Iterative Reference Selection for Normalizing Compositional Microbiome Data

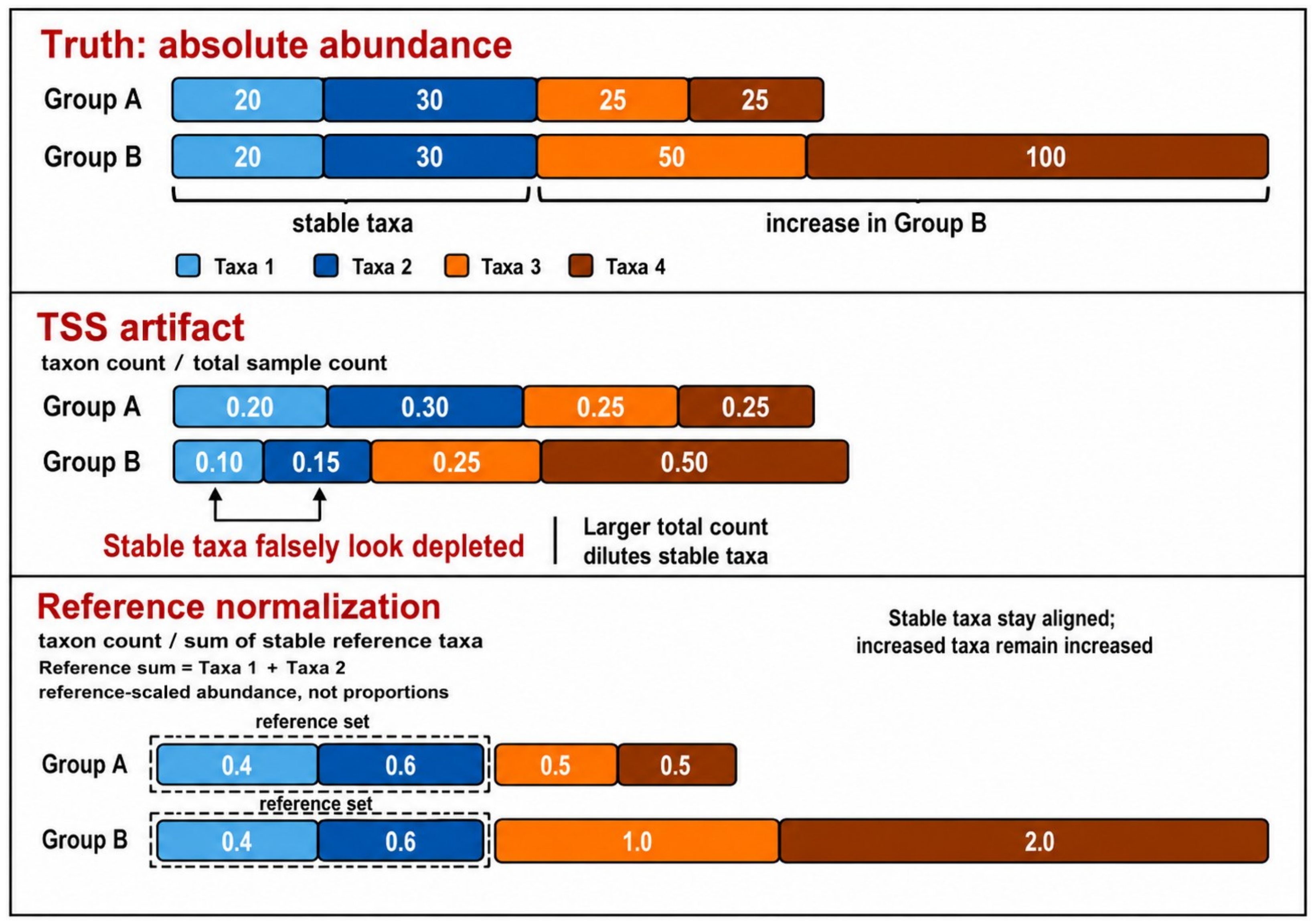
Washington University in St. Louis School of Medicine

Yiming Shi[1], Lili Liu[1], Jun Chen[2], Kristine M. Wylie[3], Todd N. Wylie[3], Sung Hee Park[1], Ruiwen Zhou[1], Yin Cao[4,5,6], Stephanie A. Fritz[3,7], Molly J. Stout[8], M. Cristina Vazquez Guillamet[7,9], Lei Liu[1]

[1] Institute for Informatics, Data Science & Biostatistics, Washington University in St. Louis; [2] Mayo Clinic; [3] Department of Pediatrics, Washington University; [4] Department of Surgery, WUSM; [5] Siteman Cancer Center; [6] Gastroenterology, WUSM; [7] Infectious Diseases, WUSM; [8] University of Michigan; [9] Pulmonary and Critical Care Medicine, WUSM

Background

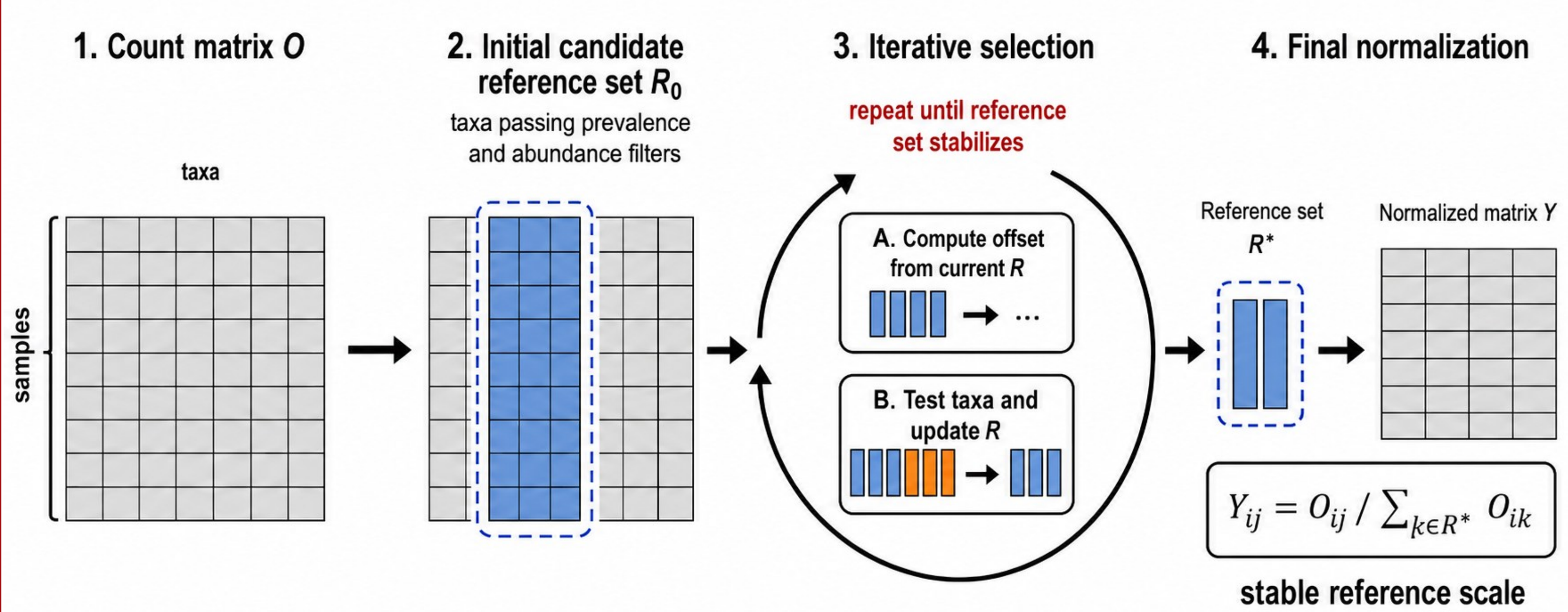
Compositional nature is a fundamental characteristic of microbiome data: the abundances are often expressed as relative proportions due to unknown sampling fraction and variable sequencing depth.



Compositional normalization can make stable taxa appear differentially abundant. When total microbial load differs across samples, standard normalization can distort fold changes: stable taxa may appear depleted and differential taxa can induce spurious negative correlations.

IRS uses stable reference taxa to preserve true relative changes while reducing compositional artifacts.

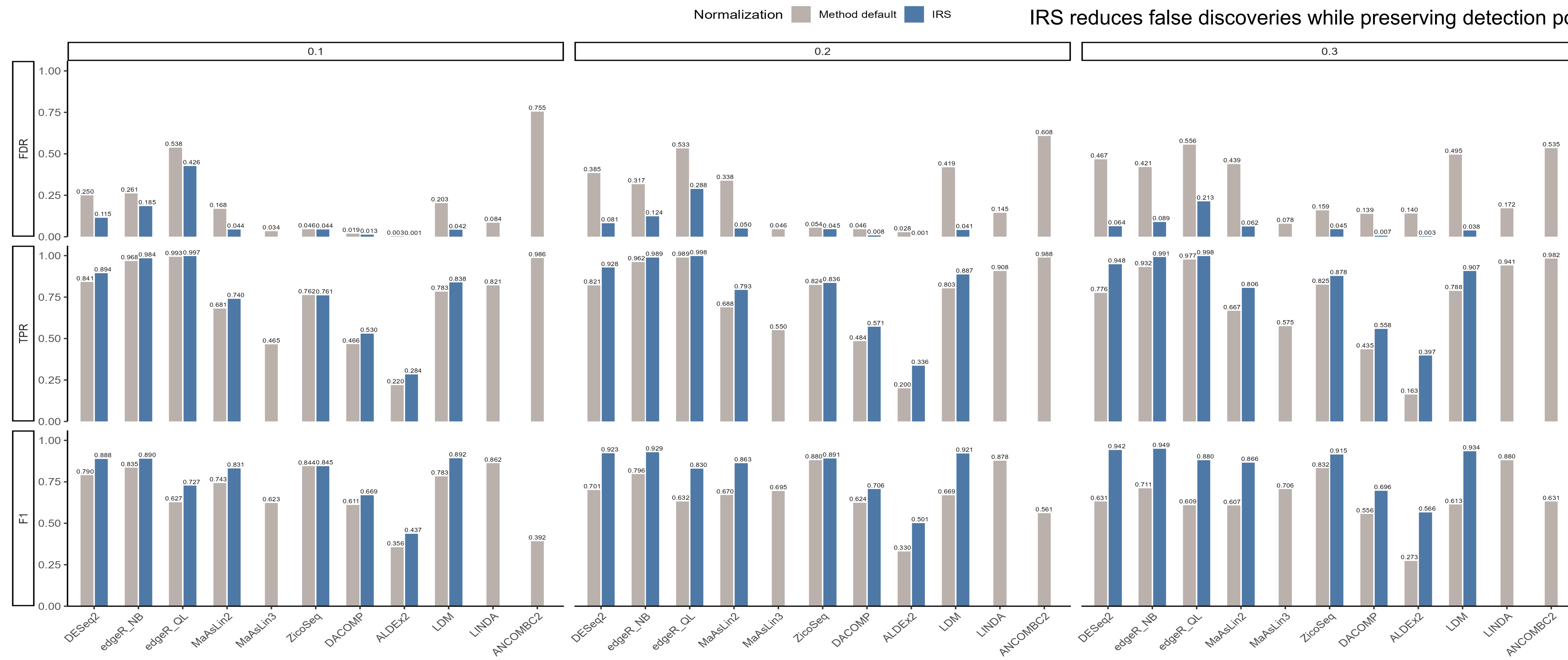
IRS: Iterative reference selection normalization



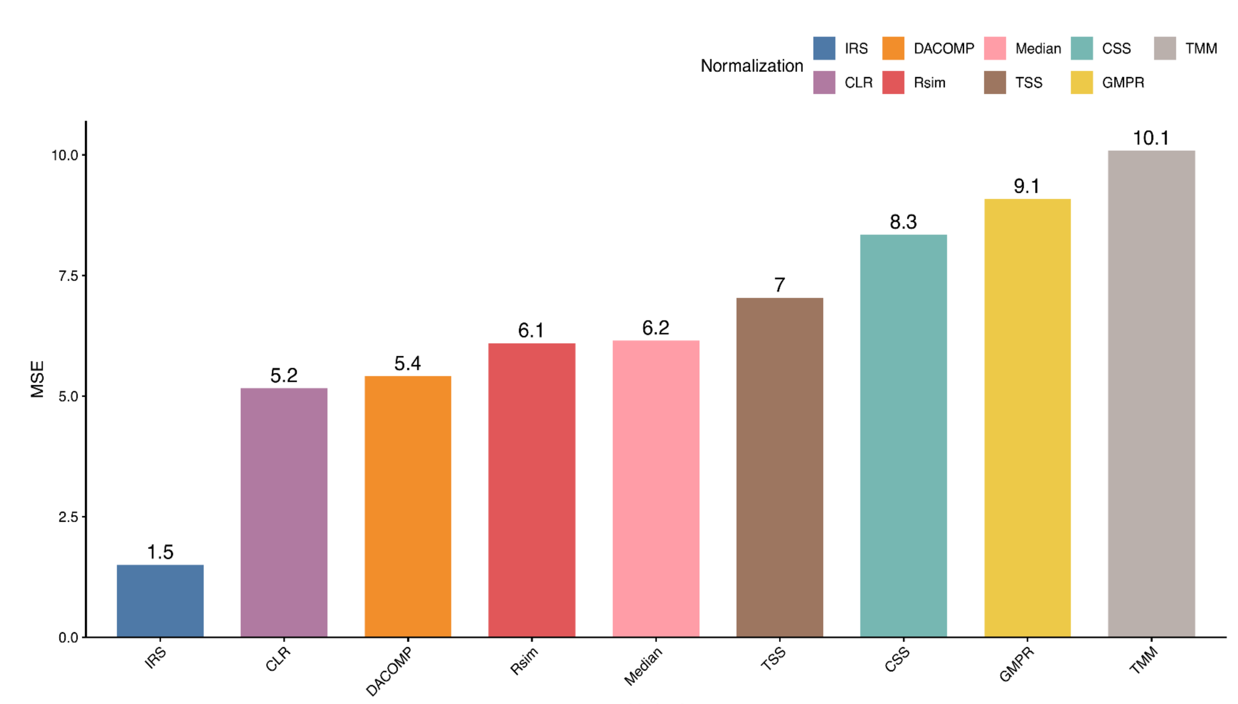
IRS iteratively selects stable taxa as an internal normalization reference.

Simulation studies

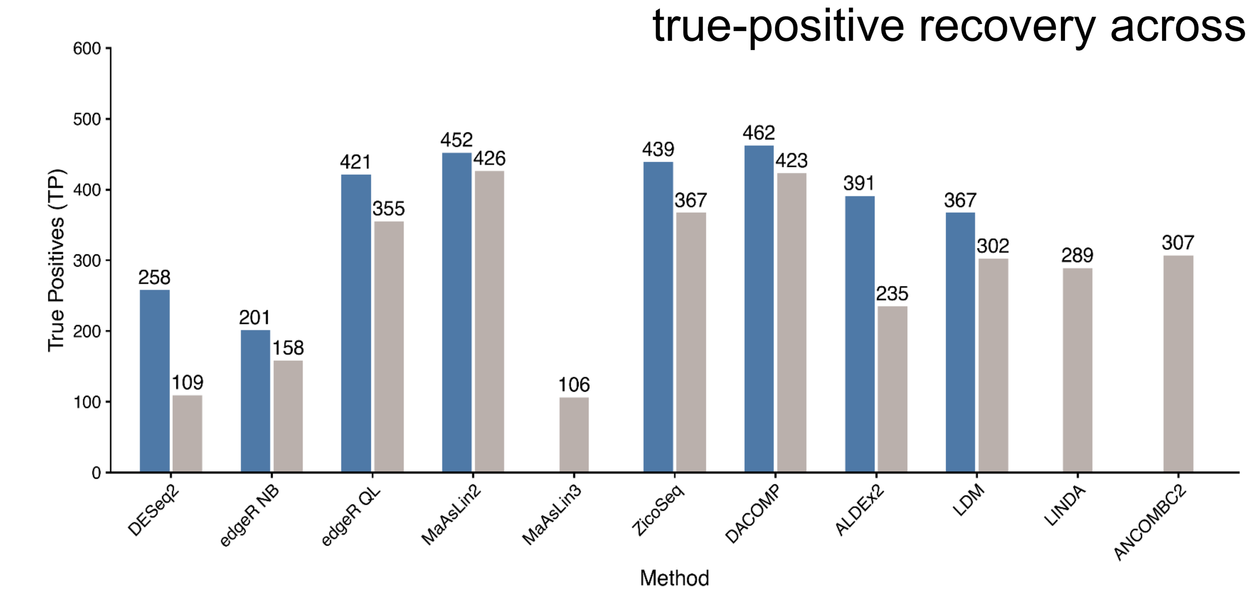
Simulated count data with known true signals benchmark normalization methods.
IRS reduces false discoveries while preserving detection power.



Real data validation



- The MetaCardis dataset with flow-cytometry microbial load was used to define an absolute-abundance truth set.
- IRS shows the lowest log2 fold-change MSE and strong true-positive recovery across DA frameworks.



Conclusions & Future Directions

- Iterative Reference selection method efficiently identifies a large, clean set of **non-differential** taxa, enabling reference-based normalization that **mitigates negative correlation bias** in compositional microbiome analyses.
- **Compared with existing normalization methods: More accurate size-factor estimates**, **smaller systematic bias**, and a **larger/cleaner reference set**.
- When used in differential abundance analysis models: **Lower FDR** with **equal or higher power** (higher TPR/F1) due to better compositional bias control.