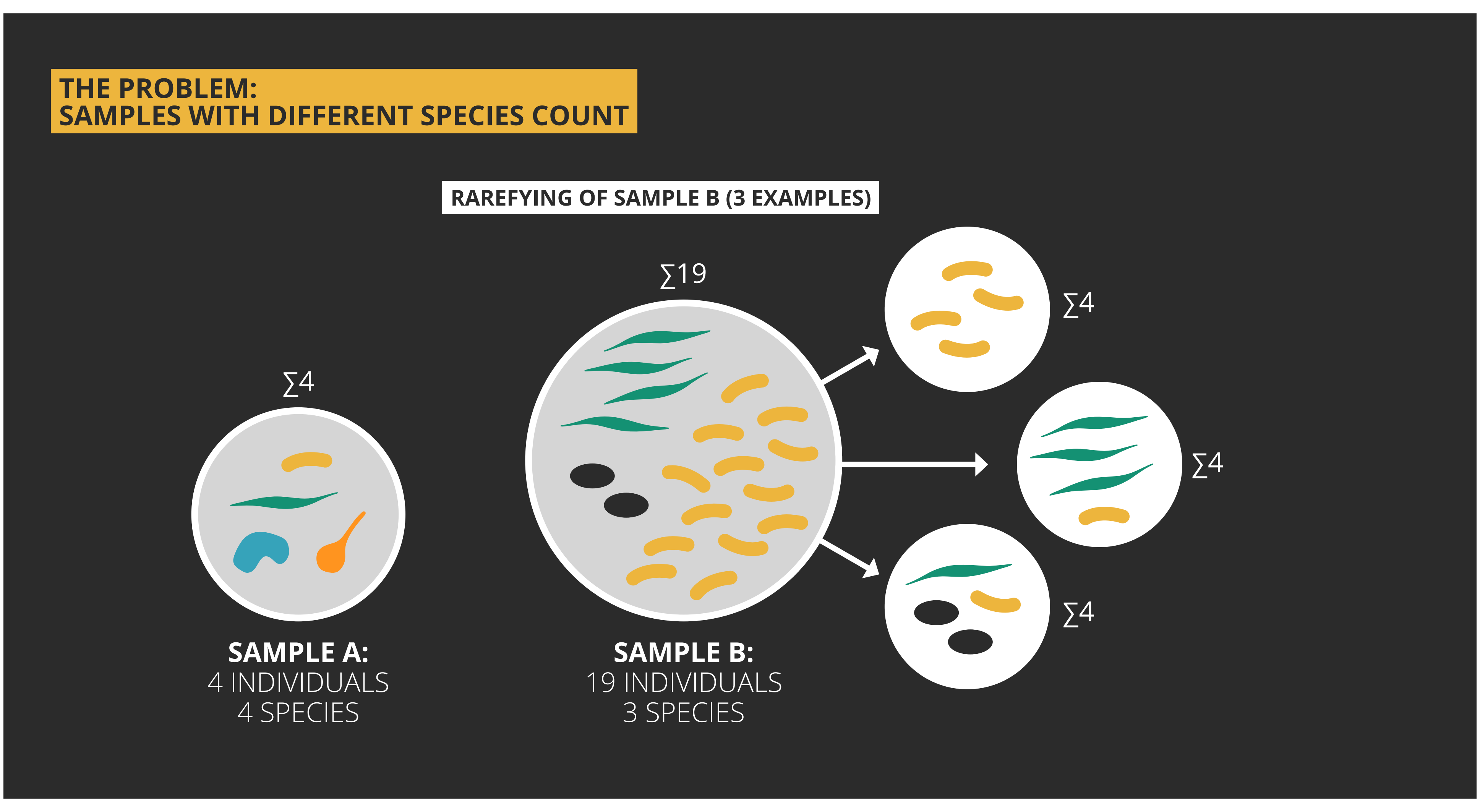
# Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities

https://peerj.com/articles/9593/

## BACKGROUND

Analysis of species count data in ecology often requires normalization to an identical sample size. **Rarefying** (random subsampling without replacement), which is the standard method for normalization, **incurs sampling error**, impairing the reproducibility and potentially distorting the community structure.

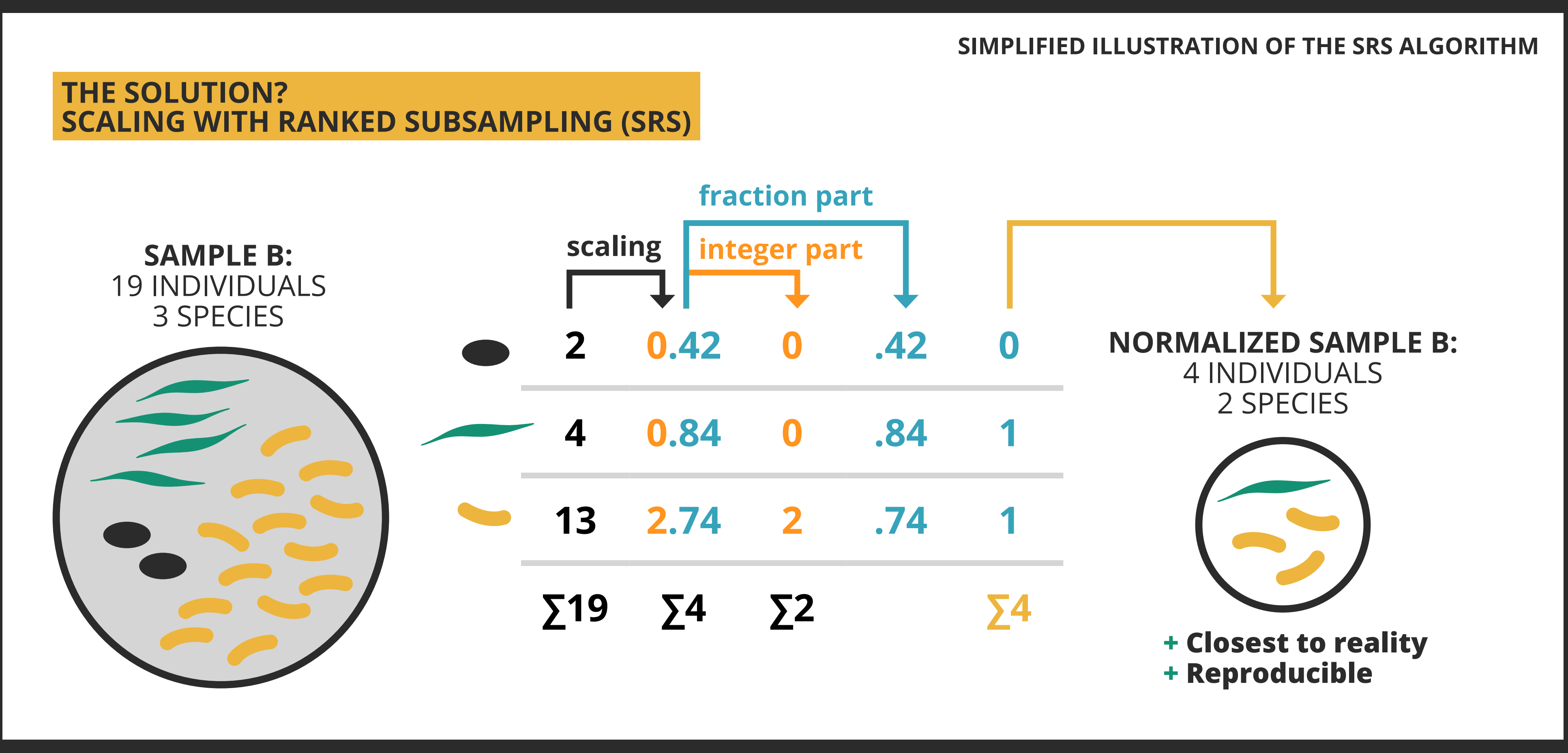


Here we introduce a **normalization method for species count data called scaling with ranked subsampling (SRS)** and demonstrate its suitability for analyzing microbial communities.

## METHODS

SRS consists of two steps:

1. **The scaling step.** The counts for all species or operational taxonomic units (OTUs) are divided by a scaling factor chosen in such a way that the sum of scaled counts equals the selected total number of counts $C_{min}$. The relative frequencies of all OTUs remain unchanged.
2. **Ranked subsampling step.** Non-integer count values are converted into integers by an algorithm that minimizes subsampling error with regard to the population structure (relative frequencies of species or OTUs) while keeping the total number of counts equal $C_{min}$.



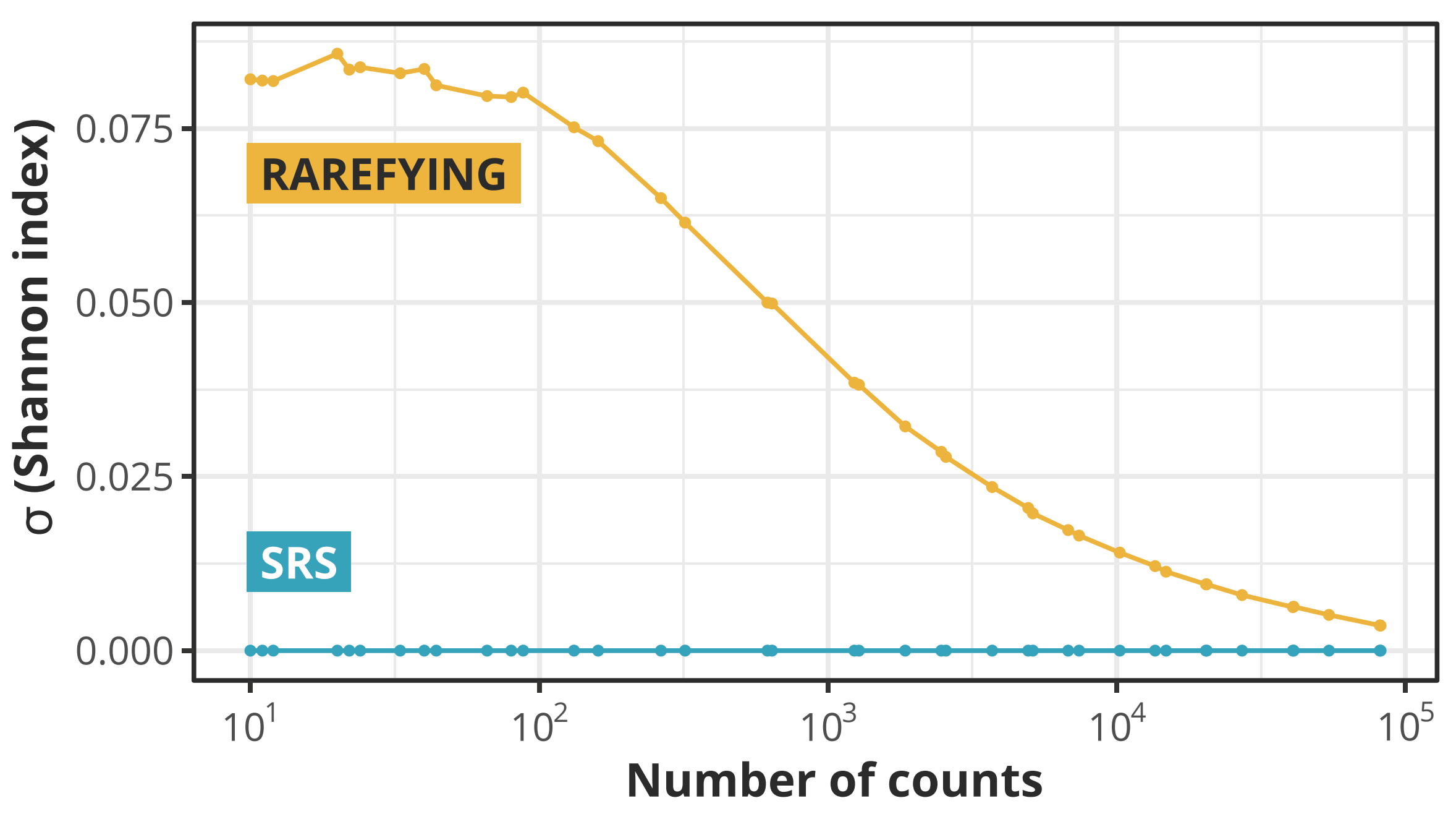
|  |  |  |  |  |
|---|---|---|---|---|
| 2 | 0.42 | 0 | .42 | 0 |
| 4 | 0.84 | 0 | .84 | 1 |
| 13 | 2.74 | 2 | .74 | 1 |
| ∑19 | ∑4 | ∑2 |  | ∑4 |

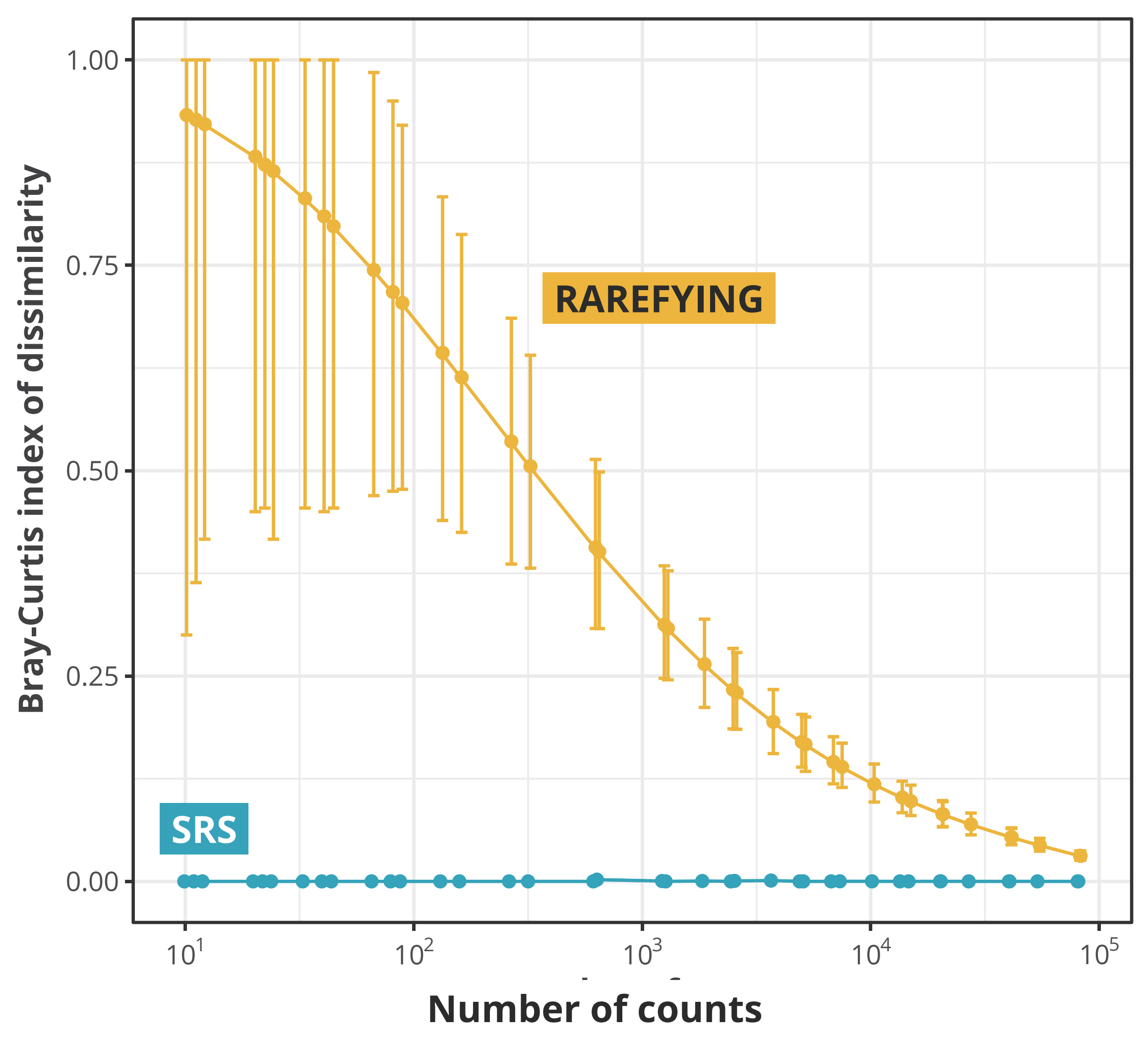
SRS and rarefying were compared by normalizing a test library representing a soil bacterial community. Common biodiversity and population structure parameters were determined for libraries normalized to different sizes by rarefying as well as SRS with 10,000 replications each.

## RESULTS

**SRS showed greater reproducibility and preserved OTU frequencies and alpha diversity better than rarefying.** The variance in Shannon diversity increased with the reduction of the library size after rarefying but remained zero for SRS. Relative abundances of OTUs strongly varied among libraries generated by rarefying, whereas libraries normalized by SRS showed only negligible variation.

Bray-Curtis index of dissimilarity revealed a large variation in species composition, which reached complete difference (not a single OTU shared) among some libraries rarefied to a small size. The dissimilarity among replicated libraries normalized by SRS remained negligibly low at each library size.





## CONCLUSIONS

**Normalization of OTU or species counts by scaling with ranked subsampling preserves the original community structure by minimizing subsampling error. We therefore propose SRS for the normalization of biological count data.**

PeerJ
Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities.
PeerJ 4:e9593 DOI: 10.7717/peerj.9593
https://peerj.com/articles/9593/