Tools for mining biomarkers from -omics data in case-control clinical studies

Host-Microbiome Factors in Cardiovascular Disease, PI Sofia Kirke Forslund[1,2,3,4]

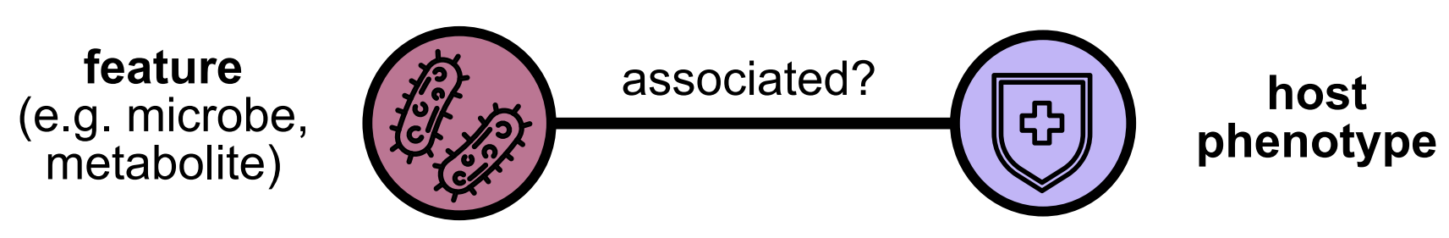
Morgan Essex,[1,2,3] Till Birkner,[1,2,3] Theda UP Bartolomaeus[1,2,3,4]

(1) Clinical Research Center of MDC and Charité (ECRC), Berlin, Germany
(2) Charité - Universitätsmedizin Berlin, a corporate member of Freie Universität and Humboldt-Universität zu Berlin, Berlin, Germany
(3) Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany
(4) German Cardiovascular Research Center (DZHK), partner site Berlin, Germany

Univariate association tests are sensitive, but susceptible to confounding

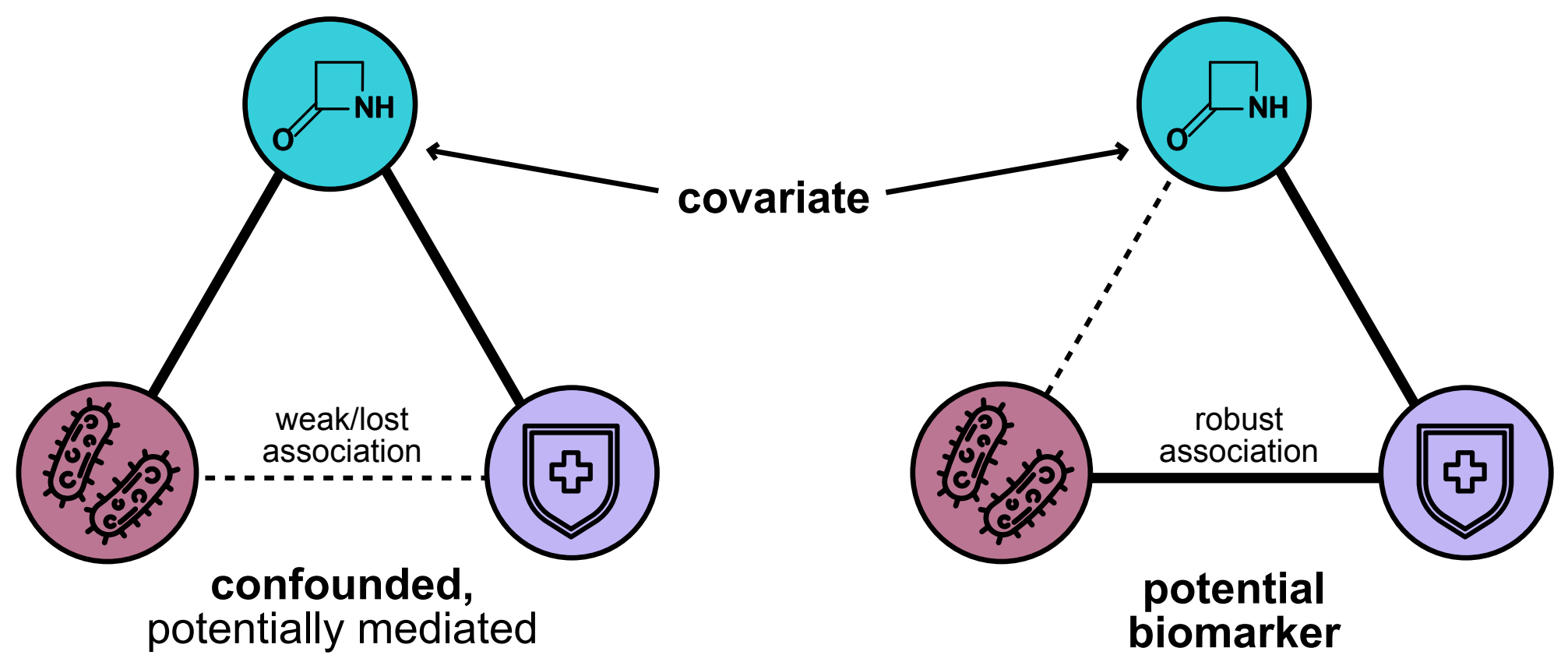
▶ Clinical microbiome research seeks to characterize **health and disease states** using **high-throughput molecular data** types, typically by comparing **case-control groups**

▶ **Univariate testing** compares groups **feature-by-feature** to identify disease- or phenotype-associated signals



▶ To ensure **robust** findings, and **generate more precise mechanistic hypotheses**, association testing frameworks must incorporate information about known **clinical covariates**

▶ If inclusion of a covariate (**adjustment**) weakens or displaces the feature-phenotype association, it is not robust!

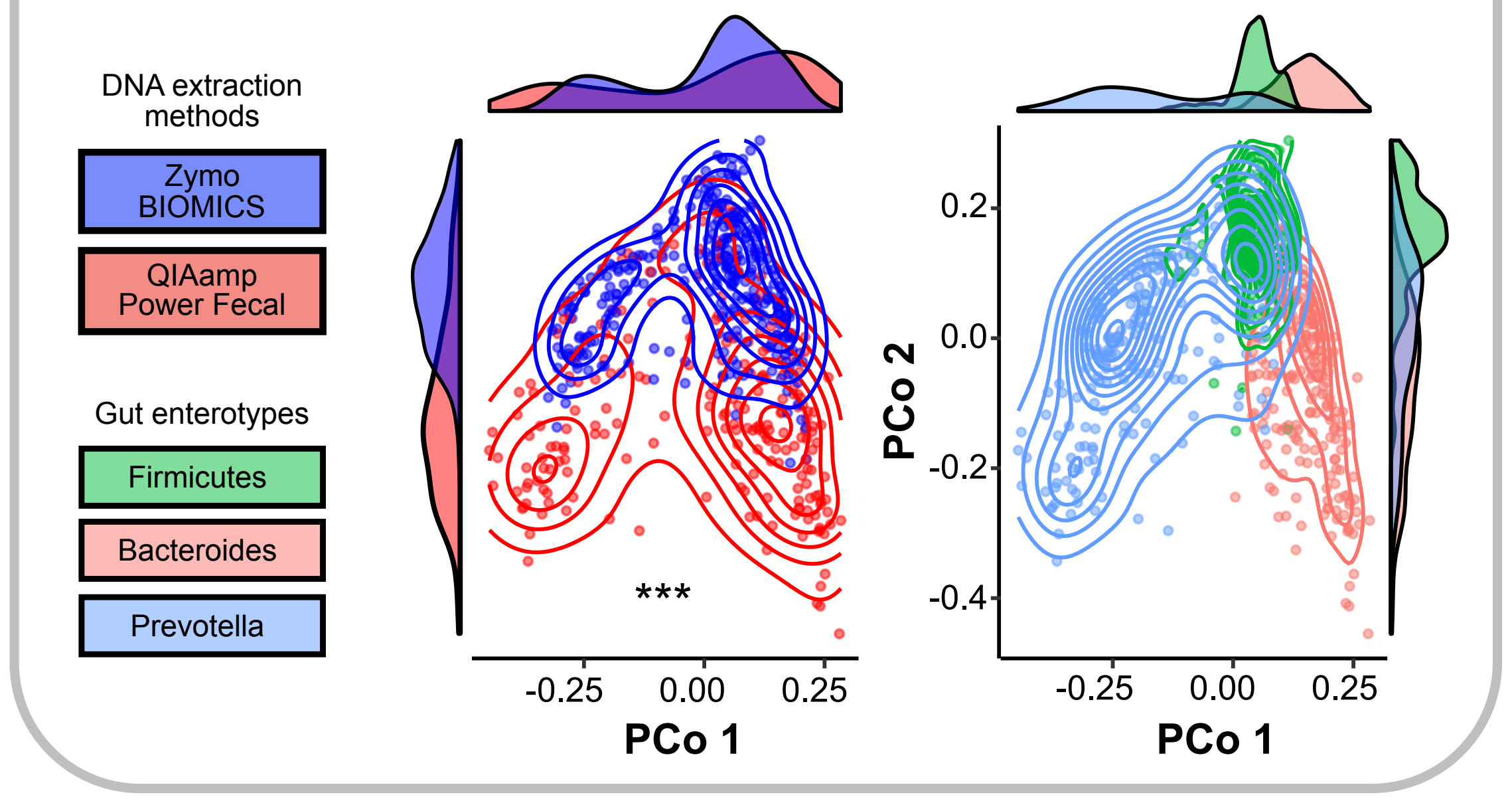


COLLECT Study¹

▶ Non-standardized **experimental protocols** can generate **technical effects** which can outweigh biological effects and obfuscate **downstream analysis**, if not accounted for

▶ We **benchmarked** several sample **storage, isolation, and extraction methods** and found the latter to explain about **6% of overall microbiome variability**

▶ Below, the extraction method dramatically impacted **enterotype determination**, with Zymo and Qiagen enriched for Firmicutes and Bacteroidetes, respectively

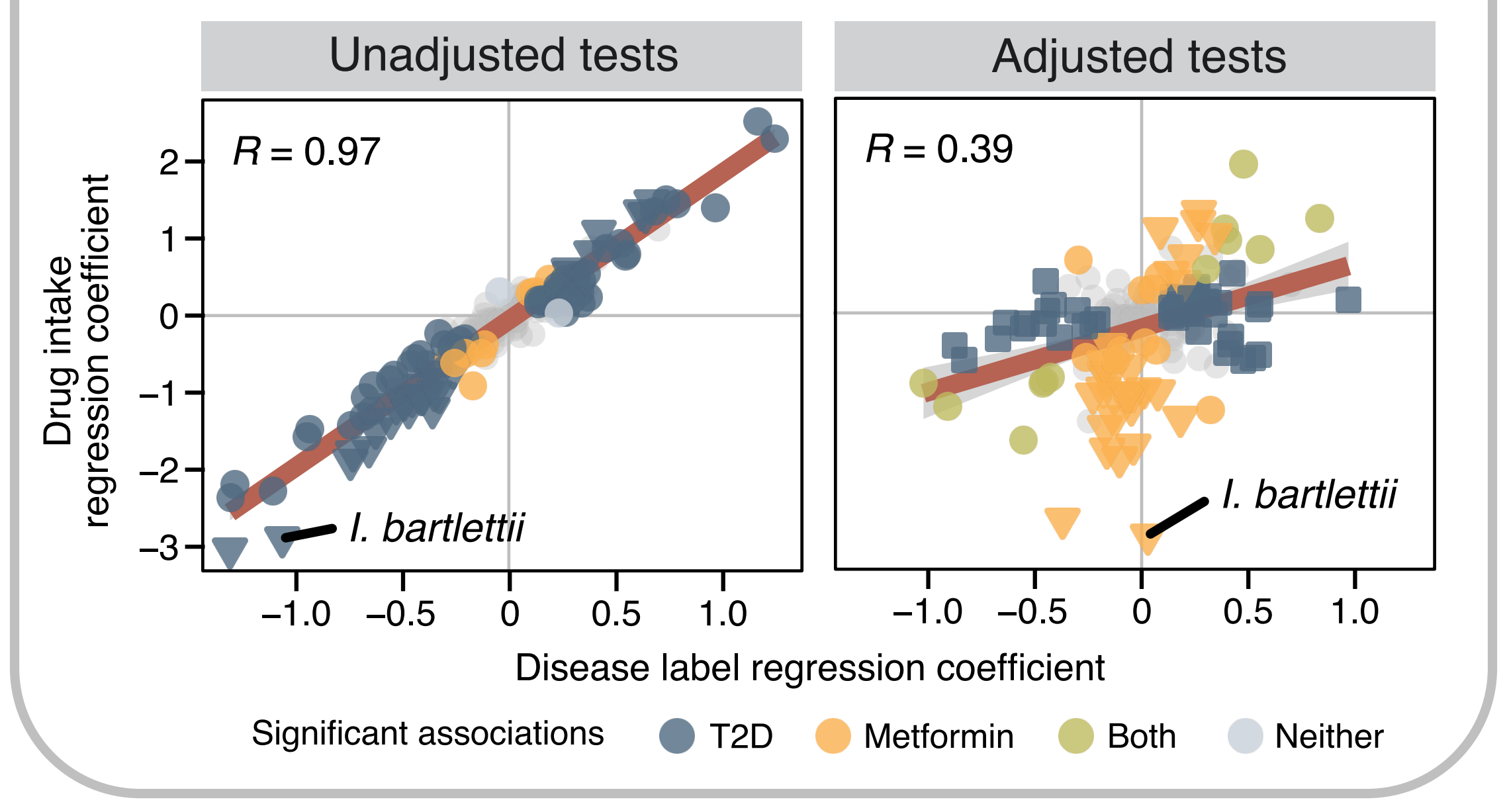


MetaCardis² and MetaDrugs

▶ **Multiple and combination drug therapies** are common in cardiometabolic disease groups, producing myriad effects on host systems and **potentially confounding** simple comparisons with **non-medicated healthy controls**

▶ ***Post hoc* stratification** and adjusted tests are able to diagnose confounding and **disentangle disease signals**

▶ Below, metformin intake is strongly correlated with T2D (ϕ=0.72), and many metformin-sensitive taxa appear **naively** T2D-associated in **unadjusted** tests

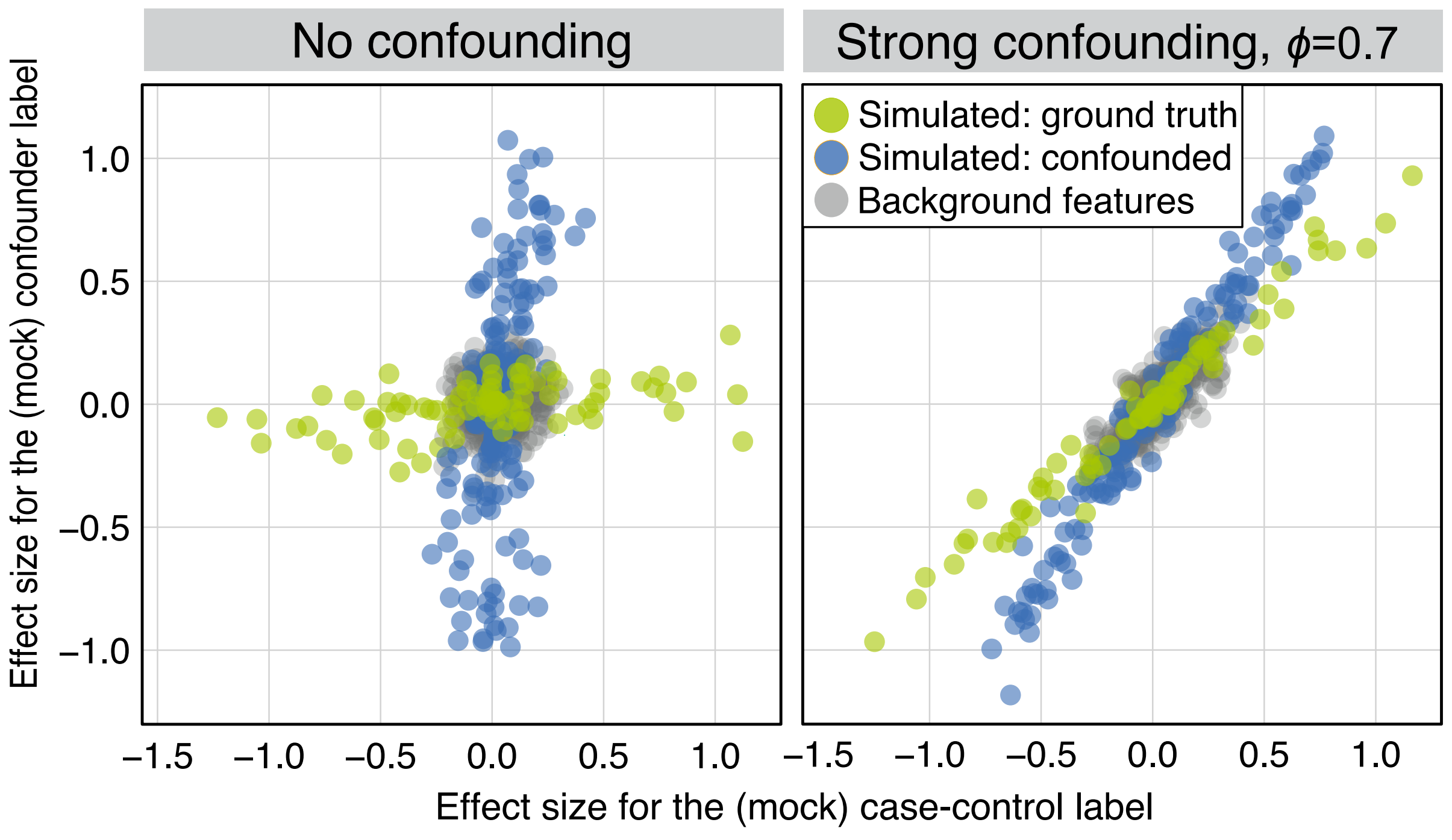


Adjusting for known confounders can disentangle true disease signals

▶ Previous differential abundance benchmarks relied on overly synthetic simulations and **oversimplified evaluations,** which generated recommendations that **do not generalize**

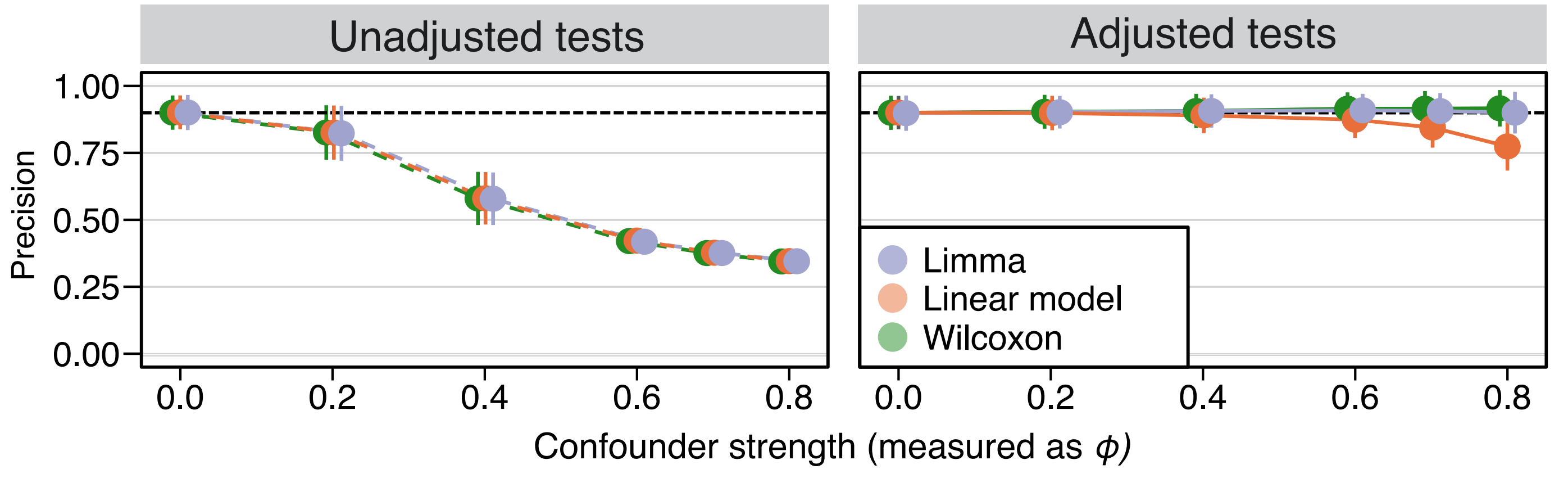
▶ We built and validated an open source **simulation framework in R** to implant calibrated signals into **real metagenomic data,** including **known confounding patterns**

▶ We used these simulations to perform a **comprehensive method benchmark** under the most **realistic conditions** possible



▶ **Unadjusted tests failed** to distinguish ground truth features and **suffered from low precision** under moderate to strong confounding, but adjusting for the (mock) confounder variable **restored good performance**

▶ **Linear models** are considerably more **flexible** than other methods tested



metadeconfoundR: a fast and flexible R package for robust association testing

▶ In a first step, metadeconfoundR calculates **standardized, non-parametric effect sizes** (Cliff's delta or Spearman correlations) paired with appropriate **statistical tests** to identify naively disease-associated features

▶ In a second step, **nested models** including covariates achieving significance in the first step are used in **likelihood ratio tests**, checking for **two conditions** needed to **classify the robustness** of feature associations

▶ **Rank-based methods** are robust to non-normal distributions; **mixed-effect models** are robust to pseudoreplication and repeated measures designs; **iterative, integrated status determination** robust to inclusion of an arbitrarily large number of covariates

▶ We have also developed a **sister R package** with similar logic geared toward **longitudinal biomarker analysis**: LongDat³

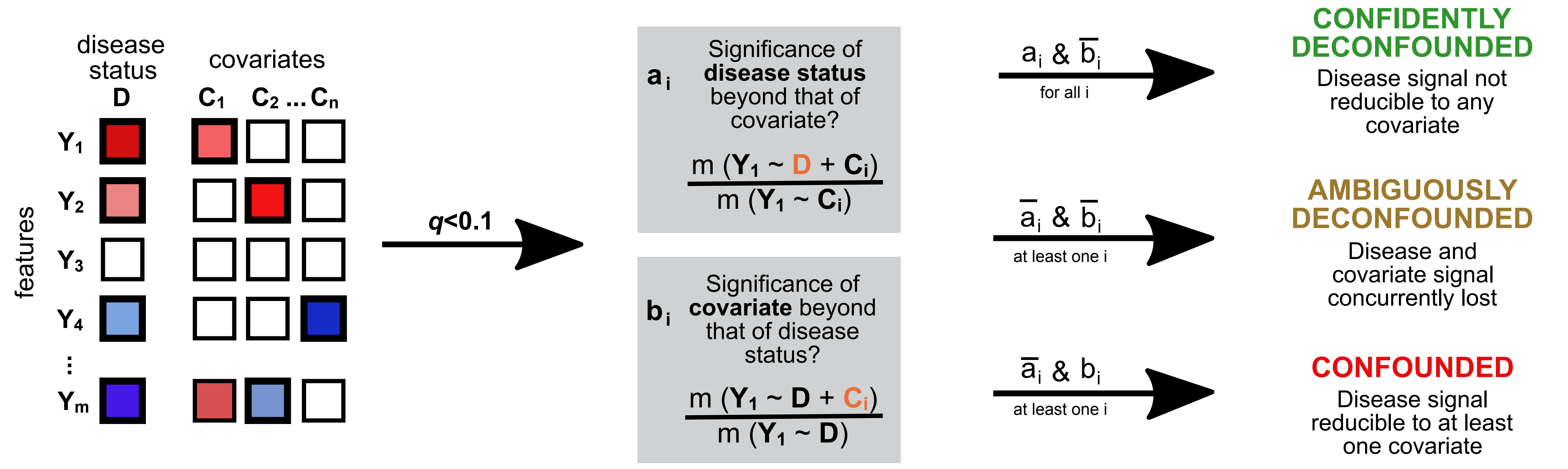
Naive association testing

between all feature-covariate pairs

Confounder modeling and *post hoc* testing

for all significant covariates, $\mathbf{C}_i$

Association status determination



disease status **D**

covariates $\mathbf{C_1}$ $\mathbf{C_2}$ ... $\mathbf{C_n}$

features $\mathbf{Y_1}$ $\mathbf{Y_2}$ $\mathbf{Y_3}$ $\mathbf{Y_4}$ ⋮ $\mathbf{Y_m}$

q<0.1

$\mathbf{a}_i$ Significance of **disease status** beyond that of covariate?

$$\frac{m\,(\mathbf{Y_1} \sim \mathbf{D} + \mathbf{C_i})}{m\,(\mathbf{Y_1} \sim \mathbf{C_i})}$$

$\mathbf{b}_i$ Significance of **covariate** beyond that of disease status?

$$\frac{m\,(\mathbf{Y_1} \sim \mathbf{D} + \mathbf{C_i})}{m\,(\mathbf{Y_1} \sim \mathbf{D})}$$

a_i & $\overline{b_i}$ for all i → **CONFIDENTLY DECONFOUNDED** Disease signal not reducible to any covariate

$\overline{a_i}$ & $\overline{b_i}$ at least one i → **AMBIGUOUSLY DECONFOUNDED** Disease and covariate signal concurrently lost

$\overline{a_i}$ & b_i at least one i → **CONFOUNDED** Disease signal reducible to at least one covariate

References

(1) Bartolomaeus TUP *et al.* Quantifying technical confounders in microbiome studies. *Cardiovascular Research* (2021); **(2) Forslund SK ... Birkner, Till** *et al.* Combinatorial, additive and dose-dependent drug–microbiome associations. *Nature* (2021); **(3) Chen CY, Löber U, and Forslund, SK**. LongDat: an R package for covariate-sensitive longitudinal analysis of high-dimensional data. *Bioinformatics Advances* (2023)

MAX DELBRÜCK CENTER
SEVENTH FRAMEWORK PROGRAMME
DFG Deutsche Forschungsgemeinschaft
EMBL
ECRC Experimental & Clinical Research Center
HELMHOLTZ
DZHK DEUTSCHES ZENTRUM FÜR HERZ-KREISLAUF-FORSCHUNG E.V.
CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN