An integrative framework of stochastic variational variable selection for joint analysis of multi-omics microbiome data

Tung Dang[1], Yushiro Fuji[2], Kie Kumaishi[3], Erika Usui[3], Shungo Kobori[3], Takumi Sato[3], Yusuke Toda[1], Kengo Sakurai[1], Yuji Yamasaki[4], Hisashi Tsujimoto[4], Masami Yokota Hirai[2], Yasunori Ichihashi[3] and Hiroyoshi Iwata[1*]

[1]Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo, Japan.
[2]RIKEN Center for Sustainable Resource Science, RIKEN, Tsurumi-ku, Yokohama, Japan.
[3]RIKEN BioResource Research Center, RIKEN, Tsukuba, Ibaraki, Japan.
[4]Arid Land Research Center, Tottori University, Tottori, Japan.

*Corresponding author(s). E-mail(s): hiroiwata@g.ecc.u-tokyo.ac.jp;

Abstract

High-dimensional multi-omics microbiome data plays an important role in elucidating microbial communities' interactions with their hosts and environment in critical diseases and ecological changes. Although Bayesian clustering methods have recently been used for the integrated analysis of multi-omics data, no method designed to analyze multi-omics microbiome data has been proposed. In this study, we propose a novel framework called integrative stochastic variational variable selection (I-SVVS), which is an extension of stochastic variational variable selection for high-dimensional microbiome data. The I-SVVS approach addresses a specific Bayesian mixture model for each type of omics data, such as an infinite Dirichlet multinomial mixture model for microbiome data and an infinite Gaussian mixture model for metabolomic data. This approach is expected to reduce the computational time of the clustering process and improve the accuracy of the clustering results. This method can also

identify a critical set of representative variables in multi-omics microbiome data. Three datasets from soybean, mice, and humans (each set integrated microbiome and metabolome) were used to demonstrate the potential of I-SVVS. The results suggest that I-SVVS achieved better accuracy and significantly faster computation than the existing methods in all cases of testing datasets and was able to identify the important microbiome species and metabolites that characterized a cluster.

Keywords: integrative analysis, stochastic variational inference; Bayesian infinite mixture model; variable selection; drought irrigation; environmental and human microbiome and metabolome

1 Introduction

Owing to the substantial development of high-throughput technologies, high-dimensional omics data have been generated in various areas, such as agriculture and medicine. For example, in agricultural crops, multi-omics datasets, including soil metabolites, minerals, and microbes, provide an opportunity to jointly analyze datasets to elucidate the network structure of an agroecosystem [1–3]. In medicine, the joint analysis of a multi-omics microbiome dataset plays an important role in investigating the influence of host genes and microbiome associations [4–10] or host metabolism and microbiome associations [11–14] on human health and diseases.

The typical challenges in the development of computational methods for conducting joint analysis of multi-omics datasets are the problems of high dimensionality, sparsity, and multicollinearity to identify biologically meaningful associations among a large number of heterogeneous biological variables in different types of omics. Recently, several integrative approaches have been developed for the joint analysis of multi-omics datasets. For example, Bayesian Consensus Clustering (BCC) [15] is a Bayesian approach that simultaneously estimates clustering specific to each omics dataset and consensus clustering by integrating all datasets. BCC introduces parameters that adjust the differences in cluster assignments between datasets, allowing for heterogeneity across different omics datasets. A consensus clustering solution was then estimated by integrating the posterior distributions of the latent cluster variables across all the datasets. However, BCC is not suitable for very large and complex datasets, because the computational complexity and memory requirements can become prohibitively high. Furthermore, the important assumption of BCC that the observed omics variables follow normal distributions limits the applicability of this method. As a more recent example, Clusternomics [16] is a probabilistic framework with hierarchical Dirichlet mixture models that rely on the existence of a consistent clustering structure across heterogeneous datasets. A context-specific cluster was developed for specific omics data to describe particular aspects of biological processes. Global clustering results from a combination of context-specific cluster assignments. Because of

the two-level cluster assignment, the number of clusters in the local or global structures can be flexibly changed. However, the biological interpretability of Clusternomics is limited because many features in all omics datasets are involved in the analysis processes. The computational burden of the Clusternomics is also prohibitive. As another example, the iCluster algorithm [17] is a dimensionality reduction approach that uses a Gaussian latent variable model to infer clusters. The LASSO penalty was proposed to identify genomic variables that play important roles in the latent process. The iCluster permits the joint modeling of discrete and continuous variables. However, iCluster assumes that different omics datasets are generated from the same biological samples, which may not always be the case. Additionally, the selection of the number of clusters and penalty parameters can significantly influence the clustering solution and biological interpretability.

Currently, many accessible multi-omics microbiome datasets are becoming available [18], and it will become increasingly common to study the microbiome in relation to other omics, such as the expression of host genes and host metabolism. By integrating microbiome data (16S rRNA sequencing) with multiple sources of omics data, we may be able to elucidate the nature of host-microbe interactions, which may ultimately lead to novel discoveries. However, most current approaches are designed for the multi-omics datasets that include mRNA, microRNAs, DNA methylation, and proteomics, such as The Cancer Genome Atlas (TCGA) [19]. The major challenges associated with microbiome datasets have not yet been fully addressed in the development of integrative frameworks. For example, given the complex nature of metagenomic data, the current BCC and Clusternomics approaches cannot cluster communities into groups with similar compositions. In contrast, the Dirichlet multinomial mixture (DMM) [20, 21] is a successful method for probabilistic modeling of microbial metagenomics data. In a previous study [21], we proposed an improved framework of the DMM model called stochastic variational variable selection (SVVS). SVVS was used to predict the number of clusters and quickly identify the core set of important microbial species that contribute to the variation in different community compositions. However, combining the DMM with other mixture models for the joint analysis of a multi-omics microbiome dataset is currently one of the most challenging questions in computational biology. Each Bayesian mixture model has its own set of parameters, such as the number of clusters and prior probabilities over the cluster parameters, which can vary widely depending on the omics data. Another key challenge of Bayesian methods is that the number of biological variables becomes very large when microbial metagenomics is integrated with other omics data such as metabolism or host genomics data. Therefore, identifying the small number of representative variables that significantly contribute to the joint analysis of multi-omics microbiome datasets is crucial. Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling, which are used in BCC and Clusternomics, however, are difficult to use, given the dimensionality of the microbiome multi-omics dataset.

To address these challenges, we propose a significantly enhanced SVVS method, called integrative stochastic variational variable selection (I-SVVS). I-SVVS identifies heterogeneous patterns of sample-to-sample variability by integrating different types of datasets: count (microbiome) and continuous (metabolome). A key aspect of this method is the use of the hierarchical Dirichlet process (HDP) approach to [22] model the relationship between clusters across microbiome and metabolome datasets. This is achieved by introducing a set of shared clusters that are present in all data types. These shared clusters were modeled using the global structure of HDP. In the local structure, HDP defines a separate Dirichlet process for each cluster. This allows the distribution of data within each cluster to be modeled separately from the distribution of data across clusters. HDP has been successfully used in a wide variety of applications to analyze large datasets related to population genetics [23], protein homology detection [24, 25] and single-cell data clustering [26–28]. For the microbiome, I-SVVS uses a modeling strategy similar to that used in our previous study, that is, SVVS. For the metabolome, I-SVVS employs an infinite Gaussian mixture model (GMM) that uses a stick-breaking representation to treat the total number of clusters as a free parameter. An indicator variable was integrated into the framework of the infinite GMM to select significant metabolomic features. I-SVVS can be used for disparate analysis tasks, including joint clustering, data integration, and the identification of significant features in multi-omics data. To highlight this functionality, we applied the I-SVVS method to an integrative dataset of the microbiome and metabolome collected in our soybean experiments as well as public datasets of mice [29] and human gut disease [30].

2 Results

2.1 The I-SVVS approach

I-SVVS uses the hierarchical Dirichlet mixture process to integrate metabolite profile data and microbiome species (or taxonomic unit) data from multi-omics experiments as a composition of biological and technical sources of variation. I-SVVS can identify important features in the metabolite profiles and microbiome species data for each cluster. The input to I-SVVS consists of matrices of metabolite profiles, which are continuous variables, and microbiome species, which are count variables (Fig. 1a). To demonstrate the versatility of I-SVVS, we applied it to data from plant multi-omics experiments as well as published mouse and human data [29, 30].

The output of the I-SVVS approach consists of two main components (Fig. 1b). The first component clustered all samples by integrating the metabolite profile and microbiome species data. The second component provides an important core set of microbiome species (or taxonomic units) and metabolite profiles that significantly contribute to the clustering process. Due to the integration of metabolite profiles and microbiome data, shared information influenced the variable selection of the two databases.

Because the distributions of latent variables are intractable when computing their posteriors, the I-SVVS approach optimizes the parameters of both of components simultaneously using the stochastic variational inference (SVI) framework [31]. To overcome the extra-dimensional problems of integrating metabolite profiles and microbiome species databases, highly efficient techniques for stochastic optimization were adopted for the computational process. Following variational inference, the I-SVVS results were used for subsequent analysis. For example, integrated clustering results can be used for low-dimensional visualizations such as principal-coordinate analysis (PCoA) and nonmetric multidimensional scaling (NMDS) [32]. The results of the variable selection of microbiome species and metabolites were used for phylogenetic and metabolic network analyses, respectively. In the Methods section, we provide detailed specifications of the model and a further description of the estimated quantities.

2.2 I-SVVS accurately clusters large datasets by integrating multiple data types

The application of I-SVVS to multi-omics microbiome data shows that the global cluster assignment variable helps share information between microbiome species and metabolite profiles to significantly improve the performance of the clustering process. Moreover, the selected features of the two datasets provide excellent interpretations of the obtained clusters for biological exploration.

To illustrate this, we applied I-SVVS to three multi-omics microbiome datasets from humans, mice, and plants, spanning thousands to tens of thousands of features from microbiome species and metabolite profiles. We compared the performances of I-SVVS with the integration of microbiome and metabolomics data with the DMM approach [20, 21] with only microbiome species data, as well as with iClusterPlus [17] which is a general-purpose clustering method commonly applied to multi-omics data. The current version of iClusterPlus has been developed in a framework that allows the integration of categorical, count, and continuous data, and can thus analyze the integration of metabolite profile data and microbiome species data. Moreover, Clusternomics, a probabilistic clustering method [16], was employed to compare the performance of I-SVVS. The datasets were as follows: (1) the soybean root microbiome taxonomic unit data and metabolite profile datasets included 186 and 191 rhizosphere samples, which contained 16,943 taxonomic units and 265 metabolome features from the drought irrigation and control conditions, respectively. We denote this as dataset A. (2) Mouse gut datasets for obstructive sleep apnea (OSA) study included 102 samples for intermittent hypoxia and hypercapnia (IHH) (treatment group) and 102 samples for air (control group), which contained 4,690 taxonomic units and 1,710 metabolome features [29]. We denote this as dataset B. (3) The human gut datasets included 89 individuals infected with *Clostridioides difficile* infection (CDI cases), 89 individuals with diarrhea who tested negative for CDI (diarrhea controls), and 155 non-diarrheal controls that contained 3,347 taxonomic units and 103

metabolome features [30]. We denote this dataset as Dataset C. The agreement between the clustering obtained with the three approaches and the ground-truth clustering was measured using the Adjusted Rand Index (ARI) used in our previous study [21]. We followed the Deviance Information Criterion (DIC) for the selection of Bayesian models and default the values of the Clusternomics 0.1.1 packages in R to determine the number of clusters [16]. iClusterPlus uses a deviance ratio metric, which can be interpreted as the percentage of the total variation, to select the number of clusters. We followed the default values of the *tune.iClusterPlus* function in the iClusterPlus 1.32.0 package in R and selected a Gaussian distribution for the metabolic data and a Poisson distribution for the microbiome data [17].

Table 1 presents the computational time required for the calculations of I-SVVS, iClusterPlus, and Clusternomics. We found that the I-SVVS was able to significantly reduce the running times for Datasets A, B, and C. The computational time was found to increase considerably when the number of taxonomic units became very large, such as in dataset A. Moreover, although the number of features of metabolome profile data in dataset B (1,710 features) was substantially larger than that in dataset C (103 features), there was a small difference in the number of taxonomic units between the two datasets. Table 1 shows that the computational time for Dataset C is trivially faster than that for Dataset B. Therefore, the high dimensionality of microbiome data is the most important factor influencing the computational burden of multi-omics analysis. We observed that the scalability of the I-SVVS approach can significantly reduce the computational time required to handle very high-dimensional datasets; for instance, processing a complete set of approximately 17,000 microbiome species and 260 metabolic features in approximately 2 h.

Table 1: Running time of the three approaches on the real data sets.

Dataset	Clusternomics	iClusterPlus	I-SVVS
A	2.35 d	1.12 d	2.18 h
B	13.12 h	6.26 h	18.42 min
C	10.54 h	4.78 h	15.36 min

Note: All algorithms were run on a personal computer (Intel® Xeon® Gold 6230 Processor 2.10 GHz × 2, 40 cores, 2 threads per core, 128 Gb RAM) under Ubuntu 20.04.1 LTS.

Next, we calculated the Adjusted Rand Index (ARI) to evaluate the performance of the three approaches with the three datasets. Table 2 shows that the I-SVVS approach achieved the best performance for datasets A, B, and C (ARI was 0.89, 0.78, and 0.73 for datasets A, B, and C respectively). The ARI value of I-SVVS was highest in Dataset A, which had the largest number of taxonomic units. Because I-SVVS integrates an infinite Dirichlet multinomial mixture, which is a specific model for analyzing microbiome OTU data [21], this approach achieved better performance than other approaches that were

not developed to analyze microbiome data in the process. Table 2 shows that the ARI value of iClusterPlus was highest for Dataset A (ARI = 0.81). In addition, I-SVVS, iClusterPlus, and Clusternomics exhibited poor performance on dataset C (ARI = 0.73, 0.6, and 0.48, respectively). The main reason for this could be that the number of features in the metabolome profile of dataset C was significantly smaller than that of the others. Figure 2 shows the confusion matrix plots for Dataset A, calculated using I-SVVS, iClusterPlus, and Clusternomics. The drought and control groups were clustered using I-SVVS with accuracies of 90% and 88%, respectively. Figures S1a-c show the estimated mixing coefficients of the clusters in datasets A, B, and C after convergence. In some clusters, the estimated mixing coefficients were close to zero after convergence. Therefore, a highly likely number of clusters can be obtained for the mixtures. Figure S1a shows the strongest support for the two clusters in dataset A because the mixing coefficients of clusters 3 and 4 have larger values, while Figure S1b shows the highest probability of the two clusters in dataset B because the mixing coefficients of clusters 1 and 4 have large values; Figure S1c shows the highest probability of 3 clusters in dataset C.

Table 2: ARI scores of the three approaches on the real data sets.

Dataset	Clusternomics	iClusterPlus	I-SVVS
A	0.722	0.815	0.891
B	0.553	0.697	0.781
C	0.482	0.602	0.732

2.3 I-SVVS identifies a core set of features for microbiome species and metabolite profiles

Several previous studies have investigated the important roles of metabolism-microbiome associations in mice [33–35], and plants [36]. To address this trend, the I-SVVS approach was used to identify a core set of microbial species and metabolic features that showed significant differences among the clusters obtained in the analysis. Figures 3a-b show the histograms of the averages of microbiome indicator variable Φ_{1ij} and metabolism indicator variable $\Phi_{2ij'}$ over i^{th} sample in dataset A. Figure 4 shows that the top 100 selected microbial species in dataset A were mapped on the 16S phylogenetic tree. The identification of group-microbiome and group-metabolism associations is based on Φ_{1ij} and $\Phi_{2ij'}$ that i^{th} sample is assigned to a specific cluster (drought and control conditions) in dataset A. As expected, our results are consistent with those of our previous study that analyzed soybean rhizosphere microbiome data. For example, I-SVVS selected important microbiome families that were significantly associated with plant growth promotion under control conditions, such as the *Chitinophagaceae* family within the order *Chitinophagales* from

the phyla *Bacteroidetes*, *Nitrosomonadaceae* and *Chromobacteriaceae* families within the order *Gammaproteobacteria* from the phylum *Proteobacteria*. These are the dominant bacterial phyla in the soybean rhizosphere [37, 38]. Several studies showed [39, 40] the *Nitrosomonadaceae* family oxidize ammonia to nitrite using the enzyme ammonia monooxygenase (AMO), which catalyzes the first step of ammonia oxidation. This family is a key group of nitrifying bacteria that plays a vital role in the conversion of nitrogen compounds in natural ecosystems and agricultural systems [41]. *Chitinophagaceae* family produces enzymes called chitinases that are responsible for the degradation of chitin [42, 43]. The enzymatic breakdown of chitin into smaller components can be further metabolized into a source of energy, carbon, and nitrogen. This capability is important for protecting plants against fungal infections [43]. Moreover, our study showed that the *Microbacteriaceae* family within the order *Micrococcales* and the *Micromonosporaceae* family within the order *Micromonosporales* from the phylum *Actinobacteriota*, *Beijerinckiaceae*, and *Hyphomicrobiaceae* families within the order *Rhizobiales* from the class *Alphaproteobacteria* were more abundant in the drought treatments. Several studies have shown that the *Microbacteriaceae* and *Micromonosporaceae* families can improve the growth of plant hosts through nitrogen fixation, which converts dinitrogen (N_2) into ammonia [44–46]. Therefore, the host plant can utilize these natural sources of nitrogen and reduce its dependence on external sources, such as fertilizers. To assess large-scale metabolite production and consumption patterns, we hierarchically clustered the top individual metabolism and microbiome that were selected by the I-SVVS approach into two groups of dataset A (figure 5). Our analysis revealed significant correlations between metabolic traits and microbiome species across subjects in the control and drought groups such as 1-aminocyclopropane-1-carboxylate (ACC), L-proline, L-tyrosine, L-aspartic acid, and L-glutamic acid. Previous studies have shown that 1-aminocyclopropane-1-carboxylate (ACC), an important intermediate in the ethylene synthesis, can change the soil microbiome to enhance plant tolerance to salinity stress. A group of beneficial bacteria can degrade ACC to ammonia and α-ketobutyrate using ACC deaminase, thereby decreasing the level of ethylene and enhancing plant growth [47, 48]. Several studies have observed the influence of glutamic acid on the plant host and its microbiome, which can restore the population of core microbes to protect plants against diseases [49].

In the mouse fecal samples of obstructive sleep apnea (OSA) in dataset B, Figure S2a-b show the histograms of the averages of the microbiome indicator variable Φ_{1ij} and the metabolism indicator variable $\phi_{2ij'}$ over i^{th} sample in dataset B. Figure S3 shows that the top 100 selected microbial species in dataset B were mapped on the 16S phylogenetic tree. The identification of group-microbiome and group-metabolism associations was based on Φ_{1ij} and $\Phi_{2ij'}$, where the ith sample was assigned to a specific cluster (intermittent hypoxia and hypercapnia (IHH) cases and air controls) in dataset B. I-SVVS selected important microbiome families that were significantly associated with

IHH exposure, such as *Lachnospiraceae* and *Ruminococcaceae* families in the phylum *Firmicutes*. These results were also reported in previous study [29]. Previous studies on sleep fragmentation showed that the growth of highly fermentative members of the *Lachnospiraceae* and *Ruminococcaceae* families lead to visceral white adipose tissue inflammation and alterations in insulin sensitivity [50, 51]. Moreover, I-SVVS identified several important metabolomic features, such as chenodeoxycholic acid and cholic acid, which were significantly associated with IHH exposure. These are primary bile acids that play important roles in facilitating the digestion and absorption of cholesterol and triglycerides. Several studies have reported alterations in response to intermittent hypoxia [52, 53].

3 Discussion

Integrative stochastic variational variable selection (I-SVVS) is a Bayesian nonparametric method that uses an integrated multi-omics dataset of the microbiome to rapidly cluster samples and select important features from multi-omics data. We applied I-SVVS to extreme-dimensional microbiome multi-omics profiles collected in our own soybean experiments and published datasets of mice and humans. I-SVVS is also unique in its ability to integrate the microbiome dataset (such as 16S rRNA) with metabolome, which is not possible with current approaches such as iClusterPlus and Clusternomics.

First, in the soybean dataset (Dataset A), we demonstrated that I-SVVS can achieve accurate clustering based on the integration of the metabolite profile and microbiome data. Owing to the hierarchical Dirichlet mixture models, I-SVVS can capture not only information shared by microbiome and metabolome datasets but also those emerging from the complementarity of these datasets. Our results also demonstrate that I-SVVS can leverage information from multiple omics layers to accurately cluster samples from sparse profiling datasets and avoid the problem of instability of inferred clusters in previous probabilistic algorithms. Most notably, I-SVVS identified an important core set of representative features that vary per sample rather than per cluster from a large number of multi-omics biological features, that is, the microbiome and metabolome. Identification cannot be conducted using the previous Bayesian methods such as BCC and Clusternomics. Our previous method (SVVS) [21] identified the important features (taxonomic units) using only microbiome data. Therefore, I-SVVS-supported selected features play an important role in significantly improving the performance of clustering analysis and interpreting shared information and interactions between different types of omics data. For example, I-SVVS can be used to investigate the relationship between metabolites and the microbiome community structure and function, which plays a crucial role in studies on human health and disease [54, 55]. Moreover, to overcome the computational burden of high-dimensional microbiome multi-omics data, I-SVVS uses stochastic variational inference to estimate a model parameters. We also applied I-SVVS to a joint analysis of

multi-omics microbiome datasets from mice and humans. In these datasets, I-SVVS achieved good performance in identifying the key features that highlighted the impact of disease on host-commensal organism co-metabolism in human and animal guts. Therefore, the flexibility and scalability of I-SVVS make it easily applicable to multiple datasets with larger dimensionality and enable extensions that incorporate additional omic technologies.

Although we proposed several solutions to overcome important challenges in microbiome multi-omics analysis, I-SVVS is not free of limitations. The model focuses on optimizing the contributions of microbiome data, which could significantly improve its performance in the joint analysis of multi-omics datasets. Although the DMM approach is the best mixture model for analyzing microbiome count data, it cannot efficiently model count data of different omics data. Future extensions of I-SVVS may address problems that develop and integrate specific Bayesian mixture models of different omics data, such as metatranscriptome RNA sequencing (MT), and shotgun mass spectrometry-based metaproteomics (MP) [56] in its framework. In addition, variations in the structure of omics data, such as an imbalance in the number of features of each omics dataset, affect the stability and optimal performance of I-SVVS. Future research should also consider this point. Finally, although I-SVVS successfully identified a small number of vital features in different omics datasets, it was difficult to infer the interactions among the selected features. Therefore, there is room for future extensions that are more efficient at enforcing the important relationships across omics than the use of correlations.

4 Methods

4.1 Study inclusion and data acquisition

Dataset A represents the environmental microbiome data from the soybean field experiments, which included 186 drought irrigation samples, 191 control samples, 16,943 microbiome species (or taxonomic units), and 265 metabolome features. The experimental explanations for Dataset A are provided in Appendix A.

We also employed case-control studies from two published gut microbiome and metabolome datasets in mice and humans. Dataset B represents the obstructive sleep apnea (OSA) dataset for mice, which includes 102 samples for intermittent hypoxia and hypercapnia (IHH), 102 samples for control, 4,690 taxonomic units and 1,710 metabolome features [29]. Dataset C represents a study on C. difficile infection (CDI) in humans. The 338 samples included 3,347 taxonomic units and 103 metabolome features [30].

4.2 The integrative framework of infinite mixture models for microbiome multi-omics data

The proposed approach integrates a diverse range of data types, including metabolomics, microbiomics, ionomics, genomics and so on. First, we introduce the following notation: X_{mij} denotes the omics variable associated with the jth ($j \in [1, ..., D_m]$) omics feature in the ith ($i \in [1, ..., N]$) sample of the mth ($m \in [1, ..., M]$) omics data type. For example, an omics feature can be either a metabolite profile or a microbial species, depending on the data type. The binary latent variable $W_{ik} \in (0, 1)$ denotes an indicator variable that assigns samples to the kth ($k \in [1, 2, ...]$) global-level cluster. In the global-level construction, we applied the conventional stick-breaking representation as follows:

$$\Psi_k^{'} \sim \text{Beta}(1, \gamma)$$
$$\Psi_k = \Psi_k^{'} \prod_{k'=1}^{k-1} (1 - \Psi_{k'}^{'})$$

where the random variables Ψ_k denote the stick-breaking weights that satisfy $\sum_{k=1}^{\infty} \Psi_k = 1$ and are generated by breaking a unit-length stick into an infinite number of pieces. The indicator variable $\boldsymbol{W} = (W_{i1}, W_{i2}, ...)$ is distributed according to $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, ...)$ in the following form:

$$p(\boldsymbol{W}) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} \Psi_k^{W_{ik}} = \prod_{i=1}^{N} \prod_{k=1}^{\infty} \left[\Psi_k^{'} \prod_{k'=1}^{k-1} (1 - \Psi_{k'}^{'}) \right]^{W_{ik}}$$

Subsequently, the indicator variable $Z_{mkt} \in (0, 1)$ assigns the kth global cluster to the tth local cluster in the mth omics data type. Similarly, we used the stick-breaking representation to construct each local-level Dirichlet process for specific omics data types as follows:

$$\Pi_{mt}^{'} \sim \text{Beta}(1, \lambda)$$
$$\Pi_{mt} = \Pi_{mt}^{'} \prod_{t'=1}^{t-1} \left(1 - \Pi_{mt'}^{'}\right)$$

where Π_{mt} denotes a set of stick-breaking weights that satisfy $\sum_{t=1}^{\infty} \Pi_{mt} = 1$. The indicator variable $\boldsymbol{Z} = (Z_{mk1}, Z_{mk2}, ...)$ is distributed according to $\boldsymbol{\Pi} = (\Pi_{m1}, \Pi_{m2}, ...)$ in the form

$$p\left(\boldsymbol{Z}\right)=\prod_{m=1}^{M}\prod_{k=1}^{\infty}\prod_{t=1}^{\infty}\Pi_{mt}^{Z_{mkt}}=\prod_{m=1}^{M}\prod_{k=1}^{\infty}\prod_{t=1}^{\infty}\left[\Pi_{mt}^{'}\prod_{t'=1}^{t-1}\left(1-\Pi_{mt'}^{'}\right)\right]^{Z_{mkt}}$$

Moreover, microbiome multi-omics data sets typically include a large number of features. However, in practice, not all omics features are significant, and a large number of them may be irrelevant and negatively influence the performance of the clustering processes. Therefore, a feature selection approach is necessary to select the best omics feature subsets. We propose a binary latent variable Φ_{mij} which represents the feature relevance indicator. Specifically, $\Phi_{mij}=1$ means that the jth feature of the mth omics data type is important, otherwise, the feature X_{mij} is irrelevant.

The corresponding likelihood function of the proposed model for samples $\mathbf{X}=\left(X_{1ij},X_{2ij},...,X_{Mij}\right)$ can be written as

$$p\left(\mathbf{X}\right)=\prod_{m=1}^{M}\prod_{i=1}^{N}\prod_{k=1}^{\infty}\prod_{t=1}^{\infty}\left[\prod_{j=1}^{D}p\left(X_{mij}\mid\Theta_{jt}\right)^{\Phi_{mij}}p\left(X_{mij}\mid\Theta_{j}^{'}\right)^{1-\Phi_{mij}}\right]^{Z_{mkt}W_{ik}} \tag{1}$$

where $p\left(X_{mij}\right)$ is the probability density selected for the specific omis data type. Therefore, the different choices for $p\left(X_{mij}\right)$ allow the modeling of different types of omics data.

Here, we describe our modeling approach for specific omics datasets. In a simple case, there are two types of omics dataset (M = 2). One is microbial metagenomic data and the other is metabolite data. If X_{1ij} denotes the microbial metagenomic variable associated with the jth $(j\in[1,...,D])$ taxonomic units (or species) in the ith $(i\in[1,...,N])$ sample, because microbial data are the count data type, we consider the infinite Dirichlet multinomial mixture (DMM) model in our previous study [21] as follow

$$p\left(\boldsymbol{X_1}\mid\boldsymbol{Z_1},\boldsymbol{\Phi_1},\boldsymbol{\alpha},\boldsymbol{\beta}\right)=\prod_{i=1}^{N}\prod_{t=1}^{\infty}\left[\begin{array}{c}\left(\frac{B\left(\overrightarrow{\alpha_t}+\overrightarrow{X_{1i}}\right)}{B\left(\overrightarrow{\alpha_t}\right)}J_i!\prod_{j=1}^{D}\frac{1}{X_{1ij}!}\right)^{\Phi_{1ij}}\\ \left(\frac{B\left(\boldsymbol{\beta}+\overrightarrow{X_{1i}}\right)}{B\left(\boldsymbol{\beta}\right)}J_i!\prod_{j=1}^{D}\frac{1}{X_{1ij}!}\right)^{1-\Phi_{1ij}}\end{array}\right]^{Z_{1it}} \tag{2}$$

where $Z_{1it}\in[0,1]$ is an allocation variable, such that $Z_{1it}=1$ if $\overrightarrow{X_{1i}}=\left(X_{1i1},...,X_{1iD}\right)$ belongs to the t^{th}cluster and 0, otherwise. Φ_{1ij} is an indicator variable, such that $\Phi_{1ij}=1$ indicates that the jth taxonomic units of ith sample are important in the tth cluster and follow a Dirichlet-multinomial distribution with $\boldsymbol{\alpha}$ parameter and $\Phi_{1ij}=0$ denotes that the jth taxonomic

units of ith sample is unimportant in the tth cluster and follows a Dirichlet-multinomial distribution with $\boldsymbol{\beta}$ parameter. Function B is a multinomial Beta function $B\left(\overrightarrow{\alpha_t}\right) = \frac{\prod_{j=1}^{D}\Gamma(\alpha_{tj})}{\Gamma\left(\sum_{j=1}^{D}\alpha_{tj}\right)}$, $B\left(\overrightarrow{\alpha_t} + \overrightarrow{X_{1i}}\right) = \frac{\prod_{j=1}^{D}\Gamma(\alpha_{tj}+X_{1ij})}{\Gamma\left(\sum_{j=1}^{D}(\alpha_{tj}+X_{1ij})\right)}$, $B\left(\boldsymbol{\beta}\right) = \frac{\prod_{j=1}^{D}\Gamma(\beta_j)}{\Gamma\left(\sum_{j=1}^{D}\beta_j\right)}$ and $B\left(\boldsymbol{\beta} + \overrightarrow{X_{1i}}\right) = \frac{\prod_{j=1}^{D}\Gamma(\beta_j+X_{1ij})}{\Gamma\left(\sum_{j=1}^{D}(\beta_j+X_{1ij})\right)}$. The total number of counts (i.e., sequence reads) from the ith community sample was $J_i = \sum_{j=1}^{D} X_{1ij}$.

If $X_{2ij'}$ is a continuous variable, we assume that it follows a normal distribution, and consider a Dirichlet process mixture model. In this study, the metabolite profile data were continuous variables. $X_{2ij'}$ denotes the metabolite profile variable associated with the $j^{'th}$ $\left(j^{'} \in [1, ..., D^{'}]\right)$ metabolite profile features in the ith sample. The likelihood function of the proposed model is expressed as follows:

$$p\left(\boldsymbol{X_2} \mid \boldsymbol{Z_2}, \boldsymbol{\Phi_2}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\mu}^{'}, \boldsymbol{\delta}^{'}\right) = \prod_{i=1}^{N}\prod_{j'=1}^{D'}\prod_{t'=1}^{\infty}\left[N\left(X_{2ij'} \mid \mu_{t'j'}, \delta_{t'j'}^{-1}\right)^{\Phi_{2ij'}} N\left(X_{2ij'} \mid \mu_{j'}^{'}, \delta_{j'}^{'\,-1}\right)^{1-\Phi_{2ij'}}\right]^{Z_{2it'}} \tag{3}$$

where $Z_{2it'} \in [0,1]$ is an allocation variable such that $Z_{2it'} = 1$ if $\overrightarrow{X_{2i}} = (X_{2i1}, ..., X_{2iD'})$ belongs to the $t^{'th}$cluster and 0, otherwise. $\Phi_{2ij'}$ is an indicator variable, such that $\Phi_{2ij'} = 1$ indicates that the $j^{'th}$ metabolite profile feature of ith sample is important in the $t^{'th}$ cluster and follows a normal distribution with $\boldsymbol{\mu}, \boldsymbol{\delta}$ parameter and $\Phi_{2ij'} = 0$ denotes that the $j^{'th}$ metabolite profile feature of ith sample is unimportant in the $t^{'th}$ cluster and follows a normal distribution with $\boldsymbol{\mu}^{'}, \boldsymbol{\delta}^{'}$ parameter.

We substitute Equations 2 and 3 into Equation 1. If integration of microbiome data with metabolomics data, the likelihood function for the samples $\boldsymbol{X} = \left(X_{1ij}, X_{2ij'}\right)$ can be written as follows:

$$p\left(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\mu}^{'}, \boldsymbol{\delta}^{'}\right) = \prod_{i=1}^{N}\prod_{k=1}^{\infty}\left[\left(\prod_{t=1}^{\infty}\left[\prod_{j=1}^{D} DM\left(X_{1ij} \mid \alpha_{jt}\right)^{\Phi_{1ij}} DM\left(X_{1ij} \mid \beta_j\right)^{1-\Phi_{1ij}}\right]^{Z_{1kt}}\right) \left(\prod_{t'=1}^{\infty}\left[\prod_{j'=1}^{D'} N\left(X_{2ij'} \mid \mu_{j't'}, \delta_{j't'}^{-1}\right)^{\Phi_{2ij'}} N\left(X_{2ij'} \mid \mu_{j'}^{'}, \delta_{j'}^{'\,-1}\right)^{1-\Phi_{2ij'}}\right]^{Z_{2kt'}}\right)\right]^{W_{ik}} \tag{4}$$

where $W_{ik} \in [0,1]$ is an allocation variable of the global cluster, such that $W_{ik} = 1$ if the ith sample of microbiomics and metabolomics data belongs to the kth global cluster and 0, otherwise. $Z_{1kt} \in [0,1]$ is an allocation variable

of the local cluster for microbiome data, such as $Z_{1kt} = 1$ the kth global cluster belongs to the tth local cluster of microbiome data and 0, otherwise. $Z_{2kt'} \in [0, 1]$ is an allocation variable of a local cluster for metabolomics data, such that $Z_{2kt'} = 1$ the kth global cluster belongs to the $t^{'th}$ local cluster of metabolomics data and 0, otherwise. $DM()$ denotes the Dirichlet-multinomial distribution and $N()$ denotes the normal distribution. The prior distributions, that were considered specifically for all variables and parameters, are explained in detail in the Supplementary Material.

4.3 The integrative stochastic variational variable selection (I-SVVS) approach for microbiome multi-omics data

Here, we propose an extension of the SVI approach, which was proposed to estimate the parameters of the infinite DMM model in our previous study, to learn the integrative framework of the proposed model. Given the observed omics datasets $\boldsymbol{X}$, the proposed model has a set of parameters ($\boldsymbol{\Xi}$), an allocation variable of global cluster ($\boldsymbol{W}$) with the unit length of the stick of the stick-breaking representation $\left(\boldsymbol{\Psi}'\right)$, an allocation variable of the local cluster for each specific omics data ($\boldsymbol{Z}$) with the unit length of the stick $\left(\boldsymbol{\Pi}'\right)$, the indicator variable of omics feature selection ($\boldsymbol{\Phi}$) and parameters ($\boldsymbol{\Theta}$) of the distribution $p\left(\boldsymbol{X} \mid \boldsymbol{\Theta}\right)$. For example, in the case of microbiome and metabolomics data, ($\boldsymbol{\Theta}$) includes the parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}$) of the Dirichlet-multinomial distributions and the parameters ($\boldsymbol{\mu}, \boldsymbol{\delta}$) of normal distributions. We then define the variational distribution of the parameters $q\left(\boldsymbol{\Xi}\right)$. In this study, we adopt the factorization assumption of mean-field variational inference, which allows for independence among the variables of the variational distribution $q\left(\boldsymbol{\Xi}\right)$. Furthermore, the proposed model integrated integrates several infinite mixture models by proposing an allocation variable for global cluster ($\boldsymbol{W}$). Thus, to obtain feasible computations, truncated stick-breaking representations are considered for the global cluster at the largest value $\mathrm{K}_{\max}$ and local cluster at the largest value $\mathrm{T}_{\max}$. The truncation levels of the global and local clusters become variational parameters that can be automatically optimized by extending the SVI approach. The variational distribution $q\left(\boldsymbol{\Xi}\right)$ can be specifically factorized into the disjoint tractable distributions as follows:

$$
\begin{aligned}
q\left(\boldsymbol{\Xi}\right) = & \prod_{i=1}^{N} \prod_{k=1}^{K_{max}} q\left(W_{ik}\right) \times \prod_{k=1}^{K_{max}} q\left(\Psi_k^{'}\right) \times \prod_{m=1}^{M} \prod_{k=1}^{K_{max}} \prod_{t=1}^{T_{max}} q\left(Z_{mkt}\right) \\
& \times \prod_{m=1}^{M} \prod_{t=1}^{T_{max}} q\left(\Pi_{mt}^{'}\right) \times \prod_{m=1}^{M} \prod_{i=1}^{N} \prod_{j=1}^{D} q\left(\Phi_{ijm}\right) \times \prod_{j=1}^{D} \prod_{t=1}^{T_{max}} q\left(\Theta_{jt}\right) \times \prod_{j=1}^{D} q\left(\Theta_j^{'}\right)
\end{aligned}
\tag{5}
$$

The distributions of the exponential families were selected for the variational distributions to guarantee a feasible computation of the expectations. The specific considerations for specific omics data, such as microbiome and metabolomics, are explained in detail in the Supplementary Material.

Then, the KL divergence is used to evaluate the distance between the true intractable posterior distributions $p(\boldsymbol{\Xi} \mid \boldsymbol{X})$ and $q(\boldsymbol{\Xi})$. In previous studies, we showed that the fact indices for the computation of KL divergence were difficult. Thus, the variational framework maximizes the Evidence Lower Bound (ELBO), which equals the minimization of KL divergence, to approximate the true posterior distribution $p(\boldsymbol{\Xi} \mid \boldsymbol{X})$. The ELBO function of the proposed method is expressed as follows:

$$\mathcal{L}\left[q\left(\boldsymbol{\Xi}\right)\right] = \mathrm{E_q}\left[\log\left(p\left(\boldsymbol{\Xi}, \boldsymbol{X}\right)\right)\right] - \mathrm{E}_q\left[\log\left(q\left(\boldsymbol{\Xi}\right)\right)\right] \tag{6}$$

In the framework of the SVI approach, it is important to divide the variational parameters into two subgroups: local variables $[\boldsymbol{\Xi_l} \in (\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Phi})]$ and global variables $\left[\boldsymbol{\Xi_g} \in \left(\boldsymbol{\Psi}', \boldsymbol{\Pi}', \boldsymbol{\Theta}, \boldsymbol{\Theta}'\right)\right]$. Expect for the global cluster variable $\boldsymbol{W}$, the numbers of local and global variables depend on the number of omics datasets used for integrated analysis. The variational parameters of local variables were optimized using a coordinate ascent algorithm. One of the most difficult problems is the intractable computation of expectations in equation 6. Depending on the framework of the mixture models specifically considered for each omics dataset, several special expectations cannot be obtained directly from analytically tractable solutions. For example, the DMM showed that the expectations of the logarithms of the multinomial beta function could not be calculated analytically. To overcome these problems, mathematical expansions such as the Taylor expansion and the delta method were adopted in this study. Mathematical explanations of these expansions and variational objective functions are provided in the Supplementary Material.

In particular, the dimensionality of integrated microbiome multi-omics data sets can rapidly increase if the number of different omics datasets increases. For example, the average number of species in the microbiome data is approximately tens of thousands, and metabolomics datasets include thousands of metabolite profile features. Thus, the variational parameters of the global variables were optimized using stochastic algorithms. The stochastic inference is much more computationally efficient because it updates variational factors by sampling the data points in each iteration and uses the natural gradient method, which can achieve faster convergence than standard gradients.

Supplementary information.

- **Supplementary Figure S1**: The estimated values of the mixing coefficients. a. dataset A; b. dataset B; c. dataset C.
- **Supplementary Figure S2**: Histogram of the average of ϕ_{ij} and $\phi'_{ij'}$ in dataset B. The dashed lines are bound to select microbiome species and metabolome profile features. a. microbiome data; b. metabolome data.

- **Supplementary Figure S3**: Microbial species selected using the integrative stochastic variational variable selection (I-SVVS) approach and mapped on the phylogenetic tree for dataset B. Blue-colored starts symbols denote intermittent hypoxia and hypercapnia (IHH) cases and Red-colored plus denote air controls.

Acknowledgments. This work was supported by JSPS KAKENHI (Grant Number JP21J21850), the JST-CRESET Program (Grant Number JPMJCR1602), and the JST-Mirai Program (Grant Number JPMJMI120C7, Japan).

Declarations

- **Ethics approval and consent to participate**
  Not applicable
- **Competing interests**
  The authors declare that they have no competing interests.
- **Consent for publication**
  Not applicable.

- **Funding**
  SPS KAKENHI (Grant Number JP21J21850), the JST-CRESET Program (Grant Number JPMJCR1602), and the JST-Mirai Program (Grant Number JPMJMI120C7, Japan)

Appendix A Soybean rhizosphere microbiome and metabolome

A.1 Field experiment

The accessions and experimental fields were identical to those used by Toda et al. and Sakurai et al [57, 58]. A diverse panel of 198 soybean accessions registered in the National Agriculture and Food Research Organization Genbank (https://www.gene.affrc.go.jp/) was used. These 198 accessions mainly consist of Japanese and global soybean minicore collections [59, 60]. The field trial was conducted in 2019 in an experimental field with sandy soil at the Arid Land Research Center, Tottori University (35°32' N, 134°12' E, 14 m above sea level). Each plot consisted of four plants, and the distances between the two rows, two plots, and two individuals were 50, 80, and 20 cm, respectively. Sowing was performed at the beginning of July, followed by thinning after two weeks. Fertilizers (13, 6.0, 20, 11, and 7.0 g m-2 of N, P, K, Mg, and Ca, respectively) were applied to the field before sowing. White mulch sheets (Dupont, Wilmington, DE, USA) were laid to prevent rainwater infiltration and to control soil conditions with artificial irrigation. Watering tubes were installed under the sheets to irrigate fields. Two watering treatments, non-watered, and well-watered, were used to evaluate the influence of drought and

control conditions. The watering treatment was started after thinning and two weeks after sowing. Artificial irrigation was applied at a flow rate of 1.1L / h m for 5 h daily (7:00–9:00, 12:00–14:00, and 16:00–17:00). Root sampling was conducted at the beginning of September. Root sampling was performed from on the 3rd individual. A root of an individual was collected and stored at -80 degrees Celsius until metabolome and microbiome analysis.

A.2 Metabolome analysis

Metabolome analysis of root samples was performed according to Sawada et al., 2009 [61] and Uchida et al., 2020 [62]. The collected samples were freeze-dried and powdered using a mixer (Shake Master NEO BMS-M10N21, Bio Medical Science, Tokyo, Japan). The powdered sample was weighed to 4 mg, 1 mL of extraction solvent (80% methanol containing 0.1% formic acid, 8.4 nM lidcain, and 210 nM 10-camphoresulfonic acid as internal standards) was added, and sonication was performed for 10 min. After centrifugation (10,000 rpm for 5 min), the collected supernatant was used as the extract. To 25 μL of the extract, 75 μL of the extraction solvent was added. Twenty-five μL of the diluted extract was dried with N2 gas, and 250 μL of deionized water was added, dissolved by shaking, and used as the test solution. The final concentrations of the internal standards in the test solutions of the root samples were 0.84 nM for lidcain and 21 nM for 10-camphoresulfonic acid. After filtration through a MultiScreen 386 well-plate filter (Merck Millipore, Billerica, MA, USA), the test solution was subjected to metabolome analysis using a Shimadzu LCMS-8050 system coupled to a Nexera X2 UPLC system (Shimadzu, Kyoto, Japan). A Waters HSS T3 column (1.0 $\times$ 50 mm, 1.8 μm) was used, and the column oven temperature was kept at 30 °C. The mobile phase consisted of 0.1% (v/v) formic acid/water (A) and 0.1% (v/v) formic acid/acetonitrile (B) with gradient elution of 0.1 to 9% (B) at 0.25–0.4 min, 9 to 17% (B) at 0.4–0.8 min, 17 to 99.9% (B) at 0.8–1.9 min, 99.9 to 99.9% (B) at 1.9–2.1 min, 99.9 to 0.1% (B) at 2.1–2.11 min, and 0.1% (B) at 2.11–2.7 min. The flow rate of the mobile phase was 0.24 mL/min and a sample injection volume of 1 μL. The mass spectrometer was operated in the positive (+) and negative (-) ESI modes. Parameters for the interface were set as follows: nebulizing gas flow 3 mL/min, heating gas flow 10 L/min, drying gas flow 10 L/min, interface temperature 300 °C, dilution line temperature 250 °C, heat block temperature 400 °C, collision-induced dissociation gas 270 kPA and capillary voltage 4 kV (+); -3.5kv (-). The mass spectrometer was operated in the scheduled multiple reaction monitoring (MRM) mode for MS/MS measurements. The MRM transitions of each compound are listed in Supplementary Table 1. Raw data (lcd file) were converted into Abf files using the Reifycs Abf (Analysis base file) Converter (https://www.reifycs.com/AbfConverter/). Peak areas of the LC–QqQ-MS data were calculated using MRMPROBS (Tsugawa et al. 2013, http://prime.psc.riken.jp/compms/mrmprobs/ main.html) [63].

A.3 Microbiome analysis

DNA extraction

Bacterial DNA was extracted from the root and rhizosphere soils, as previously described [64, 65]. Briefly, 100 mg of the powdered sample was dispensed into a 1.5 mL tube and chilled in liquid nitrogen. Subsequently, 200 μL of Lysis/binding buffer (LBB) containing 1 M LiCl (Cat. #L7026-500ML, Sigma-Aldrich, St. Louis, MO, USA), 100 mM Tris-HCl (Cat. #318-90225, Wako Pure Chemical Corporation, Osaka, Japan), 1% SDS (Cat. #313-90275, Wako Pure Chemical Corporation), 10 mM EDTA pH 8.0 (Cat. #311-90075, Wako Pure Chemical Corporation), Antifoam A (Cat. #A5633-25G, Sigma-Aldrich), 5 mM dithiothreitol (Cat. #048-29224, Wako Pure Chemical Corporation), 11.2 M 3-Mercapto-1,2-propanediol (Cat. #139-16452, Wako Pure Chemical Corporation) and DNase/RNase-free H2O (Cat. #10977015, Thermo Fisher Scientific, Waltham, MA, USA) [1] were added to the tubes, mixed by vortexing, incubated at room temperature (24 °C) for 5 min. The tube of each sample was centrifuged at 13,000 rpm for 10 min at RT and 50 μL of supernatant was mixed with an equal amount of AMPure XP beads (Cat. #A63881, Beckman Coulter, Inc.) in new tube to clean and size selection of DNA. The mixture was vortexed, incubated for 5 min at RT and then placed on a magnetic station for 5 min. After removing the supernatant, the magnetic beads were washed twice with 200 μL of 80% ethanol. Finally, the DNA was eluted from the magnetic beads using 15 μL of 10 mM Tris-HCl (pH 7.5).

Bacterial 16S rRNA gene sequencing

Libraries were prepared by modifying a previously described two-step PCR amplification protocol [1, 64]. Briefly, the DNA library was constructed using primers (515f: 5'- TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG - [3-6-mer Ns] - GTG YCA GCM GCC GCG GTA A -3', 806rB: 5'- GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G - [3-6-mer Ns] - GGA CTA CNV GGG TWT CTA AT -3') for amplifying the V4 region of the bacterial 16S rRNA [66, 67]. The V4 region of 16S rRNA was amplified in a 10 μL reaction containing 1 μL of 10-fold diluted sample, 2× KAPA HiFi HotStart ReadyMix (Cat. #07958935001, KAPA Biosystems Inc., USA), 0.2 μM forward and reverse primers, and 1 μM blocking primers (mPNA and pPNA; PNA BIO, Inc., Newbury Park, CA, USA) in a thermal cycler as follows: 1 cycle for 3 min at 95 °C, 35 cycles for 20 sec at 98 °C, 10 sec at 78 °C, 30 sec at 50 °C, 90 sec at 72 °C, and a final extension for 5 min at 72 °C (ramp rate = 1 °C/s).

The first PCR product was purified using ExoSAP-IT Express (Cat #75001.1.EA; Thermo Fisher Scientific) to remove excess primer and dNTPs. The PCR product (5 μL) was mixed with 2 μL of ExoSAP-IT Express and incubated at 37 °C for 4 min, followed by inactivation at 80 °C for 1 min.

The second PCR was performed using the following index primers: forward primer (5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC AC - [8-mer

index] - TCG TCG GCA GCG TC -3') and reverse primer (5'- CAA GCA GAA GAC GGC ATA CGA GAT - [8-mer index] - GTC TCG TGG GCT CGG -3') [68]. The second PCR reaction was performed in a 10 μL reaction containing 0.8 μL of the purified first amplification, 2× KAPA HiFi HotStart ReadyMix, 1 μM forward and reverse primers, and 1 μM blocking primers (mPNA and pPNA) in a thermal cycler as follows: 1 cycle for 3 min at 95 °C, 8 cycles for 20 sec at 98 °C, 10 sec at 78 °C, 30 sec at 55 °C, 90 sec at 72 °C, and a final extension for 5 min at 72 °C (ramp rate = 1 °C/s). For the second PCR amplification, each sample was mixed with AMPure XP beads, vortexed and left at room temperature for 5 min. The samples were then placed in a magnetic station for 5 min. After removing the supernatant, the magnetic beads were washed twice with 200 μL of 80% ethanol. The DNA libraries were eluted from the magnetic beads with 15 μL of 10 mM Tris-HCl (pH 7.5) and quantified using Quant-iT PicoGreen dsDNA reagent (Cat. #P7581, Invitrogen) and a microplate photometer (Infinite 200 PRO M Nano+, TECAN Japan Co., Ltd., Kanagawa, Japan). DNA libraries were diluted and pooled with equimolar concentrations of each library. The pooled libraries were quantified by qPCR using the NEBNext Library Quant Kit for Illumina (Cat. #E7630L, New England BioLabs, Ipswich, MA, USA) and diluted to 3 nM. Denatured libraries were spiked with 20% PhiX Control v3 (Cat. #FC-110-3001, Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq platform using a 2× 300-bp MiSeq Reagent Kit v3 (Illumina).

16S V4 rRNA data preprocessing

Raw paired-end reads were processed using the Quantitative Insight Into Microbial Ecology 2 program (QIIME2, ver. 2020.6.0). Raw FASTQ files were imported into QIIME2, sequencing primers were trimmed using the Cutadapt plugin. For ASV-based analysis, the primer-free sequencing reads were analyzed using the DADA2 pipeline [69] to truncate forward and reverse reads (length of 240 for forward reads and 180 for reverse reads), filter, denoise, merge reads, and detect and remove chimeras. The taxonomy of all ASVs was annotated using SILVA (ver. 138) reference database [70, 71]. Sequences derived from Archaea, Eukaryota, mitochondria, or chloroplasts were removed.

References

[1] Ichihashi, Y., Date, Y., Shino, A., Shimizu, T., Shibata, A., Kumaishi, K., Funahashi, F., Wakayama, K., Yamazaki, K., Umezawa, A., *et al.*: Multi-omics analysis on an agroecosystem reveals the significant role of organic nitrogen to increase agricultural crop yield. Proceedings of the National Academy of Sciences **117**(25), 14552–14560 (2020)

[2] Xu, L., Pierroz, G., Wipf, H.M.-L., Gao, C., Taylor, J.W., Lemaux, P.G., Coleman-Derr, D.: Holo-omics for deciphering plant-microbiome interactions. Microbiome **9**(1), 1–11 (2021)

[3] Fujiwara, F., Miyazawa, K., Nihei, N., Ichihashi, Y.: Agroecosystem engineering extended from plant-microbe interactions revealed by multi-omics data. Bioscience, Biotechnology, and Biochemistry **87**(1), 21–27 (2023)

[4] Spor, A., Koren, O., Ley, R.: Unravelling the effects of the environment and host genotype on the gut microbiome. Nature Reviews Microbiology **9**(4), 279–290 (2021)

[5] Kurilshikov, A., Medina-Gomez, C., Bacigalupe, R., Radjabzadeh, D., Wang, J., Demirkan, A., Le Roy, C.I., Raygoza Garay, J.A., Finnicum, C.T., Liu, X., *et al.*: Large-scale association analyses identify host factors influencing human gut microbiome composition. Nature genetics **53**(2), 156–165 (2021)

[6] Priya, S., Burns, M.B., Ward, T., Mars, R.A., Adamowicz, B., Lock, E.F., Kashyap, P.C., Knights, D., Blekhman, R.: Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. Nature Medicine **7**(1), 1 (2022)

[7] Sanna, S., Kurilshikov, A., van der Graaf, A., Fu, J., Zhernakova, A.: Challenges and future directions for studying effects of host genetics on the gut microbiome. Nature genetics **54**(2), 100–106 (2022)

[8] Di Simone, S.K., Rudloff, I., Nold-Petry, C.A., Forster, S.C., Nold, M.F.: Understanding respiratory microbiome–immune system interactions in health and disease. Science Translational Medicine **15**(678), 5126 (2023)

[9] O'Donnell, J.A., Zheng, T., Meric, G., Marques, F.Z.: The gut microbiome and hypertension. Nature Reviews Nephrology **19**, 153–167 (2023)

[10] Valles-Colomer, M., Menni, C., Berry, S.E., Valdes, A.M., Spector, T.D., Segata, N.: Cardiometabolic health, diet and the gut microbiome: a meta-omics perspective. Nature Medicine **29**, 551–561 (2023)

[11] Visconti, A., Le Roy, C.I., Rosa, F., Rossi, N., Martin, T.C., Mohney, R.P., Li, W., de Rinaldis, E., Bell, J.T., Venter, J.C., *et al.*: Interplay between the human gut microbiome and host metabolism. Nature communications **10**(1), 1–10 (2019)

[12] Asnicar, F., Berry, S.E., Valdes, A.M., Nguyen, L.H., Piccinno, G., Drew, D.A., Leeming, E., Gibson, R., Le Roy, C., Khatib, H.A., *et al.*: Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. Nature Medicine **27**(2), 321–332 (2021)

[13] Cheney, A.M., Costello, S.M., Pinkham, N.V., Waldum, A., Broadaway, S.C., Cotrina-Vidal, M., Mergy, M., Tripet, B., Kominsky, D.J., Grifka-Walk, H.M., *et al.*: Gut microbiome dysbiosis drives metabolic dysfunction

in familial dysautonomia. Nature Communications **14**(1), 218 (2023)

[14] Xiong, R., Gunter, C., Fleming, E., Vernon, S.D., Bateman, L., Unutmaz, D., Oh, J.: Multi-'omics of gut microbiome-host interactions in short-and long-term myalgic encephalomyelitis/chronic fatigue syndrome patients. Cell Host & Microbe **31**(2), 273–287 (2023)

[15] Lock, E.F., Dunson, D.B.: Bayesian consensus clustering. Bioinformatics **29**(20), 2610–2616 (2022)

[16] Gabasova, E., Reid, J., Wernisch, L.: Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLoS computational biology **13**(10), 1005781 (2017)

[17] Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., Shen, R.: Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences **110**(11), 4245–4250 (2013)

[18] Integrative, H.: The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell host & microbe **16**(3), 276–289 (2014)

[19] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)

[20] Holmes, I., Harris, K., Quince, C.: Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS one **7**(2), 30126 (2012)

[21] Dang, T., Kumaishi, K., Usui, E., Kobori, S., Sato, T., Toda, Y., Yamasaki, Y., Tsujimoto, H., Ichihashi, Y., Iwata, H.: Stochastic variational variable selection for high-dimensional microbiome data. Microbiome **10**(1), 1–14 (2022)

[22] Yee Whye Teh, M.J.B. Michael I Jordan, Blei, D.M.: Hierarchical dirichlet processes. Journal of the American Statistical Association **101**(476), 1566–1581 (2006)

[23] Shringarpure, S., Won, D., Xing, E.P.: Structhdp: automatic inference of number of clusters and population structure from admixed genotype data. Bioinformatics **27**(13), 324–332 (2011)

[24] Margelevičius, M.: Bayesian nonparametrics in protein remote homology search. Bioinformatics **32**(18), 2744–2752 (2016)

[25] Margelevičius, M.: A low-complexity add-on score for protein remote homology search with comer. Bioinformatics **34**(12), 2037–2045 (2018)

[26] duVerle, D.A., Yotsukura, S., Nomura, S., Aburatani, H., Tsuda, K.: Celltree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data. BMC bioinformatics **17**(1), 1–17 (2016)

[27] Adossa, N.A., Rytkönen, K.T., Elo, L.L.: Dirichlet process mixture models for single-cell rna-seq clustering. Biology Open **11**(4), 059001 (2022)

[28] Yang, Q., Xu, Z., Zhou, W., Wang, P., Jiang, Q., Juan, L.: An interpretable single-cell rna sequencing data clustering method based on latent dirichlet allocation. Briefings in Bioinformatics, 199 (2023)

[29] Tripathi, A., Melnik, A.V., Xue, J., Poulsen, O., Meehan, M.J., Humphrey, G., Jiang, L., Ackermann, G., McDonald, D., Zhou, D., *et al.*: Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome. Msystems **3**(3), 00020–18 (2018)

[30] Schubert, A.M., Rogers, M.A., Ring, C., Mogle, J., Petrosino, J.P., Young, V.B., Aronoff, D.M., Schloss, P.D.: Microbiome data distinguish patients with clostridium difficile infection and non-c. difficile-associated diarrhea from healthy controls. MBio **5**(3), 01021–14 (2014)

[31] Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. Journal of Machine Learning Research (2013)

[32] McMurdie, P.J., Holmes, S.: phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. PloS one **8**(4), 61217 (2013)

[33] Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., *et al.*: Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science **341**(6150), 1241214 (2013)

[34] Lacroix, S., Pechereau, F., Leblanc, N., Boubertakh, B., Houde, A., Martin, C., Flamand, N., Silvestri, C., Raymond, F., Di Marzo, V., *et al.*: Rapid and concomitant gut microbiota and endocannabinoidome response to diet-induced obesity in mice. MSystems **4**(6), 00407–19 (2019)

[35] Blakeley-Ruiz, J.A., McClintock, C.S., Shrestha, H.K., Poudel, S., Yang, Z.K., Giannone, R.J., Choo, J.J., Podar, M., Baghdoyan, H.A., Lydic, R., *et al.*: Morphine and high-fat diet differentially alter the gut microbiota composition and metabolic function in lean versus obese mice. ISME

Communications **2**(1), 66 (2022)

[36] Trivedi, P., Leach, J.E., Tringe, S.G., Sa, T., Singh, B.K.: Plant–microbiome interactions: from community assembly to plant health. Nature reviews microbiology **18**(11), 607–621 (2020)

[37] Mendes, L.W., Kuramae, E.E., Navarrete, A.A., Van Veen, J.A., Tsai, S.M.: Taxonomical and functional microbial community selection in soybean rhizosphere. The ISME journal **8**(8), 1577–1587 (2014)

[38] Sugiyama, A., Ueda, Y., Zushi, T., Takase, H., Yazaki, K.: Changes in the bacterial community of soybean rhizospheres during growth in the field. PloS one **9**(6), 100709 (2014)

[39] Cai, F., Luo, P., Yang, J., Irfan, M., Zhang, S., An, N., Dai, J., Han, X.: Effect of long-term fertilization on ammonia-oxidizing microorganisms and nitrification in brown soil of northeast china. Frontiers in Microbiology **11**, 622454 (2021)

[40] Prosser, J.I., Head, I.M., Stein, L.Y.: The family nitrosomonadaceae. In: The Prokaryotes: Alphaproteobacteria and Betaproteobacteria, pp. 901–918. Springer, ??? (2014)

[41] Clark, I.M., Hughes, D.J., Fu, Q., Abadie, M., Hirsch, P.R.: Metagenomic approaches reveal differences in genetic diversity and relative abundance of nitrifying bacteria and archaea in contrasting soils. Scientific Reports **11**(1), 1–9 (2021)

[42] Bailey, V.L., Fansler, S.J., Stegen, J.C., McCue, L.A.: Linking microbial community structure to β-glucosidic function in soil aggregates. The ISME journal **7**(10), 2044–2053 (2013)

[43] Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., De Hollander, M., Ruiz-Buck, D., Mendes, L.W., van Ijcken, W.F., Gomez-Exposito, R., Elsayed, S.S., *et al.*: Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science **366**(6465), 606–612 (2019)

[44] Hamedi, J., Mohammadipanah, F.: Biotechnological application and taxonomical distribution of plant growth promoting actinobacteria. Journal of industrial microbiology and biotechnology **42**(2), 157–171 (2015)

[45] Han, Q., Ma, Q., Chen, Y., Tian, B., Xu, L., Bai, Y., Chen, W., Li, X.: Variation in rhizosphere microbial communities and its association with the symbiotic efficiency of rhizobia in soybean. The ISME journal **14**(8), 1915–1928 (2020)

[46] Chukwuneme, C.F., Ayangbenro, A.S., Babalola, O.O.: Metagenomic analyses of plant growth-promoting and carbon-cycling genes in maize rhizosphere soils with distinct land-use and management histories. Genes **12**(9), 1431 (2021)

[47] Gamalero, E., Glick, B.R.: Bacterial modulation of plant ethylene levels. Plant physiology **169**(1), 13–22 (2015)

[48] Liu, H., Khan, M.Y., Carvalhais, L.C., Delgado-Baquerizo, M., Yan, L., Crawford, M., Dennis, P.G., Singh, B., Schenk, P.M.: Soil amendments with ethylene precursor alleviate negative impacts of salinity on soil microbial properties and productivity. Scientific Reports **9**(1), 6892 (2019)

[49] Kim, D.-R., Jeon, C.-W., Cho, G., Thomashow, L.S., Weller, D.M., Paik, M.-J., Lee, Y.B., Kwak, Y.-S.: Glutamic acid reshapes the plant microbiota to protect plants against pathogens. Microbiome **9**(1), 1–18 (2021)

[50] Poroyko, V.A., Carreras, A., Khalyfa, A., Khalyfa, A.A., Leone, V., Peris, E., Almendros, I., Gileles-Hillel, A., Qiao, Z., Hubert, N., *et al.*: Chronic sleep disruption alters gut microbiota, induces systemic and adipose tissue inflammation and insulin resistance in mice. Scientific reports **6**(1), 35405 (2016)

[51] Allaband, C., Lingaraju, A., Martino, C., Russell, B., Tripathi, A., Poulsen, O., Dantas Machado, A.C., Zhou, D., Xue, J., Elijah, E., *et al.*: Intermittent hypoxia and hypercapnia alter diurnal rhythms of luminal gut microbiome and metabolome. Msystems **6**(3), 00116–21 (2021)

[52] Zhang, Y., Luo, H., Niu, Y., Yang, X., Li, Z., Wang, K., Bi, H., Pang, X.: Chronic intermittent hypoxia induces gut microbial dysbiosis and infers metabolic dysfunction in mice. Sleep Medicine **91**(1), 84–92 (2022)

[53] Xue, J., Allaband, C., Zhou, D., Poulsen, O., Martino, C., Jiang, L., Tripathi, A., Elijah, E., Dorrestein, P.C., Knight, R., *et al.*: Influence of intermittent hypoxia/hypercapnia on atherosclerosis, gut microbiome, and metabolome. Frontiers in Physiology **12**(1), 663950 (2021)

[54] Collins, S.L., Stine, J.G., Bisanz, J.E., Okafor, C.D., Patterson, A.D.: Bile acids and the gut microbiota: metabolic interactions and impacts on disease. Nature Reviews Microbiology **21**(4), 236–247 (2023)

[55] Zhang, Y., Chen, R., Zhang, D., Qi, S., Liu, Y.: Metabolite interactions between host and microbiota during health and disease: Which feeds the other? Biomedicine & Pharmacotherapy **160**(1), 114295 (2023)

[56] Hultman, J., Waldrop, M.P., Mackelprang, R., David, M.M., McFarland, J., Blazewicz, S.J., Harden, J., Turetsky, M.R., McGuire, A.D., Shah, M.B., *et al.*: Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature **521**(7551), 208–212 (2015)

[57] Sakurai, K., Toda, Y., Kajiya-Kanegae, H., Ohmori, Y., Yamasaki, Y., Takahashi, H., Takanashi, H., Tsuda, M., Tsujimoto, H., Kaga, A., *et al.*: Time-series multispectral imaging in soybean for improving biomass and genomic prediction accuracy. The Plant Genome **15**(4), 20244 (2022)

[58] Toda, Y., Sasaki, G., Ohmori, Y., Yamasaki, Y., Takahashi, H., Takanashi, H., Tsuda, M., Kajiya-Kanegae, H., Lopez-Lozano, R., Tsujimoto, H., et al.: Genomic prediction of green fraction dynamics in soybean using unmanned aerial vehicles observations. Frontiers in Plant Science **13** (2022)

[59] Kaga, A., Shimizu, T., Watanabe, S., Tsubokura, Y., Katayose, Y., Harada, K., Vaughan, D.A., Tomooka, N.: Evaluation of soybean germplasm conserved in nias genebank and development of mini core collections. Breeding science **61**(5), 566–592 (2012)

[60] Kajiya-Kanegae, H., Nagasaki, H., Kaga, A., Hirano, K., Ogiso-Tanaka, E., Matsuoka, M., Ishimori, M., Ishimoto, M., Hashiguchi, M., Tanaka, H., *et al.*: Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections. DNA Research **28**(1), 032 (2021)

[61] Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T., Saito, K., Hirai, M.Y.: Widely targeted metabolomics based on large-scale ms/ms data for elucidating metabolite accumulation patterns in plants. Plant and Cell Physiology **50**(1), 37–47 (2009)

[62] Uchida, K., Sawada, Y., Ochiai, K., Sato, M., Inaba, J., Hirai, M.Y.: Identification of a unique type of isoflavone o-methyltransferase, gmiomt1, based on multi-omics analysis of soybean under biotic stress. Plant and Cell Physiology **61**(11), 1974–1985 (2020)

[63] Tsugawa, H., Arita, M., Kanazawa, M., Ogiwara, A., Bamba, T., Fukusaki, E.: Mrmprobs: a data assessment and metabolite identification tool for large-scale multiple reaction monitoring based widely targeted metabolomics. Analytical chemistry **85**(10), 5191–5199 (2013)

[64] Kumaishi, K., Usui, E., Suzuki, K., Kobori, S., Sato, T., Toda, Y., Takanashi, H., Shinozaki, S., Noda, M., Takakura, A., *et al.*: High throughput method of 16s rrna gene sequencing library preparation for plant root microbial community profiling. Scientific Reports **12**(1), 19289 (2022)

[65] Ichihashi, Y., Fukushima, A., Shibata, A., Shirasu, K.: High impact gene discovery: simple strand-specific mrna library construction and differential regulatory analysis based on gene co-expression network. Plant Transcription Factors: Methods and Protocols, 163–189 (2018)

[66] Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R.: Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. Proceedings of the national academy of sciences **108**(supplement_1), 4516–4522 (2011)

[67] Lundberg, D.S., Yourstone, S., Mieczkowski, P., Jones, C.D., Dangl, J.L.: Practical innovations for high-throughput amplicon sequencing. Nature methods **10**(10), 999–1002 (2013)

[68] Toju, H., Baba, Y.G.: Dna metabarcoding of spiders, insects, and springtails for exploring potential linkage between above-and below-ground food webs. Zoological letters **4**, 1–12 (2018)

[69] Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature methods **13**(7), 581–583 (2016)

[70] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic acids research **41**(D1), 590–596 (2012)

[71] Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O.: The silva and "all-species living tree project (ltp)" taxonomic frameworks. Nucleic acids research **42**(D1), 643–648 (2014)

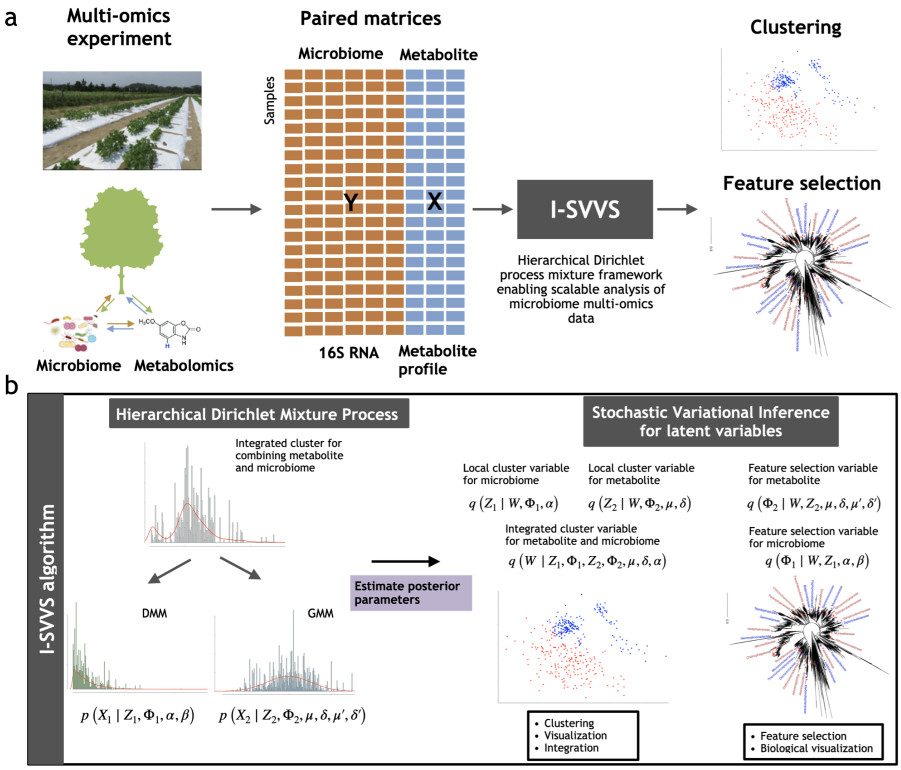


Fig. 1: **Schematic of a microbiome multi-omics data analysis pipeline with I-SVVS**. **a**, An example of a microbiome multi-omics experiment on a plant simultaneously measures metabolite profile and microbiome species in each sample, producing paired matrices for metabolite and microbiome. These matrices include a matrix of count values for microbiome species data and another matrix of continuous values for metabolite profile data. They are the input to the I-SVVS, which integrates all databases to cluster samples and selects the important features of each database. **b**, Schematic of the I-SVVS approach. Firstly, the infinite Dirichlet multinomial mixture (DMM) model with variable selection is proposed for microbiome data, Z_1 is an allocation variable, Φ_1 is a feature selection variable, α is parameter of Dirichlet-multinomial distribution for the selected group of microbiome species and β is parameter of Dirichlet-multinomial distribution for the rejected group of microbiome species. The infinite Gaussian mixture model (GMM) with variable selection is proposed for the metabolite profile data, Z_2 is an allocation variable, Φ_2 is a feature selection variable, μ, δ are the parameters of normal distribution captured the selected group of metabolite profiles and μ', δ' are parameters of normal distribution captured the rejected group of metabolite profiles. Then the framework of the hierarchical Dirichlet mixture process integrates the information of GMM and DMM approaches that allow the shared information between metabolite profile and microbiome data for the clustering process. W denotes the integrated (global) allocation variable. Next, the stochastic variable inference is proposed to estimate the posterior distributions of the latent variables for clustering and variable selection (Methods).

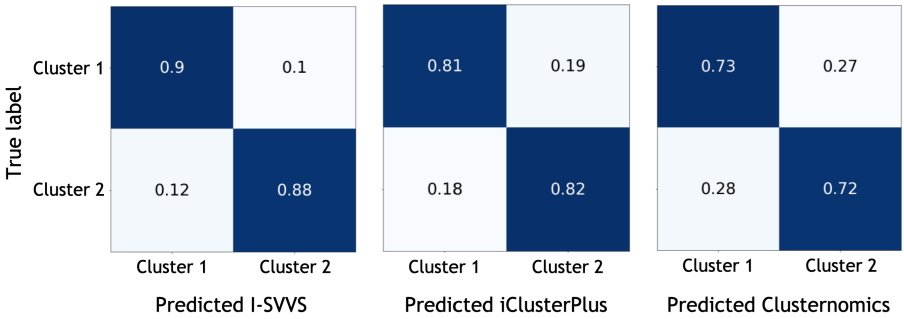


Fig. 2: Confusion matrix plots of dataset A with labels indicating predicted class using the three approaches and true group. Cluster 1 denotes the drought and cluster 2 denotes the control. (Left) I-SVVS, (Middle) iClusterPlus, (Right) Clusternomics.

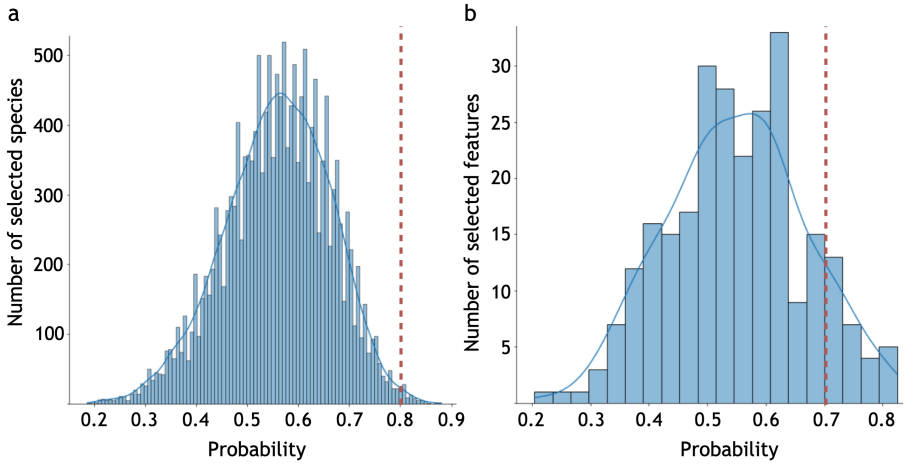


Fig. 3: Histogram of the average of Φ_{1ij} and $\Phi_{2ij'}$ in dataset A. The dashed lines are bound to select microbiome species and metabolome profile features. a. microbiome data; b. metabolome data

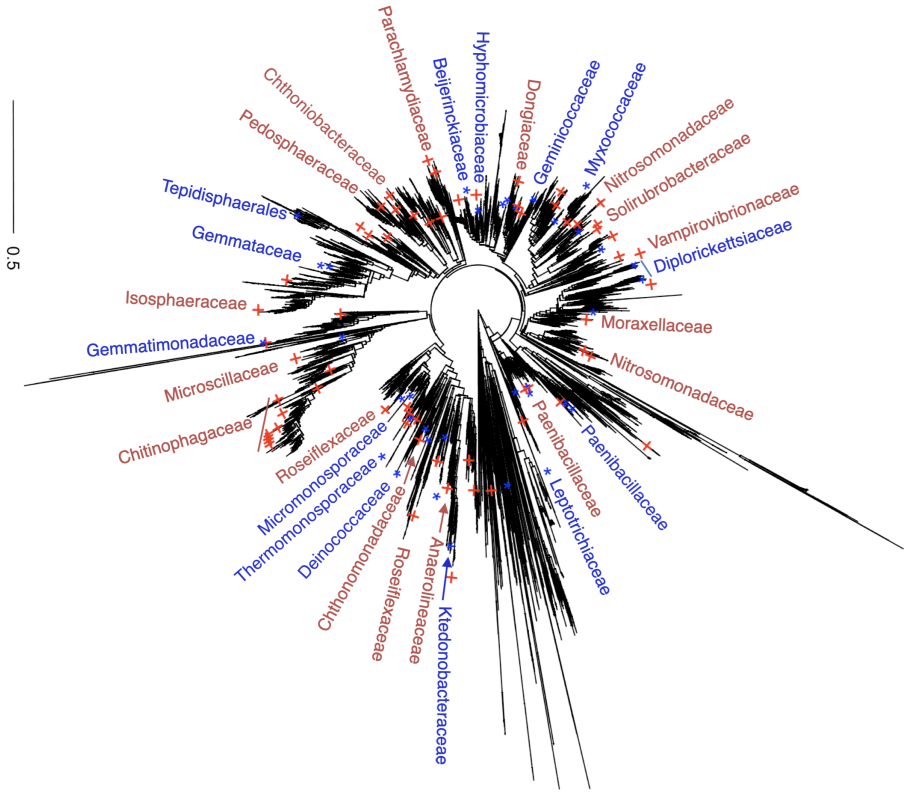
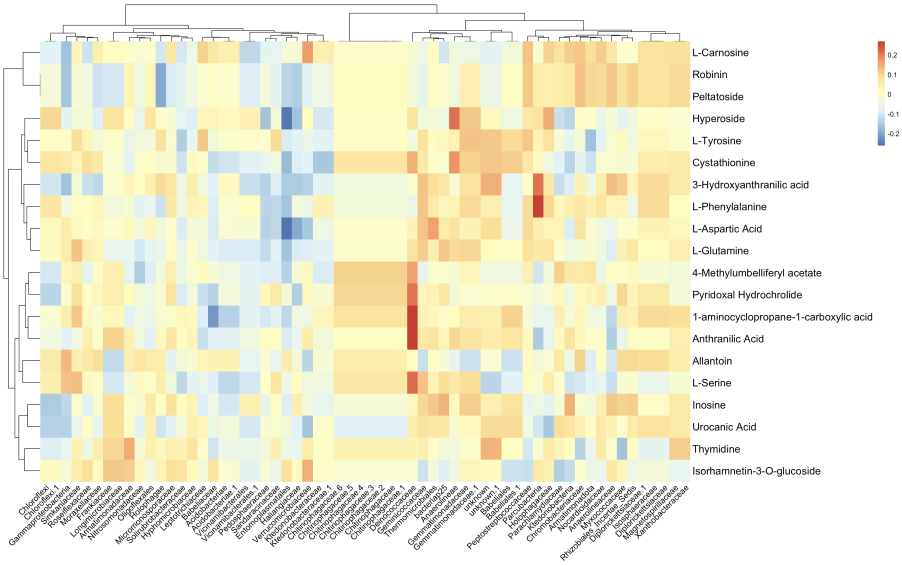
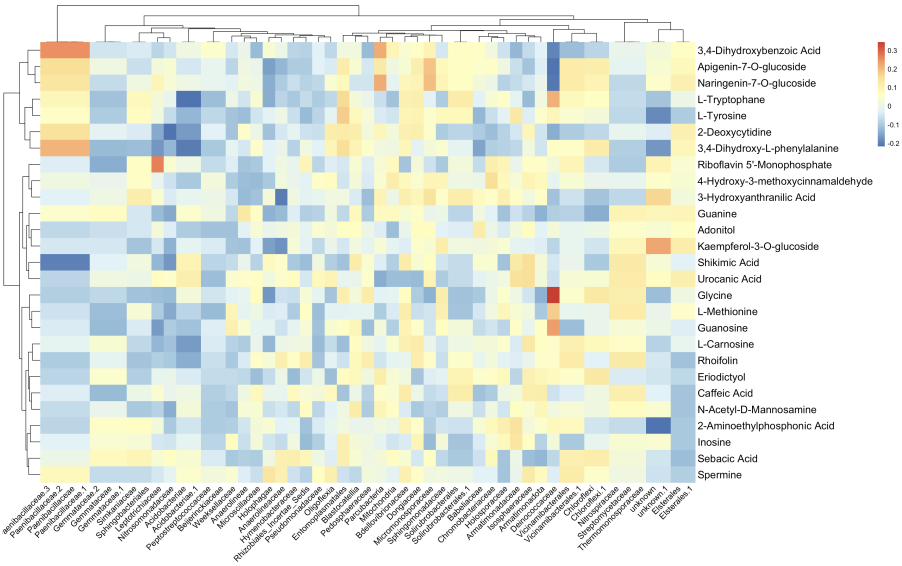


Fig. 4: Microbial species selected using the integrative stochastic variational variable selection (I-SVVS) approach and mapped on the phylogenetic tree for dataset A. Red-colored plus symbols denote the control and blue-colored starts denote drought.



(a) Control group

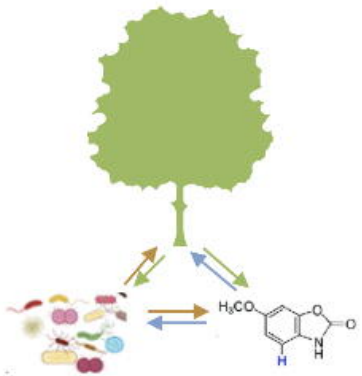


(b) Drought group

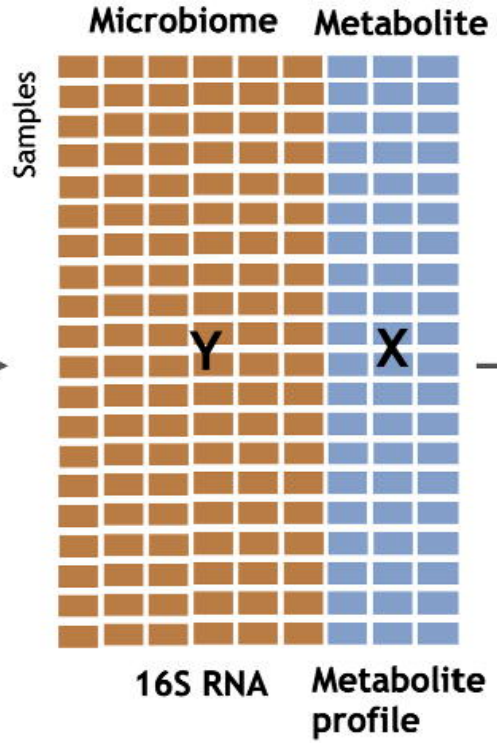
Fig. 5: Heat map of metabolomic profiles and microbiome species that were selected by the I-SVVS approach in two groups. Individual metabolites and microbiomes were hierarchically clustered (Ward's method) using Euclidean distance.

a

Multi-omics experiment



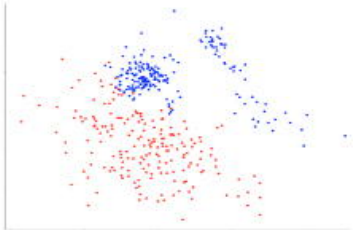
Paired matrices



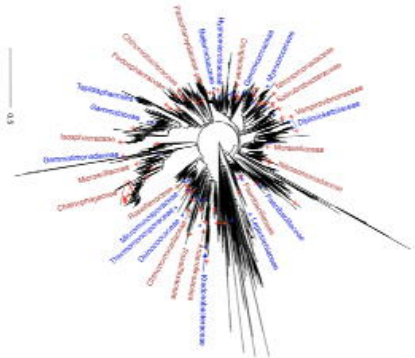
I-SVVS

Hierarchical Dirichlet process mixture framework enabling scalable analysis of microbiome multi-omics data

Clustering



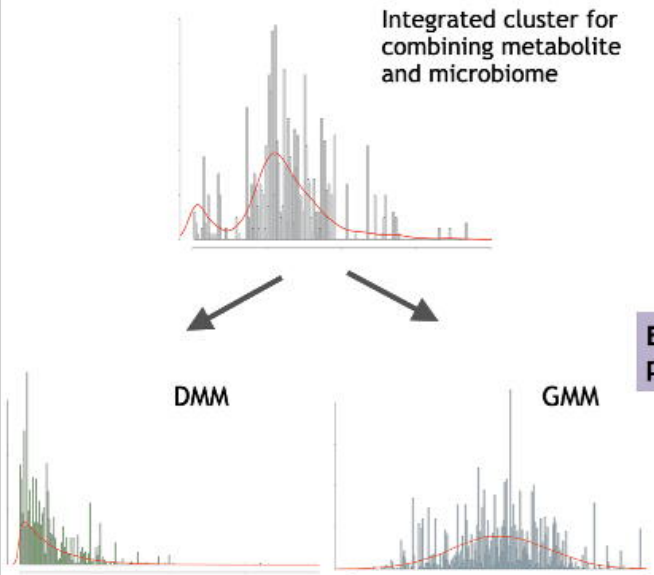
Feature selection



b

I-SVVS algorithm

Hierarchical Dirichlet Mixture Process



$p\left(X_1 \mid Z_1, \Phi_1, \alpha, \beta\right)$ $p\left(X_2 \mid Z_2, \Phi_2, \mu, \delta, \mu', \delta'\right)$

Estimate posterior parameters

Stochastic Variational Inference for latent variables

Local cluster variable for microbiome

$q\left(Z_1 \mid W, \Phi_1, \alpha\right)$

Local cluster variable for metabolite

$q\left(Z_2 \mid W, \Phi_2, \mu, \delta\right)$

Feature selection variable for metabolite

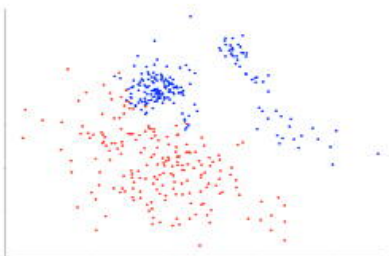
$q\left(\Phi_2 \mid W, Z_2, \mu, \delta, \mu', \delta'\right)$

Integrated cluster variable for metabolite and microbiome

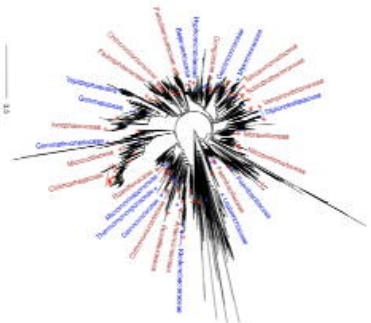
$q\left(W \mid Z_1, \Phi_1, Z_2, \Phi_2, \mu, \delta, \alpha\right)$

Feature selection variable for microbiome

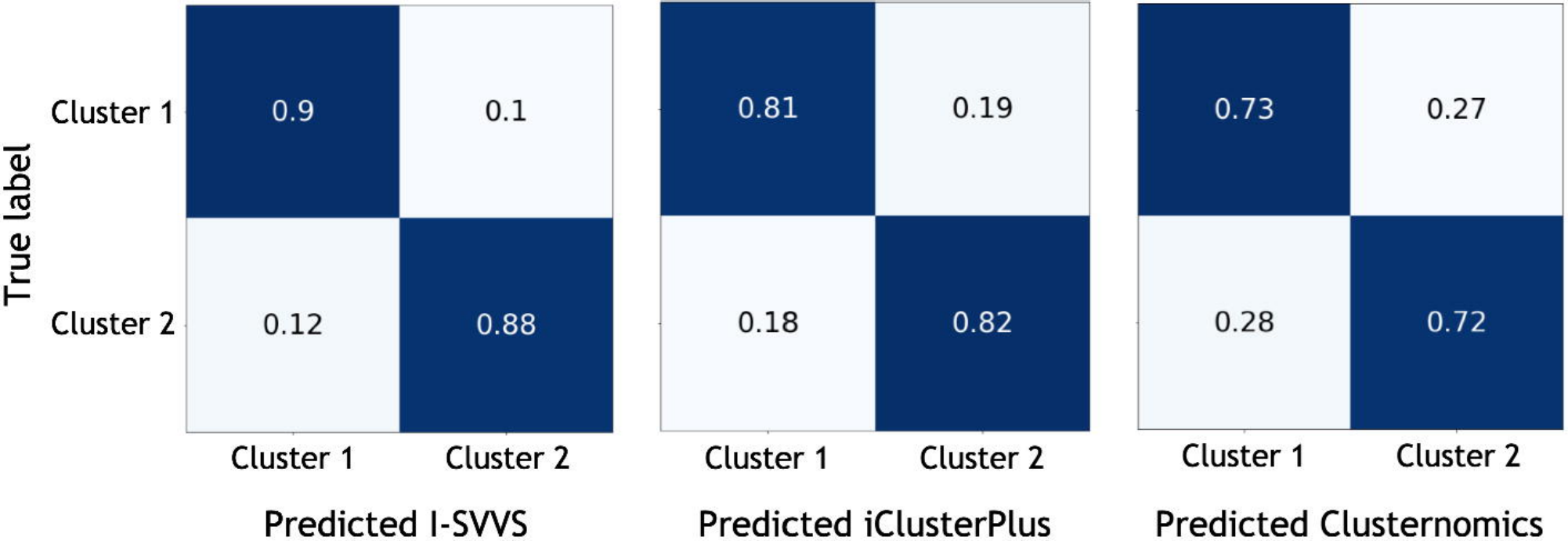
$q\left(\Phi_1 \mid W, Z_1, \alpha, \beta\right)$

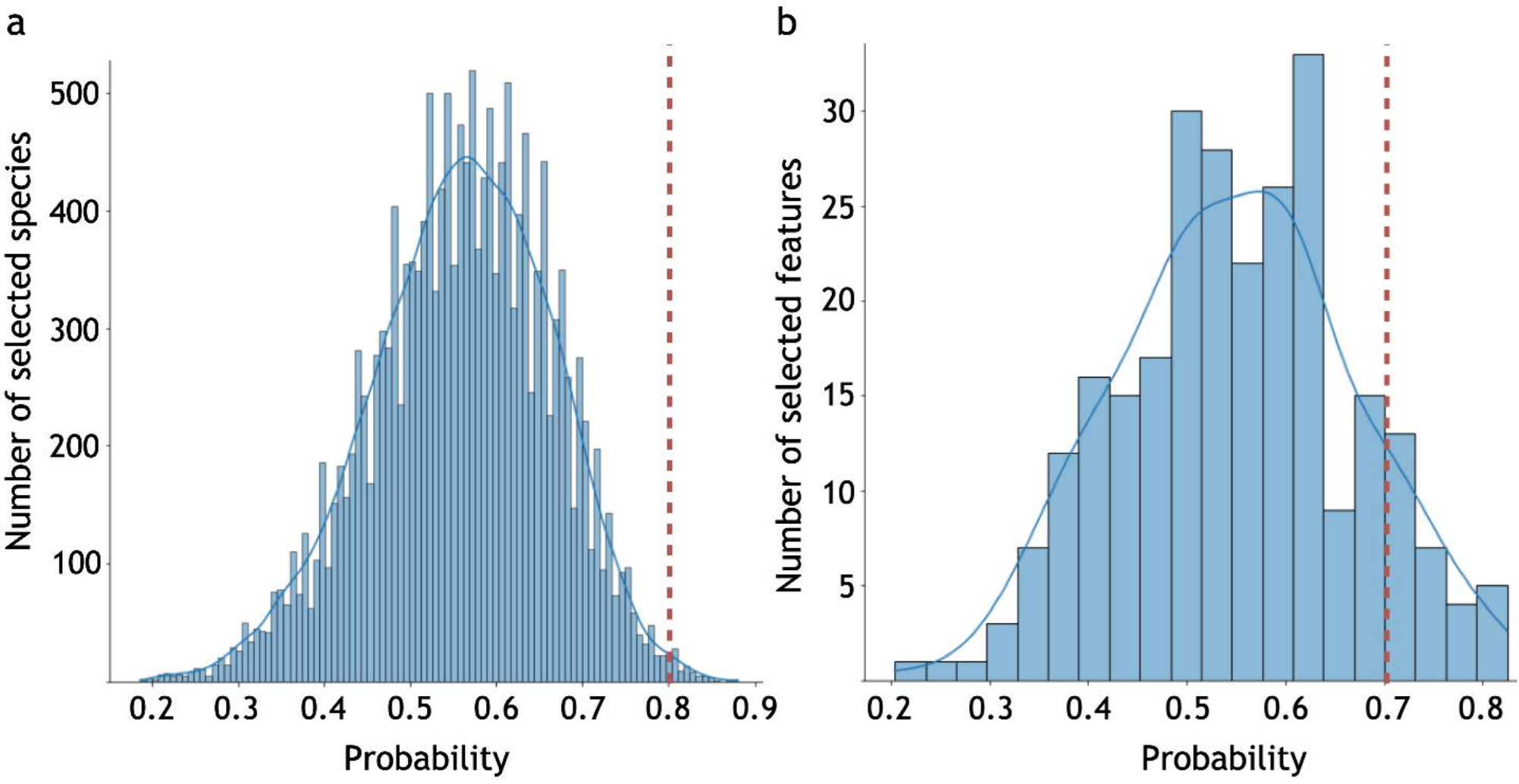


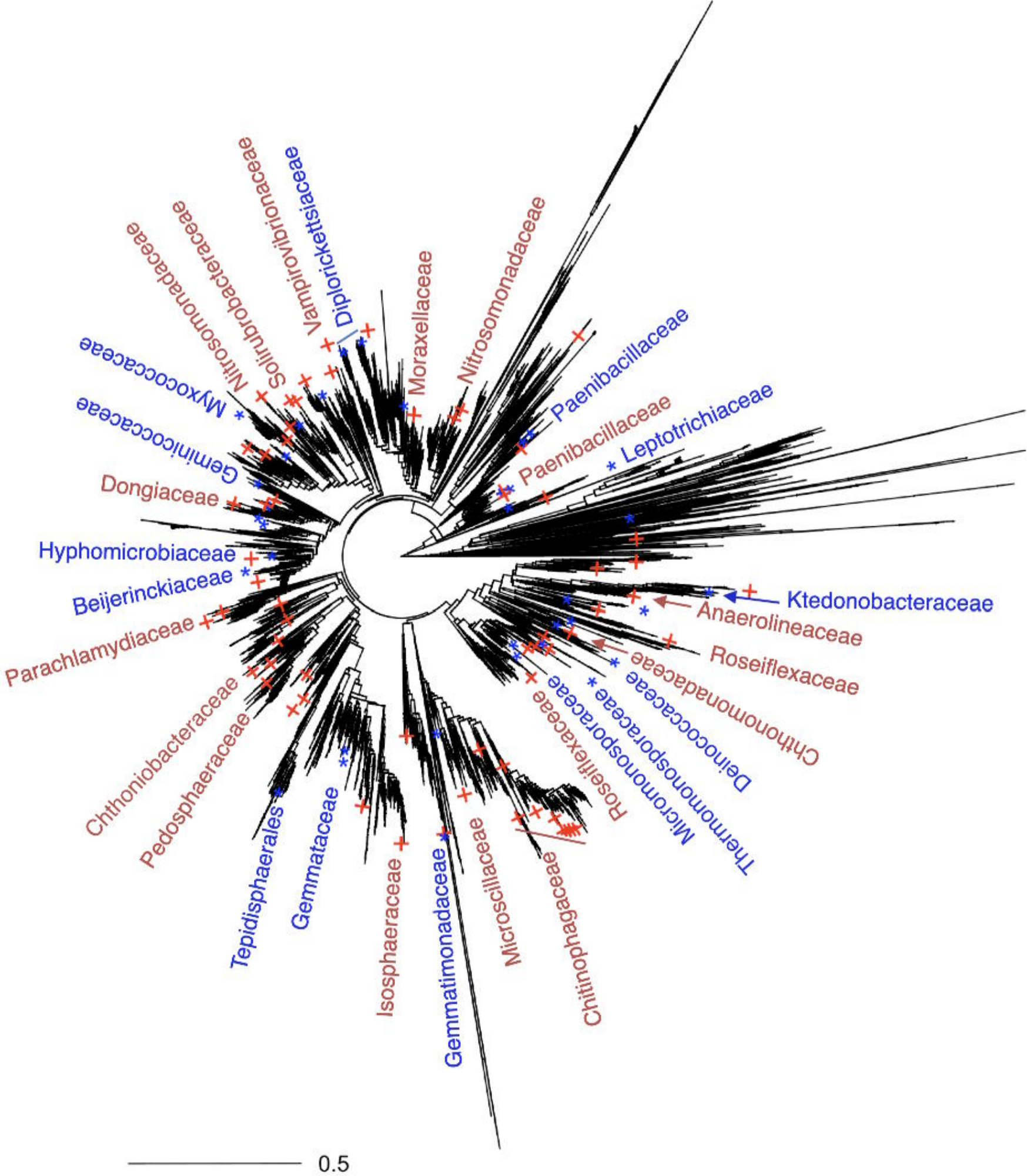
- Clustering
- Visualization
- Integration

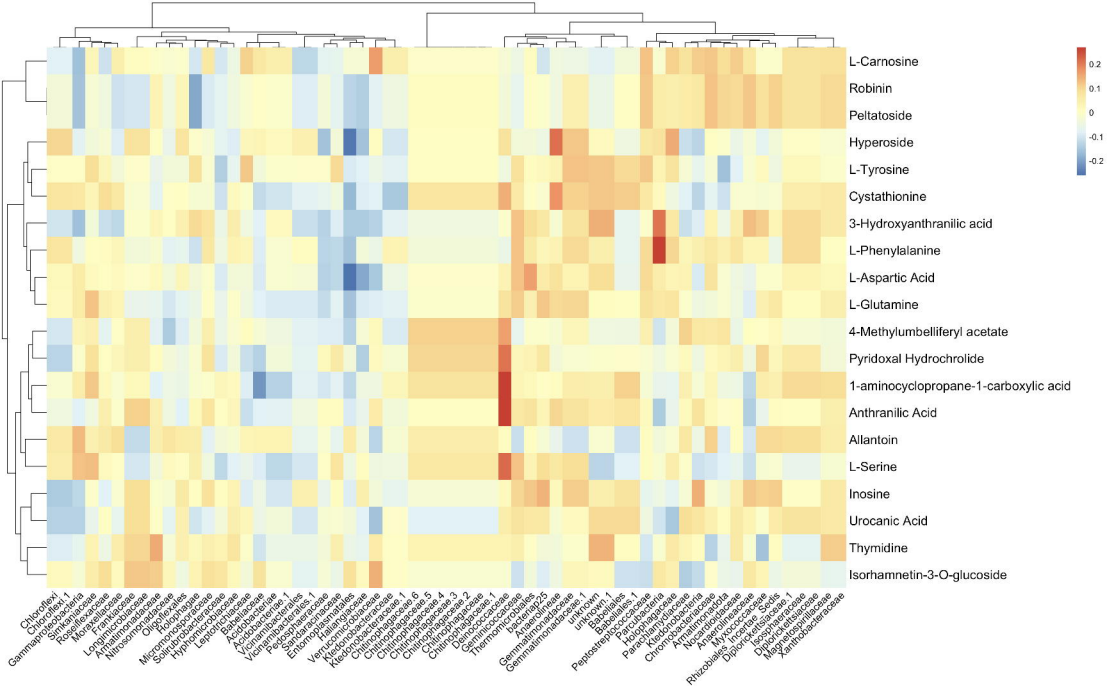


- Feature selection
- Biological visualization

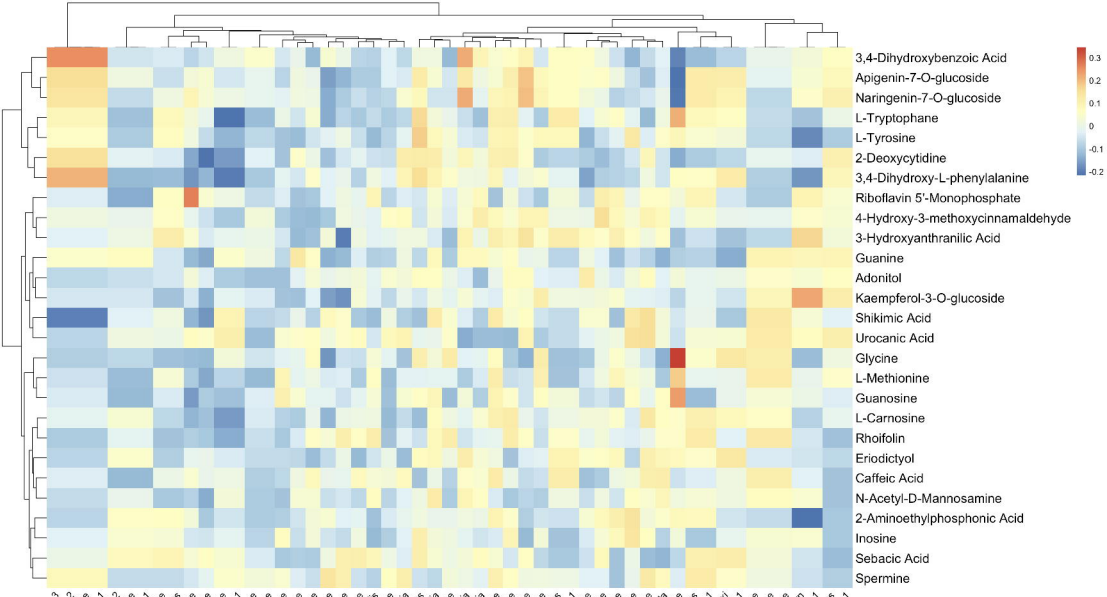

True label
Cluster 1
Cluster 2
0.9
0.1
0.12
0.88
Cluster 1
Cluster 2
Predicted I-SVVS
0.81
0.19
0.18
0.82
Cluster 1
Cluster 2
Predicted iClusterPlus
0.73
0.27
0.28
0.72
Cluster 1
Cluster 2
Predicted Clusternomics

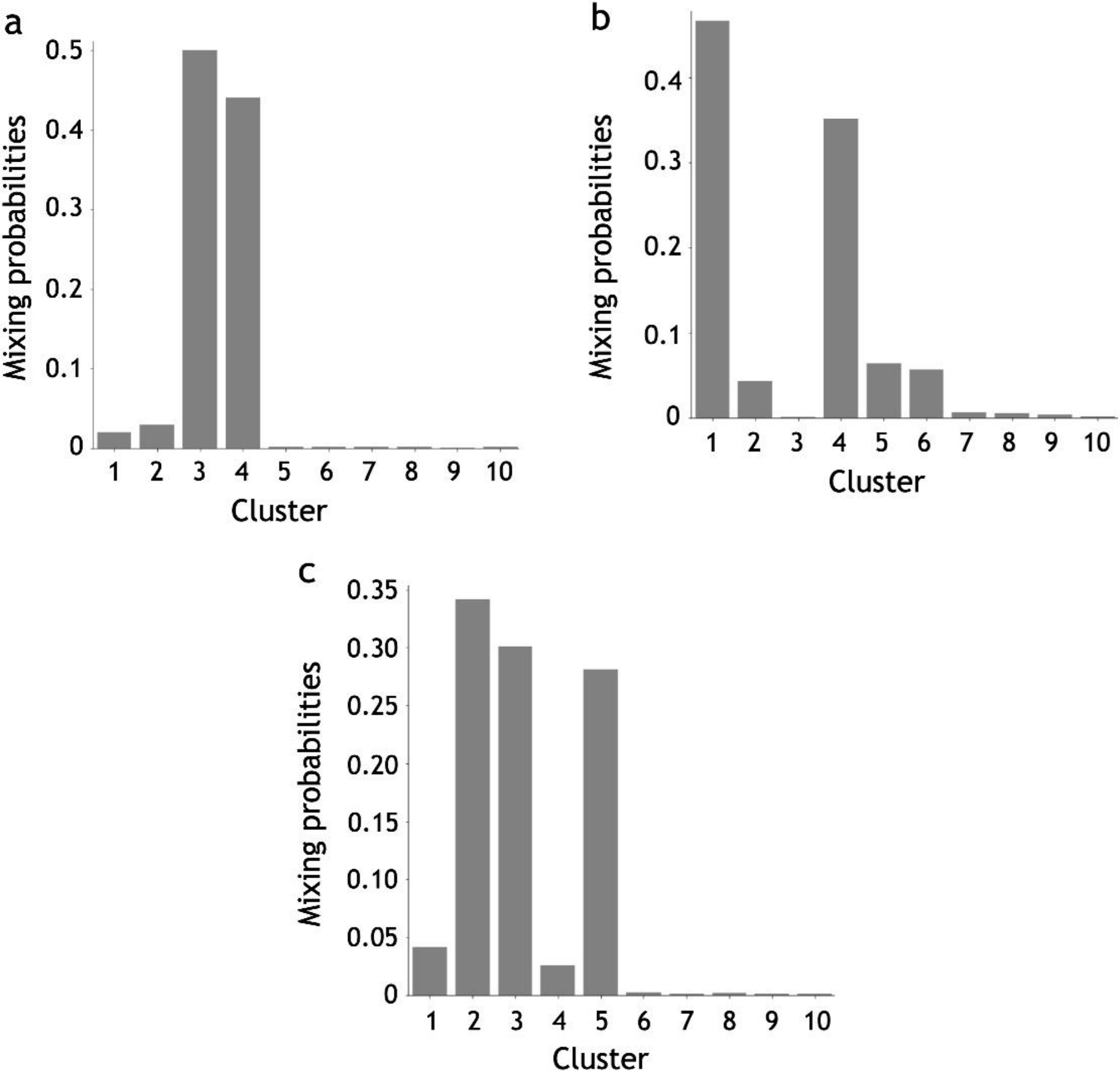

a
Number of selected species
500
400
300
200
100
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Probability
b
Number of selected features
30
25
20
15
10
5
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Probability

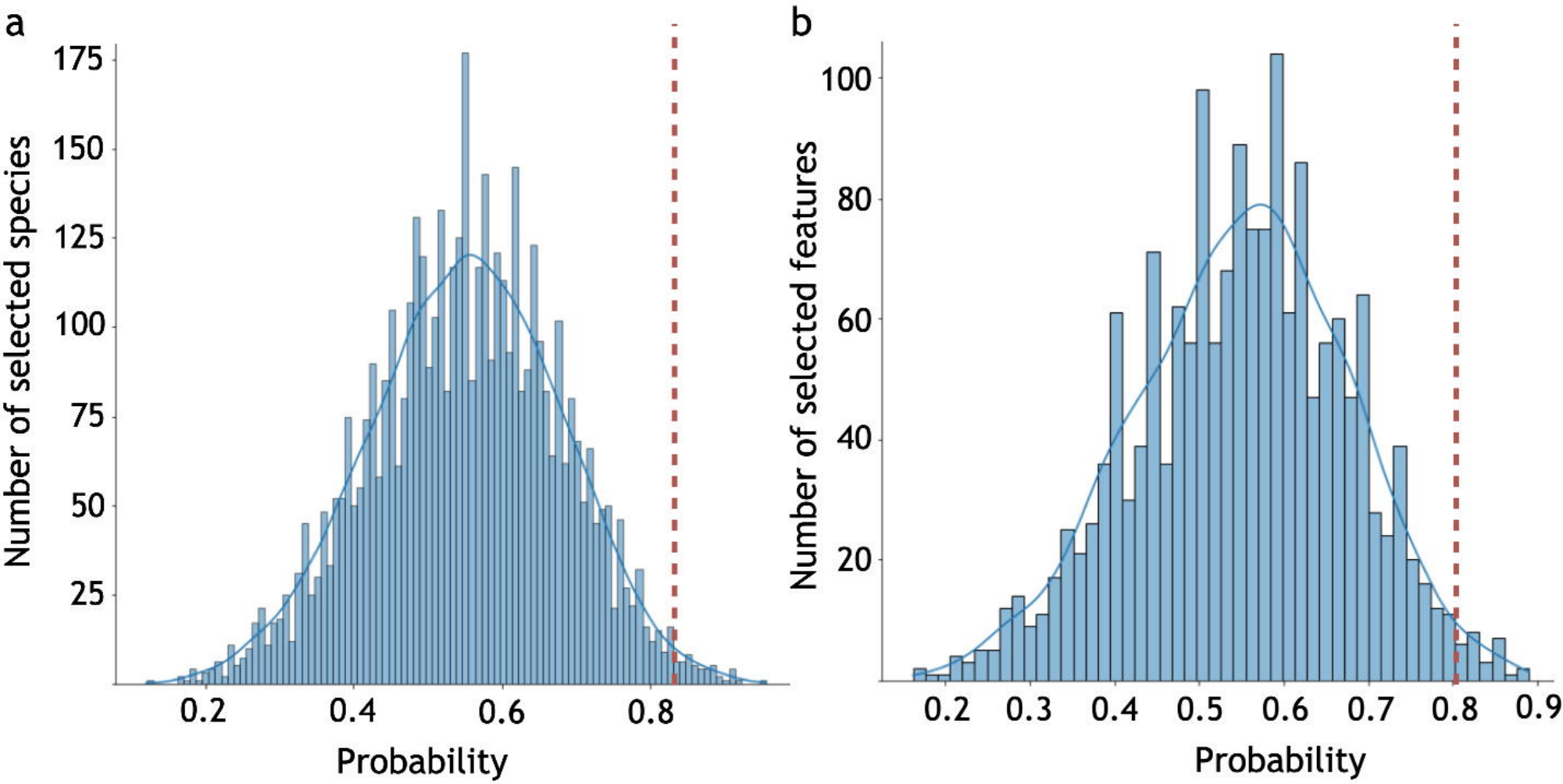

Vampirovibrionaceae
Diplorickettsiaceae
Solirubrobacteraceae
Nitrosomonadaceae
Moraxellaceae
Nitrosomonadaceae
Myxococcaceae
Gemmicoccaceae
Dongiaceae
Hyphomicrobiaceae
Beijerinckiaceae
Parachlamydiaceae
Chthoniobacteraceae
Pedosphaeraceae
Tepidisphaerales
Gemmataceae
Isosphaeraceae
Gemmatimonadaceae
Microscillaceae
Chitinophagaceae
Roseiflexaceae
Micromonosporaceae
Thermomonosporaceae
Deinococcaceae
Chthonomonadaceae
Roseiflexaceae
Anaerolineaceae
Ktedonobacteraceae
Leptotrichiaceae
Paenibacillaceae
Paenibacillaceae
0.5

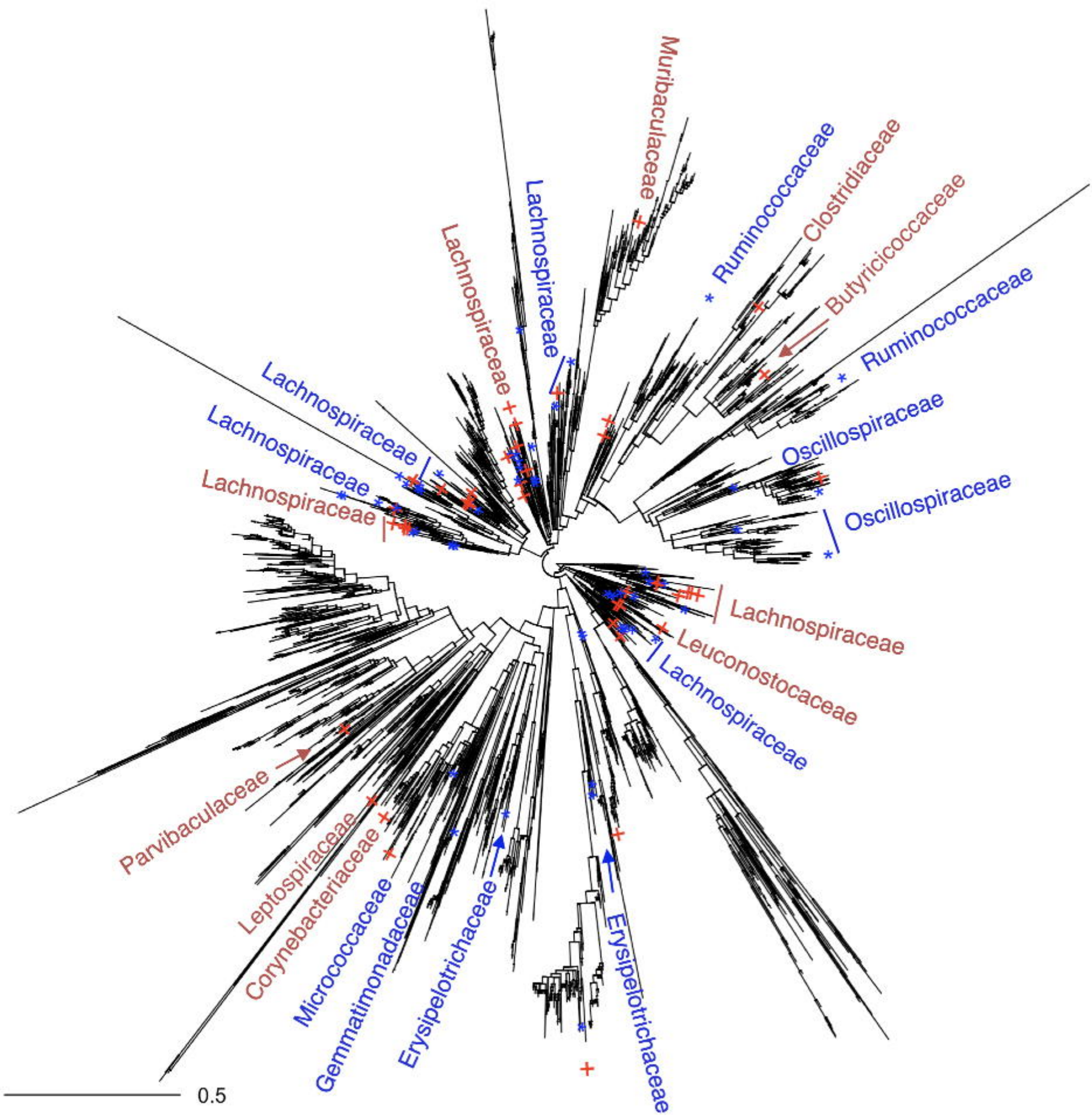


L-Carnosine
Robinin
Peltatoside
Hyperoside
L-Tyrosine
Cystathionine
3-Hydroxyanthranilic acid
L-Phenylalanine
L-Aspartic Acid
L-Glutamine
4-Methylumbelliferyl acetate
Pyridoxal Hydrochrolide
1-aminocyclopropane-1-carboxylic acid
Anthranilic Acid
Allantoin
L-Serine
Inosine
Urocanic Acid
Thymidine
Isorhamnetin-3-O-glucoside
0.2
0.1
0
-0.1
-0.2
Chloroflexi
Chloroflexi_1
Gammaproteobacteria
Simkaniaceae
Rosiflexaceae
Morawellaceae
Frankiaceae
Longimicrobiaceae
Armatimonadaceae
Nitrosomonadaceae
Oligoflexales
Hyphomicrobiaceae
Microlunatiaceae
Solirubrobacteraceae
Hyphomicrobiaceae
Leptotrichiaceae
Babeliaceae
Acidobacteriae
Acidobacteriae_1
Vicinamibacterales
Vicinamibacterales_1
Pedosphaeraceae
Sandaracinaceae
Entomoplasmatales
Haliangiaceae
Verrucomicrobiaceae
Ktedonobacteraceae
Ktedonobacteraceae_1
Chitinophagaceae_6
Chitinophagaceae_5
Chitinophagaceae_4
Chitinophagaceae_3
Chitinophagaceae_2
Chitinophagaceae_1
Chitinophagaceae
Dermatococcaceae
Geminicoccaceae
Thermomicrobiales
bacteria25
Anaerolineae
Gemmatimonadaceae
Gemmatimonadaceae_1
unknown
unknown_1
Babeliales
Babeliales_1
Peptostreptococcaceae
Parcubacteria
Holophagaceae
Paraphaeroidaceae
Ktedonobacteria
Chromobacteriaceae
Armatimonadota
Nocardioidaceae
Anaerolineaceae
Myxococcaceae
Rhizobiales_Incertae_Sedis
Diplorickettsiaceae
Isosphaeraceae
Diplorickettsiaceae
Magnetospirillaceae
Xanthobacteraceae


3,4-Dihydroxybenzoic Acid
Apigenin-7-O-glucoside
Naringenin-7-O-glucoside
L-Tryptophane
L-Tyrosine
2-Deoxycytidine
3,4-Dihydroxy-L-phenylalanine
Riboflavin 5'-Monophosphate
4-Hydroxy-3-methoxycinnamaldehyde
3-Hydroxyanthranilic Acid
Guanine
Adonitol
Kaempferol-3-O-glucoside
Shikimic Acid
Urocanic Acid
Glycine
L-Methionine
Guanosine
L-Carnosine
Rhoifolin
Eriodictyol
Caffeic Acid
N-Acetyl-D-Mannosamine
2-Aminoethylphosphonic Acid
Inosine
Sebacic Acid
Spermine
0.3
0.2
0.1
0
-0.1
-0.2
aenibacillaceae.3
Paenibacillaceae.2
Paenibacillaceae
Paenibacillaceae.1
Gemmatacea.2
Gemmatacea.1
Gemmatacea
Simkaniaceae
Sphingobacteriales
Leptotrichiaceae
Nitrosomonadaceae
Acidobacteriae
Acidobacteriae.1
Peptostreptococcaceae
Beijerinckiaceae
Weeksellaceae
Anaerolineae
Microscillaceae
Holophagae
Anaerolineaceae
Hymenobacteraceae
Rhizobiales_Incertae_Sedis
Pseudomonadaceae
Oligoflexia
Entomoplasmatales
Blastocatellia
Pedosphaeraceae
Parcubacteria
Mitochondria
Bdellovibrionaceae
Dongiaceae
Micromonosporaceae
Sphingomonadaceae
Solirubrobacterales
Solirubrobacterales.1
Babeliaceae
Chromobacteriaceae
Holosporaceae
Armatimonadaceae
Isosphaeraceae
Armatimonadota
Deinococcaceae
Vicinamibacterales
Vicinamibacterales.1
Chloroflexi
Chloroflexi.1
Nitrosporaceae
Streptomycetaceae
Thermomonosporaceae
unknown
unknown.1
Elsterales
Elsterales.1


a
Mixing probabilities
0.5
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10
Cluster
b
Mixing probabilities
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7 8 9 10
Cluster
c
Mixing probabilities
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
1 2 3 4 5 6 7 8 9 10
Cluster


a
Number of selected species
175
150
125
100
75
50
25
0.2
0.4
0.6
0.8
Probability
b
Number of selected features
100
80
60
40
20
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Probability


Muribaculaceae
Ruminococcaceae
Clostridiaceae
Butyricicoccaceae
Ruminococcaceae
Oscillospiraceae
Oscillospiraceae
Lachnospiraceae
Lachnospiraceae
Lachnospiraceae
Lachnospiraceae
Lachnospiraceae
Lachnospiraceae
Leuconostocaceae
Lachnospiraceae
Parvibaculaceae
Leptospiraceae
Corynebacteriaceae
Micrococcaceae
Gemmatimonadaceae
Erysipelotrichaceae
Erysipelotrichaceae
0.5