

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Metataxonomic comparison between internal transcribed spacer and 26S ribosomal large subunit (LSU) rDNA gene

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1683068> since 2019-12-30T13:57:45Z

Published version:

DOI:10.1016/j.ijfoodmicro.2018.10.010

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

**Metataxonomic comparison between internal transcribed spacer and 26S ribosomal large
subunit (LSU) rDNA gene**

Authors: Jatziri Mota-Gutierrez, Ilario Ferrocino, Kalliopi Rantsiou, Luca Cocolin*

Affiliations:

Department of Agricultural, Forest, and Food Science, University of Turin, Largo Paolo Braccini 2,
10095, Grugliasco, Torino, Italy

Corresponding Author:

Luca Cocolin, Ph.D., Department of Agricultural, Forest and Food Science of the University of
Turin (DISAFA), University of Turin, Grugliasco, Torino, 10095, Italy

E-mail: lucasimone.cocolin@unito.it

19 **Highlights**

- 20 - Primer selection plays critical roles in the sensitivity and tracking fungal communities to
- 21 assess reliable and accurate ecological populations
- 22 - The 26S target region exploited in rRNA sequencing demonstrated greater taxonomical
- 23 depth for fungal communities
- 24 - Preferential amplification phenomenon contributes to underestimations and overestimations
- 25 of fungal species
- 26 - The limited availability of updated databases to assess ecological populations.

27

28 **Abstract**

29 Next-generation sequencing has been used to strengthen knowledge about taxonomic diversity and
30 ecology of fungi within food ecosystems. However, primer amplification and identification bias
31 could edge our understanding into the fungal ecology. The aim of this study is to compare the
32 performance of two primer pairs over two nuclear ribosomal RNA (rRNA) regions of the fungal
33 kingdom, namely the ITS2 and 26S regions. Fermented cocoa beans were employed as biological
34 material and the fungal ecology during fermentation was studied using amplicon-based sequencing
35 tools, making use of a manually curated 26S database constructed in this study, and validated with
36 SILVA database. To explore potential biases introduced by PCR amplification of fungal
37 communities, a mock community of known composition was prepared and tested. The relative
38 abundances observed for ITS2 suggest that species with longer amplification fragments are
39 underestimated and concurrently species that render shorter amplification fragments are
40 overestimated. However, this correlation between amplicon length and estimation is not valid for all
41 the species analysed. Variability in the amplification lengths contributed to the preferential
42 amplification phenomenon. DNA extracted from twenty fermented cocoa bean samples were used
43 to assess the performance of the two target regions. Overall, the metataxonomic data set recovered
44 similar taxonomic composition and provided consistent results in OTU richness among biological
45 samples. However, 26S region provided higher alpha diversity index and greater fungal rRNA
46 taxonomic depth and robustness results compared with ITS2. Based on the results of this study we
47 suggest the use of the 26S region for targeting fungi. Furthermore, this study showed the efficacy of
48 the manually curated reference database optimized for annotation of mycobiota by using the 26S as
49 a gene target.

50 **Keywords:** Amplicon sequencing; fungal ecology; primer bias; Illumina; fungal database.

51 **1. Introduction**

52 Fungi are eukaryotic microorganisms which belong to one of the most diverse kingdoms on
53 Earth (Blackwell, 2011). They play an important role in the safety, quality, and stability of all
54 foodstuff to some degree, whether they are required during processing or whether they have a
55 negative impact during shelf life. Therefore, tracking fungal communities of food systems has been
56 a concern in food research. To date, most recent studies on the microbial diversity of fermented
57 food such as vegetable, seafood, beverages, cheese, olives and spontaneously fermented American
58 cool ship ale fermentations have employed amplicon sequencing approaches (Bokulich et al., 2012;
59 Cocolin et al., 2013; Ercolini et al., 2012; Li et al., 2011; Roh et al., 2010).

60 Illumina sequencing platform has been currently providing a sensitive description of the
61 microbial dynamics within food ecosystems. Some of the advantages of using this technology is
62 that it yields greater sequencing coverage and increased sample throughput at lower cost *per* sample
63 compared to other platforms (Caporaso et al., 2011; Quail et al., 2012). The sensitivity of this
64 approach relies on the high coverage and accurate taxonomic resolution of short amplicon length
65 (Quail et al., 2012). Recent advances in the microbial diversity using next-generation sequencing
66 technologies (NGS) have underlined the importance of the reliability of PCR primers targeting a
67 specific genetic marker (Bokulich and Mills, 2013). In spite of the importance of the amplification
68 of shorter fragments amplified by PCR in NGS, recent studies described a more reliable community
69 of fungi using shorter Internal Transcribed Spacer amplicons (ITS) of the nuclear ribosomal RNA
70 (rRNA) (Bokulich and Mills, 2013; Ihrmark et al., 2012).

71 The ITS region is considered the universal barcode for identification of fungi and includes
72 the ITS1 and ITS2 regions, separated by the 5.8S gene. These two regions (ITS1-2) are
73 characterized by high evolutionary rates and are edged by highly conserved regions with suitable
74 target sites for universal primers (Begerow et al., 2010). However, the complete ITS region located
75 between the 18S and 28S genes in the nuclear ribosomal RNA is considered too long for 454

76 sequencing or other NGS (Bellemain and Carlsen, 2010). Therefore, various primers are used to
77 amplify parts of the ITS region. In this study, we selected the primer ITS3ngs that targets a site in
78 the ITS1 and the degenerate reverse primer ITS4 which targets an ITS-flanking site in the ribosomal
79 large subunit (LSU) encoding regions (White TJ, Bruns T, Lee S, 1990) based on their ability to
80 amplify fungal species through *in silico* analysis (Bellemain and Carlsen, 2010; Tedersoo et al.,
81 2015). Nevertheless, the nuclear rRNA large subunit (LSU/28S/26S) and small subunit (SSU/18S)
82 genes have also been often used to address fungal diversity (Bonanomi et al., 2016; David et al.,
83 2014; De Filippis et al., 2017b, 2017a; Garofalo et al., 2015; Stellato et al., 2015; Wang et al.,
84 2015). To bring an overall perspective, most yeasts have been identified from sequence divergence
85 in the D1/D2 domain of LSU rRNA (Kurtzman and Robnett, 1997). Despite the great resolution to
86 recognized yeast species through 26S rRNA sequencing reactions, little is known about the
87 potential uses and bias that can be introduced when using this target region in NGS. In this context,
88 it is necessary for ecological studies to compare different targeting regions to describe the most
89 accurate and reliable ecological populations in a food system. Given the nature of current
90 challenges, the selection of a suitable genetic marker for the identification of fungi will help
91 researchers to clear current issues insight into the selection of primer sets.

92 The main focus of this study is to address sequencing target regions and primer biases on
93 one of the dominating taxonomic groups of fungi in the Dikarya, Ascomycota, which represents 53
94 % of the described species of true Fungi (Koljalg et al., 2013). This phylum is important in the food
95 industry and serves as a source for biomass production, but also includes known human and plant
96 pathogens (Bekatorou et al., 2006; Berbee, 2001). The present research focused on the assessment
97 of two different targeting sites for amplicon-based Illumina NGS studies. We tested the 26S primer
98 set, delivering high coverage and accurate taxonomic assignment of short (~ 400 bp) fungal
99 amplicon *versus* the performance with the ITS2 region. This research intends to bring new insights
100 in the field of taxonomic assignment, validation and resolution of uncertainties on using amplicon-
101 sequencing approaches for fungi identification by using mock samples as well as fermented

102 samples. Attention was paid for monitoring fungi in mock communities and biological samples,
103 where taxonomic assurance of the technique, and mapping and monitoring fungi dynamics are
104 investigated for food applications.

105 **2. Materials and methods**

106 *2.1. Primer selection and in silico analysis*

107 Primer pairs targeting the ITS2 region (Tedersoo et al., 2015), and the D1 domain of 26S
108 rRNA gene (Cocolin et al., 2000), were selected and reported in Table 1. For the amplification of
109 the D1 domain of the 26S, we modified the LS2-MF primer sequence position from reverse to
110 forward, corresponding to nucleotide position 266 of *Saccharomyces cerevisiae* 26S gene as
111 described by Cocolin *et al.*, (2000) and a reverse primer NL4 (Jespersen et al., 2005). The Illumina
112 overhang adapter sequences were added to locus-specific sequences. The D1 region from the 26S
113 gene was amplified *in silico* to compare primer specificity and taxonomic coverage of both LS2-MF
114 and NL4 by using Primer Prospector (Walters et al., 2011) against the constructed 26S databases
115 and SILVA's database.

116 *2.2. Mock community preparation, DNA extraction, and PCR amplification*

117 Strains of yeast and filamentous fungi listed in Table 2 (DISAFA collection, Torino) were
118 used and cultured on Malt Extract Agar (Oxoid, Milan, Italy) plus 25 mg l⁻¹ streptomycin (Sigma,
119 Milan, Italy) incubated at 28 °C for 72 h for yeast and 7 to 10 days for fungi. DNA extraction from
120 yeast was carried out from a loopful of grown culture while 250 mg of mycelium was scraped from
121 the plate for filamentous fungi. DNA extraction was carried out as described by Cocolin et al.,
122 (2000). DNA from each strain was quantified by using the Qubit dsDNA assay kit (Thermo Fisher
123 Scientific, Milan, Italy) and standardized at 5 ng/μl. A pool (Mock-DNA) containing each of the
124 standardized strain DNA was then obtained and subject to amplification of the ITS2 and the 26S
125 regions. PCR was carried out for the two target regions using a PCR mixture prepared with 12.5 μl

126 of the 2X Kapa HiFi HotStart ReadyMix Taq (Roche, Milan, Italy), 1 μ M each primer, 2.5 μ l of
127 DNA template, and PCR-grade water. Each PCR were subject to the following amplification
128 conditions: thirty cycles of 30 s of denaturation (95 $^{\circ}$ C), 30 s of primer annealing (55 $^{\circ}$ C), and 30 s
129 of primer elongation (72 $^{\circ}$ C), followed by a final elongation step (72 $^{\circ}$ C) of 10 min.

130 The amplification of each fungal strain was carried out by using the same couple of primers,
131 each amplicon was then purified using the Agencourt AMPure XP beads (Beckman Coulter
132 Genomics) and quantified using Qubit dsDNA assay kit. Based on the amplicon size of DNA
133 assessed by using a Biorad experion workstation (Biorad, Milan, Italy), amplicons concentration
134 was determined. Amplicons were diluted at 20 mM and aliquots of 10 μ l were pooled together to
135 construct a Mock-Amp. In total, two independent Mock-DNA and Mock-Amp were obtained by
136 two independent DNA extraction, quantification and pooling procedure.

137 2.3. DNA extraction and PCR amplification of fermented cocoa beans

138 A total of twenty fermented cocoa beans samples were collected and DNA extracted as
139 following original study (Mota-Gutierrez et al., 2018). Samples were collected during a
140 fermentation period of 0, 48, 96 and 120 h. Detailed information of samples is reported in Mota-
141 Gutierrez et al., 2018 and in supplementary table S1. Briefly, total DNA was extracted from the
142 pellet of cocoa matrix by using the MasterPure Complete DNA & RNA Purification kit (Illumina
143 Inc, San Diego, CA) following the manufacturer's instructions. DNA was quantified by using the
144 Qubit dsDNA assay kit (Thermo Fisher Scientific), standardized at 5 ng/ μ l and subject to
145 amplification of the two target regions using primers and procedure as described above.

146 2.4. Library preparation and sequencing

147 Sequencing was performed for the two target regions and for the three target samples
148 (Mock-DNA, Mock-Amp and cocoa samples). After the first purification step following the
149 Illumina sample preparation procedure, the library was combined with the sequencing adapters and

150 dual indices using the Nextera XT Index Kit (Illumina, San Diego, USA), obtaining the multiplexed
151 paired-end libraries. Individual libraries concentration in nM were calculated based on the size of
152 amplicons by using a Biorad Experion workstation (Biorad) and diluted to 4 nM, denaturated with
153 0.2 N NaOH and spiked with 20 % (v/v) of PhiX. The combination of pool library and PhiX were
154 diluted to 12 pM and paired-end sequencing was performed on the MiSeq platform, using MiSeq
155 Reagent Kit V3 (2 x 250bp) (Illumina, San Diego, USA), following the standard Illumina
156 sequencing protocol.

157 *2.5. Constructed 26S rRNA sequence database*

158 The construction database of fungal rRNA gene sequence of the 26S gene was used to select
159 primers, which amplify the D1 region of a broad fungal taxa. The sequences were downloaded from
160 the Nucleotide database of the National Center for Biotechnology Information (NCBI;
161 <https://www.ncbi.nlm.nih.gov/nucleotide/>; accessed March 07, 2018). The database was constructed
162 using the large subunit rRNA gene sequences, 23,381 sequences were downloaded using diverse
163 taxonomic ID and the query word “26S rRNA”. The final constructed database consisted of 4 phyla,
164 27 classes, 172 families, and 442 fungal strains. Incomplete sequences or sequences with absent
165 taxonomies were removed. Duplicate sequences and sequences that clustered together at 99 % of
166 similarity were discarded by using Prinseq and USEARCH respectively (Schmieder and Edwards,
167 2011). A taxonomy file, matching exactly seven taxonomic levels (root, subphylum, class, order,
168 family, genus and species) was generated from the corresponding taxonomy strings to be
169 compatible with implementation in the NGS analysis pipeline QIIME. Both files were manually
170 curated for accuracy and consistency. Sequences obtained from the constructed 26S database from
171 biological and mock samples were compared using SILVA database. All sequences identified by
172 D1 domain of 26S rRNA sequence analysis from biological samples and mock communities were
173 compared with our constructed database.

174 *2.6. Bioinformatics*

175 Paired-end reads (2x250 bp) were first merged using the FLASH software (Magoč and Salzberg,
176 2011), with default parameters. Joint reads were further quality filtered (Phred < Q20) using the
177 QIIME 1.9.0 software (Caporaso et al., 2011). Chimeras were then removed with the adopted
178 USEARCH version 8.1 software. Lastly, OTUs were picked at 99 % of similarity by means of
179 UCLUST clustering methods (Edgar, 2010) and representative sequences of each cluster were used
180 to assign taxonomy. For the 26S data, each cluster was used to assign taxonomy using the
181 Constructed 26S rRNA gene database and SILVA, while for the ITS dataset the UNITE rRNA ITS
182 database version 2012, by means of the RDP Classifier. Sequences were double-checked using the
183 BlastN search tool (<http://www.ncbi.nlm.nih.gov/blast/>) to confirm the taxonomy assignment.
184 Cocoa samples datasets (ITS and 26S) were rarefied at 10,018 reads after raw read quality filtering,
185 and OTU tables were filtered for OTUs occurring at 1 % of the relative abundance in at least 2
186 samples. While for mock community reads from the two target regions were rarefied at 17,313
187 reads.

188 2.7. Statistical analyses

189 Statistics and plotting were carried out in the R environment (www.r-project.org). Alpha
190 diversity indices were calculated using the diversity function of the vegan package (Dixon, 2007).
191 OTUs table were used to find differences between target regions by Anosim statistical test in R
192 environment. A two-sided permutation test with 999 permutations was performed to compare the
193 OTUs distribution and alpha diversity between the two datasets. Pairwise Kruskal-Wallis Wilcoxon
194 test or one way- ANOVA coupled with the Duncan honestly significant difference (HSD) test were
195 used as appropriate to determine significant differences in alpha diversity or OTU abundance from
196 mock communities and biological samples. Statistical analysis was acquired through the function
197 *aov* through the *stats* package and principal component analysis were plotted using the function
198 *dudi.pca* through the *made4* package using R version 3.3.2

199 2.8. Accession numbers

200 The ITS and 26S rRNA gene sequences are available at the Sequence Read Archive of the
201 National Center for Biotechnology Information (NCBI), under the SRA accession number
202 SRP126081 (fermented cocoa samples ITS) and SRP150401 (fermented cocoa samples 26S and
203 mock sequences data).

204

205 3. Results

206 3.1. *In silico* performance of 26S primers

207 We performed an *in-silico* analysis of the 26S primer set against our constructed database
208 and SILVA using Primer Prospector. LS2-MF primer showed the lowest weighted score (Fig. S1A)
209 indicating higher coverage across the database sequences and lower number of mismatches if
210 compared with NL4 (Fig. S1B). Comparing the taxonomic coverage of LS2-MF and NL4 against
211 *Zygomycota*, *Glomeromycota*, *Ascomycota* and *Basidiomycota* sequence, LS2-MF showed the best
212 performance with a coverage higher than 80 % for all the phyla except for *Glomeromycota* (Fig.
213 S2A), while NL4 account for the 20 % of the coverage against our constructed database (Fig. S2B).
214 Regarding the performance of the primer sets against the SILVA's database, the score of the
215 primers was higher compared with our database (data not shown).

216 3.2. *Performance of primers by mock community analysis*

217 A mock community containing twenty fungal species (Table 2) was prepared to validate the
218 performance of the two target regions. The possible effect of the bias introduced by PCR (Mock-
219 DNA) and that of sequencing (Mock-Amp) was then evaluated. Amplicon length of the single
220 species showed little variation when the 26S gene was amplified (461 ± 30 bp) while for ITS2 we
221 observed greater dispersion in size (445 ± 55 bp) (Table 2). Significant difference in mycobiota
222 composition (Anosim statistical test, $P < 0.05$) by using the two target regions or mock
223 communities (DNA or AMP) was observed by Principal Component Analysis (Fig. 1).

224 In both samples (Mock-DNA and Mock-Amp), the target region ITS2 showed similar
225 abundances with respect to the theoretical value for two fungal species, namely *Torulaspora*
226 *delbrueckii* and *Plectosphaerella cucumerina* (Table 3). Similarly, with the 26S target gene, for six
227 species, abundances retrieved from both Mock-DNA and Mock-Amp samples were comparable to
228 the theoretical values (*Aspergillus fumigatus*, *Pichia membranifaciens*, *Pichia kudriavzevii*,
229 *Penicillium glabrum*, *Penicillium brevicompactum* and *Starmerella bacillaris*). Furthermore, for the
230 26S region, the species *Alternaria alternata*, *Aspergillus flavus* and *Fusarium oxysporum* rendered
231 different abundances in the Mock-DNA and the Mock-Amp but in both samples the values were
232 comparable to the theoretical.

233 For 18 out of the 20 fungal species tested, the ITS2 region resulted in underestimation or
234 overestimation with respect to the theoretical value in the Mock-DNA, Mock-Amp or both ($P <$
235 0.05). Four species were significantly overestimated (*A. fumigatus* (439bp), *F. verticillioides*
236 (415bp), *K. marxianus* (521bp) and *P. brevicompactum* (428bp)) while other 4 were significantly
237 underestimated (*Galactomyces geotrichum* (324bp), *Hanseniaspora opuntiae* (484bp),
238 *Schizosaccharomyces pombe* (562bp) and *Starmerella bacillaris* (361bp)), in both samples.
239 Interestingly, *G. geotrichum* and *S. pombe* were not detected in any of the two samples. Nine more
240 species resulted to be significantly different from the theoretical value (either higher or lower) in the
241 Mock-DNA or Mock-Amp sample only (Table 3).

242 When the 26S region was targeted, 11 out of the 20 species were either underestimated or
243 overestimated in the Mock-DNA, Mock-Amp or both. More specifically, 4 species were
244 underestimated in both types of samples (*Candida sake* (467bp), *H. opuntiae* (435 bp),
245 *Kluyveromyces marxianus* (427 bp), *T. delbrueckii* (461 bp)) and only one (*G. geotrichum* (502bp))
246 was overestimated. Five more species were significantly different from the theoretical value (either
247 higher or lower) in the Mock-DNA or Mock-Amp sample only (Table 2 and 3). It also should be

248 pointed out that we did not observe a clear correlation between amplicon size and over or
249 underestimation.

250 In the Mock-Amp samples, correct relative quantification was obtained for 13 out of the 20
251 species targeting the 26S region and for 10 out of 20 species with the ITS2 region. In the Mock-
252 DNA samples, correct relative quantification was obtained for 10 out of 20 species in the 26S
253 region and for 4 out of 20 species targeting the ITS2 (Table 3).

254 Remarkably, *G. geotrichum* and *S. pombe* were only detected when the 26S region was
255 targeted while *H. opuntiae* was the only species that was consistently underestimated,
256 independently from the target region or sample. Overall, 26S sequencing data aligned better to
257 theoretical abundance values for the fungal species tested than did the ITS sequencing data.

258 3.3. Mycobiota in biological samples

259 Sequencing of twenty fermented cocoa beans samples collected during a previous
260 experiment, after amplification with the primers ITS2 and 26S showed a mean sequence length of
261 412 and 390 bp. respectively and an estimated sample coverage of 97.73 and 95.87 %, respectively
262 (See Table S1). The 26S target region revealed greater OTU richness compared to the ITS2 region
263 ($P < 0.05$) as shown in Fig. 2. Overall, 20 and 37 fungal OTUs were identified during the
264 fermentations using the primer set ITS2 and 26S, respectively. In addition, we observed differences
265 in length distribution across the two target genes. Histogram of reads length of 26S showed that the
266 higher reads proportion were around 380 bp while for ITS2 we observed a varied distribution of the
267 reads length around 370bp, 400bp, 420bp, and 450bp (See Fig. S3). Eleven OTUs, namely *Candida*
268 *jaroonii*, *Candida tallmaniae*, *Fusarium*, *Hanseniaspora*, *H. opuntiae*, *Hanseniaspora uvarum*, *K.*
269 *marxianus*, *S. cerevisiae*, *Saccharomycopsis crataegens*, *T. delbrueckii* and *Pichia pijperi* were
270 detected by both targeting regions (Fig. 3). The relative abundance of several fungal species was
271 significantly different according to the type of amplicon used ($P < 0.05$, Fig. 4), in which
272 significantly higher relative abundance was found for *S. cerevisiae*, *P. pijperi*, and *H. uvarum* ($P <$

0.05) using ITS2 target gene, while *Hanseniaspora* showed higher abundance when using the 26S
($P < 0.05$, Fig. 4).

3.4. Performance of the new constructed 26S database against SILVA

To validate the new 26S database, biological samples and mock communities identified by
D1 domain of 26S rRNA sequence analysis were compared with SILVA's database (Quast et al.,
2013). Significant difference in mycobiota composition of the two databases was observed in mock
communities (Anosim statistical test, $P < 0.05$). In detail, the constructed 26S database assigned
successfully the twenty fungal species, while SILVA's database assigned only ten (*A. alternata*, *H.*
osmophila, *K. marxianus*, *P. kudriavzevii*, *P. membranifaciens*, *S. ludwigii*, *S. pombe*, *S. bacillaris*,
T. delbrueckii, Table 4). Interestingly, *S. cerevisiae* was not detected by using the SILVA's
database from the mock communities. In contrast, no significant differences in the OTU
distributions and the alpha diversity calculations were observed between the two datasets from
biological samples. In depth, the constructed 26S database assigned 37 fungal species, described
above, while SILVA's database assigned 35 (data not shown). However, we observed that *S.*
cerevisiae and *H. uvarum* were not detected by using the SILVA's database.

4. Discussion

New tools and molecular techniques have been used to detect microbial ecology in the past
decades. Recently, the interest in the use of amplicon sequencing to identify taxonomically relevant
taxa in food has increased. However, this approach has potential biases as previously described
(Bowers et al., 2015; Fouhy et al., 2016) where primer selection is considered one of the most
important sources of biases (Bokulich et al., 2014; Bonanomi et al., 2016; David et al., 2014; De
Filippis et al., 2017a; Ercolini, 2013; Garofalo et al., 2015; Stellato et al., 2015; Stielow et al., 2015;
Wang et al., 2015). The 26S region (D1 domain) of the rRNA encoding gene and the ITS2 region
have been proposed as good candidates for identifying fungal species when using NGS technologies
due to the high taxonomic resolution (Tedersoo et al., 2015). In this study, we performed a

298 comparative evaluation of two regions as amplicon sequencing targets for the identification of fungi
299 and it also describes the mycobiota community in food matrices.

300 Recently the 26S region has been studied using the Roche 454 technology (De Filippis et al.,
301 2017b). However, this platform has been shown to result in high sequencing errors due to A and T
302 rich homopolymers (Luo et al., 2012), while Illumina does not present this sequencing error (Erich
303 et al., 2008). Our results and a previous study (De Filippis et al., 2017b) reveal that 26S gene is a
304 reliable target site for both NGS technologies (Roche 454 and Illumina) for eukaryotic species. In
305 order to evaluate the effect of the target gene used we compared the sequencing results of both
306 DNA samples and mock communities. In our study, different relative abundances were obtained for
307 both mock communities and biological samples and these differences were based on the PCR target
308 used. We observed in such cases that ITS2 target region led to underestimations of species with
309 longer fragments (*S. pombe*, and *H. opuntiae*), while an overestimation of shorter fragments
310 occurred (*F. vericillioides*, *A. fumigatus* and *P. brevicompactum*). However, it should be pointed out
311 that this correlation between amplicon length and estimation is not valid for all the species
312 analysed. Apart from amplification length, other parameters that influence relative abundance
313 calculations of taxa within samples could be considered. Sampling errors, different primers
314 alignment efficiencies during PCR amplification, performance of degenerate primers used during
315 PCR amplification, result in wrong representation in terms of relative abundance of microbial
316 populations (Ihrmark et al., 2012; Polz and Cavanaugh, 1998).

317 In addition to underestimation of abundances, “identification bias” is also common to
318 amplicon-based analyses, where minor groups are poorly represented (Koljalg et al., 2013). The
319 lack of updated reliable public reference data set and the discrepancies to refer to fungal species
320 have been recently demonstrated for the ITS sequences (Koljalg et al., 2013). This is also in
321 accordance with our results, suggesting that our new database for the 26S, validated by the widely
322 used SILVA, proved to be a curated and rich database to be used. Differences between the two

323 databases regarding taxonomic classification of sequences were obtained. The newly constructed
324 26S database delivered a more precise taxonomic assignment of the sequences. This could be due to
325 the fact that SILVA database comprised also non-microbial sequences, incomplete sequences and
326 sequences with unassigned taxonomy. In contrast, each taxonomy in our database was double
327 checked to get the higher taxonomic resolution, obtaining clearly more robustness results in terms
328 of taxonomic assignment from the biological samples. Special attention must be paid on the mis-
329 identification of fungal strain on the current available database. This current issue is pointed out in
330 this study, in which *S. cerevisiae* and *H. uvarum* were misidentified from fermented samples, using
331 SILVA's database.

332 Given the intricate nature of PCR, the amplification of biological samples has been
333 problematic (Polz and Cavanaugh, 1998). Our results exhibited high proportion of fungal coverage
334 (98 - 96 %) by both primer pair sets, which suggested that fungi account for roughly the complete
335 eukaryote rRNA in the studied fermented cocoa beans on average. The results also highlight a
336 lower biodiversity of fungal communities for ITS2 compared with the 26S region in fermented
337 cocoa beans, which contradicts with previous studies where ITS region has been used in NGS
338 studies as the universal primer set for fungi (Bokulich et al., 2014, 2012; Tedersoo et al., 2015).
339 Such discrepancies between outcomes of different studies may arise on account of the biased
340 quantification of relative abundance of taxa due to the uneven length of ITS fragments (Bellemain
341 and Carlsen, 2010), the preferential amplification of rRNA genes for certain taxa by PCR,
342 sequencing bias due to unequal amplification of the target gene or due to inaccurate taxonomic
343 classification of reference databases (Simon and Daniel, 2011). Despite these challenges, the greater
344 recovery trends in the community composition in the 26S target region observed here, have been
345 supported from previous studies, where higher discrimination power of species identification in
346 early diverging lineages of LSU compared with ITS was reported (Schoch et al., 2012).

347 Our study suggested that the 26S as a target showed greater biodiversity in biological
348 samples compared with the universal primer ITS. However, it should be noted that the present study
349 shows the performance of a new pair of primers targeting the 26S region for fungal strains.
350 Therefore, the novelty of these primer sets is also our limitation, that can be successfully overcome
351 through future research focusing on the use of small fragments of the LSU region to target fungal
352 species, that could support our observations. Therefore, the combination of both target genes, where
353 species identification can be performed applying ITS and phylogenetic analysis with 26S, is highly
354 recommended and the use of both will depend on the purpose of taxa investigation (Klaubauf et al.,
355 2010; Schoch et al., 2012). From a molecular microbial ecology perspective, in terms of
356 classification of marker-gene sequences, there is evidently a need for more extensive testing of
357 primers targeting different genes and *loci*, to support and identify all fungal species in NGS studies.
358 Clearly, the benefits of characterizing fermented microbial diversity may bring important
359 advancements to the food industry, such as discrimination of starter culture to improve food quality
360 or to accelerate processes. This study provides new insight into the selection of better primers and
361 taxonomical assignment to study fungal ecology, which should enable food research to gain better
362 view of the microbial diversity present in a range of fermentations avoiding biases. One also notes,
363 the limited availability of updated databases to assess ecological populations.

364 **Acknowledgments**

365 We would like to thank Dr. Sara Franco Ortega and Simona Prencipe for providing fungal
366 strains.

367 **Funding sources**

368 This research did not receive any specific grant from funding agencies in the public,
369 commercial, or not-for-profit sectors.

370 Reference

- 371 **Begerow, D., Nilsson, H., Unterseher, M., Maier, W.,** 2010. Current state and perspectives of
372 fungal DNA barcoding and rapid identification procedures. *Appl. Microbiol. Biotechnol.* 87,
373 99–108. <https://doi.org/10.1007/s00253-010-2585-4>
- 374 **Bekatorou, A., Psarianos, C., Koutinas, A.A.,** 2006. Production of food grade yeasts. *Food*
375 *Technol. Biotechnol.* 44, 407–415.
- 376 **Bellemain, E., Carlsen, T.,** 2010. ITS as an environmental DNA barcode for fungi: an in silico
377 approach reveals potential PCR biases. *BMC Microbiol.* 10, 1–9.
- 378 **Berbee, M.L.,** 2001. The phylogeny of plant and animal pathogens in the Ascomycota. *Physiol.*
379 *Mol. Plant Pathol.* 59, 165–187. <https://doi.org/10.1006/pmpp.2001.0355>
- 380 **Blackwell, M.,** 2011. The fungi: 1, 2, 3 ... 5.1 million species? *Am. J. Bot.* 98, 426–438.
381 <https://doi.org/10.3732/ajb.1000298>
- 382 **Bokulich, N.A., Bamforth, C.W., Mills, D.A.,** 2012. Brewhouse-resident microbiota are
383 responsible for multi-stage fermentation of American coolship ale. *PLoS One* 7.
384 <https://doi.org/10.1371/journal.pone.0035507>
- 385 **Bokulich, N.A., Mills, D.A.,** 2013. Improved selection of internal transcribed spacer-specific
386 primers enables quantitative, ultra-high-throughput profiling of fungal communities. *Appl.*
387 *Environ. Microbiol.* 79, 2519–2526. <https://doi.org/10.1128/AEM.03870-12>
- 388 **Bokulich, N.A., Thorngate, J.H., Richardson, P.M., Mills, D.A.,** 2014. Microbial biogeography
389 of wine grapes is conditioned by cultivar, vintage, and climate. *Proc. Natl. Acad. Sci.* 111,
390 E139–E148. <https://doi.org/10.1073/pnas.1317377110>
- 391 **Bonanomi, G., De Filippis, F., Cesarano, G., La Stora, A., Ercolini, D., Scala, F.,** 2016.
392 Organic farming induces changes in soil microbiota that affect agro-ecosystem functions. *Soil*
393 *Biol. Biochem.* 103, 327–336. <https://doi.org/10.1016/j.soilbio.2016.09.005>
- 394 **Bowers, R.M., Clum, A., Tice, H., Lim, J., Singh, K., Ciobanu, D., Ngan, C.Y., Cheng, J.F.,**
395 **Tringe, S.G., Woyke, T.,** 2015. Impact of library preparation protocols and template quantity
396 on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 16, 1–
397 12. <https://doi.org/10.1186/s12864-015-2063-6>
- 398 **Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh,**
399 **P.J., Fierer, N., Knight, R.,** 2011. Global patterns of 16S rRNA diversity at a depth of
400 millions of sequences per sample. *Proc. Natl. Acad. Sci.* 108, 4516–4522.
401 <https://doi.org/10.1073/pnas.1000080107>
- 402 **Cocolin, L., Alessandria, V., Botta, C., Gorra, R., De Filippis, F., Ercolini, D., Rantsiou, K.,**
403 2013. NaOH-debittering induces changes in bacterial ecology during table olives fermentation.
404 *PLoS One* 8. <https://doi.org/10.1371/journal.pone.0069074>
- 405 **Cocolin, L., Bisson, L.F., Mills, D.A.,** 2000. Direct profiling of the yeast dynamics in wine
406 fermentations. *FEMS Microbiol. Lett.* 189, 81–87. [https://doi.org/10.1016/S0378-1097\(00\)00257-3](https://doi.org/10.1016/S0378-1097(00)00257-3)
- 408 **David, V., Terrat, S., Herzine, K., Claisse, O., Rousseaux, S., Tourdot-Maréchal, R., Masneuf-**
409 **Pomarede, I., Ranjard, L., Alexandre, H.,** 2014. High-throughput sequencing of amplicons
410 for monitoring yeast biodiversity in must and during alcoholic fermentation. *J. Ind. Microbiol.*
411 *Biotechnol.* 41, 811–821. <https://doi.org/10.1007/s10295-014-1427-2>

412 **De Filippis, F., La Stora, A., Blaiotta, G.,** 2017a. Monitoring the mycobiota during Greco di Tufo
413 and Aglianico wine fermentation by 18S rRNA gene sequencing. *Food Microbiol.* 63, 117–
414 122. <https://doi.org/10.1016/j.fm.2016.11.010>

415 **De Filippis, F., Laiola, M., Blaiotta, G., Ercolini, D.,** 2017b. Different amplicon targets for
416 sequencing-based studies of fungal. *Appl. Environ. Microbiol.* 83, 1–9.

417 **Dixon, P.,** 2007. Vegan: Community ecology package for R. *J. Veg. Sci.* 14, 927–930.
418 [https://doi.org/10.1658/1100-9233\(2003\)014\[0927:vaporf\]2.0.co;2](https://doi.org/10.1658/1100-9233(2003)014[0927:vaporf]2.0.co;2)

419 **Edgar, R.C.,** 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*
420 26, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>

421 **Ercolini, D.,** 2013. High-throughput sequencing and metagenomics: Moving forward in the culture-
422 independent analysis of food microbial ecology. *Appl. Environ. Microbiol.* 79, 3148–3155.
423 <https://doi.org/10.1128/AEM.00256-13>

424 **Ercolini, D., De Filippis, F., La Stora, A., Iacono, M.,** 2012. “Remake” by high-throughput
425 sequencing of the microbiota involved in the production of water Buffalo mozzarella cheese.
426 *Appl. Environ. Microbiol.* 78, 8142–8145. <https://doi.org/10.1128/AEM.02218-12>

427 **Erlich, Y., Mitra, P.P., delaBastide, M., McCombie, W.R., Hannon, G.J.,** 2008. Alta-Cyclic: A
428 self-optimizing base caller for next-generation sequencing. *Nat. Methods* 5, 679–682.
429 <https://doi.org/10.1038/nmeth.1230>

430 **Fouhy, F., Clooney, A.G., Stanton, C., Claesson, M.J., Cotter, P.D.,** 2016. 16S rRNA gene
431 sequencing of mock microbial populations-impact of DNA extraction method, primer choice
432 and sequencing platform. *BMC Microbiol.* 16, 1–13. [https://doi.org/10.1186/s12866-016-](https://doi.org/10.1186/s12866-016-0738-z)
433 [0738-z](https://doi.org/10.1186/s12866-016-0738-z)

434 **Garofalo, C., Osimani, A., Milanović, V., Aquilanti, L., De Filippis, F., Stellato, G., Di Mauro,
435 S., Turchetti, B., Buzzini, P., Ercolini, D., Clementi, F.,** 2015. Bacteria and yeast microbiota
436 in milk kefir grains from different Italian regions. *Food Microbiol.* 49, 123–133.
437 <https://doi.org/10.1016/j.fm.2015.01.017>

438 **Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J.,
439 Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D.,** 2012.
440 New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial
441 and natural communities. *FEMS Microbiol. Ecol.* 82, 666–677. [https://doi.org/10.1111/j.1574-](https://doi.org/10.1111/j.1574-6941.2012.01437.x)
442 [6941.2012.01437.x](https://doi.org/10.1111/j.1574-6941.2012.01437.x)

443 **Jespersen, L., Nielsen, D.S., Hønholt, S., Jakobsen, M.,** 2005. Occurrence and diversity of yeasts
444 involved in fermentation of West African cocoa beans. *FEMS Yeast Res.* 5, 441–453.
445 <https://doi.org/10.1016/j.femsyr.2004.11.002>

446 **Klaubauf, S., Inselsbacher, E., Zechmeister-Boltenstern, S., Wanek, W., Gottsberger, R.,
447 Strauss, J., Gorfer, M.,** 2010. Molecular diversity of fungal communities in agricultural soils
448 from Lower Austria. *Fungal Divers.* 44, 65–75. <https://doi.org/10.1007/s13225-010-0053-1>

449 **Koljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F.S., Bahram, M.,** 2013.
450 Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–
451 5277. <https://doi.org/10.1111/mec.12481>

452 **Kurtzman, C.P., Robnett, C.J.,** 1997. Identification of clinically important ascomycetous yeasts
453 based on nucleotide divergence in the 5' end of the large-subunit (26S) ribosomal DNA gene.
454 *J. Clin. Microbiol.* 35, 1216. <https://doi.org/0095-1137/97>

455 **Li, X.R., Ma, E.B., Yan, L.Z., Meng, H., Du, X.W., Zhang, S.W., Quan, Z.X.**, 2011. Bacterial
456 and fungal diversity in the traditional Chinese liquor fermentation process. *Int. J. Food*
457 *Microbiol.* 146, 31–37. <https://doi.org/10.1016/j.ijfoodmicro.2011.01.030>

458 **Luo, C., Tsementzi, D., Kyrpides, N., Read, T., Konstantinidis, K.T.**, 2012. Direct comparisons
459 of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA
460 sample. *PLoS One* 7, 30087. <https://doi.org/10.1371/journal.pone.0030087>

461 **Magoč, T., Salzberg, S.L.**, 2011. FLASH: Fast length adjustment of short reads to improve
462 genome assemblies. *Bioinformatics* 27, 2957–2963.
463 <https://doi.org/10.1093/bioinformatics/btr507>

464 **Mota-Gutierrez, J., Botta, C., Ferrocino, I., Giordano, M., Bertolino, M., Dolci, P., Cannoni,
465 M., Cocolin, L.**, 2018. Dynamics and biodiversity of bacterial and yeast communities during
466 the fermentation of cocoa beans. *Appl. Environ. Microbiol.* AEM.01164-18.
467 <https://doi.org/10.1128/AEM.01164-18>

468 **Polz, M.F., Cavanaugh, C.M.**, 1998. Bias in template-to-product ratios in multitemplate PCR.
469 *Appl. Environ. Microbiol.* 64, 3724–3730. <https://doi.org/10.1007/s00253-012-4244-4>

470 **Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A.,
471 Swerdlow, H.P., Gu, Y.**, 2012. A tale of three next generation sequencing platforms:
472 comparison of Ion Torrent, PacificBiosciences and Illumina MiSeq sequencers. *BMC*
473 *Genomics* 13, 1–13. <https://doi.org/10.1186/1471-2164-13-341>

474 **Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner,
475 F.O.**, 2013. The SILVA ribosomal RNA gene database project: Improved data processing and
476 web-based tools. *Nucleic Acids Res.* 41, 590–596. <https://doi.org/10.1093/nar/gks1219>

477 **Roh, S.W., Kim, K.H., Nam, Y. Do, Chang, H.W., Park, E.J., Bae, J.W.**, 2010. Investigation of
478 archaeal and bacterial diversity in fermented seafood using barcoded pyrosequencing. *ISME J.*
479 4, 1–16. <https://doi.org/10.1038/ismej.2009.83>

480 **Schmieder, R., Edwards, R.**, 2011. Quality control and preprocessing of metagenomic datasets.
481 *Bioinformatics* 27, 863–864. <https://doi.org/10.1093/bioinformatics/btr026>

482 **Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., et al.**, 2012. Nuclear ribosomal internal
483 transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl.*
484 *Acad. Sci.* 109, 6241–6246. <https://doi.org/10.1073/pnas.1117018109>

485 **Stellato, G., De Filippis, F., La Storia, A., Ercolini, D.**, 2015. Coexistence of lactic acid bacteria
486 and potential spoilage microbiota in a dairy processing environment. *Appl. Environ.*
487 *Microbiol.* 81, 7893–7904. <https://doi.org/10.1128/AEM.02294-15>

488 **Stielow, J.B., Lévesque, C.A., Seifert, K.A., Meyer, W., et al.**, 2015. One fungus, which genes?
489 Development and assessment of universal primers for potential secondary fungal DNA
490 barcodes. *Persoonia - Mol. Phylogeny Evol. Fungi* 35, 242–263.
491 <https://doi.org/10.3767/003158515X689135>

492 **Tedersoo, L., Anslan, S., Bahram, M., Pölme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V.,
493 Nilsson, H., Hildebrand, F., Bork, P., Abarenkov, K.**, 2015. Shotgun metagenomes and
494 multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding
495 analyses of fungi. *MycoKeys* 10, 1–43. <https://doi.org/10.3897/mycokeys.10.4852>

496 **Walters, W.A., Caporaso, J.G., Lauber, C.L., Berg-Lyons, D., Fierer, N., Knight, R.**, 2011.
497 Primer Prospector: de novo design and taxonomic analysis of barcoded polymerase chain
498 reaction primers. *Bioinformatics* 27, 1159–1161. <https://doi.org/10.1093/bioinformatics/btr087>

499 **Wang, C., García-Fernández, D., Mas, A., Esteve-Zarzoso, B.,** 2015. Fungal diversity in grape
500 must and wine fermentation assessed by massive sequencing, quantitative PCR and DGGE.
501 Front. Microbiol. 6, 1–8. <https://doi.org/10.3389/fmicb.2015.01156>
502 **White TJ, Bruns T, Lee S, T.J.,** 1990. Amplification and direct sequencing of fungal ribosomal
503 RNA genes for phylogenetics, In: PCR Pr. ed. Academic Press, Inc., New York.
504
505

506 **Table legend**

507

508 **Table 1.** Primers used for Illumina MiSeq sequencing

509 **Table 2.** Fungal strains used for sequencing analysis and respective amplicon length

510 **Table 3.** Relative abundance (%) of the fungal species identified in mock communities amplified
511 using two different target regions. The expected concentration is referred to as theoretical

512

513

514

515 **Figure legend**

516 **Fig 1.** Principal component analysis (PCA) based on mock mycobiota composition.

517 **Fig 2.** Boxplots describe α -diversity measures (Chao1, Shannon index and number of observed
518 species) of fermented cocoa bean samples. Individual points and brackets represent the richness
519 estimate and the theoretical standard error range, respectively.

520 **Fig 3.** Distribution of OTUs in fermented cocoa bean samples in the amplicon datasets divided into
521 26S (upper figure) and ITS (lower figure). Only OTUs with an incidence above 1 % in at least 2
522 samples are shown.

523 **Fig 4.** Boxplots describe statistically different species detected in fermented cocoa bean samples
524 analysed with two different target genes. Individual points and brackets represent the relative
525 abundance and the theoretical standard error range, respectively.

526

527

528

529

530 **Supplementary table**

531 **Table S1.** Estimation of sample coverage of fungal community of fermented cocoa beans using 26S
532 target gene (A). Estimation of sample coverage of fungal community of fermented cocoa beans
533 using ITS2 target region (B)

534 **Table S2.** Relative abundance of mock communities identified by two different fungal databases

535

536 **Supplementary figure**

537 **S1.** Bar chart showing the evaluation of primer efficiency. Overall matches and weight score of 26S
538 primer pair against our constructed database 26S. Assessment of LS2-MF forward primer and NL4
539 reversed primer

540 **S2.** Predictive taxonomic coverage of 26S primers. Numeric values above bins represent total
541 sequence counts for each set, LS2-MF forward primer and NL4 reversed primer

542 **S3.** Histograms of read length of fungal communities using 26S target region and ITS2 target region

543

544

545 **Table 1**

Primer	Features	Primer sequence	Target region	Amplicon length	Reference
ITS3tagmix1	Fwd	5'-CTAGACTCGTCACCGATGAAGAACGCAG-3'	ITS2	385	Tedersoo <i>et al.</i> , 2015
ITS4ngs	Rev	5'- TTCCTSCGCTTATTGATATGC-3'	ITS2		Tedersoo <i>et al.</i> , 2015
LS2-MF	Fwd	5'-GAGTCGAGTTGTTTGGGAAT-3'	LSU D1	369	This study
NL-4	Rev	5'-GGTCCGTGTTTCAAGACGG-3'	LSU D1		Jespersen <i>et al.</i> , 2005

546

547

548

549 **Table 2**

Fungal species	Size bp (26S)	Size bp (ITS2)
<i>Alternaria alternata</i>	454	414
<i>Aspergillus flavus</i>	455	430
<i>Aspergillus fumigatus</i>	459	439
<i>Candida sake</i>	467	393
<i>Fusarium oxysporum</i>	462	405
<i>Fusarium verticillioides</i>	458	415
<i>Galactomyces geotrichum</i>	502	324
<i>Hanseniaspora opuntiae</i>	435	484
<i>Hanseniaspora osmophila</i>	462	520
<i>Kluyveromyces marxianus</i>	427	521
<i>Penicillium brevicompactum</i>	456	428
<i>Penicillium glabrum</i>	458	426
<i>Pichia kudriavzevii</i>	472	431
<i>Pichia membranifaciens</i>	466	398
<i>Plectosphaerella cucumerina</i>	457	441
<i>Saccharomyces cerevisiae</i>	460	496
<i>Saccharomycodes ludwigii</i>	452	470
<i>Schizosaccharomyces pombe</i>	488	562
<i>Starmerella bacillaris</i>	389	361
<i>Torulaspora delbrueckii</i>	461	518

550

551

552 **Table 3**

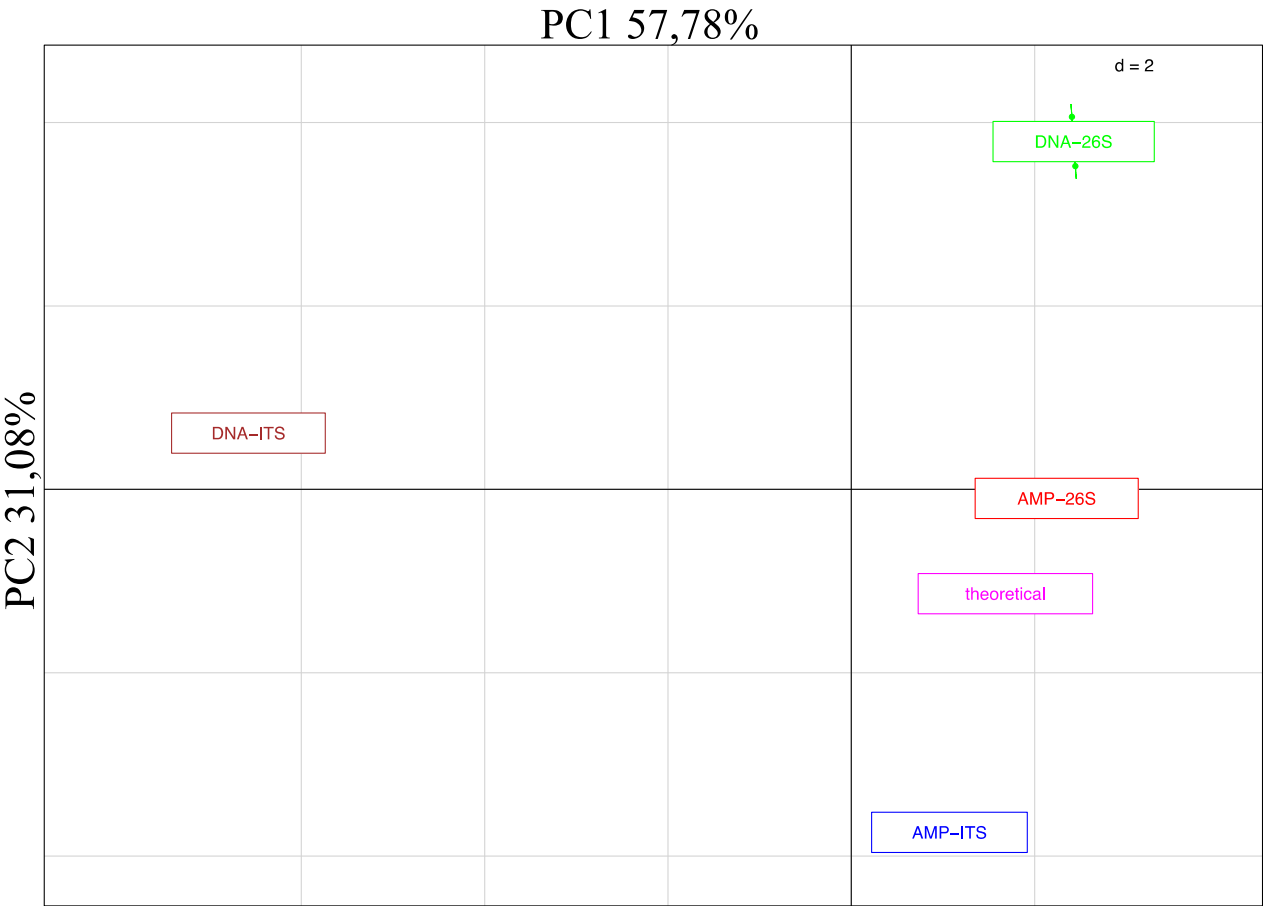
OTU	Theoretical	26S				Theoretical	ITS2				553
		DNA		AMP			DNA		AMP		
<i>Alternaria alternata</i>	5 ^{ab}	3.62	± 0.34 ^b	5.71	± 0.47 ^a	5 ^a	7.67	± 0.94 ^b	6.63	± 0.09 ^{ab}	
<i>Aspergillus flavus</i>	5 ^{ab}	5.78	± 0.46 ^a	4.03	± 0.07 ^b	5 ^a	9.39	± 0.17 ^b	7.19	± 0.26 ^a	
<i>Aspergillus fumigatus</i>	5 ^a	4.82	± 1.34 ^a	3.71	± 0.09 ^a	5 ^a	8.41	± 0.20 ^b	6.14	± 0.01 ^c	
<i>Candida sake</i>	5 ^a	0.24	± 0.20 ^b	0.06	± 0.02 ^b	5 ^a	4.57	± 0.38 ^a	6.19	± 0.03 ^b	
<i>Fusarium oxysporum</i>	5 ^{ab}	3.99	± 0.25 ^b	5.08	± 0.22 ^a	5 ^a	8.11	± 0.00 ^b	6.33	± 0.04 ^a	
<i>Fusarium verticillioides</i>	5 ^a	6.27	± 0.04 ^c	4.63	± 0.02 ^b	5 ^a	10.44	± 0.12 ^c	6.60	± 0.27 ^b	
<i>Galactomyces geotrichum</i>	5 ^a	13.39	± 0.61 ^c	6.94	± 0.11 ^b	5 ^b	0.00	± 0.00 ^a	0.00	± 0.00 ^a	
<i>Hanseniaspora opuntiae</i>	5 ^b	2.94	± 0.16 ^a	3.46	± 0.33 ^a	5 ^c	0.70	± 0.07 ^a	3.67	± 0.01 ^b	
<i>Hanseniaspora osmophila</i>	5 ^a	10.63	± 0.56 ^b	4.73	± 0.13 ^a	5 ^b	1.63	± 0.34 ^a	4.78	± 0.08 ^b	
<i>Kluyveromyces marxianus</i>	5 ^b	3.68	± 0.08 ^a	3.33	± 0.16 ^a	5 ^a	8.55	± 0.33 ^b	6.71	± 0.08 ^b	
<i>Penicillium brevicompactum</i>	5 ^a	6.00	± 0.92 ^a	3.55	± 0.29 ^a	5 ^a	14.39	± 0.09 ^c	9.04	± 0.08 ^b	
<i>Penicillium glabrum</i>	5 ^a	4.86	± 0.19 ^a	4.88	± 0.22 ^a	5 ^b	7.08	± 0.11 ^a	5.69	± 0.07 ^b	
<i>Pichia kudriavzevii</i>	5 ^a	7.24	± 1.89 ^a	10.76	± 0.45 ^a	5 ^b	0.12	± 0.14 ^a	2.11	± 0.16 ^b	
<i>Pichia membranifaciens</i>	5 ^a	6.63	± 1.83 ^a	4.82	± 0.28 ^a	5 ^b	2.58	± 0.39 ^a	3.57	± 0.13 ^b	
<i>Plectosphaerella cucumerina</i>	5 ^b	1.11	± 0.13 ^a	4.33	± 0.02 ^b	5 ^b	3.51	± 0.62 ^a	4.45	± 0.26 ^b	
<i>Saccharomyces cerevisiae</i>	5 ^b	1.86	± 0.09 ^a	4.20	± 0.49 ^b	5 ^a	4.74	± 0.03 ^a	5.44	± 0.10 ^b	
<i>Saccharomycodes ludwigii</i>	5 ^b	3.51	± 0.00 ^a	4.49	± 0.00 ^b	5 ^b	0.68	± 0.03 ^a	4.66	± 0.03 ^b	
<i>Schizosaccharomyces pombe</i>	5 ^b	5.00	± 0.11 ^b	3.71	± 0.19 ^a	5 ^b	0.01	± 0.00 ^a	0.00	± 0.00 ^c	
<i>Starmerella bacillaris</i>	5 ^a	4.21	± 1.68 ^a	5.91	± 0.11 ^a	5 ^b	0.16	± 0.07 ^a	0.19	± 0.10 ^b	
<i>Torulaspora delbrueckii</i>	5 ^c	2.46	± 0.02 ^a	3.90	± 0.17 ^b	5 ^a	5.38	± 0.11 ^a	5.30	± 0.52 ^a	

554 Values are expressed as the mean from duplicate determinations (%). Different letters indicate statistical difference related to relative abundances of
555 mock communities using least significant difference test ($P < 0.05$). P -values were adjusted using Bonferroni's method. Different colour showed no
556 difference (grey), underestimation (light green) or overestimation (light blue) between mock samples and theoretical data.

557

558

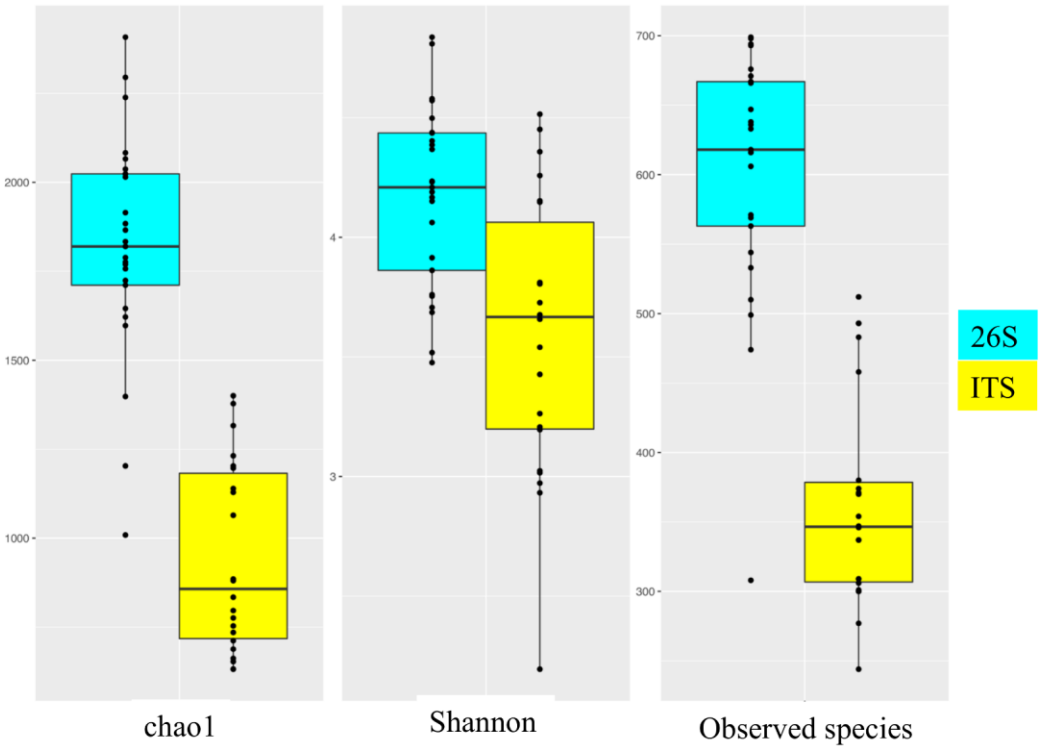
559 **Figure 1.**



560

561

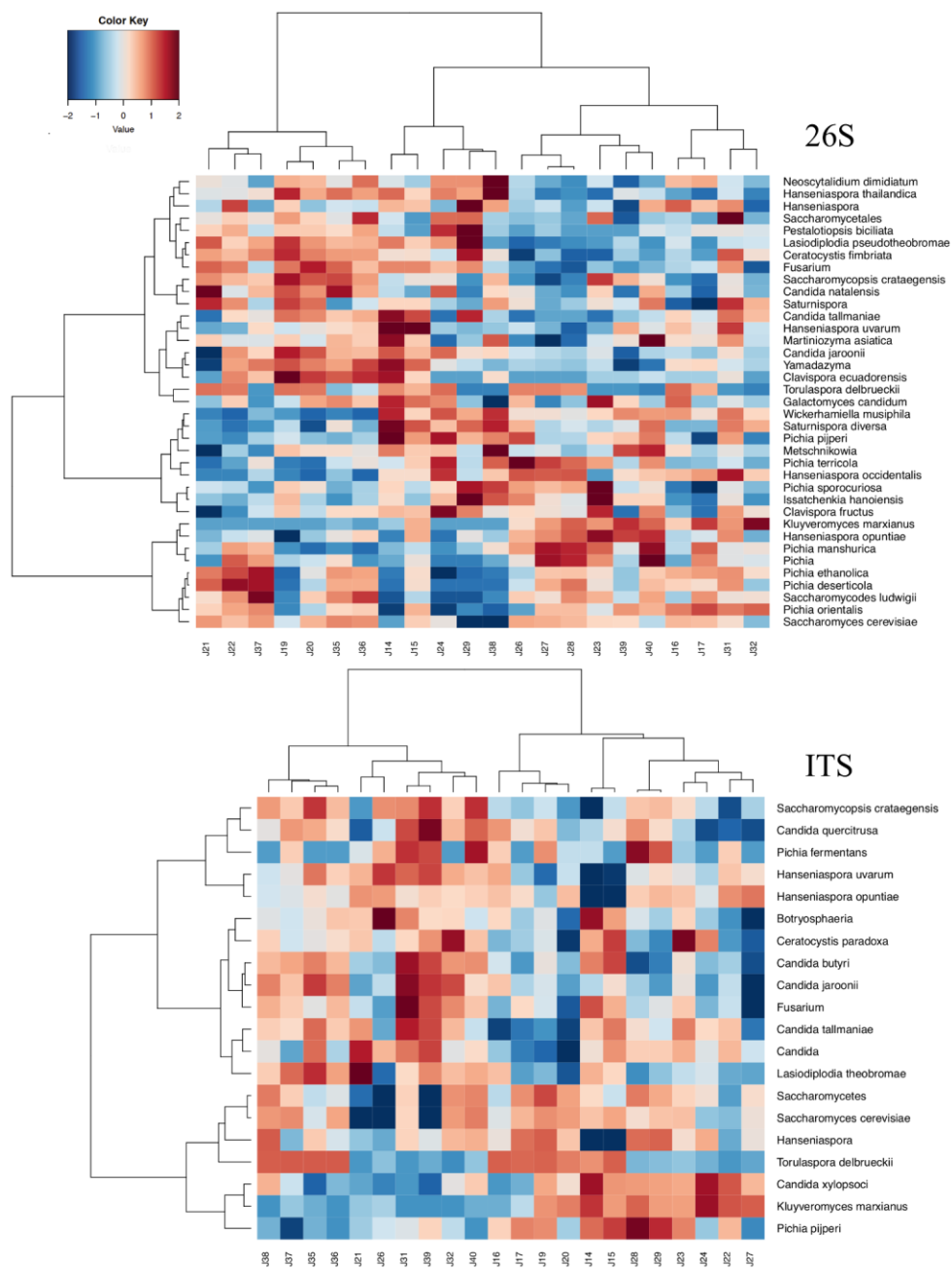
562 **Figure 2.**



563

564

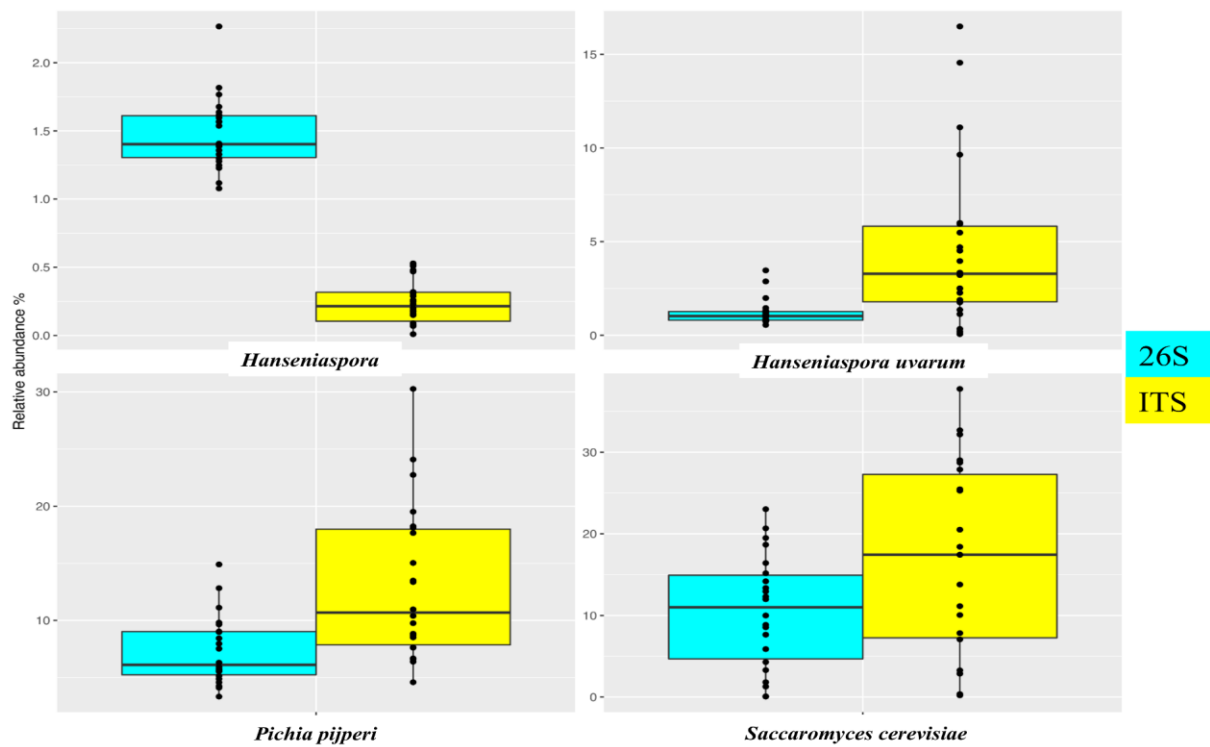
565 **Figure 3.**



566

567

568 **Figure 4.**



569