Title: *A probabilistic model to identify the core microbial community*

Running header*: Identifying the probable core microbial community*

Authors: *[1*]Thiago Gumiere, [2]Kyle M. Meyer, [2]Adam R. Burns, [3]Silvio J. Gumiere, [2]Brendan J. M. Bohannan, [1]Fernando D. Andreote*

[1]Department of Soil Science, "Luiz de Queiroz" College of Agriculture, University of São Paulo. Av. Pádua Dias, 11 – Piracicaba, São Paulo, Brazil 13418-900.

[2]Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon, United States of America 97403-5789.

[3] Department of Soil and Agricultural Engineering, 2480 Hochelaga Boulevard, Laval University, Quebec City, QC, Canada, G1V 0A6.

****Corresponding author:*** *Thiago Gumiere, Department of Soil Science, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Av Pádua Dias, 11, Piracicaba, SP, Brazil. 13418-900. Phone: +55 19 3417-2123, E-mail:* thiago.gumiere@gmail.com

Conflict of Interest statement: The authors declare no conflict of interest.

Originality-Significance Statement

More rigorous and less arbitrary statistical methods could increase knowledge regarding the role of microorganisms and their interactions. Here, we suggest a probabilistic method to identify the microbial core community across systems. Our method identifies a large proportion of the rare community that likely belongs to the microbial core community, which was not identified by conventional methods. Our probabilistic model is a non-arbitrary approach to defining the microbial core community, which may help in the next step of the microbial core community studies.

ABSTRACT

The core microbial community has been hypothesized to have essential functions ranging from maintaining health in animals to protection against plant disease. However, the identification of the core microbial community is frequently based on arbitrary thresholds, selecting only the most abundant microorganisms. Here, we developed and tested an approach to identify the core community based on a probabilistic model. The Poisson distribution was used to identify OTUs with a probable occurrence in every sample of a given dataset. We identified the core communities of four extensive microbial datasets, and compared the results with conventional, but arbitrary, methods. The datasets were composed of the microbiomes of humans (tongue, gut, and skin), mice (gut), plant (grapevine) tissue, and the maize rhizosphere. Our proposed method revealed core microbial communities with higher richness and diversity than those previously described. This method also includes a greater number of rare taxa in the core, which are often neglected by arbitrary threshold methods. We demonstrated that our proposed method revels a probable core microbial community for each different habitat, which extend our knowledge about shared microbial communities. Our proposed method may help the next steps proving the essential functions of core microbial communities.

INTRODUCTION

The composition of microbial communities can vary greatly even over fine spatial and temporal scales, making it difficult to identify the drivers of community dynamics and the link between composition and function. To overcome the obfuscating effects of this variation, researchers often limit their focus to the 'core' community, which is defined as organisms that are ubiquitous in a given habitat, despite environmental fluctuation (Hamady and Knight, 2009). In microbial ecology, the core community refers to microbial taxa (Shade and Handelsman, 2012), or genes (Turnbaugh *et al.*, 2007), shared across a set of samples in a given ecosystem.

There are considerable attempts to identify the core community across different hosts including corals (Ainsworth *et al.*, 2015), zebrafish (Roeselers *et al.*, 2011), mice (Pédron *et al.*, 2012), ruminants (Henderson *et al.*, 2015), *Arabidopsis thaliana* (Lundberg *et al.*, 2012) and sugarcane plants (Yeoh *et al.*, 2015). It has been suggested that the core microbial community could play essential roles in ecosystem functioning, and may also be useful as indicators of system perturbation (Shade and Handelsman, 2012; Saunders *et al.*, 2015). For example, an abundant microbial core was identified across 210 human adult fecal samples, varying substantially in geographic origin, ethnic background and diet (Sekelja *et al.*, 2011). The authors suggested that this core has an important role in gut homeostasis and health. Other studies have suggested roles for the core in plant growth promotion and the maintenance of plant health (Schlaeppi *et al.*, 2014). However, few studies have been successful in directly linking the core microbial community to important community or ecosystem functions.

The lack of evidences for the importance of the core community may be due to how the core is identified. Since the core is defined to be ubiquitous in a habitat, it is assumed that the microbial taxa or genes belonging to the core should be found in every sample collected from a given habitat. The core microbial community is identified by

identifying shared microorganisms or genes across a collection of samples (*discussed by* Shade and Handelsman, 2012). In this approach, the core is represented by taxa found in every sample analyzed (100% frequency across samples). However, to date no methodological approach has fully assessed the microbial diversity of any environmental sample (Kanagawa, 2003; Feinstein *et al.*, 2009; Prosser, 2015). Current sequencing methods used to survey complex microbial communities tend to target the most abundant groups of microorganisms (Caporaso *et al.*, 2011). Consequently, the rare component of the core microbial community is missed in these studies. The most commonly used approach to circumvent this problem is the definition of cutoffs for the frequency of microbes or genes to be classified as a member of the core microbial. For instance, researchers have used cutoff values ranging from 30% to 99% frequency across samples (Li *et al.*, 2013; Ainsworth *et al.*, 2015) to define the core community in environmental samples. However, these cutoffs still do not include rare taxa and also could result in false assignments to the core, thus influencing inferences about its function and composition.

Given the numerous difficulties associated with sampling and fully sequencing microbial communities, one solution to identify core community members is to use a probabilistic model to assign members of the microbial community to the core community. Here, we develop and test an approach to identifying the core community based on the Poisson distribution. Given the occurrence distribution of an event, *i.e.* a microorganism, in a group of samples, this model estimates the probability of this event in a group of samples (Rao and Rubin, 1964). Among discrete probability models, we selected the Poisson distribution because it is particularly suitable for large count datasets, *e.g.* a high number of events, and the occurrence of small or rare probabilities (Karlis, 2003), situations common when using microbial datasets to estimate a core

community. Unlike other attempts to define the core community (e.g. Turnbaugh *et al.*, 2009) there is no abundance threshold in our proposed method, which allows inclusion of rare taxa as possible members of the core microbial community.

We tested our proposed method using several previously published datasets, and compared our results to those obtained using conventional (i.e. arbitrary threshold) approaches. These datasets included human, mice, plant (grapevine tissue and maize rhizosphere), and soil data, and were obtained from the Earth Microbiome Project (EMP; http://www.earthmicrobiome.org). We hypothesized that our approach would lead to the identification of a probable core community that would be a higher proportion of the microbial community, and would also be composed of more microorganisms with low abundances (rare community members), than the core community identified using conventional approaches.

RESULTS AND DISCUSSION

Testing the distribution models and rarefaction effect

The first step was to select the most appropriate probabilistic method that fitted in OTU distributions. We tested 13 different models (described in Supplementary Material), and in Figure S1, we can observe the fourth best distribution models (Poisson, Chi-squared, Gamma and Beta) fitted on each dataset (Human, Grape, Maize and Mice). The Poisson distribution showed the higher and significant fit on OTU distribution, which is indicated by R^2 and p-value < 0.05 in Table S1. We also observed that the Poisson distribution indicated lesser value of RMSE. Models based on 'Poissonization' arguments has also been indicated as good predictor of microbial unknown (Lladser *et al.*, 2011).

The use of rarefaction, normalization method which equalizes the number of sequences (or reads) per sample, is discussed in the literature. According to McMurdie

and Holmes (2014), the rarefaction increases the number of false positives species, and also with different abundance across sample classes. However, other simulation studies indicated that the rarefication is better than other normalization methods, clustering samples as biological origin (Weiss *et al.*, 2017). As probability models requires the normalization, we evaluated the effect of the rarefaction on our proposed.

It can be observed in Figures S2, S3, S4 and S5 that the rarefaction method affects the line of Poisson distribution identification. We also observed that the values of R^2 decreases with the increase of rarefication levels. However, the number of OTU's identified as probable members of the core microbial community did not present a significant variation in general (Table S2). In grape dataset only the two highest rarefication levels, and in maize and human dataset only the lesser rarefication level showed a significant different number of core OTU's identified. As indicated in Figures S6, S7, S8 and S9, the taxonomic composition at the phyla level was not significant affect by the most of rarefication levels. We verified the similarity of core community composition by different rarefication levels using NMDS analyses (Jaccard similarity). In Figure S10, we can observe that only the lowest level of rarefaction for the grape (Core_500), maize (Core_100), and human (core_100) datasets showed a significant difference from the other rarefaction levels. For the mice dataset, we observe the lower variation than the other datasets, but with the same pattern (lowest rarefaction level is not grouped). Considering this normalization effect, we decided to maintain the same method (rarefaction level) used by the authors of each published datasets for the next steps.

A probabilistic method to identify the core microbial community

Using this probabilistic model, we identified core microbial communities for each dataset selected for analysis with R^2 varying between 0.46 (mice) and 0.91 (grape), and

with *p-values* as lower than 0.05. The obtained curves indicated the occurrence of OTUs with distinct values of frequency occurrence as components of the core microbial communities, which is not observed when other approaches are used (Figure 1 and Supplementary Figures S11, S12, and S13). As the results were based on a probabilistic method, we expected that our proposed method would identify a group closer to the real core community than the group identified by conventional methods.

We observed that our probabilistic method reveals a rich and diverse group of microorganism which has not been identified by conventional methods, but belong to the probable core microbial community. For example, the core microbial community identified in the mice database is composed of 170 OTUs using an arbitrary threshold of 30% detection frequency, and 1,717 OTUs using the method based on the Poisson distribution (Table 2). In particular, these differences were found for the occurrence of OTUs with low abundance, much more pronounced in the core community obtained by the method based on the Poisson distribution (e.g. Figure 1).

In the literature, the microorganisms with low abundance are frequently referred to as the "rare biosphere" (Sogin *et al.*, 2006). The rare biosphere was first described as microorganisms with low growth rates, which could act as a "seed bank" of species or genes important in maintaining the functional redundancy of a system (Pedrós-Alió, 2006). These taxa could become dominant (in high abundance) under certain conditions (Shade *et al.*, 2014). Following this view, members of the rare community can be classified as conditionally rare taxa (CRT), suggested to be ubiquitous in some systems (Shade and Gilbert, 2015). As members of a core microbial community, the CRT could be important to the stability and functional resilience of a system. Using our methodology, these groups could be properly classified within the core community, while the arbitrarily defined core rarely included these putative CRTs, likely due to their

lower frequency (e.g. Figure 1B). The cut-offs for the core may fail to identify members of the core microbial community, *i.e.* this method may produce "false negatives". By failing to include members in the core (*e.g.* low abundance taxa that are ubiquitous), researchers may be underestimating the contribution of the core to ecosystem function. Data from the mice dataset (Turnbaugh *et al.*, (2009) did not identify a core microbial community across 100% of samples, or also using the PSM with abundance threshold. The probabilistic method identified the same three phyla as the arbitrary cutoff method (*Actinobacteria*, *Bacterioidetes*, and *Firmicutes*), but also recovered an additional eight phyla (*Cyanobacteria*, *Fusobacteria*, *Lentisphaerae*, *Proteobacteria*, *Synergistetes*, *Tenericutes*, *TM7*, and *Verrucomicrobia*) as members of the core microbial community (Figure 2). The authors also indicated the distinct proportions of the *Bacteroidetes* and *Actinobacteria* phyla associated to obese and lean mice. Both phyla were also detected by our probabilistic method, with OTUs affiliated with these groups as components of the core microbial community.

Rather than defining a specific, core cutoffs, some researchers have used the term 'persistent' – referring to taxa with a high (but below 100%) occurrence frequency, or 'transient' referring to taxa with low occurrence frequency. For example, Caporaso *et al.*, (2011) have identified a persistent and transient communities, which are classified as OTUs occurring in 60% or 20% of samples, respectively. Using this dataset (Caporaso *et al.* 2011), we identified a probable core community, also based on OTUs, across all of the human site samples made of 8,751 OTUs (Supplementary Figure S10). The authors identified classes belonging to the phyla *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, and *Tenericutes* in the human gut. Similar results were obtained by our approach, with the major affiliation of the OTUs to the phyla *Firmicutes*, *Proteobacteria, and Bacteroidetes* (Supplementary Figure S14). We believe that our

206 approach better succeeds to identify the core community for two reasons. First, our
207 method identified core communities across assessments previously identified as not
208 having a core community (as determined by 100% frequency occurrence). Second, our
209 method offers a complement to other terms as “persistent’ and “transient” communities,
210 e.g. indicating the rare microorganisms that could be classified in persistent group.

211 Same results were observed applying our proposed method to grapevine (leaves,
212 flowers, grapes, and roots), and the maize rhizosphere. For example, Zarraonaindia *et*
213 *al.*, (2015) suggested a bacterial core community identified by three OTUs across 75%
214 of samples from grape (leaves, flowers, grapes, and roots) and soils, over two growing
215 seasons. These OTUs belonged to the genera *Bradyrhizobium*, *Steroidobacter* and
216 *Acidobacteria*. By using our proposed method on the same dataset, 5,039 OTUs were
217 identified as belonging to the core community (Supplementary Figure S12A and S12B).
218 In addition, members of the *Cyanobacteria* phylum - which was a dominant group
219 identified by the arbitrary methods (90% of relative abundance; Supplementary Figure
220 S15) – comprised only a small component of the core microbial community using the
221 probabilistic method. This variation in dominance could directly affect the conclusions
222 about microbial composition across the system and may also affect the correlations with
223 environmental drivers.

224 Here, we demonstrate the use of a probabilistic model to identify the core microbial
225 communities. By applying a probabilistic model, our results suggest that the core
226 microbial community may be higher in richness and diversity than previously
227 demonstrated using other methods. Our method also allowed us to include rare (low
228 abundance) members in the core microbial community, which would otherwise be a
229 challenge using an arbitrary core cutoff. The use of a probabilistic model can extend our
230 detection of the core microbial community, and could potentially help researchers to

231 better connect the core community to ecosystem functions. An increased understanding
232 of core microbial functions could support more robust studies in several fields, from
233 human health (Zaura *et al.*, 2009) to increased crop production. The microbial core
234 community could also be used as an indicator of system perturbations (Shade and
235 Handelsman, 2012) such as disease occurrence. This new approach could provide future
236 studies a more realistic strategy to define calculate the core community, and could help
237 to investigate the role of core microbial community in ecosystem function, or to
238 elucidate the drivers of its composition. The probabilistic model is a new tool to step
239 forward in the microbial community investigation. Only with the use of more rigorous
240 and less arbitrary statistical methods it will be possible to understand the microbial
241 ecology and its interactions.

242 **EXPERIMENTAL PROCEDURES**

243 We selected four datasets composed of microbiomes from human samples
244 (tongue, gut, and palms), mice (gut), grapevines (plant organs and bulk soil), and the
245 maize rhizosphere to study the core microbial community identified using arbitrary
246 cutoffs and a probabilistic method based on the Poisson distribution (Table 1).

247 The mice dataset was used to evaluate how the gut microbiome influences host
248 adiposity (Turnbaugh *et al.*, 2009). The data are from fecal samples from 154
249 individuals (mice) divided into adult females, monozygotic or dizygotic twin pairs, and
250 their mothers. The core microbial community was identified using the *Phylotype*
251 *Sampling Model* (PSM), which by Poisson distribution estimates the failures to observe
252 microbial groups possibly belonging to the core community. The authors established a
253 threshold value for abundance, considering only the OTUs with more than 0,5% of
254 relative abundance as members of the core microbial community.

255 The human microbiome database consists of 396 samples, collected along a time

256 series of two individuals at four body sites, including gut, tongue, and left and right
257 palm (Caporaso *et al.*, 2011). In the original study, the authors aimed to evaluate the
258 temporal variation in the human microbiome. The authors used the terms persistent
259 (microbial taxa with high levels of occurrence across samples), and transient (taxa with
260 low levels of occurrence across samples) community, because it identified a very small
261 temporal core across all samples. The core was defined as the taxa found across 100%
262 of the samples.

263 In the grapevine database, Zarraonaindia *et al.*, (2015) identify the OTUs shared
264 across grapevine organs (flower, leaves, grapes, root), the root zone, and bulk soil over
265 two growing seasons. The authors reduced the cutoff to 75% occurrence across samples
266 to determine the core community. This decision was justified by the authors due to the
267 lack of OTUs occurring across all samples.

268 The maize database is the only study included in our dataset that did not attempt
269 to identify the core community. The authors aimed to determine the impact of genetic
270 variation on the composition of bacterial communities inhabiting the maize rhizosphere
271 (Peiffer *et al.*, 2013).

272 The biological observation matrices (BIOM) derived from these data were
273 obtained from the Earth Microbiome Project (EMP; http://www.earthmicrobiome.org),
274 available on the *Qiita* platform (https://qiita.ucsd.edu). We used the BIOM files due to
275 the similar treatment of data by bioinformatics, including quality filters and assignment
276 of OTU taxonomy (Elli *et al.*, 2010; Caporaso *et al.*, 2011; Peiffer *et al.*, 2013;
277 Zarraonaindia *et al.*, 2015). We used the software *QIIME* (Chen and Lifschitz, 1989) to
278 convert the BIOM files into text files, which were further imported into the *R* software
279 (Team 2016), where we analyzed it using the packages *'RAM'* (Chen *et al.*, 2016),
280 *'vegan'* (Oksanen *et al.*, 2016) and *'Hmisc'* (Harrell Jr *et al.*, 2016).

The identification of the core microbial community is conventionally obtained by defining limits of frequency across the samples, *i.e.* a core community could be defined as microorganisms occurring in all samples (100% of occurrence frequency) or in a part of the samples (varying from 30% to 90% of frequency). For example, Ainsworth *et al.*, (2015) identified the ubiquitous endosymbiont bacterial community (or core community) associated with corals using a 30% occurrence frequency cut-off. Similarly, the human and grapevine studies were used determined the core community, respectively at levels of 100%, 100% and 75% occurrence frequency across the samples. We used a range of limits - 30, 40, 50, 60,70, 80, 90 and 100% occurrence frequency - based on the OTU tables across the samples to verify the difference in the core microbial community selected by these methods.

The method proposed here is based on the probability test for the distribution of each microbial taxon (OTU) among samples. This probability test is based on the Poisson distribution, which is a discrete random probability regression model. The Poisson distribution expresses the probability of an event taking place at a given point in time (Rao and Rubin, 1964). Here we treat events as OTUs across a series of collected samples. The Poisson distribution has previously been used in biogeographic studies to predict the abundance of species in a given ecosystem (Vincent and Haworth, 1983; Guisan and Zimmermann, 2000).

Following the idea proposed in the Phylotype Sampling Model (Turnbaugh *et al.*, 2009), the Poisson distribution was used to verify the sampling error expected given the sample size and the probability of observing the minimum abundance of a microorganism in any sample. However, the major difference from the previously methods including the Phylotype Sampling Model is that our proposed method does not present abundance or frequency thresholds. The probability (***P***) of Poisson distribution

306 is obtained by $\boldsymbol{P}(\boldsymbol{x}) = \boldsymbol{\lambda}^{\boldsymbol{x}}\, \boldsymbol{e}^{-\boldsymbol{\lambda}}/\boldsymbol{x}!$, where the lambda ($\boldsymbol{\lambda}$) and **x** represent the average of
307 relative abundance and the occurrence frequency of each taxon across the communities,
308 respectively. Using this formula, we have tested two hypotheses: $\boldsymbol{H_0}$ – the individual
309 (OTU) fits in the Poisson distribution and thus likely occurs in every sample (95% of
310 confidence), indicating that it cannot be excluded from the core microbial community;
311 $\boldsymbol{H_1}$ – the individual does not fit in the Poisson distribution, and thus is unlikely to occur
312 in every sample, supporting its exclusion from the core microbial community.

313 The calculation starts with the determination of the average of sequences per
314 community source ($\boldsymbol{N}$), the average relative abundance of each taxon across
315 communities ($\boldsymbol{p}$) and the occurrence frequency of each taxon across communities ($\boldsymbol{f}$).
316 The $\boldsymbol{p}$ and $\boldsymbol{f}$ are calculated with values of $\boldsymbol{A}$ and $\boldsymbol{rich} > 0$, and they are used in the
317 Poisson distribution, where the $\boldsymbol{\lambda}$ is obtained per OTU by the formula $\boldsymbol{\lambda} = \boldsymbol{N} \times \boldsymbol{p}$.

318 The goodness-of-fit of the Poisson model to distribution of OTUs were
319 determined from the R^2 (adjusted) and *p-value*. The goodness of fit (R^2) indicates the
320 level of variance of an OTU's relative abundance explained by the Poisson distribution,
321 which in this case is correlated with the proportion of microbial community that could
322 be not excluded as possible member of the core microbial group. The *p-value* is used to
323 calculate the significance of OTUs predicted as probable core members by the Poisson
324 distribution.

325 The arbitrary (thresholds of 30, 40, 50, 60, 70, 80, 90 and 100%) and the
326 proposed (Poisson distribution) methods resulted in OTU tables for the core microbial
327 community and the "variable" community (made of those that do not belong to the core
328 community). The statistical analyses comparing the results were performed using the R
329 software version 3.2.2 (R Core Team, 2015), including the Shannon index. We also
330 developed a function in R, which identifies a core microbial community by the method

331 based on the Poisson distribution. The R script of this function is available in
332 Supplementary Code Simplified file, and the description is available in Supplementary
333 Code Description file.

334

335 **ACKNOWLEDGEMENTS**

336 We thank FAPESP for the projects funding, 2014/22845-5 and 2013/18529-8. We are
337 grateful to the students in Brendan J.M. Bohannan's laboratory and the students from the
338 Institute of Ecology and Evolution for support. We thank the members of the Earth Microbiome
339 Project and the authors that made the database available. We also acknowledge the comments
340 and discussion provided by Annelise Mendes Nascimento, Clarisse Betancourt, Ademir Durrer,
341 and Trish Pasby throughout the manuscript preparation.

342

343 **References**

344 Ainsworth, T.D., Krause, L., Bridge, T., Torda, G., Raina, J.-B., Zakrzewski, M., et al.
345 (2015) The coral core microbiome identifies rare bacterial taxa as ubiquitous
346 endosymbionts. *ISME J.* **9**: 2261–2274.

347 Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A.,
348 Stombaugh, J., et al. (2011) Moving pictures of the human microbiome. *Genome*
349 *Biol.* **12**: R50.

350 Chen, H.M. and Lifschitz, C.H. (1989) Preparation of fecal samples for assay of volatile
351 fatty acids by gas-liquid chromatography and high-performance liquid
352 chromatography. *Clin. Chem.* **35**: 74–76.

353 Chen, W., Simpson, J., and Levesque, C.A. (2016) RAM: R for Amplicon-Sequencing-
354 Based Microbial-Ecology.

355 Elli, M., Colombo, O., and Tagliabue, a. (2010) A common core microbiota between
356 obese individuals and their lean relatives? Evaluation of the predisposition to
357 obesity on the basis of the fecal microflora profile. *Med. Hypotheses* **75**: 350–352.

358 Feinstein, L.M., Woo, J.S., Blackwood, C.B., Sul, W.J., Blackwood, C.B., Woo, J.S., et
359 al. (2009) Assessment of bias associated with incomplete extraction of microbial
360 DNA from soil. *Appl. Environ. Microbiol.* **75**: 5428–5433.

361 Guisan, A. and Zimmermann, N.E. (2000) Predictive habitat distribution models in
362 ecology. *Ecol. Modell.* **135**: 147–186.

363 Hamady, M. and Knight, R. (2009) Microbial community profiling for human
364 microbiome projects: Tools, techniques, and challenges. *Genome Res.* **19**: 1141–
365 1152.

366 Harrell Jr, F.E., with contributions from Charles Dupont, and many others. (2016)

Hmisc: Harrell Miscellaneous.

Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Abecia, L., et al. (2015) Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Sci. Rep.* **5**: 14567.

Kanagawa, T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**: 317–323.

Karlis, D. (2003) An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* **30**: 63–77.

Li, K., Bihan, M., and Methé, B.A. (2013) Analyses of the stability and core taxonomic memberships of the human microbiome. *PLoS One* **8**: e63139.

Lladser, M.E., Gouet, R., and Reeder, J. (2011) Extrapolation of urn models via Poissonization: Accurate measurements of the microbial unknown. *PLoS One* **6**: e21105.

Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**: 86–90.

McMurdie, P.J. and Holmes, S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput. Biol.* **10**: e1003531.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., et al. (2016) Vegan: Community Ecology Package.

Pédron, T., Mulet, C., Dauga, C., Frangeul, L., Chervaux, C., Grompone, G., and Sansonettia, P.J. (2012) A crypt-specific core microbiota resides in the mouse colon. *MBio* **3**: 1–7.

Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.* **14**: 257–263.

Peiffer, J.A., Spor, A., Koren, O., Jin, Z., Tringe, S.G., Dangl, J.L., et al. (2013) Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci.* **110**: 6548–6553.

Prosser, J.I. (2015) Dispersing misconceptions and identifying opportunities for the use of "omics" in soil microbial ecology. *Nat. Rev.* **13**: 439.

R Core Team (2015) R: A Language and Environment for Statistical Computing.

Rao, C.R. and Rubin, H. (1964) On a characterization of the poisson distribution. *Indian J. Stat.* **26**: 295–298.

Roeselers, G., Mittge, E.K., Stephens, W.Z., Parichy, D.M., Cavanaugh, C.M., Guillemin, K., and Rawls, J.F. (2011) Evidence for a core gut microbiota in the zebrafish. *ISME J.* **5**: 1595–608.

Saunders, A.M., Albertsen, M., Vollesen, J., and Nielsen, P.H. (2015) The activated sludge ecosystem contains a core community of abundant organisms. *ISME J.* **10**: 11–20.

Schlaeppi, K., Dombrowski, N., Oter, R.G., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2014) Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 585–92.

Sekelja, M., Berget, I., Næs, T., and Rudi, K. (2011) Unveiling an abundant core microbiota in the human adult colon by a phylogroup-independent searching approach. *ISME J.* **5**: 519–531.

Shade, A. and Gilbert, J.A. (2015) Temporal patterns of rarity provide a more complete view of microbial diversity. *Trends Microbiol.* **23**: 335–340.

Shade, A. and Handelsman, J. (2012) Beyond the Venn diagram: The hunt for a core microbiome. *Environ. Microbiol.* **14**: 4–12.

Shade, A., Jones, S.E., Caporaso, J.G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, J.A. (2014) Conditionally Rare Taxa Disproportionately Contribute to Temporal Changes in Microbial Diversity. *MBio* **5**: e01371-14.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proc. Natl. Acad. Sci.* **103**: 12115–12120.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., et al. (2009) A core gut microbiome in obese and lean twins. *Nature* **457**: 480–484.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007) The Human Microbiome Project. *Nature* **449**: 804–810.

Vincent, P.J. and Haworth, J.M. (1983) Poisson regression models of species abundance. *J. Biogeogr*. **10**: 153–160.

Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**: 27.

Yeoh, Y.K., Paungfoo-Lonhienne, C., Dennis, P.G., Robinson, N., Ragan, M.A., Schmidt, S., and Hugenholtz, P. (2015) The core root microbiome of sugarcanes cultivated under varying nitrogen fertiliser application. *Environ. Microbiol.* **18**: 1338–1351.

Zarraonaindia, I., Owens, S.M., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., et al. (2015) The Soil microbiome influences grapevine-associated microbiota. *MBio* **6**: e02527-14.

Zaura, E., Keijser, B.J.F., Huse, S.M., and Crielaard, W. (2009) Defining the healthy "core microbiome" of oral microbial communities. *BMC Microbiol.* **9**: 259.

441 **Table legends**

442 **Table 1** – Databases selected from EMP on the *Qiita platform.*

443

444 **Table 2** – Number of OTU's identified by the arbitrary and proposed method (based on

445 the Poisson distribution) across the datasets

446

447

448 **Figure legends**

449 **Figure 1** – The core and variable communities of the mice microbiome

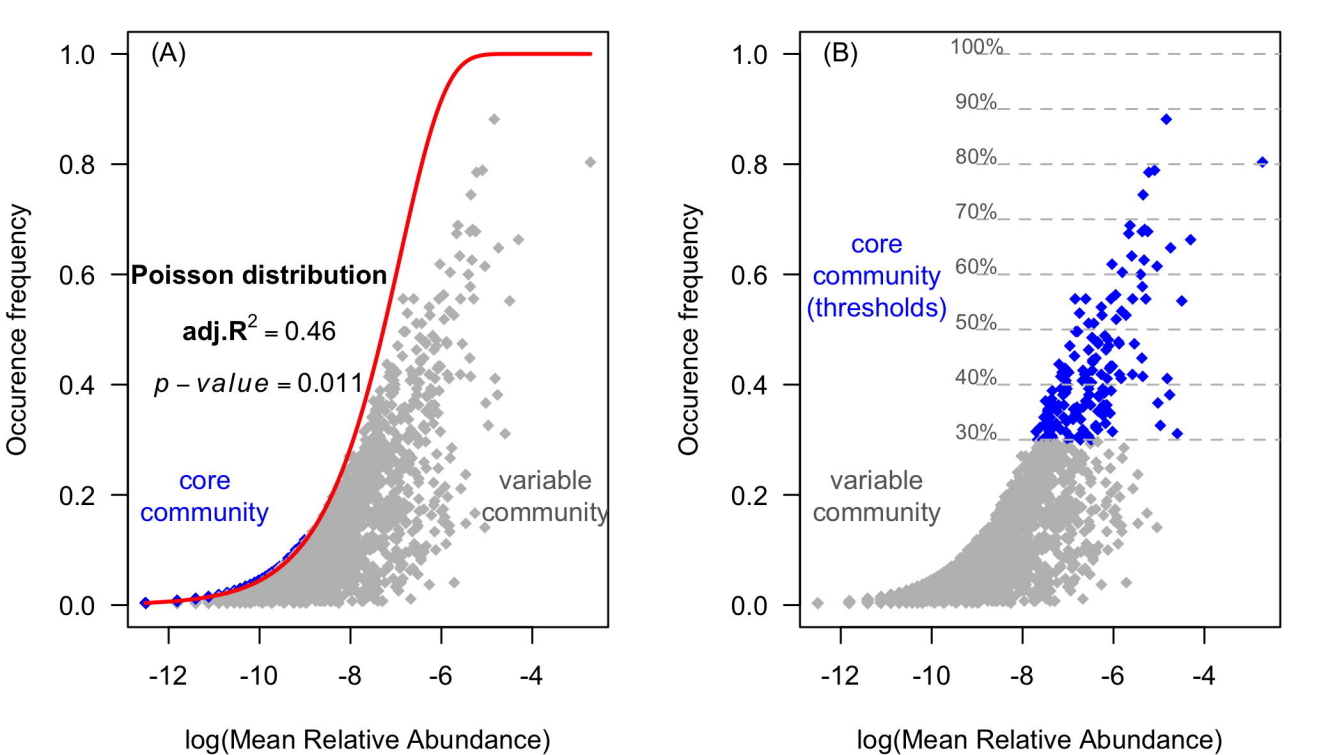
450 determined by (A) our proposed method based on the Poisson distribution and (B) an

451 arbitrary, threshold-based method.

452 **Figure 2** – Percentage of the relative abundance of the core communities of the

453 mice database determined by arbitrary methods (thresholds of 30,40,50,60,70,80,90 and

454 100%) and by our proposed method (Core Poisson).


(A)
Occurence frequency
1.0
0.8
0.6
0.4
0.2
0.0
Poisson distribution
adj.R2 = 0.46
p − value = 0.011
core
community
variable
community
-12
-10
-8
-6
-4
log(Mean Relative Abundance)
(B)
100%
90%
80%
70%
60%
50%
40%
30%
core
community
(thresholds)
variable
community

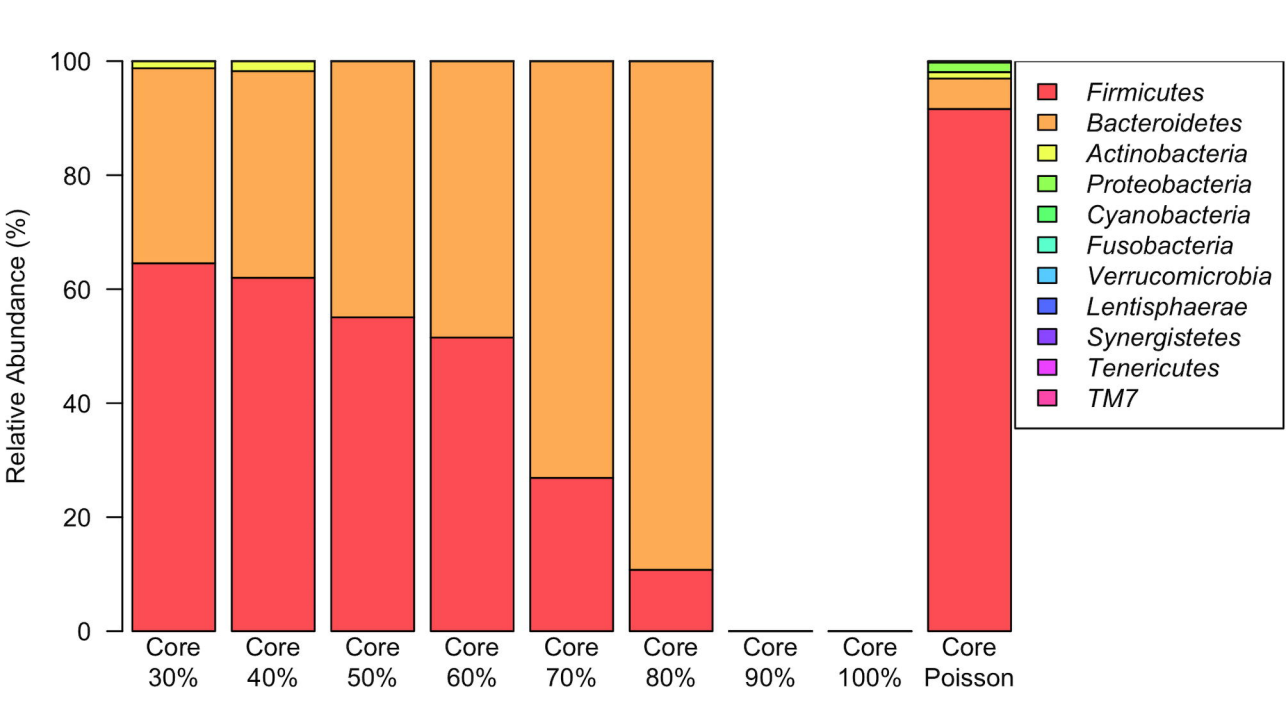

Relative Abundance (%)
100
80
60
40
20
0
Core 30%
Core 40%
Core 50%
Core 60%
Core 70%
Core 80%
Core 90%
Core 100%
Core Poisson
Firmicutes
Bacteroidetes
Actinobacteria
Proteobacteria
Cyanobacteria
Fusobacteria
Verrucomicrobia
Lentisphaerae
Synergistetes
Tenericutes
TM7

Table 1 – Databases selected from EMP on Qiita plataform

	Databases selected from EMP			
	Grape	**Maize**	**Human**	**Mice**
Study EMP – ID	2382	1792	550	77
***Qiita* Link**	https://qiita.ucsd.edu/study/description/2382	https://qiita.ucsd.edu/study/description/1792	https://qiita.ucsd.edu/study/description/550	https://qiita.ucsd.edu/study/description/77
Title	The Soil Microbiome Influences Grapevine-Associated Microbiota	Diversity and heritability of the maize rhizosphere microbiome under field conditions	Moving pictures of the human microbiome	A core gut microbiome in obese and lean twins
Number of samples	401	442	1,736	271
Data Type	16S - HiSeq	16S – 454 FLX	16S – 454 FLX	16S – 454 FLX
Number of reads / sample	1,000	2,080	5,000	1,000
OTUs	8,583	10,747	16,129	4,495
Reference	(Zarraonaindia *et al.*, 2015)	(Peiffer *et al.*, 2013)	(Caporaso *et al.*, 2011)	(Turnbaugh *et al.*, 2009)

Table 2 – Number of OTU's identified by the arbitrary and proposed method (based on the Poisson distribution) across the datasets

Methods		Databases							
		Grapevine		Maize		Human		Mice	
		Core community	Variable community	Core community	Variable community	Core community	Variable community	Core community	Variable community
Conventional method	30%	211	8,372	272	10,475	206	15,923	170	4,325
	40%	109	8,474	145	10,602	93	16,036	82	4,413
	50%	40	8,543	80	10,667	42	16,087	35	4,460
	60%	15	8,568	39	10,708	24	16,105	19	4,476
	70%	5	8,578	19	10,728	12	16,117	5	4,490
	80%	0	8,583	5	10,742	2	16,127	2	4,493
	90%	0	8,583	3	10,744	0	16,129	0	4,495
	100%	0	8,583	0	10,747	0	16,129	0	4,495
Proposed method		5,039	3,544	5,294	5,453	8,751	7,378	1,717	2,778