

# SCIENTIFIC REPORTS



OPEN

## DNA extraction and amplicon production strategies deeply influence the outcome of gut mycobiome studies

Alessandra Frau<sup>1</sup>, John G. Kenny<sup>2,7</sup>, Luca Lenzi<sup>2</sup>, Barry J. Campbell<sup>1</sup>, Umer Z. Ijaz<sup>3</sup>, Carrie A. Duckworth<sup>1</sup>, Michael D. Burkitt<sup>1,4</sup>, Neil Hall<sup>5</sup>, Jim Anson<sup>6</sup>, Alistair C. Darby<sup>2</sup> & Christopher S. J. Probert<sup>1</sup>

Microbial ecology studies are often performed through extraction of metagenomic DNA followed by amplification and sequencing of a marker. It is known that each step may bias the results. These biases have been explored for the study of bacterial communities, but rarely for fungi. Our aim was therefore to evaluate methods for the study of the gut mycobiome. We first evaluated DNA extraction methods in fungal cultures relevant to the gut. Afterwards, to assess how these methods would behave with an actual sample, stool from a donor was spiked with cells from the same cultures. We found that different extraction kits favour some species and bias against others. In terms of amplicon sequencing, we evaluated five primer sets, two for ITS2 and one for ITS1, 18S and 28S rRNA. Results showed that 18S rRNA outperformed the other markers: it was able to amplify all the species in the mock community and to discriminate among them. ITS primers showed both amplification and sequencing biases, the latter related to the variable length of the product. We identified several biases in the characterisation of the gut mycobiome and showed how crucial it is to be aware of these before drawing conclusions from the results of these studies.

The microbiota inhabiting the gut is mainly made of prokaryotes, but micro-eukaryotes are also present, including fungi. Recently, the gut mycobiome in health and disease has been explored and its role in some chronic diseases, such as inflammatory bowel disease (IBD), has been proposed<sup>1–4</sup>. As has been the case for prokaryotes, the analysis of the gut mycobiome has mainly been based upon amplicon sequencing. However, unlike the techniques used to evaluate structure and diversity of bacterial communities, those applied to the analysis of the mycobiome have not, as yet, been thoroughly validated. A call has been made to address this issue<sup>5</sup> but this has only been given attention by a few recent studies<sup>6,7</sup>.

Amplicon sequencing requires the extraction of DNA from a diverse and complex mixture of microorganisms whose identity is determined at the end of the experiment. Therefore, DNA extraction needs to be effective for a wide range of taxa; efficient extraction is crucial and its influence on the outcome is significant<sup>8–10</sup>. As for prokaryotic cells, the main challenge during the extraction of DNA from fungi is the lysis of the cell wall. In fungi, the cell wall is made of several polysaccharides, proteins and glycoproteins. Polysaccharides make up around 80% of the wall, while proteins comprise 20%. Lipid and wax can also be found in low concentration and these are mainly involved in preventing desiccation<sup>11</sup>. Generally, there are three layers: the outermost layer is composed of glycoproteins with glucan or mannan polysaccharides, followed by a layer of  $\beta$ -1,3 glucans and the innermost layer is made from chitin<sup>11</sup>. The structure of the cell wall varies among different taxa: ratios of the components listed

<sup>1</sup>Gastroenterology Research Unit, Department of Cellular & Molecular Physiology, Institute of Translational Medicine, University of Liverpool, Ashton Street, Liverpool, L69 3GE, UK. <sup>2</sup>Centre for Genomic Research (CGR), University of Liverpool, Crown Street, Liverpool, L69 7ZB, UK. <sup>3</sup>School of Engineering, University of Glasgow, Oakfield Avenue, Glasgow, G12 8LT, UK. <sup>4</sup>Division of Diabetes, Endocrinology and Gastroenterology, University of Manchester, Dover Street, Manchester, M13 9PT, UK. <sup>5</sup>Earlham Institute, Colney Ln, Norwich, NR4 7UZ, UK. <sup>6</sup>Liverpool Clinical Laboratories Directorate of Infection and Immunity, Royal Liverpool and Broadgreen University Hospitals NHS Trust, Prescot Street, Liverpool, L7 8XP, UK. <sup>7</sup>Teagasc Food Research Centre, Moorepark, Cork, Ireland. Correspondence and requests for materials should be addressed to C.S.J.P. (email: [mdcsjp@liverpool.ac.uk](mailto:mdcsjp@liverpool.ac.uk))

above can change and their composition may vary<sup>11</sup>. The cell wall also varies according to the cell status within the same species (e.g. spore vs hyphae)<sup>12</sup>. This variation influences the extraction efficiency of a given DNA extraction procedure in various species<sup>13,14</sup>. Cell lysis can be achieved by chemical treatment (with detergents, acids or alkaline agents), enzymatic degradation of cell wall components (with enzymes such as proteinase K, lyticase and chitinase) or using mechanical/physical approaches (e.g. sonication, bead beating, high temperature etc.)<sup>15</sup>. Often, a combination of these may be applied, especially if the goal is to broaden the range of species extracted. In early studies describing the gut mycobiome, three main approaches have been applied to extract DNA from faecal samples: (i) published protocols, which have often included bead beating<sup>4,16,17</sup>, (ii) commercial kits used as per manufacturer defined protocols<sup>18–23</sup> and (iii) commercial kit protocols which have been modified with the addition of extra initial steps, such as bead beating and/or enzymatic degradation, to improve efficiency of cell lysis<sup>24–30</sup>.

To date, no unique genetic biomarker for the study of fungal communities has been identified. The entire internal transcribed spacer (ITS) has been proposed as the “universal barcode marker for fungi”<sup>31</sup>. However, whether the ITS1 or ITS2 region of fungal ribosomal DNA (rDNA) is better for metabarcoding complex fungal communities is unclear: ITS1 was thought to be more variable and hence should allow better distinction among species than ITS2<sup>32</sup>, but the opposite has been shown<sup>33,34</sup>. ITS allows classification at species level, but the high variability in sequence and in length does not enable phylogenetic analysis or classification at high taxonomic ranks<sup>35,36</sup>. Genes encoding for the large (LSU) and small (SSU) rRNA subunit (28S rRNA and 18S rRNA) have also been considered as biomarkers for fungi<sup>31</sup>. These are more conserved, of regular length and can be used to make phylogenetic analysis<sup>35</sup>. However, the variability is considered too low to enable classification at species level, especially for 18S rRNA<sup>31</sup>. Higher resolution may be possible with 28S rRNA<sup>31</sup>, comparable to ITS<sup>37,38</sup>, although even this has been questioned<sup>34</sup>. A negative aspect of using LSU is that the reference sequence database is limited<sup>6</sup>. In order to achieve classification at species level and to enable phylogenetic analysis, the use of two markers, one from ITS and another from a region of the large or small rRNA subunit, could be considered; this would also allow us to understand amplification biases between these approaches<sup>39</sup>. The choice of primer set is also crucial, as it influences the community profile markedly<sup>5,34</sup>. It has been shown that ITS1F<sup>40</sup> and ITS2<sup>41</sup> primers, commonly used to characterise the gut mycobiome<sup>3,28–30</sup>, have mismatches to species relevant to the human mycobiome, including yeasts such as *Saccharomycetes* (ITS1F) and *Basidiomycetes* (ITS2) (i.e. *Malassezia* and *Cryptococcus*)<sup>5</sup>. Therefore, validation is needed before they are used to characterise metagenomes from the human body. In this context, it is clear that the choice of a suitable primer set is complicated. Because different markers and/or primers seem to bias against specific taxonomic groups, the environment analysed and its likely mycobiome must be taken into account when designing the experiment<sup>34</sup>.

This study has two main aims. First, we wanted to understand if DNA extraction kits specific for stool and developed for the extraction of bacterial metagenomes were also suitable for the extraction of fungal DNA. To do this, we first tested the PSP<sup>®</sup> Spin Stool DNA kit (Strattec) on seven pure fungal strains (*Candida albicans*, *Candida tropicalis*, *Saccharomyces cerevisiae*, *Cryptococcus neoformans*, *Malassezia furfur*, *Aspergillus fumigatus* and *Penicillium chrysogenum*), each of which has been detected in the gut in three or more studies<sup>1,19,23,28</sup>. The PSP<sup>®</sup> Spin Stool DNA kit was used as per manufacturer’s protocol and compared with a modification of the same, i.e. inclusion of an extra initial bead beating step, to determine whether this facilitated an increase in fungal DNA extraction. Secondly, the QIAamp<sup>®</sup> Fast DNA Stool Mini Kit (Qiagen) was tested on pure cultures of two of the seven fungal strains (*C. albicans* and *A. fumigatus*) with the goal of comparing two different DNA extraction kits. Extractions were compared by gene quantification of 18S rRNA<sup>42</sup>. Following this, and to evaluate the outcomes of the extraction on fungal cells in actual stool matrix, a human faecal sample from a healthy donor was spiked with cells from the seven fungal species listed above, and the extraction methods used for the pure cultures applied. These tests were also made on the non-spiked sample in order to exclude and evaluate the community already present in the sample as a background matrix for all other experiments. The outcomes were evaluated by 18S rRNA gene quantification<sup>42</sup> and amplicon sequencing on the Illumina HiSeq platform. Finally, to test the performance of the extraction kits in a real scenario, 24 stool samples from donors were extracted with the two kits under evaluation (i.e. the PSP kit with additional bead beating and the Qiagen kit) and the fungal load (18S rRNA qPCR<sup>42</sup>) was measured. The second aim was to evaluate five primer sets to assess their performance and eventual biases on the characterisation of the gut mycobiome. The markers evaluated were 18S and 28S rRNA, ITS1 and ITS2. For ITS2, two primer sets were evaluated. A mock community made of the seven fungal species listed above and the templates obtained with the multiple extractions of the spiked/non-spiked stool sample (from a healthy donor) were used as a template for amplicon production and sequencing. To further test this in a real scenario, two primer sets were used to amplify and sequence stool DNA from 24 donors.

## Results

**DNA extraction from pure fungal cultures.** The approach and outcome of the extraction of DNA from fungal isolates are shown in Table 1 and Fig. 1 respectively. The addition of an initial 6-minute bead beating step prior to extraction using the PSP kit improved the yield of DNA from two of the isolates: *C. tropicalis* and *M. furfur*. However, the yield of *C. neoformans* DNA was reduced following bead beating (Fig. 1A). Comparing the PSP kit and the Qiagen kit (Fig. 1C) showed these to have significant and opposite results. The first improved the extraction of *C. albicans*, while the second greatly improved the extraction of *A. fumigatus*.

**DNA extraction from stool.** A flowchart summarising the aliquoting, spiking and extractions performed in stool is shown in Fig. 1B. In general, the methods behaved similarly, each retrieving approximately 10<sup>6</sup> copies (Fig. 1D). However, the PSP kit did recover more fungal DNA than the Qiagen kit (Fig. 1E). As expected, the stool samples that were spiked with each of the seven fungal strains contained fungal DNA. Non-spiked samples (ns) also showed a conspicuous number of copies (from 1 × 10<sup>6</sup> up to 5 × 10<sup>6</sup> gene copies) that masked the spike with 10<sup>4</sup> cells, especially if we consider that sequences assigned to the pure cultures were found in very low abundance in these samples (Fig. 2B).

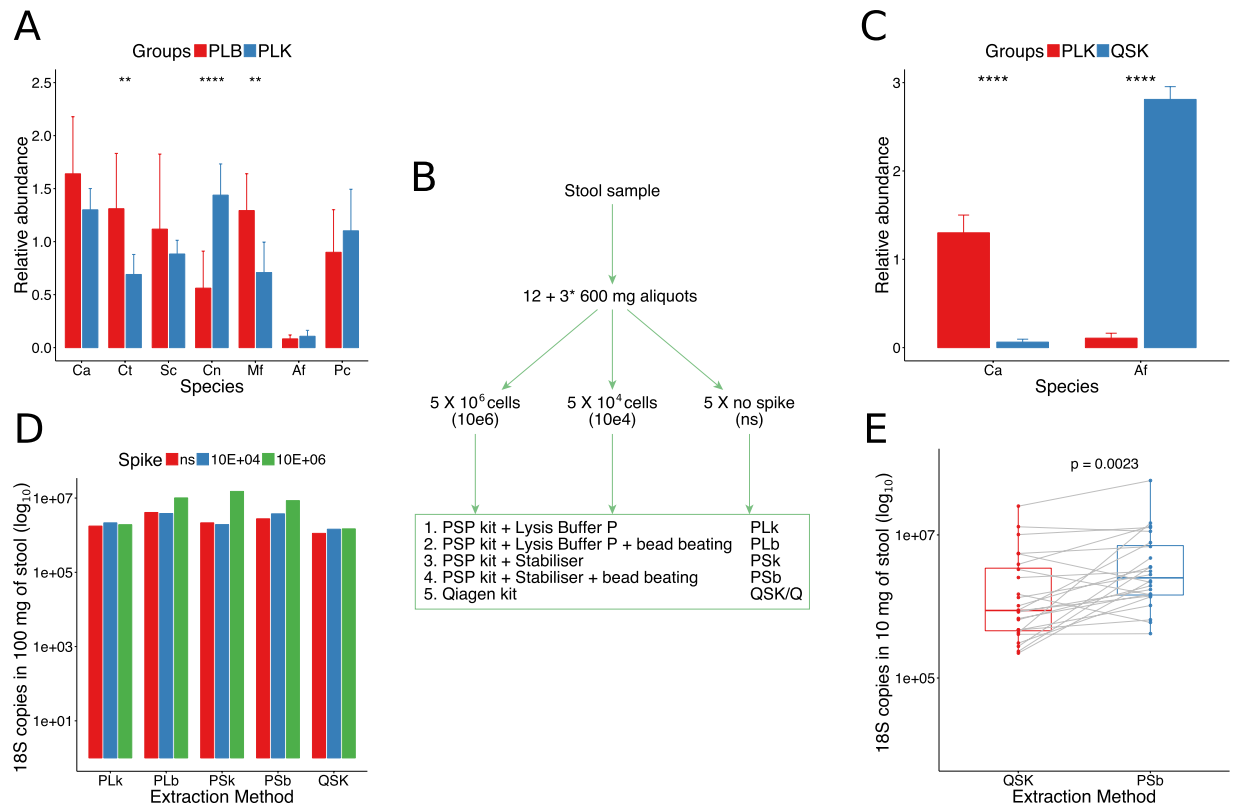
| Name   | Extraction method   |                       |
|--|---|-----------------------|
| <b>Pure cultures and stool samples extractions</b> |   |                       |
| PLk  | PSP kit + Lysis Buffer P  |                       |
| PLb  | PSP kit + Lysis Buffer P + bead beating   |                       |
| PSk  | PSP kit + Stabilizer  |                       |
| PSb  | PSP kit + Stabilizer + bead beating   |                       |
| QSK/Q  | Qiagen Kit  |                       |
| Name   | Extraction method   | Spike                 |
| <b>Stool sample extractions and spike</b>          |   |                       |
| PL10e6K  | PSP kit + Lysis Buffer P  | 10 <sup>6</sup> cells |
| PL10e4K  |   | 10 <sup>4</sup> cells |
| PLnsK  |   | Non-spiked            |
| PL10e6b  | PSP kit + Lysis Buffer P + bead beating   | 10 <sup>6</sup> cells |
| PL10e4b  |   | 10 <sup>4</sup> cells |
| PLnsb  |   | Non-spiked            |
| PS10e6K  | PSP kit + Stabilizer  | 10 <sup>6</sup> cells |
| PS10e4K  |   | 10 <sup>4</sup> cells |
| PSnsK  |   | Non-spiked            |
| PS10e6b  | PSP kit + Stabilizer + bead beating   | 10 <sup>6</sup> cells |
| PS10e4b  |   | 10 <sup>4</sup> cells |
| PSnsb  |   | Non-spiked            |
| Q10e6  | Qiagen Kit  | 10 <sup>6</sup> cells |
| Q10e4  |   | 10 <sup>4</sup> cells |
| Qns  |   | Non-spiked            |
| Name   | Preparation   |                       |
| <b>Mock Community</b>                              |   |                       |
| PreMC  | Genomic DNA from each fungal isolate was pooled then amplified (when replicates were processed, these are numbered 1 and 2) |                       |
| POSTMC   | DNA was amplified separately for each species, then pooled before sequencing  |                       |
| Name   | Sample type   |                       |
| <b>Donors</b>                                      |   |                       |
| D  | Stool samples from donors collected in the context of the SysMedIBD project   |                       |

**Table 1.** Names of samples as used in Figs 1–5.

Sequencing results of spiked stool samples are shown in Fig. 2. The taxa summary (Fig. 2B) shows that extraction with different methods generated different taxonomic profiles. The difference between the PSP kit with stabilizer and bead beating (PSb) and the Qiagen kit for the spike with 10<sup>6</sup> cells was striking. The PSb method showed a higher relative abundance of *C. albicans* and less *A. fumigatus*, whereas the Qiagen kit showed the opposite. The reduced *C. albicans* and increased *A. fumigatus* in Q10e6 confirm the qPCR results (Fig. 1C). However, the sequencing results of PLk, the third extraction method tested, did not entirely confirm the qPCR results carried out in pure cultures. In PL10e6k there was a reduction of *A. fumigatus* compared to Q10e6, but it was not coupled with an increase in *C. albicans*. Intriguingly, when samples extracted with the PSP kit with Lysis Buffer P (PL10e6k and PL10e6b) were compared with those extracted with the Stabilizer (PS10e6k and PS10e6b), beta diversity calculated by Bray-Curtis distance (Fig. 2C) showed that the PS10e6k and PS10e6b samples clustered together, whereas the PL10e6k and PL10e6b were separate and closer to Q10e6. The Stabilizer seemed to be a key factor in the extraction. Moreover, when samples were treated with Stabilizer and then bead beaten (PS10e6b), a greater abundance of both *Candida* species and *M. furfur* was obtained. *S. cerevisiae* DNA was extracted by all methods and while it was already present in the sample, its abundance was much higher when using the PSP kit with the Stabilizer. The extraction of *C. neoformans* was challenging regardless of the method used: the greatest relative abundance was found with the Qiagen kit (Fig. 2B). The results for *P. crysogenum* were also interesting: its abundance was higher with the Qiagen kit and PSP with the lysis buffer than with the PSP kit with the Stabilizer buffer.

OTU richness results (Fig. 2B) showed that the highest number of species was identified with the highest spike (10<sup>6</sup> cells) as expected. This result was consistent when analysing the samples with ITS2 primer sets (Fig. 3). Beta diversity results for 18S rRNA amplicons (Fig. 2C) showed that the samples clustered according to the spike; in the unweighted UniFrac chart, a gradient from the non-spiked samples to the samples spiked with 10<sup>6</sup> cells was observed, with the 10<sup>4</sup>-spiked samples in the centre. When abundance was taken into account (Bray-Curtis distance), the non-spiked and the 10<sup>4</sup>-spiked samples clustered more tightly.

**Primer set performance.** The 5 primer sets gave different taxonomic profiles. The ITS primers lacked some of the species of the mock community. This was observed more in the pre-PCR mock community (Table 2), showing that this was related to amplification biases. Moreover, primers amplifying the small and large rRNA subunits showed a higher number of extra OTUs in the pre-PCR mock community (Table 2), suggesting the amplification

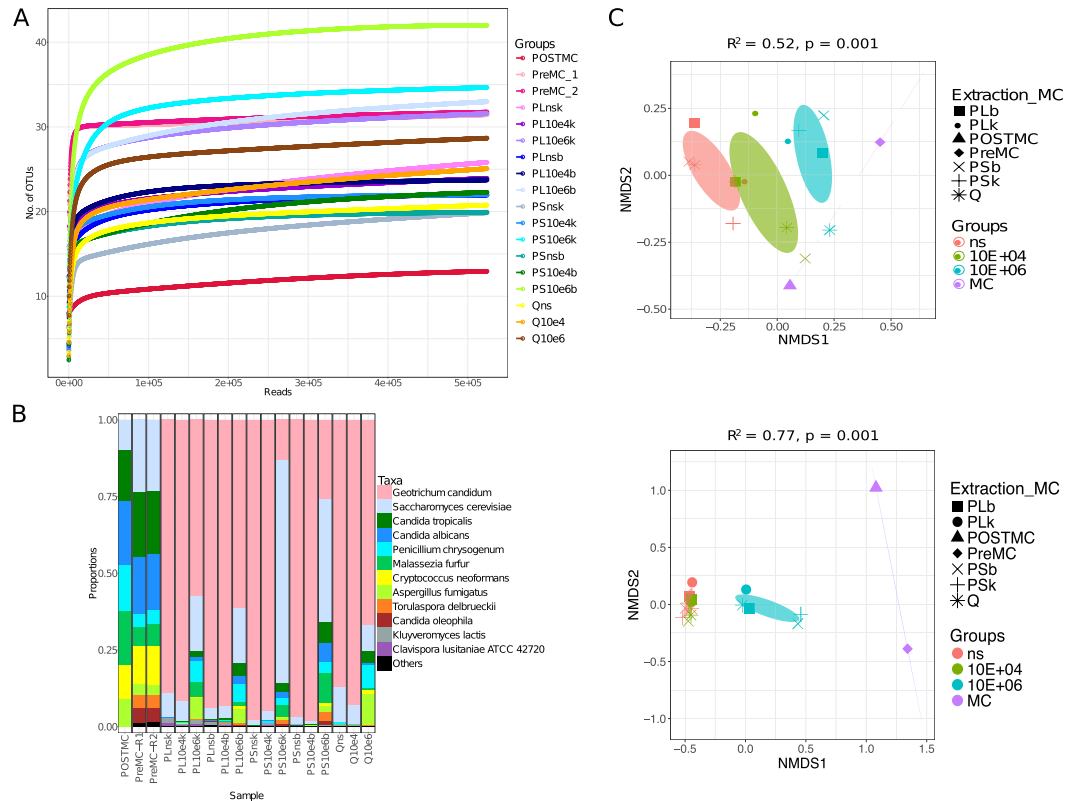


**Figure 1.** 18S rRNA qPCR results (A,C,D,E) and flowchart (B) showing the aliquoting, spiking and extraction of the stool sample. This was aliquoted (line 2 of the flowchart), then a third of the aliquots was spiked with  $10^6$  cells of each species of the mock community, a third was spiked with  $10^4$  cells and the final third was analysed as it was (non-spiked). These were then extracted using 5 different methods (box). (A) Comparison of the PSP kit performance with extra bead beating (PLb) and without (PLk). (C) Comparison of the PSP kit (PLk) performance versus the Qiagen kit (QSK). (A,C) The relative abundance (DNA concentration) was normalised to the mean of all values. A Welch t-test was used to compare the means of the two extraction methods for each species; where results were found to be significant, this is shown in the figure (ns:  $p > 0.05$ ; \* $p \leq 0.05$ ; \*\* $p \leq 0.01$ ; \*\*\* $p \leq 0.001$ ; \*\*\*\* $p \leq 0.0001$ ). The error bars denote standard deviation of the mean (SD). Fungal strains: Ca = *C. albicans*, Ct = *C. tropicalis*, Sc = *S. cerevisiae*, Cn = *C. neoformans*, Mf = *M. furfur*, Af = *A. fumigatus*, Pc = *P. crysogenum*. (D) Bar chart showing the number of 18S rRNA copies in 100 mg of spiked/non-spiked stools. The samples are grouped according to the extraction method, as shown in the flowchart. The colours indicate the spiking (ns = non-spiked,  $10^4$  and  $10^6$  cells). (E) Pair-wise comparison of 18S rRNA gene copies (method = wilcox.test) in 10 mg of stool from donors ( $n = 24$ ) extracted with PSb and QSK kits. A list of sample names with related descriptions can be found in Table 1. \*Sample preparation for the QSK extraction was done separately following the first aliquoting.

of spurious products in mixed samples. These observations are described in detail for each primer set in the next sections; moreover, a summary of sequencing results with the number of raw and quality filtered reads, chimeras and number of OTUs and reads in the OTU tables can be found in Supplementary Table S1.

**18S rRNA.** The two most significant outcomes of using this primer set were (i) The primers retrieved all the species in the mock community samples (Fig. 2B) and assignment was possible at genus/species level, even though this marker has been considered too conserved for such a deep classification in the taxonomic hierarchy; and (ii) the rarefaction curve in Fig. 2A shows that the number of OTUs is inflated; only 7 species were expected in the mock community. This number is close for the post-PCR mock community (13 instead of 7) and may be related to the presence of two different OTUs assigned to *C. tropicalis* and a few other OTUs with a very low number of reads ( $< 60$ ) assigned to *Saccharomyces*. The pre-PCR mock community was more inflated (33 OTUs were recorded). It was speculated that the extra OTUs were spurious products created during the amplification process.

To further explore these two outcomes, the OTUs were aligned and analysed. The 8 OTUs of the post-PCR mock community (3 OTUs had  $< 10$  reads and were excluded) showed that the OTU assigned to *C. albicans* differed only by 1 base with one *C. tropicalis* OTU and by 2 bases with the other OTU assigned to this species. By contrast, the OTU assigned to *S. cerevisiae* differed by 10 bases (almost 3%), showing why a discrimination was obtained among these Saccharomycetales. The alignment showed that *A. fumigatus* and *P. chrysogenum* were also very similar. These belong to the same family (Trichocomaceae) and their sequences differed by 3 bases.

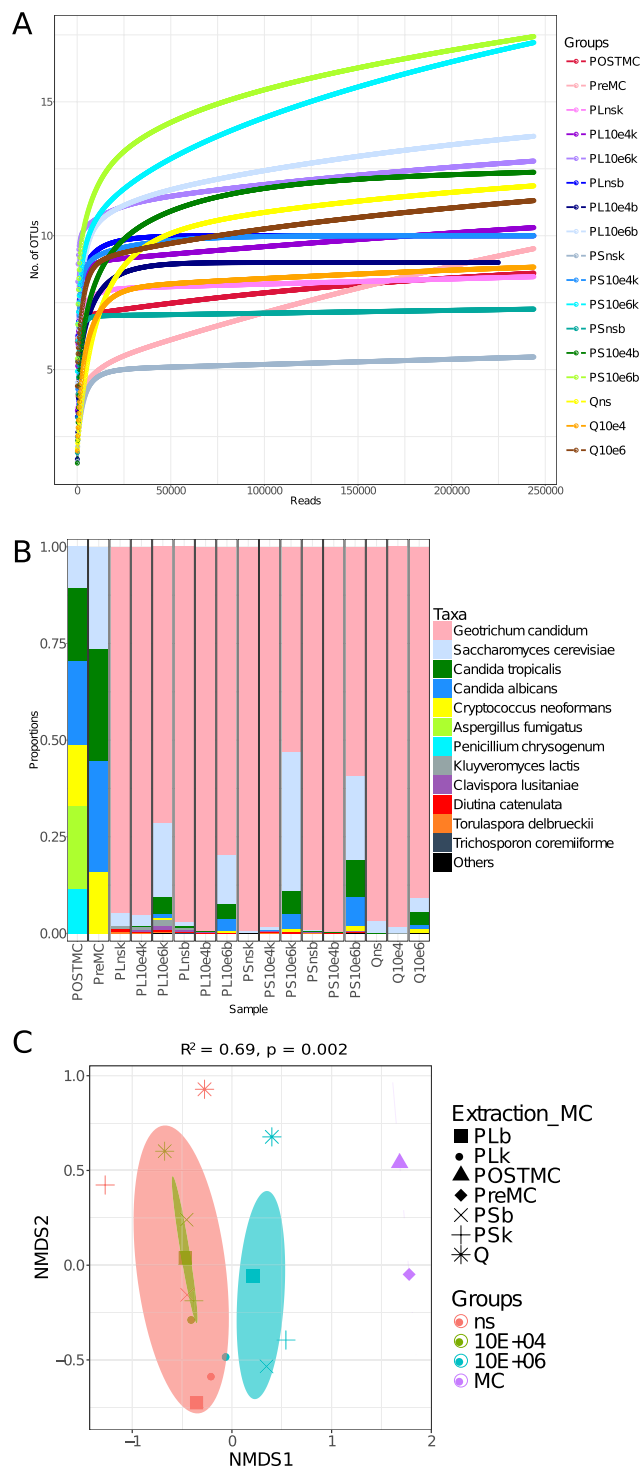


**Figure 2.** Use of the 18S rRNA primer set to define fungal diversity in stool ( $n = 15$ ) and mock community ( $n = 3$ ). **(A)** Rarefaction curves showing the number of OTUs versus the number of reads per sample. **(B)** Taxa summary showing the relative abundance of species, with each bar representing a sample. **(C)** Ordination of samples according to their community calculated through Non-Metric Distance Scaling (NMDS) was produced using unweighted UniFrac (top) and Bray-Curtis (bottom) distances. The ellipses were drawn at the 95% confidence interval of standard error and the mean value of the groups. PERMANOVA was used to assess if the clustering was significant. The  $R^2$  refers to the percentage of variability explained by the groups in terms of microbial community structure. A list of sample names, each with a related description, can be found in Table 1.

To understand the inflation in the number of OTUs, the 8 OTUs found in the post-PCR mock community were used as a gold standard and compared with the 22 OTUs found in excess in the pre-PCR mock community. The alignment showed that these were indeed chimeras, sharing their sequence with 2 or more of the gold standard OTUs. For example, the chart (Fig. 2) shows that two extra species (not included in the mock community) were rather abundant: *Torulospira delbrueckii* and *Candida oleophila*. Their alignment was analysed against the gold standard OTUs and we found that these were in fact chimeras (see Supplementary Table S3).

During the preparation of the pre-PCR mock community, an attempt was made to balance the amount of the 7 species added by adjusting the concentration according to the Cp value. Despite this effort, the proportion of some species was lower, especially *A. fumigatus*, *M. furfur* and *P. crysogenum*. This discrepancy may be explained by the fact that different amplification conditions between the qPCR and amplicon production were used. Moreover, other factors such as the different number of gene copies in the genome of the seven strains may have contributed to this discrepancy. No mismatches were detected for this primer set (see Supplementary Table S2), indicating this was not the reason for the imbalance observed in pre-PCR community profile. As shown in Fig. 2B, the stool sample is rich in a single species (*G. candidum*); this species was assigned to the OTU using BLASTN (see Methods section for details), as SILVA assigned it to an “uncultured Geotrichum”.

**ITS2 (ITS3tagmix/ITS4NGS set).** While 18S rRNA primers enabled the assessment of diversity in the mock community, this was not the case for the ITS3tagmix/ITS4NGS primer set. The post-PCR mock community showed all of the species. However, *M. furfur* was present in much lower abundance (0.04% for the post-PCR and 0.01% for the pre-PCR mock community) compared to the other species (see Fig. 3B and Supplementary Table S1). Because its amplification product was longer than that of other species (>600 bp), we initially assumed that *M. furfur* was missing because the reads were too long to be paired during data analysis. To explore this further, the unpaired reads (R1 and R2) for this sample were analysed separately. Despite this, *M. furfur* reads were very low. The most likely explanation is that the longer amplicons which were already present at a low abundance compared to shorter reads (see fragment analyser output: Supplementary Fig. S1) were biased against during clustering on the sequencing flowcell because of their length<sup>43</sup>. Secondly, the pre-PCR mock community lacked *A. fumigatus* and *P. crysogenum* (Fig. 3B). The pre-PCR mock community results were confirmed by the



**Figure 3.** Use of the ITS2 (ITS3tagmix/ITS4NGS) primer set to define fungal diversity in stool ( $n = 15$ ) and mock community ( $n = 2$ ). **(A)** Rarefaction curves showing the number of OTUs versus the number of reads per sample. **(B)** Taxa summary showing the relative abundance of species, with each bar representing a sample. **(C)** Ordination of samples according to their community, calculated through NMDS, was produced using Bray-Curtis distance. The ellipses were drawn at the 95% confidence interval of standard error and the mean value of the groups. PERMANOVA was used to assess if the clustering was significant. The  $R^2$  refers to the percentage of variability explained by the groups in terms of microbial community structure. A list of sample names, each with a related description, can be found in Table 1.

results obtained from stool samples, where moulds were not observed in any of the samples spiked with  $10^6$  cells, indicating that in a mixed sample, the primers seem to bias against these species. Primer mismatches were explored also for this set (see Supplementary Table S2) and again no mismatches were detected. Other factors,

| Species              | 18S rRNA |        | ITS2 (ITS3tagmix/ITS4NGS) |        | ITS2 (gITS7/ITS4NGS) |        | ITS1  |        | 28S rRNA |        |
|----------------------|----------|--------|---------------------------|--------|----------------------|--------|-------|--------|----------|--------|
|                      | PreMC    | POSTMC | PreMC                     | POSTMC | PreMC                | POSTMC | PreMC | POSTMC | PreMC    | POSTMC |
| <i>C. albicans</i>   | ✓        | ✓      | ✓                         | ✓      | ✓                    | ✓      | ✓     | ✓      | ✓        | ✓      |
| <i>C. tropicalis</i> | ✓        | ✓      | ✓                         | ✓      | ✓                    | ✓      | ✓     | ✓      | ✓        | ✓      |
| <i>C. neoformans</i> | ✓        | ✓      | ✓                         | ✓      | ✓                    | ✓      | ✓     | ✓      | ✓        | ✓      |
| <i>M. furfur</i>     | ✓        | ✓      | ✓                         | ✓      | ✓                    | ✓      | ✓     | ✓      | ✓        | ✓      |
| <i>S. cerevisiae</i> | ✓        | ✓      | ✓                         | ✓      | ✓                    | ✓      | ✓     | ✓      | ✓        | ✓      |
| <i>A. fumigatus</i>  | ✓        | ✓      | X                         | ✓      | X                    | ✓      | X     | ✓      | ✓        | ✓      |
| <i>P. crysogenum</i> | ✓        | ✓      | X                         | ✓      | X                    | ✓      | X     | ✓      | ✓        | ✓      |
| Extra OTUs           | ✓        | ✓      | X                         | X      | X                    | X      | ✓     | ✓      | ✓        | ✓      |
| N of OTUs            | 33*      | 13     | 14                        | 9      | 14*                  | 16     | 27    | 15     | 58       | 29     |

**Table 2.** Performance summary for each primer set in terms of MC species identified and inflation of species. The lack of some species in the pre-PCR compared to the post-PCR mock community using ITS primers suggests that the primer sets bias against these species during their amplification in mixed samples. In the same way, the observation of extra OTUs in the pre-PCR mock community with 18S and 28S rRNA primers suggests that spurious products were produced by these primers when using a template made of DNA from multiple species. NOTE: To be considered present (✓), OTUs had to have more than 10 reads. However, all the OTUs were considered in the OTU total count. Details of OTU numbers are in Supplementary Table S1. \*Average of the two replicates.

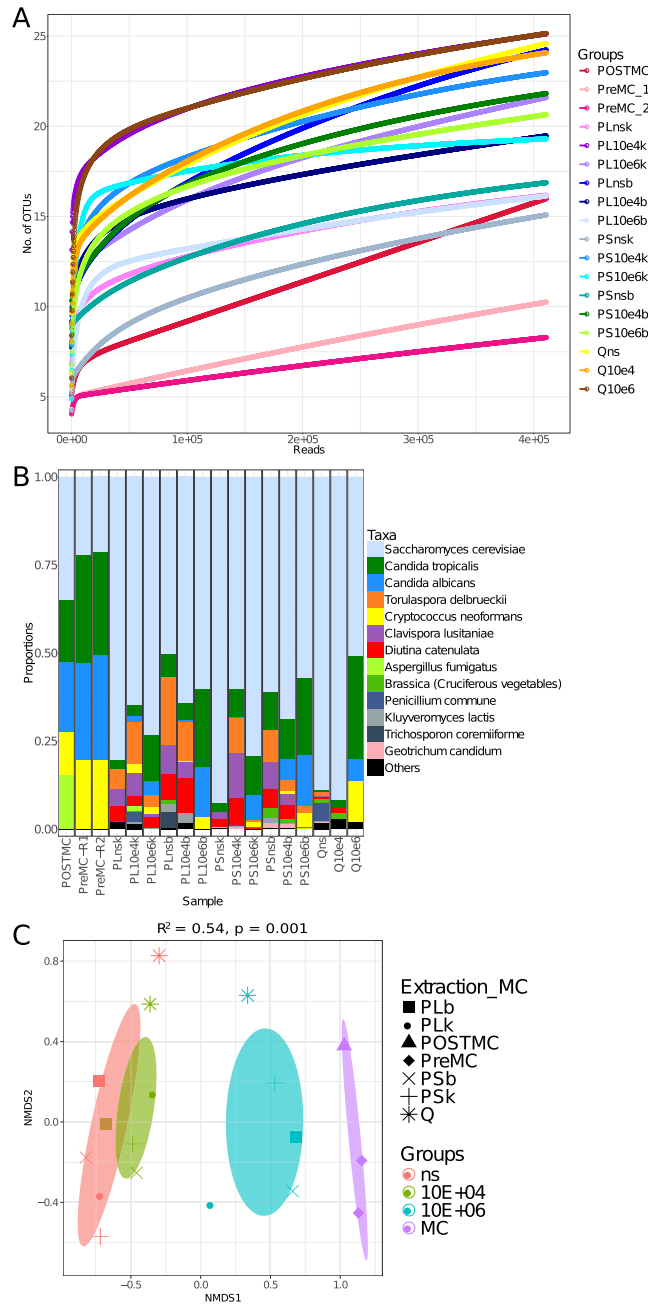
such as the different number of template copies in the genome of the species and the use of different PCR protocols for the qPCR assay to prepare the mock community and the amplicons production, which in this case also included different primers, could have contributed to this observation. In terms of number of OTUs amplified by these primers (see Fig. 3A), the number of species was less inflated than that seen for 18S rRNA (Table 2). The OTU table (Supplementary Table S1) showed that the abundance of the extra 2 OTUs in the post-PCR and 8 OTUs in the pre-PCR mock communities was very low (<10 reads) and could be excluded as noise or contamination. As observed with 18S rRNA, the community from the sample had a very high relative abundance of a single species/OTU (*G. candidum*) (Fig. 3B). This was originally assigned to *Dipodascus australiensis*; however, when running the read in BLASTN, it was assigned to *G. candidum* (consistent with the 18S rRNA marker). Finally, beta diversity results (Fig. 3C) were similar to those found with 18S rRNA primers.

**ITS2 (gITS7/ITS4NGS set).** As with the first ITS2 primer set (ITS3tagmix/ITS4NGS), very little *M. furfur* was observed (Supplementary Table S1). *Aspergillus* abundance was also very low. Moreover, the beta diversity results (Fig. 4C) showed similar clustering to that observed with 18S rRNA and ITS3tagmix/ITS4NGS sets. Apart from these similarities, several discrepancies were found with this set: reads assigned to *Penicillium* were very low in the POSTMC sample (41 reads; see Supplementary Table S1). However, this genus was found in some of the stool samples, both spiked and non-spiked (see Supplementary Table S1). The stool sample had a very low relative abundance of *G. candidum* (which dominated with 18S rRNA and ITS3tagmix) but showed a very high abundance of *Saccharomyces*. Moreover, this primer set gave extra taxa compared to the previous two primer sets: i.e. *C. tropicalis* was recorded in most of the samples, along with *Torulospora*, *Clavispora* and others (Fig. 4B). The number of OTUs observed in the non-spiked samples was higher than that which was observed when amplified using the ITS3tagmix/ITS4NGS primer set (Fig. 4A and Table 2).

**ITS1.** This set showed similar biases to the two ITS2 primers. However, in this case, *Saccharomyces* was in very low abundance because of the length of its product (>600 bp). Two OTUs were assigned to *S. cerevisiae* and a total of 269 reads were assigned to this species in the POSTMC and 31 in the PreMC sample (Supplementary Table S1). As shown with the other ITS primer sets, when in a mixed sample, the primers appeared to bias against *A. fumigatus* and *P. crysogenum* amplification (Fig. 5C). Contrary to what was observed for the three primer sets previously described, mismatches were observed for this primer set; ITS1-F had mismatches against *S. cerevisiae*, *A. fumigatus* and *P. crysogenum*, while ITS2r showed mismatches against *C. neoformans* and *M. furfur* (see Supplementary Table S2).

**28S rRNA.** This primer set gave interesting results: it was able to amplify most of the species in both mock community types. Nevertheless, it seemed to bias against *M. furfur*, as only a small amount was observed in the PreMC sample (Fig. 5D and Supplementary Table S1). This may be related to the fact that the reverse primer (LF402) showed a mismatch against this species (see Supplementary Table S2). However, the main issue was the lack of a comprehensive database, which prevented the classification of OTUs. *Candida* spp. were not classified with the database, but manually using BLASTN.

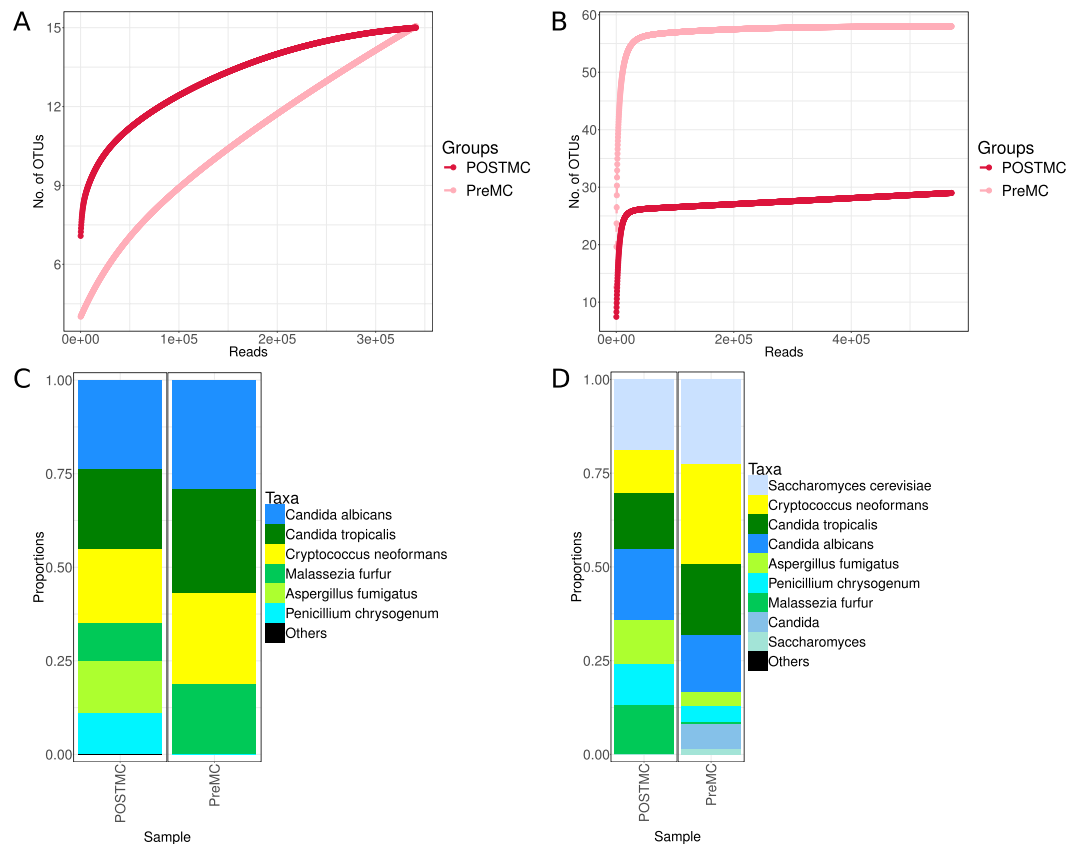
As observed for 18S rRNA, more spurious products were observed than with ITS, confirmed by an inflation in the number of OTUs (Fig. 5B, Table 2 and Supplementary Table S1). To explore this further, the OTUs' sequences were aligned. The post-PCR mock community showed 7 OTUs (one for each strain of the mock community) that were very abundant (>10<sup>4</sup> reads) and 13 OTUs had between 100 reads and up to 264 reads. The first, very abundant, OTUs were used as gold standards and the second were considered errors. Analysis of the alignment



**Figure 4.** ITS2 (gITS7/ITS4NGS) primer set performance in defining fungal diversity in stool ( $n = 15$ ) and mock community ( $n = 3$ ). **(A)** Rarefaction curves showing the number of OTUs versus the number of reads per sample. **(B)** Taxa summary showing the relative abundance of species, with each bar representing a sample. **(C)** Ordination of samples according to their community, calculated through NMDS, was produced using Bray-Curtis distance. The ellipses were drawn at the 95% confidence interval of standard error and the mean value of the groups. PERMANOVA was used to assess if the clustering was significant. The  $R^2$  refers to the percentage of variability explained by the groups in terms of microbial community structure. A list of sample names, each with a related description, can be found in Table 1.

showed that these were not chimeras but errors. For example, for some OTUs (i.e. OTU5, OTU6, OTU8, OTU12, OTU19 and OTU124) the primers were not correctly trimmed, causing an excess in the OTUs picked. In other cases, up to two base differences with the gold standard caused the detection of extra OTUs (i.e. OTU7, OTU51, OTU38, OTU121, OTU144 and OTU150). The pre-PCR mock community OTU table showed the same 7 gold standard OTUs with a high abundance, the lowest having 4255 reads (*M. furfur*) and the highest 299,044 reads (*C. neoformans*). Of the 49 OTUs in excess with more than 50 reads, 6 had more than 5000 reads (up to 17590 reads). All these were assigned to Saccharomycetales species belonging to genera such as *Candida* and *Saccharomyces*. The alignment of these OTUs showed that these were chimeras of the gold standard OTUs (see Supplementary Table S3).

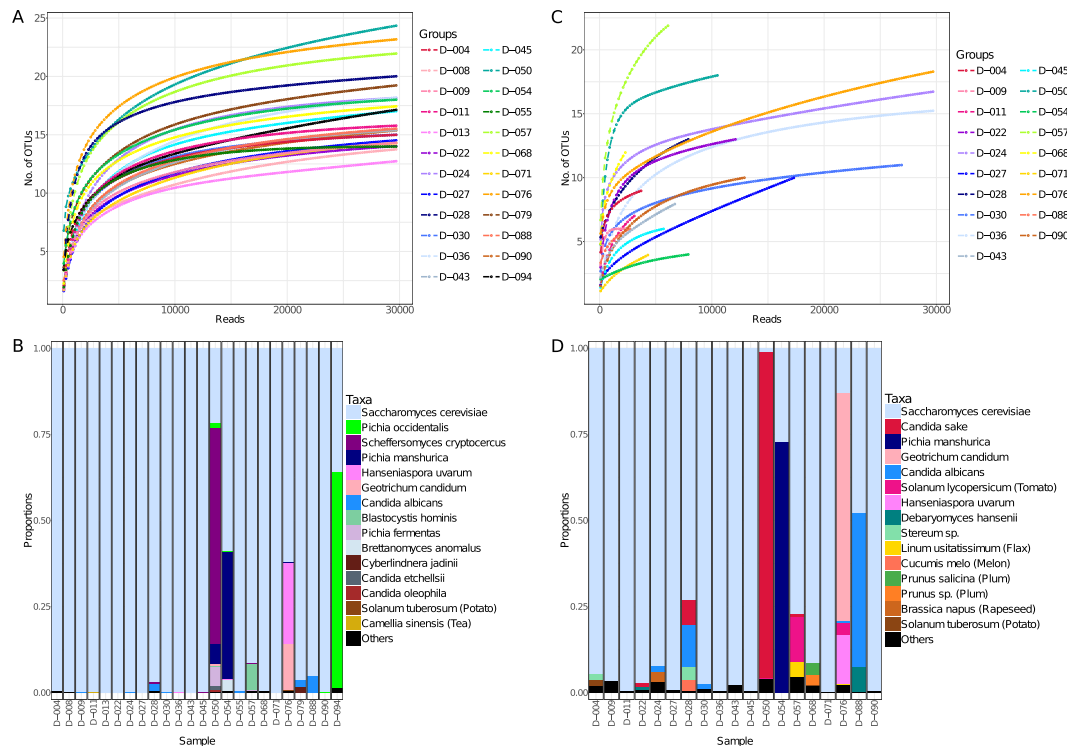




**Figure 5.** Performance of the ITS1 primer set in the mock community ( $n = 2$ ) (A,C) and 28S rRNA in the mock community ( $n = 2$ ) (B,D). ITS1 (A) and 28S rRNA (B) rarefaction curves showing the number of OTUs versus the number of reads per sample. ITS1 (C) and 28S rRNA (D) taxa summaries showing the relative abundance of species with each bar representing a sample. A list of sample names, each with a related description, can be found in Table 1.

**Negative controls.** Most of the negative controls were clean (i.e. the number of reads was below our threshold of 1000). Two exceptions were observed: the amplification negative control for the primer set ITS3tagmix/ITS4NGS and the DNA extraction negative control for the method PLb (Table 1) amplified with the gITS7/ITS4NGS primer set. The former gave 3436 high-quality reads, 98.75% of which were assigned to a single OTU, classified as *Aureobasidium pullulans*. This is a ubiquitous Ascomycetes yeast<sup>44</sup>. Considering that all the other PCR controls were clean, a contamination of the reagents can be excluded; therefore, this likely occurred during the preparation of the PCR. The latter negative control gave 1284 reads, of which 99.5% were assigned to a single OTU classified as *Malassezia restricta*, a species that is commonly found on the skin of healthy people<sup>45</sup>. While the contamination may have occurred during the extraction, it cannot be excluded that the kit reagents were contaminated.

**Primer comparison with stool samples.** The first observation made when comparing the performance of the 18S rRNA and ITS3tagmix/ITS4 sets was that the latter produced a very low amount of amplification product. When comparing the fragment analyser traces (Supplementary Fig. S2) of the 24 samples tested, only 4 samples gave a signal comparable to that observed with the 18S rRNA primer set. Twelve samples gave smaller peaks and the rest did not generate any product. All of the samples showed a large signal for a short (<200 bp) nonspecific product, likely a primer dimer. In spite of the poor PCR yield of some samples, all were sequenced. During the filtering of reads, samples with <2000 reads were discarded, leaving only 19 samples for analysis. The low PCR yield was reflected in the low number of reads gathered, as shown in the rarefaction curve in Fig. 6C. Taxonomy profiles were similar (Fig. 6B for 18S rRNA and Fig. 6D for ITS2 (ITS3tagmix/ITS4 set)): while most of the samples contained mainly *Saccharomyces cerevisiae*, 6 donor stool samples showed a more diverse profile. These were: D-028, D-050, D-054, D-057, D-076 and D-088. The assignment of the OTUs and the diversity were not always consistent between the two primer sets. For certain OTUs, this was due to discrepancies between the databases used for ITS2 (UNITE) and 18S rRNA (SILVA). For example, OTU5 (ITS2) and OTU49 (18S rRNA) were the same species, but UNITE assigned OTU5 to *Pichia mandshurica*, while SILVA assigned OTU49 to *Candida ethanolica*. This was solved when running the OTUs in BLASTN and both were assigned to *P. mandshurica*. For other OTUs, we could not explain these differences, which may be due to the different taxonomical resolution between the two markers or they may be related to biases during the amplification or sequencing steps, as described in previous sections. Nevertheless, some species seem to be consistent between samples. For example, *C. albicans* was recorded in samples D-024, D-028, D-030 and D-088 by both primer sets. It must be noted that its relative



**Figure 6.** Sequencing results of 18S rRNA (A,B) from donors ( $n = 24$ ) and ITS2 (C,D) ( $n = 19$ ). (A) Rarefaction curves showing the number of OTUs versus the number of reads per sample. (B) Taxa summary at species level, with each column representing a sample. (C) Rarefaction curves showing the number of OTUs versus the number of reads per sample. (D) Taxa summary at species level, with each column representing a sample.

abundance was higher when using ITS2 primer set. For this set, several OTUs were not assigned to any taxa. When these were evaluated in BLASTN, it was found that they belonged to plants: *Solanum* (tomato and potato), *Linum* (pale flax), *Cucumis* (melon), *Prunus* (plum) and *Lactuca* (lettuce), and were very likely to have originated in the subjects' diet. The 18S rRNA primers also amplified other eukaryotic DNA (e.g. *Blastocystis*), but in a lower relative abundance than that observed using the ITS2 primers.

## Discussion

Microbial ecology studies are known to be influenced by the methods used. What we observe is a function of what we are able to extract, amplify and sequence. These biases have been extensively evaluated for the study of bacterial communities of the human GI tract<sup>8,46</sup>, although the methods to resolve bias have yet to be fully established. Regarding fungal communities, only a few studies have reported the effectiveness of the methods used<sup>6,7,25</sup>. Our study evaluated several DNA extraction approaches of pure fungal cultures and of fungal strain-spiked stool samples and analysed stool samples from a cohort of donors. We have established that DNA extraction techniques can bias the community profile by being more effective with certain species than others. We also tested five primer sets that amplify fungal rDNA and have demonstrated how primer selection can significantly influence outcomes, both through preferential amplification and with respect to the relative quality of reference databases available. Finally, we have shown that when using markers with a variable length, the sequencing method also influences the outcome by biasing against longer reads.

The extraction of metagenomic DNA from stool is made challenging by the nature of the stool matrix and requires the use of specific reagents to reduce DNA degradation and neutralise inhibitors<sup>47</sup>. The PSP kit with the Stabilizer buffer addresses both of these issues. Its effectiveness for the extraction of fungal DNA has been shown in one study<sup>25</sup>, and our initial hypothesis was that by adding an additional bead beating step, we could improve the extraction of fungal DNA. The effectiveness of bead beating on sturdy fungal structures including spores was observed in other studies (in combination with other methods/kits)<sup>48,49</sup> and it is considered a better alternative than the use of an enzymatic step, because of the lower cost and shorter processing time<sup>50</sup>. It was interesting to see that this was observed only for two species (*C. tropicalis* and *M. furfur*) out of seven, and that the bead beating actually decreased the extraction of a fungus (*C. neoformans*). More striking was the comparison of the PSP kit and the Qiagen kit. On the one hand, the fact that the Qiagen kit was effective for the extraction of *A. fumigatus* spores was not surprising, as this has been already shown in studies focused on the diagnosis of invasive aspergillosis<sup>49</sup>. We confirmed this result through quantification, but also with a sequencing-based experiment, showing that a higher relative abundance of *A. fumigatus* is gathered with the Qiagen kit. The PSP kit, on the other hand, was more efficient in extracting *Candida* and *Malassezia* DNA, especially when bead beating was combined with

the Stabilizer buffer. Our work confirms that DNA extraction methods influence the community profile, and in order to maximise the diversity of species extracted, we took the informed decision to use both methods (PSB and QSK) to extract DNA from stool. A limitation of our experimental design is the use of spores for *A. fumigatus* instead of hyphae. This fungus is normally found in the latter state when active<sup>14</sup>. However, spores were chosen because these allow an exact quantification of the number of cells used during the extraction and therefore enable the standardisation of the extractions from pure cultures and the spiking of stool.

The comparison of four markers (18S and 28S rRNA, ITS1 and ITS2) and five primer sets produced some intriguing results. First, in contrast to a previous study<sup>25</sup> and the general view of this marker<sup>31</sup>, 18S rRNA performed similarly to the other markers in terms of depth of classification. This set was not only able to distinguish all of the mock community species, but also those that were phylogenetically close (i.e. *C. albicans*, *C. tropicalis* and *S. cerevisiae*). Moreover, when comparing its performance using stool samples, this set gave a higher yield of PCR product and more of its reads were assigned to fungi than ITS2, which amplified more plant/food DNA. 18S rRNA was also able to gather a diversity profile similar to ITS2. These outcomes may be related to the use of a primer set specific for fungi and thoroughly tested with human pathogens<sup>42</sup>. Two limitations of using 18S rRNA are that (i) it is difficult to assess if any diversity (outside the mock community) is lost because of the high level of conservation of the gene, and therefore it is not recommended to rely on a classification at species level in “black box” samples, whose taxonomic profile is discovered at the end of the experiment; and (ii) the use of an rRNA gene (18S or 28S for fungi or 16S for bacteria) made of variable and conserved regions increases the chance of producing spurious products, i.e. chimeras generated during amplification that inflate diversity<sup>46,51</sup>. This issue may be attenuated by reducing the number of cycles during PCR<sup>46</sup>, along with the use of bioinformatic tools to filter out chimeras. Production of spurious products is reduced with ITS primer sets, as by being hypervariable in its entire length, the formation of chimeras during amplification is more unlikely<sup>51</sup>. This was confirmed in our study: when taking abundance into account, the percentage of chimeras obtained with ITS primers was at least 16 times lower compared to the rRNA primer sets; this was 3.2% for 18S rRNA, 0.2% and 0.1% for the two ITS2 primer sets respectively, 0% for ITS1 and 5.5% for 28S rRNA (see Supplementary Table S1). Despite these limitations, the use of this 18S rRNA primer set was very promising for the study of the gut mycobiome; also, compared to priming for ITS, it allowed phylogenetic analysis, the product had a regular length (reducing PCR and sequencing biases) and it gave a higher PCR yield when amplifying DNA from a matrix of stool. Moreover, it can be used when RNA (or cDNA) is the template in amplicon sequencing studies. The 18S rRNA database from SILVA is very comprehensive and allowed us to classify at genus/species level for most OTUs. This was not the case for 28S rRNA.

If 18S rRNA gave us the mock community profile closest to that which we expected, this was not the case for the ITS sets. All three primer sets tested showed two principal drawbacks: (i) preferential amplification and (ii) sequencing bias due to the variable length of the ITS amplicons. The first explains why *A. fumigatus* and *P. crysogenum* were not found in the pre-PCR mock community with any of the ITS primer sets. This indicates that when these are present in a mixed sample, their amplification is biased against. This also occurred with *Geotrichum candidum* in stool amplified with gITS7/ITS4NGS. This OTU was detected at much lower levels with this primer set, giving taxonomic profiles for the stool sample (spiked and non-spiked) that were radically different from those we obtained with 18S rRNA and ITS3tagmix/ITS4NGS-targeted primer sets.

The second issue was related to the sequencing process. Certain species (*M. furfur* with ITS2 and *S. cerevisiae* with ITS1) gave amplicons that were lost because of their length. This raises the question of how much diversity is unknowingly lost in metagenomic metabarcoding studies (with actual environmental samples) when ITS is sequenced using an Illumina platform (HiSeq or MiSeq). With this work, we confirmed the issue related to preferential amplification and ITS markers<sup>6</sup> and uncovered, for the first time, to our knowledge, ITS issues related to the length of the amplicons and sequencing platforms.

A final issue was related to the analysis of the data. We showed that, by using different markers and therefore databases, different results were obtained in terms of taxonomical assignment, with some markers (i.e. 28S rRNA) being particularly weak in this aspect<sup>6</sup>. We harmonised the outcomes of different markers in this study by evaluating all of the OTUs in the NCBI web tool BLASTN. This issue can make the comparison of different studies misleading, particularly when different markers and databases are used and the taxonomy assignment is not cross-validated. Moreover, an attempt to unify the nomenclature of all databases should be made, along with a thorough evaluation of spurious products, especially for SSU and LSU markers.

In conclusion, we have demonstrated how three main steps (DNA extraction, choice of primer and taxonomy assignment) of amplicon sequencing for the characterisation of the gut mycobiome can bias results. It is therefore important to test these methods, select them according to the environment studied and be aware of their limitations before drawing conclusions. Based on the observations presented in this study, we believe that combining two extraction methods (PSB and QSK) allows the extraction of fungal DNA from stool to be maximised. For the characterization of the gut mycobiome we recommend the 18S rRNA primer set. This primer set is highly specific towards fungi, allowing phylogenetic analysis and producing a classification comparable to ITS.

## Methods

**Preparation of cells.** Seven fungal strains were tested: *Candida albicans*, *Candida tropicalis*, *Saccharomyces cerevisiae*, *Cryptococcus neoformans*, *Malassezia furfur*, *Aspergillus fumigatus* and *Penicillium chrysogenum*. All were clinical isolates (Liverpool Clinical Laboratories of Infection and Immunity, Royal Liverpool and Broadgreen University Hospitals, NHS Trust), except for *Malassezia furfur*, which was from the CBS collection (CBS 1878). The identity of all strains was confirmed by Sanger sequencing (GATC Biotech Ltd.) of ITS2 with ITS3tagmix/ITS4NGS PCR product of each of these (for sequences, see Supplementary Data S1). *C. albicans*, *C. tropicalis*, *S. cerevisiae* and *C. neoformans* were grown in Yeast extract-peptone-dextrose agar (YPD; Sigma-Aldrich) containing 26 mg/L chloramphenicol (Sigma-Aldrich) for 2 days at 30 °C. Afterwards, a single colony of each strain was inoculated in 8 mL of YPD broth at 30 °C, with overnight shaking at 200 rpm. *M. furfur* was

grown in modified Dixon's (mDixon) agar<sup>52</sup> containing 26 mg/L chloramphenicol for 3–4 days at 32 °C. An 8 mL volume of mDixon broth<sup>52</sup> was then inoculated with a single colony and incubated at 32 °C, shaking at 200 rpm for 48 hours. To assess cell concentration, the  $OD_{600}$  of a ten-fold dilution from each culture broth was measured with a GeneQuant Pro RNA/DNA calculator spectrophotometer (GE Healthcare). A standard curve of  $OD_{600}$ /CFU was previously produced for each strain. Quarter strength sterile Ringer solution (Sigma-Aldrich) was used as diluent. *A. fumigatus* and *P. chrysogenum* were inoculated in Sabouraud dextrose agar (SDA) containing chloramphenicol (Sigma-Aldrich); *A. fumigatus* for three days at 37 °C, and *P. chrysogenum* for 5–7 days at room temperature. Spores were harvested from the SDA plate with a sterile swab and suspended in 10 mL of sterile PBS pH 7.3 containing 0.05% v/v Tween 20 (Sigma-Aldrich). Following incubation for 10 minutes to allow the hyphae to deposit in the bottom, a 2 mL volume was then taken from the top, and spores were counted with a Neubauer improved haemocytometer. Yeast cell and spore concentrations were adjusted to spike 1.2 mL of Lysis Buffer P (PSP®Spin Stool DNA Kit, Stratec) with  $1 \times 10^6$  cells/spores. Because the QIAamp®Fast DNA Stool Mini Kit (Qiagen) protocol carries a lower volume of starting material than the PSP kit (200  $\mu$ L instead of 800  $\mu$ L), the number of cells of *C. albicans* and spores of *A. fumigatus* were increased to use the same amount of material. DNA was extracted immediately (see DNA extraction section).

**Stool samples.** A stool sample from a healthy volunteer was collected and immediately brought to the laboratory for processing. The spiking and aliquoting processes are summarised in Fig. 1B. Briefly, twelve aliquots were made; four were spiked with  $10^6$  cells of each species, four with  $10^4$  cells and four were not spiked. On the day of the extraction, 3.6 mL of Lysis Buffer P (PSP®Spin Stool DNA Kit, Stratec) was added to two aliquots of each spike ( $10^6$ ,  $10^4$ , non-spiked) and DNA was extracted immediately (see next section for details). 3.6 mL of Stabilizer (PSP Spin Stool DNA Plus Kit, Stratec) was added to a further two aliquots of each spike ( $10^6$ ,  $10^4$ , non-spiked), which were vortexed briefly and incubated at RT overnight. Each tube was split in three and three extractions for each extraction method and spike were carried out in parallel. Because the extraction with the Qiagen kit was performed at a later timepoint, the sample was frozen, not fresh, when spiked. Nevertheless, the sample was aliquoted, spiked and stored as described above.

Stool samples used to compare the performance of 18S rRNA and ITS2 primers were collected at the Royal Liverpool University Hospital (UK). In accordance with ethical approval from 15/NW/0045, as part of the EU-funded SysMedIBD (Systems medicine of chronic Inflammatory Bowel Disease) Project. Written, informed consent was obtained from volunteers before they donated a faecal sample. Samples were brought to the laboratory the same day of clinic and stored at  $-80$  °C until extraction. DNA was extracted from 220 mg of stool.

**DNA extraction.** Half of the samples extracted with the PSP kit underwent a bead beating step. For this step, 0.5 g of 0.1 and 0.5 mm zirconia/silica beads (BioSpec) and 5 Zirconia II beads from the PSP®Spin Stool DNA Kit were used. Briefly, all the samples were first heated at 95 °C, with mixing at 900 rpm for 10 minutes in a Thermomixer (Eppendorf). Half of the samples were then processed following the kit protocol (Protocol 2<sup>53</sup>); the other half were bead beaten for a total of 6 minutes at 30 Hz in a TissueLyser II (Qiagen®). After every 1.5 minutes of bead beating, the samples were incubated in ice for one minute. After this step, all the samples were processed in parallel as per kit protocol. The extraction with the QIAamp®Fast DNA Stool Mini Kit (Qiagen®) was performed following the kit protocol "Isolation of DNA from Stool for Pathogen Detection"<sup>54</sup>.

A negative control (without starting material) was also extracted each time. For the extraction from cells or spores, DNA was eluted from the column in 100  $\mu$ L buffer; when extracting from stool samples, the elution was repeated for a total of 200  $\mu$ L. Extractions from pure cultures were done in triplicate and each replicate was analysed separately. For the spiked stool sample, extractions were carried out in triplicate, and 20  $\mu$ L from each replicate was pooled and used for downstream analysis. Following extraction, DNA was quantified with a Qubit (Qubit dsDNA HS assay Kit, Life Technologies). Stool samples from donors were extracted with the methods PSb and Q (Table 1). The Qubit assay was used to quantify DNA, which was normalised to 10 ng/ $\mu$ L and pooled.

**Gene quantification.** 18S rRNA gene quantification was achieved using the FungiQuant Assay<sup>42</sup>. PCR reactions were made in a total volume of 10  $\mu$ L using 1X PrimeTime® Gene Expression Master Mix (IDT), 1.8  $\mu$ M of each primer<sup>42</sup> (Table 3), 225 nM of LNA PrimeTime probe (5'-6-FAM/TG + GTG + CATGG + CC + GTT/3 IABkFQ) (IDT), 3.5  $\mu$ L of template and Ultrapure DNase/RNase-free water (Life Technologies). The reactions were run in a Roche LightCycler 480 with the following program: 95 °C for 3 minutes, then 45 cycles of 95 °C for 15 seconds and 65 °C for 1 minute. To assess assay efficiency, a calibration curve was made for each species where a quantified amount of genomic DNA was diluted, and serial ten-fold dilutions made. To quantify the number of gene copies in stool samples, a calibration curve was made with known concentration of a gBlock (IDT) of the 18S rRNA amplicon sequence from *C. albicans* SC5314. Cp values and efficiencies were calculated with the LightCycler 480 Software (v 1.5.0)<sup>55</sup> and gene quantification analyses were made using the Second Derivative Maximum Method<sup>56</sup>. Samples were run in triplicate along with triplicates of a negative control (no template) and a positive control (calibrator). All negatives controls were clean (no signal), or when a signal was detected this had a Cp value  $>3$  compared to the most diluted point of the calibration curve. Efficiencies and calibration curves are shown in the Supplementary Fig. S3.

**Amplicon production and sequencing.** The primers tested are shown in Table 3. A universal tail tag dual index barcoding approach<sup>46</sup> was used. In the first PCR round, reactions were made in a total volume of 20  $\mu$ L, with 0.02 U/ $\mu$ L Q5 High-Fidelity DNA Polymerase (NEB), 1X Q5 Reaction Buffer (NEB), 0.125  $\mu$ M of each primer (overhang + specific primer, Table 3) (HPLC grade, IDT), 200  $\mu$ M of dNTPs (NEB) and Ultrapure DNase/RNase-free water (Life Technologies) to reach the volume. For the stool samples, 10 ng of DNA was used per reaction. Meanwhile, two types of mock community amplicons were produced. In the first, DNA from each

| Name                   | Sequence 5' to 3'                             | Marker   | References |
|------------------------|---|----------|------------|
| Forward overhang       | ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT   | na       | 46         |
| Reverse overhang       | GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC T |          | 46         |
| FungiQuant-F (forward) | GGR AAA CTC ACC AGG TCCA G                    | 18S rRNA | 42         |
| FungiQuant-R (reverse) | GSW CTA TCC CCA KCA CGA                       |          | 42         |
| ITS1-F (forward)       | CTT GGT CAT TTA GAG GAA GTA A                 | ITS1     | 40         |
| ITS2r (reverse)        | GCT GCG TTC TTC ATC GAT GC                    |          | 41         |
| ITS3tagmix (forward)   | CTA GAC TCG TCA NCG ATG AAG AAC GYR G         | ITS2     | 34         |
| gITS7 (forward)        | GTG ART CAT CGA RTC TTT G                     |          | 86         |
| ITS4NGS (reverse)      | TTC CTS CGC TTA TTG ATA TGC                   | 28S rRNA | 34         |
| LR0Rngs (forward)      | ACS CGC TGA ACT TAA GC                        |          | 34         |
| LF402 (reverse)        | TTC CCT TTY ARC AAT TTC AC                    |          | 34         |

**Table 3.** Primers used for amplicon production and qPCR.

species was adjusted to the same Cp value (see Gene quantification section for details) and pooled (PreMC). In the second, DNA from each species was amplified separately and the same amount was pooled after the second PCR (POSTMC). Samples were amplified in a thermocycler (Multigene® and Multigene® OptiMax, Labnet International) with the following program: 98 °C for 30 seconds, then 15 cycles at 98 °C for 10 seconds, followed by 55 °C (ITS2 and 28S primer sets)/56.1 °C (ITS1)/62 °C (18S) for 30 seconds and 72 °C for 20 seconds, followed by a final extension at 72 °C for 2 minutes. Amplicons were then purified with the AxyPrep Mag PCR Clean-up kit (Axygen) and eluted in 10 µL.

The second PCR was carried out using the same amount of polymerase, buffer and dNTPs as per the first PCR. However, this time 0.25 µM of each index primer<sup>46</sup> (TruGrade, IDT) was used. The whole volume of purified amplicons was used as template. The amplification program was 98 °C for 30 seconds, then 20 cycles at 98 °C for 10 seconds, followed by 65 °C for 30 seconds and 72 °C for 20 seconds, followed by a final extension at 72 °C for 2 minutes. Negative controls of the stool DNA extraction and of the PCR (no template) were also amplified; these did not give a visible product when run on an agarose gel. However, a few µL were included in the pool and sequenced. All of the amplifications were performed in triplicate, and the replicates were pooled for a total of 60 µL which was purified with the AxyPrep Mag PCR Clean-up kit and eluted in 25 µL. Amplicons were quantified with a Qubit dsDNA HS assay Kit and electrophoresis was performed using a 1.3% w/v agarose gel (UltraPure™ Agarose, Invitrogen). Amplicons were then pooled equimolarly.

Samples were then sent to the University of Liverpool CGR for sequencing. There, the pool of amplicons was quantified by Qubit and the fragment distribution was assessed using an Agilent DNA high sensitivity kit (2100 Bioanalyzer, Agilent). The pool of amplicons was size selected within the 250–750 bp range using a 1.5% w/v agarose cassette on a PippinPrep instrument. The size-selected pool was quantified again by Qubit. The fragment distribution was checked and further assessed by qPCR, using the Illumina Library Quantification Kit (KAPA, cat. no. KK4854; Roche) on a Roche Light Cycler LC480II according to manufacturer instructions. The concentration obtained by qPCR was used in combination with Bioanalyzer data to prepare the template DNA for the sequencing run. To increase library diversity, 15% PhiX was used as spike-in. The sequencing was carried out on one lane of an Illumina HiSeq. 2500 with v2 2 × 300 bp paired-end sequencing chemistry. Stool samples from donors were prepared similarly, except that the fragment distribution of individual amplicons was assessed with an Agilent DNA high sensitivity kit before pooling. The length and the concentration were then used to pool the samples equimolarly. Libraries were sequenced with a MiSeq with v2 2 × 250 bp paired-end sequencing chemistry.

**Data analysis.** qPCR data were analysed in R (v 3.4.2)<sup>57</sup>. Barplots and a boxplot were produced with ggbarplot() and ggpaired() functions respectively.

Demultiplexing, adaptor and quality trimming (Cutadapt v 1.2.1<sup>58</sup> and Sickle v 1.2.59) of reads was performed by the University of Liverpool CGR. Reads were then error-corrected with BayesHammer<sup>60,61</sup> in SPAdes (v 3.7.0)<sup>62</sup>, merged with PEAR (v 0.9.10)<sup>63</sup> and PhiX was removed. Primers were trimmed and samples with less than 1000 reads were excluded. Clustering was done with SWARM 2.0 (d = 3 for 18S and 28S rRNA and d = 2 for ITS)<sup>64</sup>. ITS reads were also processed with ITSx (v 1.0.11)<sup>65</sup>. Chimeras were filtered out with UCHIME<sup>66</sup> ref mode. Taxonomy was assigned with BLAST<sup>67</sup>, with assign\_taxonomy.py in MacQIIME (v 1.9.1)<sup>68</sup>. Reference databases used were SILVA<sup>69</sup> (SILVA\_123) for 18S rRNA, a curated database for 28S rRNA<sup>70</sup>, UNITE<sup>71</sup> (01.12.2017 release) for ITS. An OTU table was produced (make\_otu\_table.py) and those of 18S and 28S rRNA amplicons were also filtered (filter\_otus\_from\_otu\_table.py) at a 0.05% threshold<sup>72</sup>. For 18S rRNA reads, alignment (align\_seqs.py) and phylogeny (make\_phylogeny.py) analyses were also carried out using MacQIIME (v 1.9.1)<sup>68</sup> default algorithms (PyNAST<sup>73</sup> and FastTree<sup>74</sup>). Aligned OTUs were filtered (filter\_alignment.py) to remove the 10% most variable positions, with an 80% gap filter threshold. It must be noted that the names of taxa assigned to OTUs were not always consistent among these databases and in some cases no name was given. Therefore, to unify the nomenclature among the taxonomy summary graphs (Figs 2B, 3B, 4B, 5C,D and 6B,D), OTUs were also run in BLASTN webtool<sup>75</sup> excluding uncultured and environmental sample sequences, and the names were then changed manually. When the identity percentage was 100%, the name of the species was used; when this was <100%, genera were reported. OTUs assigned to the same species (or genus) are merged in the taxonomy summary. Mismatches between the primers and the reference sequences of the species of the mock community were

assessed with Primer Prospector<sup>76</sup> (analyze\_primers.py function). Reference sequences were downloaded from NCBI<sup>77</sup>. If a reference sequence was not available for a species, the sequence of another species from the same genus was used. 18S rRNA, 28S rRNA and ITS2 sequences were aligned to explore the similarity among and inflation of OTUs. This alignment was made with MEGA X<sup>78</sup>, using clustalW<sup>79</sup> (default parameters).

Statistical analyses were performed in R (v 3.4.2)<sup>57</sup> using the OTU tables, phylogenetic trees (where applicable) and metadata associated with this study. Samples with less than 2000 reads were excluded from analysis. This threshold was justified by the need to find a minimum number of reads that would allow most of the alpha diversity to be gathered, as we found that samples with a number of reads lower than this threshold do not reach a plateau in the rarefaction analysis. In general, the majority of analyses were performed using the Vegan<sup>80</sup> package including rarefaction curves (rarefy() function) and beta diversity analysis. For beta diversity, non-metric distance scaling (NMDS) was performed using two different distance measures in metacmds() function: (i) Bray-Curtis and (ii) unweighted UniFrac<sup>81</sup>. For Unifrac distances, the Phyloseq<sup>82</sup> package was employed. Additionally, ordiellipse() function was used to draw the 95% confidence interval of standard errors on NMDS plots. Analysis of variance (PERMANOVA) against sources of variations (groups in this study) was performed using Vegan's adonis() function against the distance matrices as described above, while aov() was used to calculate pair-wise ANOVA. To give an account of abundant OTUs, proportional abundances of top N OTUs were plotted using a custom script. These scripts and workflows are available at Dr. Umer Z. Ijaz webpage<sup>83</sup>.

**Ethical approval.** Ethical approval for the use of stool samples (SysMedIBD; European Union Seventh Framework Programme (FP7) under grant agreement no. 305564) was granted by NRES Committee North West - Liverpool East, reference (15/NW/0045).

### Data Availability

The datasets are available at the European Nucleotide Archive<sup>84</sup> under study accession numbers ERP111304 and ERP111352. The first includes the reads produced during the testing phase of the 5 primer sets, when mock community and spiked stool sample were used as template. The second includes the amplicons produced with 18S rRNA and ITS2 primers, in which the templates were stool samples from the SysMed IBD project.

Raw data related to gene quantification experiments can be found in the platform Figshare<sup>85</sup>.

### References

- Richard, M. L., Lamas, B., Liguori, G., Hoffmann, T. W. & Sokol, H. Gut fungal microbiota: the Yin and Yang of inflammatory bowel disease. *Inflammatory Bowel Diseases* **21**, 656–665, <https://doi.org/10.1097/MIB.0000000000000261> (2014).
- Liguori, G. *et al.* Fungal dysbiosis in mucosa-associated microbiota of Crohn's disease patients. *Journal of Crohn's and Colitis* **10**, 296–305, <https://doi.org/10.1093/ecco-jcc/jjv209> (2015).
- Hoarau, G. *et al.* Bacteriome and mycobiome interactions underscore microbial dysbiosis in familial Crohn's disease. *Mbio* **7**, e01250–16, <https://doi.org/10.1128/mBio.01250-16> (2016).
- Sokol, H. *et al.* Fungal microbiota dysbiosis in IBD. *Gut* **66**, 1039–1048, <https://doi.org/10.1136/gutjnl-2015-310746> (2017).
- Tedersoo, L. & Lindahl, B. Fungal identification biases in microbiome projects. *Environmental Microbiology Reports* **8**, 774–779, <https://doi.org/10.1111/1758-2229.12438> (2016).
- De Filippis, F., Laiola, M., Blaiotta, G. & Ercolini, D. Different amplicon targets for sequencing-based studies of fungal diversity. *Applied and Environmental Microbiology* **83**, <https://doi.org/10.1128/AEM.00905-17> (2017).
- Huseyin, C. E., Rubio, R. C., O'Sullivan, O., Cotter, P. D. & Scanlan, P. D. The fungal frontier: a comparative analysis of methods used in the study of the human gut mycobiome. *Frontiers in Microbiology* **8**, 1432, <https://doi.org/10.3389/fmicb.2017.01432> (2017).
- Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19, <https://doi.org/10.1186/2049-2618-2-19> (2014).
- Guo, F. & Zhang, T. Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Applied Microbiology and Biotechnology* **97**, 4607–4616, <https://doi.org/10.1007/s00253-012-4244-4> (2013).
- Kim, D. *et al.* Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **5**, 52, <https://doi.org/10.1186/s40168-017-0267-5> (2017).
- Moore, D., Robson, G. D. & Trinci, A. P. *21st century guidebook to fungi* (Cambridge University Press, 2011).
- Erwig, L. P. & Gow, N. A. Interactions of fungal pathogens with phagocytes. *Nature Reviews Microbiology* **14**, 163, <https://doi.org/10.1038/nrmicro.2015.21> (2016).
- Karakousis, A., Tan, L., Ellis, D., Alexiou, H. & Wormald, P. An assessment of the efficiency of fungal DNA extraction methods for maximizing the detection of medically important fungi using PCR. *Journal of Microbiological Methods* **65**, 38–48, <https://doi.org/10.1016/j.mimet.2005.06.008> (2006).
- Fredricks, D. N., Smith, C. & Meier, A. Comparison of six DNA extraction methods for recovery of fungal DNA as assessed by quantitative PCR. *Journal of Clinical Microbiology* **43**, 5122–5128, <https://doi.org/10.1128/JCM.43.10.5122-5128.2005> (2005).
- Saunders, G. C. & Rossi, J. M. DNA extraction. In Keer, J. T. & Birch, L. (eds) *Essentials of Nucleic Acid Analysis: A Robust Approach*, chap. 4, 59–82, <https://doi.org/10.1039/9781847558213-00059> (Royal Society of Chemistry, 2008).
- Scanlan, P. D. & Marchesi, J. R. Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *The ISME Journal* **2**, 1183–1193, <https://doi.org/10.1038/ismej.2008.76> (2008).
- Ukhanova, M. *et al.* Effects of almond and pistachio consumption on gut microbiota composition in a randomised cross-over human feeding study. *British Journal of Nutrition* **111**, 2146–2152, <https://doi.org/10.1017/S0007114514000385> (2014).
- Ott, S. J. *et al.* Fungi and inflammatory bowel diseases: alterations of composition and diversity. *Scandinavian Journal of Gastroenterology* **43**, 831–841, <https://doi.org/10.1080/00365520801935434> (2008).
- David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563, <https://doi.org/10.1038/nature12820> (2014).
- Li, Q. *et al.* Dysbiosis of gut fungal microbiota is associated with mucosal inflammation in Crohn's disease. *Journal of Clinical Gastroenterology* **48**, 513–523, <https://doi.org/10.1097/MCG.0b000000000000035> (2014).
- Chehoud, C. *et al.* Fungal signature in the gut microbiota of pediatric patients with inflammatory bowel disease. *Inflammatory Bowel Diseases* **21**, 1948–1956, <https://doi.org/10.1097/MIB.0000000000000454> (2015).
- Heisel, T. *et al.* Complementary amplicon-based genomic approaches for the study of fungal communities in humans. *PLoS One* **10**, 1–20, <https://doi.org/10.1371/journal.pone.0116705> (2015).
- Qiu, X. *et al.* Changes in the composition of intestinal fungi and their role in mice with dextran sulfate sodium-induced colitis. *Scientific Reports* **5**, 10416, <https://doi.org/10.1038/srep10416> (2015).

24. Chen, Y. *et al.* Correlation between gastrointestinal fungi and varying degrees of chronic hepatitis B virus infection. *Diagnostic Microbiology and Infectious Disease* **70**, 492–498, <https://doi.org/10.1016/j.diagmicrobio.2010.04.005> (2011).
25. Doolive, S. *et al.* A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biology* **13**, R60, <https://doi.org/10.1186/gb-2012-13-7-r60> (2012).
26. Iliev, I. D. *et al.* Interactions between commensal fungi and the C-type lectin receptor Dectin-1 influence colitis. *Science* **336**, 1314–1317, <https://doi.org/10.1126/science.1221789> (2012).
27. Pandey, P. K. *et al.* Molecular typing of fecal eukaryotic microbiota of human infants and their respective mothers. *Journal of Biosciences* **37**, 221–226, <https://doi.org/10.1007/s12038-012-9197-3> (2012).
28. Hallen-Adams, H. E., Kachman, S. D., Kim, J., Legge, R. M. & Martinez, I. Fungi inhabiting the healthy human gastrointestinal tract: a diverse and dynamic community. *Fungal Ecology* **15**, 9–17, <https://doi.org/10.1016/j.funeco.2015.01.006> (2015).
29. Luan, C. *et al.* Dysbiosis of fungal microbiota in the intestinal mucosa of patients with colorectal adenomas. *Scientific Reports* **5**, 7980, <https://doi.org/10.1038/srep07980> (2015).
30. Tang, J., Iliev, I. D., Brown, J., Underhill, D. M. & Funari, V. A. Mycobiome: approaches to analysis of intestinal fungi. *Journal of Immunological Methods* **421**, 112–121, <https://doi.org/10.1016/j.jim.2015.04.004> (2015).
31. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* **109**, 6241–6246, <https://doi.org/10.1073/pnas.1117018109> (2012).
32. Nilsson, R. H. *et al.* variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics* **4**, EBO–S653, <https://doi.org/10.4137/EBO.S653> (2008).
33. Bazzicalupo, A. L., Bálint, M. & Schmitt, I. Comparison of ITS1 and ITS2 rDNA in 454 sequencing of hyperdiverse fungal communities. *Fungal Ecology* **6**, 102–109, <https://doi.org/10.1016/j.funeco.2012.09.003> (2013).
34. Tedersoo, L. *et al.* Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycKeys* **1–43**, <https://doi.org/10.3897/mycokeys.10.4852> (2015).
35. Bruns, T. D. & Shefferson, R. P. Evolutionary studies of ectomycorrhizal fungi: recent advances and future directions. *Canadian Journal of Botany* **82**, 1122–1132, <https://doi.org/10.1139/b04-021> (2004).
36. O'Brien, H. E., Parrent, J. L., Jackson, J. A., Moncalvo, J.-M. & Vilgalys, R. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology* **71**, 5544–5550, <https://doi.org/10.1128/AEM.71.9.5544-5550.2005> (2005).
37. Brown, S. P., Rigdon-Huss, A. R. & Jumpponen, A. Analyses of ITS and LSU gene regions provide congruent results on fungal community responses. *Fungal Ecology* **9**, 65–68, <https://doi.org/10.1016/j.funeco.2014.02.002> (2014).
38. Mueller, R. C., Gallegos-Graves, L. V. & Kuske, C. R. A new fungal large subunit ribosomal RNA primer for high-throughput sequencing surveys. *FEMS Microbiology Ecology* **92**, <https://doi.org/10.1093/femsec/fiv153> (2016).
39. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology* **19**, 21–31, <https://doi.org/10.1111/j.1365-294X.2009.04480.x> (2010).
40. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for basidiomycetes-application to the identification of mycorrhizae and rusts. *Molecular Ecology* **2**, 113–118, <https://doi.org/10.1111/j.1365-294X.1993.tb00005.x> (1993).
41. White, T. J., Bruns, T., Lee, S. & Taylor, J. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (eds) *PCR protocols: a guide to methods and applications*, chap. 38, 315–322 (Academic Press, San Diego, CA, 1990).
42. Liu, C. M. *et al.* FungiQuant: a broad-coverage fungal quantitative real-time PCR assay. *BMC Microbiology* **12**, 255, <https://doi.org/10.1186/1471-2180-12-255> (2012).
43. Illumina. Technical Note: DNA Sequencing. Nextera® Library Validation and Cluster Density Optimization. Tech. Rep., Illumina (Pub. No. 770-2013-003, 23 November 2014).
44. Zalar, P. *et al.* Redefinition of *Aureobasidium pullulans* and its varieties. *Studies in Mycology* **61**, 21–38, <https://doi.org/10.3114/sim.2008.61.02> (2008).
45. Sugita, T., Tajima, M., Amaya, M., Tsuboi, R. & Nishikawa, A. Genotype analysis of *Malassezia restricta* as the major cutaneous flora in patients with atopic dermatitis and healthy subjects. *Microbiology and Immunology* **48**, 755–759, <https://doi.org/10.1111/j.1348-0421.2004.tb03601.x> (2004).
46. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55, <https://doi.org/10.1186/s12864-015-2194-9> (2016).
47. Thomas, V., Clark, J. & Doré, J. Fecal microbiota analysis: an overview of sample collection methods and sequencing strategies. *Future Microbiology* **10**, 1485–1504, <https://doi.org/10.2217/fmb.15.87> (2015).
48. Griffiths, L. J. *et al.* Comparison of DNA extraction methods for *Aspergillus fumigatus* using real-time PCR. *Journal of Medical Microbiology* **55**, 1187–1191, <https://doi.org/10.1099/jmm.0.46510-0> (2006).
49. Nawrot, U. *et al.* Comparison of the utility of five commercial kits for extraction of DNA from *Aspergillus fumigatus* spores. *Acta Biochimica Polonica* **57**, 567–571, [https://doi.org/10.18388/abp.2010\\_2445](https://doi.org/10.18388/abp.2010_2445) (2010).
50. White, P. L. *et al.* *Aspergillus* PCR: one step closer to standardization. *Journal of Clinical Microbiology* **48**, 1231–1240, <https://doi.org/10.1128/JCM.01767-09> (2010).
51. Bjørnsgaard Aas, A., Davey, M. L. & Kausrud, H. ITS all right mama: investigating the formation of chimeric sequences in the ITS 2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* **17**, 730–741, <https://doi.org/10.1111/1755-0998.12622> (2017).
52. Guého-Kellermann, E., Boekhout, T. & Begerow, D. Biodiversity, phylogeny and ultrastructure. In *Malassezia and the skin*, 17–63, [https://doi.org/10.1007/978-3-642-03616-3\\_2](https://doi.org/10.1007/978-3-642-03616-3_2) (Springer, 2010).
53. STRATEC Molecular GmbH, Berlin, Germany. *User manual PSP® Spin Stool DNA Kit/User manual PSP® Spin Stool DNA Plus Kit* (2013).
54. QIAGEN®. *QIAamp® Fast DNA Stool Mini Handbook 03/2014* (2014).
55. Roche. *LightCycler® 480 Instrument Operator's Manual, Software Version 1.5* (2008).
56. Rasmussen, R. Quantification on the LightCycler. In Meuer, S., Wittwer, C. & Nakagawara, K.-I. (eds) *Rapid Cycle Real-Time PCR: Methods and Applications*, 21–34, [https://doi.org/10.1007/978-3-642-59524-0\\_3](https://doi.org/10.1007/978-3-642-59524-0_3) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2001).
57. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12, <https://doi.org/10.14806/ej.17.1.200> (2011).
59. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)[Software]. link to Version 1.2, <https://github.com/najoshi/sickle/releases/tag/v1.2>, Online; accessed 30 January 2019 (2011).
60. Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. In *BMC Genomics*, vol. 14, S7, <https://doi.org/10.1186/1471-2164-14-S1-S7> (BioMed Central, 2013).
61. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research* **43**, e37, <https://doi.org/10.1093/nar/gku1341> (2015).
62. Nurk, S. *et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology* **20**, 714–737, <https://doi.org/10.1089/cmb.2013.0084> (2013).

63. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620, <https://doi.org/10.1093/bioinformatics/btt593> (2013).
64. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, e1420, <https://doi.org/10.7717/peerj.1420> (2015).
65. Bengtsson-Palme, J. *et al.* Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution* **4**, 914–919, <https://doi.org/10.1111/2041-210X.12073> (2013).
66. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200, <https://doi.org/10.1093/bioinformatics/btr381> (2011).
67. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410, [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) (1990).
68. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336, <https://doi.org/10.1038/nmeth.f.303> (2010).
69. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596, <https://doi.org/10.1093/nar/gks1219> (2012).
70. Liu, K.-L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A. & Xie, G. Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. *Applied and Environmental Microbiology* **78**, 1523–1533, <https://doi.org/10.1128/AEM.06826-11> (2012).
71. Kõljalg, U. *et al.* Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**, 5271–5277, <https://doi.org/10.1111/mec.12481> (2013).
72. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**, 57–60, <https://doi.org/10.1038/nmeth.2276> (2013).
73. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267, <https://doi.org/10.1093/bioinformatics/btp636> (2009).
74. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, 1–10, <https://doi.org/10.1371/journal.pone.0009490> (2010).
75. NCBI. Nucleotide Blast (blastn), <https://blast.ncbi.nlm.nih.gov> (Online; accessed August–September 2018).
76. Walters, W. A. *et al.* Primerprospector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161, <https://doi.org/10.1093/bioinformatics/btr087> (2011).
77. NCBI. Nucleotide database, <https://blast.ncbi.nlm.nih.gov> (Online; accessed February 2019).
78. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution* **35**, 1547–1549, <https://doi.org/10.1093/molbev/msy096> (2018).
79. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948, <https://doi.org/10.1093/bioinformatics/btm404> (2007).
80. Oksanen, J. *et al.* *vegan: Community Ecology Package*, R package version 2.5-1 (2018).
81. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228–8235, <https://doi.org/10.1128/AEM.71.12.8228-8235.2005> (2005).
82. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217, <https://doi.org/10.1371/journal.pone.0061217> (2013).
83. Ijaz, U. Z. Dr Umer Zeeshan Ijaz bioinformatic resources, <http://userweb.eng.gla.ac.uk/umer.ijaz#bioinformatics> (Online; accessed 8 October 2018).
84. EMBL-EBI. European Nucleotide Archive (ENA), <https://www.ebi.ac.uk/ena> (Online; accessed 8 October 2018).
85. Frau, A. *et al.* qPCR raw data relevant to this study are available in Figshare, <https://doi.org/10.6084/m9.figshare.8138018> (Online; accessed 16 May 2019).
86. Ihrmark, K. *et al.* New primers to amplify the fungal ITS2 region—evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* **82**, 666–677, <https://doi.org/10.1111/j.1574-6941.2012.01437.x> (2012).

## Acknowledgements

This work was funded by Crohn's and Colitis UK (CCUK) (grant number M/15/4, principal investigator CSJP). UZI was funded by the NERC IRF NE/L011956/1. The authors would also like to thank: the technical support of CGR, in particular Dr Anita Lucaci for carrying out the experiments concerning the sequencing of the samples and Dr Richard Gregory for the bioinformatics support; Dr Andrea Ockhardt from Stratec for her advice on DNA extraction from stool samples and for providing a free DNA extraction kit; Prof William Hope from the University of Liverpool for sharing *A. fumigatus* and *C. neoformans* strains, and his group for the advice and support on setting up the cultures; Diana Cox from Unilever for sharing the *M. furfur* strain and for her advice on setting up the cultures; Jennifer C Cornell from Reads Well Support Services for the professional editing and proofreading; and Dave Berry, the technical support in Gastroenterology Research Unit (University of Liverpool), for his constant support in the lab.

## Author Contributions

A.F., J.G.K., A.C.D., N.H. and C.S.J.P. conceptualized this work. A.F., J.G.K., B.J.C., C.A.D., M.D.B., N.H., J.A., A.C.D. and C.S.J.P. were involved in the development and design of the methods. A.F., J.G.K., L.L., C.A.D., M.D.B., A.C.D. and C.S.J.P. were also involved in the validation of the work. A.F. carried out the investigation, formal analysis and wrote the original manuscript. This was then edited and reviewed by A.F., J.G.K., L.L., B.J.C., U.Z.I., C.A.D., M.D.B., N.H., A.C.D. and C.S.J.P. Scripts for data analysis were created by L.L. and U.Z.I. and marginally by A.F. A.F., L.L. and U.Z.I. were also involved in data curation. Visualization and data presentation were carried out by A.F. with the contribution of U.Z.I. C.S.J.P., B.J.C. and A.C.D. supervised the work. C.S.J.P. was also administrator of the project. B.J.C., N.H., J.A. and C.S.J.P. were involved in resources and funding acquisition.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-019-44974-x>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019