Quantitative metabarcoding of soil fungi and bacteria

Miguel Camacho-Sanchez (miguel.camacho@juntadeandalucia.es)
Instituto Andaluz de Investigación y Formación Agraria, Alimentaria y de la Producción Ecológica (IFAPA) Centro Las Torres

Research Article

Keywords: metabarcoding, spike-in, internal standards, compositional data, soil microbiology

Posted Date: May 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2885222/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Metabarcoding is a powerful tool to characterize biodiversity in biological samples. The interpretation of taxonomic profiles from metabarcoding data has been hindered by their compositional nature. Several strategies have been proposed to transform compositional data into quantitative, with their intrinsic limitations. Here, I propose a workflow based on bacterial and fungal cellular internal standards (spike-ins) for absolute quantification of the microbiota in soil samples. These standards were added to the samples before DNA extraction in amounts estimated after qPCRs, to target around 1-2% coverage in the sequencing run. In bacteria, proportions of spike-in reads in the sequencing run were very similar (< 2-fold change) to those predicted by the qPCR assessment, but for fungi they differed up to 40-fold. The little variation between replicated samples highlights the reproducibility of the method. Estimates based on multiple bacterial spike-ins were highly correlated ($r = 0.99$). Procrustes analysis evidenced significant biological effects on the community composition when normalizing compositional data. A protocol based on qPCR estimation of input amounts of cellular spikes is proposed as a cheap and reliable strategy for quantitative metabarcoding of biological samples.

Introduction

DNA metabarcoding is an efficient way to assess the biodiversity of complex bulk samples via the sequencing of a DNA region shared among a phylogenetic group of taxa (Taberlet et al. 2012; Ruppert et al. 2019). A main constraint of metabarcoding data relates to its compositional nature, i.e., the sum of all its elements adds up to a constant (Tsilimigras and Fodor 2016; Gloor et al. 2017). This intrinsic characteristic dictates the number of reads output by the sequencing instrument is not correlated to the counts of features present in the biological sample. Thus, read counts only reflect within sample relative abundances, and their frequencies cannot be compared between samples (McMurdie and Holmes 2014). A set of tools for analyzing compositional data coming from other science domains has been adapted to metabarcoding studies (Tsilimigras and Fodor 2016; Gloor et al. 2017; Quinn et al. 2019). Despite their power to address relevant biological questions (Morton et al. 2019), they fail to assess quantitative changes between samples.

To overcome this limitation, many strategies have been proposed to get quantitative data from compositional data by normalizing with an internal reference, also called “internal standard” or “spike-in”. Most strategies propose adding a known amount of cells, DNA molecules, or genomic DNA, to the sample (Harrison et al. 2021). There are methodological difficulties and limitations to the use of these internal standards (reviewed in Harrison et al., 2021). First, they must capture the variation introduced by the methodological workflow: i.e. DNA extraction, library preparation and sequencing. By adding the internal standard before DNA extraction, it incorporates the variation introduced by different DNA extraction yields (Harrison et al. 2021; Haro et al. 2021). Second, during the workflow, internal standards must behave similarly to the native targeted DNA molecules in the samples. So, the target DNA region from internal standards should have similar characteristics (%GC content, length, primer binding affinity) to those in the target biological species, so that PCR-related biases are minimized (Jones et al. 2015; Bonk et al. 2018;
Shelton et al. 2022). Third, internal standard DNA sequences must be known, to filter them bioinformatically, and must not occur naturally in the biological sample. This can be achieved by using extremophile organisms or synthetic DNA standards (Smets et al. 2016; Stämmler et al. 2016; Hardwick et al. 2018). Forth, the amount of internal standard added to the sample should represent a small fraction (0.1-3%) of the reads in the sequencing run, enough to normalize the reads accurately without wasting sequencing coverage (Jones et al. 2015; Smets et al. 2016; Harrison et al. 2021).

The amount of internal standard to add to samples is difficult to predict due to heterogeneous starting biological samples, template-specific PCR efficiencies (Jones et al. 2015; Bonk et al. 2018), differential yields in DNA extraction across taxa (Dopheide et al. 2019; Haro et al. 2021), and copy number variation (CNV) of target loci (Stoddard et al. 2015; Lofgren et al. 2019). Therefore, the input amount of internal standard must be determined empirically. This has been approximated by counting cells on the microscope (Yang et al. 2018), with the inherent biases pointed out above, or achieved by sequencing a fraction of the samples with increasing dilutions of the internal standard (Yang et al. 2018; Lin et al. 2019). The latter strategy, despite being accurate, involves long return times between library preparation to data analysis, and can be relatively expensive.

Here, I propose a practical workflow for implementing absolute quantification in metabarcoding protocols of bacteria and fungi from soil samples by adding a known number of cells from bacteria and from a fungus. It is hypothesized the proposed workflow provides a fast, cheap, and reliable way to implement quantitative metabarcoding in the study of soil samples.

**Material and methods**

The proposed protocol estimates the amount of internal standard (“spike-in”, hereon) to add to biological samples for a desired coverage in the sequencing run. This is achieved by measuring via quantitative PCR (qPCR) the copy number of each marker in a few representative samples and in a spike-in solution. Then, the input amount of spike-in added to the samples is adjusted based on these quantifications.

**Sampling and sample processing**

A total of twelve soil samples were collected from strawberry fields at the IFAPA experimental station “El Cebollar”, located at Moguer (Huelva, Spain) (37°14′25.4” N 6°48′09.2” W): four different time points, with three replicas at each time, along two consecutive campaigns (Table 1). An auger of ~ 4 cm in diameter was used to sample cores of soil at 5–6 points within the sampling unit. Soil was placed in sterile plastic bags and homogenized manually (Rodriguez-Mena et al. 2022). Samples were lyophilized in the Telstar® LyoQuest freeze-dryer for a minimum of 8h at 0.1 mbar in a chamber at -80°C, until complete dehydration. Dried soil was disaggregated and homogenized by shaking. A homemade setup was used to stream sieving and weighting efficiently (Fig. 1A), and to remove small stones and other organic structures from the sample. Around .25 mg were weighted using a precision scale. The exact weight was recorded for later normalization of the read count.
Table 1

<table>
<thead>
<tr>
<th>samples</th>
<th>date collected</th>
<th>timepoint</th>
<th>campaign</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-C</td>
<td>2019-09-11</td>
<td>beginning</td>
<td>1</td>
</tr>
<tr>
<td>D-F</td>
<td>2019-10-17</td>
<td>beginning</td>
<td>1</td>
</tr>
<tr>
<td>G-I</td>
<td>2020-05-14</td>
<td>end</td>
<td>1</td>
</tr>
<tr>
<td>J-L</td>
<td>2021-05-25</td>
<td>end</td>
<td>2</td>
</tr>
</tbody>
</table>

DNA extraction and quantification of the copy number for the target markers

DNA extractions were done with the DNeasy PowerSoil Kit, following manufacturer’s instructions, in two batches: a first one with a non-spiked subset of samples for qPCRs, and a second one with spiked soil samples for sequencing. Both included a positive control containing only the spike-in (no soil) (Fig. 1B) and blanks of extraction. The spike-in sample consisted of 10 µL of the commercial spike-in (ZymoBIOMICS™ Spike-in Control I # D6320) and 20 µL of a $10^6$ cells/µL solution of the yeast *Yarrowia lipolytica*. It was added directly to the homogenization tube from the DNA extraction kit. The bacterial spike-in had cells from the halophilic bacteria *Imtechella halotolerans* ($10^6$ cells/µL) and *Allobacillus halotolerans* ($10^6$ cells/µL), unlikely to be found in these agricultural soils. For the fungus, the yeast *Yarrowia lipolytica* strain CECT 1240 was selected from the Spanish Collection of Type Cultures (www.uv.es) because it is a unicellular organism that can be counted in plates, and it is easy to grow and resuspend in desired concentrations to be used as a cellular spike-in. *Yarrowia lipolytica* spike-in stock solutions were prepared at concentrations of $10^6$ and $10^7$ ufc/ml (S1). DNA concentrations were measured with the NanoDrop 1000. The number of 16S copies per genome varied in the spiked bacteria: 7 copies/genome in *A. halotolerans* and 3 copies/genome in *I. halotolerans*. ITS2 copy number for *Y. lipolytica* was unknown, so copy estimates were related to the number of cells.

A qPCR was run on twelve DNA extractions corresponding to four selected representative samples (A, D, G and J) that had been extracted in three technical replicas; first batch of DNA extractions with no spike-in. The DNA extract from the spike-in solution was included as a standard. Ribosomal regions from fungi and bacteria commonly used in metabarcoding were selected. For fungi, primers ITS3 (5’ GCATCGATGAAGACGCGGC 3’) and ITS4R (5’ TCCTCCGCTTATGGATATGC 3’) (White et al. 1990) were selected to amplify the ITS2 region, whereas primers 341F (5’ CCTACGGGNGGCWGCAG 3’) and 805R (5’ GACTACHVGGGTATCTAATCC 3’) (Herlemann et al. 2011) were chosen to target the V3-V4 region of bacterial 16S rRNA. Before the qPCR, all soil DNA extracts were diluted to 1/25 to prevent PCR inhibition (Sidstedt et al. 2020), which had been previously detected in this sample set. qPCRs were performed in final volumes of 10 µL, containing 1x of SensiFAST SYBR, 0.5 µM of each primer, 0.2 mg/mL of BSA, and 3.9 µL of template DNA (12–60 ng). The cycling conditions consisted of an initial denaturation at 95°C
for 5 min, followed by 35 cycles of 95°C for 5 s, 58°C for 20 s, and 72°C for 40 s. The plate was read at
the end of each extension step. A melting curve analysis was added after the 35 cycles, from 65 °C to 95
°C, + 5 °C / 5 s to verify a unimodal peak in fragment sizes. All samples were included in technical
duplicates for the 16S and the ITS primer sets, together with standards for relative quantification. The
standard consisted of the DNA extraction from the spike-in in serial dilutions by a factor of 5: 1, 1/5, 1/25,
1/125, 1/625, and 1/3125. The qPCR was carried out on the Bio-Rad CFX Connect Real-Time System, and
relative quantification was obtained with the software Bio-Rad CFX Manager 3.1. The number of copies in
the samples was estimated from the linear regression of the Ct against the natural logarithm of starting
quantities of the known dilutions for the spike-in.

Estimation of the input spike-in to target a given proportion of sequencing reads

The amount of spike-in to add to the soil samples to target a given ratio of the sequencing reads can be
approximated from the qPCR results and the relations below (Eqs. 1–6). Locus-specific copies in the DNA
extract (C) is the sum of spike-in (C_{s1}) plus the wild copies (C_w) (Eq. 1); and C is proportional to wild
copies (C_w) divided by their proportion in the pool (1-R), where R is the ratio of spike-in copies with respect
to the total (Eq. 2):

(Eq. 1) \[ C = C_{s1} + C_w \]

(Eq. 2) \[ C = \frac{C_w}{1 - R} \]

Eq. 1 and Eq. 2 can be equalized:

(Eq. 3) \[ C_{s1} + C_w = \frac{C_w}{1 - T}; \ C_{s1} = \frac{C_w}{1 - R} - C_w \]

The yield from a DNA extraction of a known amount of pure spike-in (no soil) can be estimated via qPCR
(output/input). Assuming constant DNA yields across a batch of DNA extractions, this ratio can be used
to approximate the number of locus copies in soil samples from the same batch. So, the ratio between
the spike-in added to a soil sample (V_0) and its estimated copies via qPCR (C_{s0}), is approximately equal
to the ratio between the spike-in needed to add to any other sample (V_1) to get the desired number of
copies (C_{s1}). V_1 can be adjusted to the desired ratio of spike-in/wild copies the DNA extract:

(Eq. 4)
\[
\frac{V_0}{C_{s0}} \approx \frac{V_1}{C_{s1}}; V_1 \approx \frac{V_0}{C_{s0}} C_{S1}
\]

Assuming the above ratio associated with the qPCR remains approximately constant in library preparation plus sequencing, the term \( C_{s1} \) from Eq. 4 can be substituted with Eq. 3:

(Eq. 5)

\[
V_1 \approx \frac{V_0}{C_{s0}} \left( \frac{C_w}{1 - R} - C_w \right)
\]

The dilution factors of the DNA templates from the spike-in \( (D_0) \) and sample \( (D_1) \) at the time of qPCR quantification often differ. These factors derive mainly from elution volumes during DNA extractions and the preparation of the qPCR reaction. These are incorporated into Eq. 6. In addition, \( V_1 \) can be normalized to the amount of soil sample used for the DNA extraction \( (M) \). So, the final units of \( V_1 \) refer to the mass of soil:

(Eq. 6)

\[
V_1 \approx \frac{V_0}{C_{S0}} \left( \frac{C_w}{1 - R} - C_w \right) \frac{D_0}{D_1 M}
\]

Summary of parameters (schematic representation in Fig. 1B):

- \( V_0 \): volume of spike-in solution added to pure spike-in DNA extraction [\( \mu L \)].
- \( C_{s0} \): spike-in copies from marker \( i \) in pure spike-in DNA extraction [copies].
- \( C \): copies from marker \( i \) in the DNA extract.
- \( C_w \): wild copies from marker \( i \) in DNA extract [copies].
- \( C_{s1} \): spike-in copies from marker \( i \) in DNA extract [copies].
- \( R \): ratio of spike-in copies with respect to the total.
- \( V_1 \): volume of spike-in necessary to get desired \( R \) [\( \mu L \)].
- \( D_0 \): dilution factor of spike-in DNA extract in qPCR.
- \( D_1 \): dilution factor of biological sample DNA extract in qPCR.
- \( M \): soil mass used in DNA extraction [g]
A ready-to-use template spreadsheet to estimate initial amounts of spike-in from qPCR results (Eq. 6) and a worked example are provided in S2. This template was used to estimate the volume of spike-in to add to the soil samples A-L (Table 1) before DNA extractions.

DNA extraction of samples with spike-in added, library preparation, and sequencing

After these calculations, a new spike-in solution was prepared containing a 2:5 vol/vol of bacterial ZymoResearch spike-in and a $10^7$ cells/mL suspension of the yeast *Y. lipolytica*. From this solution, 5–7 µL were added to each soil sample in the homogenization tube (first step in the DNA extraction), to a total of twelve soil samples from the same plot in a strawberry plantation from two consecutive years (Table 1). DNA of extraction was carried out as in section 2.1. DNA from two of these samples was extracted in three technical replicas. Two blanks of extraction and a sample containing only the spike-in solution were included. A total of nineteen bacterial 16S and fungal ITS2 amplicon libraries were prepared and sequenced in Illumina NovaSeq PE250, as in Rodriguez-Mena et al. (2022), targeting 100K reads for 16S libraries and 50K reads for fungal libraries.

Determination of sequence variants and taxonomic profiling

Locus-specific primers and Illumina adapters were trimmed with *cutadapt* 2.1 (Martin 2011). Sequencing reads were further filtered with *dada2* (Callahan et al. 2016) in R 4.1.3 (R Core Team 2022), with minimum filtering parameters: `filterAndTrim(maxN = 0, maxEE = 3, truncQ = 2, minLen = 50)`. Bacterial 16S reads were truncated at 220 nt. However, no truncation length was applied to fungal ITS2 sequences, since ITS is very variable in size and biologically relevant variants below the read length had been detected in preliminary analysis. A custom R script was used to compute error rates from binned quality scores of the NovaSeq system (github.com/ErnakovichLab/dada2_ernakovichlab). ASV determination was carried out on forward and reverse reads with *dada* function, followed by merging paired-end reads and ending with the removal of chimeric sequences (details in github.com/csmiguel/spike-in). ASVs were iteratively clustered into KTUs (K-mer Taxonomic Units), reducing the sparsity of the taxon-abundance matrix and the number of variants into optimal biological relevant units (Liu et al. 2022). A newer beta version of the KTU package (KTU2: https://github.com/poyuliu/KTU2) was used to stream the *dada2* workflow in R directly into the clustering algorithm and reduce memory burden. Taxonomic ranks were assigned to each ASV (i.e. KTU) down to “genus” level using *idTaxa* (Murali et al. 2018) from DECIPHER 2.22.0 package (Wright 2016). For bacterial reads, the SILVA 138 prokaryotic small subunit rRNA taxonomic training set (Quast et al. 2012) was used, whereas for fungal reads I used the UNITE v2020 database (Kõljalg et al. 2020), which contains the eukaryotic nuclear ribosomal ITS region. KTU taxon-abundance matrix (OTU table hereon), sample metadata, and DNA sequences were stored in *phyloseq* 1.38.0 objects (McMurdie and Holmes 2013) in R. KTUs with no phylum assigned or classified as eukaryotic organelle were discarded. Potential contaminant KTUs in blanks were removed with *decontam* 1.10.0 (Davis et al. 2018). Low abundant KTUs (relative abundances < 5e-5) and/or present in only one sample were also removed.

Estimation of absolute abundances
At this point in the bioinformatics workflow, the data is compositional. A central goal is to transform data from compositional to quantitative. The amount of spike-in added to each sample and their fraction of reads in the filtered ASVs can be used to normalize counts. A custom R script was used to detect the ASVs of the spiked organisms. Assuming the ratio of spike-in in sequencing reads is similar to its ratio in the sample (Eq. 7), the number of wild (non spike-in) biological units \(C_w\) can be approximated, and normalized to the initial amount of biological sample (Eq. 8). If \(N\) internal standards are used for the same marker, the average \(C_w\) \((C_{wav})\) can be approximated with Eq. 8bis.

(Eq. 7)

\[
\frac{C_w}{C_{s1}} = \frac{R_w}{R_s}
\]

(Eq. 8)

\[
C_w = \frac{R_w C_{s1}}{M R_s}
\]

(Eq. 8 bis)

\[
C_{wav} = \frac{R_w}{M} \sum_{i=1}^{N} \frac{C_s_i}{R_{s_i} N}
\]

\(R_s\), reads in sample classified as spike-in \(i\).

\(C_s\), spike-in units in inoculated sample.

\(R_w\), reads in sample classified as wild reads (non spike-in).

\(C_w\), wild units in sample.

Disregarding CNV, quantitative biological units in the sample \((OTU_{abs})\) can be calculated as the total estimated counts \((C_{wav})\) multiplied by a numeric vector with their proportions \((OTU_{rel})\) (Eq. 9).

(Eq. 9)

\[
\overrightarrow{OTU_{abs}} = \overrightarrow{OTU_{rel}} \cdot C_{wav}
\]

Quantitative data for bacteria referred to 16S copies per gram of soil, whereas for fungi, units referred to the unknown ITS2 copy number present in \(Y.\ lipolytica\) per gram of soil.

Evaluation of internal standards
The correlation between absolute frequencies estimated from both bacteria spike-ins (*I. halotolerans* and *A. halotolerans*) was evaluated with Pearson’s *r*.

To further evaluate the predictability of the proposed workflow, observed proportions of spike-in in the sequencing reads (*R* observed) were compared to the expected ones (*R* expected) from the suspension added to the sample (*V*). Observed proportions were calculated dividing spike-in reads by total filtered reads. Expected *R*’s were calculated from qPCR results after clearing the term *R* from *Eq. 6*. They were calculated for each sample independently since its value depends on the exact amount of the spike-in suspension added to the sample and approximate marker wild copies (*C*<sub>w</sub>) guessed from the qPCR. For samples that had not been quantified by qPCR, expected target ratios were computed by extrapolating average *C*<sub>w</sub> from the measured ones. In the case of bacteria, *R* was split among the two species in the spike-in based on the CNV of the 16S rRNA: 3 copies in *I. halotolerans* and 7 copies in *A. halotolerans*. Comparisons were expressed symmetrically around 0 by taking base 2 logarithm of the fold change of observed proportions divided by expected proportions. Log2 ratios of fold change from samples measured by qPCR were compared to the extrapolated ones with *t*-student tests.

Alpha and beta diversity. Procrustes superimposition

Alpha diversity was assessed with *phyloseq::estimate_richness*. Absolute abundances of bacterial phyla and fungus classes were visualized with stacked bar plots.

Beta diversity was evaluated with Multidimensional Scaling (MDS) on Bray-Curtis (BC) distances (*phyloseq::ordinate*) on the OTU-tables with compositional and absolute data of bacteria and fungi. Composite data was transformed by multiplying feature proportions (sum 1) by the geometric mean of absolute counts per sample. The logarithm was applied to both, transformed compositional and quantitative data. Then, symmetric Procrustes analyses with *vegan::procrustes* (Oksanen et al. 2020) were run on the first two dimensions of the MDS to compare the effect of using compositional versus quantitative abundance in beta diversity. The non-randomness (significance) of the MDS configuration between relative and absolute beta diversities was tested with the *vegan::protest* function with 999 permutations, and the *m*<sub>12</sub> correlation was reported (Peres-Neto and Jackson 2001).

**Results**

Estimation of the input spike-in to target a given proportion of sequencing reads

The coefficient of variation for qPCR technical replicas was low: 0.005–0.081 for 16S and 0.006–0.104 for ITS2. From the qPCR estimation of copy number for 16S and ITS2, and according to *Eq. 6*, for bacteria, 2.3 µL (*s.d.*, 0.72; range, 1.1–3.8) of spike-in solution would be needed to target 2% of spike-in reads in the sequencing reads (S3). For fungi, 4.7 µL (*s.d.*, 1.9; range, 1.3–8.2) of a 10<sup>7</sup> cells/mL solution of *Y. lipolytica* would be needed to reach a 1% of spike in reads (S3). The estimations from the qPCR derived from four random samples whose DNA had been extracted in triplicate. The associated
coefficient of variation between extraction replicas ranged from 0.03–0.32 for 16S and 0.09–0.36 for ITS2.

Sequencing results and alpha diversity

A total of 1.87 and 1.09 million raw sequences from sixteen 16S and ITS libraries, respectively, from soil samples were processed, averaging 117,000 (s.d. ±11,400) for 16s and 68,000 (s.d. ±8,200) for ITS2. Filtered sequences per sample averaged 79,600 (s.d. ±8,600) for 16S and 38,000 (± 6,800) for ITS2 (Fig. 2; S4). Corresponding NCBI Sequence Read Archives and BioSample data are listed in S5.

A total of 1,766 bacterial and 120 fungal KTUs were determined, averaging 1,125 (s.d. ±73) per sample for bacteria and 72 (s.d. ±10) for fungi (S4).

Estimation of absolute abundances and reliability of the method

Absolute counts estimated independently from both bacteria spike-ins (I. halotolerans and A. halotolerans) were highly correlated, r(14) = .99; P < .001 (Fig. 3). Absolute estimations yielded up to 5x and 10x fold changes between samples for bacteria and fungi, respectively (Fig. 4). However, between technical triplicates (samples A and D), the coefficient of variation of absolute counts was low for bacteria, 0.10, and 0.16–0.26 for fungi. The number of wild 16S units estimated for 1 g of soil ranged from 0.53 to 2.6 x10^9 (mean ± s.d., 1.44 ± .64 x10^9), while the number of ITS2 units ranged from 0.37 to 3.80 x10^6 (mean ± s.d., 1.53 ± 1.20 x10^6) (Fig. 4).

Observed proportions of spike-in reads were very similar to those expected in bacteria, as reflected by values of log2 fold changes close to 0, whereas for the fungus Y. lipolytica, observed reads were on average 39 times more (log2 fold change around 5) than expected (Fig. 5). Specifically, average and s.d. log2 fold change for the three spike-ins ranged from 0.32 ± 0.44 in A. halotolerans, 0.88 ± 0.51 in I. halotolerans and 5.28 ± 0.92 for Y. lipolytica (Fig. 5). Log2 extrapolated ratios to samples not measured in the qPCR were in the range of those measured by qPCR (A. halotolerans, t(14) = 0.24, p = .81; I. halotolerans, t(14) = -.11, p = .92; Y. lipolytica, t(13) = -1.57, p = 0.14).

Effect of absolute frequencies on beta diversity

Community differentiation increased in most cases when considering quantitative data (Fig. 6). This pattern was evidenced by larger dissimilarities for bacteria (mean ± s.d.: .004 ± .005) and fungi (mean ± s.d.: .012 ± .02) and changes in the Procrustes superimposition. For bacteria, both datasets was highly correlated but significantly different (m_{12} = .0014, p = .001), whereas for fungi the matrices were different and little correlated (m_{12} = .32, p = .001).

Discussion

Here, it is described a simple and reliable strategy for quantitative metabarcoding profiling of bacteria and fungi from soil samples using cellular spike-ins, which are added before DNA extraction. Adjustment
of input spike-ins based on qPCRs yielded sequencing coverages near those predicted for bacteria but not for fungi. The reliability of this method was further supported by the correlation of the results from multiple bacterial spike-ins.

qPCRs served to estimate the input spike-in to get a small fraction of the reads from the sequencing run (1–2%), in which the observed coverages were close to those predicted. For the two bacterial spikes, average deviations were below a 2-fold change. However, coverage observed for the fungal spike-in in the sequencing reads reached up to almost 40 times more than the expected one based on qPCR estimations with the yeast Y. lipolytica (Fig. 5). In any case, normalization of compositional data was not altered by deviations in predicted versus observed spike-in coverage. These deviations only affect the total reads of the spike-in in the runs. In both, bacteria and fungi libraries, spike-in reads (Fig. 2) were sufficient for normalization of the compositional data by consuming only a fraction of the sequencing run while leaving a large fraction of the wild reads allocated for biodiversity analysis (Fig. 4). Thus, even if mistaking input amounts of spike-ins by an order of magnitude, sequencing results can be useful and robust to normalization within an ample range of spike-in coverage. These results depict a practical path for implementing absolute quantification in metabarcoding studies. Up to date, the spread of quantitative metabarcoding has probably been hindered by the ignorance of treating compositional data as quantitative (Chen et al. 2016; Gloor et al. 2017), or by methodological constraints related to the questions of what spike-in to use, at which step in the protocol, in which input amounts, or how to normalize the data (Harrison et al. 2021). In the absence of a gold-standard, most of the literature reporting the use of internal standards in the form of cells, genomic DNA and synthetic fragments, often developed their own protocols (Harrison et al. 2021). The question related to “how much input to add” has been approximated by counting cells on the microscope (Yang et al. 2018), sequencing libraries with a dilution series of the spike-ins (Smets et al. 2016; Stämmler et al. 2016; Lin et al. 2019), or counting cells using flow cytometry (Props et al. 2017). When using only cell count data, variation can arise from microbial groups having different DNA yields depending on the extraction method (Haro et al. 2021). Also, genome copies of the rRNA operon vary between species, from a handful in bacteria (Stoddard et al. 2015) to hundredths in fungal genomes (Lofgren et al. 2019). Consequently, empirical qPCR estimation of locus copy number from spiked cells provides superior accuracy since much of the methodological and biological variation (extraction yields, amplicon specific PCR efficiency, CNV) is already included in this measurement. Concordantly, low coefficients of variation between the absolute counts estimated for technical replicas were found (same soil sample extracted in different DNA extractions): 0.10 for bacteria and from 0.16 to 0.26 for fungi. The homogenization step of the biological sample through sieving to remove small stones and organic particles, and the subsequent lyophilization were probably responsible for much of these low coefficients of variation. Reproducibility was further supported by the high correlation of the absolute counts estimated independently with two bacterial spikes (Fig. 3) (r = .99, p < .001), which has already been discussed to reduce estimation errors (Stämmler et al. 2016).

Transformation of composite data to quantitative in the literature is often based on a straightforward transformation by a division of taxon i reads by spike-in reads (Chen et al. 2016; Lin et al. 2019). So, counts after normalization are directly proportional, but not absolute frequencies of the taxonomic
profiles in the biological sample (Stämmler et al. 2016). This approach is a closer attempt at an absolute quantification of soil microbiota, as it incorporates CNV data from the bacterial spike-in and soil mass. However, these numbers are still far from real taxonomic profiles since they do not take into account CNV of the microbiota, taxon-specific PCR efficiencies (Shelton et al. 2022), or other PCR biases (Jones et al. 2015). Deviations in spike-in coverage in the ITS2 data could be associated with PCR biases in library preparation or sequencing (Jones et al. 2015; Bonk et al. 2018), or to the quantification of non-target molecules during qPCRs. Especially, ITS2 region can be subjected to these effects due to its variation in size and %GC composition (Quast et al. 2012). This result points to the use of multiple spike-ins in quantitative metabarcoding (Stämmler et al. 2016).

Sequencing multiple samples with a wide range of spike-in concentrations provides accurate estimations of input standards, but typically involves large prices for library preparation and sequencing (~$50/library), plus around 2-months of turnaround time and the burden of bioinformatics analysis to detect the spike-in reads within the read pool. The alternative here proposed, co-amplification in a qPCR of a subset of biological samples plus the pure spike-in itself, is a cheap approach to approximate starting inputs of internal standards, which can be routinely done in a molecular laboratory in one day at the price of ~$2/reaction, data being easily interpreted.

Quantitative abundances improve biological interpretations of microbial diversity (Gloor et al. 2017; Hardwick et al. 2018; Kong et al. 2021) (Fig. 4). A practical workflow is proposed for normalizing compositional data based on qPCR measurements of spikes and sample DNA extracts that can be applied in any molecular lab. The study was focused on soil samples, but the overall strategies and mathematical framework can be applied to different types of biological samples.

**Declarations**

**Acknowledgements**

Funding came from the IFAPA grant PP.AVA.AVA2019.034, financed by the Junta de Andalucía with 80% FEDER funds from the European Union. The Spanish Collection of Type Cultures (CECT; www.uv.es), Valencia, supplied *Yarrowia lipolytica* strain CECT 1240. MC-S has a postdoctoral contract from the Plan Andaluz de Investigación, Desarrollo e Innovación (PAIDI 2020). Part of the analysis was done in the High-Performance Computing cluster hosted by the Centro Informático Científico de Andalucía, CICA (https://www.cica.es/). Dr. Nieves Capote facilitated access to part of the laboratory equipment. María Camacho provided comments to the manuscript. Berta de los Santos and María Camacho and Luis Miranda led the field work.

**Data availability**

Raw FASTQ files and assembled sequences have been deposited in the NCBI under the BioProject PRJNA883611, with the accessions KHUY00000000 for fungal ITS2 data and KHUZ00000000 for
bacterial 16S rRNA. Code and *phyloseq* objects with data were deposited in https://github.com/csmiguel/spike-in (zenodo. xxx).

**References**


https://doi.org/10.3390/microorganisms8121910

Bacteria in Surface Waters of the Subtropical Northwest Pacific Ocean Estimated With High-
Throughput Quantification Sequencing. Front Microbiol 11:.
https://doi.org/10.3389/fmicb.2020.599614


15. Liu P, Yang S, Yang S (2022) KTU: K-mer Taxonomic Units improve the biological relevance of
https://doi.org/10.1111/2041-210X.13758

https://doi.org/10.1111/mec.14995

EMBnet.journal 17:10–12. https://doi.org/http://dx.doi.org/10.14806/ej.17.1.200


0521-5

version 2.5-6. 2019

https://doi.org/10.1007/S004420100720/METRICS

abundances. ISME J 11:584–587. https://doi.org/10.1038/ismej.2016.117

https://doi.org/10.1093/nar/gks1219

27. R Core Team (2022) R: A Language and Environment for Statistical Computing
38. Wright ES (2016) Using DECIPHER v2. 0 to analyze big biological sequence data in R. R J 8:

Figures
Figure 1

(A) set-up to couple sample sieving with weighing in precision scale. (B) Schematic representation of the sample and spike-in DNA extraction workflow and associated parameters for estimating the amount of spike-in solution needed for each sample to target a given coverage. The DNA extractions 1-3 represent samples with no spike-in (1), spike-in solution with no soil (2) and soil sample with spike-in added (3).
Figure 2

Filtered reads assigned to wild (non-spike-in taxa) and spike-in taxa in all samples (libraries in parenthesis) for bacteria (green scale, left side) and fungi (orange scale, right side).
Correlation between the absolute wild 16S copies in soil samples estimated with the two bacterial species in the spike-in. The solid line is an adjusted regression line forced at origin, and the dashed line is a 0-intercept with slope = 1 line.

**Figure 3**
Figure 4

Alpha diversity and absolute counts of selected taxonomic ranks in one gram of soil sample (library in parenthesis) for bacteria (A) and fungi (B).
Figure 5

Observed (sequencing run) versus expected (amount added to DNA extractions) proportions of the different spike-ins across samples. Comparisons are expressed in the base 2 logarithm of the fold change.
Figure 6

Procrustes errors from superimposition of matrices derived from MDS Dimensions 1 and 2 from Bray-Curtis (BC) distances for bacteria (A) and fungi (B). Embedded histograms (red, negative values; green, positive values) represent the distribution of BC dissimilarities from quantitative OTU-matrices minus those from relative frequencies.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S1Ylipolytica.docx
- S2spikeintemplate.xlsx
- S3calculationsspikein.xlsx
- S4readsalphadiv.xlsx
- S5ncbicodes.xlsx