Finer Taxonomic Resolution Achieved with a Custom Database
Introducing TaxAss

Robin R. Rohwer[1], Joshua J. Hamilton[1], Ryan J. Newton[2], Katherine D. McMahon[1]

[1]University of Wisconsin-Madison Department of Bacteriology, [2]University of Wisconsin-Milwaukee School of Freshwater Sciences
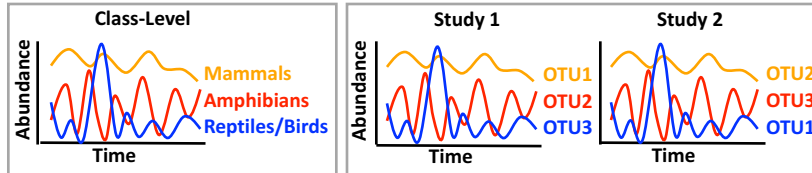
github.com/McMahonLab/TaxAss

TaxAss Enables Cross-Study Comparisons

- TaxAss is a Taxonomy Assignment workflow that uses both a custom, ecosystem-specific database and a comprehensive database such as Greengenes or SILVA.
- TaxAss improves taxonomic resolution while maintaining alpha diversity.

Coarse taxa levels lack ecological relevance

OTU-based analyses lack comparability

Class-Level — Abundance / Time: Mammals, Amphibians, Reptiles/Birds

Study 1 / Study 2 — Abundance / Time: OTU1, OTU2, OTU3 / OTU2, OTU3, OTU1
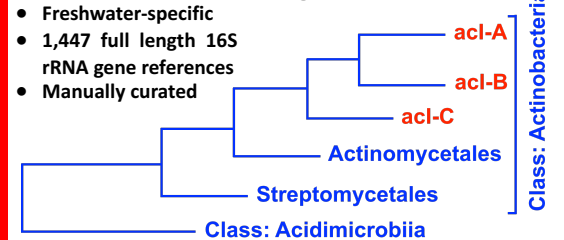
Fine-level taxonomy assignments bridge this gap- they are both comparable between datasets and ecologically relevant.

Hop on the FreshTrain!

The Freshwater Training Set:
- Freshwater-specific
- 1,447 full length 16S rRNA gene references
- Manually curated

Class: Actinobacteria — acI-A, acI-B, acI-C, Actinomycetales, Streptomycetales
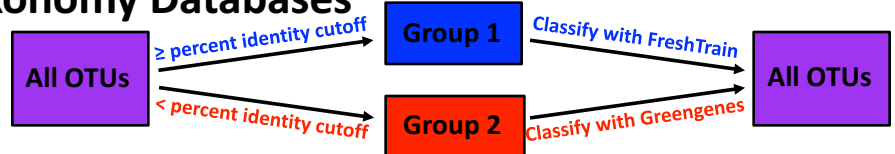Class: Acidimicrobiia

- The FreshTrain's monophyletic branches called Lineage, Clade, and Tribe approximate Family, Genus, and Species.
- Lineage acI dominates most freshwater lake epilimnia.

Newton *et al.* A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. Rev.* 75, 14–49 (2011).

TaxAss Leverages Two Taxonomy Databases

It's not as simple as you think!

- You can't classify with only the custom database
  - → OTUs are "forced" into misclassifications!
  - → Diversity is lost!
- You can't concatenate the databases
  - → Replicate references confuse the classifier!
- You can't combine the databases with curation
  - → That would take forever!
- You CAN use TaxAss
  - → Easy to follow step-by-step instructions!
  - → Generalizable to your favorite databases!
  - → Still use your favorite curation pipeline!
  - → Faster than forever!

github.com/McMahonlab/TaxAss

All OTUs → ≥ percent identity cutoff → Group 1 → Classify with FreshTrain → All OTUs
All OTUs → < percent identity cutoff → Group 2 → Classify with Greengenes → All OTUs

BLAST splits OTUs into 2 groups that are classified separately.

Group 1 has high percent identity to a reference sequence in the FreshTrain.
Group 2 is not similar to anything in the FreshTrain.

BLAST Hit #	BLAST pident	length	qlen	TaxAss pident
1	98.97	97	100	96.00
2	98.00	100	100	98.00
3	98.00	100	100	98.00
4	97.94	97	100	95.00
5	97.94	97	100	95.00

BLAST's alignment percent identity is recalculated into a percent identity for the full length OTU.

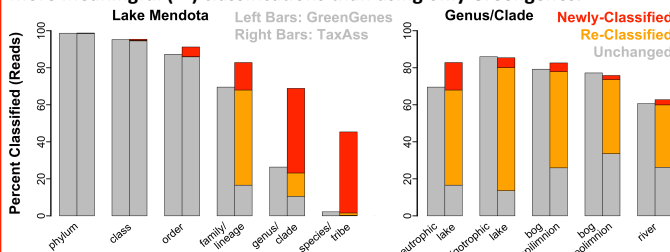Percent Classified by the FreshTrain — Reads (%) / Percent Identity Cutoff
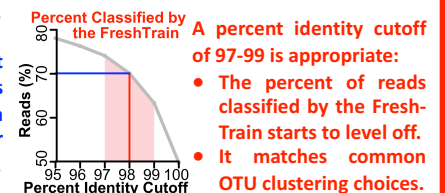
A percent identity cutoff of 97-99 is appropriate:
- The percent of reads classified by the FreshTrain starts to level off.
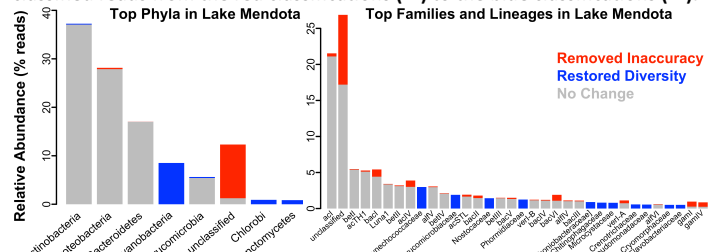- It matches common OTU clustering choices.

TaxAss Improves Classifications

Leveraging both the FreshTrain and Greengenes resulted in more (■) and more meaningful (■) classifications than using only Greengenes.

Lake Mendota — Left Bars: GreenGenes, Right Bars: TaxAss
Percent Classified (Reads): phylum, class, order, family/lineage, genus/clade, species/tribe

Genus/Clade — Newly-Classified, Re-Classified, Unchanged
eutrophic lake, oligotrophic lake, bog epilimnion, bog hypolimnion, river

Comparing the percentage of datasets that get classified shows TaxAss improves fine-level taxonomy assignments. Improvement is highest in ecosystems most similar to the origins of FreshTrain references (lakes).
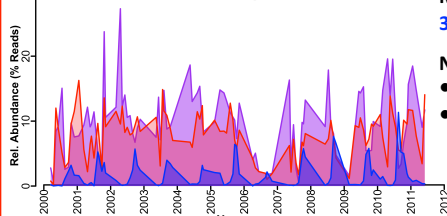
TaxAss Maintains Diversity

Leveraging both the FreshTrain and Greengenes moved FreshTrain-only classified reads from the red classifications (■) to the blue classifications (■).

Top Phyla in Lake Mendota — Relative Abundance (% reads): Actinobacteria, Proteobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia, unclassified, Chlorobi, Planctomycetes

Top Families and Lineages in Lake Mendota

Removed Inaccuracy
Restored Diversity
No Change

Rank abundance curves of Lake Mendota show that TaxAss maintains diversity (■). At fine taxa levels accuracy is improved as "forced" reads are removed from top taxa whose dynamics they would have obscured. (■)

Up Next: Use TaxAss to Compare Datasets!

Clades Under acI Lineage in Lake Mendota — Rel. Abundance (% Reads) / Year (2000–2012): acI-A, acI-B, acI-C

Much is known about the ecology and genetics of abundant FreshTrain taxa (poster 394A). TaxAss is a practical way to leverage this information in community analyses.

Next I will use TaxAss to compare taxa distributions among freshwater ecosystems.
- Distribution patterns over diverse environments can teach us about taxa functions.
- Knowledge about taxa functions can teach us about their preferred environments.

Rel. Abundance (% Reads)
acI-A: Eutrophic, Oligotrophic, Bog Epi, Bog Hypo, River
acI-B: Eutrophic, Oligotrophic, Bog Epi, Bog Hypo, River
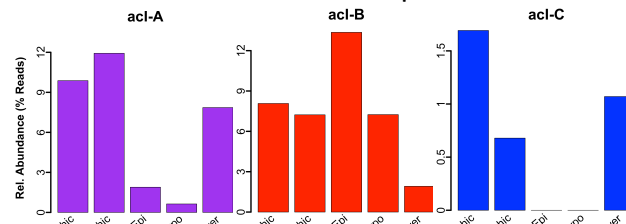acI-C: Eutrophic, Oligotrophic, Bog Epi, Bog Hypo, River

Lineage acI includes three abundant clades, shown in a 10-year Lake Mendota time series. Clade acI-C's seasonal pattern in Aug – Sept matches annual cyanobacteria blooms. Across datasets acI-C is abundant in eutrophic Lake Mendota and the eutrophic Danube River, but not in oligotrophic Lake Michigan or Dystrophic bogs where there are few cyanobacteria.

Thanks:

NSF, LTER, USDA, Applied Microbial Ecology McMahon Lab, The University of Wisconsin Madison