Title

Poisson hurdle model-based method for clustering microbiome features

Journal

Bioinformatics, 39(1)

ISSN

1367-4803

Authors

Qiao, Zhili

Barnes, Elle

Tringe, Susannah

et al.

Publication Date

2023

DOI

10.1093/bioinformatics/btac782

System Biology

Poisson hurdle model-based method for clustering microbiome features.

Zhili Qiao [1], Elle Barnes [2], Susannah Tringe [2,3], Daniel P. Schachtman [4], and Peng Liu [1]*

[1]Department of Statistics, Iowa State University, Ames, IA, USA; [2]Department of Energy, Joint Genome Institute, Berkeley, CA, USA; [3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [4]Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE, USA

*Peng Liu or Zhili Qiao.

Abstract

Motivation: High-throughput sequencing technologies have greatly facilitated microbiome research and have generated a large volume of microbiome data with the potential to answer key questions regarding microbiome assembly, structure, and function. Cluster analysis aims to group features that behave similarly across treatments, and such grouping helps highlight the functional relationships among features and may provide biological insights into microbiome networks. However, clustering microbiome data is challenging due to the sparsity and high-dimensionality.

Results: We propose a model-based clustering method based on Poisson hurdle models for sparse microbiome count data. We describe an expectation-maximization algorithm and a modified version using simulated annealing to conduct the cluster analysis. Moreover, we provide algorithms for initialization and choosing the number of clusters. Simulation results demonstrate that our proposed methods provide better clustering results than alternative methods under a variety of settings. We also apply the proposed method to a sorghum rhizosphere microbiome dataset that results in interesting biological findings.

Availability: R package is freely available for download at https://cran.r-project.org/package=PHclust.

Contact: pliu@iastate.edu or zlqiao@iastate.edu

Supplementary information: Supplementary Materials are available at *Bioinformatics* online.

1 Introduction

Over the past two decades, the number of microbiome studies has grown rapidly due to the advancement of next generation sequencing (NGS) technologies. With lower cost and increasing computational power, we are able to obtain tremendous amounts of data regarding the diversity and function of the microbiome from a host or a habitat. One of the most popular NGS approaches is the amplicon-based sequencing (Poretsky *et al.*, 2014) which generates data matrices of amplicon sequence variants (ASVs) or operational taxonomic units (OTUs) where ASVs and OTUs are unique taxonomic features. The ASV/OTU/taxa table is the starting point for most statistical analysis. However, these datasets present some challenges: they are high-dimensional and sparse (i.e., contain many zeroes), and there is high variability in sequencing depth across different samples (Cullen *et al.*, 2020). These data characteristics make many classic and popular data analysis approaches not directly applicable to microbiome data, and call for development of new statistical methods.

Cluster analysis has been a popular method of multivariate data analysis that help identify relationships among high-dimensional variables. It has been widely applied to high-dimensional gene expression data (Yeung *et al.*, 2001). Applied to microbiome data, cluster analysis can help identify potential microbiome sub-communities, which give insights into how features (ASV/OTU/taxa) with similar abundance levels are grouped together. With cluster analysis, researchers can more easily identify potential species patterns from highly diverse datasets. For example, clusters may represent taxa (or strains) that are functionally related to each other (i.e., guilds) or that share sensitivity to certain environmental conditions (i.e., niche selection), which can be further probed by downstream metagenomic techniques. When applied to time series data, clustering approaches could also help identify changes in microbial community states, which could provide important insights into microbiome assembly and manipulation.

Despite the increasing demand for methods for microbiome data analysis including cluster analysis, there are not many clustering algorithms developed specifically for microbiome data. To cluster microbial features, Gloor *et al.* (2016) applied the K-means clustering, Badri *et al.* (2020) used spectral clustering, while Casero *et al.* (2017) applied the model-based Poisson clustering developed by Si *et al.* (2014) for RNA-sequencing data. To cluster samples or microbial communities, Zhang *et al.* (2017) and Lonèar-Turukalo *et al.* (2019) applied spectral clustering while the latter paper proposed an implementation of kernel PCA for data reduction. None of the above-mentioned methods take into account the excessive zeros (sparsity) in microbiome data, which is a common issue especially after rarefaction (McMurdie and Holmes, 2014; Gloor *et al.*, 2017). Such sparsity has made more and more researchers believe that the excessive zero counts in microbiome data need to be treated differently (Xu *et al.*, 2015).

In this manuscript, we propose a model-based algorithm for clustering microbiome features based on Poisson hurdle models. The hurdle models, introduced by Cragg (1971), separately model the zero part and the non-zero part of a random variable and hence naturally allow zero inflation that often occur in microbiome count data. Hurdle model also automatically deals with the issue of dropout events. Based on mixtures of Poisson hurdle models, we developed clustering methods to group microbial features sharing similar patterns of change across different treatments/conditions.

Section 2 presents our method. We describe Poisson hurdle models for microbiome count data in Section 2.1 and propose our clustering algorithm based on mixture of Poisson hurdle models, including the expectation-maximization (EM) algorithm in Section 2.2, a stochastic modified EM algorithm in Section 2.3, an initialization method based on Kendall's τ correlation in Section 2.4, and a hierarchical merging algorithm for determining number of clusters in Section 2.5. In Section 3, we compare the performance of our algorithms and other methods under a variety of simulation settings. In Section 4, we apply our proposed method to a sorghum microbiome dataset. We conclude this paper with some discussion in Section 5.

2 Poisson hurdle model-based clustering

Model-based clustering methods assume that data are generated by a mixture of probability distributions where each component corresponds to one cluster. Compared to traditional clustering methods such as K-means or hierarchical clustering, model-based clustering automatically offers quantitative measure of the uncertainty of the clustering results, i.e., the probability of each feature belonging to each cluster. Extensive research has been done in model-based clustering with multivariate normal mixture distributions, see Fraley and Raftery (2002) for an excellent review. However, the count data with excessive zeros cannot be modelled directly using normal distributions. To handle the zero-inflated microbiome data, we propose a model-based clustering algorithm based on Poisson hurdle distribution.

2.1 Poisson hurdle distribution

Two types of statistical models have been commonly applied to modeling count data with extra zeros: zero-inflated models and hurdle models (also known as two part models) (Hilbe, 2011). In fact, zero-inflated models are special cases of hurdle models: hurdle models can handle both zero-inflation and zero-deflation. Although many features of microbiome have a lot of zeros, there are also features that are not zero-inflated and should not be modeled by zero-inflated distributions. In addition, estimates based on hurdle models tend to be more computationally stable, especially for data with small amounts of zeros (Xu *et al.*, 2015). Hence, We propose to use Poisson hurdle models for microbiome count data.

Suppose we have a microbiome dataset with G features and I treatment groups. Let $N_{gij}, g = 1, \cdots, G, i = 1, \cdots, I, j = 1, \cdots, n_i$ denote the count data for feature g in replicate j of treatment i. The Poisson hurdle distribution models data by two parts separately: the zero part and the zero-truncated Poisson part. If N_{gij} follows a Poisson hurdle distribution corresponding to a cluster k, then its probability mass function (pmf) is:

$$f(N_{gij}) = \begin{cases} 1 - q_{kij}, N_{gij} = 0 \\ \dfrac{q_{kij}}{1 - \exp(-\lambda_{kgij})} \dfrac{\lambda_{kgij}^{N_{gij}} \exp(-\lambda_{kgij})}{N_{gij}!}, N_{gij} > 0 \end{cases} \tag{1}$$

$$\log(\lambda_{kgij}) = s_{ij} + \alpha_{gk} + \mu_{ki} \tag{2}$$

$$q_{kij} = \frac{1}{1 + \exp[-(\gamma_{0ki} + \gamma_{1ki} s_{ij})]}, \gamma_{1ki} > 0 \tag{3}$$

where q_{kij} is the probability of N_{gij} in cluster k being positive (non-zero), and λ_{kgij} is the mean of the Poisson distribution before zero-truncation.

In expression (2) of the Poisson mean, s_{ij} is a normalization factor that adjusts for technical variations in sequencing depth across samples. In this manuscript, we use the log upper-quartile estimator. This normalization method, originally proposed for RNA-seq analysis (Bullard *et al.*, 2010), has been shown to work well in microbiome datasets (Weiss *et al.*, 2017). Once estimated, s_{ij} is treated as known. The parameter α_{gk} represents the geometric mean abundance level in the Poisson part across all treatments for feature g in cluster k, and μ_{ki} (with $\sum_{i=1}^{I} \mu_{ki} = 0$) represents the i-th treatment effect in abundance level for features in cluster k.

In expression (3), we model q_{kij} as a logistic function of the normalization factor s_{ij} and allow different intercepts and slopes $\gamma_{0ki}, \gamma_{1ki}$ for different combinations of cluster and treatment (k, i). We further constraint $\gamma_{1ki} \geq 0$ because samples with larger sequencing depths tend to have larger non-zero proportions.

Note that in model (1), features in the same cluster have the same treatment effects (μ_{ki}) but we allow different geometric means (α_{gk}) across features in the same cluster. The reason is to cluster treatment effects, i.e., changes in abundance levels across treatments. Alternatively, we can also cluster features according to their abundance levels by assuming a reduced model with both the same geometric mean α_k and the same treatment effects (μ_{ki}) for all features in the same cluster. We present more details about this reduced model in Section 1 of the Supplementary Materials and also have functions to implement it in our R package. The remaining part of the main text deals with the full model with α_{gk}.

Assuming a total of K clusters, we model each cluster by Poisson hurdle models with cluster-specific parameter vectors $\underset{\sim}{\mu}_k$ and $\underset{\sim}{\gamma}_k$, where $\underset{\sim}{\mu}_k = (\mu_{k1}, \mu_{k2}, \cdots, \mu_{kI})$, with $\sum_{i=1}^{I} \mu_{ki} = 0$, models the pattern of changes in abundance level across treatments and $\underset{\sim}{\gamma}_k = (\gamma_{0k1}, \gamma_{1k1}, \cdots, \gamma_{0kI}, \gamma_{1kI})$ is the vector modeling q_{kij} the probabilities of being positive counts for features in cluster k. Based on a mixture of Poisson hurdle models, the likelihood given observations of feature g can be expressed by $L_g = \sum_{k=1}^{K} p_k f(\alpha_{gk}, \underset{\sim}{\mu}_k, \underset{\sim}{\gamma}_k | \underset{\sim}{N}_g)$, where $\underset{\sim}{N}_g$ represents the count vector for g^{th} feature across all samples, f is the probability mass function for Poisson hurdle model (1), and p_k is the mixing proportion corresponding to component k with $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$.

The high dimensionality of microbiome data and complex relationships among microbial features makes it nearly impossible to model the dependency among features. In this manuscript, we assume independence among features but we evaluate the performance of our procedure under more realistic compositional structures among features. With

independence assumption, the total likelihood can be expressed by

$$L = \prod_g L_g = \prod_{g=1}^{G} \sum_{k=1}^{K} p_k f(\alpha_{gk}, \underset{\sim}{\mu}_k, \underset{\sim}{\gamma}_k | \underset{\sim}{N}_g).$$

2.2 Poisson hurdle clustering via the EM algorithm

We apply an Expectation-Maximization (EM) algorithm to obtain parameter estimates and clustering result. The EM algorithm for model-based clustering introduces a latent variable Z_{gk} as the indicator that feature g belongs to cluster k for each combination of g and k. These indicators are treated as missing data, and the conditional expectation $E[Z_{gk}|\underset{\sim}{\theta}]$, where $\underset{\sim}{\theta} = (p_k, \alpha_{gk}, \underset{\sim}{\mu}_k, \underset{\sim}{\gamma}_k)$, gives the conditional probability that feature g belongs to cluster k. EM algorithm proceeds by iteratively calculating conditional expectations (E-step) and updating all unknown parameters by maximizing the likelihood function (M-step) until convergence. Clustering results are obtained based on the final conditional expectations. EM algorithm is a common approach in model-based clustering (Fraley and Raftery, 2002). Si *et al.* (2014) and Rau *et al.* (2011) implemented EM algorithm in Poisson model-based clustering.

Compared to Poisson models, Poisson hurdle models are much more complicated due to the two-part structure. The total log-likelihood for mixture of Poisson hurdle models involve an extremely high dimension of parameters, and there are no closed-form solutions for the maximum likelihood estimator (MLE) of μ_{ki} and α_{gk} in the M-step. This poses extra challenges for the EM algorithm. In this manuscript, we propose to utilize a numerical method based on coordinate descent in the M step.

Algorithm 1: EM algorithm for Poisson hurdle clustering.

1.Initialization ($m = 1$):
Set $p_k^{(1)} = 1/K, k = 1, \cdots, K$.
For each cluster k, obtain the initial values for parameters: $\underset{\sim}{\gamma}_k^{(1)}$, $\underset{\sim}{\mu}_k^{(1)}$ with $\sum_{i=1}^{I} \mu_{ki}^{(1)} = 0$, and $\underset{\sim}{\alpha}_k^{(1)}$, where $\underset{\sim}{\alpha}_k = (\alpha_{1k}, \cdots, \alpha_{Gk})$.
See Section 2.4 and Section 3 of the Supplementary Materials for our proposed initialization method.

2.E-step:
In the m^{th} iteration, calculate the conditional expectation $E[Z_{gk}|\underset{\sim}{\theta}^{(m)}]$, denoted as $\hat{Z}_{gk}^{(m)}$ by:

$$\hat{Z}_{gk}^{(m)} = \frac{p_k^{(m)} f(\underset{\sim}{N}_g | \alpha_{gk}^{(m)}, \underset{\sim}{\mu}_k^{(m)}, \underset{\sim}{\gamma}_k^{(m)})}{\sum_{l=1}^{K} p_l^{(m)} f(\underset{\sim}{N}_g | \alpha_{gl}^{(m)}, \underset{\sim}{\mu}_l^{(m)}, \underset{\sim}{\gamma}_l^{(m)})} \tag{4}$$

3.M-step:
Given $\hat{Z}_{gk}^{(m)}$, the mixing proportion p_k is updated by

$$p_k^{(m+1)} = \frac{\sum_g \hat{Z}_{gk}^{(m)}}{G}$$

Maximizing the likelihood function is equivalent to maximizing the following log-likelihood for each cluster k, with the constraint $\sum_{i=1}^{I} \mu_{ki} = 0$.

$$\begin{aligned} & l_k(\underset{\sim}{\mu}_k, \underset{\sim}{\gamma}_k, \underset{\sim}{\alpha}_k) \\ =& \sum_g \hat{Z}_{gk}^{(m)} * \log f(\underset{\sim}{N}_g | \underset{\sim}{\mu}_k, \underset{\sim}{\gamma}_k, \alpha_{gk}) \\ =& \sum_g \hat{Z}_{gk}^{(m)} * \{ \sum_{i,j \in C_g} \log(1 - q_{kij}) + \sum_{i,j \notin C_g} [\log q_{kij} + \\ & N_{gij} \log \lambda_{gij} - \lambda_{gij} - \log(1 - e^{-\lambda_{gij}})] \} \end{aligned}$$

where $C_g = \{i, j : N_{gij} = 0\}$, $\lambda_{gij} = \exp(s_{ij} + \alpha_{gk} + \mu_{ki})$, $q_{kij} = \frac{1}{1+\exp[-(\gamma_{0ki} + \gamma_{1ki} s_{ij})]}$.

We can maximize over the two set of parameters $\gamma_{0ki}, \gamma_{1ki}$ and α_{gk}, μ_{ki} separately. But still, there are no closed-form solutions for maximum likelihood estimate (MLE) of those parameters. Hence, we propose to use a one-step coordinate descent algorithm to obtain numerical solutions for MLEs, which greatly reduce computation. Please see Section 2 in the Supplementary Materials for more details.

4.Iterate the E-step and M-step until convergence, i.e. when the change in total likelihood is relatively small.

5.Obtain $\hat{Z}_{gk}$ from the last iteration, and assign feature g into cluster k where $k = \underset{l}{\text{argmax}} \hat{Z}_{gl}$.

2.3 Simulated annealing modification

As a strictly ascending algorithm, EM algorithm can be trapped in local maximum. Various methods for adding randomness to help EM algorithm escape from local maximum have been introduced, and simulated annealing (SA) is one of them. The SA algorithm modifies the way to obtain conditional expectation in (3) by introducing a "temperature" $t^{(m)} > 0$ and a "cooling rate" $c \in (0, 1)$ as follows:

$$\begin{aligned} \tilde{Z}_{gk}^{(m)} &= \frac{[p_k^{(m)} f(\underset{\sim}{N}_g | \alpha_{gk}, \underset{\sim}{\mu}_k^{(m)}, \underset{\sim}{\gamma}_k^{(m)})]^{1/t^{(m)}}}{\sum_{l=1}^{K} [p_l^{(m)} f(\underset{\sim}{N}_g | \alpha_{gl}, \underset{\sim}{\mu}_l^{(m)}, \underset{\sim}{\gamma}_l^{(m)})]^{1/t^{(m)}}}, \\ t^{(m+1)} &= c \times t^{(m)}. \end{aligned}$$

Given $\tilde{Z}_{gk}^{(m)}$, the SA algorithm clusters each feature g into class k with multinomial probability $\tilde{Z}_{gk}^{(m)}$ and generates an indicator matrix with entries 0 or 1 that replaces the $\hat{Z}_{gk}$ in the M step of the original EM algorithm. This clustering step of SA introduces more randomness (Celeux and Govaert, 1992) that is controlled by the temperature t, and larger t leads to larger randomness. SA usually starts with a relatively high temperature $t^{(0)}$ and slowly reduces it to 0 as the algorithm proceeds, and the cooling rate c controls the speed of reduction. van Laarhoven and Aarts (1987) recommended $t^{(0)} = 2$ and $c = 0.9$ which is what we use. Simulation results in Section 3.3 show that SA algorithm yields competitive result compared with the original EM algorithm (Algorithm 1).

2.4 Initialization

EM algorithm is an iterative, strictly ascending algorithm whose convergence rate and final results are significantly influenced by the initialization (McLachlan *et al.*, 2008; Melnykov and Melnykov, 2012; Biernacki *et al.*, 2003). Commonly used approaches to initialize the EM algorithm start by picking K observations that are far from each other regarding some distance measure, such as $(1-$ Pearson's correlation), Euclidean distance (Arthur and Vassilvitskii, 2007), ranked Euclidean distance (Melnykov and Melnykov, 2012), and likelihood ratios (Si *et al.*, 2014), etc. In this manuscript, we propose to use $(1 - \tau)$ as the distance measure, where τ is the *Kendall's τ correlation*.

We gave a detailed description about the Kendall's τ correlation and our proposed initialization algorithm in Section 3 of the Supplementary Materials. The main idea is that, we first select K observations that are well separated measured by $(1 - \tau)$, and then we obtain MLEs of the model parameters based on the K selected observations and use these MLEs as the starting values for our EM algorithm.

In practice, we also recommend utilizing multiple sets of starting values to further avoid being trapped in local maximum, i.e. run the entire algorithm multiple times with different starting values and pick the result with the largest likelihood. Increasing the number of starts tends to provide a better performance at the cost of linearly increasing computation time. In the simulation section, we will show how this affect the clustering results.

2.5 Determining number of clusters

In real data analysis, the number of clusters K is typically not available and needs to be selected. There are existing methods developed for this purpose, but they don't fit into the scenario we are dealing with. In this manuscript we propose a hybrid clustering method with the application of likelihood ratio tests to select the number of clusters K. The likelihood ratio tests determine the optimal number of K by testing, at each step of merging clusters, whether the merging will result in a reduced model that no longer fits data well.

Please refer to Section 4 of the Supplementary Materials for the discussion of existing methods, the reason why we don't use them, and our complete algorithm for choosing K.

3 Simulation studies

In this section, we present simulation studies to evaluate the performance of our proposed clustering methods and several existing clustering methods, including Poisson model-based clustering, negative binomial model-based clustering (Si *et al.*, 2014), and the K-means clustering that has been applied to microbiome data (Gloor *et al.*, 2016).

3.1 Simulation settings

We consider an experiment with $I = 3$ treatment groups, $J = 5$ replicates in each treatment group, and a total of $G = 1000$ features that belong to $K = 7$ true clusters. We assume equal mixing proportions of $p_k = 1/7$. A case of unequal mixing proportion is presented in Section 5 of the Supplementary Materials.

Data is simulated by a zero-inflated Negative Binomial model which introduces overdispersion, i.e., the variability is more than what is expected by our Poisson model. Each observation, $N_{gij} \in$ class k, is a product of a Bernoulli(q_{kgij}) random variable and a Negative Binomial random variable with mean $E(N_{gij}) = \exp(s_{ij} + \alpha_g + \mu_{ki})$ and variance $(1 + \beta \exp(\alpha_g))E(N_{gij})$

For the negative binomial random variable:

1. Overdispersion rate is $\beta \exp(\alpha_g)$, which depends on feature abundance level α_g.
2. The geometric mean abundance levels α_g's and sequencing depth factors s_{ij}'s are drawn from a Uniform$(0.8, 1.2)$ distribution.
3. β controls the overdispersion and ranges from 0 to 0.5. This allows the overdispersion rate to range from 0 to $0.5 * e^{1.2} = 1.66$, a reasonably large value for overdispersion.
4. The cluster-specific profile across treatment groups, μ_{ki}, is generated by $\mu_{ki} = \eta_\mu \delta_{ki}$ where η_μ determines the magnitude of changes across treatments, and larger η_μ results in better separation of clusters. $\underset{\sim}{\delta}_k = (\delta_{k1}, \delta_{k2}, \delta_{k3})$ characterizes the treatment effects in cluster k and is generated as follows:

Cluster k	1	2	3	4	5	6	7
δ_{k1}	0	0	1	-1	1	-1	0
δ_{k2}	1	-1	0	0	-1	1	0
δ_{k3}	-1	1	-1	1	0	0	0

Note that the first six clusters cover different abundance profiles across treatments by including all 6 permutations of 3 different treatment effects: positive effect (1), no effect (0), and negative effect (-1). The last cluster corresponds to non-differential features whose abundance levels don't change across treatments. Such microbes are typically not of interest, but exist in real data and affect the clustering performances.

For the Bernoulli(q_{kgij}) random variable that controls zero-inflation, we generate it as $q_{gkij} = \frac{1}{1+\exp[-(\gamma_{0ki}+\gamma_{1ki}s_{ij}+\gamma_{2ki}\alpha_g)]}$. For each combination of cluster k and treatment i, we independently draw $\gamma_{1ki}, \gamma_{2ki} \sim$ Uniform$(0, 0.5)$ and set γ_{0ki} such that the average zero-inflation rate $\bar{q}_{ki}$ in cluster k treatment i is a specific value in each of the following two scenarios:

1. Set the matrix $\{\bar{q}_{ki}\}_{7\times 3} = \begin{pmatrix} 0.9 & 0.9 & 0.3 & 0.6 & 0.3 & 0.6 & 0.5 \\ 0.6 & 0.3 & 0.9 & 0.9 & 0.6 & 0.3 & 0.5 \\ 0.3 & 0.6 & 0.6 & 0.3 & 0.9 & 0.9 & 0.5 \end{pmatrix}^T$.
2. Set $\bar{q}_{ki} = \phi \in (0, 1]$ for all k and i.

Note that in the first scenario, all six permutations of high, medium and low zero-inflation are present along with a group with equal mean zero-inflation rate $(0.5, 0.5, 0.5)$ across treatment groups. This is a case where zero-inflation structure varies significantly among clusters and is more aligned with our model assumptions. In the second scenario, all clusters share the same mean zero-inflation rate with $\phi = 1$ corresponding to no zero-inflation, and the difference among clusters are only reflected through the Poisson mean structure. This is a less desirable circumstance for our method because zero-inflation rate does not distinguish different clusters. In the main text, we will present simulation results for the second scenario that demonstrate our method performs better than others even in this unfavorable situation. Results for the first scenario are presented in the Supplementary Figure 2.

Finally, after a raw dataset is generated from the above procedure, we do an additional multinomial resampling on each column, with total counts $C * G = 1000C$ and probability vector proportional to each column of this raw dataset. This is to mimic the sequencing procedure and generate a compositional dataset with equal column sums. It brings in extra randomness and deviation from our assumed models, thus can test the robustness of clustering methods.

We set the default simulation setting with $\beta = 0.02, \eta_\mu = 1, \phi = 0.4$, and average sequencing depth (total count/G) $C = 10$. By varying each parameter at a time, we generate data for a variety of different simulation settings. For each simulation setting, 1000 datasets are simulated.

3.2 Simulation results

For each simulated dataset, we cluster the 1000 features into 7 clusters using five different methods under comparison: (i) Poisson hurdle model-based clustering with EM algorithm (PH-EM), (ii) Poisson hurdle model-based clustering with simulated annealing (PH-SA), (iii) Poisson model-based clustering (MB-Poisson), (iv) negative binomial model-based clustering (MB-NB), and (v) K-means clustering with Euclidean distance (other popular non-model-based methods such as spectral clustering and hierarchical clustering produce similar or worse results and are not presented). The first four model-based methods are applied to the count data, while K-means is applied to the data after centered log ratio (clr) transformation (Aitchison, 1982), as was done by Gloor *et al.* (2016).

The clustering results are evaluated by three criteria: purity, adjusted Rand index (ARI), and normalized mutual information (NMI). All three criteria measure the agreement between clustering results and true clusters used to generate data, within the range of [0,1] with higher values indicating better performance. Purity measures how "pure" the clusters are, i.e. to what extent each resulting cluster contains a single true cluster. Adjusted Rand index (Rand, 1971; Hubert and Arabie, 1985) measures similarity between two partitions (clustering results and true clusters) based on the proportion of pairs of features that are "correctly" assigned. Mutual information (MI) measures the shared information between two partitions. The normalized MI (Strehl and Ghosh, 2002) adjusts the values so that NMI is in $[0, 1]$. See Section 6 in the Supplementary Materials for definitions with mathematical expressions of all three criteria.

The results from all three criteria are consistent and show the same relative ranking of methods. In the main text, we present results based on NMI. Results on purity and ARI are presented in Supplementary Figure 1. We also evaluated the efficiency of the EM algorithm by checking the

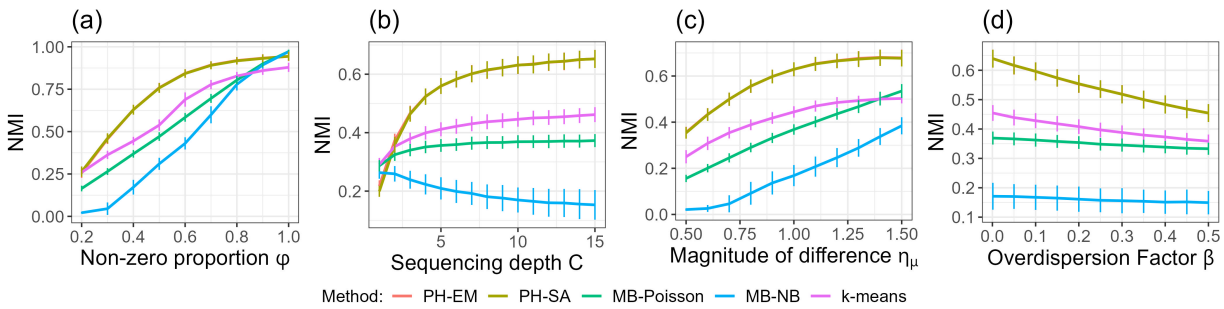

Fig. 1: Simulation results for the second simulation scenario. For each setting, we plot the NMI score averaged across the 1000 simulated datasets with the vertical bar representing standard error. The default setting is $\phi = 0.4, C = 5, \eta_\mu = 1, \beta = 0.02$. Each of (a)-(d) varies one parameter at a time. The curves for $PH - EM$ and $PH - SA$ almost overlap with each other.

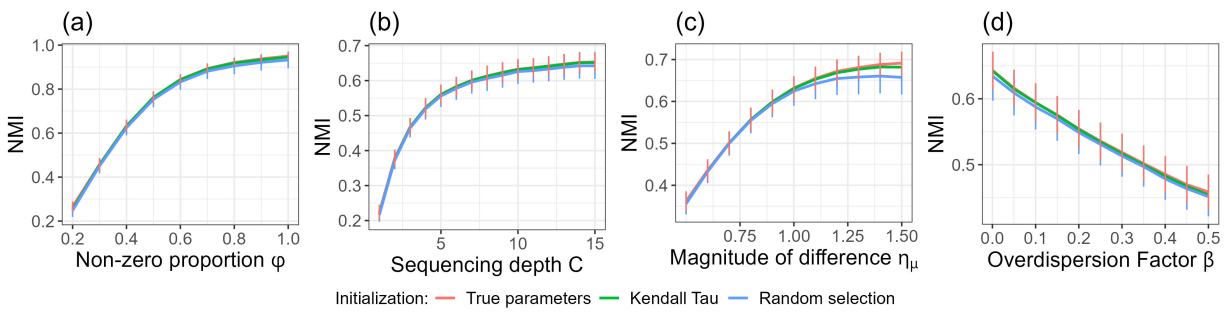

Fig. 2: Comparison of initialization methods. For each setting, we plot the NMI score averaged across the 1000 simulated datasets, with the vertical bar representing standard deviation.

computational time and the number of iterations for convergence, which is described in Section 7 of the Supplementary Materials.

3.2.1 Clustering results

Figure 1 presents results for a variety of settings of the second simulation scenario when the zero-inflation rates are constant across treatment groups. For all simulation settings, our proposed algorithms, PH-EM and PH-SA are almost indistinguishable from each other and they are the best performers among all methods under comparison. Figure 1(a) shows that both PH-EM and PH-SA perform much better than the other methods when zero-inflation rate is between 0.2 and 0.8, a range commonly encountered in real data. When there is no zero-inflation at all ($\phi = 1$), our algorithms still performs similarly to the other two model-based clustering methods that do not model zero-inflation and better than K-means. As sequencing depth grows (Figure 1b), the magnitude of treatment differences increases (Figure 1c), or the level of fluctuation decreases (Figure 1d), most methods tend to perform better. When sequencing depth is low, our methods are among the top-performing methods. For all other settings (Figures 1(b-d)), our methods are the the best with obvious advantage over the other methods . These results show the consistency and robustness of the Poisson hurdle clustering algorithms. Results from the first simulation scenario (different zero-inflation structure among clusters) presented in Supplementary Figure 2 give the same conclusion.

3.2.2 Evaluation of initialization

We also compare three initialization methods: (i) using true parameters, (ii) using the MLEs based on K randomly selected observations, and (iii) using the Kendall's τ based initialization algorithm we propose in Section 2.4. As shown by the NMI results in Figure 2, our proposed initialization method uniformly outperforms random selection, and is close to the result using true parameters which is the best we can get in simulation but not available in real data analysis. Results of purity and ARI as performance measure are in Supplementary Figure 3 and give the same conclusion.

In Supplementary Figure 4 we present the results of using different number of starts when we utilize multiple starting strategy. When we use 5 starting values, we get significant improvement in performance, while increasing to 10 starting values doesn't seem to have additional notable effect. Therefore we choose 5 starts in all simulation studies.

3.2.3 Determining the number of clusters

In Section 2.5 and Section 4 of the Supplementary Materials, we describe a method for determining the number of clusters. Here, we evaluate this method using 2000 datasets generated from the default simulation setting where the true number of clusters is $K = 7$. Table 1 shows the proportions of those datasets being identified with certain number clusters using 4 different methods: our proposed Hybrid method; AIC; BIC; and Gap statistic (Tibshirani *et al.*, 2001). We can see that while those three methods tend to greatly underestimate the number of clusters, results from our hybrid method remain close to the true value $K = 7$.

# of clusters (%)	1	2	3	4	5	6	7	8
Hybrid	0	0	0	0	0.60	11.45	86.75	1.20
AIC	0	0	94.60	4.30	1.05	0.05	0	0
BIC	0	0.7	98.95	0.35	0	0	0	0
Gap	91.65	3.75	3.85	0.40	0.20	0.15	0	0

Table 1. The percentage of 2000 simulated datasets for each number of clusters chosen by 4 different methods. The true number of clusters is 7.

4 Real data analysis

A microbiome study was carried out in Nebraska where sorghum plants of the genotype Grassl were grown with two varying nitrogen levels (Low/High). Rhizosphere microbiome samples were collected on four different dates throughout the growing season between June and September of 2017 and analyzed by 16S rRNA amplicon sequencing (Qi *et al.*,

2021). Applying the persistence method (Shade and Handelsman, 2012) to identify core ASVs among the eight treatments, we obtained a subset of 449 ASVs with average read count of 44.21, and 38% of counts in this subset are zero. Our proposed method for determining the number of clusters chose $K = 30$.

We applied all four model-based clustering algorithms to the original count data, and also applied the K-means clustering to the centered log-ratio (clr)-transformed data to group the 449 ASVs into 30 clusters. Figure 3 plots the abundance profiles of all 30 clusters identified by our PH-EM method (Figure 3a) and model-based NB clustering method (Figure 3b), respectively. The Poisson model-based clustering produced worse results than the NB model-based method and is not shown here. Results for the PH-SA method and K-means are presented in Supplementary Figure 5 and yield similar conclusions.

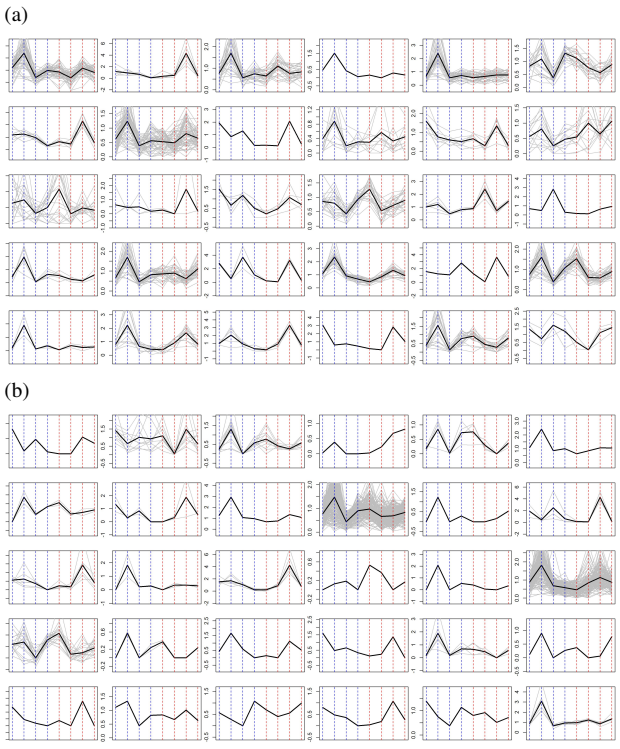

Fig. 3: Results from sorghum data analysis based on (a) PH-EM algorithm and (b) model-based NB clustering. Each subplot corresponds to one cluster, with x-axis corresponding to the 8 treatment groups (2 nitrogen levels with 4 chronically ordered dates, 1∼4 correspond to high nitrogen, 5∼8 correspond to low nitrogen) and y-axis corresponding to the abundance profile. Each grey line corresponds to the abundance profile estimated by the method of moments for an ASV, and the black line plots the geometric mean within each cluster. For each method, clusters will be referred to as cluster No. 1-6 for the first row, 7-12 for the second row, $\cdots$, and 25-30 for the last row.

As shown in Figure 3(a), the PH-EM method resulted in better separation of different clusters and more similar profiles within each cluster. In contrast, in Figure 3(b) model-based NB clustering clusters over 80% of ASVs into 2 huge clusters, and 28 small clusters of which 15 are singletons . To parse finer-scale relationships between taxa in those two huge clusters, further clustering would be needed, complicating interpretation and placing additional bioinformatic burden on the researcher. This is not the case with our method, making it more useful for researchers looking to gain exploratory insight into the biological or ecological structure of microbial communities.

Our method revealed several bacterial clusters whose abundance was sensitive to plant developmental stage and nitrogen concentration. For example, PH-EM clustering revealed several clusters consisting of plant-growth promoting taxa (e.g., *Pseudomonas*, *Sphingomonas*, *Rhizobium*, *Arthrobacter*, and *Streptomyces*) whose abundance remained relatively stable under high nitrogen, but increased dramatically under low nitrogen (Figure 3(a), clusters 2, 7, 9, 11, 14, and 27). These patterns suggest that under low nitrogen a putative guild of nitrogen-fixing taxa is selected for, at least partly driven by root exudation later in sorghum development. These results were corroborated by our analysis of amplicon sequences from 2016, as well as metagenomes assembled from both 2016 and 2017 samples that showed a significant increase in nitrogen-fixing genes under low nitrogen (Supplementary Figure 6). Similar patterns have been identified in other studies of the sorghum rhizosphere (Yu *et al.* 2011, Hara *et al.* 2019, Lopes *et al.* 2021, Wu *et al.* 2021). Interestingly, this nitrogen-fixing guild seems to be preceded by other putatively plant-growth promoting taxa, such as *Massilia* and *Bacillus* (clusters 12 and 17), whose abundances were also higher under low nitrogen. Importantly, these patterns were masked in the model-based NB clusters 10 and 18 whose average trends suggest the opposite: low abundance under low nitrogen and high abundance under high nitrogen. Based on these findings, we believe our clustering method could be particularly useful for parsing dynamic ecological relationships from datasets consisting of times-series and/or many treatments.

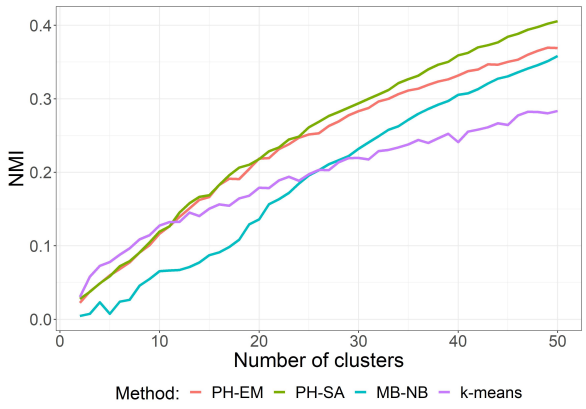

Fig. 4: Clustering results for the rhizosphere microbiome dataset. Clustering results obtained by PH-EM, PH-SA, model-based negative binomial clustering (MB-NB), and K-means are compared with genera categories for each number of clusters K, ranging from 2 to 50. The NMI values shown are averages from 100 clustering results at each K.

To provide a quantitative evaluation of the clustering results, we measured the concordance between clustering results and taxonomic categories at the genus level. With the number of clusters, K, ranging from 2 to 50, we performed cluster analysis with different methods and calculated NMI based on the concordance between each clustering result and genera. Figure 4 shows that both PH-EM and PH-SA produced higher NMI values than K-means and model-based NB clustering for K larger than 7. When K is small (2 to 7), model-based NB method gave slightly higher NMI values than our algorithms. Thus, our method outperformed other clustering methods when it came to clustering microbes based on microbial taxonomy, a proxy for ecological similarity. It is important to acknowledge that we did not expect our model to find perfect agreement between taxonomy and abundance (i.e., NMI values

near 1). This is because microbiomes represent complex communities that can display large variation across individuals which cannot be explained by deterministic assembly mechanisms alone. Absence of perfect agreement between taxonomy and abundance could represent an alternative perspective: the combined influence of deterministic and stochastic assembly factors characteristic of competitive lottery models or priority effects (Sale, 1979). In situations where a particular niche space can support only a single species, stochastic dispersal followed by high competition among related species can result in only one "winning" species—the identity of which can vary independently of any niche effect (Peay *et al.* 2012, Lee *et al.* 2013, Verster and Borenstein 2018). This could explain why ASVs of a particular genus group into more than one cluster.

For example, we found that *Pseudomonas sp.* from the sorghum rhizosphere grouped into 12 distinct clusters as compared to just 4 clusters with the model-based NB method. Our amplicon datasets from 2015 suggested that the sorghum rhizosphere exhibits a significant reduction in community diversity due to a "bloom" of a diverse collection of *Pseudomonas* ASVs. Further analysis of isolate genomes and rhizosphere metagenomes from these samples revealed that this community represents several distinct *Pseudomonas* lineages that vary in their functional capacity especially regarding their commensal or pathogenic relationship with the host plant (Chiniquy *et al.*, 2021). This diversity resulted in subtle, but distinct abundance patterns between lineages. As mentioned above, our PH-EM method captured these abundance patterns in our 2017 dataset, while the model-based NB method failed to identify this heterogeneity, instead clustering most of these *Pseudomonas* sp. into a single large cluster, 18. Further, the PH-EM method confirmed that the abundances of many (although not all) of these *Pseudomonas* sp. are significantly higher under low nitrogen as compared to high nitrogen (Figure 3(a), clusters 2, 7, 14, 23, and 28). These patterns are undiscernible from the model-based NB results which clustered Pseudomonas ASVs with contrasting high and low nitrogen abundance patterns into the same cluster (Figure 3(b), cluster 18). Thus, in addition to identifying niche effects in microbial communities, our clustering method may also help researchers hone in on more complex ecological relationships that can be isolated, tested, and confirmed via manipulative studies.

5 Discussion

Many features of microbiome data are zero-inflated while others do not exhibit zero-inflation. Traditional clustering methods do not consider such characteristic. In this manuscript, we model microbiome data with Poisson hurdle distributions that can fit features with or without zero-inflation. To cluster features, we fit a mixture Poisson hurdle model with an EM algorithm (PH-EM) or another algorithm with a SA modification (PH-SA). Both algorithms have superior performances over other algorithms in both simulation studies and real data analysis. We also propose an initialization method and a method for determining optimal number of clusters, which are shown to be effective in our simulation studies. Compared with non-model-based clustering algorithms such as K-means method, our clustering algorithms also provide the uncertainties of the clustering results.

We also considered further extending our algorithms to Negative Binomial hurdle distributions to accommodate potential over-dispersion that may exist in real dataset. We decided not to do so, due to the inefficiency in estimating the overdispersion parameter. Detailed explanations and simulation results are provided in Section 8 of the Supplementary Materials.

6 Acknowledgements

This research was partially supported by the Office of Science (BER) US Department of Energy (DE-SC0014395), the Iowa State University Plant Sciences Institute Scholars Program, the Laurence H. Baker Center of Iowa State University, and the Nonclinical Biostatistics Scholarship from the Biopharmaceutical Section of the American Statistical Association.

The work (proposal: 10.46936/10.25585/60001066) conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

The authors would like to thank three anonymous reviewers, whose feedback greatly improved this work.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 139–160.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. volume 8, pages 1027–1035.

Badri, M., Kurtz, Z. D., Bonneau, R., and Müller, C. L. (2020). Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genomics and Bioinformatics*, **2**(4). lqaa100.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, **41**(3), 561 – 575. Recent Developments in Mixture Model.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**(1), 94.

Casero, D., Gill, K., Sridharan, V., Koturbash, I., Nelson, G., Hauer-Jensen, M., Boerma, M., Braun, J., and Cheema, A. K. (2017). Space-type radiation induces multimodal responses in the mouse gut microbiome and metabolome. *Microbiome*, **5**(1), 105.

Celeux, G. and Govaert, G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, **14**(3), 315 – 332.

Chiniquy, D., Barnes, E. M., Zhou, J., Hartman, K., Li, X., Sheflin, A., Pella, A., Marsh, E., Prenni, J., Deutschbauer, A. M., Schachtman, D. P., and Tringe, S. G. (2021). Microbial community field surveys reveal abundant pseudomonas population in sorghum rhizosphere composed of many closely related phylotypes. *Frontiers in Microbiology*, **12**, 349.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**(5), 829–844.

Cullen, C. M., Aneja, K. K., Beyhan, S., Cho, C. E., Woloszynek, S., Convertino, M., McCoy, S. J., Zhang, Y., Anderson, M. Z., Alvarez-Ponce, D., Smirnova, E., Karstens, L., Dorrestein, P. C., Li, H., Sen Gupta, A., Cheung, K., Powers, J. G., Zhao, Z., and Rosen, G. L. (2020). Emerging priorities for microbiome research. *Frontiers in Microbiology*, **11**, 136.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**(458), 611–631.

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016). It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, **26**(5), 322 – 329. The Microbiome and Epidemiology.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, **8**, 2224.

Hara, S., Morikawa, T., Wasai, S., Kasahara, Y., Koshiba, T., Yamazaki, K., Fujiwara, T., Tokunaga, T., and Minamisawa, K. (2019). Identification of nitrogen-fixing bradyrhizobium associated with roots of field-grown sorghum by metagenome and proteome analyses. *Frontiers in Microbiology*, **10**.

Hilbe, J. (2011). *Hilbe, Joseph M (2011), Negative Binomial Regression, second edition, Cambridge University Press*.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Lee, S. M., Donaldson, G. P., Mikulski, Z., Boyajian, S., Ley, K., and Mazmanian, S. K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*, **501**(7467), 426–429.

Lonèar-Turukalo, T., Lazić, I., Maljković, N., and Brdar, S. (2019). Clustering of microbiome data: Evaluation of ensemble design approaches. In *IEEE EUROCON 2019 -18th International Conference on Smart Technologies*, pages 1–6.

Lopes, L. D., Chai, Y. N., Marsh, E. L., Rajewski, J. F., Dweikat, I., and Schachtman, D. P. (2021). Sweet sorghum genotypes tolerant and sensitive to nitrogen stress select distinct root endosphere and rhizosphere bacterial communities. *Microorganisms*, **9**(6).

McLachlan, G., McLachlan, G., and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLOS Computational Biology*, **10**(4), 1–12.

Melnykov, V. and Melnykov, I. (2012). Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, **56**(6), 1381 – 1395.

Peay, K. G., Belisle, M., and Fukami, T. (2012). Phylogenetic relatedness predicts priority effects in nectar yeast communities. *Proceedings. Biological sciences*, **279**, 749–58.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., and Konstantinidis, K. T. (2014). Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PLOS ONE*, **9**(4), 1–12.

Qi, M., Berry, J. C., Veley, K., O'Connor, L., Finkel, O. M., Salas-González, I., Kuhs, M., Jupe, J., Holcomb, E., del Rio, T. G., Creech, C., Liu, P., Tringe, S., Dangl, J. L., Schachtman, D., and Bart, R. S. (2021). Identification of beneficial and detrimental bacteria that impact sorghum responses to drought using multi-scale and multi-system microbiome comparisons. *bioRxiv*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Rau, A., Celeux, G., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2011). Clustering high-throughput sequencing data with poisson mixture models.

Sale, P. F. (1979). Recruitment, loss and coexistence in a guild of territorial coral reef fishes. *Oecologia*, **42**, 159–177.

Shade, A. and Handelsman, J. (2012). Beyond the venn diagram: the hunt for a core microbiome. *Environmental Microbiology*, **14**(1), 4–12.

Si, Y., Liu, P., Li, P., and Brutnell, T. P. (2014). Model-based clustering for RNA-seq data. *Bioinformatics*, **30**(2), 197–205.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, **3**, 583–617.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 411–423.

van Laarhoven, P. J. M. and Aarts, E. H. L. (1987). *Simulated annealing*, pages 7–15. Springer Netherlands, Dordrecht.

Verster, A. J. and Borenstein, E. (2018). Competitive lottery-based assembly of selected clades in the human gut microbiome. *Microbiome*, **6**(1), 186.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**(1), 27.

Wu, A.-L., Jiao, X.-Y., Wang, J.-S., Dong, E.-W., Guo, J., Wang, L.-G., Sun, A.-Q., and Hu, H.-W. (2021). Sorghum rhizosphere effects reduced soil bacterial diversity by recruiting specific bacterial species under low nitrogen stress. *Science of The Total Environment*, **770**, 144742.

Xu, L., Paterson, A. D., Turpin, W., and Xu, W. (2015). Assessment and selection of competing models for zero-inflated microbiome data. *PLOS ONE*, **10**(7), 1–30.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data . *Bioinformatics*, **17**(10), 977–987.

Yu, H., Yuan, M., Lu, W., Yang, J., Dai, S., Li, Q., Yang, Z., Dong, J., Sun, L., Deng, Z., Zhang, W., Chen, M., Ping, S., Han, Y., Zhan, Y., Yan, Y., Jin, Q., and Lin, M. (2011). Complete genome sequence of the nitrogen-fixing and rhizosphere-associated bacterium pseudomonas stutzeri strain dsm4166. *Journal of Bacteriology*, **193**(13), 3422–3423.

Zhang, Y., Hu, X., and Jiang, X. (2017). Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **14**(2), 264–271.