TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution

RUNNING TITLE

TaxAss achieves fine-scale taxonomic resolution

AUTHORS

Robin R. Rohwer[a#], Joshua J. Hamilton[b], Ryan J. Newton[c], Katherine D. McMahon[b,d]

[a]Environmental Chemistry and Technology Program, University of Wisconsin-Madison
[b]Department of Bacteriology, University of Wisconsin-Madison
[c]School of Freshwater Sciences, University of Wisconsin-Milwaukee
[d]Department of Civil and Environmental Engineering, University of Wisconsin-Madison

[#]Robin.Rohwer@gmail.com

WORD COUNTS:

Abstract (250)
Importance (147)
Text (5325)

ABSTRACT

Taxonomy assignment for microbial community composition studies can be limited by a lack of relevant reference organisms in large taxonomy databases. TaxAss is a taxonomy assignment workflow to classify 16S rRNA gene amplicon data using two taxonomy reference databases: a large comprehensive database such as Greengenes or SILVA, and a small ecosystem-specific database curated by scientists within a field. To test TaxAss performance, we classified five different freshwater datasets using the comprehensive Greengenes database and the freshwater-specific FreshTrain database. TaxAss increased the percent of the dataset classified compared to using only Greengenes. The increase in classifications was highest at fine-resolution taxa levels, where across the freshwater test-datasets the classifications at species-level increased by 24-40 percent reads. A similar increase in classifications was not observed in a control mouse gut dataset, which was not expected to contain freshwater bacteria. TaxAss maintained taxonomic richness compared to using only the FreshTrain. Richness was maintained across all taxa-levels from phylum to species. Without TaxAss, the majority of organisms not represented in the FreshTrain were unclassified, but at finer taxa levels incorrect classifications were also significant. TaxAss splits a dataset's sequences into two groups based on their percent identity to reference sequences in the ecosystem-specific database. Highly similar sequences are classified using the ecosystem-specific database and the others are classified using the comprehensive database. TaxAss metrics help users choose a percent identity cutoff

appropriate for their data. TaxAss is free and open source, and available at www.github.com/McMahonLab/TaxAss.

IMPORTANCE

Microbial communities drive ecosystem processes, but microbial community composition analyses using 16S rRNA gene amplicon datasets is limited by the lack of fine-resolution taxonomy classifications. Course taxonomic groupings at phylum, class, and order level lump ecologically distinct organisms together. To avoid this, many researchers cluster similar sequences into operational taxonomic units (OTUs). These fine-resolution groupings are more ecologically relevant, but OTU definitions are dataset-dependent and therefore cannot be compared between datasets. Microbial ecologists studying a variety of environments have curated small, ecosystem-specific taxonomy databases to provide consistent and up-to-date terminology. We created TaxAss, a workflow that leverages these ecosystem-specific databases to assign taxonomy. We found that TaxAss improves fine-resolution taxonomic classifications (family, genus and species). Fine taxonomic groupings are also ecologically relevant, so they provide an alternative to OTU-based analyses that is consistent and comparable between datasets. TaxAss enables researchers to compare data using ecologically relevant terminology.

INTRODUCTION

Microbial communities form the foundations of all ecosystems, yet interpretation of community data is limited by the difficulty of comparing across datasets. With the rapid development of massively parallel sequencing technology, scientists are increasingly able to fingerprint microbial communities using amplicon sequencing of marker genes such as the 16S rRNA gene**.** The resulting sequences are typically clustered into sequence identity-based Operational Taxonomic Units (OTUs). Comparison between amplicon datasets is difficult because sequence identity-based OTUs are specific to each analysis. For clarity, this paper refers to 16S rRNA gene amplicon sequencing datasets as “datasets” and defines OTUs as a dataset’s sequence unit of measure, irrespective of whether those units represent clustered sequences or unique sequences.

Taxonomy Allows Cross-Study Analyses

Sequence identity-based OTUs are widely accepted as good representations of coherent ecological entities (1), however these OTUs represent study-specific phylotypes that cannot be compared between datasets. This is because sequence identity-based OTU clustering is specific to each analysis, resulting in arbitrary OTU names. Additionally, the clustering algorithm and the similarity cutoff chosen for curation of each dataset impacts the ecological relationships observed (2). For these reasons, direct comparisons of sequence identity-based OTUs between multiple datasets are impossible without the computationally expensive task of combining and re-curating.

87

88 Taxonomic naming systems allow comparisons between datasets by creating consistent
89 terminology and consistent phylogeny-determined boundaries between organisms.
90 However, taxonomic naming is most useful when sequences can be classified to a fine
91 level (e.g. family, genus, or species). Many abundant taxa are poorly represented in
92 reference taxonomy databases (hereafter "databases"), resulting in only coarse
93 classifications for large proportions of amplicon datasets (e.g. phylum, class, or order).
94 Coarse taxonomic groupings often include diverse organisms with differing ecological
95 roles, so analyses at coarse taxa levels miss underlying ecological dynamics (3). Fine-
96 resolution taxonomic names are required to bridge the gap between ecologically
97 relevant OTU-based analyses and consistent, comparable taxonomy-based analyses.
98

99 **Ecosystem-Specific Taxonomy Databases**

100 Microbial ecologists from diverse ecosystems have created fine-resolution reference
101 taxonomies by curating databases specific to their ecosystems. These ecosystem-
102 specific databases are small compared to the large comprehensive databases compiled
103 by Greengenes (4), Silva (5), and the Ribosomal Database Project (6), but they are
104 generally well-curated with more finely resolved phylogenies for ecosystem-specific
105 lineages. Examples of ecosystems with curated databases include the human oral
106 cavity (7), the cow rumen (8), the honey bee gut (9), the cockroach and termite gut (10),
107 activated sludge (11), and freshwater lakes (12). Ecosystem-specific databases are
108 created to establish consistent vocabulary for common uncultured bacteria, create

monophyletic reference structures, incorporate new reference information, and understand what the "typical" organisms are in a given ecosystem. Additionally, ecosystem-specific databases can be used to assign taxonomy to a finer resolution than can be achieved with a large comprehensive database.

Taxonomy Assignment Algorithm

Classification algorithms assign taxonomic names to OTUs based on their similarity to reference sequences in a database. The most commonly used classification algorithm was developed by Wang *et. al* (13) for the Ribosomal Database Project and is implemented in both mothur (14) and QIIME (15)**.** This naïve Bayesian classifier (hereafter "Wang classifier") assigns taxonomy to OTUs based on 8-mer signatures and reports a bootstrap confidence estimate for each assignment (13). This bootstrap confidence value is based on the repeatability of the OTU's assignment with subsampled 8-mers, not on an absolute similarity measure. In a large database an OTU dissimilar to any reference sequences will not be classified repeatably as any one taxon, resulting in a low bootstrap confidence. However, in a small database an OTU dissimilar to any reference sequences nevertheless can be classified repeatably because there are fewer references from which to choose. This pitfall is referred to as "forcing" in this paper, because OTUs are forced into incorrect classifications by the small database.

130 **Introducing TaxAss**

131 We aimed to obtain fine-level taxonomy classifications by leveraging an ecosystem-
132 specific database, while at the same time maintaining the full biological diversity of an
133 amplicon dataset. To this end, we developed an open source taxonomy assignment
134 workflow (TaxAss) that uses the popular Wang classifier as implemented in mothur and
135 employs both an ecosystem-specific database and a comprehensive database. The
136 workflow maintains taxonomic richness and accuracy by only classifying OTUs that
137 share high percent identity with ecosystem-specific reference sequences using the
138 ecosystem-specific database. The remaining OTUs are classified using the
139 comprehensive database.

140

141 Ideally, an ecosystem-specific database would be merged with a comprehensive
142 database and TaxAss would not be necessary. However in reality, updates to
143 comprehensive databases usually occur only every few years, during which time both
144 databases diverge further from each other as they continue to be curated. TaxAss
145 allows researchers to effectively leverage the most current version of both databases to
146 achieve the maximum, most accurate, and finest-resolution classifications possible.

147

148 **The FreshTrain**

149 This paper demonstrates TaxAss's efficacy using a variety of freshwater amplicon
150 datasets, the comprehensive Greengenes database (4), and the ecosystem-specific
151 Freshwater Training Set (FreshTrain) (12). The FreshTrain database was created in

152 2012 and was originally curated alongside Greengenes. It matches Greengenes at the
153 phylum, class, and order levels, but at finer taxonomic levels it was curated based on
154 additional information such as the geographical distribution of sequences. These finer
155 levels are referred to as lineage, clade, and tribe and approximate the Linnaean family,
156 genus, and species (12). This workflow and a newly updated version of the FreshTrain
157 are available online at www.github.com/McMahonLab/TaxAss.
158

159 # RESULTS

160 ## Methods Summary

161 TaxAss uses both an ecosystem-specific database and a large comprehensive
162 database to improve taxonomic assignment resolution while maintaining richness. To
163 classify the maximum possible number of OTUs and avoid forcing inaccurate
164 ecosystem-specific classifications onto OTUs, the amplicon dataset is split into two
165 groups prior to classification: OTUs with high percent identity to ecosystem-specific
166 reference sequences and OTUs with low percent identity to ecosystem-specific
167 reference sequences. The two groups are then classified separately using the Wang
168 classifier and the appropriate database (Figure 1).
169

170 ## Fine-Resolution Classifications Increased

171 To test whether TaxAss improved taxonomic classification over solely using a
172 comprehensive database, we assigned taxonomy to a Lake Mendota amplicon dataset
173 first by using Greengenes alone and then by using TaxAss to leverage both

174 Greengenes and the FreshTrain (Figure 2a and Table 1). We compared the percent of
175 reads classified by both methods and observed a marked improvement in the percent of
176 the dataset classified to the fine taxa levels of family/lineage, genus/clade, and
177 species/tribe. At species/tribe level, the percent of reads classified increased from 1.4%
178 to 41%, at genus/clade level they increased from 23% to 63%, and at family/lineage
179 level they increased from 64% to 78%. In addition to these increases in classifications,
180 TaxAss also improved the quality of classifications because the FreshTrain is curated
181 with terminology consistent with the freshwater microbial ecology literature. At
182 family/lineage level, Greengenes alone could classify a majority of the dataset, but 75%
183 of those Greengenes-classified reads received more meaningful ecosystem-specific
184 nomenclature when using TaxAss.
185
186 The FreshTrain reference sequences come exclusively from temperate lake epilimnia,
187 and many of them come from Lake Mendota itself. Lake Mendota is a eutrophic,
188 temperate lake in Wisconsin, USA; and the Lake Mendota amplicon dataset consists of
189 95 epilimnetic samples collected by the North Temperate Lakes Microbial Observatory
190 over 11 years. To test TaxAss's efficacy when the ecosystem-specific database is less
191 representative of the ecosystem under investigation, we classified amplicon datasets
192 from a range of freshwater ecosystems first by using Greengenes alone and then by
193 using TaxAss to leverage Greengenes and the FreshTrain (Figure 2b). The additional
194 ecosystems we chose included the epilimnion of oligotrophic Lake Michigan (16), the
195 eutrophic Danube River (17), and the epilimnion and hypolimnion of dystrophic Trout

196 Bog (WI, USA) (18). We also used a mouse gut dataset (19) as a negative control to
197 ensure that TaxAss would not assign FreshTrain classifications erroneously. All
198 freshwater datasets showed improvements at all fine taxa levels (Figure 2b and
199 Supplemental Figure 2), with the amount of improvement reflecting the similarity of each
200 ecosystem to the FreshTrain reference sequences. For example, the temperate lake
201 Mendota and Lake Michigan epilimnia improved most (54 and 52% at genus/clade
202 level), while the dystrophic bog hypolimnion improved least (28% at genus/clade level).
203 Only 0.1, 0.3, and 0.2% of the mouse gut control dataset received FreshTrain
204 classifications at the species, genus, and family levels.
205

206 **Richness Maintained**

207 To test whether TaxAss improved taxonomic classification over solely using an
208 ecosystem-specific database, we assigned taxonomy to the Lake Mendota dataset first
209 by using the FreshTrain alone and then by using TaxAss to leverage both the
210 FreshTrain and Greengenes. TaxAss maintained taxonomic richness at all taxa levels
211 by classifying OTUs into a larger variety of taxonomic names (Figure 3 and Table 1). At
212 the same time TaxAss prevented inaccuracies in fine-resolution classifications that were
213 introduced by FreshTrain-only classifications (Figure 3b).
214

215 The FreshTrain is a more specific database with less taxonomic richness than
216 Greengenes, so a decrease in taxonomic richness in a FreshTrain-only classification
217 was expected. For example, the FreshTrain focuses on heterotrophic bacteria and does

218 not include any Cyanobacteria, which comprised 8.5% of the Lake Mendota dataset. All
219 of Lake Mendota's cyanobacterial OTUs were classified as something else (99% as
220 unclassified), which resulted in a loss of phylum-level richness in the FreshTrain-only
221 classification (Figure 3a). In contrast, TaxAss maintained the taxonomic richness of a
222 Greengenes-only classification and reduced unclassified reads at all taxa levels (Table
223 1).
224
225 In the FreshTrain-only classification scheme, we also observed occasions when an
226 OTU was incorrectly given a taxonomic assignment when it should have been labeled
227 unclassified (Figure 3b). This "forcing" of OTUs into incorrect classifications by the small
228 FreshTrain database was less common than OTUs being called "unclassified," but had
229 significant effects on taxa relative abundances at finer-resolution taxa levels. Lake
230 Mendota's 5th most abundant lineage, the *Bacteroidetes* bacI, gained 27% more reads
231 in a FreshTrain-only classification compared to using TaxAss. The forcing errors
232 TaxAss prevented were significant enough to change basic attributes such as rank
233 abundances of top taxa, and had an even larger impact on the freshwater test-datasets
234 that differed more from the FreshTrain references (Supplemental Figure 3).
235

236 **Percent Identity Cutoff**

237 An OTU is classified taxonomically in the ecosystem-specific database only if it matches
238 a sequence in that database at a sequence identity above the threshold set by the user.
239 Therefore, the percent identity cutoff choice for taxonomic classification is central to the

240 proper functioning of TaxAss because it determines which OTUs are classified in each
241 database (ecosystem-specific vs comprehensive). If the percent identity cutoff is set too
242 high, ecosystem-specific OTUs are passed to the comprehensive database for
243 classification; while if it is set too low, non-ecosystem-specific OTUs are passed to the
244 ecosystem-specific database for classification.
245
246 We found that a percent identity cutoff of 98-99% was appropriate for the analyzed
247 freshwater datasets, and we applied a cutoff of 98% when processing all data used in
248 this paper. We chose the percent identity cutoff to maximize the percent of each dataset
249 receiving taxonomic classification at every taxa level (Figure 4 and Supplemental Figure
250 4). TaxAss allows users to choose a cutoff specific to their data by generating the plots
251 shown in Figure 4, but users who wish to save computational time can simply choose a
252 percent identity cutoff and only run the classification once.
253

254 **BLAST Conversion Validated**

255 The calculation of percent identity for use in database selection is based on the percent
256 identity returned by The National Center for Biotechnology Information's Basic Local
257 Alignment Search Tool (BLAST) (20). The default megablast settings are appropriate
258 for our application because they have been highly optimized to find short, highly similar
259 alignments. However, BLAST finds areas of local similarity and there is no way to
260 require BLAST to align the entire length of a query OTU's sequence. 16S rRNA gene
261 amplicon sequences are highly similar, and differences in taxonomic classification can

262 be based on even a single mismatch in the amplified region. Therefore, we recalculated
263 the percent identities BLAST returned into “full-length” percent identities for the entire
264 query OTU’s sequence (Supplemental Document 1 and Equation 1).
265
266 We found that recalculating percent identity was necessary to prevent dissimilar OTUs
267 from inclusion in the ecosystem-specific classification. For example, the FreshTrain
268 does not include any reference sequences from the major freshwater phylum
269 *Cyanobacteria*, so no cyanobacterial OTUs have high true percent identities to any
270 references in the FreshTrain. We found that the percent identity recalculation was
271 necessary to prevent some cyanobacterial OTUs from meeting the percent identity
272 cutoff due to the original BLAST percent identities being based on only a short aligned
273 section of the OTU sequence (Supplemental Figure 1).
274
275 We also found that it was necessary to recalculate the percent identity from several
276 BLAST alignments (“hits”) for each OTU because the best BLAST hit did not always
277 have the highest recalculated percent identity. TaxAss examines the top five BLAST
278 hits, recalculates the percent identity of each, and then uses the highest recalculated
279 percent identity to determine if an OTU meets the cutoff. To ensure enough BLAST hits
280 were examined to consistently arrive at the highest possible recalculated percent
281 identity, we calculated the proportion of times each BLAST hit number had the highest
282 recalculated percent identity. In the Lake Mendota amplicon dataset the first BLAST hit
283 almost always also had the best recalculated score, and the contribution of additional

284 BLAST hits was very low, especially when only "good" hits above a stringent percent
285 identity cutoff were considered (Table 2). In the Lake Mendota dataset at the chosen 98
286 percent identity cutoff, 99.69% of the best hits found by BLAST were also the best re-
287 calculated hits and only 0.01% of BLAST's 5th hits were used. TaxAss generates a
288 version of Table 2 for users' individual datasets, and if they observe more high-number
289 BLAST hits contributing to the best re-calculated hit they can increase the number of
290 BLAST results used for the calculation.
291

292 # DISCUSSION

293 ## Ecosystem-Specific Databases

294 The need for curated ecosystem-specific databases has been recognized by microbial
295 ecologists studying many ecosystems. TaxAss was developed specifically to leverage
296 the Freshwater Training Set (FreshTrain) (12), but it could be applied to custom
297 databases curated for other ecosystems: the dictyopteran gut microbiota reference
298 database (DictDb) (10), the rumen and intestinal methanogen database (RIM-DB) (8),
299 the honey bee database (HBDB) (9), the microbial database for activated sludge
300 (MiDAS) (11), and the human oral microbiome database (HOMD) (7). These databases
301 were created by starting with a comprehensive database such as SILVA or Greengenes
302 and then re-curating the reference sequences from the study ecosystem, sometimes
303 also incorporating new reference sequences. Often during curation, phylogenies were
304 collapsed to be monophyletic and incorporate new organisms, and abundant but
305 unnamed organisms were given placeholder names to allow for consistent terminology
306 among researchers.

307

308 DictDB, HBDB, and MiDAS are fully integrated with modified versions of the entire
309 SILVA database, so a workflow like TaxAss that leverages two databases is not needed
310 because the single merged database can be used in one step for taxonomy
311 assignment. However fully integrated databases can be difficult to maintain over time
312 because new versions of each database will diverge from each other, and TaxAss
313 provides a means to circumvent this divergence. The FreshTrain is an example of this
314 divergence in action. The FreshTrain was originally integrated into the Hugenholtz
315 database that eventually became Greengenes, and Greengenes was last updated in
316 May 2013. In addition, SILVA now contains more total references and has been
317 updated as recently as September 2016, so some researchers prefer to use the more
318 recently updated SILVA as their comprehensive database. Similarly, the FreshTrain has
319 been updated almost annually since its creation as new full-length 16S rRNA gene
320 sequences from freshwater ecosystems became available. TaxAss allows microbial
321 ecologists to use the most up-to-date versions of their preferred databases without
322 performing or waiting for reconciliation of each release.

323

324 Once an ecosystem-specific database has diverged from the comprehensive database,
325 as occurred with the FreshTrain, leveraging the ecosystem-specific database for
326 taxonomy assignment is no longer straightforward. Reintegrating the ecosystem-specific
327 database into the comprehensive database is more involved than simply concatenating
328 databases and removing duplicated references because conflicting phylogenetic

329 structures must be resolved. Analysis of community amplicon data is a fairly routine part
330 of many studies for which extensive phylogenetic curation would fall outside the scope.
331 The FreshTrain has been used in a variety of ways since it diverged from the current
332 version of Greengenes, and it is often difficult to discern the specifics from cursory
333 sentences in a paper's methods section. TaxAss provides a well-documented and
334 rigorously tested workflow to leverage two conflicting databases without extensive
335 curation.
336

337 **Current FreshTrain Usage**

338 The simplest way the FreshTrain has been used to assign taxonomy to amplicon
339 datasets is as part of a separate, complementary analysis. For example, in a study of
340 the River Thames Basin (21), FreshTrain and Greengenes classifications were
341 displayed side by side and separate metrics such as diversity indices were calculated
342 for each. However, the bulk of the taxonomic analyses were carried out at the coarse
343 phylum level despite most abundant OTUs having FreshTrain nomenclature. When the
344 FreshTrain is used independently, the loss of richness in taxonomic classifications
345 (Figure 3) makes it difficult to use ecosystem-specific classifications for entire-dataset
346 analyses. TaxAss provides ecosystem-specific classifications without loss of taxonomic
347 richness, thus allowing for a single comprehensive analysis.
348

349 Another straightforward approach has been to classify amplicon datasets sequentially,
350 first using the FreshTrain and then re-classifying the unclassified sequences using a

351 comprehensive database. For example, in a study of Lake Erken, Sweden (22), OTUs
352 were first classified with the FreshTrain, and then unclassified OTUs were reclassified
353 using SILVA. While this approach allows for a single analysis, the initial classification of
354 all sequences with the small FreshTrain database can cause forcing of OTUs into
355 incorrect classifications (Figure 3b). TaxAss prevents forcing by splitting the OTUs into
356 two groups prior to classification.
357
358 This forcing of OTUs into incorrect classifications when using the FreshTrain was
359 observed in a study of cyanobacterial blooms in Yanga Lake, Australia (23), where the
360 authors observed that cyanobacterial OTUs were forced into heterotrophic
361 classifications. To prevent this, Greengenes was used for an initial classification, then
362 only OTUs assigned to phyla included in the FreshTrain were reclassified and renamed
363 with confidently assigned FreshTrain nomenclature. This approach prevented the
364 forcing of Yanga Lake's abundant cyanobacterial OTUs, but it would not prevent forcing
365 of OTUs that belong to phyla included in the FreshTrain (Verrucomicrobia,
366 Bacteroidetes, Proteobacteria, and Actinobacteria). In freshwater datasets such as
367 bogs, rivers, and lake hypolimnia many organisms belonging to FreshTrain phyla differ
368 significantly from the lake epilimnion references included in the FreshTrain. TaxAss
369 prevents forcing classifications onto OTUs of any phyla.
370
371 Another way to avoid the forcing observed with the Wang classifier is to use BLAST-
372 based taxonomy assignment algorithms that determine assignments based on

373 sequence similarity. Since the BLAST algorithm calculates an absolute similarity instead
374 of a relative one, a similarity cutoff prevents classifications to dissimilar sequences. The
375 BLAST method to assign taxonomy has been used with the FreshTrain to classify
376 sequences from boreal lakes in Quebec, Canada (24). However, unlike the Wang
377 classifier, BLAST only takes into account individual reference sequences and ignores
378 their encompassing phylogenetic structure. The BLAST-based algorithm from
379 Classification Resources for Environmental Sequence Tags (CREST) (25) addresses
380 this by taking a lowest common ancestor approach. Each query OTU is classified to the
381 finest taxa level that its top BLAST hits share. The CREST algorithm also has been
382 used to assign taxonomy using the FreshTrain to sequences obtained from the Danube
383 River in southeastern Europe (17). This approach avoided forcing and incorporated
384 phylogenetic information in the taxonomy assignments, however it does not maintain
385 diversity by also leveraging a comprehensive database. Additionally, the Wang classifier
386 is more robust at coarser taxa levels and for shorter sequences (25), and it is
387 implemented in common tools like mothur and QIIME. TaxAss allows users to leverage
388 both ecosystem-specific and comprehensive databases using the highly trusted and
389 conveniently implemented Wang classifier.
390

391 **Future TaxAss Usage**

392 We recommend all microbial ecologists studying freshwater systems use the FreshTrain
393 and TaxAss to classify their 16S rRNA gene amplicon datasets. This will result in a
394 consistent, specific, and comparable vocabulary throughout the field, and will improve

395 classification for analysis of individual datasets. We also recommend that microbial
396 ecologists with different ecosystem-specific databases consider TaxAss when their
397 databases diverge from the most up-to-date comprehensive database and phylogenetic
398 curation is outside the scope of their project.
399
400 We recommend microbial ecologists create ecosystem-specific databases if one does
401 not already exist, since they provide improved analysis and enhanced collaboration for
402 the entire field. Phylogenies must be created from full-length 16S rRNA gene
403 sequences, which are not collected as routinely as short amplicon sequences. However,
404 we believe the benefit of these databases as a reference for the field and to improve
405 taxonomic classification of amplicon sequences justifies the effort to create them,
406 especially since TaxAss allows their use without constant re-curation. Additional full-
407 length sequences to flesh-out the existing phylogenetic structure of organisms can be
408 created with clone libraries, as was done for the FreshTrain. New sequencing
409 technologies, such as the long reads produced by Nanopore (26) and PacBio (27, 28)
410 instruments, promise even easier reference sequence generation in the future.
411

412 **Practical Guidance for Using TaxAss**

413 TaxAss includes detailed descriptions of its constituent scripts including argument
414 options and descriptions, so users are able to customize their analyses. The most
415 important decision users make is the cutoff percent identity that determines which
416 database is used to classify each OTU. If an OTU is above the cutoff (i.e. has high

417 percent identity to an ecosystem-specific reference sequence) then it will be classified
418 with the ecosystem-specific database. When the cutoff is higher, fewer OTUs are
419 classified with the ecosystem-specific database and users run the risk of leaving some
420 ecosystem-specific OTUs poorly classified by the comprehensive database. If the cutoff
421 is lower, more OTUs will be classified with the ecosystem-specific database and users
422 run the risk of forcing incorrect classifications onto OTUs and losing taxonomic richness.
423 Users can decide on a percent identity cutoff at the beginning and run only one
424 classification, or they can run TaxAss with several cutoffs and generate versions of
425 Figure 4 to help guide their choice.
426
427 We found that a percent identity cutoff of 98-99% optimized classifications in our test-
428 datasets. The finding that most OTUs match their ecosystem-specific reference
429 sequences with such high percent identity suggests that the commonly chosen 97%
430 sequence identity clustering is too coarse to observe fine taxa level dynamics. This is
431 supported by previous findings that sequence identity-based OTUs can impose artificial
432 delineations between organisms that affect results differently depending on the lineage
433 (29), and that sequence identity-based OTUs can contain temporally discordant
434 sequences (30). We recommend that users planning a taxonomy-centric analysis
435 classify unique sequences without sequence identity-based clustering and use fine-level
436 taxonomic assignment to group their data. This will make the classification take longer,
437 but likely users will save computational time overall by not clustering. For users who
438 also want to emphasize OTU-based analyses and whose datasets are too large for

439 unclustered analyses, we recommend choosing a finer sequence identity-based OTU
440 definition such as 98 or 99% to best leverage the fine-level classification provided by
441 TaxAss and a detailed ecosystem-specific database. When OTUs have been clustered
442 based on sequence identity, we recommend that users choose the same or lower
443 percent identity cutoff in TaxAss to prevent OTUs with constituent sequences falling on
444 either side of the percent identity cutoff.
445

446 **TaxAss Improves Analyses**

447 Taxonomy-based analyses allow researchers to compare results across datasets.
448 Leveraging an ecosystem-specific database for taxonomy assignment results in a high
449 proportion of fine resolution classifications, and grouping sequences based on these
450 classifications is a dataset-independent way to describe community composition. The
451 resulting taxonomic terminology is consistent and comparable between analyses, and
452 the finely resolved taxonomic groupings rival the resolution of sequence identity-based
453 OTUs. TaxAss can replace or complement sequence identity-based OTU clustering,
454 which results in dataset-dependent OTU clusters and names. Reclustering sequence
455 identity-based OTUs to compare across datasets is computationally expensive, and is
456 not possible for datasets created with differing amplification primers. When researchers
457 use TaxAss to assign fine-level taxonomy to their datasets, colleagues can compare
458 their results directly, without re-analysis and regardless of primer set. Additionally,
459 taxonomic nomenclature can also bridge amplicon-based analyses and genomic
460 analyses.

461

462 Leveraging ecosystem-specific databases for taxonomy assignment also improves
463 researchers' interpretations of individual datasets. Ecosystem-specific terminology is
464 more meaningful because ecosystem-specific databases incorporate additional
465 reference sequences, finer phylogenetic delineations, consistent nomenclature for
466 uncultured organisms, and monophyletic structures. For example, the dominant lineage
467 in freshwater is the FreshTrain's actinobacterial lineage acI, which in Greengenes is
468 usually classified as Ack-M1. Although a classification exists for this organism in both
469 databases, the Greengenes lineage includes all sequences that match a single
470 historically used primer, while the FreshTrain classification includes a slightly different
471 subset of actinobacterial sequences and is based on a phylogeny manually curated
472 from full-length 16S rRNA genes. Additionally, the FreshTrain includes finer-level
473 phylogenetic information on acI's constituent subgroups known as clades and tribes,
474 which prior work suggests are ecologically and metabolically differentiated (12). The
475 fine-resolution taxonomy assignments provided by TaxAss and an ecosystem-specific
476 database allow researchers to link their amplicon datasets with known ecophysiological
477 traits.

478

479 Ecosystem-specific phylogenies that are not fully incorporated into a comprehensive
480 database are not straightforward to leverage for taxonomy assignment. The FreshTrain,
481 for example, has diverged from Greengenes since it was created, and it has been used
482 for taxonomy assignment with inconsistent and sometimes unreliable methods. TaxAss

483 is a well-documented, open source, and rigorously tested workflow that avoids the
484 pitfalls of using a small database: forcing incorrect classifications onto sequences and
485 losing taxonomic richness by leaving unrepresented organisms unclassified. At the
486 same time, TaxAss achieves the benefits of an ecosystem-specific database: more
487 meaningful nomenclature, larger proportions of the dataset classified, and finer-
488 resolution classifications.
489

490 # METHODS

491 ## How to Use TaxAss

492 TaxAss replaces only the taxonomy assignment step of users' preferred amplicon
493 dataset processing pipeline such as mothur or qiime. TaxAss consists of a series of
494 scripts using R, Python, bash, mothur, and BLAST that are run from the terminal
495 command line. The input to TaxAss is a quality controlled fasta file, and if users opt to
496 run the optional percent identity cutoff metrics a relative abundance table is also
497 required. The output of TaxAss is the fasta file's sequence IDs followed by their 7-level
498 taxonomy assignments. Scripts, step-by-step instructions, and detailed explanations of
499 script argument options are available online at https://github.com/McMahonLab/TaxAss.
500

501 ## Percent Identity Recalculation

502 The naïve Bayesian algorithm used for taxonomy assignment (the Wang classifier) (13)
503 can "force" incorrect classifications onto OTUs when a close match does not exist in a
504 small reference database. TaxAss uses the well-accepted Wang classifier, but avoids

505 forcing when classifying with a small ecosystem-specific database by only classifying
506 sequences for which a close reference exists. The National Center for Biotechnology
507 Information's Basic Local Alignment Search Tool (BLAST) (20) is utilized to split the
508 amplicon dataset into two groups prior to classification: one is classified with the
509 ecosystem-specific database, the other with the comprehensive database.
510
511 Blastn queries each OTU sequence against the ecosystem-specific database using the
512 default megablast settings, which are optimized to find highly similar matches between
513 sequences longer than 30 bp (31). However, blast returns the percent identity of the
514 highest scoring pair (the "pident"), which does not necessarily include the full length of
515 the query OTU sequence. OTU sequences are highly similar; a single mismatch can
516 change a classification, so mismatches at the ends of OTU sequences (in the
517 "overhang") that BLAST leaves out of the alignment must be included in the percent
518 identity cutoff used for classification. Therefore, the BLAST pident is recalculated to a
519 full length percent identity with the following equation:
520
521 $$\text{percent identity} = \frac{\text{pident} \times \text{length}}{\text{qlen}+(\text{length}-(\text{qend}-\text{qstart}+1)} \quad \text{(Equation 1)}$$
522
523 where "pident" is the percent identity returned by BLAST, "length" is the length of the
524 alignment, "qlen" is the query length, "qend" is the query end, and qstart is the query
525 start. All of these parameters are returned by BLAST output format 6, and detailed

526 descriptions of what they are, the equation derivation, and an example alignment and
527 calculation are included in Supplemental Document 1.
528
529 The recalculation to full length percent identity is conservative; all query nucleotides not
530 included in the alignment (nucleotides in the “overhang”) are considered mismatches.
531 This means that it would be possible to exclude an OTU from the ecosystem-specific
532 classification when its true percent identity is above the cutoff due to matches in
533 unaligned overhangs. An example of this situation is illustrated in Supplemental
534 Document 1. When the highest scoring BLAST alignments contain matches on the
535 overhangs, some of the lower-scoring alignments will be longer, and therefore have a
536 higher recalculated percent identity. To correct for this TaxAss recalculates the percent
537 identity of the top five blast hits and uses the best one for the cutoff decision. TaxAss
538 also shows users the distribution of chosen hits, so that settings can be re-evaluated if
539 BLAST is not primarily returning hits that have the best recalculated percent identities.
540

541 **Cutoff Choice**

542 OTUs with percent identities greater than or equal to the users’ specified cutoff are
543 classified with the ecosystem-specific database using the Wang classifier as
544 implemented by mothur. The remaining OTUs are classified with the comprehensive
545 database, also using the Wang classifier. The choice of a percent identity cutoff is left to
546 users so that they can balance their choices based on the structures of their datasets
547 and their plans for analysis. If the percent identity cutoff is too low, dissimilar OTUs will

548 be classified in the ecosytem-specific database and may be left unclassified or forced
549 into incorrect classifications, but if the percent identity cutoff is too high, OTUs similar to
550 the ecosystem-specific database will be classified by the comprehensive database and
551 may end up poorly classified.
552
553 Users have the option to choose a cutoff percent identity at the start, or they can
554 classify their datasets with multiple cutoffs and TaxAss will provide metrics to guide their
555 decisions. These metrics include versions of Figure 4, which shows the cutoff choices
556 that maximize the proportion of dataset classified at different taxa levels.
557
558 As an additional optional check, users can also classify their datasets with only the
559 comprehensive database and then compare the classifications. TaxAss provides
560 metrics to check for forcing incorrect coarse-resolution classifications onto OTUs
561 (Supplemental Table 1). Phylum- and class-level classifications are more reliable when
562 assigned by a large comprehensive database that includes more diversity, so if
563 ecosystem-specific classifications at these coarse taxa levels disagree with the
564 comprehensive database's assignments it suggests that the percent identity cutoff is too
565 low. Only these coarse levels can be used to check for forcing, because at finer
566 taxonomic levels too many OTUs end up unclassified by the comprehensive database
567 to compare assignments.
568

Data Availability

The freshwater datasets used in this paper are all publicly available on the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA). The accession numbers are: Lake Mendota (ERP016591), Trout Bog (ERP016854), Lake Michigan (SRP056973), and Danube River (SRP045083). The Lake Michigan and bog project accessions include additional sample types, so only the Lake Michigan and Trout Bog samples were used. The mouse gut dataset is the full version of the example data used by mothur's miSeq SOP, and is available on the mothur website (https://www.mothur.org/wiki/MiSeq_SOP).

Quality control of fastq files was performed according to mothur's MiSeq SOP (19, accessed September 2017) through the chimera checking step. The resulting unique sequences were used as OTUs in all further analyses. The single-end sequencing datasets (Lake Mendota and Trout Bog) were also pre-processed with VSEARCH (32) to trim to uniform lengths and remove low quality sequences with > 0.5 expected errors. During TaxAss, the percent identity cutoff used for all datasets was 98%, and the Wang classifier's bootstrap confidence was set at 80% for all classifications. Batch files that reproduce all download, quality control, and TaxAss processing for each dataset are available in Supplemental Document 2.

The taxonomy databases used in this paper are also publicly available. The Freshwater Training Set (FreshTrain) (12) is available from www.github.com/McMahonLab/TaxAss.

591 The Greengenes database (4) is available from

592 greengenes.secondgenome.com/downloads. Further details on download, versions,

593 and formatting can be found in Supplemental Document 2 and in the detailed directions

594 provided at www.github.com/McMahonLab/TaxAss.

595 ACKNOWLEDGEMENTS

596

597 We thank the North Temperate Lakes Long Term Ecological Research Program (NTL-

598 LTER) and the Microbial Observatory for their years of support collecting the Lake

599 Mendota and Trout Bog time series. We also thank the Earth Microbiome Project for

600 generously sequencing our Lake Mendota and Trout Bog datasets. We thank Benjamin

601 Crary and Jason Woodhouse for sharing insights from their efforts to combine taxonomy

602 databases for classification, and the many users of TaxAss who have reported bugs

603 during development.

604 REFERENCES

605

606 1. Schmidt TSB, Matias Rodrigues JF, von Mering C. 2014. Ecological Consistency of

607 SSU rRNA-Based Operational Taxonomic Units at a Global Scale. PLoS Comput

608 Biol 10:e1003594.

609 2. Schmidt TSB, Matias Rodrigues JF, von Mering C. 2015. Limits to robustness and

610 reproducibility in the demarcation of operational taxonomic units: Robustness and

611 reproducibility in OTU demarcation. Environ Microbiol 17:1689–1706.

612 3. Ruiz-González C, Salazar G, Logares R, Proia L, Gasol JM, Sabater S. 2015.
613 Weak Coherence in Abundance Patterns Between Bacterial Classes and Their
614 Constituent OTUs Along a Regulated River. Front Microbiol 6.

615 4. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T,
616 Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a Chimera-Checked 16S rRNA
617 Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol
618 72:5069–5072.

619 5. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner
620 FO. 2013. The SILVA ribosomal RNA gene database project: improved data
621 processing and web-based tools. Nucleic Acids Res 41:D590–D596.

622 6. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro
623 A, Kuske CR, Tiedje JM. 2014. Ribosomal Database Project: data and tools for
624 high throughput rRNA analysis. Nucleic Acids Res 42:D633–D642.

625 7. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. 2010. The
626 Human Oral Microbiome Database: a web accessible resource for investigating
627 oral microbe taxonomic and genomic information. Database 2010.

628 8. Seedorf H, Kittelmann S, Henderson G, Janssen PH. 2014. RIM-DB: a taxonomic
629 framework for community structure analysis of methanogenic archaea from the
630 rumen and other intestinal environments. PeerJ 2:e494.

631 9. Newton IL, Roeselers G. 2012. The effect of training set on the classification of
632 honey bee gut microbiota using the Naïve Bayesian Classifier. BMC Microbiol
633 12:221.

634 10. Mikaelyan A, Köhler T, Lampert N, Rohland J, Boga H, Meuser K, Brune A. 2015.
635 Classifying the bacterial gut microbiota of termites and cockroaches: A curated
636 phylogenetic reference database (DictDb). Syst Appl Microbiol 38:472–482.

637 11. McIlroy SJ, Saunders AM, Albertsen M, Nierychlo M, McIlroy B, Hansen AA, Karst
638 SM, Nielsen JL, Nielsen PH. 2015. MiDAS: the field guide to the microbes of
639 activated sludge. Database 2015:bav062.

640 12. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. 2011. A Guide to the
641 Natural History of Freshwater Lake Bacteria. Microbiol Mol Biol Rev 75:14–49.

642 13. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian Classifier for
643 Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl
644 Environ Microbiol 73:5261–5267.

645 14. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski
646 RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van
647 Horn DJ, Weber CF. 2009. Introducing mothur: Open-Source, Platform-
648 Independent, Community-Supported Software for Describing and Comparing
649 Microbial Communities. Appl Environ Microbiol 75:7537–7541.

650 15. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK,
651 Fierer N, Pena AG, Goodrich JK, Gordon JI, others. 2010. QIIME allows analysis of
652 high-throughput community sequencing data. Nat Methods 7:335–336.

653 16. Newton RJ, McLellan SL. 2015. A unique assemblage of cosmopolitan freshwater
654 bacteria and higher community diversity differentiate an urbanized estuary from
655 oligotrophic Lake Michigan. Front Microbiol 6.

656 17. Savio D, Sinclair L, Ijaz UZ, Parajka J, Reischer GH, Stadler P, Blaschke AP,
657 Blöschl G, Mach RL, Kirschner AKT, Farnleitner AH, Eiler A. 2015. Bacterial
658 diversity along a 2600 km river continuum: River bacterioplankton diversity. Environ
659 Microbiol 17:4994–5007.

660 18. Linz AM, Crary BC, Shade A, Owens S, Gilbert JA, Knight R, McMahon KD. 2017.
661 Bacterial Community Composition and Dynamics Spanning Five Years in
662 Freshwater Bog Lakes. mSphere 2.

663 19. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013.
664 Development of a Dual-Index Sequencing Strategy and Curation Pipeline for
665 Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform.
666 Appl Environ Microbiol 79:5112–5120.

667 20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment
668 search tool. J Mol Biol 215:403–410.

669 21. Read DS, Gweon HS, Bowes MJ, Newbold LK, Field D, Bailey MJ, Griffiths RI.
670 2015. Catchment-scale biogeography of riverine bacterioplankton. ISME J 9:516–
671 526.

672 22. Eiler A, Heinrich F, Bertilsson S. 2012. Coherent dynamics and association
673 networks among lake bacterioplankton taxa. ISME J 6:330–342.

674 23. Woodhouse JN, Kinsela AS, Collins RN, Bowling LC, Honeyman GL, Holliday JK,
675 Neilan BA. 2016. Microbial communities reflect temporal changes in cyanobacterial
676 composition in a shallow ephemeral freshwater lake. ISME J 10:1337–1351.

677 24. Cottrell M, del Giorgio P, Kirchman D. 2015. Biogeography of globally distributed
678 bacteria in temperate and boreal Québec lakes as revealed by tag pyrosequencing
679 of 16S rRNA genes. Aquat Microb Ecol 76:175–188.

680 25. Lanzén A, Jørgensen SL, Huson DH, Gorfer M, Grindhaug SH, Jonassen I, Øvreås
681 L, Urich T. 2012. CREST – Classification Resources for Environmental Sequence
682 Tags. PLoS ONE 7:e49334.

683 26. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M.
684 2016. Thousands of primer-free, high-quality, full-length SSU rRNA sequences
685 from all domains of life. bioRxiv 070771.

686 27. Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. 2016.
687 Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene
688 classification. BMC Microbiol 16:274.

689 28. Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK. 2016.
690 Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing
691 system. PeerJ 4:e1869.

692 29. Youngblut ND, Shade A, Read JS, McMahon KD, Whitaker RJ. 2013. Lineage-
693 Specific Responses of Microbial Communities to Environmental Change. Appl
694 Environ Microbiol 79:39–47.

695 30. Hall MW, Rohwer RR, Perrie J, McMahon KD, Beiko RG. 2017. Ananke: temporal
696 clustering reveals ecological dynamics of microbial communities. PeerJ 5:e3812.

697 31. Camacho C, Madded T, Tao T, Agarwala R, Morgulis A. 2008. BLAST® Command
698 Line Applications User Manual [Internet]. Bethesda (MD): National Center for
699 Biotechnology Information (US).

700 32. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile
701 open source tool for metagenomics. PeerJ 4:e2584.

702
703
704
705
706
707

TABLES

Table 1. Classification of the Lake Mendota dataset

	Percent Classified[a]			Taxonomic Richness[b]		
	Greengenes	TaxAss	FreshTrain	Greengenes	TaxAss	FreshTrain
Phylum	98	98	87	68	68	5
Class	93	93	79	209	207	17
Order	83	87	74	423	419	35
Family/Lineage	64	78	68	699	733	65
Genus/Clade	23	63	56	1161	1206	102
Species/Tribe	1	41	42	1322	1405	154

[a]Percent classified = total reads classified / total reads in dataset *x* 100%

[b]Taxonomic Richness = total unique classifications

Table 2. Agreement between BLAST and recalculated percent identities.

BLAST result	Percent Identity Cutoff Applied[a]					
	95	**96**	**97**	**98**	**99**	**100**
Hit 1	97.8[b]	98.3	99.0	99.69	99.905	100
Hit 2	1.3	1.0	0.6	0.22	0.066	0
Hit 3	0.4	0.3	0.2	0.05	0.014	0
Hit 4	0.2	0.2	0.1	0.03	0.009	0
Hit 5	0.3	0.3	0.1	0.01	0.006	0

[a]Calculations performed only on sequences above listed recalculated percent identities

[b]*e.g.* 97.8% of BLAST's first hits also had the highest recalculated percent identity

FIGURES

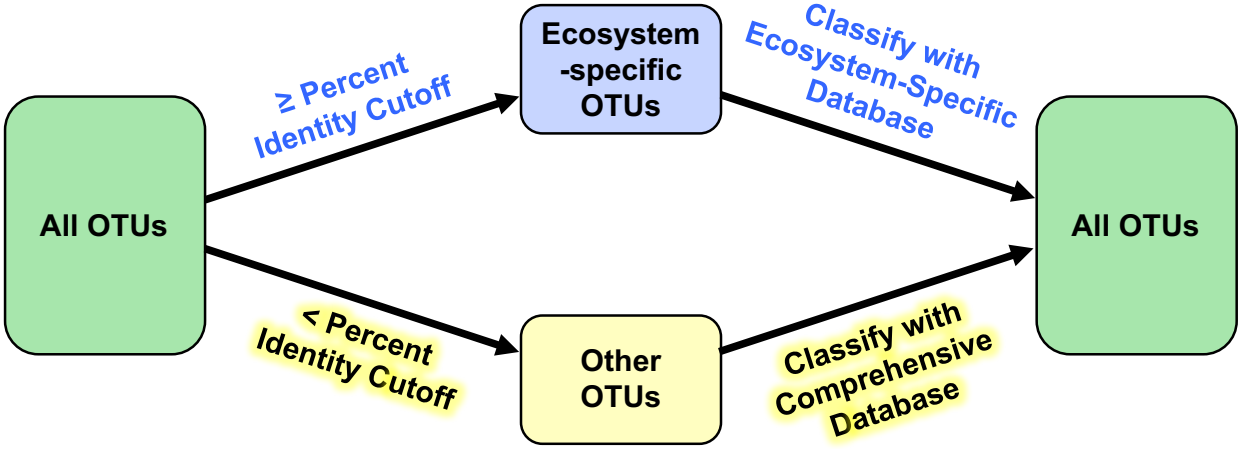

Figure 1. TaxAss Conceptual Diagram

TaxAss separates OTUs into two groups that are classified separately and then recombined. OTUs similar to any ecosystem-specific reference sequences are classified using the ecosystem-specific database, otherwise they are classified by the comprehensive database.

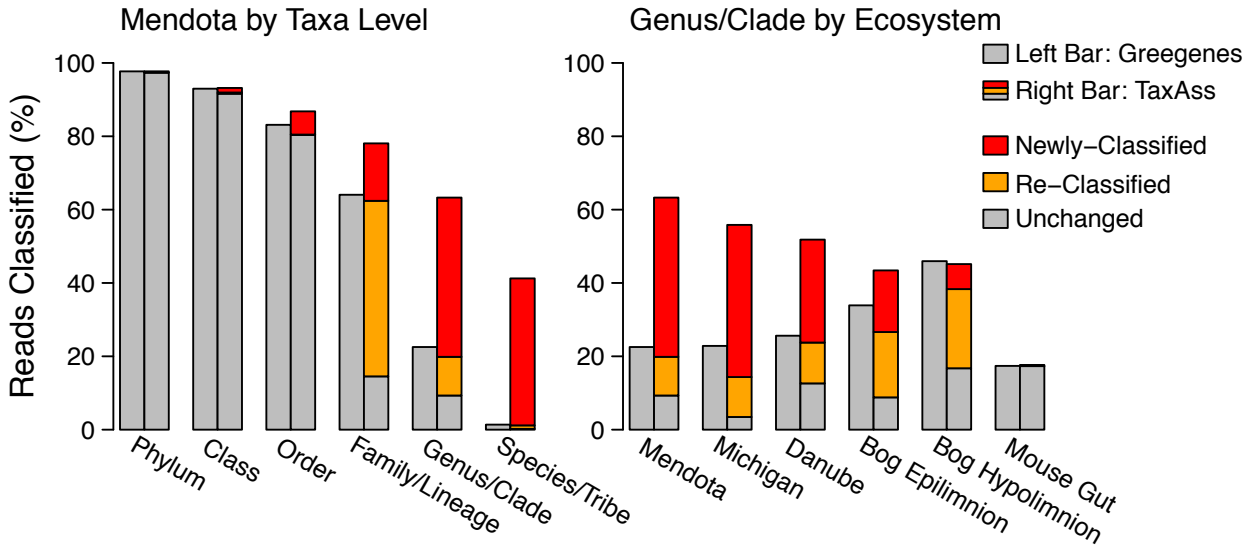

Figure 2. TaxAss performance compared to Greengenes-only performance. **A and B:** Left bars represent the Greengenes-only classification and right bars represent the TaxAss classification that leveraged both Greengenes and the FreshTrain. Within the right bars, red reads were classified by the FreshTrain using TaxAss and were left unclassified using only Greengenes; yellow reads were classified by the FreshTrain using TaxAss but received Greengenes classifications using only Greengenes, and grey reads were classified by Greengenes when using TaxAss. **A:** In the Lake Mendota dataset, TaxAss leveraged the FreshTrain and Greengenes to achieve improved fine-resolution classifications. **B:** TaxAss achieved improvements in a range of freshwater datasets despite the FreshTrain's primary focus on temperate lake epilimnia. Few changes in classification were observed in the mouse gut control. Versions of this figure across all datasets and taxa levels can be found in Supplemental Figure 2.

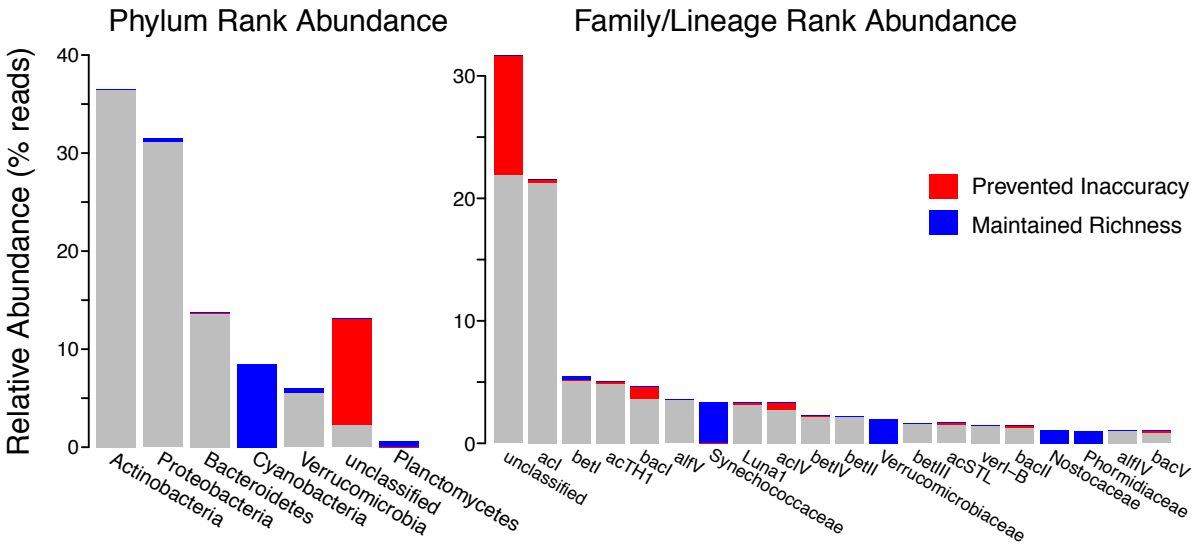

Figure 3. TaxAss performance compared to FreshTrain-only performance. **A and B:** Lake Mendota **r**eads represented by blue bars were incorrectly classified as red bars in the FreshTrain-only classification. Rank order of the bars follows the TaxAss-classification rank abundances. Only taxa with at least 0.5% relative abundance are included, and at lineage level the number of bars displayed is further truncated to 20. **A:** TaxAss maintained phylum richness by classifying the phyla (blue) that are not included in the FreshTrain using Greengenes. **B**: TaxAss similarly maintained lineage richness, and also prevented inaccuracies from forced taxonomic classifications (red bars). Versions of this figure across all test-ecosystems can be found in Supplemental Figure 3.

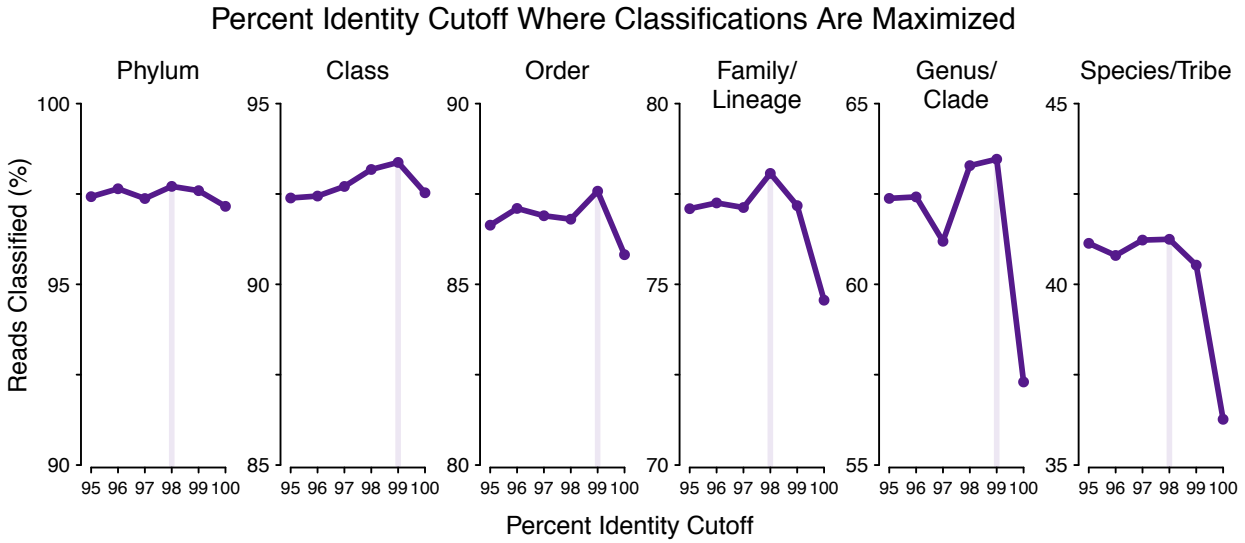

Figure 4. Percent identity cutoff choice.

The percent of reads classified in the Lake Mendota dataset at each taxa level. Faint vertical lines highlight the maxima. TaxAss users can use this visualization to identify a percent identity cutoff appropriate for their dataset. Versions of this figure across all test-ecosystems can be found in Supplemental Figure 4.

783 # LIST OF SUPPLEMENTAL DOCUMENTS

784

785 **Supplemental Table 1. Coarse-level classification disagreements in Lake**

786 **Mendota dataset at 98 percent identity cutoff.**

787

788 **Supplemental Figure 1. Importance of percent identity recalculation.**

789 The phylum Cyanobacteria exemplifies why the percent identity recalculation is

790 necessary. The ecosystem-specific FreshTrain database contains no cyanobacterial

791 references, so cyanobacterial reads serve as a control for something that should be

792 classified in the comprehensive Greengenes group. However, BLAST returned hits with

793 high percent identities for cyanobacterial OTUs due to including partial alignments (red

794 bars). After the TaxAss percent identity recalculation, the cyanobacterial OTUs had

795 lower percent identities and none were included in the FreshTrain classification group

796 (grey bars).

797

798 **Supplemental Figure 2. TaxAss compared to Greengenes-only**

799 **performance.**

800 Versions of Figure 2 for each test-dataset and taxa level.

801

802 **Supplemental Figure 3. TaxAss compared to FreshTrain-only performance.**

803 Versions of Figure 3 for each test-dataset.

804

805 **Supplemental Figure 4. Percent Identity Cutoff Where Classifications Are**
806 **Maximized.**
807 Versions of Figure 4 for each test-dataset.
808

809 **Supplemental Document 1. BLAST percent identity recalculation.**
810 This 3-page document defines the blast terminology, derives the equation to recalculate
811 percent identity, and provides an example alignment and calculation.
812

813 **Supplemental Document 2. Data Processing Batch Files.**
814 This zip file contains directions for reproducing all data processing in this paper. It
815 includes batch scripts that download each dataset, batch scripts that quality control
816 each dataset, and batch scripts that perform TaxAss on each dataset. It also contains
817 the versions of TaxAss scripts used in this paper. A README file includes directions for
818 how to download the versions of databases and programs used in processing, and how
819 to source the included batch scripts. The zip file does not include the data itself, just the
820 folder structure and scripts for download and processing; the uncompressed folder is 1
821 MB.