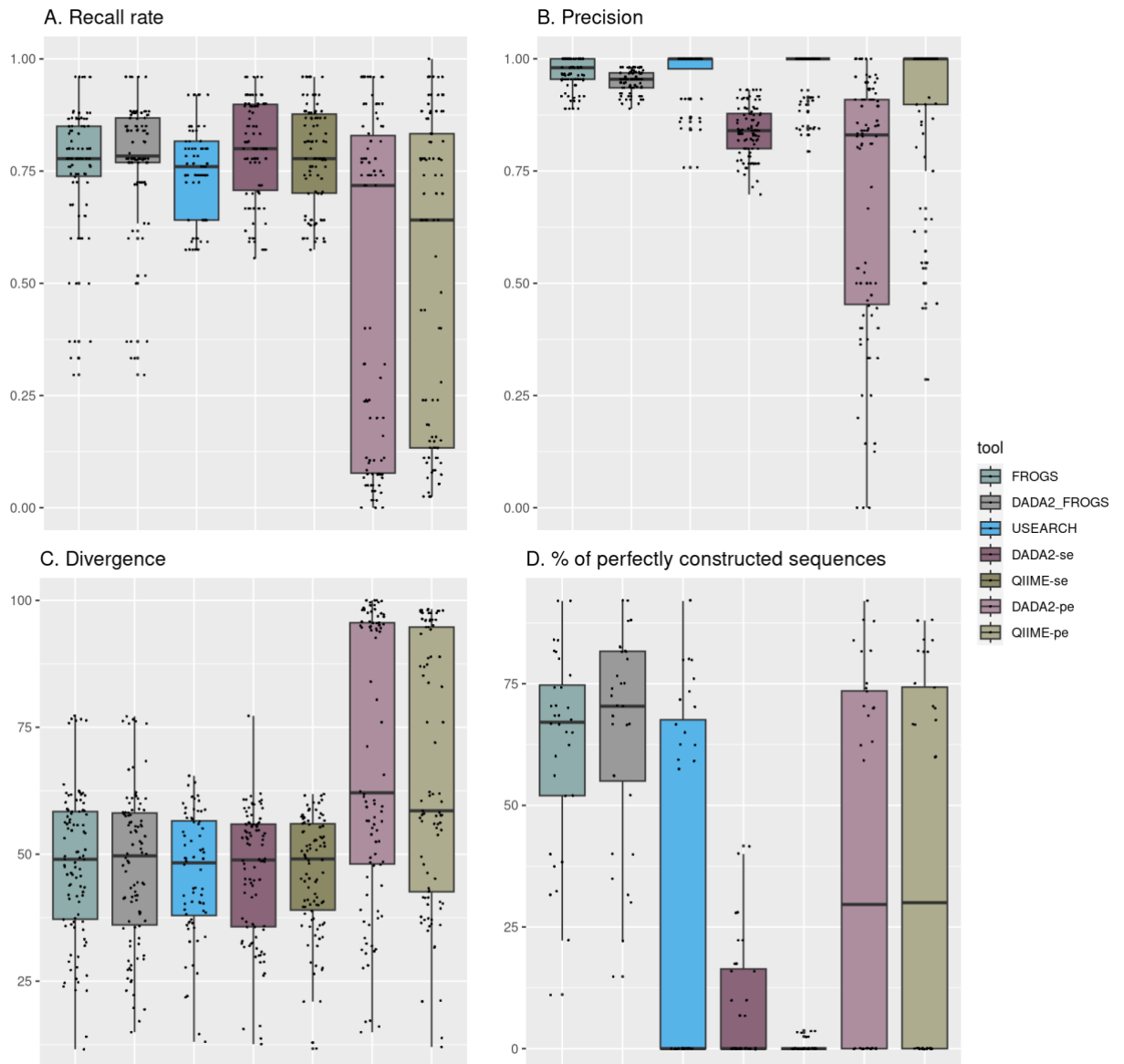
370 **Divergence rate**

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

376 **Reconstruction of sequences**

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

380
381
382
383



384

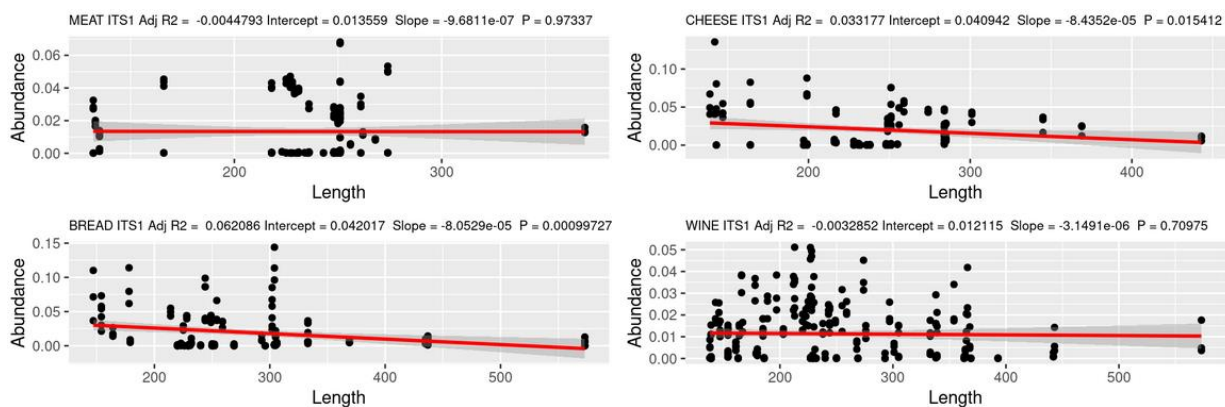
385 **Figure 5** - Quality parameters obtained with the seven bioinformatics pipelines. A) Recall rate ($TP/(TP+FN)$) reflects the
386 capacity of the tools to detect expected species. B) Precision ($TP/(TP+FP)$) shows the fraction of relevant species among the
387 retrieved species. C) Divergence rate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

390
391 Overall, the results obtained with the four indicators reveal that the DADA2_FROGS approach performs
392 the best for analyzing ITS1, ITS2, D1/D2 and RPB2 mock samples. We thus selected this approach for all
393 subsequent analysis. Nevertheless, it should be noted that the FROGS tool also performs well as it yields
394 indicator values that are similar to DADA2_FROGS. The main difference is due to species harboring very similar
395 sequences, such as those belonging to *Penicillium* spp.
396

397 **Effect of amplicon length on the detected relative abundance**

398 The ITS1 and ITS2 amplicon size is highly variable depending on the considered fungal species, as observed
399 for those included in our mock communities (Figure 2). The effect of amplicon size on the relative abundance
400 of the different species was evaluated using the PCR mock dataset (Figure 6) (code and figures are available
401 on Recherche Data Gouv platform: <https://doi.org/10.57745/APNOH8>).
402
403

PCR



404
405 **Figure 6** - Effect of ITS1 amplicon size on the relative abundance of the It is rather likely that all primers have mismatches
406 with certain groups of fungi.detected ASVs in the PCR mock datasets.

407 A significant negative relationship is observed between amplicon length and relative abundance in two out
408 of four tested mock communities for ITS1 (cheese and bread, but not meat and wine) and ITS2 (meat and wine,
409 but not cheese and bread). Furthermore, the determination coefficient (R2), which indicates the proportion of
410 variation in the relative abundance data that is predictable from the amplicon length, is comprised between
411 0.013 and 0.085. This parameter therefore only has a limited impact on the observed proportions when using
412 ITS1 and ITS2 as barcode markers.
413

414 **Comparison of markers**

415 To compare the capacity of each marker to correctly reflect the fungal diversity present in fermented food
416 samples, we only focused on the results obtained with the DADA2_FROGS approach. Code and figures are

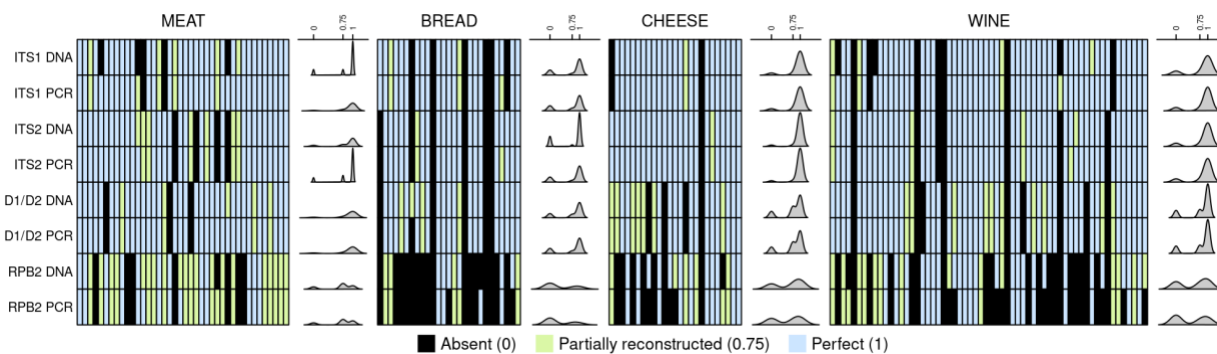
417 available on Recherche Data Gouv platform (<https://doi.org/10.57745/X6AXA6>). Figure 7 shows the
 418 divergence rate obtained for each marker and mock community type (DNA or PCR mock communities).
 419



420

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

423 Figure 8 shows the amount of absent, partially and perfectly reconstructed sequences for each ecosystem.
 424



425

426 **Figure 8** - Heatmap of expected species for each mock community and barcode marker using the DADA2_FROGS approach.
 427 Dark green represents perfectly reconstructed sequences (score of 1), light green partially reconstructed sequences (score of
 428 0.75) and orange species that are missing (0). Density of results is indicated for each line, representing the ability of markers to
 429 be efficient.
 430

431 *Bread*

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

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

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

444 *Cheese*

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

452 *Fermented meat*

453 Regarding the fermented meat mock community, all tested markers show comparable performance in
454 terms of divergence, ranging from 52 to 56% and 23 to 35% for the mock DNA and mock PCR mixture,
455 respectively. As observed for the cheese mock community, divergence is 1.5-2 times higher for the mock DNA
456 mixture than for the mock PCR mixture, the lowest divergence being observed with ITS1 marker in the mock
457 PCR mixture (23% divergence). D1/D2 exhibits the highest number of false negatives (n=8) as compared to
458 other markers while for PCR mock community mixture, ITS1 and RPB2 exhibit 4 and 5 false negatives,
459 respectively. Concerning the true positive metric, all markers perform equally well with between 32 and 33
460 true positives out of 40 expected species for both mock DNA and PCR mixtures with the exception of ITS1
461 marker in the mock PCR mixture which yields 36 true positives. Based on the above mentioned results, we
462 conclude that the ITS1 barcode is slightly more accurate for profiling fermented meat fungal communities,
463 although ITS2 and RPB2 also performed well.

464 *Wine*

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

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

472

473 **Analysis of real samples**

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

477 *Bread*

478 Metabarcoding results from the wheat sourdough sample analyses showed that hits with identities above
479 80% were not detected using the RPB2 marker. Besides fungal DNA, the other three markers amplified plant
480 DNA. The number of plant DNA hits was much higher for ITS2 and D1/D2 than ITS1. Moreover, ITS2 and D1/D2
481 markers amplified DNA from *Triticum* species (*Triticum aestivum*, *Triticum monococcum*, *Triticum durum*), and
482 crop weeds, such as *Viciae* sp., *Gallium* sp. and *Calystegia* sp., often found in cereal fields. The ITSx tool
483 automatically removed plant-derived ASVs in the final ITS1 and ITS2 ASV table whereas those from the D1D2
484 dataset had to be manually removed.

485 Regarding filamentous fungi, several genera were not detected with D1/D2 including *Aspergillus*,
486 *Aureobasidium* and *Tilletia*. Regarding mycotoxin-producing wheat pathogens (e.g., *Fusarium* and *Penicillium*

487 spp.) or species involved in negatively impacting grain and flour quality for bread making (e.g., rotten fish smell
488 due to *Tilletia* sp.), results showed that the ITS1 barcode did not detect *Penicillium* spp. contrary to D1/D2 and
489 ITS2 barcode markers. On the other hand, ITS1 allowed a better resolution to the species level within the
490 *Fusarium* and *Tilletia* genus.

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

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

499 *Cheese*

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

516 *Fermented meat*

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

530 Wine

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

534 When comparing the main species encountered, striking differences were observed between ITS1, ITS2
535 and D1D2 results. *Aspergillus* and *Aureobasidium* were detected in all samples, although with large differences
536 according to barcode. Nearly 50% was identified in the triplicates of sample 1 using ITS2 while only 25% with
537 ITS1 and about 40% with D1D2 (higher abundances of *Aspergillus* than *Aureobasidium* were also noted). An
538 opposite situation was observed for *Botrytis*. Indeed, this genus was much more abundant using ITS1 versus
539 ITS2 and D1D2. In agreement with mock analysis results, *Starmerella bacillaris*, a major component of grape
540 must microbiota, was detected at very low abundance in the different samples with ITS1, whereas this species
541 was detected in all samples with ITS2 and D1D2, and represented up to 40% in sample 2 according to ITS2 read
542 counts. In a similar manner, *Hanseniaspora uvarum* was poorly detected with ITS1, well detected with ITS2
543 and most abundant with D1D2. In contrast, *Metschnikowia* species were detected in a similar manner with
544 ITS1 and ITS2 (more than 40% abundance in samples 3.1 and 3.2) despite the short length and the
545 polymorphism of the amplified sequences.

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

547

Discussion

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

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

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

575 limitation, it would be worth designing new consensus primers suitable for Basidiomycota, Ascomycota and
576 Mucoromycotina. Alternative protein-coding genes also need to be tested.

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

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

594 The ITS1 and ITS2 markers performed better than D1D2 but their performance varied according to the
595 tested fermented foods. Indeed, while ITS2 performed better for cheese, meat and wine, ITS1 seems better
596 for bread. Concerning the latter, ITS2 primers amplified the ITS from wheat and several weeds which hampered
597 its efficiency for cereal-based products. Besides the fermented product being studied, the choice of the ITS
598 barcode also depends on the expected species diversity in the ecosystem and may be driven by fungal species
599 of interest. Moreover, choice of ITS primers can also be adapted to targeted species. Indeed, some ITS
600 barcoding primers may have mismatches with the sequence of fungal species of interest, such as *Yarrowia*
601 species (Ihrmark et al. 2012, Tedersoo and Lindahl, 2016). Finally, although ITS1 and ITS2 seem to be the best
602 barcodes for distinguishing between species and, according to our results, their variation in size does not
603 appear to introduce a large bias, their difference in size may hinder sequence alignment and therefore beta
604 diversity estimates that take phylogenetic distances into account.

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

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

616
617 In conclusion, although ITS2 appears as the most accurate barcode marker for fermented meat, cheese
618 and wine samples and ITS1 for sourdough bread, no generic recommendation for all fermented food types can
619 be made. This is mainly due to the fact that taxonomic resolution within some genera is not efficient which
620 highlights the need to combine metabarcoding with culture-dependent analysis such as culturomic
621 approaches. The availability of long-read technologies, like Oxford Nanopore Technologies or PacBio

622 technology, provides the opportunity to sequence longer fragments of the fungal ribosomal operon, up to 6
623 Kb (18S-ITS1-5.8S-ITS2-28S) and to improve the taxonomy assignment of the communities up to species level
624 (D'Andrea et al., 2021) but their current cost is still a brake to replace short reads technologies. Shotgun
625 metagenomics sequencing is also an alternative or a complementary method to study food fermentations
626 (Leech et al., 2020). It may provide a less biased vision of food microbiota than metabarcoding (Sternes et al.,
627 2017), a more comprehensive insight into the microbial composition, and functional potential but at a much
628 higher cost for low abundant species.

629 **Acknowledgements**

630 We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics
631 Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources. We thank
632 Julien Lebrat, Stéphane Guezenc, Anne-Sophie Sarthou for technical assistance and Claire Vincent for her
633 participation in building the reference databank. We also thank Mary and Loulou Bonneau, Stéphane Marrou,
634 Anna and Maximilien for providing security for their sourdough.

635 **Data, scripts, code, and supplementary information availability**

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

641

642 **Conflict of interest disclosure**

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

645 **Funding**

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

648 **References**

- 649 Anslan, S., Nilsson, R.H., Wurzbacher, C., Baldrian, P., Tedersoo, L., Bahram, M., 2018. Great differences in
650 performance and outcome of high-throughput sequencing data analysis platforms for fungal
651 metabarcoding. *MycKeys* 39, 29–40. <https://doi.org/10.3897/mycokeys.39.28109>
- 652 Bernard, M., Rué, O., Mariadassou, M., Pascal, G., 2021. FROGS: a powerful tool to analyse the diversity of
653 fungi with special management of internal transcribed spacers. *Briefings in Bioinformatics* 22, bbab318.
654 <https://doi.org/10.1093/bib/bbab318>
- 655 Berni, E., 2014. Molds, in: *Handbook of Fermented Meat and Poultry*. John Wiley & Sons, Ltd, pp. 147–153.
656 <https://doi.org/10.1002/9781118522653.ch17>

- 657 Bokulich, N.A., Mills, D.A., 2013. Improved Selection of Internal Transcribed Spacer-Specific Primers Enables
658 Quantitative, Ultra-High-Throughput Profiling of Fungal Communities. *Applied and Environmental*
659 *Microbiology* 79, 2519–2526. <https://doi.org/10.1128/AEM.03870-12>
- 660 Bokulich, N.A., Ohta, M., Richardson, P.M., Mills, D.A., 2013. Monitoring Seasonal Changes in Winery-Resident
661 Microbiota. *PLoS One* 8:e66437. <https://doi.org/10.1371/journal.pone.0066437>
- 662 Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J.,
663 Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T.,
664 Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C.,
665 Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M.,
666 Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H.,
667 Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R.,
668 Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciulek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-
669 X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V.,
670 Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B.,
671 Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S.,
672 Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson,
673 L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hoof, J.J.J., Vargas, F., Vázquez-
674 Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C.,
675 Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019.
676 Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature*
677 *Biotechnology* 37, 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- 678 Brandon-Mong, G.-J., Gan, H.-M., Sing, K.-W., Lee, P.-S., Lim, P.-E., Wilson, J.-J., 2015. DNA metabarcoding of
679 insects and allies: an evaluation of primers and pipelines. *Bulletin of Entomological Research* 105, 717–727.
680 <https://doi.org/10.1017/S0007485315000681>
- 681 Brown, S.P., Rigdon-Huss, A.R., Jumpponen, A., 2014. Analyses of ITS and LSU gene regions provide congruent
682 results on fungal community responses. *Fungal Ecology* 9, 65–68.
683 <https://doi.org/10.1016/j.funeco.2014.02.002>
- 684 Callahan, B.J., McMurdie, P.J., Holmes, S.P., 2017. Exact sequence variants should replace operational
685 taxonomic units in marker-gene data analysis. *ISME J* 11, 2639–2643.
686 <https://doi.org/10.1038/ismej.2017.119>
- 687 Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P., 2016. DADA2: High-
688 resolution sample inference from Illumina amplicon data. *Nature Methods* 13, 581–583.
689 <https://doi.org/10.1038/nmeth.3869>
- 690 Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser,
691 L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G., Knight, R., 2012. Ultra-high-throughput microbial
692 community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6, 1621–1624.
693 <https://doi.org/10.1038/ismej.2012.8>
- 694 Caruso, V., Song, X., Asquith, M., Karstens, L., 2019. Performance of Microbiome Sequence Inference Methods
695 in Environments with Varying Biomass. *mSystems* 4, e00163-18. <https://doi.org/10.1128/mSystems.00163-18>
- 697 Coton, M., Deniel, F., Mounier, J., Joubrel, R., Robieu, E., Pawtowski, A., Jeuge, S., Taminiau, B., Daube, G.,
698 Coton, E., Frémaux, B., 2021. Microbial Ecology of French Dry Fermented Sausages and Mycotoxin Risk
699 Evaluation During Storage. *Frontiers in Microbiology* 12.
- 700 D’Andrea, S., Cuscó, A., Francino, O., 2021. Rapid and real-time identification of fungi up to species level
701 with long amplicon nanopore sequencing from clinical samples. *Biology Methods and Protocols* 6, bpaa026.
702 <https://doi.org/10.1093/biomethods/bpaa026>

- 703 De Filippis, F., Laiola, M., Blaiotta, G., Ercolini, D., 2017. Different Amplicon Targets for Sequencing-Based
704 Studies of Fungal Diversity. *Appl Environ Microbiol* 83, e00905-17. <https://doi.org/10.1128/AEM.00905-17>
- 705 Dugat-Bony, E., Straub, C., Teissandier, A., Onésime, D., Loux, V., Monnet, C., Irlinger, F., Landaud, S., Leclercq-
706 Perlat, M.-N., Bento, P., Fraud, S., Gibrat, J.-F., Aubert, J., Fer, F., Guédon, E., Pons, N., Kennedy, S.,
707 Beckerich, J.-M., Swennen, D., Bonnarne, P., 2015. Overview of a surface-ripened cheese community
708 functioning by meta-omics analyses. *PLoS ONE* 10, e0124360.
709 <https://doi.org/10.1371/journal.pone.0124360>
- 710 Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461.
711 <https://doi.org/10.1093/bioinformatics/btq461>
- 712 Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G.,
713 Combes, S., Pascal, G., 2018. FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics* 34, 1287–
714 1294. <https://doi.org/10.1093/bioinformatics/btx791>
- 715 Floren, C., Wiedemann, I., Brenig, B., Schütz, E., Beck, J., 2015. Species identification and quantification in meat
716 and meat products using droplet digital PCR (ddPCR). *Food Chemistry* 173, 1054–1058.
717 <https://doi.org/10.1016/j.foodchem.2014.10.138>
- 718 Franciosa, I., Coton, M., Ferrocino, I., Corvaglia, M.R., Poirier, E., Jany, J.-L., Rantsiou, K., Cocolin, L., Mounier,
719 J., 2021. Mycobiota dynamics and mycotoxin detection in PGI Salame Piemonte. *Journal of Applied*
720 *Microbiology* 131, 2336–2350. <https://doi.org/10.1111/jam.15114>
- 721 Gardes, M., Bruns, T.D., 1993. ITS primers with enhanced specificity for basidiomycetes--application to the
722 identification of mycorrhizae and rusts. *Mol. Ecol.* 2, 113–118.
- 723 Garofalo, C., Tristezza, M., Grieco, F., Spano, G., Capozzi, V., 2016. From grape berries to wine: population
724 dynamics of cultivable yeasts associated to “Nero di Troia” autochthonous grape cultivar. *World J Microbiol*
725 *Biotechnol* 32, 59. <https://doi.org/10.1007/s11274-016-2017-4>
- 726 Glassman, S.I., Martiny, J.B.H., 2018. BROADSCALE Ecological Patterns Are Robust to Use of Exact Sequence
727 Variants versus Operational Taxonomic Units. *mSphere* 3, e00148-18.
728 <https://doi.org/10.1128/mSphere.00148-18>
- 729 Gweon, H.S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D.S., Griffiths, R.I., Schonrogge, K., 2015. PIPITS:
730 an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina
731 sequencing platform. *Methods in Ecology and Evolution* 6, 973–980. <https://doi.org/10.1111/2041-210X.12399>
- 732
- 733 Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J.,
734 Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D., 2012. New primers to amplify the fungal ITS2
735 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology*
736 82, 666–677. <https://doi.org/10.1111/j.1574-6941.2012.01437.x>
- 737 Irlinger, F., Layec, S., Hélinck, S., Dugat-Bony, E., 2015. Cheese rind microbial communities: diversity,
738 composition and origin. *FEMS microbiology letters*, 362(2), 1-11. <https://doi.org/10.1093/femsle/fnu015>
- 739 Jolly, N.P., Augustyn, O.P.H., Pretorius, I.S., 2003. The Occurrence of Non-Saccharomyces cerevisiae Yeast
740 Species Over Three Vintages in Four Vineyards and Grape Musts From Four Production Regions of the
741 Western Cape, South Africa. *South African J. Enol. Vitic.* 24:8–10. <https://doi.org/10.21548/24-2-2640>
- 742 Katoh, K., Asimenos, G., Toh, H., 2009. Multiple Alignment of DNA Sequences with MAFFT, in: Posada, D. (Ed.),
743 *Bioinformatics for DNA Sequence Analysis, Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp.
744 39–64. https://doi.org/10.1007/978-1-59745-251-9_3
- 745 Lavrinienko, A., Jernfors, T., Koskimäki, J.J., Pirttilä, A.M., Watts, P.C., 2021. Does Intraspecific Variation in rDNA
746 Copy Number Affect Analysis of Microbial Communities? *Trends in Microbiology* 29, 19–27.
747 <https://doi.org/10.1016/j.tim.2020.05.019>
- 748 Leech, J., Cabrera-Rubio, R., Walsh, A.M., Macori, G., Walsh, C.J., Barton, W., Finnegan, L., Crispie, F., O’Sullivan,
749 O., Claesson, M.J., Cotter, P.D., 2020. Fermented-Food Metagenomics Reveals Substrate-Associated

- 750 Differences in Taxonomy and Health-Associated and Antibiotic Resistance Determinants. *mSystems* 5,
751 e00522-20. <https://doi.org/10.1128/mSystems.00522-20>
- 752 Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T., Machida, R.J., 2013. A new
753 versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding
754 metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in Zoology* 10, 34.
755 <https://doi.org/10.1186/1742-9994-10-34>
- 756 Loos, D., Zhang, L., Beemelmans, C., Kurzai, O., Panagiotou, G., 2021. DANIEL: A User-Friendly Web Server for
757 Fungal ITS Amplicon Sequencing Data. *Frontiers in Microbiology* 12.
- 758 Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
759 *EMBnet.journal* 17, 10–12. <https://doi.org/10.14806/ej.17.1.200>
- 760 Matheny, P.B., 2005. Improving phylogenetic inference of mushrooms with RPB1 and RPB2 nucleotide
761 sequences (Inocybe; Agaricales). *Molecular Phylogenetics and Evolution* 35, 1–
762 20. <https://doi.org/10.1016/j.ympev.2004.11.014>
- 763 Montel, M. C., Buchin, S., Mallet, A., Delbes-Paus, C., Vuitton, D. A., Desmasures, N., Berthier, F., 2014.
764 Traditional cheeses: Rich and diverse microbiota with associated benefits. *International journal of food*
765 *microbiology*, 177, 136-154. <http://dx.doi.org/10.1016/j.ijfoodmicro.2014.02.019>
- 766 Mota-Gutierrez, J., Ferrocino, I., Rantsiou, K., Cocolin, L., 2019. Metataxonomic comparison between internal
767 transcribed spacer and 26S ribosomal large subunit (LSU) rDNA gene. *International Journal of Food*
768 *Microbiology* 290, 132–140. <https://doi.org/10.1016/j.ijfoodmicro.2018.10.010>
- 769 Nilsson, R. Henrik, Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., Tedersoo, L., 2019. Mycobiome
770 diversity: high-throughput sequencing and identification of fungi. *Nat Rev Microbiol* 17, 95–109.
771 <https://doi.org/10.1038/s41579-018-0116-y>
- 772 Nilsson, Rolf Henrik, Larsson, K.-H., Taylor, A.F.S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P.,
773 Picard, K., Glöckner, F.O., Tedersoo, L., Saar, I., Kõljalg, U., Abarenkov, K., 2019. The UNITE database for
774 molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids*
775 *Research* 47, D259–D264. <https://doi.org/10.1093/nar/gky1022>
- 776 O'Donnell, K., 1993. *Fusarium and its near relatives*, in: Reynolds, D., Taylor, J. (Eds.), . Presented at the The
777 fungal holomorph: mitotic, meiotic and pleomorphic speciation in fungal systematics, pp. 225–233.
- 778 Özkurt, E., Fritscher, J., Soranzo, N., Ng, D.Y.K., Davey, R.P., Bahram, M., Hildebrand, F., 2022. LotuS2: an
779 ultrafast and highly accurate tool for amplicon sequencing analysis. *Microbiome* 10, 176.
780 <https://doi.org/10.1186/s40168-022-01365-1>
- 781 Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large
782 Alignments. *PLoS ONE* 5, e9490. <https://doi.org/10.1371/journal.pone.0009490>
- 783 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA
784 ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*
785 41, D590–D596. <https://doi.org/10.1093/nar/gks1219>
- 786 Rossouw, D., Bauer, F., 2016. Exploring the phenotypic space of non-Saccharomyces wine yeast biodiversity.
787 *Food Microbiol.* 55:32–46. <http://dx.doi.org/10.1016/j.fm.2015.11.017>
- 788 Rué, Olivier, 2022a, "METABARFOOD - Description of the sequences used in mock communities",
789 <https://doi.org/10.57745/AZNJFE>, Recherche Data Gouv, V1
- 790 Rué, Olivier, 2022b, "METABARFOOD - Results on mock communities, choice of the bioinformatics solution",
791 <https://doi.org/10.57745/109NNP>, Recherche Data Gouv, V1
- 792 Rué, Olivier, 2022c, "METABARFOOD - Results on real communities", <https://doi.org/10.57745/ENE09G>,
793 Recherche Data Gouv, V1
- 794 Rué, Olivier, 2022d, "METABARFOOD - Comparison of amplicons", <https://doi.org/10.57745/X6AXA6>,
795 Recherche Data Gouv, V1

- 796 Rué, Olivier, 2022e, "METABARFOOD - Complementary analyses", <https://doi.org/10.57745/APNOH8>,
797 Recherche Data Gouv, V1
- 798 Sayers, E.W., Bolton, E.E., Brister, J.R., Canese, K., Chan, J., Comeau, D.C., Connor, R., Funk, K., Kelly, C., Kim, S.,
799 Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L.,
800 Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B.W., Pruitt, K.D., Sherry, S.T., 2022. Database
801 resources of the national center for biotechnology information. *Nucleic Acids Research* 50, D20–D26.
802 <https://doi.org/10.1093/nar/gkab1112>
- 803 Schliep, K.P., 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27, 592–593.
804 <https://doi.org/10.1093/bioinformatics/btq706>
- 805 Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Consortium, F.B.,
806 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,
807 F.-Y., Barreto, R.W., Begerow, D., Bergeron, M.-J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N.,
808 Burgaz, A.R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P.,
809 Cummings, C., Damm, U., Beer, Z.W. de, Hoog, G.S. de, Del-Prado, R., Dentinger, B., Diéguez-Uribeondo, J.,
810 Divakar, P.K., Douglas, B., Dueñas, M., Duong, T.A., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova,
811 K., Furtado, M., García, M.A., Ge, Z.-W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M.,
812 Grube, M., Gryzenhout, M., Guo, L.-D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P.,
813 Heller, G., Herrera, C., Hirayama, K., Hirooka, Y., Ho, H.-M., Hoffmann, K., Hofstetter, V., Högnabba, F.,
814 Hollingsworth, P.M., Hong, S.-B., Hosaka, K., Houbraeken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T.,
815 Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B.G., Kelly, L.J., Kirk, P.M., Knapp, D.G., Kõljalg, U.,
816 Kovács, G.M., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Ligginstoffer, A.S., Liimatainen, K., Lombard, L.,
817 Luangsa-ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S.N., Martin, M.P., May, T.W.,
818 McTaggart, A.R., Methven, A.S., Meyer, W., Moncalvo, J.-M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H.,
819 Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas,
820 R., Quaedvlieg, W., Raja, H.A., Redecker, D., Rintoul, T.L., Ruibal, C., Sarmiento-Ramírez, J.M., Schmitt, I.,
821 Schüßler, A., Shearer, C., Sotome, K., Stefani, F.O.P., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S.,
822 Suh, S.-O., Sung, G.-H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W.A.,
823 Urbina, H., Vágvölgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.-M., Wang, Y., Weir, B.S., Weiß, M., White,
824 M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.-C., Zhang, N., Zhuang, W.-Y., Schindel, D., 2012.
825 Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi.
826 *PNAS* 109, 6241–6246. <https://doi.org/10.1073/pnas.1117018109>
- 827 Selgas, M. d., García, M. I., 2014. Yeasts, in: *Handbook of Fermented Meat and Poultry*. John Wiley & Sons, Ltd,
828 pp. 139–146. <https://doi.org/10.1002/9781118522653.ch16>
- 829 Setati, M.E., Jacobson, D., Andong, U.C., Bauer, F., 2012. The Vineyard Yeast Microbiome, a Mixed Model
830 Microbial Map. *PLoS One* 7:e52609. [https://doi.org/10.1371/annotation/b9d307d9-f5c1-4e0d-8945-](https://doi.org/10.1371/annotation/b9d307d9-f5c1-4e0d-8945-c5a747b6f58e)
831 [c5a747b6f58e](https://doi.org/10.1371/annotation/b9d307d9-f5c1-4e0d-8945-c5a747b6f58e)
- 832 Setati, M.E., Jacobson, D., Bauer, F., 2015. Sequence-based analysis of the *Vitis vinifera* L. cv cabernet
833 sauvignon grape must mycobiome in three South African vineyards employing distinct agronomic systems.
834 *Front. Microbiol.* 6:1–12. <https://doi.org/10.3389/fmicb.2015.01358>
- 835 Shen, W., Ren, H., 2021. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *Journal of Genetics and*
836 *Genomics*, Special issue on Microbiome 48, 844–850. <https://doi.org/10.1016/j.jgg.2021.03.006>
- 837 Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for
838 environmental DNA research. *Molecular Ecology* 21, 1794–1805. [https://doi.org/10.1111/j.1365-](https://doi.org/10.1111/j.1365-294X.2012.05538.x)
839 [294X.2012.05538.x](https://doi.org/10.1111/j.1365-294X.2012.05538.x)
- 840 Sternes, P.R., Lee, D., Kutyna, D.R., Borneman, A.R., 2017. A combined meta-barcoding and shotgun
841 metagenomic analysis of spontaneous wine fermentation. *GigaScience* 6, gix040.
842 <https://doi.org/10.1093/gigascience/gix040>

- 843 Tedersoo, L., Lindahl, B., 2016. Fungal identification biases in microbiome projects. *Environmental*
844 *microbiology reports*, 8(5), 774-779. <https://doi.org/10.1111/1758-2229.12438>
- 845 Toju, H., Tanabe, A.S., Yamamoto, S., Sato, H., 2012. High-Coverage ITS Primers for the DNA-Based
846 Identification of Ascomycetes and Basidiomycetes in Environmental Samples. *PLoS One* 7.
847 <https://doi.org/10.1371/journal.pone.0040863>
- 848 Větrovský, T., Kolařík, M., Žifčáková, L., Zelenka, T., Baldrian, P., 2016. The rpb2 gene represents a viable
849 alternative molecular marker for the analysis of environmental fungal communities. *Mol Ecol Resour* 16,
850 388–401. <https://doi.org/10.1111/1755-0998.12456>
- 851 von Gastrow, L., Gianotti, A., Vernocchi, P., Serrazanetti, D.I., Sicard, D., 2023. Taxonomy, Biodiversity, and
852 Physiology of Sourdough Yeasts. In: Gobbetti, M., Gänzle, M. (eds) *Handbook on Sourdough Biotechnology*.
853 Springer, Cham. https://doi.org/10.1007/978-3-031-23084-4_7
- 854 von Gastrow, L., Michel, E., Legrand, J., Amelot, R., Segond, D., Guezenc, S., Rué, O., Chable, V., Goldringer, I.,
855 Dousset, X., Serpolay-Bessoni, E., Taupier-Letage, B., Vindras-Fouillet, C., Onno, B., Valence, F., Sicard, D.,
856 2022. Microbial community dispersal from wheat grains to sourdoughs: A contribution of participatory
857 research. *Molecular Ecology* mec.16630. <https://doi.org/10.1111/mec.16630>
- 858 Vu, D., Groenewald, M., Szöke, S., Cardinali, G., Eberhardt, U., Stielow, B., de Vries, M., Verkleij, G.J.M., Crous,
859 P.W., Boekhout, T., Robert, V., 2016. DNA barcoding analysis of more than 9 000 yeast isolates contributes
860 to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology* 85, 91–105.
861 <https://doi.org/10.1016/j.simyco.2016.11.007>
- 862 White, T., Bruns, T., Lee, S., Taylor, J., 1990. Amplification and direct sequencing of fungal ribosomal RNA genes
863 for phylogenetics, in: Innis, M., Gelfand, D., Shinsky, J., White, T. (Eds.), *PCR Protocols: A Guide to Methods*
864 *and Applications*. Academic Press, pp. 315–322.
- 865 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
866 internal transcribed spacer (ITS), specifically ITS1 and ITS2, for the analysis of fungal diversity by deep
867 sequencing. *PLOS ONE* 13, e0206428. <https://doi.org/10.1371/journal.pone.0206428>
- 868 Zemb, O., Achard, C.S., Hamelin, J., De Almeida, M.-L., Gabinaud, B., Cauquil, L., Verschuren, L.M.G., Godon, J.-
869 J., 2020. Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of
870 relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard. *MicrobiologyOpen*
871 9, e977. <https://doi.org/10.1002/mbo3.977>
- 872 Zhang, J., Ding, X., Guan, R., Zhu, C., Xu, C., Zhu, B., Zhang, H., Xiong, Z., Xue, Y., Tu, J., Lu, Z., 2018. Evaluation
873 of different 16S rRNA gene V regions for exploring bacterial diversity in a eutrophic freshwater lake. *Science*
874 *of The Total Environment* 618, 1254–1267. <https://doi.org/10.1016/j.scitotenv.2017.09.228>
- 875 Zimmer-Faust, A.G., Steele, J.A., Xiong, X., Staley, C., Griffith, M., Sadowsky, M.J., Diaz, M., Griffith, J.F., 2021.
876 A Combined Digital PCR and Next Generation DNA-Sequencing Based Approach for Tracking Nearshore
877 Pollutant Dynamics Along the Southwest United States/Mexico Border. *Frontiers in Microbiology* 12.

