Evaluation of microbiome association models under realistic and confounded conditions

Morgan Essex, Jakob Wirbel*, Sofia Forslund, Georg Zeller*

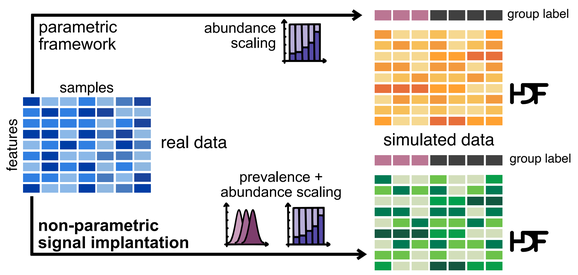
*both authors contributed equally

1. Introduction

- Testing for differential abundance is a fundamental task in metagenome-wide association studies, yet **there is no consensus on statistical methodology**
- Benchmarks typically use reshuffled real data, which lacks a ground truth, or **simulated data** generated by parametric methods, which **lacks evaluation** on its resemblance to real metagenomic data
- **Technical or biological confounders** further hamper reliability and reproducibility of findings in clinical applications of differential abundance testing, yet have thus far been largely ignored in benchmarks

2. Methods and Study Design

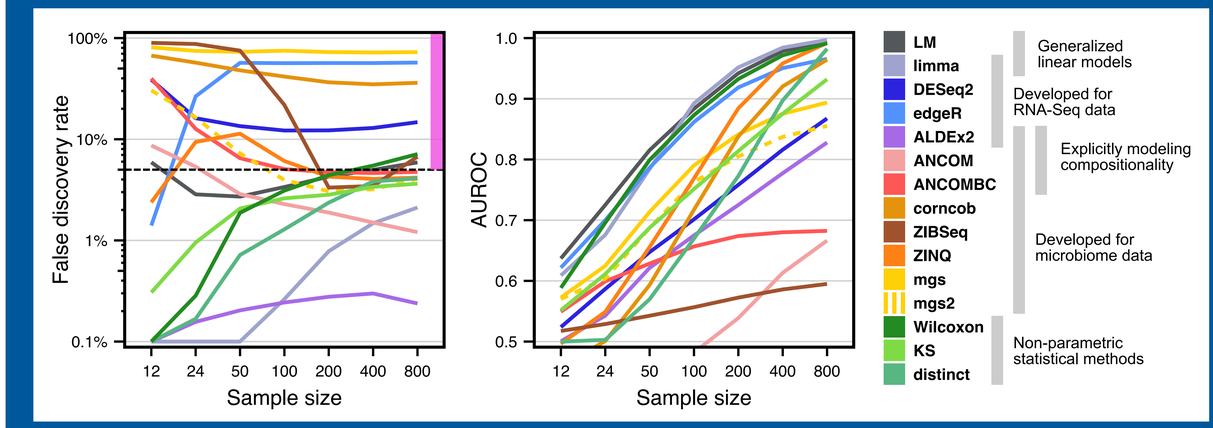
- We developed and validated a **novel simulation framework** mimicking case-control study designs that implants signals into **real metagenomic data**[1]



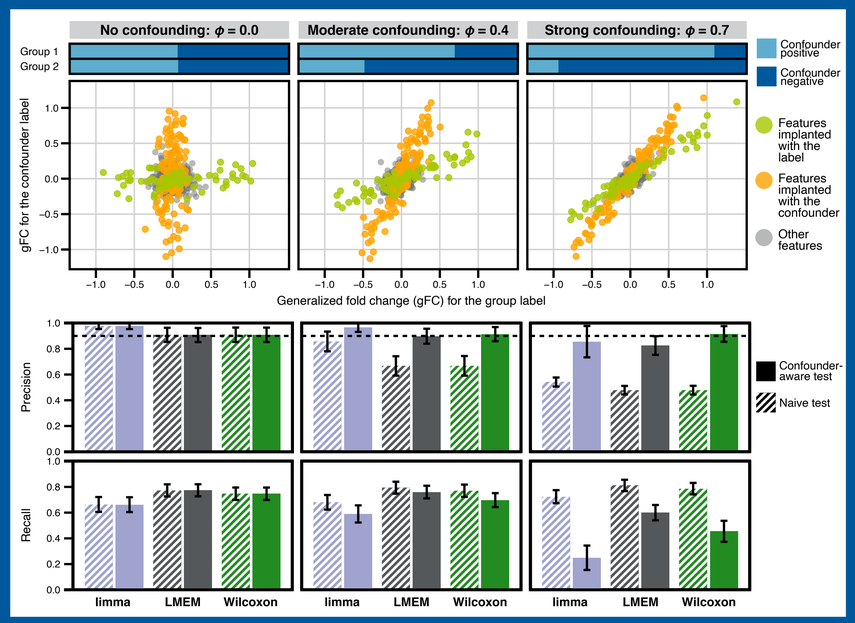
- We performed a large **benchmarking study,** including 6 simulation frameworks (5 parametric + ours) with varying effect sizes, and 15 differential abundance methods with different normalization techniques
- We extended our framework to simulate known **patterns of confounding**, mimicking e.g. technical batch effects or medication intake in disease cohorts
- We further benchmarked methods that performed well in our first evaluation on their ability to detect differentially abundant features under **varying strengths of confounding effects**
- Our code is **open source** to facilitate community development of e.g. other study designs, method benchmarks, or power analysis applications

SIMBA (**si**mulation of **m**icrobiome data with **b**iological **a**ccuracy)
BAMBI (**b**enchmarking **a**nalysis of **m**icro**b**iome **i**nference methods)

Many **differential abundance methods** applied in metagenome-wide association studies **fail to control the false discovery rate** or **offer limited sensitivity** to detect potential biomarkers



These issues are **exacerbated in the presence of confounding factors** known to plague real-world studies

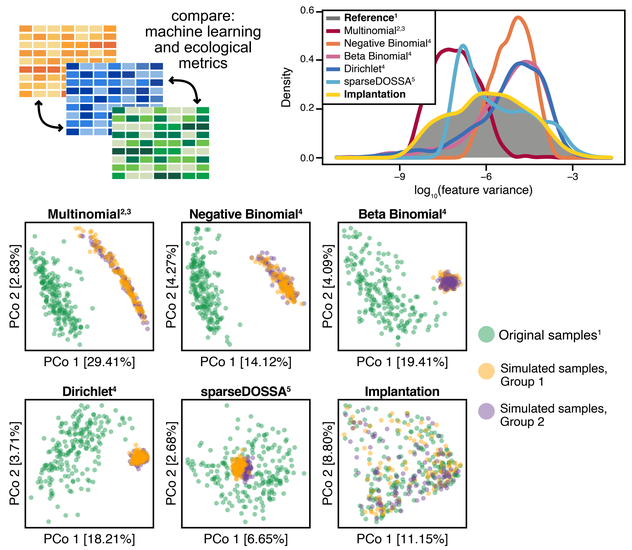


Linear models and the **Wilcoxon test** suffer the least from these issues, can be **adjusted for covariates,** and **should be preferred** over other statistical methods for robust differential abundance analysis of **microbiome data**

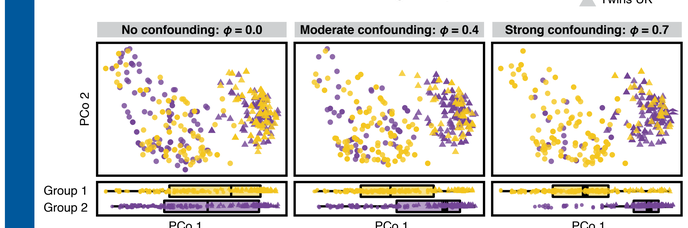
ECRC Experimental & Clinical Research Center
MDC MAX DELBRÜCK CENTER FOR MOLECULAR MEDICINE IN THE HELMHOLTZ ASSOCIATION
CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN
EMBL

3. Main Findings and Implications

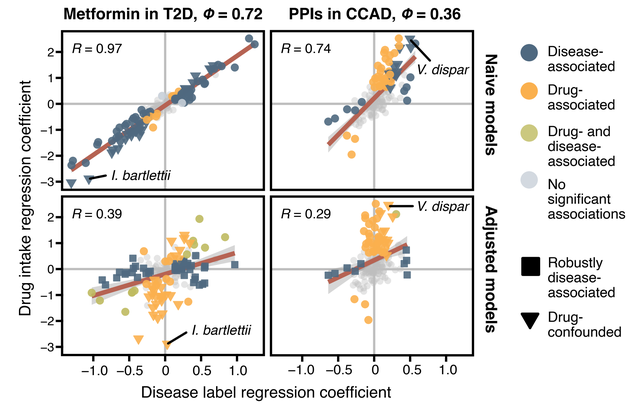
- Signal implantation, but not parametric simulations, can reproduce key characteristics of metagenomic data critical for robust method benchmarking



- Biased resampling produces simulated case-control groups aligned with a secondary variable to a pre-calibrated degree (below: study of origin, left: simulated confounder group)



- Linear mixed-effect models offer the most flexibility and scalability in their implementation, and are capable of disentangling drug- and disease-associated microbial features in real clinical data[7]



References

1) Zeevi et al., Personalized nutrition by prediction of glycemic responses. *Cell* (2015). **2) McMurdie and Holmes,** Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology* (2014). **3) Weiss et al.,** Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* (2017). **4) Hawinkel et al.,** A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics* (2019). **5) Ma et al.,** A statistical model for describing and simulating microbial community profiles. *bioRxiv* (2021). **6) Xie et al.,** Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Systems* (2016). **7) Forslund et al.,** Combinatorial, additive, and dose-dependent drug-microbiome associations. *Nature* (2021).