Supplementary data

Lauren Alteio1, Joana Séneca1, Alberto Canarini1, Roey Angel2, Ksenia Guseva1, Jan Jansa3, Christina Kaiser1, Andreas Richter1, Hannes Schmidt1

1Centre for Microbiology and Environmental Systems Science, University of Vienna, Vienna, Austria

2 Soil and Water Research Infrastructure and Institute of Soil Biology, Biology Centre CAS, České Budějovice, Czechia

3 Institute of Microbiology, Czech Academy of Sciences, Praha, Czechia

# Description of Web of Science search

We investigated the number of articles in soil science journals that have applied amplicon sequencing in soil microbiome research using a keyword search in Web of Science. The complete list of soil science journals (as defined by Web of Science)  is shown in Supplementary Table 1. The article search was carried out by selecting only original research articles which contained at least one of the following words: “amplicon”, “sequencing”, “454”, “pyrosequencing”, “Ion torrent” or “sanger”. The number of published articles in the soil science journals that have used amplicon sequencing between 1990 and 2020 is shown in Figure 1, including also their percentage relative to the total of published manuscripts. The number of published articles and their relative percentage for the top ten journals in 2020 is also shown in Figure 1 within the pie chart.

## Table S1. List of journals included in the Web of Science search for articles using amplicon sequencing

**Hosted file**

Table S1.xlsx available at <https://authorea.com/users/243794/articles/496867-supplementary-data>

## 

## Table S2. Potentials and Limitations of Marker Gene Approaches

**Hosted file**

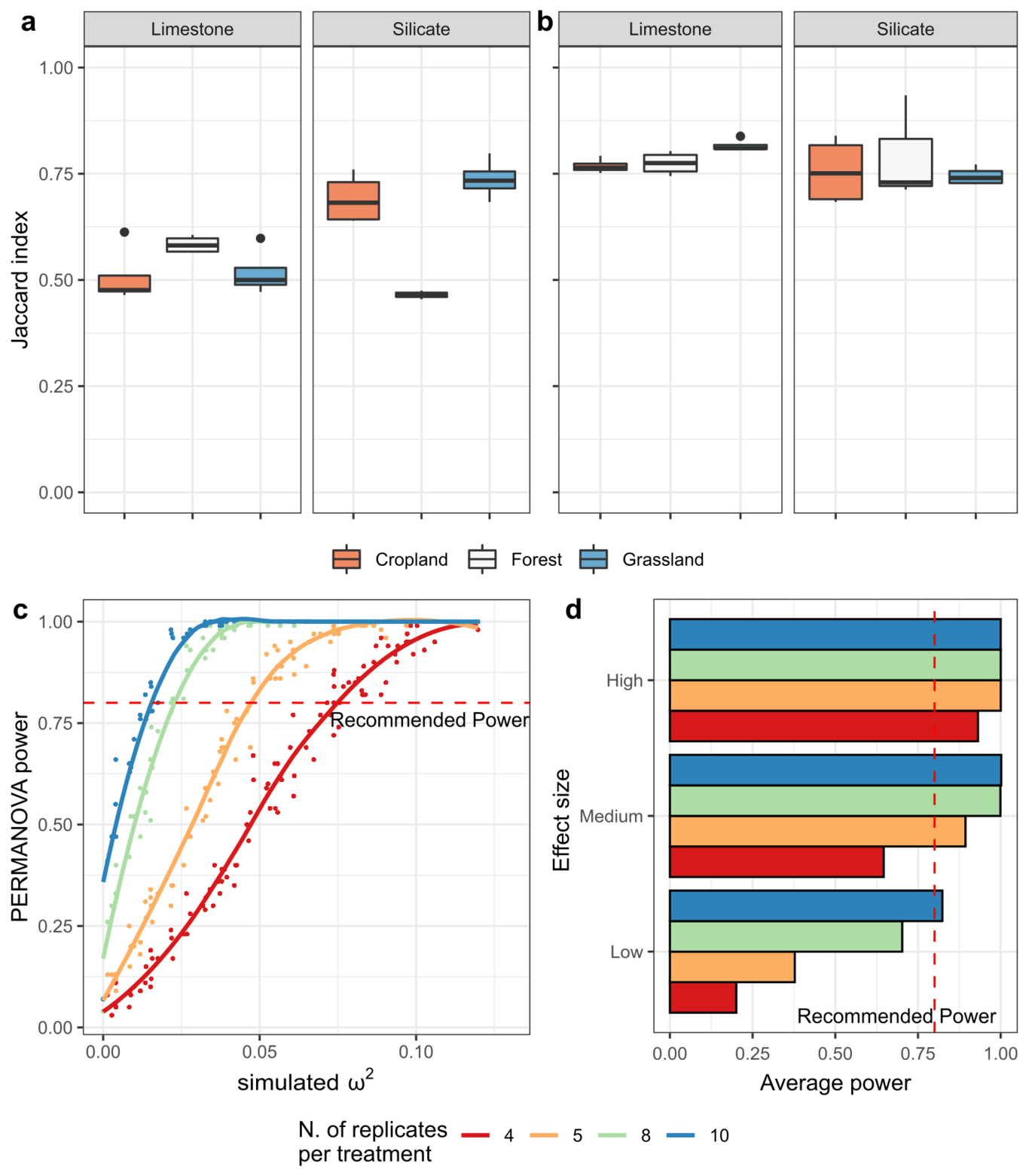
Table2\_AmpliconSeqPotentialandLimitations1.xlsx available at <https://authorea.com/users/243794/articles/496867-supplementary-data>

## Power and sample-size estimation for soil microbiome data

In this analysis we aimed at reproducing results obtained and reported previously (Kelly et al., 2015) from the Human Microbiome Project (HMP) dataset to a soil microbial data set. To achieve this aim we used available supplementary dataset accompanying a previous publication (Zheng et al., 2019) available on NCBI under the BioProject accession number PRJNA551019. This dataset contains 16S and ITS1 amplicon sequencing data from a multifactorial experiment that aimed to assess the effect of land use and bedrock type on soil microbial communities. The dataset included a total of 24 samples per functional gene, representative of three land use types (grassland, cropland and forest) and two bedrock types (silicate and limestone), and therefore representative of different soil types and environmental conditions. Three biological replicates per condition were sequenced. The paired-end MiSeq reads were processed as described previously  (Zheng et al., 2019) and OTU tables were  generated at 97% and 99% sequence identity for the 16S and ITS datasets, respectively. The R package ‘phyloseq’  (McMurdie and Holmes, 2013) was used for all the initial filtering steps. The OTU table for 16S rRNA gene sequences was taxonomically filtered in order to remove Eukaryotic (e.g., chloroplastidial and mitochondrial) reads and reads unclassified at the domain level using the function *subset\_taxa()*. Similarly, the OTU table for ITS1 region was filtered in order to remove fungal OTUs unclassified at the Phylum level. Finally, both OTU tables were filtered in order to only keep OTUs with an overall abundance above 0.1% across all samples. The sequencing of one biological replicate of the ‘silicate x forest’ condition failed to produce sufficient number of reads and was therefore excluded from the analysis. All metadata, code, original and filtered OTU tables are publicly  available under All metadata, code, original and filtered OTU tables are publicly  available under <https://github.com/joanaseneca/use_misuse.git>. The data set was used only to obtain representative Jaccard indices (Fig. S1a)  for soil microbiome, in order to generate distance matrices with similar parameters. The data set was used only to obtain representative Jaccard indices (Fig. S1a)  for soil microbiome, in order to generate distance matrices with similar parameters.

We used the R package “micropower” (Kelly et al., 2015) which allows to simulate distance matrices from a set of parameters to generate available PERMANOVA power or necessary sample size for a planned microbiome analysis. The Jaccard similarity index was calculated to simulate OTU/ASV tables (with the function *calcWJstudy*) for both ITS and 16S rRNA gene sequences, obtaining an average and standard deviation across all samples (Fig. S1a).  We then computed the dependency of statistical power of permutational multivariate analysis of variance (PERMANOVA) on the effect size for four datasets with varying replicate numbers (4, 5, 8 and 10 replicates; Fig. 5). The analysis followed the description in Kelly et al. (2015) and we refer the reader to this article for further details. Briefly,  we first simulated a set of distance matrices separately for the 16S and the ITS dataset, with the within-group pairwise–distance distributions set to be the same across the entire set of distance matrices, whereas each simulated distance matrix differed in its simulated effect size.  We then selected bootstrap samples of subjects from each of the simulated distance matrices and performed PERMANOVA testing, comparing the PERMANOVA *P*-value with the pre-specified threshold for type I error (0.05). PERMANOVA power is calculated as the proportion of bootstrap distance matrices for which PERMANOVA *P*-values are less than the 0.05 threshold for type I error (Kelly et al., 2015). We performed the bootstrap procedure with 4, 5, 8 and 10 subjects per group.

From the data obtained above, we further calculated the average statistical power for a range of effect sizes  (  *ω*2 ) defined as ‘Low’ (0.001-0.04), ‘Medium’ (0.04-0.08) and ‘High’ (0.08-0.12). The definition of low, medium and high were chosen as a rule of thumb and should not be interpreted as a defined category.



Top panels: the Jaccard index calculated from the data set used (Zheng et al., 2019) for 16S (a) and ITS (b). Bottom panels: c) the calculated PERMANOVA power for a range of  simulated effect (quantified by the adjusted coefficient of determination omega-squared (*ω*2) and divided by number of replicates per treatment); d) the average PERMANOVA power of panel ‘a’, grouped by number of replicates per treatment and into three effect size ranges: Low (0.001-0.04), Medium (0.04-0.08) and high (0.08-0.12).  PERMANOVA power was calculated as the proportion of bootstrap distance matrices for which PERMANOVA *P*-values are less than the pre-specified threshold for type I error (0.05).

# References

Kelly, B.J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J.D., Collman, R.G., Bushman, F.D., Li, H., 2015. Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. Bioinformatics 31, 2461–2468. doi:10.1093/bioinformatics/btv183

McMurdie, P.J., Holmes, S., 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8, e61217. doi:10.1371/journal.pone.0061217

Zheng, Q., Hu, Y., Zhang, S., Noll, L., Böckle, T., Dietrich, M., Herbold, C.W., Eichorst, S.A., Woebken, D., Richter, A., Wanek, W., 2019. Soil multifunctionality is affected by the soil environment and by microbial community composition and diversity.. Soil Biol Biochem 136, 107521.