Figure: Scatter plots of mean relative abundances (log scale) versus log2 fold change in PD across three columns (Dataset 1, Dataset 2, Replicated) for multiple methods. Each panel reports MRA and FC odds ratios and P-values.

Columns: Dataset 1, Dataset 2, Replicated

Rows (methods): GLM CLR, t-test CLR, t-test TSS, limma-voom, GLM TSS, KW TSS, LEfSe, KW CLR, ANCOM-BC, SAMseq, fitFeatureModel, ALDEx2 Wilcoxon, ALDEx2 t-test, DESeq2, fitZIG, edgeR RLE, edgeR TMM, t-test rCLR, GLM NBZI, GLM rCLR, ANCOM, KW rCLR, baySeq

GLM CLR — Dataset 1: MRA: OR=1.31, P=0.5; FC: OR=7.28, P=5e-04. Dataset 2: MRA: OR=1.77, P=0.1; FC: OR=3.51, P=0.001. Replicated: MRA: OR=3.33, P=0.03; FC: OR=19.05, P=9e-05.

t-test CLR — Dataset 1: MRA: OR=1.31, P=0.5; FC: OR=7.28, P=5e-04. Dataset 2: MRA: OR=1.77, P=0.1; FC: OR=3.51, P=0.001. Replicated: MRA: OR=3.33, P=0.03; FC: OR=19.05, P=9e-05.

t-test TSS — Dataset 1: MRA: OR=0.88, P=0.8; FC: OR=6.56, P=2e-04. Dataset 2: MRA: OR=1.71, P=0.1; FC: OR=4.47, P=2e-04. Replicated: MRA: OR=1.98, P=0.2; FC: OR=23.52, P=1e-05.

limma-voom — Dataset 1: MRA: OR=1.42, P=0.4; FC: OR=12.36, P=3e-05. Dataset 2: MRA: OR=1.4, P=0.3; FC: OR=4.48, P=2e-04. Replicated: MRA: OR=2.31, P=0.1; FC: OR=Inf, P=9e-07.

GLM TSS — Dataset 1: MRA: OR=1.1, P=0.8; FC: OR=6.24, P=4e-04. Dataset 2: MRA: OR=1.97, P=0.06; FC: OR=4.14, P=7e-04. Replicated: MRA: OR=1.81, P=0.2; FC: OR=21.97, P=2e-05.

KW TSS — Dataset 1: MRA: OR=1.42, P=0.4; FC: OR=12.36, P=3e-05. Dataset 2: MRA: OR=2.42, P=0.02; FC: OR=8.15, P=5e-06. Replicated: MRA: OR=2.52, P=0.06; FC: OR=Inf, P=4e-07.

LEfSe — Dataset 1: MRA: OR=1.42, P=0.4; FC: OR=12.36, P=3e-05. Dataset 2: MRA: OR=4.53, P=2e-04; FC: OR=6.4, P=1e-04. Replicated: MRA: OR=2.99, P=0.03; FC: OR=Inf, P=9e-07.

KW CLR — Dataset 1: MRA: OR=2.09, P=0.2; FC: OR=3.72, P=0.04. Dataset 2: MRA: OR=2.3, P=0.02; FC: OR=2.65, P=0.02. Replicated: MRA: OR=3.14, P=0.06; FC: OR=Inf, P=1e-04.

ANCOM-BC — Dataset 1: MRA: OR=1.84, P=0.2; FC: OR=3.76, P=0.01. Dataset 2: MRA: OR=0.98, P=1; FC: OR=12.83, P=7e-13. Replicated: MRA: OR=2.36, P=0.07; FC: OR=26.81, P=2e-06.

SAMseq — Dataset 1: MRA: OR=1.33, P=0.5; FC: OR=3.31, P=0.004. Dataset 2: MRA: OR=1.79, P=0.06; FC: OR=5.75, P=2e-07. Replicated: MRA: OR=1.42, P=0.4; FC: OR=20.08, P=1e-07.

fitFeatureModel — Dataset 1: MRA: OR=1.62, P=0.3; FC: OR=9.46, P=3e-04. Dataset 2: MRA: OR=2.42, P=0.03; FC: OR=5.12, P=0.001. Replicated: MRA: OR=1.6, P=0.4; FC: OR=13.79, P=0.0002.

ALDEx2 Wilcoxon — Dataset 1: MRA: OR=4.69, P=0.01; FC: OR=4.12, P=0.08. Dataset 2: MRA: OR=4.09, P=6e-04; FC: OR=2.31, P=0.05. Replicated: MRA: OR=5.76, P=0.02; FC: OR=Inf, P=5e-04.

ALDEx2 t-test — Dataset 1: MRA: OR=2.92, P=0.1; FC: OR=6.21, P=0.1. Dataset 2: MRA: OR=5, P=3e-04; FC: OR=1.87, P=0.2. Replicated: MRA: OR=5.1, P=0.03; FC: OR=Inf, P=0.001.

DESeq2 — Dataset 1: MRA: OR=2.95, P=0.03; FC: OR=7.94, P=0.002. Dataset 2: MRA: OR=2.12, P=0.02; FC: OR=4.45, P=5e-05. Replicated: MRA: OR=3.33, P=0.03; FC: OR=8.69, P=9e-04.

fitZIG — Dataset 1: MRA: OR=0.27, P=4e-04; FC: OR=10.91, P=1e-09. Dataset 2: MRA: OR=1.11, P=0.8; FC: OR=6.5, P=5e-09. Replicated: MRA: OR=0.59, P=0.2; FC: OR=24.43, P=2e-11.

edgeR RLE — Dataset 1: MRA: OR=0.67, P=0.3; FC: OR=108.79, P=4e-16. Dataset 2: MRA: OR=0.7, P=0.3; FC: OR=8.72, P=9e-12. Replicated: MRA: OR=0.71, P=0.4; FC: OR=63.86, P=1e-11.

edgeR TMM — Dataset 1: MRA: OR=0.59, P=0.3; FC: OR=Inf, P=8e-07. Dataset 2: MRA: OR=1.43, P=0.3; FC: OR=5.73, P=4e-06. Replicated: MRA: OR=0.79, P=0.8; FC: OR=Inf, P=1e-05.

t-test rCLR — Dataset 1: MRA: OR=2.25, P=0.1; FC: OR=5.28, P=0.02. Dataset 2: MRA: OR=2.05, P=0.08; FC: OR=1.98, P=0.2. Replicated: MRA: OR=5.36, P=0.1; FC: OR=Inf, P=0.03.

GLM NBZI — Dataset 1: MRA: OR=1.02, P=1; FC: OR=Inf, P=5e-09. Dataset 2: MRA: OR=1.71, P=0.1; FC: OR=Inf, P=7e-12. Replicated: MRA: OR=0.63, P=0.5; FC: OR=Inf, P=2e-06.

GLM rCLR — Dataset 1: MRA: OR=2.02, P=0.2; FC: OR=4.88, P=0.03. Dataset 2: MRA: OR=1.86, P=0.1; FC: OR=2.09, P=0.1. Replicated: MRA: OR=Inf, P=0.03; FC: OR=Inf, P=0.06.

ANCOM — Dataset 1: MRA: OR=6.52, P=0.06; FC: OR=3.54, P=0.4. Dataset 2: MRA: OR=Inf, P=3e-04; FC: OR=2.95, P=0.2. Replicated: MRA: OR=6.54, P=0.06; FC: OR=Inf, P=0.02.

KW rCLR — Dataset 1: MRA: OR=1.24, P=0.6; FC: OR=2.64, P=0.01. Dataset 2: MRA: OR=0.63, P=0.1; FC: OR=2.67, P=0.002. Replicated: MRA: OR=1.27, P=0.6; FC: OR=6.61, P=0.002.

baySeq — Dataset 1: MRA: OR=0.18, P=3e-04; FC: OR=3.55, P=0.02. Dataset 2: MRA: OR=0.74, P=0.5; FC: OR=17.38, P=1e-07. Replicated: MRA: OR=1.02, P=1; FC: OR=Inf, P=0.008.

Y-axis: Mean relative abundances (log scale). X-axis: log2 fold change in PD.

Fig. S2. Detection of differentially abundant genera as a function of mean relative abundance and effect size when performing differential abundance testing on filtered taxonomic data.

To observe what type of differentially abundant genera were being detected by methods based on mean relative abundance and effect size when performed on filtered taxonomic data, the mean relative abundances of tested genera (on log scale) and log2 fold change of genera in PD were plotted for dataset 1 (left column), dataset 2 (middle column), and replicated signatures (right column), and repeating these plots for each DA method (rows). Fisher's exact test was used to determine if DA signatures were enriched for more or less common genera, and genera with absolute fold change of 1.3 or higher. Individual points in the plots represent each tested genus, and is plotted according to that specific genus' mean relative abundance in cases and controls and log2 fold change in PD. Composite mean relative abundances and fold changes were used to plot points for replicated DA signatures by taking the average between datasets. For each method, points are shaded black if a method detected a particular genus as differentially abundant, and grey if not detected as such. Horizontal dashed lines represent the median mean relative abundance for either dataset 1, dataset 2, or replicated signatures. Vertical dashed lines represent a fold change of ~1.3 in positive and negative directions. Blue labeled methods are methods who were found to have consistently higher than average concordances in both datasets and replicated signatures, while red labeled methods were found to have consistently lower than average concordances. Grey labeled methods are those that were found to have varied mean concordances. MRA: mean relative abundance, results from Fisher's exact test testing enrichment of more or less common genera in detected DA signatures; FC: fold change, results from Fisher's exact test testing enrichment of genera with absolute fold changes > or < 1.3 in detected DA signatures; OR: odds ratio; P: P value of OR; GLM: generalized linear model; CLR: centered log-ratio; KW: Kruskal-Wallis; TSS: total sum scaling (relative abundances); rCLR: robust centered log-ratio transformation with matrix completion; RLE: relative log expression; TMM: trimmed mean of M-values; NBZI: negative binomial zero-inflated.