

# I-SVVS: Integrative stochastic variational variable selection to explore joint patterns of multi-omics microbiome data

*Tung Dang<sup>1,2</sup>, Yushiro Fuji<sup>3</sup>, Kie Kumaishi<sup>4</sup>, Erika Usui<sup>4</sup>, Shungo Kobori<sup>4</sup>, Takumi Sato<sup>4</sup>, Megumi Narukawa<sup>4</sup>, Yusuke Toda<sup>2</sup>, Kengo Sakurai<sup>2</sup>, Yuji Yamasaki<sup>5</sup>, Hisashi Tsujimoto<sup>5</sup>, Masami Hirai<sup>3</sup>, Yasunori Ichihashi<sup>4</sup> and Hiroyoshi Iwata<sup>2,\*</sup>*

<sup>1</sup> Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo, Tokyo, Japan

<sup>2</sup> Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo, Japan

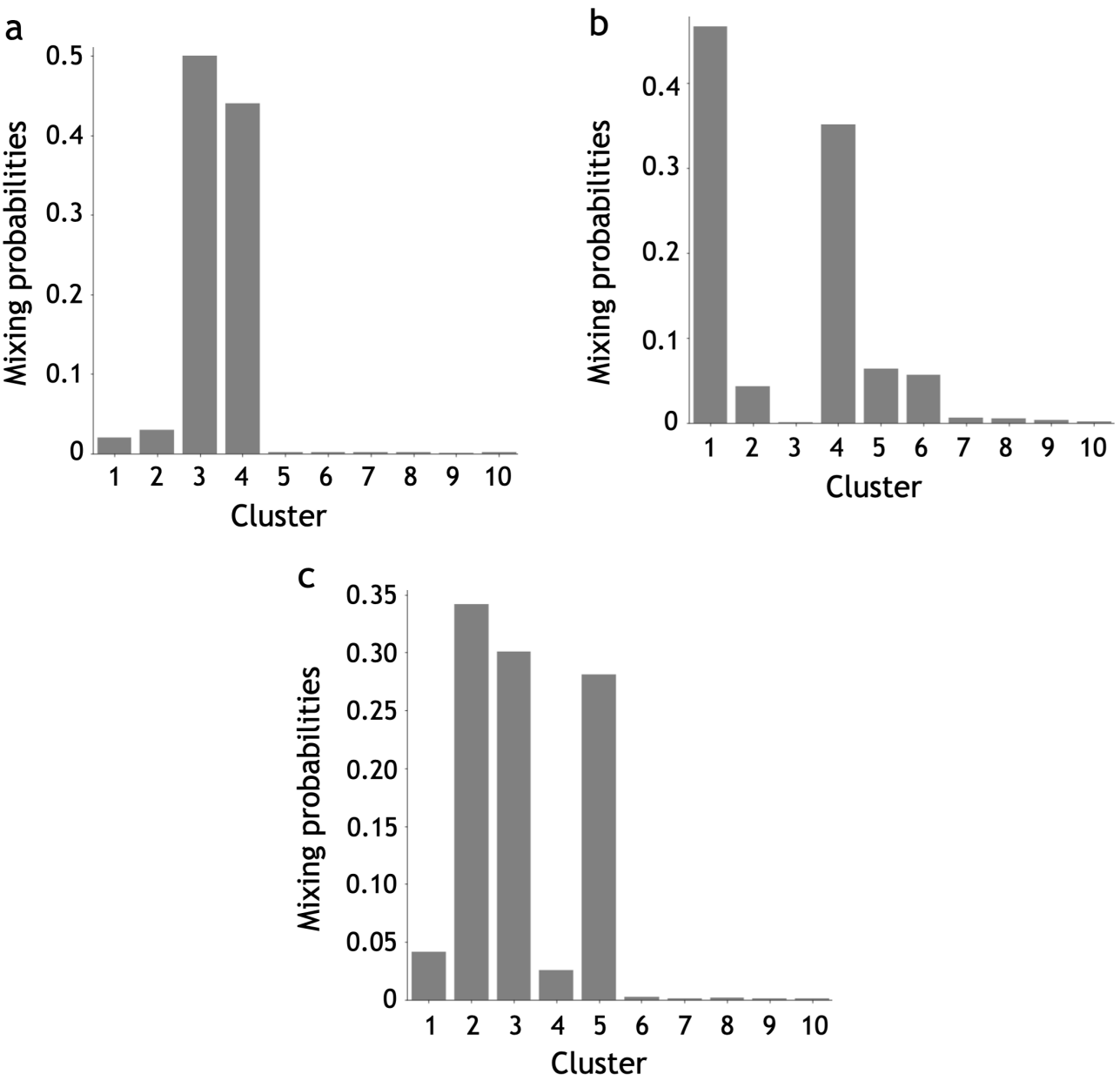
<sup>3</sup> RIKEN Center for Sustainable Resource Science, Tsurumi-ku, Yokohama, Japan

<sup>4</sup> RIKEN BioResource Research Center, Tsukuba, Ibaraki, Japan

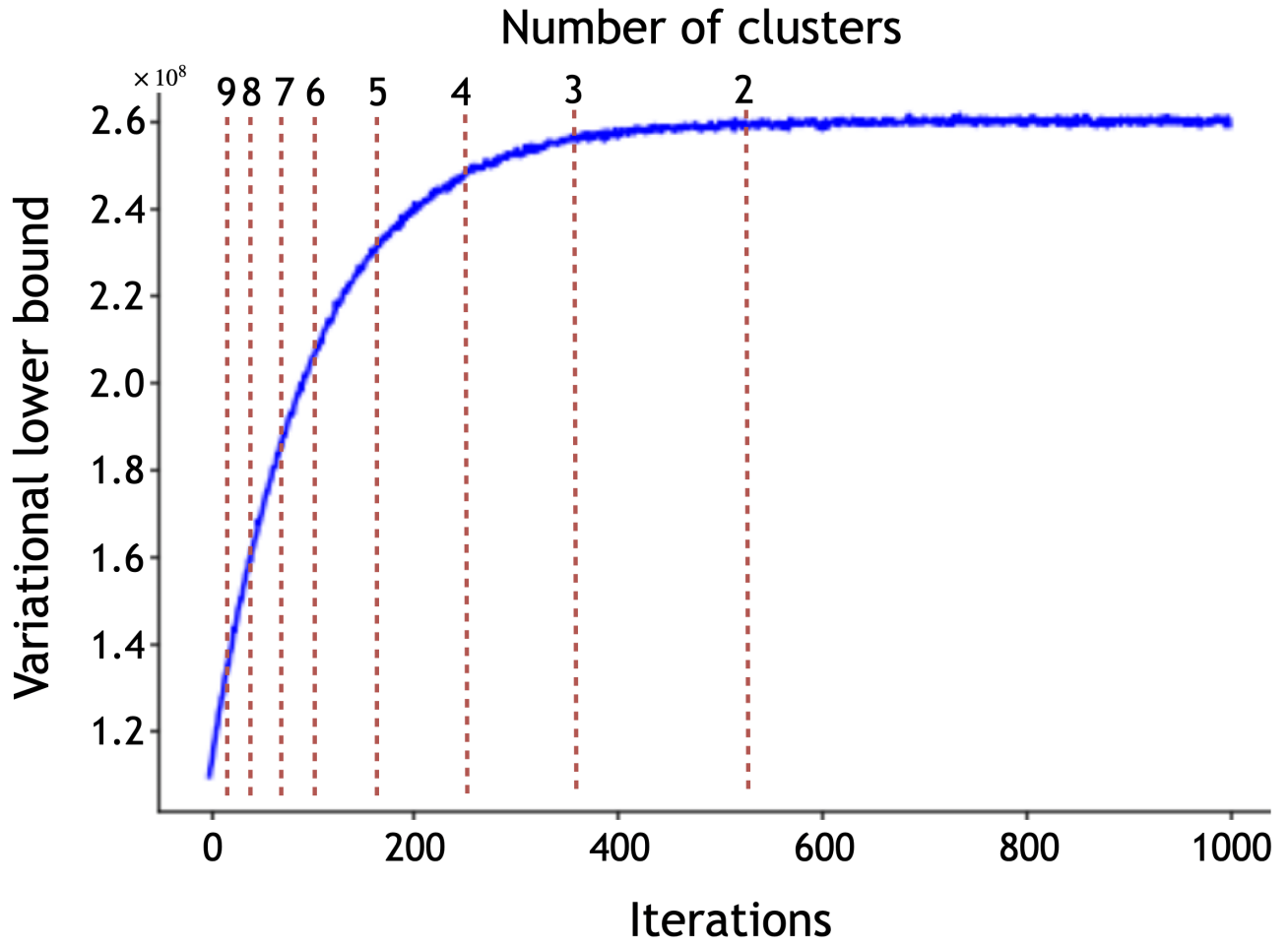
<sup>5</sup> Arid Land Research Center, Tottori University, Tottori, Japan

\* Corresponding author: hiroiwata@g.ecc.u-tokyo.ac.jp

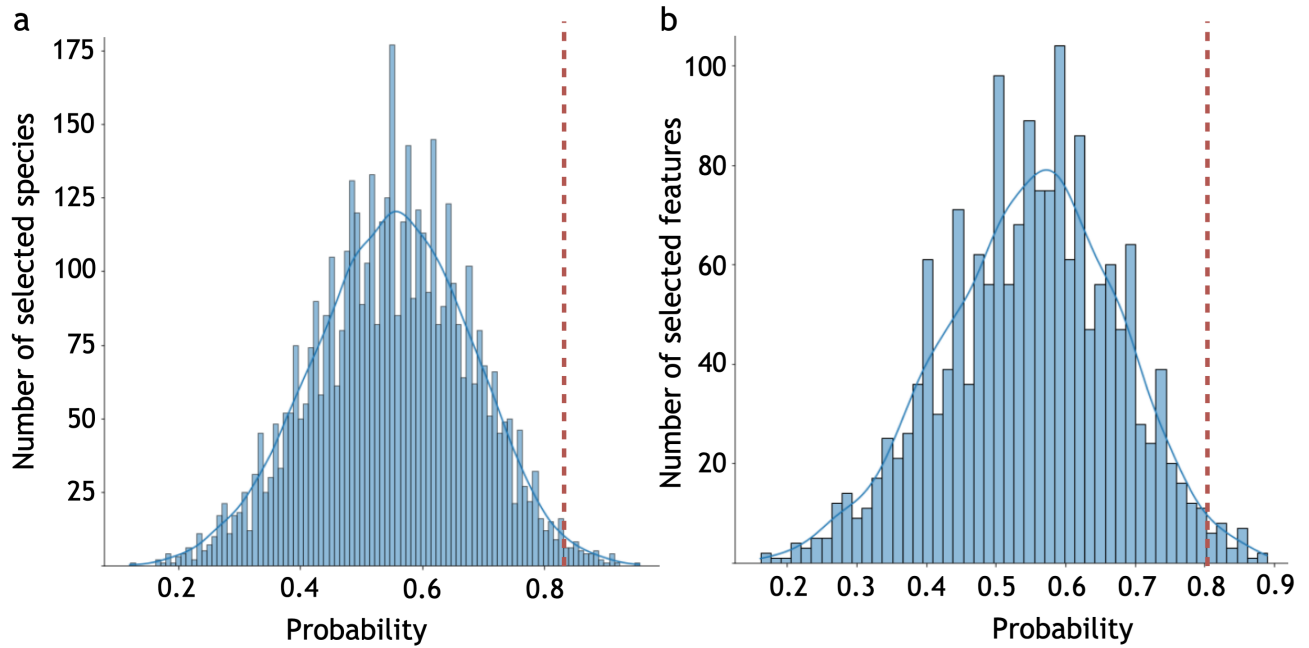
# 1 Supplementary Figures



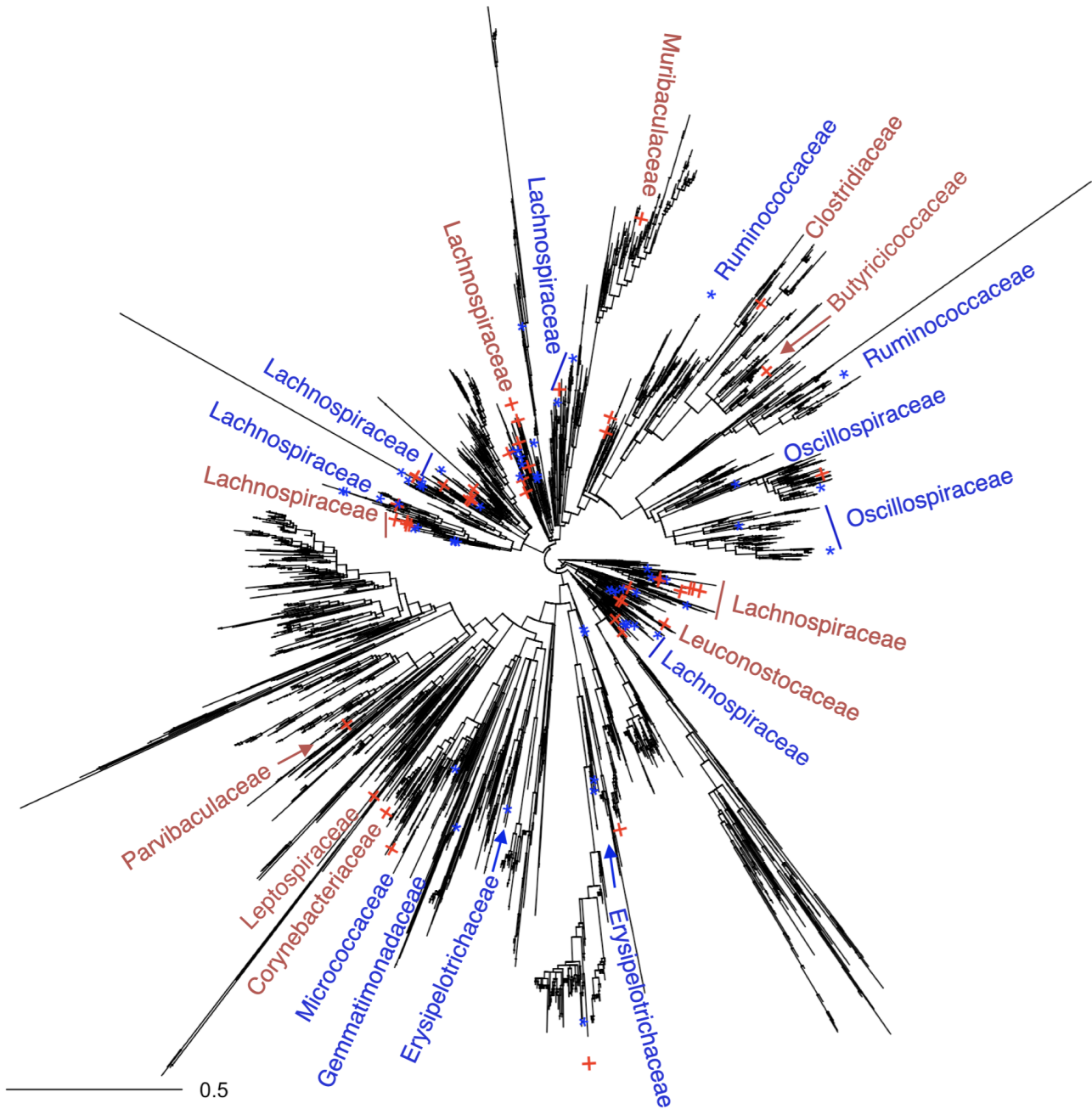
**Supplementary Figure S1:** The estimated values of the mixing coefficients. a. dataset A; b. dataset B; c. dataset C



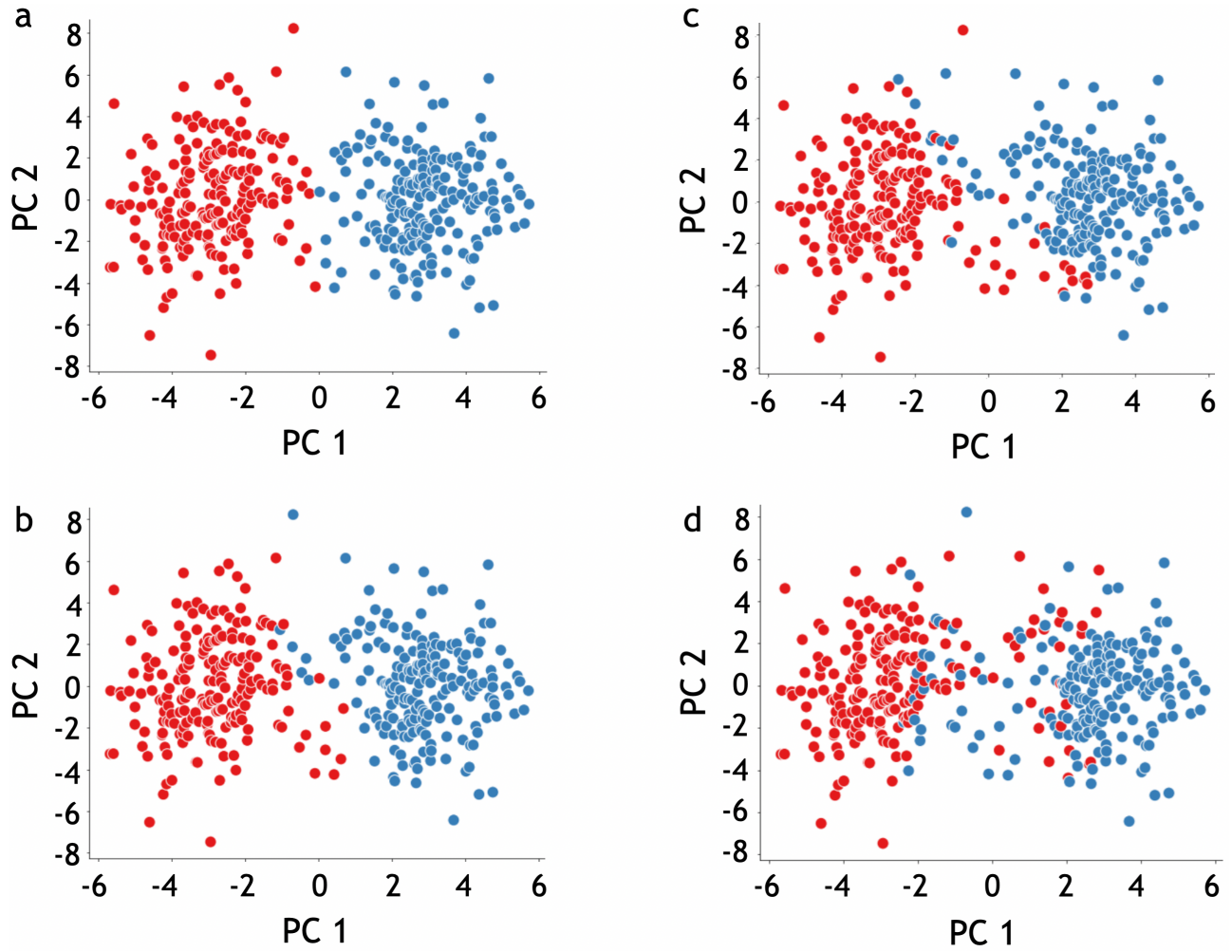
**Supplementary Figure S2:** Variational lower bound function values obtained by I-SVVS during iterations for dataset A. We take the setting of an initial number of the clusters as 10, as an example. The numbers on top of the vertical dashed line are the present number of clusters at the current iteration.



**Supplementary Figure S3:** Histogram of the average of  $\Phi_1$  and  $\Phi_2$  in dataset B. The dashed lines are bound to select microbiome species and metabolome profile features. a. microbiome data; b. metabolome data.



**Supplementary Figure S4:** Microbial species selected using the integrative stochastic variational variable selection (I-SVVS) approach and mapped on the phylogenetic tree for dataset B. Blue-colored stars symbols denote intermittent hypoxia and hypercapnia (IHH) cases and Red-colored plus denote air controls.



**Supplementary Figure S5:** Principal component analysis (PCA) plots of the second scenario with labels indicating predicted class using the three approaches and true group. a True labels. b Integrative stochastic variational variable selection (I-SVVS). c iCluster algorithm. d Clusternomics algorithm. Red-colored circles denote the cluster 1 and blue-colored circles denote cluster 2.

## 2 Mean-Field Variational Inference for a integration of Gaussian mixture model (GMM) and Dirichlet multinomial mixtures (DMM) Model

We expand specifically the variational lower bound equation as follows:

$$\begin{aligned}
\mathcal{L} = & \sum_{k=1}^{K_{max}} E_q \left[ \log \left( p \left( \Psi'_k | 1, \kappa \right) \right) \right] - \sum_{k=1}^{K_{max}} E_q \left[ \log \left( q \left( \Psi'_k | \vartheta_k, \vartheta'_k \right) \right) \right] \\
& + \sum_{i=1}^N \sum_{k=1}^{K_{max}} E_q \left[ \log p(W_i^k | \Psi'_1, \Psi'_2, \dots, \Psi'_{K_{max}}) \right] - \sum_{i=1}^N \sum_{k=1}^{K_{max}} E_q \left[ \log \left( q \left( W_i^k | g_i^k \right) \right) \right] \\
& + \sum_{t=1}^{T_{max}} E_q \left[ \log \left( p \left( \Pi'_{1t} | 1, \kappa^* \right) \right) \right] - \sum_{t=1}^{T_{max}} E_q \left[ \log \left( q \left( \Pi'_{1t} | \vartheta_t^*, \vartheta_t'^* \right) \right) \right] \\
& + \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} E_q \left[ \log p(Z_{1t}^k | \Pi'_{11}, \Pi'_{12}, \dots, \Pi'_{1T_{max}}) \right] - \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} E_q \left[ \log \left( q \left( Z_{1t}^k | l_{1t}^k \right) \right) \right] \\
& + \sum_{i=1}^N \sum_{j=1}^D E_q \left[ \log p(\Phi_{1ij} | \epsilon_{j1}, \epsilon_{j2}) \right] - \sum_{i=1}^N \sum_{j=1}^D E_q \left[ \log \left( q \left( \Phi_{1ij} | f_{ij} \right) \right) \right] \\
& + \sum_{j=1}^D E_q \left[ \log \left( p \left( \epsilon_j | \xi \right) \right) \right] - \sum_{j=1}^D E_q \left[ \log \left( q \left( \epsilon_j | \xi^* \right) \right) \right] \\
& + \sum_{j=1}^D \sum_{t=1}^{T_{max}} E_q \left[ \log \left( p \left( \alpha_{jt} | \lambda_{jt} \right) \right) \right] - \sum_{j=1}^D \sum_{t=1}^{T_{max}} E_q \left[ \log \left( q \left( \alpha_{jt} | \lambda_{jt}^* \right) \right) \right] \\
& + \sum_{j=1}^D E_q \left[ \log \left( p \left( \beta_j | \iota_j \right) \right) \right] - \sum_{j=1}^D E_q \left[ \log \left( q \left( \beta_j | \iota_j^* \right) \right) \right] \\
& + \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( p \left( \Pi'_{2t'} | 1, \kappa^+ \right) \right) \right] - \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( q \left( \Pi'_{2t'} | \vartheta_{t'}^+, \vartheta_{t'}'^+ \right) \right) \right] \\
& + \sum_{k=1}^{K_{max}} \sum_{t'=1}^{T'_{max}} E_q \left[ \log p(Z_{2t'}^k | \Pi'_{21}, \Pi'_{22}, \dots, \Pi'_{2T'_{max}}) \right] - \sum_{k=1}^{K_{max}} \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( q \left( Z_{2t'}^k | l_{2t'}^k \right) \right) \right] \\
& + \sum_{i=1}^N \sum_{j'=1}^{D'} E_q \left[ \log p(\Phi_{2ij'} | \epsilon'_{j'1}, \epsilon'_{j'2}) \right] - \sum_{i=1}^N \sum_{j'=1}^{D'} E_q \left[ \log \left( q \left( \Phi_{2ij'} | f'_{ij'} \right) \right) \right] \\
& + \sum_{j'=1}^{D'} E_q \left[ \log \left( p \left( \epsilon'_{j'} | \xi' \right) \right) \right] - \sum_{j'=1}^{D'} E_q \left[ \log \left( q \left( \epsilon'_{j'} | \xi'^* \right) \right) \right] \\
& + \sum_{j'=1}^{D'} \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( p \left( \mu_{j't'} | \varphi_{j't'}, \varsigma_{j't'} \right) \right) \right] - \sum_{j'=1}^{D'} \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( q \left( \mu_{j't'} | \varphi_{j't'}^*, \varsigma_{j't'}^* \right) \right) \right] \\
& + \sum_{j'=1}^{D'} \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( p \left( \delta_{j't'} | \varpi_{j't'}, \chi_{j't'} \right) \right) \right] - \sum_{j'=1}^{D'} \sum_{t'=1}^{T'_{max}} E_q \left[ \log \left( q \left( \delta_{j't'} | \varpi_{j't'}^*, \chi_{j't'}^* \right) \right) \right] \\
& + \sum_{j'=1}^{D'} E_q \left[ \log \left( p \left( \mu'_{j'} | \varphi'_{j'}, \varsigma'_{j'} \right) \right) \right] - \sum_{j'=1}^{D'} E_q \left[ \log \left( q \left( \mu'_{j'} | \varphi_{j'}^*, \varsigma_{j'}^* \right) \right) \right] \\
& + \sum_{j'=1}^{D'} E_q \left[ \log \left( p \left( \delta'_{j'} | \varpi'_{j'}, \chi'_{j'} \right) \right) \right] - \sum_{j'=1}^{D'} E_q \left[ \log \left( q \left( \delta'_{j'} | \varpi_{j'}^*, \chi_{j'}^* \right) \right) \right]
\end{aligned} \tag{1}$$

To compute the variational expectations  $E_q[\cdot]$  in equation (1), we use the properties of exponential family distribution. If variational distributions for  $q(\alpha_{jt} | \lambda_{jt})$  and  $q(\beta_j | \iota_j)$  are Dirichlet distributions, then the exponential family representations are given by

$$\begin{aligned}
q(\alpha_t | \lambda_t) &= \exp \left[ \left( \sum_{j=1}^D (\lambda_j^t - 1) \log(\alpha_j^t) \right) + \log \Gamma \left( \sum_{j=1}^D \lambda_j^t \right) - \sum_{j=1}^D (\log \Gamma(\lambda_j^t)) \right] \\
q(\beta_j | \iota_j) &= \exp \left[ \left( \sum_{j=1}^D (\iota_j - 1) \log(\beta_j) \right) + \log \Gamma \left( \sum_{j=1}^D \iota_j \right) - \sum_{j=1}^D (\log \Gamma(\iota_j)) \right]
\end{aligned}$$

So the natural parameters and sufficient statistics of the Dirichlet distributions for  $\alpha$  and  $\beta$  are  $\eta_{\alpha_{jt}} = \lambda_{jt} - 1$ ,  $T(\alpha_{jt}) = \log(\alpha_{jt})$  and  $\eta_{\beta_j} = \iota_j - 1$ ,  $T(\beta_j) = \log(\beta_j)$ , respectively. The approximation expectations are

$$\begin{aligned}
E[\alpha_{jt}] &= \frac{\lambda_{jt}}{\sum_{j'=1}^D \lambda_{jt'}} & E[\beta_j] &= \frac{\iota_j}{\sum_{j'=1}^S \iota_{j'}} \\
E[\log(\alpha_{jt})] &= \psi(\lambda_{jt}) - \psi \left( \sum_{j'=1}^D \lambda_{jt'} \right) & E[\log(\beta_j)] &= \psi(\iota_j) - \psi \left( \sum_{j'=1}^S \iota_{j'} \right)
\end{aligned}$$

where  $\psi(\cdot)$  is the digamma function.

Similarity, if variational distributions for  $q\left(\mu_{j't'}|\varphi_{j't'},\varsigma_{j't'}\right)$  and  $q\left(\delta_{j't'}|\varpi_{j't'},\chi_{j't'}\right)$  are norm and Wishart distributions, the approximation expectations are

$$\begin{aligned} \mathbb{E}\left[\mu_{j't'}\right] &= \varphi_{j't'} & \mathbb{E}\left[\delta_{j't'}\right] &= \varpi_{j't'}\chi_{j't'} \\ \mathbb{E}\left[\log\left(\delta_{j't'}\right)\right] &= \psi\left(\frac{\chi_{j't'}}{2}\right)D' \log 2 \log\left(\varpi_{j't'}\right) \end{aligned}$$

where  $\psi(\cdot)$  is the digamma function.

### 3 Stochastic Optimization of the Variational Parameters for global clustering

With a truncated stick-breaking representation of the hierarchical Dirichlet process (HDP) [1], variational factors for the stick lengths  $q\left(\Psi'_k|\vartheta_k,\vartheta'_k\right)$  are Beta distributions and the assignment variable  $W_i^k = \mathbb{I}[W_i = k]$  for the  $i^{th}$  sample allocation is governed by a multinomial distribution indexed by a variational parameter  $g_{ik}$ . The computation of the approximate expectation for the parameters of truncated stick-breaking representation has been considered carefully by [2]. The results of variational expectation for both variables are as follows:

$$\begin{aligned} q(W_i = k) &= g_i^k \\ q(W_i > k) &= \sum_{k'=k+1}^{K_{\max}} g_i^{k'} \\ \mathbb{E}_q\left[\log\left(\Psi'_k\right)\right] &= \psi(\vartheta_k) - \psi(\vartheta_k + \vartheta'_k) \\ \mathbb{E}_q\left[\log\left(1 - \Psi'_k\right)\right] &= \psi(\vartheta'_k) - \psi(\vartheta_k + \vartheta'_k) \end{aligned}$$

Following the principles of the variational inference [1, 2], we consider the derivation of the update equation of variational parameters of  $W_i^k$  for only one variable by fixing the others' distribution. The optimal coordinate updates are derived for these variational parameters. The log of the optimized factor to the posterior distribution of  $W_i^k$  is

$$\log Q^*(W) = E_{Z,\Phi,\Psi,\epsilon,\epsilon',\alpha,\beta,\mu,\delta}\left[\log p\left(W, Z, \Phi, \Psi, \epsilon, \epsilon', \alpha, \beta, \mu, \delta\right)\right] + const = \sum_{i=1}^N \sum_{k=1}^K W_i^k \log(g_i^k) + const$$

$$\begin{aligned} \log g_i^k &= \\ T_{max} \sum_{t=1}^D \mathbb{E}_Q[\Phi_{1ij}] \mathbb{E}_Q[Z_{1kt}] &\left( \mathbb{E}_Q\left[\log\left(\frac{\Gamma\left(\sum_{j=1}^D \alpha_{tj}\right)}{\Gamma\left(\sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \alpha_{tj}\right)}\right)\right] + \sum_{j=1}^D \mathbb{E}_Q\left[\log\left(\frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})}\right)\right] + const \right) + \\ T'_{max} \sum_{t'=1}^{D'} \mathbb{E}_Q[\Phi_{2ij'}] \mathbb{E}_Q[Z_{2kt'}] &\left[ \frac{1}{2} \mathbb{E}_Q[\log(\delta_{t'j'})] + \left(\frac{-1}{2} (X_{2ij'} - \mathbb{E}_Q[\mu_{t'j'}])^T \mathbb{E}_Q[\delta_{t'j'}] (X_{2ij'} - \mathbb{E}_Q[\mu_{t'j'}])\right) \right] \\ + \mathbb{E}_Q[\log(\Psi'_k)] &+ \sum_{k'=1}^{k-1} \mathbb{E}_Q[\log(1 - \Psi'_{k'})] \end{aligned} \quad (2)$$

The expected logarithm of the functions  $\mathbb{E}_Q\left[\log\left(\frac{\Gamma(\sum_{j=1}^D \alpha_{tj})}{\Gamma(\sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \alpha_{tj})}\right)\right]$  and  $\mathbb{E}_Q\left[\log\left(\frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})}\right)\right]$  in equation (2) have not the closed form. Thus, the calculation of these equations are analytically intractable. We need a closed-form expression to use the standard form of the variational inference or the stochastic variational inference [2]. In order to overcome this problem, we adopt the Taylor solutions in our previous paper [3]. The variational parameters are updated based on first-order Taylor expansion as follows:

$$\begin{aligned} \log\left(\frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})}\right) &\geq \log\left(\frac{\Gamma(X_{1ij} + \overline{\alpha}_{tj})}{\Gamma(\overline{\alpha}_{tj})}\right) + \frac{\partial F(\alpha_{tj})}{\partial \alpha_{tj}} \frac{\partial \alpha_{tj}}{\partial \log(\alpha_{tj})} \Big|_{\alpha_{tj} = \overline{\alpha}_{tj}} (\log(\alpha_{tj}) - \log(\overline{\alpha}_{tj})) \\ \mathbb{E}_Q\left[\log\left(\frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})}\right)\right] &\geq \log\left(\frac{\Gamma(X_{1ij} + \overline{\alpha}_{tj})}{\Gamma(\overline{\alpha}_{tj})}\right) + \overline{\alpha}_{tj} [\psi(\overline{\alpha}_{tj} + X_{1ij}) - \psi(\overline{\alpha}_{tj})] (\mathbb{E}_Q[\log(\alpha_{tj})] - \log(\overline{\alpha}_{tj})) \\ &\geq \log\left(\frac{\Gamma(X_{1ij} + \overline{\alpha}_{tj})}{\Gamma(\overline{\alpha}_{tj})}\right) + \overline{\alpha}_{tj} [\psi(\overline{\alpha}_{tj} + X_{1ij}) - \psi(\overline{\alpha}_{tj})] (\psi(\lambda_{kj}^*) - \psi(\sum_{j=1}^D \lambda_{tj}^*) - \log(\overline{\alpha}_{tj})) \\ \mathbb{E}_Q\left[\log\left(\frac{\Gamma(\sum_{j=1}^D \alpha_{tj})}{\Gamma(\sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \alpha_{tj})}\right)\right] &\geq \log\left(\frac{\Gamma(\sum_{j=1}^D \overline{\alpha}_{tj})}{\Gamma(\sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha}_{tj})}\right) \\ &+ \sum_{j=1}^D [\psi(\sum_{j=1}^D \overline{\alpha}_{tj}) - \psi(\sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha}_{tj})] \overline{\alpha}_{tj} (\psi(\lambda_{tj}^*) - \psi(\sum_{j=1}^D \lambda_{tj}^*) - \log(\overline{\alpha}_{tj})) \end{aligned}$$

Based on this approach, the variational parameters of  $W_{ik}$  are updated as follow:

$$\begin{aligned}
\log g_i^k = & \sum_{t=1}^{T_{max}} \sum_{j=1}^D f_{ij} l_{1tk} \log \left( \frac{\Gamma \left( \sum_{j=1}^D \overline{\alpha_{tj}} \right)}{\Gamma \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha_{tj}} \right)} \right) \\
& + \sum_{t=1}^{T_{max}} \sum_{j=1}^D f_{ij} l_{1tk} \left[ \psi \left( \sum_{j=1}^D \overline{\alpha_{tj}} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha_{tj}} \right) \right] \overline{\alpha_{tj}} \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\overline{\alpha_{tj}}) \right) \\
& + \sum_{t=1}^{T_{max}} \sum_{j=1}^D f_{ij} l_{1tk} \left[ \log \left( \frac{\Gamma(X_{1ij} + \overline{\alpha_{tj}})}{\Gamma(\overline{\alpha_{tj}})} \right) + \overline{\alpha_{tj}} [\psi(\overline{\alpha_{tj}} + X_{1ij}) - \psi(\overline{\alpha_{tj}})] \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\overline{\alpha_{tj}}) \right) \right] \\
& + \sum_{t'=1}^{T'_{max}} \sum_{j'=1}^{D'} f'_{ij'} l_{2kt'} \left( \frac{1}{2} \left[ \psi \left( \frac{X_{t'j'}^* + 1 - j'}{2} \right) + D' \log 2 + \log |\varpi_{t'j'}^*| \right] + const \right) \\
& + \sum_{t'=1}^{T'_{max}} \sum_{j'=1}^{D'} f'_{ij'} l_{2kt'} \left( \frac{-1}{2} \left[ \chi_{t'j'}^* (X_{2ij'} - \varphi_{t'j'}^*) \varpi_{t'j'}^* (X_{2ij} - \varphi_{t'j'}^*)^T + \frac{D'}{\varsigma_{t'j'}^*} \right] \right) \\
& + \psi(\vartheta_k) - \psi(\vartheta_k + \vartheta'_k) + \sum_{k'=1}^{k-1} \psi(\vartheta'_{k'}) - \psi(\vartheta_{k'} + \vartheta'_{k'})
\end{aligned} \tag{3}$$

Similarly, the optimized solution to the posterior distribution of unit-length sticks  $\Psi'_{ik}$

$$\begin{aligned}
\log Q^* \left( \Psi'_{ik} \right) = & (1 - 1) \log \left( \Psi'_{ik} \right) + (\kappa - 1) \log \left( 1 - \Psi'_{ik} \right) \\
& + \sum_{i=1}^N E_Q [W_{ik}] \log \left( \Psi'_{ik} \right) + \sum_{i=1}^N \sum_{k'=k+1}^{K_{max}} E_Q [W_{ik}] \log \left( 1 - P s'_{ik} \right) + const
\end{aligned} \tag{4}$$

which has the logarithmic form of the Beta distribution and the corresponding variational distributions of breaking proportions  $q \left( \Psi'_{ik} \right)$  are considered to be the Beta distribution. Then, the standard conditions are satisfied for a closed form coordinate update for local parameters. Based on the stochastic natural gradient, the variational parameters of  $\Psi'_{ik}$  are updated by computing variational expectations in equation (11) [4, 2],

$$\begin{aligned}
(\vartheta_t)^{(t+1)} = & (1 - \rho^{(t)}) (\vartheta_t)^{(t)} + \rho^{(t)} \left\{ 1 + \sum_{i=1}^N g_{ik} \right\} \\
(\vartheta'_t)^{(t+1)} = & (1 - \rho^{(t)}) (\vartheta'_t)^{(t)} + \rho^{(t)} \left\{ \kappa + \sum_{i=1}^N \sum_{k'=k+1}^{K_{max}} g_{ik} \right\}
\end{aligned}$$

## 4 Stochastic Optimization of the Variational Parameters for Dirichlet multinomial mixtures (DMM) Model

### 4.1 Compute the variational function parameters for $Z_{1kt}$

Following the principles of the variational inference [4, 2], variational parameters of  $Z_{1kt}$  is the local parameters, we consider the derivation of the update equation for only one variable by fixing the others' distribution. The optimal coordinate updates are derived for local parameters. The log of the optimized factor to the posterior distribution of  $Z_{1kt}$  is

$$\begin{aligned}
\log Q^* (Z_1) = & E_{W, \Phi_1, \Pi_1, \epsilon, \alpha, \beta} [\log p(W, Z_1, \Phi_1, \Pi_1, \epsilon, \alpha, \beta)] + const = \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} Z_{1kt} \log(l_{1kt}) + const \\
\log l_{1tk} = & \sum_{i=1}^N \sum_{j=1}^D E_Q [W_{ik}] E_Q [\Phi_{1ij}] \left( E_Q \left[ \log \left( \frac{\Gamma \left( \sum_{j=1}^D \alpha_{tj} \right)}{\Gamma \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \alpha_{tj} \right)} \right) \right] + \sum_{j=1}^D E_Q \left[ \log \left( \frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})} \right) \right] + const \right) \\
& + E_Q [\log(\Pi_{1t})] + \sum_{t'=1}^{t-1} E_Q [\log(1 - \Pi_{1t'})]
\end{aligned} \tag{5}$$

$$E_Q [Z_{1kt}] = \exp \{ \log(l_{1kt}) \}$$

The variational parameters of  $Z_{ik}$  are updated as follow:



$$\begin{aligned}
\log l_{1kt} &= \sum_{i=1}^N \sum_{j=1}^D g_{ik} f_{ij} \log \left( \frac{\Gamma \left( \sum_{j=1}^D \bar{\alpha}_{tj} \right)}{\Gamma \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \bar{\alpha}_{tj} \right)} \right) \\
&+ \sum_{i=1}^N \sum_{j=1}^D g_{ik} f_{ij} \left[ \psi \left( \sum_{j=1}^D \bar{\alpha}_{tj} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \bar{\alpha}_{tj} \right) \right] \bar{\alpha}_{tj} \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\bar{\alpha}_{tj}) \right) \\
&+ \sum_{i=1}^N \sum_{j=1}^D g_{ik} f_{ij} \left[ \log \left( \frac{\Gamma(X_{1ij} + \bar{\alpha}_{tj})}{\Gamma(\bar{\alpha}_{tj})} \right) + \bar{\alpha}_{tj} [\psi(\bar{\alpha}_{tj} + X_{1ij}) - \psi(\bar{\alpha}_{tj})] \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\bar{\alpha}_{tj}) \right) \right] \\
&+ \psi(\vartheta_t^*) - \psi(\vartheta_t^* + \vartheta_t^{*'}) + \sum_{t'=1}^{t-1} \psi(\vartheta_{t'}^*) - \psi(\vartheta_{t'}^* + \vartheta_{t'}^{*'})
\end{aligned} \tag{6}$$

## 4.2 Compute the variational function parameters of indicator variables for microbiome selection

Similarity, the optimized solution to the posterior distribution of variable selection  $\Phi_{1ij}$

$$\begin{aligned}
\log Q^*(\Phi_1) &= E_{W, Z_1, \Pi_1, \epsilon, \alpha, \beta} [\log p(W, Z_1, \Phi_1, \Pi_1, \epsilon, \alpha, \beta)] + \text{const} = \sum_{i=1}^N \sum_{j=1}^S \Phi_{1ij} \log(f_{ij}) + \text{const} \\
\log f_{ij}^{\Phi_{1ij}} &= \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} g_{ik} l_{1tk} \log \left( \frac{\Gamma \left( \sum_{j=1}^D \bar{\alpha}_{tj} \right)}{\Gamma \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^S \bar{\alpha}_{tj} \right)} \right) \\
&+ \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} g_{ik} l_{1tk} \left[ \psi \left( \sum_{j=1}^D \bar{\alpha}_{tj} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^S \bar{\alpha}_{tj} \right) \right] \bar{\alpha}_{tj} \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\bar{\alpha}_{tj}) \right) \\
&+ \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} g_{ik} l_{1tk} \left[ \log \left( \frac{\Gamma(X_{1ij} + \bar{\alpha}_{tj})}{\Gamma(\bar{\alpha}_{tj})} \right) + \bar{\alpha}_{tj} [\psi(\bar{\alpha}_{tj} + X_{1ij}) - \psi(\bar{\alpha}_{tj})] \left( \psi(\lambda_{tj}^*) - \psi \left( \sum_{j=1}^D \lambda_{tj}^* \right) - \log(\bar{\alpha}_{tj}) \right) \right] \\
&+ [\psi(\xi_1^*) - \psi(\xi_1^* + \xi_2^*)]
\end{aligned} \tag{7}$$

$$\begin{aligned}
\log f_{ij}^{1-\Phi_{1ij}} &= \log \left( \frac{\Gamma \left( \sum_{j=1}^D \bar{\beta}_j \right)}{\Gamma \left( \sum_{j=1}^S X_{1ij} + \sum_{j=1}^D \bar{\beta}_j \right)} \right) \\
&+ \sum_{j=1}^S \left[ \psi \left( \sum_{j=1}^D \bar{\beta}_j \right) - \psi \left( \sum_{j=1}^S X_{1ij} + \sum_{j=1}^D \bar{\beta}_j \right) \right] \bar{\beta}_j \left( \psi(\iota_j^*) - \psi \left( \sum_{j=1}^D \iota_j^* \right) - \log(\bar{\beta}_j) \right) \\
&+ \log \left( \frac{\Gamma(X_{1ij} + \bar{\beta}_j)}{\Gamma(\bar{\beta}_j)} \right) + \bar{\beta}_j [\psi(\bar{\beta}_j + X_{1ij}) - \psi(\bar{\beta}_j)] \left( \psi(\iota_j^*) - \psi \left( \sum_{j=1}^D \iota_j^* \right) - \log(\bar{\beta}_j) \right) \\
&+ [\psi(\xi_2^*) - \psi(\xi_1^* + \xi_2^*)]
\end{aligned} \tag{8}$$

## 4.3 Updating variational parameters of $\alpha$

Based on the principal framework of the stochastic variational inference [2], we consider parameters of  $\alpha$  as global parameters. These parameters are updated by a stochastic gradient step and noisy estimations of the natural gradient of the variational objective with respect to  $\alpha_{tj}$ . Following the computational method of the natural gradient (Hoffman et al. 2013), we compute the natural gradient of equation (1) with respect to the global variational parameters of  $\alpha_{tj}$ .

We consider Dirichlet distribution as a prior distribution with concentration parameters  $\lambda_{tj}$ , so a conditional distribution of profile given the observation data has form of the Dirichlet distribution. We consider the representation of the exponential family,

$$p(\alpha|W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta) = h(\alpha) \exp \left( \eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)^T t(\alpha) - a(\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)) \right)$$

where:  $h(\cdot)$  is the base measure;  $a(\cdot)$  is the log-normalize;  $\eta(\cdot)$  is the natural parameter;  $t(\cdot)$  is the sufficient statistics.

As above assumption of the variational parameters, we set  $q(\alpha|\lambda^*)$  to be Dirichlet distribution as the complete conditional distributions. So we get

$$q(\alpha|\lambda^*) = h(\alpha) \exp \left( (\lambda^*)^T t(\pi) - a(\lambda^*) \right)$$

We consider the lower bound for only  $\alpha$ ,

$$\begin{aligned}\mathcal{L}(\alpha) &= \mathbb{E}_Q [\log p(\alpha|W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)] - \mathbb{E}_Q [q(\alpha|\lambda^*)] \\ &= \mathbb{E}_Q \left[ \log \{h(\alpha)\} + \eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)^T t(\alpha) - a(\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)) - \log \{h(\alpha)\} - (\lambda^*)^T t(\alpha) + a(\lambda^*) \right] \\ &= \mathbb{E}_Q \left[ \eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)^T t(\alpha) - a(\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)) - (\lambda^*)^T t(\alpha) + a(\lambda^*) \right] \\ &= \mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)]^T [\nabla_{\lambda^*} \{a(\lambda^*)\}] - a(\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)) - (\lambda^*)^T [\nabla_{\lambda^*} \{a(\lambda^*)\}] + a(\lambda^*)\end{aligned}$$

where, the expected value of the sufficient statistics is the gradient of log normalizer  $\mathbb{E}_Q [t(\alpha)] = \nabla_{\lambda^*} \{a(\lambda^*)\}$ . If the classical principal of the gradient method for maximization is used directly to find a maximum of  $\mathcal{L}(\alpha)$  based on taking step of size  $\rho$  in direction of the gradient, the optimized results is

$$\lambda_{(t+1)}^* = \lambda_{(t)}^* + \rho \nabla_{(\lambda^*)} L(\alpha) = \lambda_{(t)}^* + \rho^{(t)} \left[ \nabla_{(\lambda^*)}^2 \{a(\lambda^*)\} \right] \left\{ \mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)]^T - (\lambda^*)^T \right\}$$

Following [2], by premultiplying the gradient by the inverse Fisher information  $G(\lambda^*)$  and apply the stochastic natural gradient of the variational objective with respect to  $\alpha$ , we get

$$\begin{aligned}G(\lambda^*) &= E_{\lambda^*} \left[ (\nabla_{(\lambda^*)} \log q(\alpha|\lambda^*)) (\nabla_{(\lambda^*)} \log q(\alpha|\lambda^*))^T \right] = \nabla_{(\lambda^*)}^2 \{a(\lambda^*)\} \\ \widehat{\nabla_{(\lambda^*)} L(\alpha)} &= \{G(\lambda^*)\}^{-1} \nabla_{(\lambda^*)} L(\alpha) = \{\mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)] - \lambda^*\}\end{aligned}$$

$$\begin{aligned}\lambda_{(t+1)}^* &= \lambda_{(t)}^* + \rho \widehat{\nabla_{(\lambda^*)} L(\alpha)} = \lambda_{(t)}^* + \rho \left\{ \mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)] - \lambda_{(t)}^* \right\} \\ &= (1 - \rho) \lambda_{(t)}^* + \rho \{\mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)]\}\end{aligned} \quad (9)$$

Based traditional variational inference, we get the conditional distribution of  $\alpha$

$$\begin{aligned}\log p(\alpha|W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta) &= \sum_{i=1}^N \sum_{j=1}^D \sum_{k=1}^{K_{max}} \sum_{t=1}^{T_{max}} \log \left( \frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})} \right) \times \log \left( \frac{\Gamma(X_{1ij} + \alpha_{tj})}{\Gamma(\alpha_{tj})} \right) \times \mathbb{E}_Q [W_{ik}] \mathbb{E}_Q [Z_{1kt}] \mathbb{E}_Q [\Phi_{1ij}] \\ &+ \log(\alpha_{tj}) (\lambda_{tj} - 1) + const\end{aligned}$$

$$\begin{aligned}\mathbb{E}_Q [\eta(W, Z_1, \Phi_1, \Pi_1, \epsilon, \beta)] &= \sum_{i=1}^N \sum_{k=1}^{K_{max}} \mathbb{E}_Q [W_{ik}] \mathbb{E}_Q [Z_{1kt}] \mathbb{E}_Q [\Phi_{1ij}] \overline{\alpha_{kj}} \times \\ &\left[ \psi \left( \sum_{j=1}^D \overline{\alpha_{tj}} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha_{tj}} \right) + \psi(\overline{\alpha_{tj}} + X_{1ij}) - \psi(\overline{\alpha_{tj}}) \right]\end{aligned} \quad (10)$$

By substituting equation (10) into equation (9), the final result of updated variational parameters is

$$\begin{aligned}(\lambda_{tj}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\lambda_{tj}^*)^{(t)} \\ &+ \rho^{(t)} \left\{ \lambda_{tj} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{1kt} f_{ij} \overline{\alpha_{tj}} \left[ \psi \left( \sum_{j=1}^D \overline{\alpha_{tj}} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\alpha_{tj}} \right) + \psi(\overline{\alpha_{tj}} + X_{1ij}) - \psi(\overline{\alpha_{tj}}) \right] \right\}\end{aligned}$$

#### 4.4 Updating variational parameters of $\beta$

Similarity, the stochastic natural gradient of the variational objective with respect to  $\beta$

$$\begin{aligned}(\iota_j^*)^{(t+1)} &= (1 - \rho^{(t)}) (\iota_j^*)^{(t)} \\ &+ \rho^{(t)} \left\{ \iota_j + \sum_{i=1}^N [1 - f_{ij}] \overline{\beta_j} \left( \left[ \psi \left( \sum_{j=1}^D \overline{\beta_j} \right) - \psi \left( \sum_{j=1}^D X_{1ij} + \sum_{j=1}^D \overline{\beta_j} \right) \right] + [\psi(\overline{\beta_j} + X_{1ij}) - \psi(\overline{\beta_j})] \right) \right\}\end{aligned}$$

#### 4.5 Updating stick-breaking representation

Similarity, the optimized solution to the posterior distribution of unit length sticks  $\Pi'_{1t}$

$$\begin{aligned}\log Q^* \left( \Pi'_{1t} \right) &= (1 - 1) \log \left( \Pi'_{1t} \right) + (\kappa^* - 1) \log \left( 1 - \Pi'_{1t} \right) \\ &+ \sum_{k=1}^{K_{max}} \mathbb{E}_Q [Z_{1tk}] \log \left( \Pi'_{1t} \right) + \sum_{k=1}^{K_{max}} \sum_{t'=t+1}^{T_{max}} \mathbb{E}_Q [Z_{1kt'}] \log \left( 1 - \Pi'_{1t} \right) + const\end{aligned} \quad (11)$$

which has the logarithmic form of the Beta distribution and the corresponding variational distributions of breaking proportions  $q(\Pi'_{1t})$  are considered to be the Beta distribution. Then, the standard conditions are

satisfied for a closed form coordinate update for local parameters. Based on the stochastic natural gradient, the variational parameters of  $\Pi_{1t}$  are updated by computing variational expectations in equation (11) [4, 2],

$$\begin{aligned}(\vartheta_t^*)^{(t+1)} &= (1 - \rho^{(t)}) (\vartheta_t^*)^{(t)} + \rho^{(t)} \left\{ 1 + \sum_{k=1}^{K_{max}} l_{1kt} \right\} \\ (\vartheta_t^{*'})^{(t+1)} &= (1 - \rho^{(t)}) (\vartheta_t^{*'})^{(t)} + \rho^{(t)} \left\{ \kappa^* + \sum_{k=1}^{K_{max}} \sum_{t'=t+1}^{T_{max}} l_{1kt'} \right\}\end{aligned}$$

#### 4.6 Updating variational parameters of $\epsilon$

Similarity, the stochastic natural gradient of the variational objective with respect to  $\epsilon$

$$\begin{aligned}(\xi_{j1}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\xi_{j1}^*)^{(t)} + \rho^{(t)} \left\{ \xi_{j1} + \sum_{i=1}^N f_{ij} \right\} \\ (\xi_{j2}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\xi_{j2}^*)^{(t)} + \rho^{(t)} \left\{ \xi_{j2} + \sum_{i=1}^N (1 - f_{ij}) \right\}\end{aligned}$$

### 5 Stochastic Optimization of the Variational Parameters for Gaussian mixture model (GMM)

#### 5.1 Compute the variational function parameters for $Z_{2kt'}$

Following the principles of the variational inference [4, 2], variational parameters of  $Z_{2kt'}$  is the local parameters, we consider the derivation of the update equation for only one variable by fixing the others' distribution. The optimal coordinate updates are derived for local parameters. The log of the optimized factor to the posterior distribution of  $Z_{2kt'}$  is

$$\begin{aligned}\log Q^*(Z_2) &= E_{W, \Phi_2, \Pi_2, \epsilon', \mu, \delta, \mu', \delta'} \left[ \log p(W, Z_2, \Phi_2, \Pi_2, \epsilon', \mu, \delta, \mu', \delta') \right] + const = \sum_{k=1}^{K_{max}} \sum_{t'=1}^{T'_{max}} Z_{2kt'} \log(l_{2kt'}) + const \\ \log l_{2kt'} &= \sum_{i=1}^N \sum_{j'=1}^{D'} E_Q[W_{ik}] E_Q[\Phi_{2ij'}] \left[ \frac{1}{2} E_Q[\log(\delta_{t'j'})] + \left( \frac{-1}{2} (X_{2ij} - E_Q[\mu_{t'j'}]) \right)^T E_Q[\delta_{t'j'}] (X_{2ij} - E_Q[\mu_{t'j'}]) \right] \\ &\quad + E_Q[\log(\Pi_{2t'})] + \sum_{t''=1}^{t'-1} E_Q[\log(1 - \Pi_{2t''})]\end{aligned}\tag{12}$$

$$E_Q[Z_{2kt'}] = \exp\{\log(l_{2kt'})\}$$

The variational parameters of  $Z_{2kt'}$  are updated as follow:

$$\begin{aligned}\log l_{2kt'} &= \sum_{i=1}^N \sum_{j'=1}^{D'} g_{ik} f'_{ij'} \left( \frac{1}{2} \left[ \psi \left( \frac{\chi_{t'j'}^{*'} + 1 - j}{2} \right) + D' \log 2 + \log |\varpi_{t'j'}^*| \right] \right) \\ &\quad + \sum_{i=1}^N \sum_{j'=1}^{D'} g_{ik} f'_{ij'} \left( \frac{-1}{2} \left[ \chi_{t'j'}^* (X_{2ij'} - \varphi_{t'j'}^*) \varpi_{t'j'}^* (X_{2ij'} - \varphi_{t'j'}^*)^T + \frac{D'}{\varsigma_{t'j'}^*} \right] \right) \\ &\quad + \psi(\vartheta_{t'}^+) - \psi(\vartheta_{t'}^+ + \vartheta_{t'}^{+'}) + \sum_{t''=1}^{t'-1} \psi(\vartheta_{t''}^{+'}) - \psi(\vartheta_{t''}^+ + \vartheta_{t''}^{+'})\end{aligned}\tag{13}$$

#### 5.2 Compute the variational function parameters of indicator variables for metabolome selection

Similarity, the optimized solution to the posterior distribution of variable selection  $\Phi_{2ij'}$

$$\log Q^*(\Phi_2) = E_{W, Z_2, \Pi_2, \epsilon', \mu, \delta, \mu', \delta'} \left[ \log p(W, Z_2, \Phi_2, \Pi_2, \epsilon', \mu, \delta, \mu', \delta') \right] + const = \sum_{i=1}^N \sum_{j'=1}^{D'} \Phi_{2ij'} \log(f'_{ij'}) + const$$

$$\begin{aligned}
\log f_{ij'}^{\Phi_{2ij'}} &= \sum_{k=1}^{K_{max}} \sum_{t'=1}^{T'_{max}} g_{ik} l_{2kt'} \left( \frac{1}{2} \left[ \psi \left( \frac{\chi_{t'j'}^* + 1 - j}{2} \right) + D' \log 2 + \log |\varpi_{t'j'}^*| \right] + const \right) \\
&+ \sum_{i=1}^N \sum_{j'=1}^{D'} g_{ik} f'_{ij'} \left( \frac{-1}{2} \left[ \chi_{t'j'}^* \left( X_{2ij'} - \varphi_{t'j'}^* \right) \varpi_{t'j'}^* \left( X_{2ij'} - \varphi_{t'j'}^* \right)^T + \frac{D'}{\varsigma_{t'j'}^*} \right] \right) \\
&+ \left[ \psi(\xi_1^*) - \psi(\xi_1^* + \xi_2^*) \right]
\end{aligned} \tag{14}$$

$$\begin{aligned}
\log f_{ij'}^{1-\Phi_{2ij'}} &= \left( \frac{1}{2} \left[ \psi \left( \frac{\chi_{j'}^* + 1 - j'}{2} \right) + D' \log 2 + \log |\varpi_{j'}^*| \right] + const \right) \\
&+ \left( \frac{-1}{2} \left[ \chi_{j'}^* \left( X_{2ij'} - \varphi_{j'}^* \right) \varpi_{j'}^* \left( X_{2ij'} - \varphi_{j'}^* \right)^T + \frac{D'}{\varsigma_{j'}^*} \right] \right) \\
&+ \left[ \psi(\xi_2^*) - \psi(\xi_1^* + \xi_2^*) \right]
\end{aligned} \tag{15}$$

### 5.3 Updating stick-breaking representation

The optimized solution to the posterior distribution of unit length sticks  $\Pi_{2t'}$

$$\begin{aligned}
\log Q^* \left( \Pi_{2t'} \right) &= (1 - 1) \log \left( \Pi_{2t'} \right) + (\kappa^+ - 1) \log \left( 1 - \Pi_{2t'} \right) \\
&+ \sum_{k=1}^{K_{max}} E_Q [Z_{2kt'}] \log \left( \Pi_{2t'} \right) + \sum_{k=1}^{K_{max}} \sum_{t''=t'+1}^{T'_{max}} E_Q [Z_{2kt''}] \log \left( 1 - \Pi_{2t''} \right) + const
\end{aligned} \tag{16}$$

Based on the stochastic natural gradient, the variational parameters of  $\Pi_{2t}$  are updated by computing variational expectations in equation (16) [4, 2],

$$\begin{aligned}
(\vartheta_{t'}^+)^{(t+1)} &= (1 - \rho^{(t)}) (\vartheta_{t'}^+)^{(t)} + \rho^{(t)} \left\{ 1 + \sum_{k=1}^{K_{max}} l_{2kt'} \right\} \\
(\vartheta_{t'}^{+'})^{(t+1)} &= (1 - \rho^{(t)}) (\vartheta_{t'}^{+'})^{(t)} + \rho^{(t)} \left\{ \kappa^+ + \sum_{k=1}^{K_{max}} \sum_{t''=t'+1}^{T'_{max}} l_{2kt''} \right\}
\end{aligned}$$

### 5.4 Updating variational parameters of $\epsilon'$

The stochastic natural gradient of the variational objective with respect to  $\epsilon'$

$$\begin{aligned}
(\xi_{j'1}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\xi_{j'1}^*)^{(t)} + \rho^{(t)} \left\{ \xi_{j'1}' + \sum_{i=1}^N f'_{ij'} \right\} \\
(\xi_{j'2}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\xi_{j'2}^*)^{(t)} + \rho^{(t)} \left\{ \xi_{j'2}' + \sum_{i=1}^N (1 - f'_{ij'}) \right\}
\end{aligned}$$

### 5.5 Updating variational parameters of $\mu$

The stochastic natural gradient of the variational objective with respect to  $\mu$

$$\begin{aligned}
(\varphi_{j't'}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\varphi_{j't'}^*)^{(t)} + \rho^{(t)} \left\{ \frac{\varphi_{j't'} \varsigma_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'} X_{2ij'}}{\varsigma_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'}} \right\} \\
(\varsigma_{j't'}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\varsigma_{j't'}^*)^{(t)} + \rho^{(t)} \left\{ \varsigma_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'} \right\}
\end{aligned}$$

### 5.6 Updating variational parameters of $\delta$

The stochastic natural gradient of the variational objective with respect to  $\delta$

$$\begin{aligned}
(\varpi_{j't'}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\varpi_{j't'}^*)^{(t)} + \\
\rho^{(t)} &\left\{ \varpi_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'} (X_{2ij'} - \varphi_{j't'}^*) (X_{2ij'} - \varphi_{j't'}^*)^T + \frac{\varsigma_{j't'} \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'}}{\varsigma_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'}} (\varphi_{j't'}^* - \varphi_{j't'})^T (\varphi_{j't'}^* - \varphi_{j't'}) \right\} \\
(\chi_{j't'}^*)^{(t+1)} &= (1 - \rho^{(t)}) (\chi_{j't'}^*)^{(t)} + \rho^{(t)} \left\{ \chi_{j't'} + \sum_{i=1}^N \sum_{k=1}^{K_{max}} g_{ik} l_{2kt'} f'_{ij'} \right\}
\end{aligned}$$

## 5.7 Updating variational parameters of $\mu'$

The stochastic natural gradient of the variational objective with respect to  $\mu'$

$$\begin{aligned} (\varphi'_{j'})^{(t+1)} &= (1 - \rho^{(t)}) (\varphi'^*_{j'})^{(t)} + \rho^{(t)} \left\{ \frac{\varphi'_{j'} \varsigma'_{j'} + \sum_{i=1}^N f'_{ij'} X_{2ij'}}{\varsigma'_{j'} + \sum_{i=1}^N f'_{ij'}} \right\} \\ (\varsigma'^*_{j'})^{(t+1)} &= (1 - \rho^{(t)}) (\varsigma'^*_{j'})^{(t)} + \rho^{(t)} \left\{ \varsigma'_{j'} + \sum_{i=1}^N f'_{ij'} \right\} \end{aligned}$$

## 5.8 Updating variational parameters of $\delta'$

The stochastic natural gradient of the variational objective with respect to  $\delta'$

$$\begin{aligned} (\varpi'^*_{j'})^{(t+1)} &= (1 - \rho^{(t)}) (\varpi'^*_{j'})^{(t)} + \\ &\rho^{(t)} \left\{ \varpi'_{j'} + \sum_{i=1}^N f'_{ij'} (X_{2ij'} - \varphi'^*_{j'}) (X_{2ij'} - \varphi'^*_{j'})^T + \frac{\varsigma'_{j'} \sum_{i=1}^N f'_{ij'}}{\varsigma'_{j'} + \sum_{i=1}^N f'_{ij'}} (\varphi'^*_{j'} - \varphi'_{j'})^T (\varphi'^*_{j'} - \varphi'_{j'}) \right\} \\ (\chi'^*_{j'})^{(t+1)} &= (1 - \rho^{(t)}) (\chi'^*_{j'})^{(t)} + \rho^{(t)} \left\{ \chi'_{j'} + \sum_{i=1}^N f'_{ij'} \right\} \end{aligned}$$

# 6 Soybean rhizosphere microbiome and metabolome

## 6.1 Field experiment

The accessions and experimental fields were identical to those used by Toda et al. and Sakurai et al [5, 6]. 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 [7, 8]. 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<sup>-2</sup> 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.

## 6.2 Metabolome analysis

Metabolome analysis of root samples was performed according to Sawada et al., 2009 [9] and Uchida et al., 2020 [10]. 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  $\mu$ L of the extract, 75  $\mu$ L of the extraction solvent was added. Twenty-five  $\mu$ L of the diluted extract was dried with N<sub>2</sub> gas, and 250  $\mu$ L 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  $\mu$ m) 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  $\mu$ L. 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>) [11].

### 6.3 Microbiome analysis

#### DNA extraction

Bacterial DNA was extracted from the root and rhizosphere soils, as previously described [12, 13]. Briefly, 100 mg of the powdered sample was dispensed into a 1.5 mL tube and chilled in liquid nitrogen. Subsequently, 200  $\mu$ L 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 H<sub>2</sub>O (Cat. #10977015, Thermo Fisher Scientific, Waltham, MA, USA) [14] 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  $\mu$ L 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  $\mu$ L of 80% ethanol. Finally, the DNA was eluted from the magnetic beads using 15  $\mu$ L 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 [12, 14]. 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 [15, 16]. The V4 region of 16S rRNA was amplified in a 10  $\mu$ L reaction containing 1  $\mu$ L of 10-fold diluted sample, 2 $\times$  KAPA HiFi HotStart ReadyMix (Cat. #07958935001, KAPA Biosystems Inc., USA), 0.2  $\mu$ M forward and reverse primers, and 1  $\mu$ M 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  $\mu$ L) was mixed with 2  $\mu$ L 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') [17]. The second PCR reaction was performed in a 10  $\mu$ L reaction containing 0.8  $\mu$ L of the purified first amplification, 2 $\times$  KAPA HiFi HotStart ReadyMix, 1  $\mu$ M forward and reverse primers, and 1  $\mu$ M 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  $\mu$ L of 80% ethanol. The DNA libraries were eluted from the magnetic beads with 15  $\mu$ L 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 $\times$  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 [18] 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 [19, 20]. Sequences derived from Archaea, Eukaryota, mitochondria, or chloroplasts were removed.

## References

- [1] M. J. B. Yee Whye Teh, Michael I Jordan and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.
- [2] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, “Stochastic variational inference,” *Journal of Machine Learning Research*, 2013.
- [3] T. Dang, K. Kumaishi, E. Usui, S. Kobori, T. Sato, Y. Toda, Y. Yamasaki, H. Tsujimoto, Y. Ichihashi, and H. Iwata, “Stochastic variational variable selection for high-dimensional microbiome data,” *Microbiome*, vol. 10, no. 1, pp. 1–14, 2022.
- [4] D. M. Blei and M. I. Jordan, “Variational inference for dirichlet process mixtures,” 2006.
- [5] K. Sakurai, Y. Toda, H. Kajiya-Kanegae, Y. Ohmori, Y. Yamasaki, H. Takahashi, H. Takanashi, M. Tsuda, H. Tsujimoto, A. Kaga *et al.*, “Time-series multispectral imaging in soybean for improving biomass and genomic prediction accuracy,” *The Plant Genome*, vol. 15, no. 4, p. e20244, 2022.
- [6] Y. Toda, G. Sasaki, Y. Ohmori, Y. Yamasaki, H. Takahashi, H. Takanashi, M. Tsuda, H. Kajiya-Kanegae, R. Lopez-Lozano, H. Tsujimoto *et al.*, “Genomic prediction of green fraction dynamics in soybean using unmanned aerial vehicles observations,” *Frontiers in Plant Science*, vol. 13, 2022.
- [7] A. Kaga, T. Shimizu, S. Watanabe, Y. Tsubokura, Y. Katayose, K. Harada, D. A. Vaughan, and N. Tomooka, “Evaluation of soybean germplasm conserved in nias genebank and development of mini core collections,” *Breeding science*, vol. 61, no. 5, pp. 566–592, 2012.
- [8] H. Kajiya-Kanegae, H. Nagasaki, A. Kaga, K. Hirano, E. Ogiso-Tanaka, M. Matsuoka, M. Ishimori, M. Ishimoto, M. Hashiguchi, H. Tanaka *et al.*, “Whole-genome sequence diversity and association analysis of 198 soybean accessions in mini-core collections,” *DNA Research*, vol. 28, no. 1, p. dsaa032, 2021.
- [9] Y. Sawada, K. Akiyama, A. Sakata, A. Kuwahara, H. Otsuki, T. Sakurai, K. Saito, and M. Y. Hirai, “Widely targeted metabolomics based on large-scale ms/ms data for elucidating metabolite accumulation patterns in plants,” *Plant and Cell Physiology*, vol. 50, no. 1, pp. 37–47, 2009.
- [10] K. Uchida, Y. Sawada, K. Ochiai, M. Sato, J. Inaba, and M. Y. Hirai, “Identification of a unique type of isoflavone o-methyltransferase, gmiomt1, based on multi-omics analysis of soybean under biotic stress,” *Plant and Cell Physiology*, vol. 61, no. 11, pp. 1974–1985, 2020.
- [11] H. Tsugawa, M. Arita, M. Kanazawa, A. Ogiwara, T. Bamba, and E. Fukusaki, “Mrmprobs: a data assessment and metabolite identification tool for large-scale multiple reaction monitoring based widely targeted metabolomics,” *Analytical chemistry*, vol. 85, no. 10, pp. 5191–5199, 2013.
- [12] K. Kumaishi, E. Usui, K. Suzuki, S. Kobori, T. Sato, Y. Toda, H. Takanashi, S. Shinozaki, M. Noda, A. Takakura *et al.*, “High throughput method of 16s rrna gene sequencing library preparation for plant root microbial community profiling,” *Scientific Reports*, vol. 12, no. 1, p. 19289, 2022.
- [13] Y. Ichihashi, A. Fukushima, A. Shibata, and K. Shirasu, “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*, pp. 163–189, 2018.
- [14] Y. Ichihashi, Y. Date, A. Shino, T. Shimizu, A. Shibata, K. Kumaishi, F. Funahashi, K. Wakayama, K. Yamazaki, A. Umezawa *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*, vol. 117, no. 25, pp. 14 552–14 560, 2020.

- [15] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight, “Global patterns of 16s rrna diversity at a depth of millions of sequences per sample,” *Proceedings of the national academy of sciences*, vol. 108, no. supplement\_1, pp. 4516–4522, 2011.
- [16] D. S. Lundberg, S. Yourstone, P. Mieczkowski, C. D. Jones, and J. L. Dangl, “Practical innovations for high-throughput amplicon sequencing,” *Nature methods*, vol. 10, no. 10, pp. 999–1002, 2013.
- [17] H. Toju and Y. G. Baba, “Dna metabarcoding of spiders, insects, and springtails for exploring potential linkage between above-and below-ground food webs,” *Zoological letters*, vol. 4, pp. 1–12, 2018.
- [18] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, “Dada2: High-resolution sample inference from illumina amplicon data,” *Nature methods*, vol. 13, no. 7, pp. 581–583, 2016.
- [19] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, “The silva ribosomal rna gene database project: improved data processing and web-based tools,” *Nucleic acids research*, vol. 41, no. D1, pp. D590–D596, 2012.
- [20] P. Yilmaz, L. W. Parfrey, P. Yarza, J. Gerken, E. Pruesse, C. Quast, T. Schweer, J. Peplies, W. Ludwig, and F. O. Glöckner, “The silva and “all-species living tree project (ltp)” taxonomic frameworks,” *Nucleic acids research*, vol. 42, no. D1, pp. D643–D648, 2014.