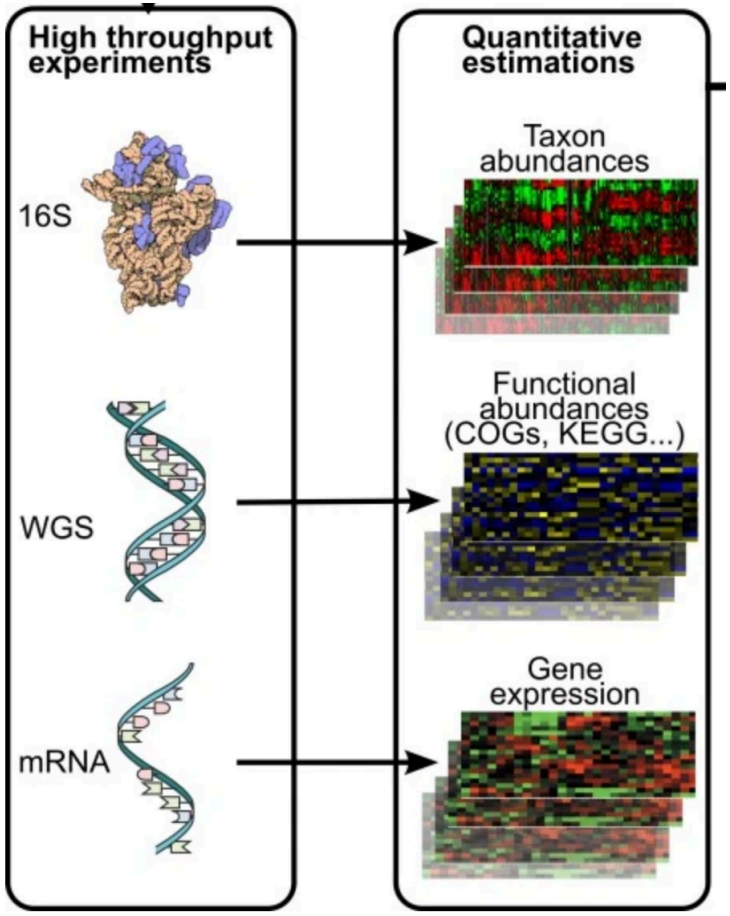
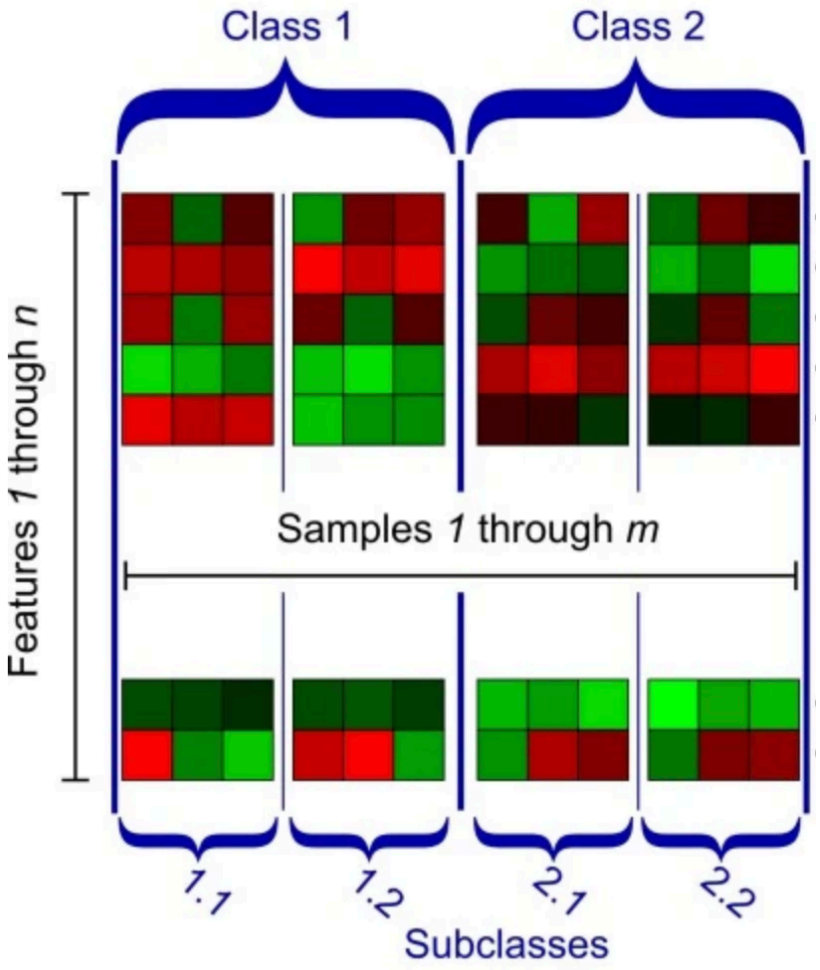
CEfSe: A Pure Python-based Re-implementation

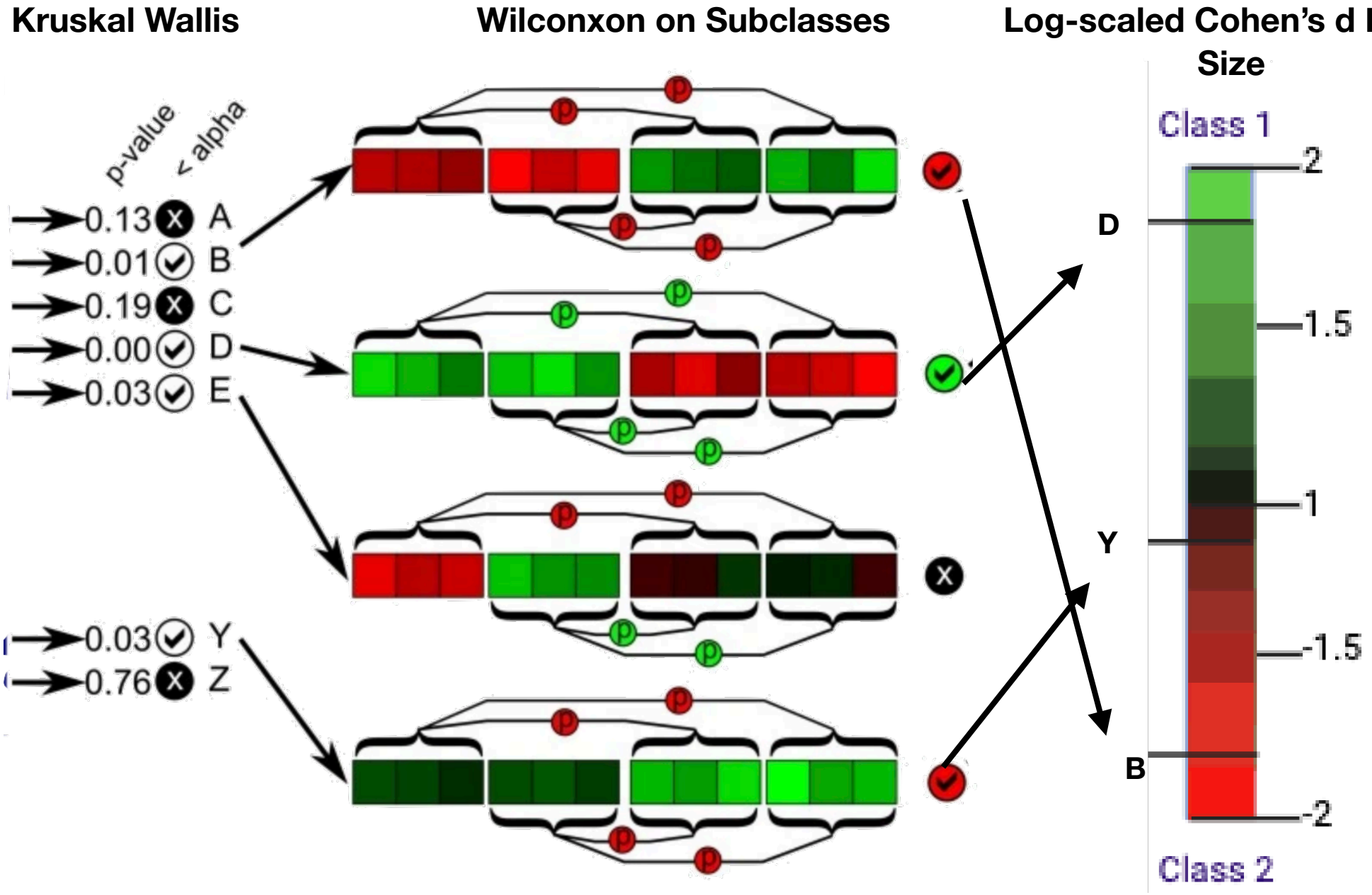
Raw Data



Input Matrix



# of LEfSe using Cohen's d effect size



LEfSe: Biomarker-discovery tool used to detect feature that differ between biological groups (1).
LDA: LEfSe, implemented using R in the LEfSe pipeline, is used to estimate how well each feature separates the classes by projecting them onto a discriminant axis (1).

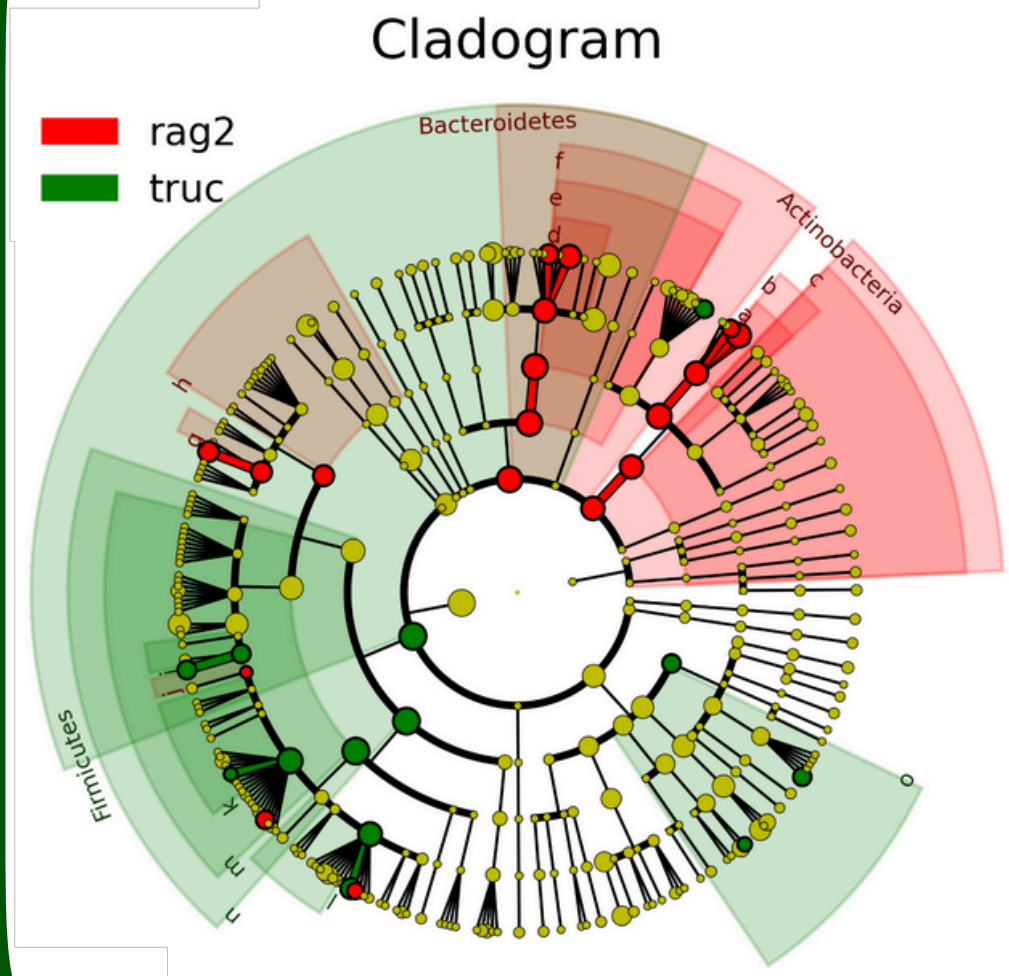
Data

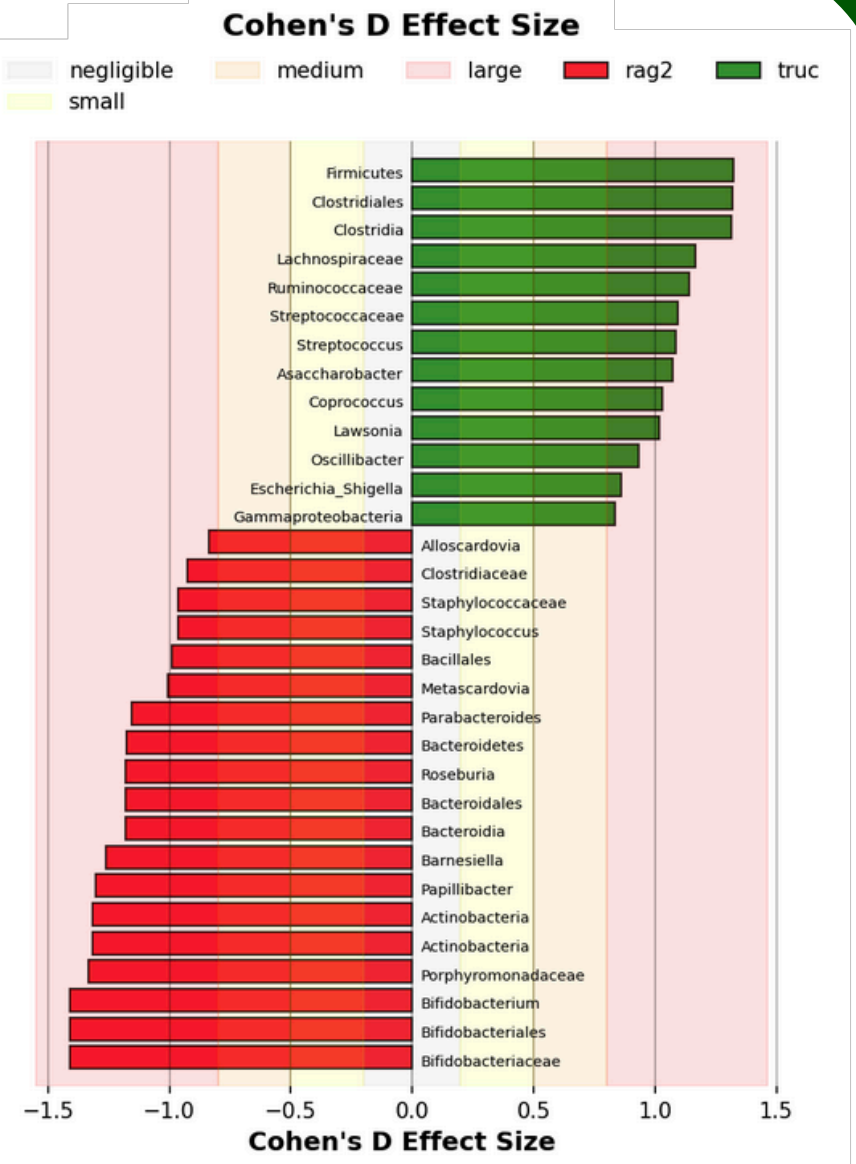
CEfSe was evaluated on the diseased vs. non-diseased mouse dataset (T-bet$^{-/-}$ × Rag2$^{-/-}$ vs. Rag2$^{-/-}$) because it provides a two-class setup for testing the pipeline and avoids the added complexity of multiclass analysis (1).

Cohen's d

- Cohen's d is a standardized effect size that measures how far apart two groups are in units of standard deviation (3).
- Every feature gets a Cohen's d score, and larger |d| means better class separation.
- Cohen's d was chosen because it fills the same role as LDA for ranking features but eliminates the need for any R dependencies.

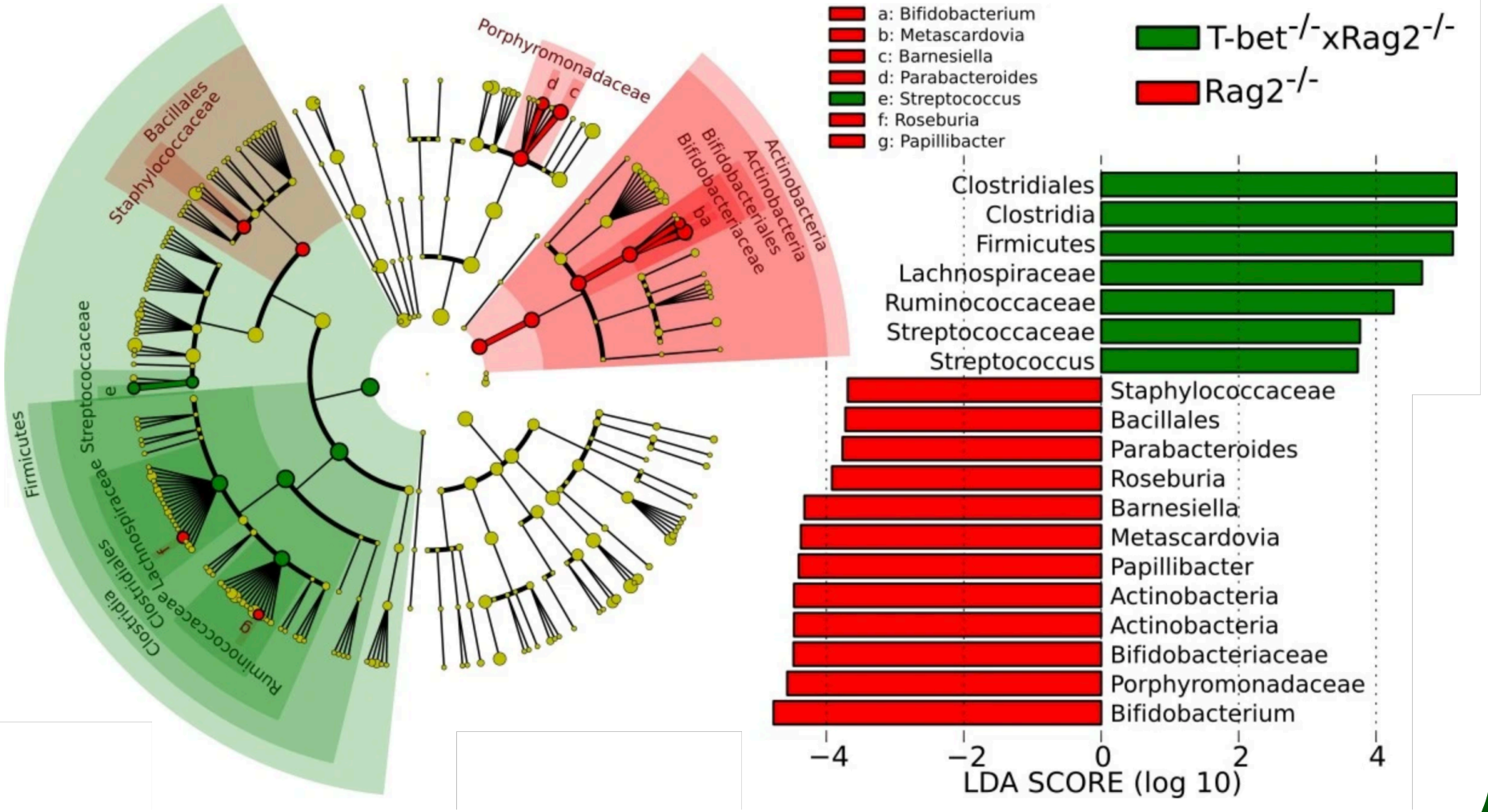
CEfSe


Cladogram
rag2
truc
Bacteroidetes
Actinobacteria
Firmicutes


Cohen's D Effect Size
negligible
small
medium
large
rag2
truc
Firmicutes
Clostridiales
Clostridia
Lachnospiraceae
Ruminococcaceae
Streptococcaceae
Streptococcus
Asaccharobacter
Coprococcus
Lawsonia
Oscillibacter
Escherichia_Shigella
Gammaproteobacteria
Alloscardovia
Clostridiaceae
Staphylococcaceae
Staphylococcus
Bacillales
Metascardovia
Parabacteroides
Bacteroidetes
Roseburia
Bacteroidales
Bacteroidia
Barnesiella
Papillibacter
Actinobacteria
Actinobacteria
Porphyromonadaceae
Bifidobacterium
Bifidobacteriales
Bifidobacteriaceae
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Cohen's D Effect Size

LEfSe

From: **Metagenomic biomarker discovery and explanation**



Results

Cladogram: shows which microbial groups differ between the two classes and how those taxa are related in the phylogenetic tree (1).

Bar Plot: shows the effect size for each taxon, indicating which class it is enriched in and how strongly it contributes to the difference between the two groups (1).

- LEfSe reported 19 differentially abundant clades, and CEfSe recovers the same major taxa.
- LEfSe uses LDA scores after KW + Wilcoxon, while CEfSe uses Cohen's d (standardized mean difference.)
- Because the effect size measures scale features differently, CEfSe includes all LEfSe-detected taxa plus additional ones that do not meet LDA's cutoff.
- CEfSe uses Python 3 and updated matplotlib defaults, which changes the spacing, colors, and styling compared to LEfSe's Python 2 plotting code.

1. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S. and Huttenhower, C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biology*, 12, R60. doi: 10.1186/gb-2011-12-6-r60.
2. Chen, Y.C., Su, Y.Y., Chu, T.Y., Wu, M.F., Huang, C.C. and Lin, C.C. (2025) PreLect: prevalence-leveraged consistent feature selection decodes microbial signatures across cohorts. *NPJ Biofilms and Microbiomes*, 11(1), 3. doi: 10.1038/s41522-024-00598-2.
3. Cumming, G. (2012) 'Cohen's d', in *Understanding the New Statistics*. 1st edition [Online]. United Kingdom: Routledge. pp. 281–320.