

Towards more accurate microbial source tracking via non-negative matrix factorization (NMF)

Ziyi Huang⁺, Dehan Cai⁺, Yanni Sun*

City University of Hong Kong

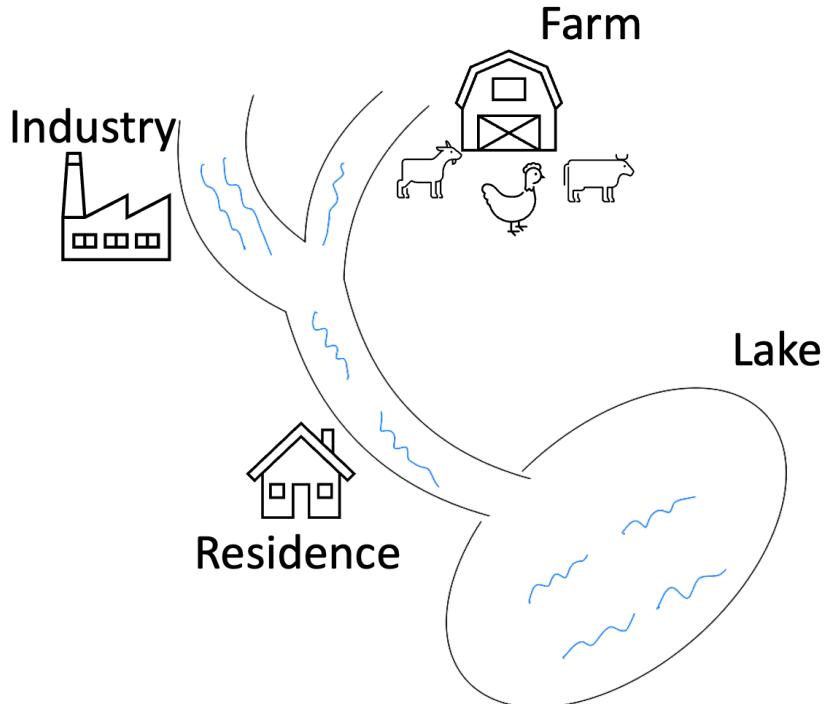
<https://yannisun.github.io/>

July 13, 2024



Sources of Lake Pollution: Who is Responsible?

2



- **Issue:** Poor water quality due to microbial pollution
- **Potential Polluters:** Industry, farms, residences
- **Action:** Sampling lake and surrounding sources
- **Goal:** Identify pollution sources (microbial source tracking, MST)

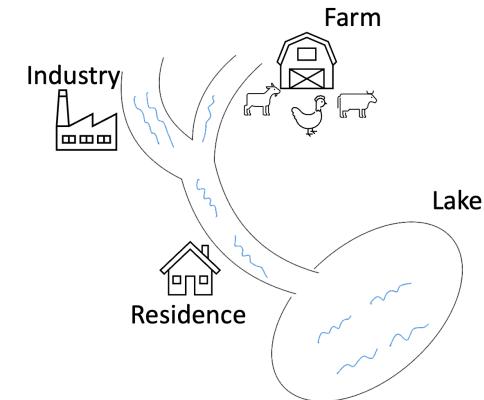
Microbial Source Tracking (MST): overview

3

► Microbial Source Tracking

- Input: a **target/sink** sample and multiple **source** samples from other microbial communities

Goal: to determine the **contribution of each source to the target**.



► Applications

- Quantifying contamination
 - Laboratory contamination: the target sample may be contaminated by sampling procedures, reagents, indoor air, etc.
 - Water contamination
- Microbial interaction analysis
 - Microbial interaction between humans and the indoor environment

microbial contamination from
the upper respiratory tract

nose

pharynx

lung



bronchoalveolar lavage fluid (BLF)



Input data: taxa abundance

4

Example: Tracking laboratory contamination

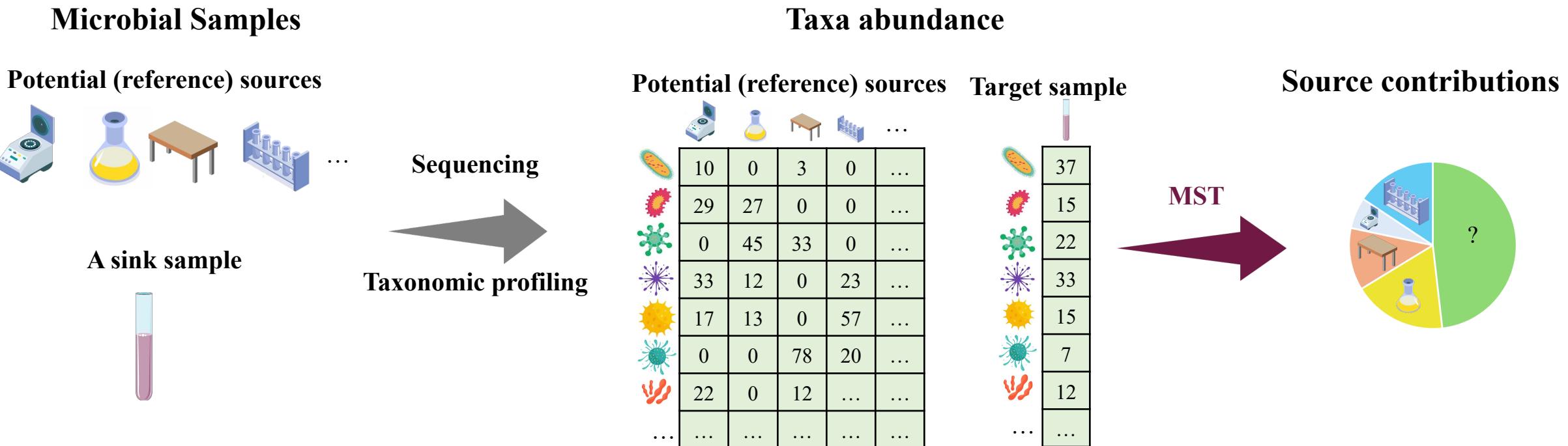


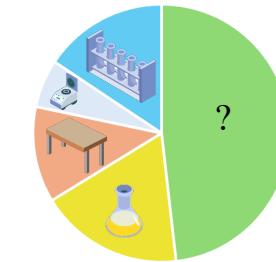
Figure sources: <https://bioicons.com/>

➤ Reference sources

- Irrelevant sources
- Unknown/unseen sources
- Similar sources
- Sequencing noises

Potential (reference) sources

| | ... | ... | ... | ... |
|-----|-----|-----|-----|-----|
| ... | 10 | 0 | 3 | 0 |
| ... | 29 | 27 | 0 | 0 |
| ... | 0 | 45 | 33 | 0 |
| ... | 33 | 12 | 0 | 23 |
| ... | 17 | 13 | 0 | 57 |
| ... | 0 | 0 | 78 | 20 |
| ... | 22 | 0 | 12 | ... |
| ... | ... | ... | ... | ... |



➤ State-of-the-art tools

- SourceTracker^[1]
- FEAST^[2]

[1] Knights, Dan, et al. "Bayesian community-wide culture-independent microbial source tracking." *Nature methods* 8.9 (2011): 761-763.

[2] Shenhav, Liat, et al. "FEAST: fast expectation-maximization for microbial source tracking." *Nature methods* 16.7 (2019): 627-632.

Method

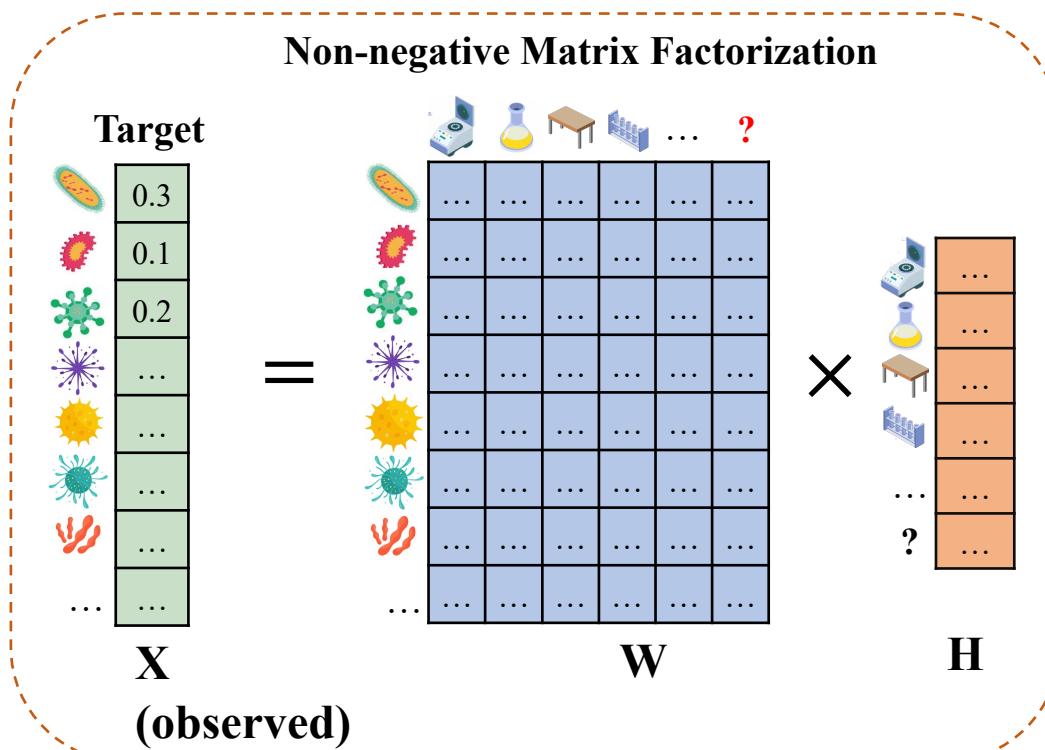
Non-negative Matrix Factorization (NMF)

7

- X: relative taxa abundance of the target sample
- W: relative taxa abundance matrix of sources
- H: proportion (contribution) of sources
- ?: Unknown sources

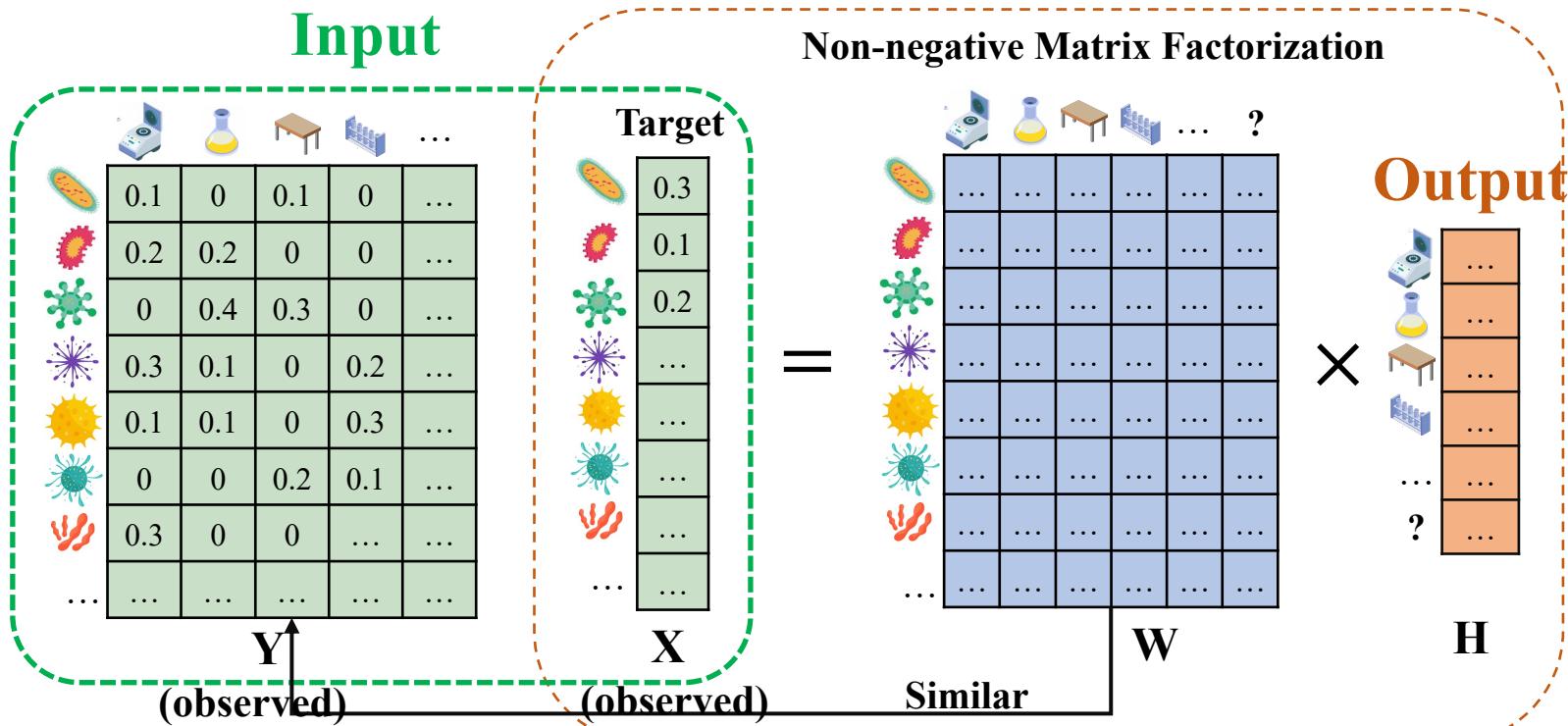
X is given.

Our goal: find W and H



Adding reference sources

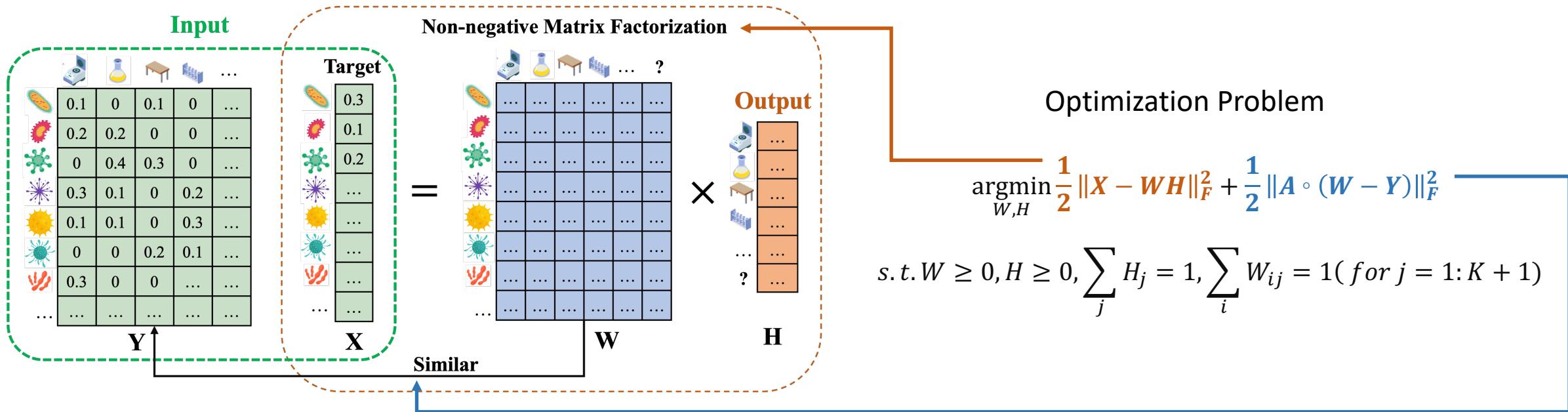
8



- Y: observed taxa abundance of reference sources

SourceID-NMF: optimization objective

9



- X : taxa abundance of a sample
- Y : observed taxa abundance of reference sources
- W : taxa abundance of sources (to be estimated)
- H : mixing proportion of sources (to be estimated)
- ?: Unknown sources
- A : the last column is all 0 and other elements are 1

SourceID-NMF: optimization method (ADMM)

10

Optimization Problem

$$\begin{aligned} & \operatorname{argmin}_{W,H} \frac{1}{2} \|X - WH\|_F^2 + \frac{1}{2} \|A \circ (W - Y)\|_F^2 \\ \text{s.t. } & W \geq 0, H \geq 0, \sum_j H_j = 1, \sum_i W_{ij} = 1 \text{ (for } j = 1:K+1) \end{aligned}$$

Optimization process of H and H^+

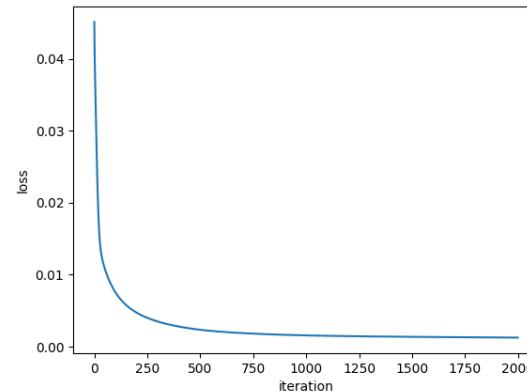
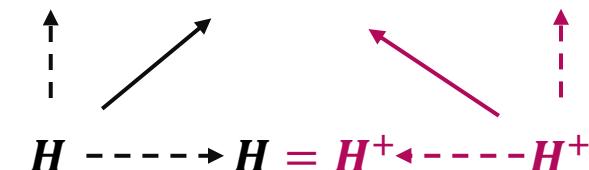
$$\frac{1}{2} \|X - WH\|_F^2$$

$$H^+ \geq 0 \sum_j H_j^+ = 1$$

Alternating Direction Method of Multipliers (ADMM)^[1]

$$\begin{aligned} & \operatorname{argmin}_{W^+, H^+, W, H} \frac{1}{2} \|X - WH\|_F^2 + \frac{1}{2} \|A \circ (W^+ - Y)\|_F^2 \\ \text{s.t. } & W = W^+, H = H^+ \\ & W^+ \geq 0, H^+ \geq 0, \sum_j H_j^+ = 1, \sum_i W_{ij}^+ = 1 \text{ (for } j = 1:K+1) \end{aligned}$$

auxiliary variables: W^+, H^+



[1] Boyd, S. et al. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), 1–122.

Result

► Testing SourceID-NMF on different datasets

- Simulated data
 - Inter-source similarity
 - Irrelevant sources
 - Unknown sources
 - Sources with noises
- Real data
 - Indoor environmental samples
 - Infant's fecal samples

► Metrics: contribution of sources (H)

Jensen-Shannon Divergence (JSD): the difference of true H and the estimated H

Pearson Correlation (PCC): the correlation between H and the estimated H

► Compare with two state-of-the-art tools

- SourceTracker^[1]
- FEAST^[2]

[1] Knights, Dan, et al. "Bayesian community-wide culture-independent microbial source tracking." *Nature methods* 8.9 (2011): 761-763.

[2] Shenhav, Liat, et al. "FEAST: fast expectation-maximization for microbial source tracking." *Nature methods* 16.7 (2019): 627-632.

Simulated data settings

13

► Simulated data process: how to obtain X (target) and Y (observed sources)

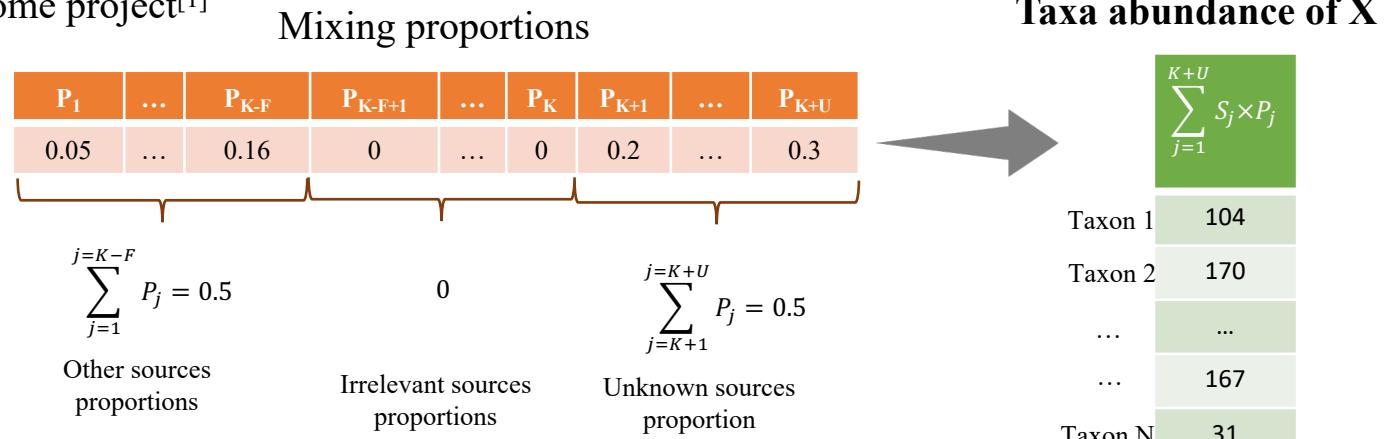
Taxa abundance (reads counts) of sources from the Earth's microbiome project^[1]

| | S_1 | ... | S_{K-F} | S_{K-F+1} | ... | S_K | S_{K+1} | ... | S_{K+U} |
|---------|-------|-----|-----------|-------------|-----|-------|-----------|-----|-----------|
| Taxon 1 | 15 | ... | 600 | 88 | ... | 130 | 234 | ... | 12 |
| Taxon 2 | 31 | ... | 34 | 993 | ... | 403 | 44 | ... | 334 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| ... | 179 | ... | 159 | 10 | ... | 299 | 39 | ... | 556 |
| Taxon N | 431 | ... | 90 | 204 | ... | 19 | 394 | ... | 20 |

↓ Multinomial sampling

Observed taxa abundance of K sources (Y)

| | S'_1 | ... | S'_{K-F} | S'_{K-F+1} | ... | S'_K |
|--|--------|-----|------------|--------------|-----|--------|
| | 10 | ... | 589 | 78 | ... | 123 |
| | 38 | ... | 50 | 948 | ... | 430 |
| | ... | ... | ... | ... | ... | ... |
| | 168 | ... | 143 | 12 | ... | 278 |
| | 456 | ... | 98 | 231 | ... | 14 |



► Simulated data settings

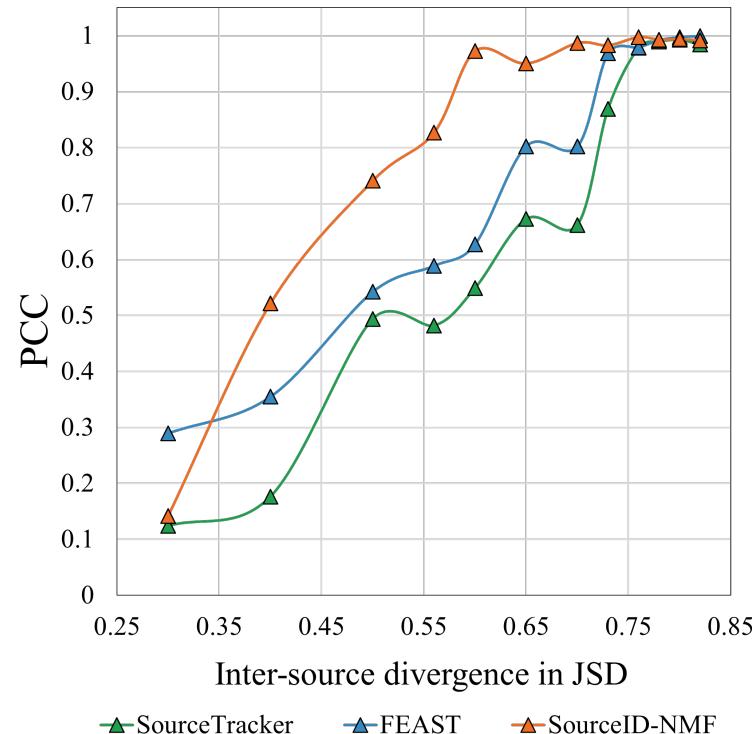
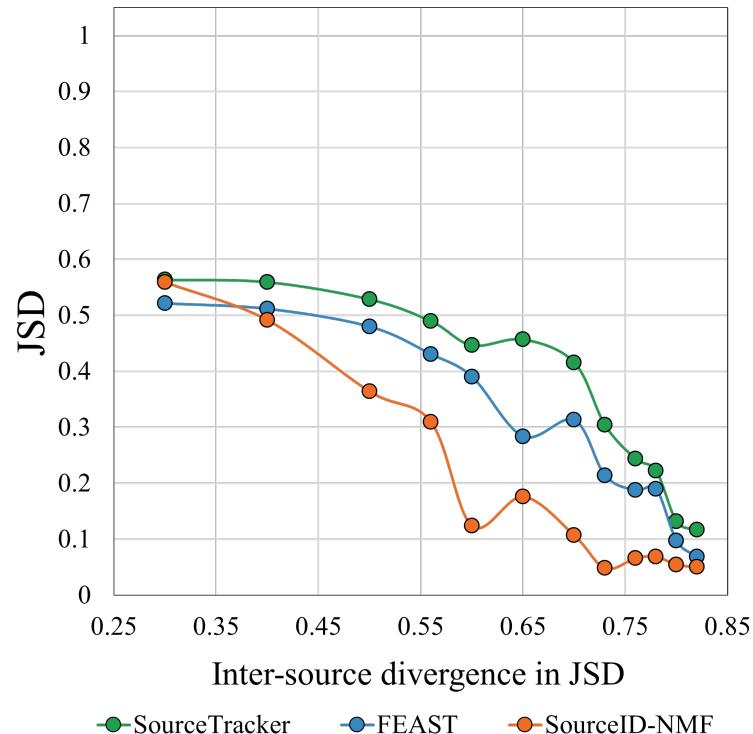
- K (15) + U (5) = 20 sources with different similarities
- Varying proportions of 5 unknown sources (0.1 to 0.9)
- 5 irrelevant sources in the 15 sources

[1] Thompson, L. R. et al. (2017). A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681), 457–463.

Simulated data experiment

14

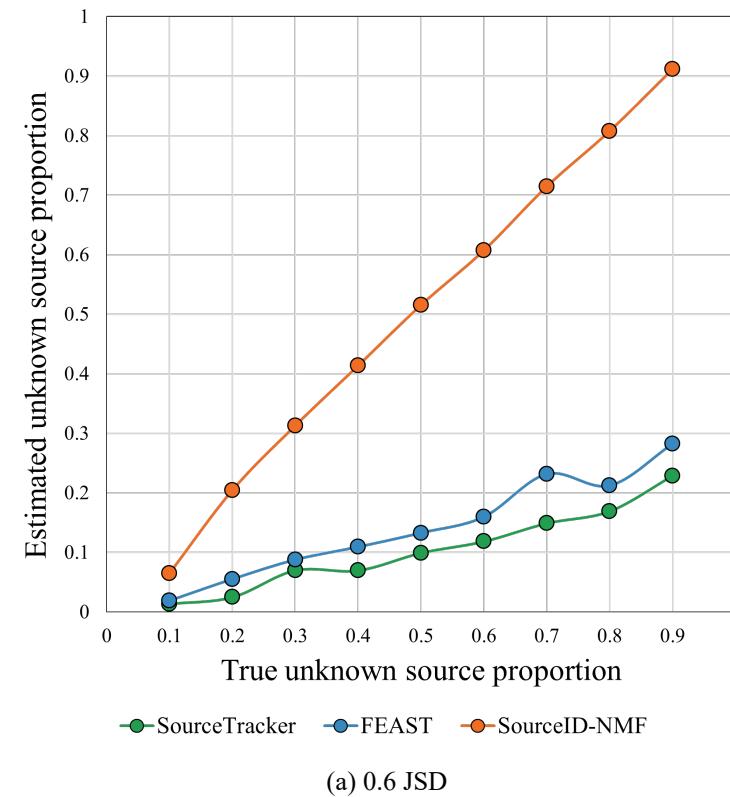
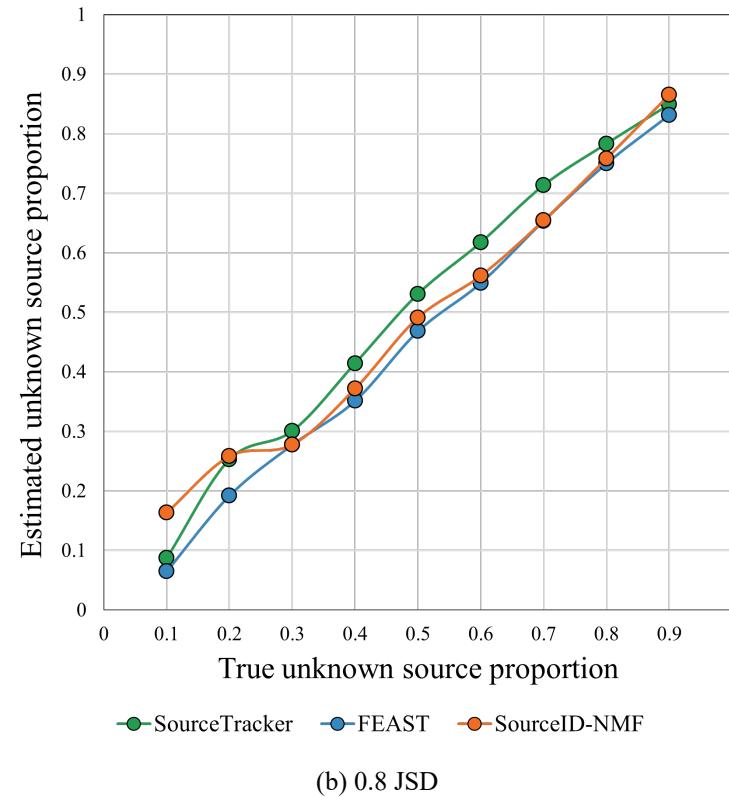
➤ Impact of inter-source similarity on source tracking: estimated H and the true H



Simulated data experiment

15

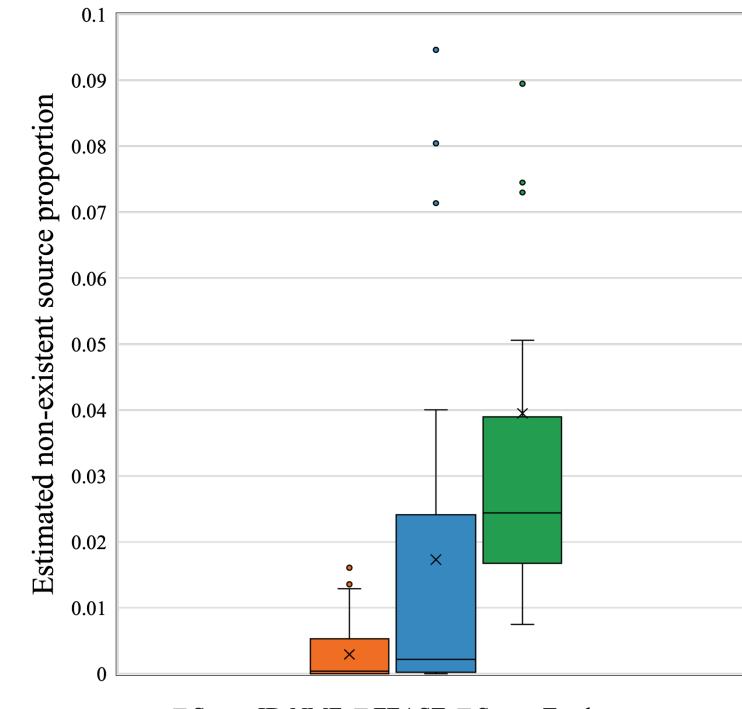
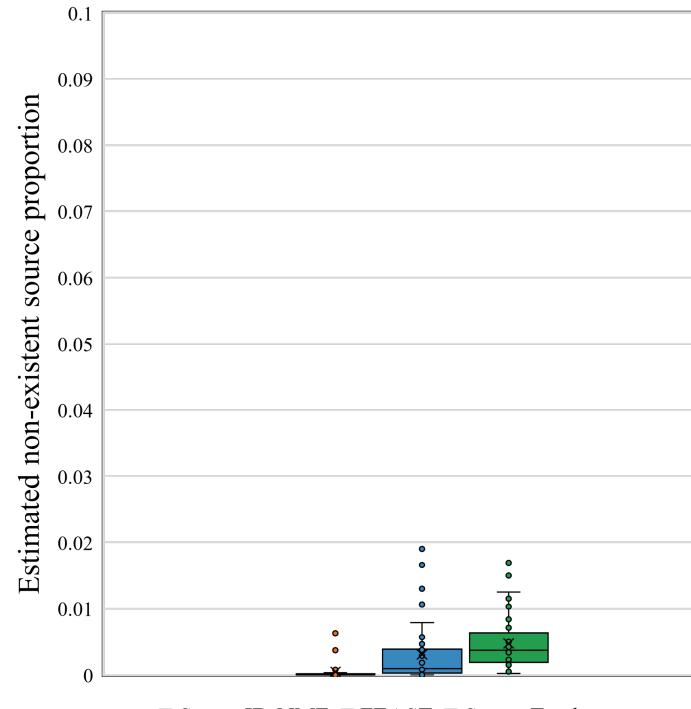
➤ SourceID-NMF has more accurate estimations of unknown sources' proportion



Simulated data experiment

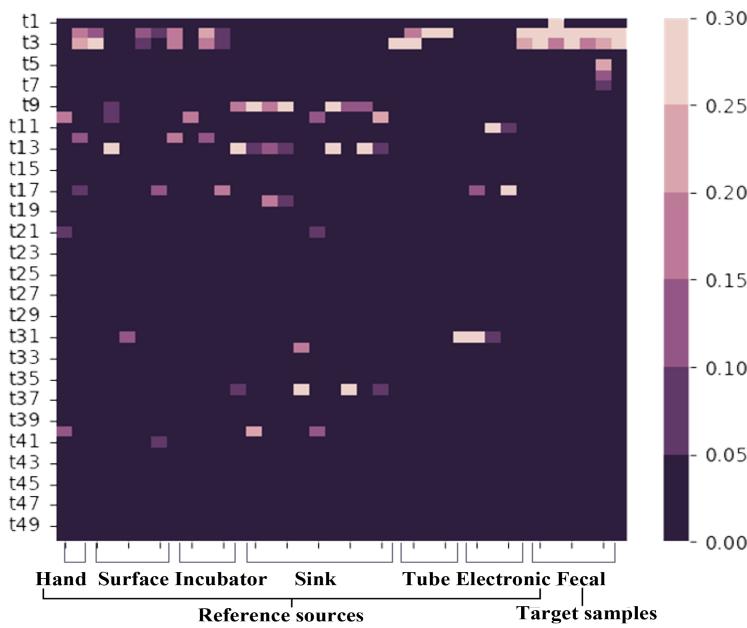
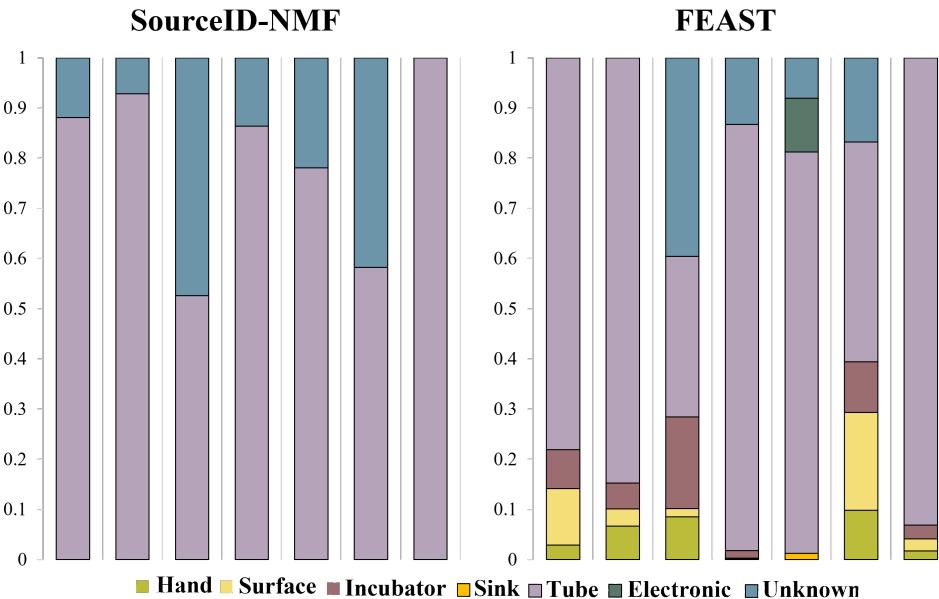
16

► More accurate identification of irrelevant sources (ideal value: 0)



► Infants' fecal samples

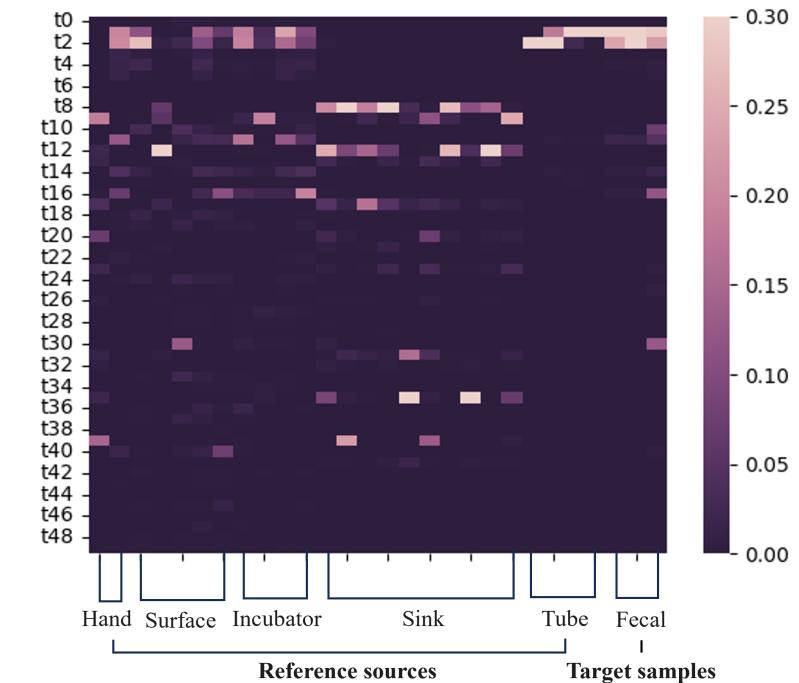
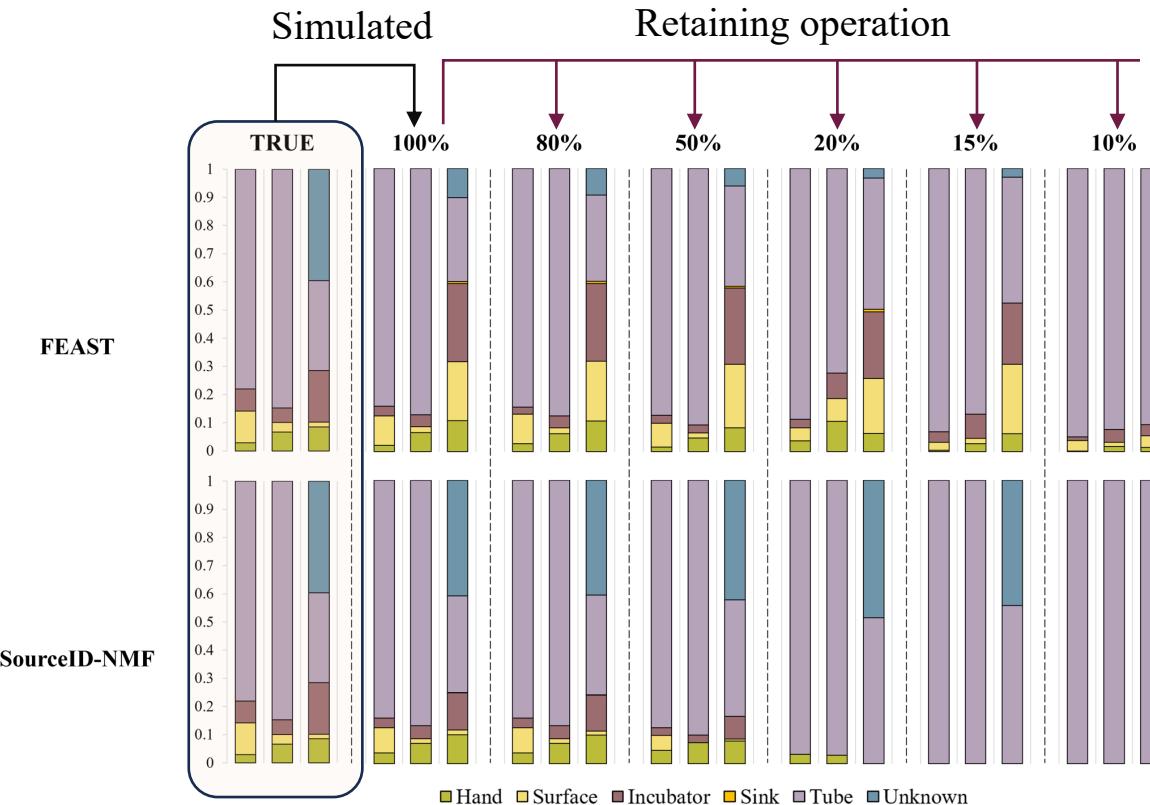
- Target samples: 7 fecal samples
- Reference sources: 29 samples from hands, environmental surfaces, incubators, sinks, tubes, and electronics



Investigate the reasons via simulation (I)

18

► Retaining Top n% of Taxa in the Target Sample While Setting the Rest to Zero

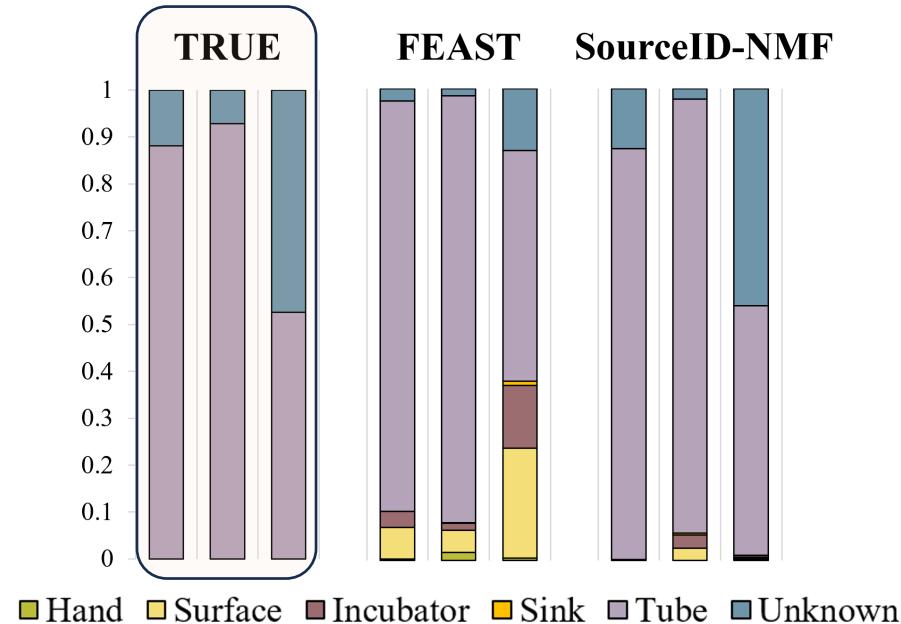


Heatmap of the taxa abundance when retaining 20%
(highly similar to the real data)

Investigate the difference via simulation (II)

19

► Re-generate the target sample using only tube and unknown sources

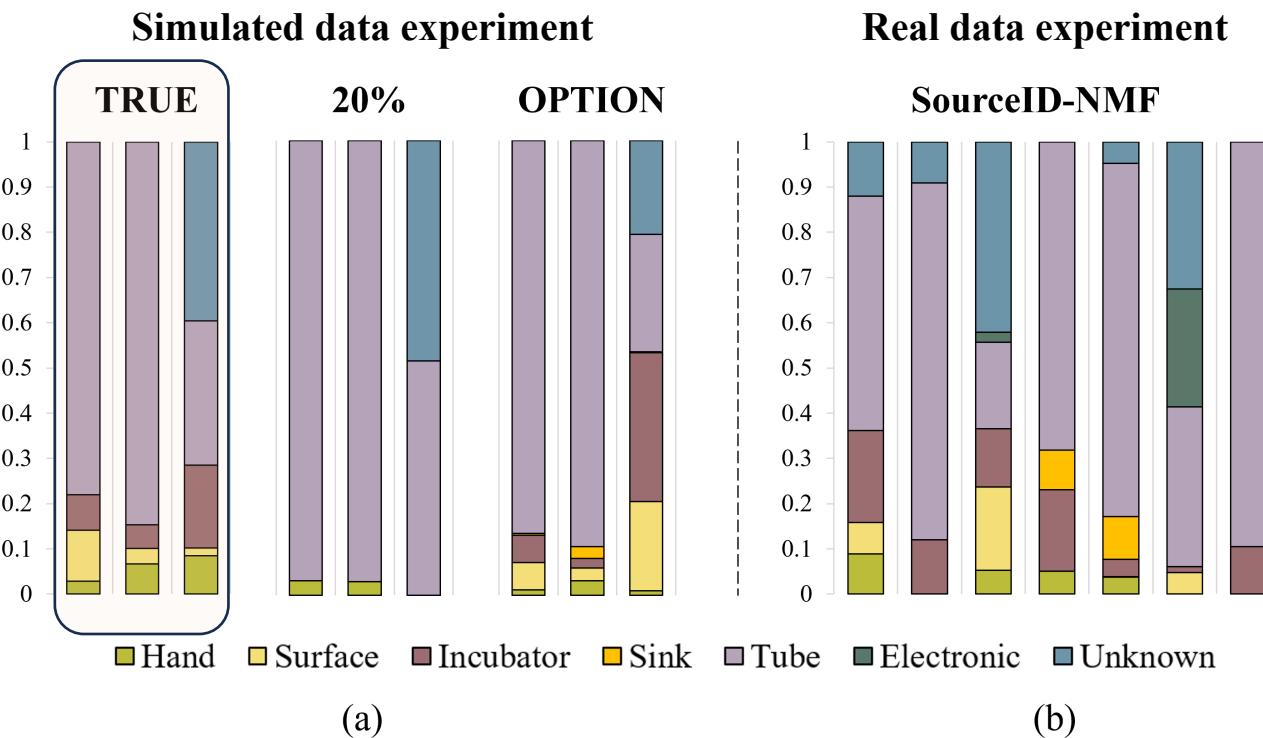


FEAST can generate false positives, consistent with previous experimental results

An option to include more sources

20

➤ Only use taxa observed in the target samples (OPTION)



➤ SourceID-NMF

- Goal: estimate the proportions of reference sources in a target microbial sample
- Input: taxa abundance of the target samples and reference sources
- Output: the proportion of reference sources and unknown source
- Good at estimating sources' proportions especially the unknown sources

Funding: GRF, ITF, and City University of Hong Kong



Ziyi Huang



Dehan Cai

GitHub of SourceID:

<https://github.com/ZiyiHuang0708/SourceID-NMF>



Group Website:

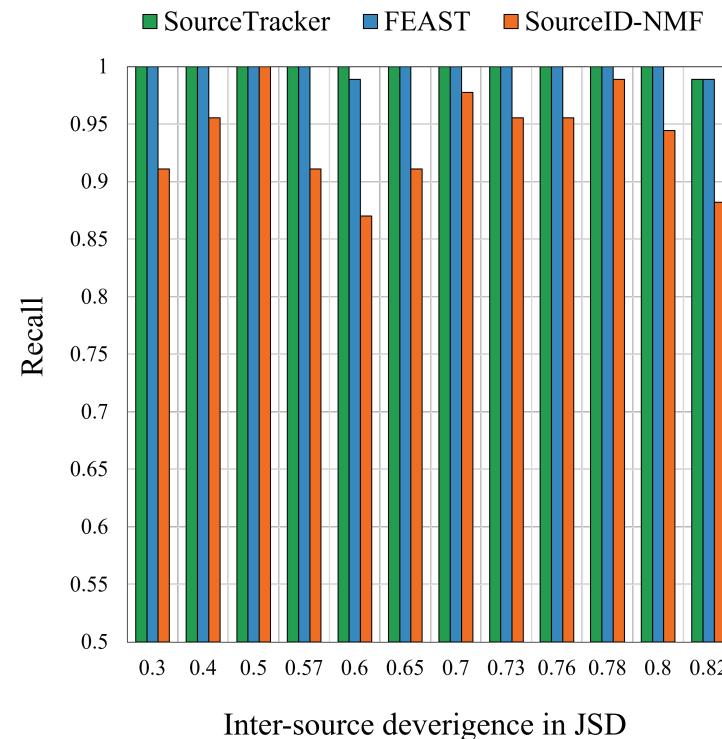
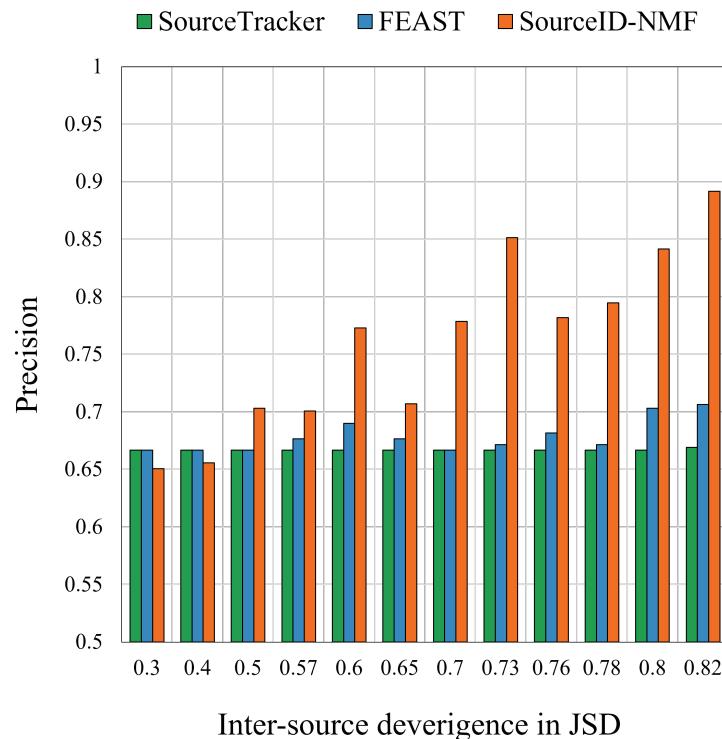
<https://yannisun.github.io/>



Simulated data experiment

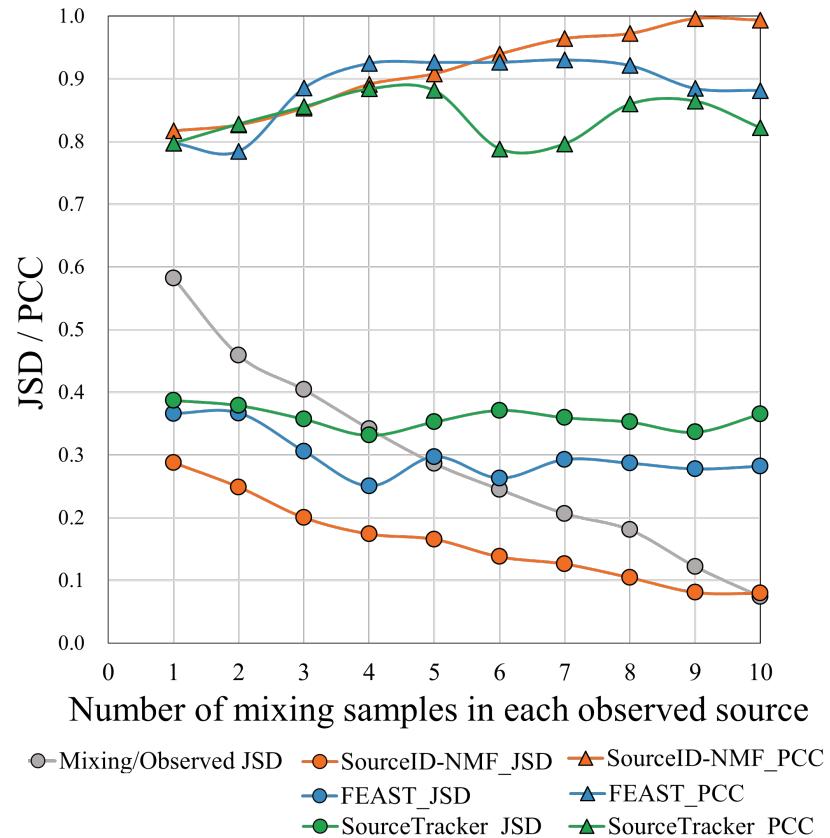
23

► Source tracking by ignoring specific proportions



A threshold of 1E-06 was set to determine source identification.

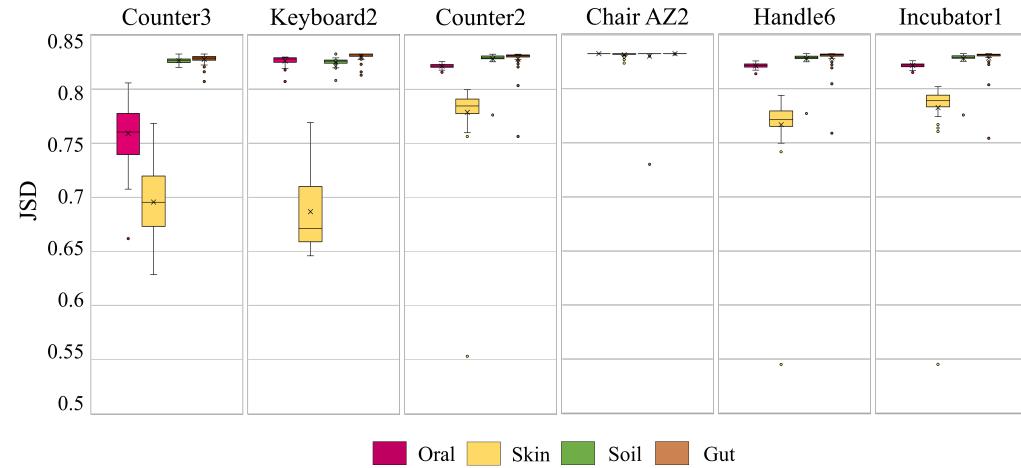
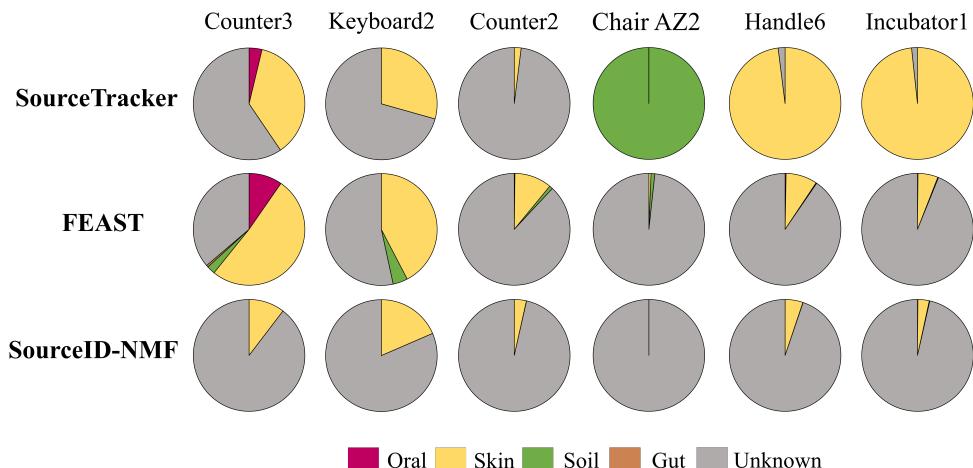
➤ Reference sources with noises



► Indoor environmental samples

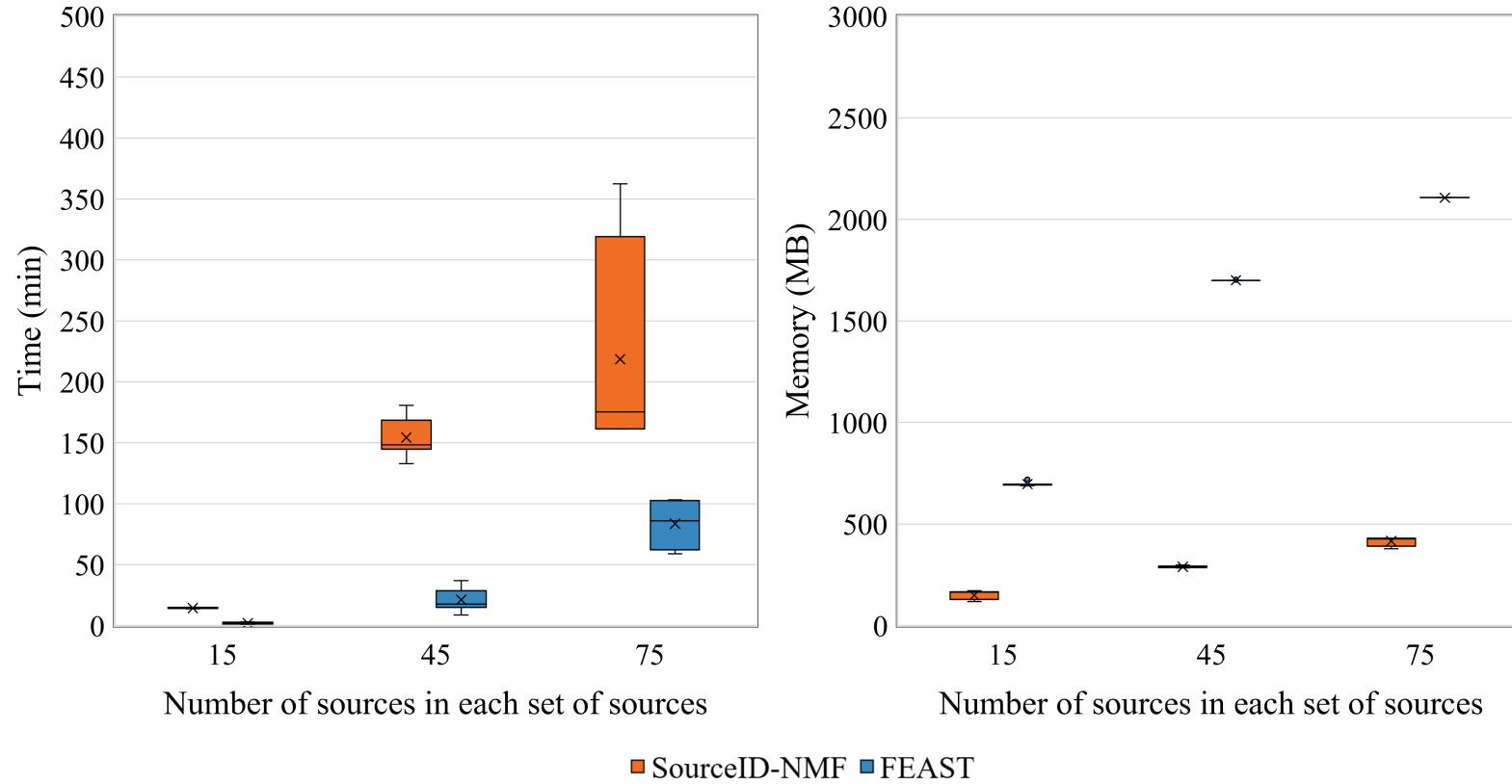
- Target samples: 6 Indoor environmental samples
- Reference sources: 180 microbial samples from gut, oral, skin and soil (45 samples for each)

No space-time correlation



Running time and Memory Usage

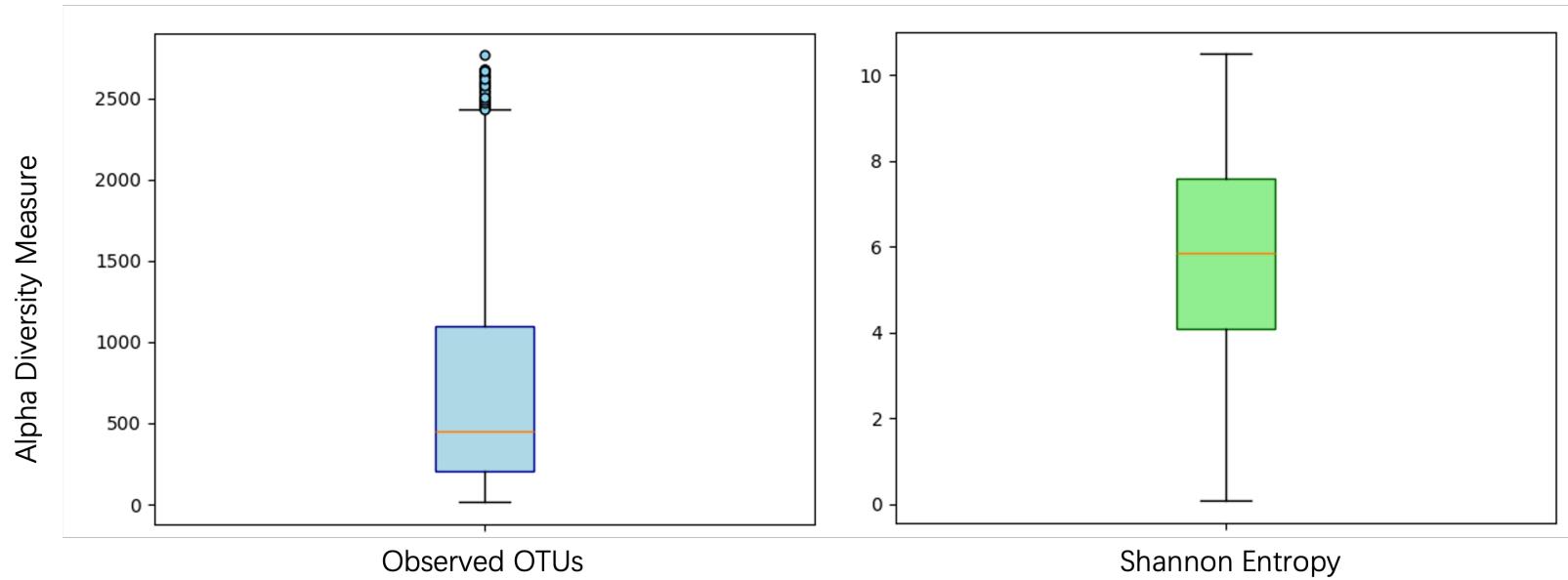
26



Alpha diversity

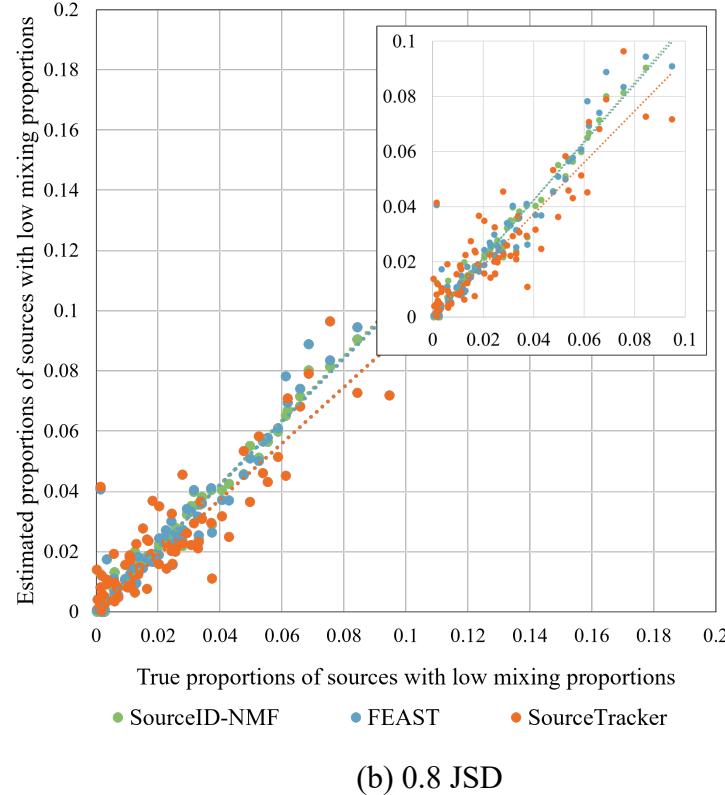
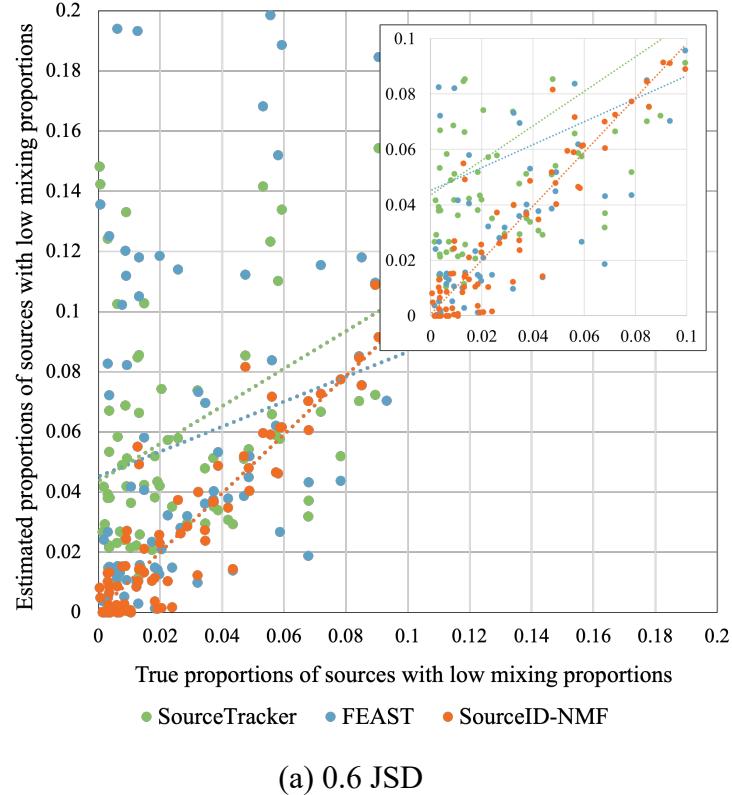
27

► Earth's microbiome project



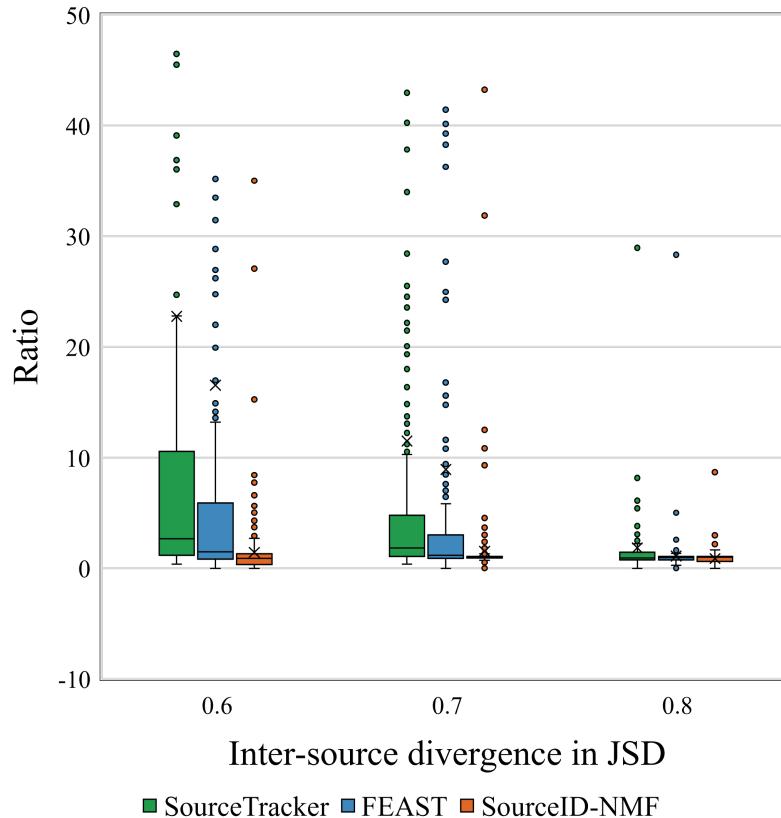
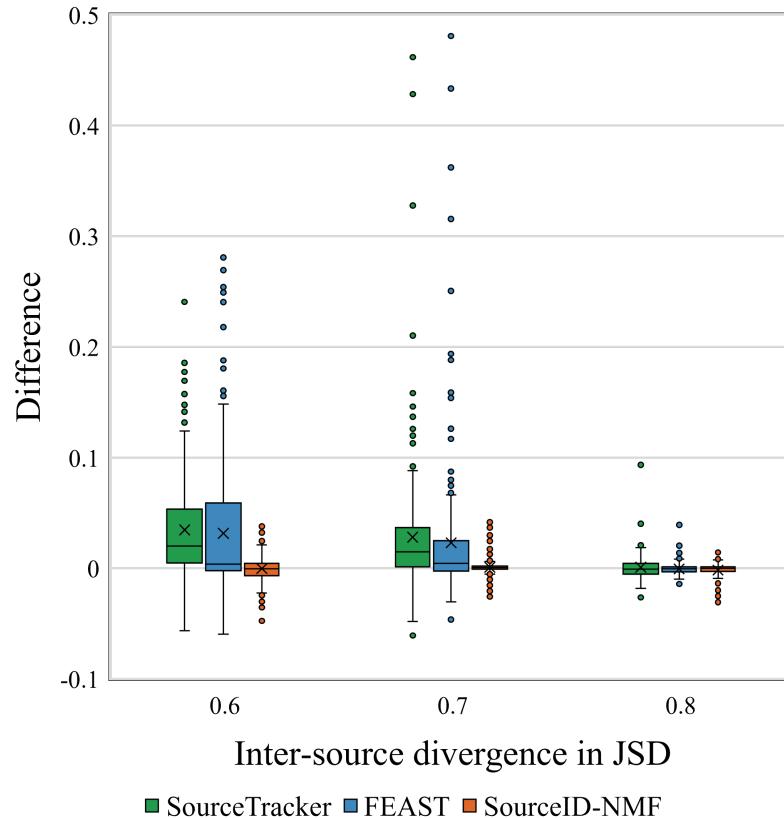
Sources with small proportions

28



Sources with small proportions

29



The boxplots of the estimated proportions for the low abundance sources (below 0.1 proportions). **Difference and Ratio (estimation divided by the ground-truth) between estimated proportions and true proportions.**