Standardizing microbial community filtering for the purpose of MWAS and mGWAS through algorithmic process

LISA PARUIT *under the supervision of* DR. PIERRE HOHMANN

*lisa.paruit@agroparistech.fr

Abstract: MWAS and mGWAS prove to be great tools for the study of plant-microbe interactions and thus be used in the development of microbiome-assisted genomic selection. Plant associate microbiota are very complex data with lots of artifacts (ie. detected taxa that are not part of the plant associated functional microbial community) that increase the dimension of the data while not being useful to the analysis. We investigate a mean to assess for the trade-off between the loss of statistical power and information gain through mathematical modeling of the information held out by microbial communities and algorithmic filtering of the experimental data. As the implemented model return satisfying results, it is therefore promising for building up a complete decision support system. However, further development and experiments still need to be carried out in order to achieve this goal.

Introduction

Plant associated microbiota have a significant impact on their host crops functions as they form complex interactions on the molecular and even the genetic level. Previous GWAS, mGWAS and MWAS have shown that plant genome, plant microbiome and plant phenotype are intertwined in a complex dependence relationship. Such analytical methods that allow us to establish coorelation between genome and phenotype for GWAS, genoma and microbiome for mGWAS and phenotype with microbiome for MWAS, pen the door to microbiome-enhanced genomic selection (MEGS), meaning using predictive models that use all three GWAS, mGWAS and MWAS results to predict the host's phenotype [1].

However, MWAS and mGWAS encounter the same pitfalls as early GWAS, namely the lack of stringent criteria for reporting association, the adoption of common data formats and that of common analytic workflow [2]. After amplicon sequencing, sequencing data must be filtered to cut out the noise from the analysis. Pipelines such as iTagger [3] to obtain OTUs, UPARSE [4] to remove chimeras or more comprehensive processing tools such as DADA2 to perform quality filtering of the sequence as well as taxonomic assignment [5]. The raw OTU table is then often filtered out again in order to remove taxa that are supposed to be too sparsely present to have a significant impact on the functional analysis of the community and whose presence reduces statistical power for useless purposes. This trade-off between dimension reduction and information loss is the main limitation of MWAS and mGWAS and resides in the complexity of the microbiome data [6]. In the literature, the late filtering process varies from one study to the other. Filtering criteria vary from red counts in a certain portion of the samples [7], [8], percentage of samples with non zero data [9], count number after normalization [10], and many more criteria to assess for the significance of the information provided by OTUs. From the perspective of developing more standard workflow, there is room for the design of a standard workflow to perform this OTU filtering for MWAS and mGWAS, which we aim to do in this study.

1. Design of the filtering process

1.1. Design of the filtering strategy

The current solution to circumvent the issue of statistical power loss due to multiple testing in high dimension data is to focus only on a subset of taxa or variants, although this can lead to getting rid of valuable information, hence limiting the possibility to discover novel associations [6]. Therefore, the subset community should discard taxa that don't have great importance on the ecological equilibrium of the whole community, meaning the filter should aim to reduce the total number of taxa within the community while :

1. **Improving sample representativeness :** In an ideal dataset, samples only contain taxa that are functionally important. The filter should therefore favor taxa that have higher relative abundance within each sample.
2. **Preserving diversity across samples :** Cutting out data should not prevent from keeping track of the microbial diversity in the analyzed community. The filter should thus assess for and limit the information loss through taxa elimination.

1.2. Design and implementation of the algorithm

When dealing with OTU datasets, scientists face the trade-off between statistical power and information loss through the process of cutting out very rare taxa. Since the appreciation of the exact definition of "rarity" for a microbial taxa depends on the dataset, the algorithm still requires human prompt for defining the level to which taxa are considered for discarding. The filter consists in an indicator (K) that varies with the the community composition and describes the sub-community quality according to the above criteria . The overall process is described in the following:

1. Rare taxa (considered for removal) are set apart from the rest of the OTUs
2. K is calculated for the primarily filtered sub-community (K_0)
3. For each taxa considered as rare:
   (a) The taxa is placed back in the sub-community
   (b) K is calculated for this modified community
   (c) If there is a rise in K value that reaches above a given threshold over K_0, the information loss through the removal of this taxa can be considered too important to legitimate the dimension reduction. The taxa is therefore reintroduced in the main sub-community for analysis.
4. The algorithm starts again until K does not reach the threshold anymore

All scripts and files can be found on : https://github.com/lisaparuit/Bona-Planta.git

2. Modeling sample representativeness

The filter should give more positive weight to the samples that contribute widely to the composition of sampled individuals. Taxa that are present in smaller amounts are more likely to be considered as artifacts or at least insignificant OTUs with regards to the rest of the microbiome.

2.1. Mathematical perspectives

From a mathematical perspective, taxa distribution does not follow any known distribution law (cf.Appendix A). Moreover, the distribution of taxa abundance differs from one sample to another. Even when comparing only root samples, one can see a difference between curves which is expected to get more conspicuous when comparing communities extracted from different ecosystems. In the absence of clear mathematical pattern, it is preferable to use ecological indicators in order to filter out sampled taxa.

2.2. Using evenness as an indicator of sample homogeneity

From an ecological perspective, increasing the representativeness of sample composition resumes to reducing the number of rare taxa to a minimum, within the limits of information loss with regards to other samples. Evenness is defined as follow according to Pielou's definition:

$$J = \frac{H'_i}{\log(S)}$$

with J = evenness of the community ; S = number of species within each sample ; H'_i = Shannon's index for sample i .

Previous modeling experiments (subsection B.1) have lead us to the conclusion that evenness is a good indicator to filter out rarer taxa within a sample only when abundant taxa are excluded from the filtered pool. However, with regards to the overall filtering strategy, this indicator proves to be a suitable one to assess for sample homogeneity.

3. Modeling functional diversity in the microbial community

We also want to keep track of the microbial diversity across the sampled plant. The microbial community used for the study should thus tend to keep a high diversity level. Microbial diversity can be understood from two perspectives: using either Bray-Curtis dissimilarity to characterize the difference between samples or Shannon's index to characterize the microbial community as an entity in itself.
However, Shannon's index has the disadvantage of giving more weight to taxa that have a high relative abundance in the overall the community since probability pi - called the evenness of species i - is given by the relative abundance of this species in the whole community. It therefore discards rare taxa that could yet be of great importance in only a few samples. It would moreover create bias in the filtering process as the relative abundance of the taxa is already taken into account by the first component of the indicator. Bray Curtis dissimilarity is therefore chose as a measure of heterogeneity between samples which is to keep at the same level or to enhance.

Bray-Curtis dissimilarity is a quantitative measure used to assess the compositional dissimilarity between two different samples based on the presence and abundance of species or taxa. It is therefore suitable to compare the similarity of species composition between different samples regardless of the taxa's relative abundance across the overall community. It is calculated through the following formula:

$$BC_{jk} = 1 - \frac{2 \times C_j k}{S_k + S_j}$$

with BC_{jk} = Bray Curtis dissimilarity between samples j and k ; C_{jk} = sum of the count of the species common to j and k ; S_j = number of species in sample j ; S_k = number of species in sample k.

4. Constructing the filtering indicator

The filtering indicator is designed by combining both the sum of evenness within samples and the sum of Bray Curtis dissimilarities between samples.

$$K = \sum_{i=1}^{N} \frac{H'_i}{\log(S)} \times \sum_{j}^{N} \sum_{k=j+1}^{N} 1 - \frac{2 \times C_{jk}}{S_k + S_j}$$

with BC_{jk} = Bray Curtis dissimilarity between samples j and k ; C_{jk} = sum of the count of the species common to j and k ; S_j = number of species in sample j ; S_k = number of species in sample k ; N = number of samples ; J = evenness of the community ; S = number of species within each sample ; H'_i = Shannon's index for sample i.

Both terms of the product are to be maximized in order to select a suitable community subset. Therefore the higher K is, the better the community it characterizes.

5. Results

Although a significance level was first though of for fixing the threshold, we couldn't agree on the right statistics. We therefore used a fixed threshold for demonstration purposes in this experiment. The pre-filtering was carried out using quite stringent but standard criteria for this type of study. One can notice on Figure 1 that there is not any taxon whose reintroduction resulted in a rise from the initial value of K (whole community). According to our hypothesis, this means that putting back any of the discarded taxa into the sub-community designed for analysis would be a sole loss in statistical power with no benefice on the information level of the community.

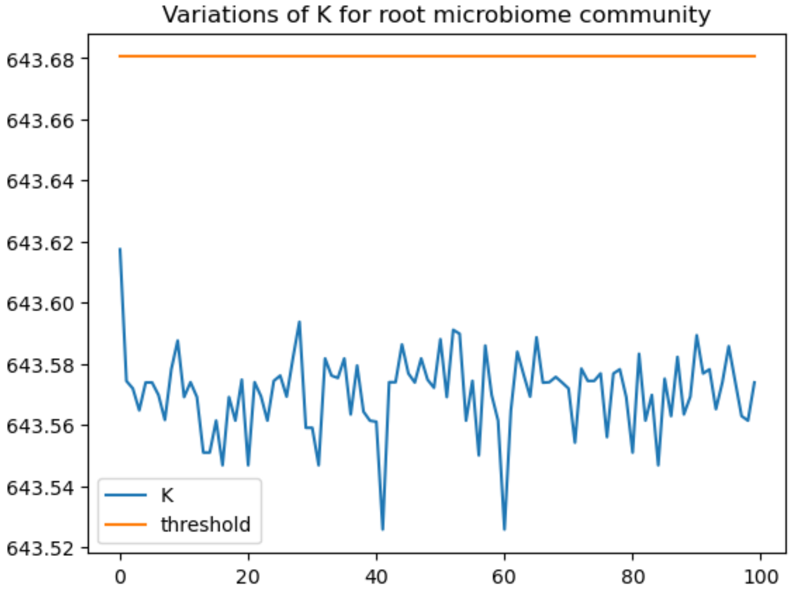
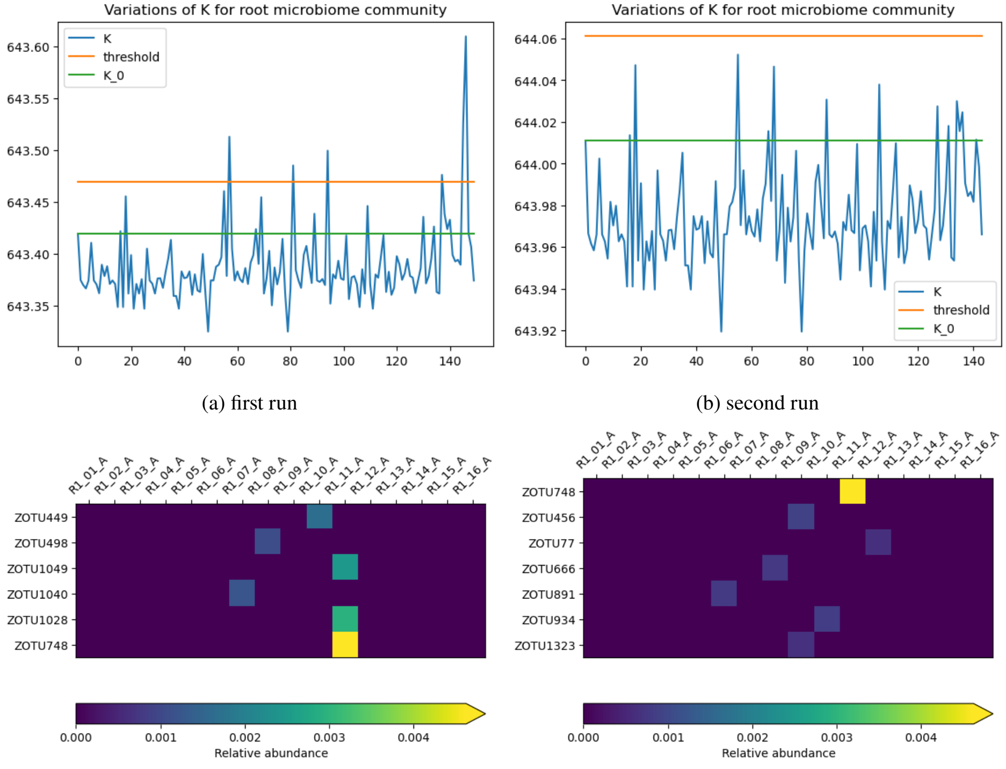


Figure 1. Display of the variations of K when placing back discarded taxa one by one into the pre-filtered sub-community.*y-axis = K value of the community* ; *x-axis = index of the OTU tested for reintroduction within the discarding bin. Chosen criteria for pre-filtering were to discard any OTU that shows less than* 10 *counts in* 10% *of the samples.*

The same experiment was carried out of the community using less stringent criteria. It is now conspicuous that some taxa are to be brought back into the main sub-community, by the presence of several upward peaks.



(a) first run

(b) second run

(c) heatmap of the taxa filtered back in during the first run.

(d) heatmap of the taxa triggering peaks in the second run

Figure 2. Display of the variations of K when placing back discarded taxa one by one into the pre-filtered sub-community. The algorithm takes 2 run with the randomly chosen threshold (T = 0.05) that are displayed on graph (a) and (b). Graphs (c) and (d) show heat maps of the taxa corresponding to the rise in K value in both graphs (a) and (b). For graph(d), the taxa represented taxa are not selected back with regards to our threshold and their relative abundance is compared to that of "ZOTU748" which was selected back in the previous run.
graphs (a) *and* (b) : *y-axis* $=$ *K value of the community* ; *x-axis* $=$ *index of the OTU tested for reintroduction within the discarding bin. Chosen criteria for pre-filtering were to discard any OTU that shows less than* 100 *counts in* 10% *of the samples. graph* (c) *and* (d) : *y-axis* $=$ *OTUs name* ; *x-axis* $=$ *sample name* ; *colorbar* $=$ *relative abundance of the taxa*

6. Next steps and discussion

6.1. Defining the acceptance threshold and improvement of the model using machine learning methods

Synthetic microbial communities (SynCom) refer to intentionally designed and constructed groups of microorganisms that are assembled to study their interactions, dynamics, and collective behaviors under controlled conditions. When transferred to a plant for colonization, the discrepancies between the original community and the sampled community after colonization indicate the colonization fails as well as the presence of artifacts [11]. They are therefore suitable to perform a systematic assessment for the effectiveness of the filter. Using an algorithm of supervised learning, we could therefore find a fixed threshold for the filtering out of taxa whose removal triggers a rise in K that would be above this limit.

Machine learning could also be used to improve the model performances. Indeed, one could think of adding modulating parameters (α and β) to each coefficient of the K indicator. Indeed, representativeness of each sample and dissimilarity between samples might not have the same weight on the information level provided by the community. We could therefore use the same process to adjust the model's parameters and define an even more accurate filter.

$$K_2 = \left(\sum_{i=1}^{N} \frac{H'_i}{\log(S)} \right)^{\alpha} \times \left(\sum_{j}^{N} \sum_{k=j+1}^{N} 1 - \frac{2 \times C_{jk}}{S_k + S_j} \right)^{\beta}$$

6.2. Taking functional interaction between taxa into account

For now, the model tests for the removal of each taxon that is considered for potential cutting off. However, microbial taxa show functional interactions within samples and therefore should be thought of as a multiple entity. One way to take this into account is to test, not only for the drop out of single taxa but of combinations of the tested taxa. This requires more computational power for one run but might *in fine* speed up the filtering process as several taxa can be removed at once.

6.3. Circumvent the need for human prompt

One could think of a trained algorithm to bypass the need for a human prompt for the definition of 'rare' taxa. An evolution of the filter could be to extend the functionalities to that of defining the pre-filtering criteria according to the composition of the overall community. This could be done through machine learning processes once the filtering threshold has been well defined.

7. Conclusion

The need for decision support software focusing of the trade-off between dimension reduction and information loss when designing microbial datasets for mGWAS and MWAS is great in the perspective of creating standardized workflow for these emerging study fields. In this paper, we suggest a way to implement such numeric tool, detail the theory and show positive results of the software's prototype. Next steps in the development process now consist in using supervised learning methods on synthetic communities in order to determine a good filtering threshold and improve the mathematical indicators underpinning the algorithmic process.

References

1. H. C. J. Y. Zhikai Yang, Tianjing Zhao, "Microbiome-enabled genomic selection improves prediction accuracy for nitrogen-related traits in maize," G3, Oxf. Univ. Press (2023).
2. B. E. S. E. Weissbrod Omer, Rothschild Daphna, "Gwas, mwas and mgwas provide insights into precision agriculture based on genotype-dependent microbial effects in foxtail millet," Curr. Opin. Microbiol. p. 19 (2018).
3. R. J. D. M. e. a. Bolyen, E., "Reproducible, interactive, scalable and extensible microbiome data science using qiime 2," Nat Biotechnol. p. 37 (2019).
4. E. RC, "Uparse: highly accurate otu sequences from microbial amplicon reads." Nat Methods p. 10 (2013).
5. M. J. R. A. W. H. A. J. A. J. S. P. H. Benjamin J Callahan, Paul J McMurdie, "Dada2: High-resolution sample inference from illumina amplicon data," Nat Methods p. 13 (2016).
6. S. D. S. H. K. S. M. N. E. T. A. G. N. M. E. R. C. Denis Awany, Imane Allali, "Host and microbiome genome-wide association studies: Current state and challenges." Front. Genet. (2018).
7. G. X. L. D. L. W. J. Y. D. C.-D. S. Deng, D.F. Caddell, "Genome wide association study reveals plant loci controlling heritability of the rhizosphere microbiome," The ISME J. (2021).
8. W. X. S. S. e. a. Wang, Y., "Gwas, mwas and mgwas provide insights into precision agriculture based on genotype-dependent microbial effects in foxtail millet," Nat. communication p. 13 (2022).

9. J. W. M. C. R. R. Y. T. G. F. M. J. S. V.-S. L. H. L. R. C. V. S. R. A. F. K. H. W. N. J. T. J. R. David A. Hughes, Rodrigo Bacigalupe, "Genome-wide associations of human gutmicrobiome variation and implications for causal inference analyses," The ISME J. (2020).
10. L. L. K. E. S. B. Jason G. Wallace, Karl A. Kremling, "Quantitative genetics of the maize leaf microbiome," Phytobiomes J. (2018).
11. I. J. O. Ambihai Shayanthan, Patricia Ann C. Ordoñez, "The role of synthetic microbial communities (syncom) in sustainable agriculture," Front. Agron.Sec. Plant-Soil Interactions (2022).

optica-article

[english]babel bookmark booktabs xcolor [table]xcolor [T1]fontenc graphicx [utf8]inputenc listings microtype multicol rotating subcaption tcolorbox tikz hyperref

Appendix

A. Descriptive analysis of the taxa distribution

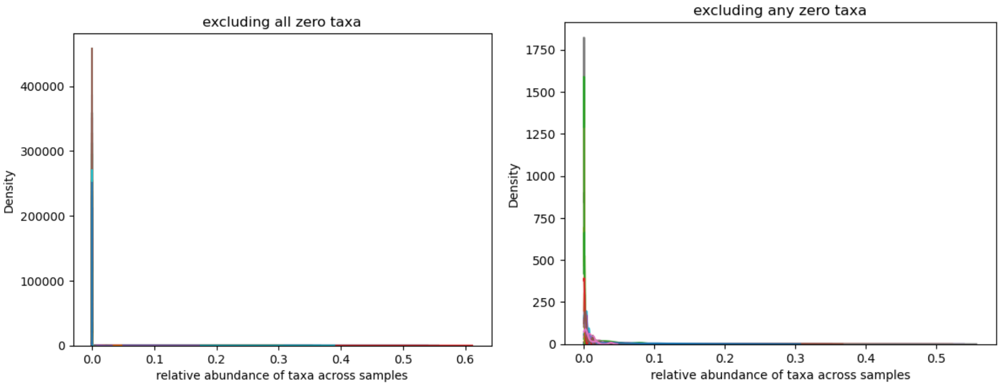


Figure 3. Distribution of taxa relative abundance for each taxa, excluding all zero taxa (left) or any zero taxa (right).*One can see that most of the taxa are present in very small relative abundance, even when this distribution is not null. This raises the question of how to handle both taxa that are highly abundant in some samples as well as taxa that show a much lower relative abundance.*

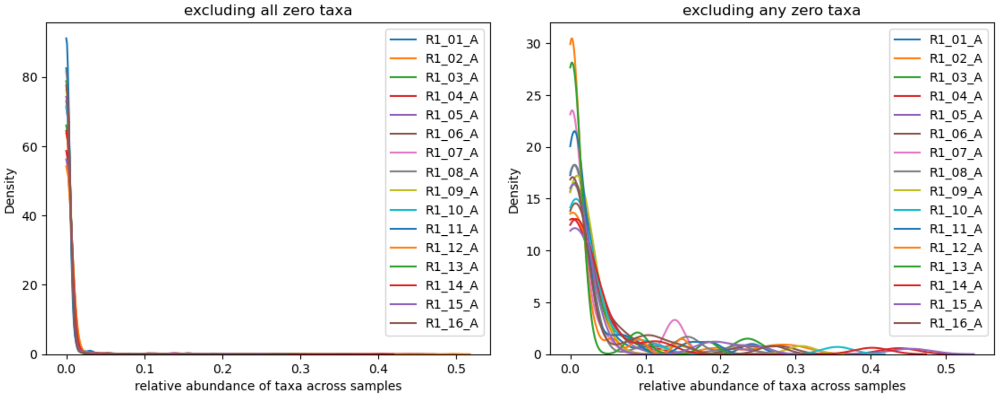


Figure 4. Distribution of taxa relative abundance for each sample, excluding all zero taxa (left) or any zero taxa (right).*Again, most of the taxa are present in very small relative abundance within samples. Most of the taxa, even when taking only taxa present in every sample into account, shows a small relative distribution. This distribution doesn't follow any well knows probability law, therefore compromising a mathematical filtering of the data.*

Samples	R1_01_A	R1_02_A	R1_03_A	R1_04_A	R1_05_A	R1_06_A	R1_07_A	R1_08_A	R1_09_A	R1_10_A	[. . .]
Count	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	1687.000000	[. . .]
Mean	0.000593	0.000593	0.000593	0.000593	0.000593	0.000593	0.000593	0.000593	0.000593	0.000593	[. . .]
Std	0.010995	0.008672	0.011986	0.012431	0.009468	0.009833	0.008480	0.008944	0.009612	0.007451	[. . .]
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	[. . .]
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	[. . .]
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	[. . .]
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	[. . .]
max	0.301169	0.241919	0.437238	0.452537	0.281010	0.316456	0.223897	0.307731	0.355284	0.188427	[. . .]

Table 1. Descriptive statistics of each sample composition in terms of relative abundance *One can see that if the mean remains the same for each sample the standard deviation changes, indicating that taxa distribution changes from one sample composition to another. This mainly is due to the variation in max values that is supposed to get lower with the number of abundant taxa found in the sample. This discrepancy between taxa distribution requires a filter that is able to deal with this diversity of cases.*

B. Explaining the effect of taxa dropout on community evenness

At this point of the project, the filtering algorithm is set to the following:

1. K is calculated for the whole community
   For each taxa in the community:
2. Removing the taxa from the community
3. Calculating K for this modified community
4. If K shows an increase that is above a given threshold that is relative to the K value of the original community, the removal of the taxa has a positive effect on the representativeness of the community and the taxa should therefore be discarded
5. Discard the taxa and start again until K doesn't reach the threshold anymore

This translates in Python as shown in Listing 1. The output of this function is displayed on Figure 5.

```
def run1 (data):
    # Remove rows with all zero values
    data = data.loc[(data!=0).any(axis=1)]
    # getting the number of OTUS
    S = data.shape[0]

    # progressbar
    bar = progressbar.ProgressBar(maxval=S+1,
                                  widgets=[progressbar.Bar('+', '[', ']'), ' ',
                                           progressbar.Percentage()])
    bar.start()

    # calculation
    kr_list = [K_calculator(data, None)[0]]
    sum_bc_list = [K_calculator(data, None)[1]]
    sum_e_list = [K_calculator(data, None)[2]]
    for i, taxa in enumerate(data.index):
        # calculating indicators without dropping taxa
        kr_list.append(K_calculator(data, taxa)[0])
        sum_bc_list.append(K_calculator(data, taxa)[1])
        sum_e_list.append(K_calculator(data, taxa)[2])
        # updating progressbar
        bar.update(i+1)
        time.sleep(0.1)

    # finishing bar
    bar.finish()

    # plotting
    X = list(range(S+1))
    #thresh = kr_list[0] + threshold
    plt.figure()
    plt.plot(X, kr_list, label='K')
    #plt.plot(X, [thresh]*(S+1), label='threshold')
    plt.title('Variations of K for root microbiome community')
    plt.legend()
    plt.show()

    # recording the K values in a DataFrame
    list_removal = np.array( [[data.index[i] , kr_list[i+1]]
                            for i in range(len(kr_list)-1) ])
    df = pd.DataFrame(list_removal, columns=["OTU_ID", "K_Value"])
    df.set_index(df["OTU_ID"], inplace=True)
    df = df.drop(labels ="OTU_ID", axis = 1)

    arr_details = np.array( [[data.index[i] ,
                              kr_list[i+1],
                              sum_bc_list[i+1],
                              sum_e_list[i+1]]
                              for i in range(len(kr_list)-1) ])
    df_details = pd.DataFrame(arr_details,
                              columns=["OTU_ID", "K_Value","Sum_BC", "Sum_E"])
    df_details.set_index(df_details["OTU_ID"], inplace=True)
    df_details = df_details.drop(labels ="OTU_ID", axis = 1)

    return df_details
```

Listing 1. First version of the model's main function.

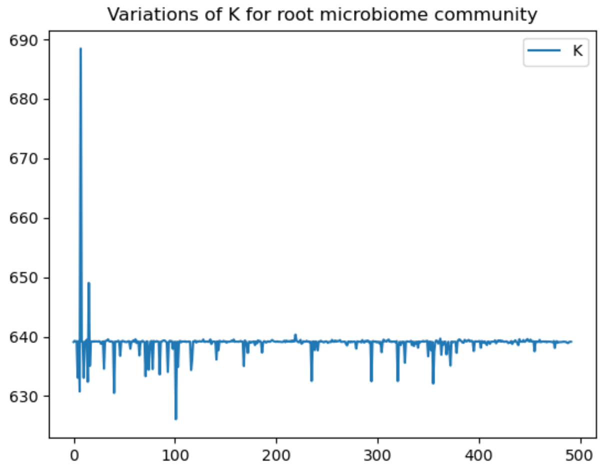


Figure 5. Variations of K with the removal of one taxa (index on the x-axis). For this first test, only the root microbiome samples (R) were used. *The model is applied to the whole community of root microbes, meaning that rare and abundant taxa dropouts are tested. One can notice the two large upward peaks that should, according to our hypothesis, indicate taxa to discard.*

A z-score test is then performed on the K value for each taxa in order to define if the value of K resulting from its removal is significantly different from the original value of K (see Table 2). A z-score measures how many standard deviations an element (in this case, an individual) is from the mean of a group (in this case, the community). It can therefore help one understand whether the individual is typical for the group or significantly different and is suitable for this purpose. The test p-value is then calculated to assess for the significance of the impact on K.

On Table 2, one can see that the two most significant variations of K value are positive variations. However, all the other significant variations are negative ones whicch is not what was expected. In order to figure out how to assess for the model's results, we want to show the distribution of these taxa across samples (Figure 6).

On Figure 6, one can see that the OTUs whose removal increases the value of K show a high relative abundance. They would hence not be meant to be discarded whereas the less significant

OTU_ID	Z score	P value	K - K_0	Significance level
ZOTU1	18.552003	0.000000	49.349180	***
ZOTU10	3.734246	0.000188	9.933265	***
ZOTU11	-4.883935	0.000001	-12.991492	***
ZOTU7	-3.221508	0.001275	-8.569360	**
ZOTU9	-3.128727	0.001756	-8.322558	**
ZOTU26	-2.630187	0.008534	-6.996418	**
ZOTU31	-2.267353	0.023369	-6.031262	**
ZOTU4	-2.255677	0.024091	-6.000204	**
ZOTU5	-2.518053	0.011801	-6.698136	**
ZOTU13	-2.486755	0.012891	-6.614882	**
ZOTU21	-2.473336	0.013386	-6.579186	**
ZOTU16	-2.461930	0.013819	-6.548846	**
ZOTU20	-2.164793	0.030404	-5.758448	**
ZOTU12	-2.057310	0.039656	-5.472538	**
ZOTU29	-1.896738	0.057862	-5.045411	*
ZOTU49	-1.773909	0.076078	-4.718681	*
ZOTU25	-1.760909	0.078254	-4.684098	*
ZOTU40	-1.717512	0.085886	-4.568660	*
ZOTU38	-1.692232	0.090602	-4.501415	*

Table 2. Details of the Z-score test for K value variation when the model in ran on the entire community unsing **run1**(). Only taxa whose removal leads to the most significant increases or decreases in the value of K for the community are shown. Columns refer to (i) the z-score, (ii) the p value associated, (iii) the difference with the value of K for the initial community (K_0 = 639.0992576612612) and (iv) the level of significance of this difference.
significance level : _p-value_ > 0.1 = ; 0.1 > _p-value_ > 0.001 = * ; 0.05 > _p-value_> 0.001 = ** ; 0.001 > _p-value_ .

OTUs would be a conspicuous choice. This run of the model therefore disproofs our hypothesis. The effect of the OTU distribution on this version of K shall thus be further investigated in order to improve either the original formula or its interpretation.

The OTUs that have a significant impact on K Table 2 don't necessarily also have a significant impact on both the sum of Bray Curtis dissimilarity (Table 3) and the sum of evenness within samples (Table 4). Moreover, the OTUs that are found in all three tables are not always the ones that show the greatest difference with the original community in terms of K value, due to the non linearity of the K filter criterion.

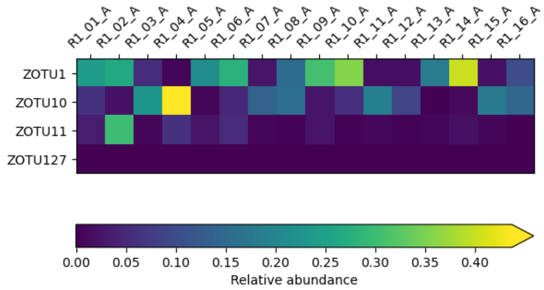


Figure 6. Heatmap of the relative abundance of OTUs that show the most significant K increase (ZOTU1 and 10), decrease (ZOTU11) and ne of the less significant OTU (ZOTU127).

OTU_ID	Z score	P value	SumBC - SumBC_0	Significance level
ZOTU1	16.281260	0.000000e+00	3.805773	***
ZOTU13	-6.338824	2.315250e-10	-1.481711	***
ZOTU3	5.426073	5.760727e-08	1.268354	***
ZOTU10	5.713030	1.109822e-08	1.335431	***
ZOTU7	-5.342623	9.161135e-08	-1.248847	***
ZOTU12	5.075410	3.866610e-07	1.186386	***
ZOTU4	4.115245	3.867681e-05	0.961946	***
ZOTU9	-4.284846	1.828658e-05	-1.001590	***
ZOTU21	-4.330396	1.488418e-05	-1.012237	***
ZOTU11	-2.444311	1.451289e-02	-0.571362	**

Table 3. Details of the Z-score test for the sum of Bray Curtis dissimilarity between samples variation when the model in ran on the entire community. Only taxa whose removal leads to the most significant increases or decreases in the value of K for the community are shown. Columns refer to (i) the z-score, (ii) the p value associated, (iii) the difference with the value of K for the initial community (SumBC_0 = 80.77275480129987) and (iv) the level of significance of this difference.
significance level : *p-value* > 0.1 = ; 0.1 > *p-value* > 0.001 = * ; 0.05 > *p-value*> 0.001 = ** ; 0.001 > *p-value* .

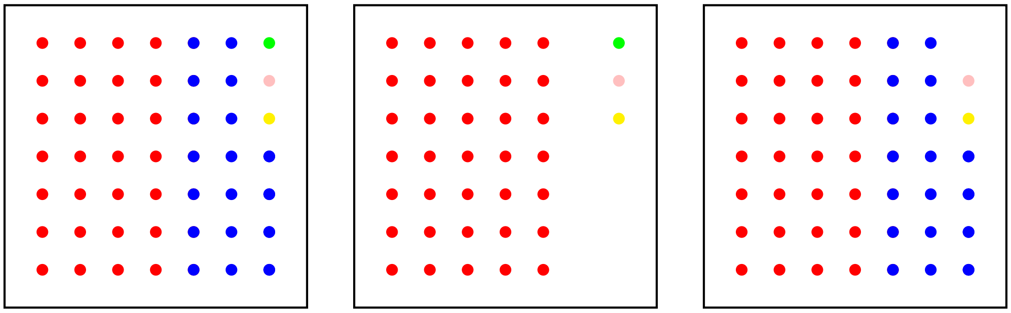
Focus on ZOTU1 This microbe is present in non-negligible relative abundance in several samples. As a significant taxa, it would therefore be expected that its removal lowers the value of K. However, both the sum of the evenness within samples and the sum of Bray Curtis dissimilarity across samples are increased by its removal. Furthermore, it is the OTU whose removal has the strongest positive effect on both indicators, individually as well as combined in the K formula. This shows that the model doesn't behave as expected, giving more positive impact on K to the removal of abundant taxa than very rare taxa which we are aiming to filter.

OTU_ID	Z score	P value	SumE - SumE_0	Significance level
ZOTU1	11.294886	0.000000e+00	0.227442	***
ZOTU12	-9.003690	0.000000e+00	-0.181305	***
ZOTU4	-8.270046	2.220446e-16	-0.166532	***
ZOTU3	-5.323187	1.019651e-07	-0.107192	***
ZOTU11	-5.245033	1.562542e-07	-0.105618	***
ZOTU31	-4.198163	2.690888e-05	-0.084537	***
ZOTU34	-3.366770	7.605416e-04	-0.067796	***
ZOTU5	-3.795982	1.470599e-04	-0.076439	***
ZOTU13	3.199724	1.375592e-03	0.064432	**
ZOTU29	-3.136417	1.710259e-03	-0.063157	**
ZOTU38	-3.135544	1.715358e-03	-0.063140	**
ZOTU20	-3.039388	2.370595e-03	-0.061203	**
ZOTU25	-3.001356	2.687804e-03	-0.060437	**
ZOTU19	-2.973709	2.942238e-03	-0.059881	**
ZOTU26	-2.878640	3.993939e-03	-0.057966	**
ZOTU16	-2.663587	7.731241e-03	-0.053636	**
ZOTU56	-2.397824	1.649277e-02	-0.048284	**
ZOTU46	-2.026834	4.267938e-02	-0.040814	**
ZOTU41	-1.707816	8.767047e-02	-0.034390	*
ZOTU49	-1.712408	8.682143e-02	-0.034482	*
ZOTU42	-1.720030	8.542704e-02	-0.034636	*

Table 4. Details of the Z-score test for the sum of evenness within each sample variation when the model in ran on the entire community. Only taxa whose removal leads to the most significant increases or decreases in the value of K for the community are shown. Columns refer to (i) the z-score, (ii) the p value associated, (iii) the difference with the value of K for the initial community (SumE_0 = 7.912312254713098) and (iv) the level of significance of this difference.
significance level : *p-value* > 0.1 = ; 0.1 > *p-value* > 0.001 = * ; 0.05 > *p-value*> 0.001 = ** ; 0.001 > *p-value* .

B.1. Limitation of evenness as an indicator

One can also observe that ZOTU10 shows a stringent significance in the increase of K value (Table 2) but seems to be of interest for mGWAS from the heatmap of its relative abundance across the community (Figure 6). Its effect comes from a strong positive effect on evenness that the fall in Bray Curtis dissimilarity sum doesn't balance. In this part, we suggest a hypothesis to explain the positive effect of the removal of ZOTU1 and 10 on evenness between samples as they are yet quite abundant taxa.



(a) Community A

(b) Community A'

(c) Community A"

Figure 7. Diagram illustrating the effect of the removal of one abundant (A') or rare (A") taxa from a community with a very heterogeneous taxa distribution (A) like it is often the case in plant microbiome communities.

Evenness within samples was chosen to screen for the sparsely distributed taxa as the evenness of the sample improves when only abundant taxa remain there. However, this case reveals a weakness of this method : when 2 species are equally frequent in a sample unit (points yellow and pink in figure 6), removing one of these two most abundant species reduces equitability but increases evenness in the sample (A to A' Figure 7). If removing all the rare taxa (green, blue and purple) at once, leaving only the most abundant species (yellow and pink), the evenness would indeed increase. However, when removing taxa one by one, the evenness increases much more when removing an abundant taxa (from A to A') than removing a rare one (from A to A").

B.2. Correcting the filtering process to prevent this unwanted effect

A way to circumvent this unwanted behavior of the model is to carry out a prefiltering of abundant taxa. Indeed, in the literature, filtering criteria for OTU datasets are more or less stringent but never concern highly abundant taxa which are always kept for analysis. From this, changing the filtering algorithm to the following would solve this issue.

1. Rare taxa (considered for removal) are set apart from the rest of the OTUs
2. K is calculated for the primarily filtered sub-community (K_0)
3. For each taxa considered as rare:
    (a) The taxa is placed back in the sub-community
    (b) K is calculated for this modified community
    (c) If there is a drop in K value that reaches below a given threshold from K_0, the information loss through the removal of this taxa can be considered too important to legitimate the dimension reduction. The taxa is therefore reintroduced in the main sub-community for analysis.
4. The algorithm starts again until K does not reach the threshold anymore

With this algorithm, the filter becomes a mean to screen for information loss through the choice of the filtering criteria by scientists for their experiment. This method requires human prompts but seem to suit more the need of MWAS and mGWAS as it gives the liberty for scientists to adapt the process to their datasets and the purpose of their experiment.

Here, instead of applying the filtering process to the entire community, only the taxa that are considered as rare would be tested. This idea arises from the virtual experiment described on Table 5.

Table 5 reflects the phenomenon explained in subsection B.1. However, when filtering is applied to rare taxa only, one can see the evenness decreases with the number of taxa and the presence of rare taxa in the community. It thus becomes a suitable indicator for the trade-off between overall taxa number reduction and abundant taxa number maximisation.

Composition	Evenness
(R2, R3)	1.0
(A1, A2)	0.995727
(R1, R2, R3)	0.864974
(R1, R2, R3)	0.864974
(R1, R2)	0.811278
(R1, R3)	0.811278
(A1, A2, R1)	0.817582
(A1, A2, R2)	0.718941
(A1, A2, R3)	0.718941
(A1, A2, R1, R2)	0.712525
(A1, A2, R1, R3)	0.712525
(A1, A2, R1, R2, R3)	0.667184
(A1, A2, R2, R3)	0.638565
(A2, R1)	0.591673
(A1, R1)	0.543564
(A2, R1, R2, R3)	0.526704

Table 5. Modeling the effect of changes in community composition on community evenness. *A simplified community composed of* 2 *abundant taxa* $(A1$ *and* $A2)$ *and* 3 *rare taxa* $(R1, R2, R3)$ *distributed according to the following count* :

$$A1 : 21, A2 : 18, R1 : 3, R2 : 1, R3 : 1$$

. *This distribution seems quite close to that in the plant microbiome community, having several abundant taxa making up from* 25 *to* 50% *of the community and appearing in* 80 *to* 100% *of the samples and very rare taxa with a relatively insignificant count.*