MEDWIN PUBLISHERS
Committed to Create Value for Researchers

International Journal of Forensic Sciences
ISSN: 2573-1734

Appendix

Nine Measurements

Assume the true proportions of evidence $Y = \left(y_1, y_2, \cdots, y_p\right)$, where p is the number of individuals in the pool, and the predicted proportions of evidence $\hat{Y} = \left(\hat{y}_1, \hat{y}_2, \ldots\ldots, \hat{y}_p\right)$. The difference between true and predicted is residuals *ri*, *i* = 1 ,2, ... , *p*. The expression for each measurement can be expressed as:

Average of absolute error (AVG):

$$AVG = \frac{1}{p}\sum_{i=1}^{p}\left|r_i\right|$$

Average relative Error (AVGER):

$$AVGER = \frac{1}{p}\sum_{i=1}^{p}\frac{\left|r_i\right|}{\max\left(y_i, \hat{} y_i\right)}$$

Total Variation Distance (DTV):

$$DTV = \frac{1}{2}\sum_{i=1}^{p}\left|r_i\right|$$

Median Absolute Deviation (MAD):

$$MAD = median_i\left(\left|r_i - median_j\left(r_i\right)\right|\right)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{p}\sum_{i=1}^{p}\left|\frac{r_i}{\max\left(y_i, \hat{} y\right)}\right|$$

The Maximum Relative Error (MAXRE):

$$MAXRE = \max_i\left(\frac{\left|r_i\right|}{\max\left(y_i, \hat{} y_i\right)}\right)$$

Mean Squared Error (MSE):

$$MSE = \frac{1}{p}\sum_{i=1}^{p}\left(r_i\right)^2$$

Relative before root Mean Squared Error (RRMSE):

$$RRMSE = \sqrt{\frac{1}{p}\sum_{i=1}^{p}\left(\frac{r_i}{\max\left(y_i, \hat{y}_i\right)}\right)^2}$$

Symmetric Mean Absolute Percentage Error (SMAPE):

$$SMAPE = \frac{1}{p}\sum_{i=1}^{p}\frac{2\times\left|r_i\right|}{\left|y_i\right| + \left|\hat{y}_i\right|}$$

Figures

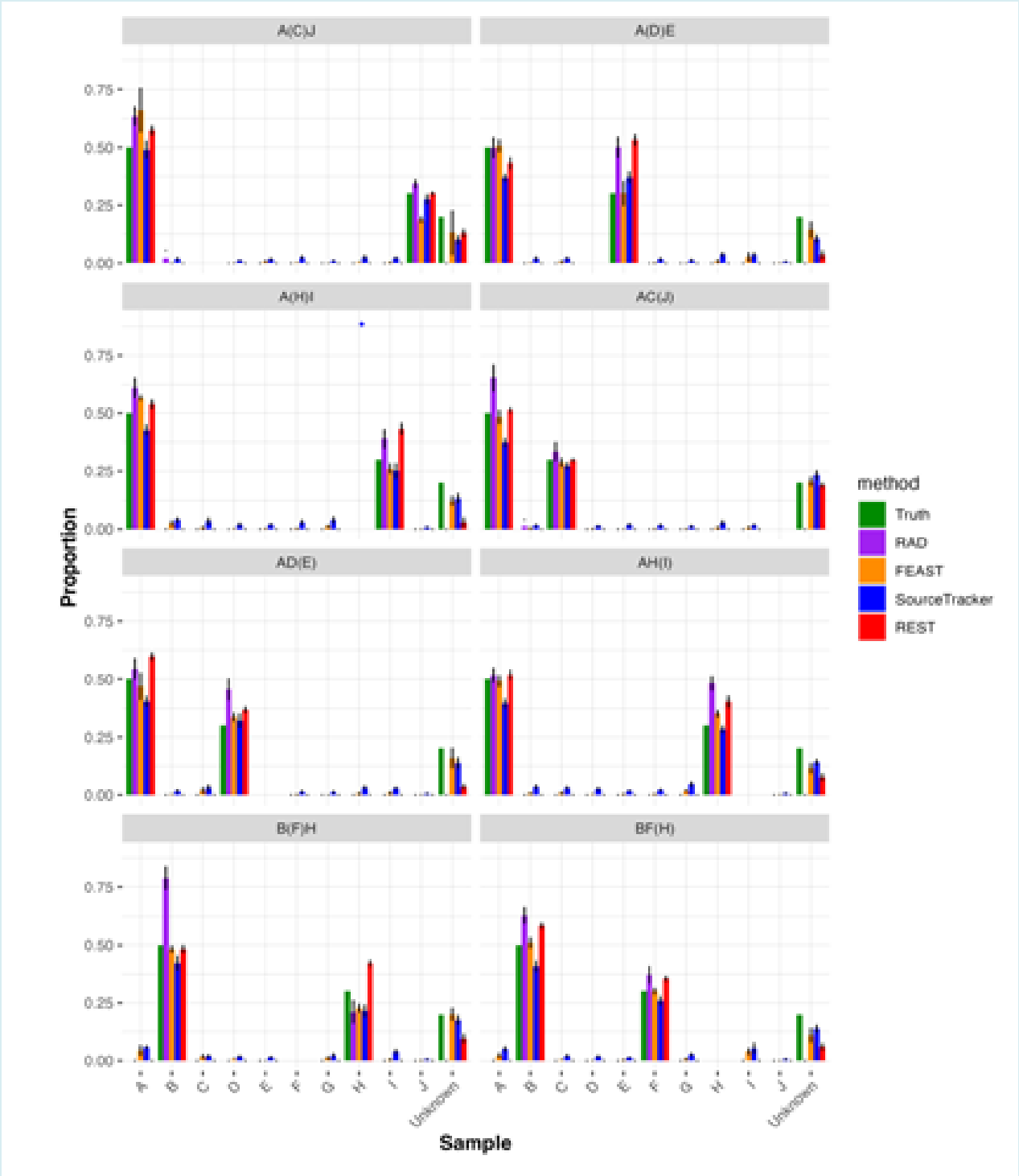

Figure S1: Barplot of the Estimated Proportions from Each Method for the Eight Experiments of Missing Source scenarios, and the True Composition of the Evidence.

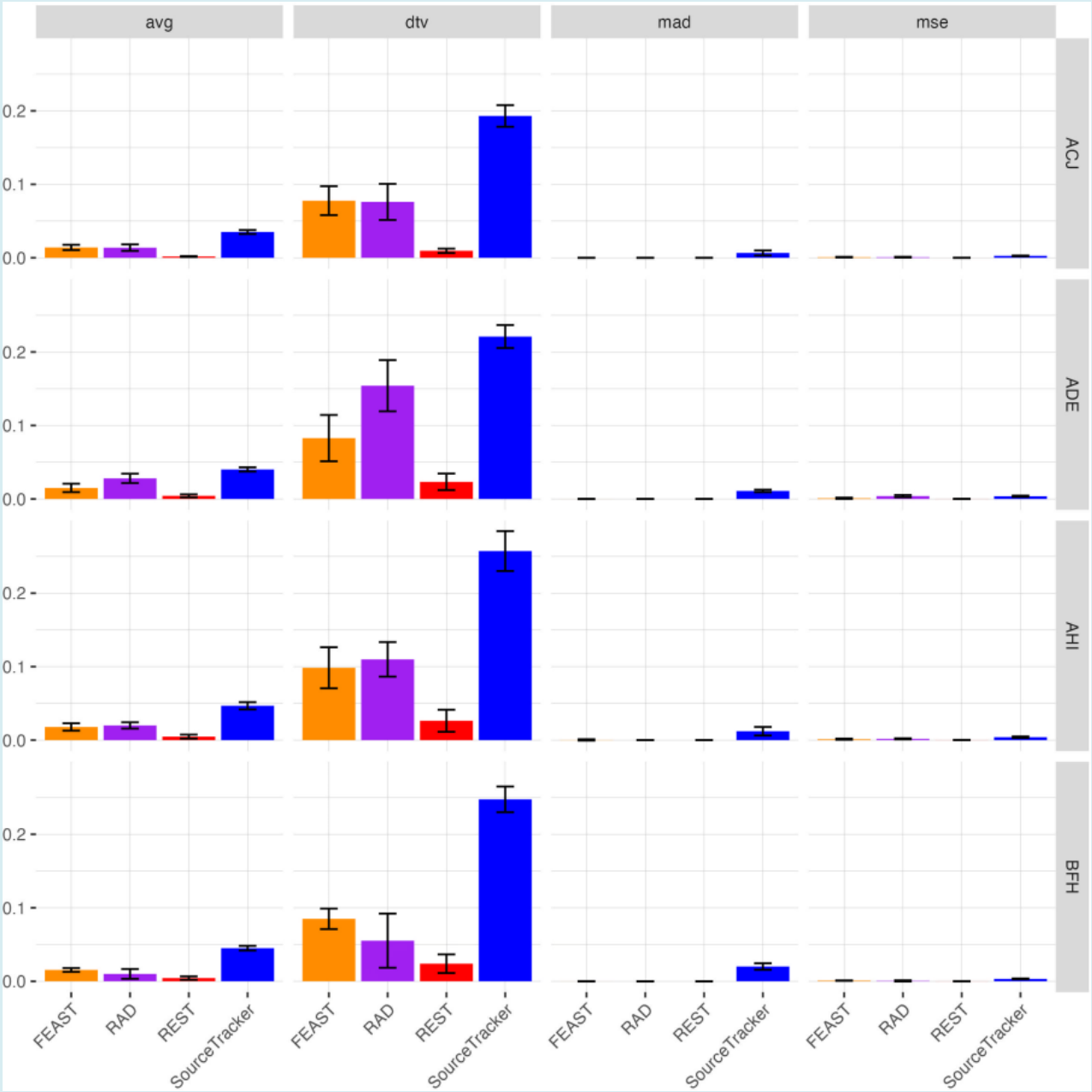

Figure S2: Bar Plot of Four Performance Measurements (on the Top) for the Non-Missing Source Scenario. Each Row Represents an Experiment Setting and Each Experiment was Replicated 10 Times. The X-Axis is the Method, and the Y-Axis is Averaged Value of Performance from 10 Replications. The Error Bars Show the Standard Deviation from the Replications. The Smaller the measurement Value, the Better the Performance. (Avg: Average Of Absolute Error; Dtv: Total Variation Distance; Mad: Median Absolute Deviation; Mse: Mean Squared Error).

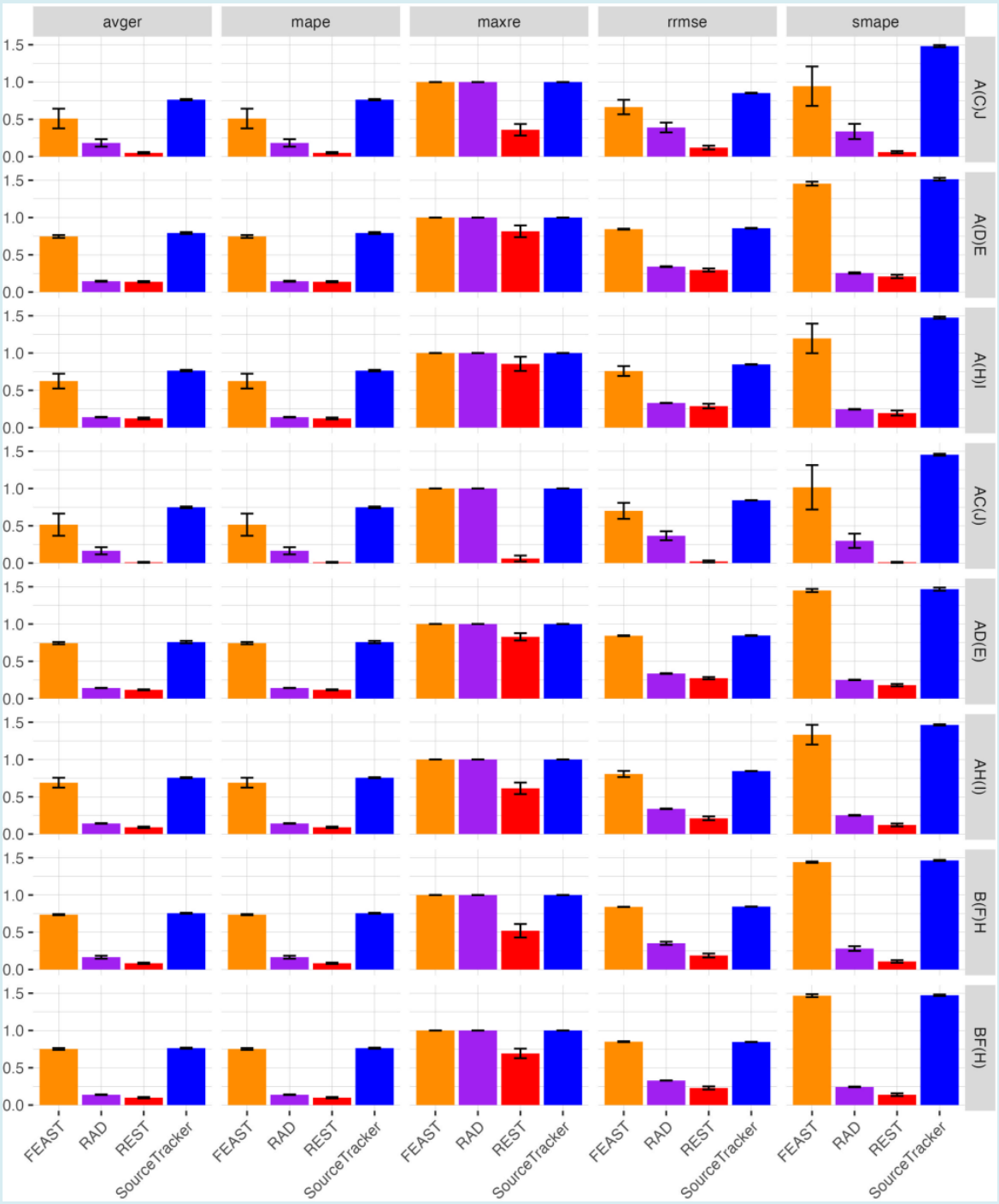

Figure S3: Bar Plot of Five Performance Measurements (on the Top) for the Missing Source Scenario. Each Row Represents an Experiment Setting and Each Experiment was Replicated 10 Times. The X-Axis is the Method, and the Y-Axis is Averaged Value of Performance from 10 Replications. The Error Bars Show the Standard Deviation from the Replications. The Smaller the measurement Value, the Better the Performance. (Avger: Average Relative Error; Mape: Mean Absolute Percentage Error; Maxre: Maximum Relative Error; Rrmse: relative before root Mean Squared Error; Smape: Symmetric Mean Absolute Percentage Error).

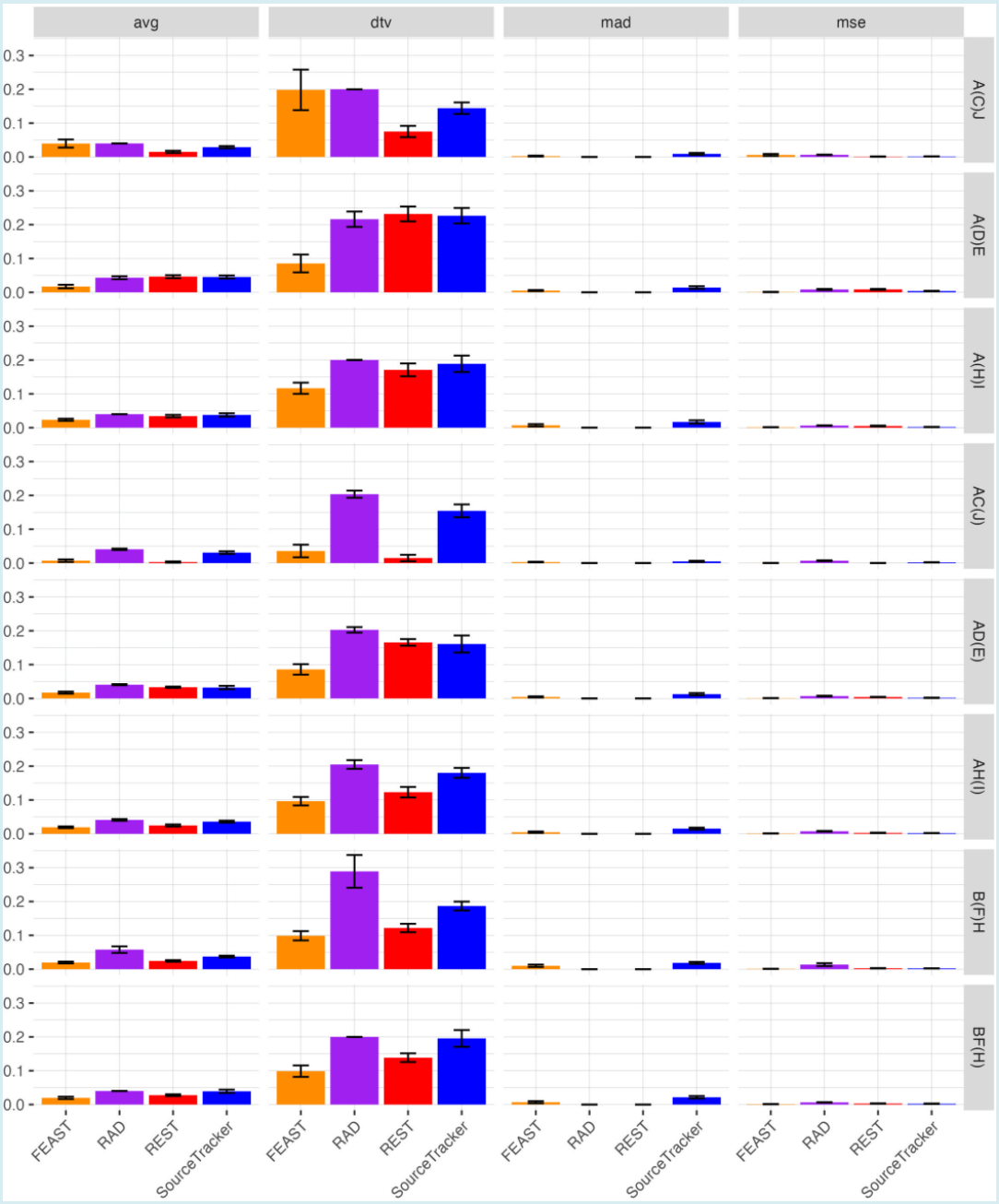

Figure S4: Bar Plot of Four Performance Measurements (on the Top) for the Missing Source Scenario. Each Row Represents an Experiment Setting and Each Experiment was Replicated 10 Times. The X-Axis is the Method, and the Y-Axis is Averaged Value of Performance from 10 Replications. The Error Bars Show the Standard Deviation from the Replications. The Smaller the measurement Value, the Better the Performance. (Avg: Average Of Absolute Error; Dtv: Total Variation Distance; Mad: Median Absolute Deviation; Mse: Mean Squared Error)

Steps of the Method REST

Step 1: Filter Contributors Via Relative Aitchison Difference Test

The first step of the REST method involves determining the important contributors among the sources to the evidence microbial sample. Let Si represent the ith source $(i = 1,..., p)$ and EV represents the evidence sample. We evaluate the dissimilarity between two samples i and j. Leveraging these pairwise dissimilarity measurements as reference or background, our objective is to identify the sources that exhibit higher similarity or lower dissimilarity with the evidence. This objective is realized through the implementation of a series of hypothesis tests.

$$H_0 : S_i \text{ is related to } EV \; vs. H_a : S_i \text{ is not related to } EV$$

Since each microbial sample represents a mixture of microbes, the data can be transformed to compositional data. In such cases, it is essential to use a distance metric that preserves the unit-sum property of compositional data. The Aitchison distance is a suitable metric for comparing microbial samples as it maintains this property. To measure the distance between each source and the evidence, the RAD test

[7] is employed.

Assuming independence among microbial species in a sample, we can demonstrate that the distances would follow an asymptotic normal distribution according to the central limit theorem. While this assumption may not hold strictly in reality, it should be sufficiently robust for testing purposes. To generate a distribution of dissimilarity we can employ bootstrap resampling with replacement. This approach allows us to build bootstrap confidence intervals for the squared Aitchison difference. Confidence intervals (e.g., 95%) that are strictly greater than zero indicate that the respective source is an important contributor to the evidence.

Step 2: Construct a Profile for Potential "Unknown" Source Via Weighted Least Squares Regression

The data used in our study consists of microbial sources and evidence, represented as operational taxonomic unit (OTU) count. We have not conducted any data transformation on these variables. However, it is important to note that the constant variance assumption of the errors is violated in our data. To address this issue, we have employed weighted least squares regression instead of ordinary least squares regression. The model expresses the following:

$$Y = XQ + \varepsilon$$

The response $\boldsymbol{Y}$ is represented as an $n \times 1$ vector, where each element corresponds to an OTU for the evidence sample. $\boldsymbol{X}$ is an $n \times p$ count matrix, where each row represents an OTU and each column is a source. ε is assumed to be normally distributed with mean 0 and non-constant variance-covariance matrix as follows

$$\begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_1^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Then we define the weight in the weighted least square regression as the reciprocal of each variance, $\sigma_i^2, w_i = \frac{1}{\sigma_i^2}$. The matrix W is expressed as:

$$W = \begin{bmatrix} W_1 & 0 & \dots & 0 \\ 0 & W_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_n \end{bmatrix}$$

After the weighted least square regression is applied to the data, the information about a potential missing or unknown source

will be mainly contained in the residuals. We define the profile of the unknown OTU as $U = u_1, \ldots, u_n$, and the *ith* element is defined as:

$$u_i - \frac{d_i}{\sum_{k=1}^{n} d_k} * S$$

where $d_i = e_i - \min(e_i)$, *ei* is the *ith* residual from the weighted least squares model fitting and S is the median value of total abundance *sj* which is defined as the sum of abundance count in n features or OTUs for source $j\,(j = 1, \ldots, p)$.

Step 3: Check the Status of Possible "Unknown" Source Via Agreement Analysis Based on Intra-Class Correlation (ICC)

In order to assess whether the "unknown" source obtained from the weighted least squares regression is noise or omitted, we utilize the intra-class correlation coefficient (ICC). The ICC, originally introduced by Fisher [8], is commonly used to evaluate reliability. In our study, we employ a random effects model to examine the consistency of the unknown profiles obtained in step 2. Assuming we have C replicates of evidence and obtain the unknown profile C times. We apply a one-way analysis of variance random effects model for the unknown profile *Zkm*:

$$Z_{ik} = \mu + \gamma_i + \omega_{ik} \;\; i = 1, .., n \;\; k = 1, \ldots, C$$

Here μ represents the overall effect, γi is a random variable and ωik is the error term. The random variable γi follows a normal distribution with a mean of zero and a variance $r2$, while the error term ωik is normally distributed with mean zero and variance φ^2. Next, we use the Intra-class correlation coefficient (ICC) to evaluate the consistency of the unknown profiles [9]. The formula for the one-way ANOVA intraclass correlation is given as:

$$\mathrm{ICC} = (\mathrm{MSB} - \mathrm{MSW}) / \mathrm{MSB} + (\mathrm{C} - 1)\mathrm{MSW}$$

Where MSB represents the mean square between subjects, MSW represents the mean square within subjects, and C represents the number of replicates of evidence. The ICC ranges from -1/(C-1) to 1, indicating the extent of consistency among the unknown profiles.

A high ICC ratio, approaching 1, indicates strong agreement among the *C* profiles, suggesting minimal within-subject variation. In such cases, *Zik* s contains valuable information regarding an omitted source. Conversely, a low ICC value indicates a weak consistency among the *C* profiles; suggesting the *Zik*s are likely influenced by independent random errors rather than a shared signal.

Step 4: Estimate Proportions Via Constrained Least Squared Linear Regression

In this step, we employ constrained least squares linear regression to estimate the coefficients that represent the probability or proportion of microbial materials composing the source relative to the evidence. With the generalization, assuming the presence of an unknown source, the constrained least squares model expresses the following:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \beta_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \cdots + \beta_{p+1} \begin{bmatrix} u_i \\ \vdots \\ u_n \end{bmatrix} + error$$

Where the response is a vector of species count in the evidence sample, and there are *n* features or OTUs. The parameter β_{p+1} is the coefficient for the "unknown" source. If the "unknown" source is detected as noise in step 3, then the "unknown" profile will be excluded from the model. If there are multiple replicates of evidence available then multiple "unknown" categories are constructed from step 2. The averaged profile from multiple "unknown" categories will be included in the model fitting to estimate proportions. In general, we assume there are $p+1$ terms in the above model.

The parameters βs are solved by calculating:

$argmin \left\| Y - X\beta \right\|^2$ With constraints $\beta \geq 0$ and $\sum_{j=1}^{p+1} \beta_j = 1$.

The above constrained least squares problem can be reformulated as a Quadratic Programming (QP) problem. The basic

algorithm for QP was introduced by E. M. L Beale in 1955 [11], and since then, numerous variant methods and numerical algorithms have been developed to tackle QP problems. In 1982, the dual algorithm for solving QP problems was introduced, offering improved efficiency and numerical stability compared to earlier versions [12].