Species-level microbial sequence classification is improved by source-environment information

Benjamin D. Kaehler[1*], Nicholas A. Bokulich[2*], J. Gregory Caporaso[2,3], Gavin A. Huttley[1]

[1]Research School of Biology, Australian National University, Canberra, Australia
[2]The Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA
[3]Department of Biological Sciences, Northern Arizona University, Flagstaff, AZ, USA

* These authors contributed equally

Abstract

Taxonomic classification of marker-gene DNA sequences is a key step in the analysis of microbial ecology data. Accurate species-level characterisation of microbial communities by sequencing standard marker-gene sequence fragments has proved elusive. We show that knowing a sample's Earth Microbiome Project Ontology (EMPO) habitat type always increases taxonomic classification accuracy, usually to the point where classification accuracy at species level exceeds the genus-level classification accuracy achieved with EMPO-ignorant methods. This improvement comes from setting EMPO habitat type-specific taxonomic class weights for a naive Bayes taxonomic classifier to the average microbial composition for that habitat. We provide q2-clawback, a QIIME 2 plugin for compiling taxonomic class weights from a given collection of samples or by directly downloading data from the Qiita microbial study management platform. The weights output by q2-clawback are compatible with the q2-feature-classifier taxonomic classifier. We show that taxonomic weights averaged across the EMPO 3 habitat types (representing the observed global frequencies of microbial taxa in the Earth Microbiome Project database) typically increase sequence classification accuracy, providing an avenue for improving classification of samples of unknown or uncharacterized sample types for which custom class weights cannot be assembled. q2-clawback also provides a utility for estimating how effectively a new set of taxonomic class weights will improve taxonomic classification accuracy. Apart from the important accuracy improvements in taxonomic classification, our results represent a meta-analysis of marker-gene amplicon sequence data across 7,204 studies to reveal useful patterns of taxonomic abundance.

Author Summary

A key step in many microbial ecology experiments is to determine the taxonomic composition of a sample. This is commonly achieved by classifying short DNA sequences according to their species of origin. In this study we assemble DNA sequence data from 7,204 studies to show

that information about a sample's source leads to more accurate classification of each constituent DNA sequence. The improvement in accuracy is sufficient to accurately identify species from DNA sequences as short as 150 nt, whereas previously only a courser level of classification could be achieved reliably using short amplicon sequences. We make these improvements accessible to all researchers by encapsulating them in q2-clawback, a software plugin to the popular QIIME 2 microbiome bioinformatics framework. The implications of this work are immediately more accurate inference wherever microbiome analysis is used, from clinical settings, to ecological research, to forensic science.

Introduction

It is difficult to overstate the importance of marker-gene amplicon sequencing-based estimation of taxonomic abundance in microbiome research. The price and accessibility of these tools make them suitable for large-scale studies and for aggregation across tens of thousands of studies, eg. [1,2,3,4]. Whereas shotgun metagenome sequencing can provide more accurate taxonomic information, the substantially greater cost and computational demands remain barriers to wider adoption and large-scale studies. In this work we show that more accurate taxonomic classification of short marker-gene sequences can be achieved if information about a sample's source environment is known, increasing the value of marker-gene sequencing for taxonomic profiling of microbial communities.

It has been known at least since the publication of the RDP Classifier [5] that precise classification of short marker-gene sequences at species level is difficult. There have been many attempts are more accurate classification, but despite prolonged effort this is still the case. See [6] for a benchmark comparison of commonly used marker-gene sequence classification methods.

RDP Classifier uses a naive Bayes classifier based on k-mer counts for classification of short reads. A key feature of the RDP Classifier calculation is that it is assumed that each taxonomic class is equally likely. q2-feature-classifier [6], a taxonomic classifier plugin for the QIIME 2 microbiome analysis platform (https://qiime2.org/), uses similar calculations using a scikit-learn [7] naive Bayes classifier. An additional feature of the scikit-learn classifier, however, is that it allows the user to specify the probability of observing each taxonomic class, and this functionality is exposed in q2-feature-classifier.

We have previously shown [6] that inputting the known taxonomic class weights for mock communities greatly increased the accuracy of q2-feature-classifier for classifying each read, beyond the capabilities of any of the other classifiers tested in that study. The vast quantity of public microbiome data that is now available provides information about the likelihood of observing specific taxa in specific environments, enabling us to move beyond current taxonomy classification methods and set probabilities of taxa for specific environments being investigated.

In this work, we show for the first time that realistic taxonomic class weights improve the accuracy of the q2-feature-classifier for real taxonomic abundances, enabling species-level inference with accuracy that was previously only achievable at genus-level. Taxonomic weights were downloaded and assembled using our new utility, q2-clawback. We created weights for 14 EMPO 3 habitat types encompassing 21,513 samples from 7,204 studies downloaded from the Qiita microbial study management platform [1,3,4]. q2-clawback can assemble weights from any appropriately curated set of samples or query Qiita by any available metadata category.

Results

Common methodology

Cross validation was used to analyse the effectiveness of setting the taxonomic weights for the q2-feature-classifier naive Bayes taxonomic classifier. Details are given in the Methods section, but a single cross-validation test follows the pattern:

1. Obtain a set of reference sequences and reference taxonomies.
2. Obtain a set of samples for a given EMPO 3 habitat type, where each sample contains the number of reads observed for each taxon.
3. Perform stratified k-fold cross validation simultaneously on reference sequences and samples.
4. For each fold:
    a. Assemble taxonomic weights by summing read counts across the training set of samples to train the classifier on the training reference sequences; or
    b. Use a preselected set of taxonomic weights to train the classifier on the training reference sequences.
    c. Simulate samples that closely match the taxonomic abundances in the test samples using the test reference sequences, then classify them using the taxonomic classifier trained on the corresponding training sets.

As detailed in the Methods section, Step 2 consisted of obtaining sequence variant abundances and classifying the sequence variants using a standard taxonomic classifier.

The above steps allowed us to determine the effect of several different options for obtaining taxonomic weights on taxonomic classification accuracy. We labelled those options as:

- *Uniform weights*: every taxonomic class is assumed to be equally likely.
- *Bespoke weights*: the weights were obtained from the training sample sets as in 4a.
- *Average weights*: the weights were obtained by averaging the class weights calculated separately for each of the 14 EMPO 3 habitat types used in the test plus Plant surface.
- *Wrong weights*: the weights for the Plant corpus EMPO 3 habitat type.

Uniform weights is the current default assumption for q2-feature-classifier and the only available option for the RDP Classifier. Average weights were tested to determine how important it is to closely match the taxonomic weights with the expected weights for a given sample, and to investigate a classification approach for uncharacterized and unknown sample types. The wrong weights were tested to determine the effect of classifying reads using weights that are clearly inappropriate. For instance, if one were to take a sample that would properly be labelled as Animal distal gut, but erroneously undertake taxonomic classification using a classifier trained using Plant corpus weights. The Plant corpus weights were chosen for this purpose as they have the highest entropy of any of the EMPO 3 habitat types that we tested, meaning that the aggregated taxa for Plant corpus were the least diverse.

We used two measures of classification accuracy: F-measure and Bray-Curtis dissimilarity, made possible because for each test sample there is an expected sample with known taxonomies for each read. Please refer to the methods section for details of how these measures were calculated and averaged across samples.

Appropriate taxonomic weights yield more accurate inference

Taxonomic classification was tested for uniform, bespoke, average, and wrong taxonomic weights across 14 EMPO 3 habitat types for classification at species level. The results are shown for F-measure in Figure 1 and Bray-Curtis dissimilarity in Figure 2. Note that larger F-measure is better and smaller Bray-Curtis dissimilarity is better. The total number of samples for each EMPO 3 habitat type are shown in Table 1. The F-measure results are also summarised in Table 2. We do not report average values across EMPO 3 habitat types as there is no clear way to weight the results.

Bespoke weights were best for every data set by both measures (Figures 1,2). Bespoke weights also offer more consistent performance than uniform weights, with a spread of F-measures across a range of ~0.10, whereas uniform weights occupy F-measures over a range of ~0.18. A similar result is true for Bray-Curtis dissimilarity (~0.11 for bespoke versus ~0.29 for uniform).

Average weights outperformed uniform weights for 10 out of 14 EMPO 3 habitat types for F-measure and 9 out of 14 types for Bray-Curtis dissimilarity. It is worth noting that average weights results are expected to be slightly inflated as the test sets contributed to the average weights.

The wrong weights were better than uniform weights for 7 out of 13 EMPO 3 habitat types for F-measure and Bray-Curtis dissimilarity. Plant corpus was excluded from the results as the Plant corpus weights were used as the wrong weights for other habitat types (and thus were not the wrong weights for Plant corpus).

Appropriate taxonomic weights are more important for finer level taxonomies

We next compare the performance of taxonomic classification using bespoke weights at species level with that using uniform weights at genus level. This is biased towards endorsing the genus level classification comparison as classification accuracy decreases rapidly with finer levels of taxonomic classification [6]. The phenomenon of decreasing accuracy with finer levels of classification is summarised in Figure 3. LOESS plots of F-measure at different taxonomic levels are shown to give a qualitative representation. Classification accuracy decreases for both bespoke and uniform weighted classifiers, but bespoke classifiers are much less prone to this effect (Figure 3).

A direct comparison between classification accuracy using bespoke weights versus uniform weights across the 14 EMPO 3 habitat types is given in Table 2. Remarkably, bespoke weights at species level have greater F-measure than uniform weights at genus level for 10 out of 14 habitat types.

Classification performance is largely predictable based on weights and reference data

Finally, we investigated what factors lead the F-measure to improve between uniform and bespoke weights for some EMPO 3 habitat types more than others. We expect that for a given habitat type these factors are likely characteristics of the distribution of taxonomic abundances across the samples.

To give an idea of the shapes of the taxonomic weights distributions, the cumulative distributions of taxonomic weights for each EMPO 3 habitat type are shown in Figure 4. The distributions are qualitatively similar, with 500 out of 5,403 species accounting for greater than 93% of the mass of each distribution. While it is not shown in Figure 4, it is also worth noting that the most common taxa for each habitat type are similar. The union of the sets of taxa that account for the first 95% of weights for each habitat type contains only 1,571 taxa.

To test the hypothesis that the improvement in F-measure was a function of the diversity of the taxa aggregated across the EMPO 3 habitat type, we regressed F-measure improvement against the entropy of the taxonomic weights for each type, but did not find a strong relationship (Pearson $r^2 \sim 0.05$, p-value ~ 0.7). We concluded that differences in diversity between habitat types were not the main factor driving the improvements in F-measure.

We next explored the hypothesis that beyond simple diversity, specific abundances were important to how much of an improvement was observed between bespoke and uniform weights. We tested this hypothesis by repeating the above cross-validation tests, but using a different machine learning classifier (a modified k nearest neighbours classifier), not performing

cross validation on samples, and performing leave-one-out cross validation on sequences (see Methods for more details). The effect was to determine what aspects of F-measure improvement are dependent on the bespoke weights, but not on weights variation between samples from the same habitat type and that are not specific to naive Bayes classification. This test also provided validation of our existing methods. The improvements in F-measure between bespoke and uniform weights for the two cross validation experiments were observed to be positively correlated (Pearson $r^2 \sim 0.68$, p-value ~ $3x10^{-4}$, Figure 5).

These findings suggest that the fine structure of the habitat-specific taxonomic weights is the main driver of improvement over uniform weights and that factors such as measures of diversity, taxonomic variation between samples, or classification method are secondary.

Discussion

The results we report here illustrate that that informed class weights can improve the accuracy and specificity of short-read taxonomic classification. Due to the commitment of many researchers in the microbiome community to make their data publicly accessible following initial publication, and efforts to provide well-curated access to these data in resources such as Qiita, better alternatives now exist for assessing taxonomic abundance distributions across a wide variety of ecosystem types. Class weights based on a sample's EMPO 3 habitat type can be obtained using q2-clawback and in our tests this information always led to improved classification performance. This is an example of how data transparency, public data repositories, and open-source software are advancing and democratizing microbiome research.

We showed that closer knowledge of a sample's expected class weights improved performance. Bespoke weights outperformed average weights, and average weights outperformed the wrong weights. Even the wrong weights slightly outperformed the uniform weights in this comparison. This may be because wrong weights still better approximate average taxonomic distribution of a sample, and the taxonomic compositions of most ecosystems are dominated by many of the same groups of taxa; uniform class weights ignore the taxonomic distribution found in nature and assume that exceedingly rare species are equally likely to be observed as ubiquitous species.

Hence, if more precise knowledge about a sample's expected class weights is available, that knowledge potentially could be used to improve classification accuracy over using its EMPO 3 habitat type. For instance, a stool sample could be classified with weights obtained from q2-clawback by requesting "sample_type" to be "stool" or "Stool" in Qiita. Alternatively, if a user has a private database of samples of the same type, then q2-clawback could be used to assemble them into an appropriate set of class weights, as shown in the online tutorial (https://forum.qiime2.org/t/using-q2-clawback-to-assemble-taxonomic-weights/5859). The EMPO 3 habitat type for which we had the fewest samples ("Surface (saline)") had 152 samples from 58 studies, demonstrating that even sparsely characterized class weights still promise performance enhancements relative to assuming uniform class weights.

q2-clawback also provides a facility for quickly predicting how much of an improvement using a specific set of class weights will provide over uniform weight assumptions, using the kNN classifier weighting F-measure improvement as a predictor (see https://forum.qiime2.org/t/using-q2-clawback-to-assemble-taxonomic-weights/5859). If users are interested in further exploring whether using bespoke weights would improve classification for their samples, they can access our full suite of cross validation tests to inform their decision. Details of how to obtain that code are provided in Methods.

While this study identifies clear improvements to the current workflows for processing microbial data, there remains scope for improvement. We have shown that knowledge of the expected taxonomic abundances of a sample improves taxonomic classification accuracy. In some sense this is a first-order approximation. More accurate information would include knowledge of how much the taxonomic abundances are expected to vary between samples, or full understanding of the distribution of taxonomic abundances. Further advances could be made by bringing this insight to a more model-driven approach.

Improvements in taxonomic composition are limited by the quality of sequence reference databases used to classify the source datasets. Incorrect annotations in reference databases can propagate these misclassifications in class weight assemblies and downstream taxonomic classifications. This reinforces the importance of ongoing curation efforts to refine these databases and additional improvements in taxonomic classification methods to detect and reclassify incorrectly annotated data. Reclassification of reference sequences and class weight taxonomies can be performed as improved methods are developed, promising continued refinement as incremental advances are implemented.

The proliferation of public microbiome data provides an opportunity to improve inferences about microbial composition in yet-uncharacterized samples. These inferences will continue to be refined as more data are collected, particularly for under-characterized ecosystem types, as sequence reference databases are improved, and as the quality and specificity of taxonomic assignments in these ecosystems evolve. Class weights can also be assembled from shotgun metagenome sequence data (or any other taxonomic profiling method), allowing potentially less biased, more specific estimates of taxonomic composition for environmental sample types as public repositories for shotgun metagenome sequencing continue to grow to encompass a range of environment types.

Methods

Data

We downloaded all public 150 nucleotide 16S v4 samples for 18 EMPO 3 habitat types from Qiita [1],[3,4]] using q2-clawback. The downloaded data consisted of sequence variant and abundance information. q2-clawback uses redbiom (https://github.com/biocore/redbiom) [8] to access Qiita [1].

The three EMPO 1 control EMPO 3 habitat types were excluded, as well as Hypersaline (saline), Aerosol (non-saline), and Plant surface, which all had fewer than nine samples in the Qiita database for 150 nt sequence variants.

For the following analysis, sequence-variant level data was discarded and only taxonomic abundance information was retained. The sequence variants were classified using the standard q2-feature-classifier naive Bayes classifier based on Greengenes 99% identity OTU reference data [9] to obtain empirical taxonomic abundance data for each sample. The naive Bayes classifier was trained using the “balanced” parameter recommendations given in [6].

The number of samples downloaded for each EMPO 3 habitat type are shown in Table 1.

Table 1. **Sample counts for each EMPO 3 habitat type.**

EMPO 3 Habitat Type	Number of Samples
Animal corpus	1,158
Animal distal gut	5,632
Animal proximal gut	903
Animal secretion	974
Animal surface	1,839
Plant corpus	543
Plant rhizosphere	328
Sediment (non-saline)	188
Surface (non-saline)	1,383
Soil (non-saline)	2,802
Water (non-saline)	4,769

Sediment (saline)	414
Surface (saline)	152
Water (saline)	428

Table 2. **Species-level classification is always better with habitat-specific taxonomic weights and often exceeds genus-level classification accuracy under the assumption that each species is equally likely. Bolded rows indicate EMPO 3 Habitat Type where species-level accuracy with the bespoke classifier is greater than genus-level accuracy with the uniform classifier.**

	F-measure*			
	Bespoke		Uniform	
EMPO 3 Habitat Type	Genus	Species	Genus	Species
Animal corpus	**0.89**	**0.87**	**0.83**	**0.76**
Animal distal gut	**0.89**	**0.87**	**0.85**	**0.81**
Animal proximal gut	**0.91**	**0.83**	**0.81**	**0.72**
Animal secretion	0.95	0.89	0.92	0.73
Animal surface	0.94	0.88	0.92	0.76
Plant corpus	0.95	0.93	0.94	0.90
Plant rhizosphere	**0.94**	**0.92**	**0.91**	**0.89**
Sediment (non-saline)	**0.91**	**0.90**	**0.88**	**0.87**
Soil (non-saline)	**0.92**	**0.90**	**0.88**	**0.85**
Surface (non-saline)	0.95	0.89	0.92	0.73
Water (non-saline)	**0.93**	**0.91**	**0.88**	**0.86**
Sediment (saline)	**0.91**	**0.90**	**0.88**	**0.87**
Surface (saline)	**0.89**	**0.85**	**0.85**	**0.79**
Water (saline)	**0.89**	**0.87**	**0.84**	**0.81**

*F-measures are for 5-fold cross validation where classifiers trained using two taxonomic weighting strategies are tested on sequences and empirical taxonomic abundances that were

not used in training. Bespoke weights are habitat-specific weights and uniform weights is the standard assumption assumes that every taxon in the reference database is equally likely to be observed. Tests were based on 21,513 empirical samples across the 14 habitat types. Bold rows indicate where bespoke species-level F-measure exceeded the uniform genus level F-measure.

Cross Validation Using Empirical Taxonomic Abundance

The basic methodology for the cross validation tests is given in the Results section. We now provide more detail. All of the code necessary is available from the paycheck repository (https://github.com/BenKaehler/paycheck/releases/tag/0.0.1) or from q2-clawback (https://github.com/BenKaehler/q2-clawback/releases/tag/0.0.2).

Data was obtained for Step 2 as detailed above.

Step 3 requires more explanation. We performed 5-fold cross validation in each instance. Standard stratification for 5-fold cross validation requires that at least five sequences exist for each taxonomy, which is not the case for the 99% identity Greengenes reference taxonomy. We therefore formed a stratum for each taxonomy for which five or more reference sequences existed (large taxonomies) and merged the remaining taxonomies (small taxonomies) into those strata. A single large taxonomy was chosen for each small taxonomy by training a naive Bayes classifier on the large taxonomies, classifying the reference sequences in the small taxonomies, then voting weighted by confidence. Shuffled stratified 5-fold cross validation was then implemented using a standard library call to scikit-learn.

Step 4 requires more detailed exposition for the same reason. Cross validation was performed simultaneously on samples and reference sequences. Each sample consisted of a set of taxonomies and their abundances. Bespoke weights were formed by aggregating those counts across the training samples. As a result of the merged strata in Step 3, some taxonomies that were present in the bespoke weights were not present amongst the taxonomies of the training sequences. Any such taxonomy was mapped to the nearest taxonomy that was present amongst the taxonomies represented by the training sequences, as measured by the voting system from Step 3. This problem also affected the wrong and average weights, and it was solved analogously in those cases.

Step 5 was complicated by the same issue. In this step, samples were simulated by drawing sequences from the test sequences in such a way as to closely resemble the taxonomic abundances of the test samples. Again as a result of the merged strata in Step 3, some taxonomies that were present in the test samples were not present in the taxonomies of the test sequences. In the same way as for Step 4a, any missing species-level taxonomy was mapped to the closest taxonomy for a sequence present in the test sequences. Once missing

taxonomies were resolved, samples were simulated by drawing test sequences as evenly as possible from each taxonomy so that any read count was a whole number.

For the q2-feature-classifier naive Bayes classifiers that were reported in this study, we used the recommended “balanced” parameters as recommended in [6] for uniform weights. That is, we used a confidence level of 0.7, even for the bespoke, average, and wrong weights. In [6] a confidence level of 0.92 was recommended for bespoke weights. We tested the classifiers at this level but in all cases the results were dominated by the less conservative confidence level of 0.7.

F-measure and Bray-Curtis dissimilarity were calculated for each sample and taxonomic level using the q2-quality-control QIIME 2 plugin (https://github.com/qiime2/q2-quality-control)[10]. F-measure for each fold was aggregated across samples by weighting by the total read count for each sample. Confidence levels were calculated using the StatsModels package [11]. Bray-Curtis dissimilarity was averaged across samples without weighting, but samples with less than 1,000 reads were filtered out.

Fast Cross Validation

The methodology for the alternative classification test was as follows:

1. Obtain a set of taxonomic weights.
2. Obtain a set of reference sequences and reference taxonomies for those reads.
3. Undersample the reference sequences so that at most 10 sequences are present for each taxonomy.
4. Extract k-mer counts for the reference sequences using the same parameters as used for the other tests in this study.
5. Use leave-one-out cross validation to predict classification for the reference sequences using weighted k nearest neighbours cross validation that is
    a. weighted by the taxonomic weights or
    b. uniformly taxonomically weighted.

Undersampling at Step 3 was performed using centroid clustering from the imbalanced-learn package [10,12].

Feature extraction and nearest neighbour discovery at Steps 4 and 5 was performed using scikit-learn [7]. Sequences were not weighted as a function of distance. Sequences were weighted for classification to achieve bespoke or uniform weights according to the scheme proposed in [7,13]. This is a simple technique where each sequence in the consensus is weighted so that the total weight for the sequences of each taxonomy is the appropriate taxonomic weight.

Finally, F-measures were calculated using the scikit-learn implementation selecting weighted, micro-averaged calculations.

Acknowledgments

This work was funded in part by National Science Foundation award 1565100 to JGC, awards from the Alfred P. Sloan Foundation to JGC, awards from the Partnership for Native American Cancer Prevention (NIH/NCI U54CA143924 and U54CA143925) to JGC, and National Health and Medical Research Council of Australia award APP1085372 to GAH and JGC. These funding bodies had no role in the design of the study, the collection, analysis, or interpretation of data, or in writing the manuscript. We would like to thank Daniel McDonald for useful comments and advice.

References

1. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017 Nov 23;551(7681):457–63.

2. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14.

3. Qiita [Internet]. [cited 2018 Jul 24]. Available from: https://qiita.ucsd.edu

4. Gonzalez, A, Navas-Molina, JA, Kosciolek, T, McDonald, D, Vázquez-Baeza, Y, Swafford, AD, Orchanian, SB, Sanders, JG, Shorenstein, J, Brislawn, CJ, Holste, H, Robbins-Pianka, A, Wang, M, Rideout, JR, Bolyen, E, Dillon, M, Caporaso, JG, Dorrestein, PC and Knight, R. Qiita's web-enabled platform accelerates microbiome meta-analyses from months to minutes. Nat Methods. in press;

5. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007 Aug;73(16):5261–7.

6. Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome. 2018 May 17;6(1):90.

7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.

8. Daniel McDonald, Benjamin Kaehler, Antonio Gonzalez, Jeff DeReus, Gail Ackermann, Rob Knight. redbiom: a rapid sample discovery and feature characterization system.

9. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006 Jul;72(7):5069–72.

10. qiime. qiime2/q2-quality-control [Internet]. GitHub. [cited 2018 Jul 24]. Available from:

https://github.com/qiime2/q2-quality-control

11. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 57–61.

12. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. J Mach Learn Res. 2017;18(17):1–5.

13. Liu W, Chawla S. Class confidence weighted kNN algorithms for imbalanced data sets. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2011. p. 345–56.

Figures

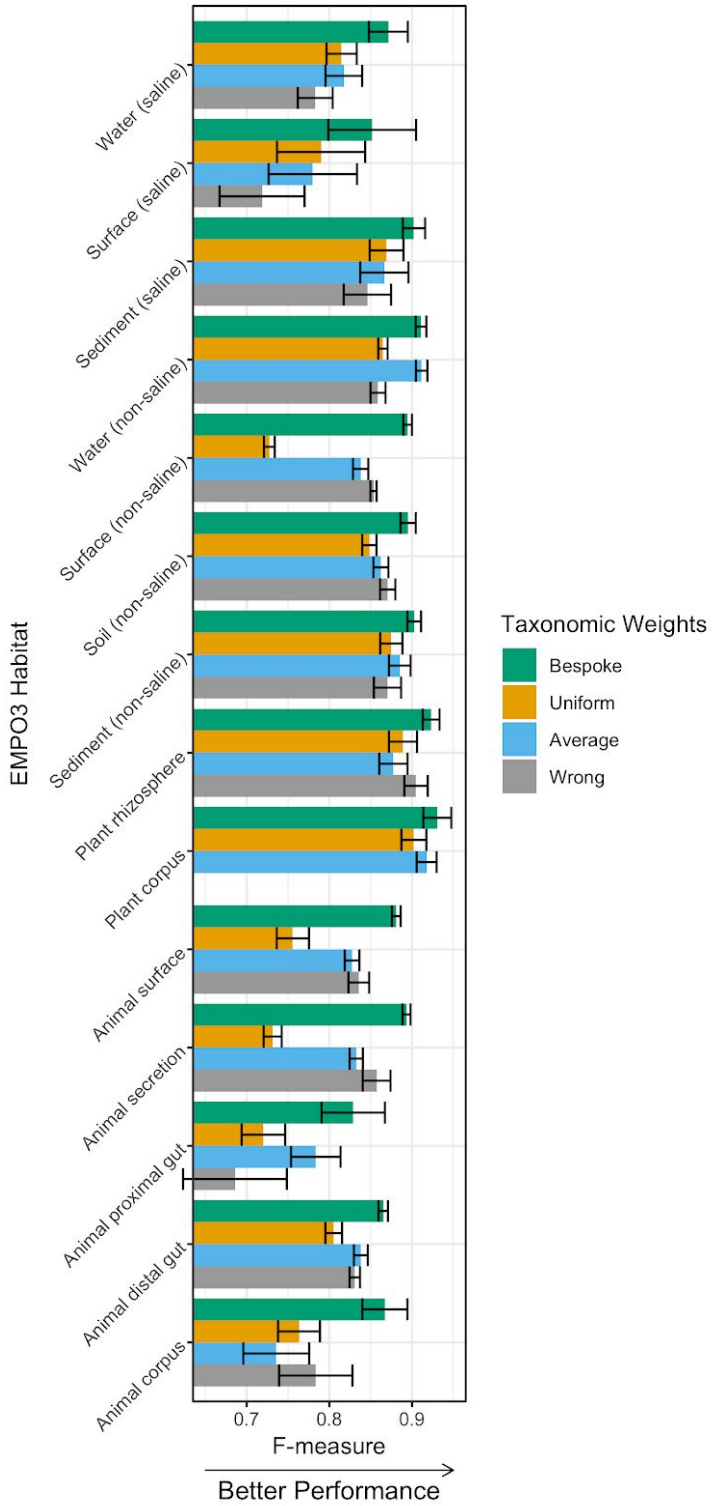

Figure 1. Species-level classification accuracy is always better with habitat-specific taxonomic weights. F-measures are for 5-fold cross validation where classifiers trained using a variety of taxonomic weighting strategies are tested on sequences and empirical taxonomic abundances that were not used in training. Bespoke weights are habitat-specific weights, average weights are averaged across 15 habitat types, wrong weights are Plant corpus taxonomic weights, and uniform weights is the standard assumption that every taxon in the reference database is equally likely to be observed. Tests were based on 21,513 empirical samples across the 14 habitat types. Error bars show 95% confidence intervals across folds. Note that wrong weights are missing for Plant corpus, as described in the text.

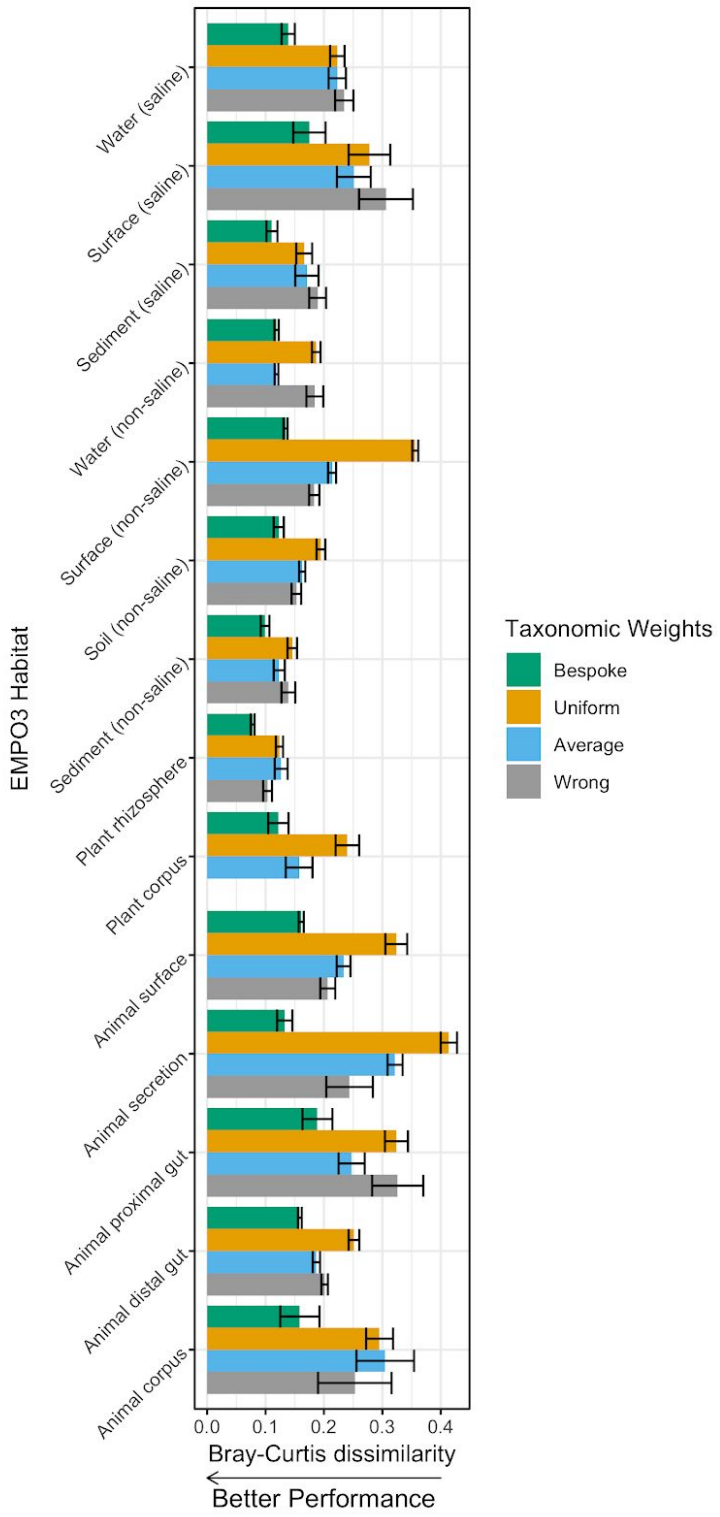

Figure 2. Beta diversity between observed and expected samples is lower with habitat-specific taxonomic weights. Bray-Curtis dissimilarities are for 5-fold cross validation where classifiers trained using a variety of taxonomic weighting strategies are tested on sequences and empirical taxonomic abundances that were not used in training. Bespoke weights are habitat-specific weights, average weights are averaged across 15 habitat types, wrong weights are Plant corpus, and uniform weights is the standard assumption that every taxon in the reference database is equally likely to be observed. Tests were based on 18,222 empirical samples across the 14 habitat types. Error bars show 95% confidence intervals across folds. Note that wrong weights are missing for Plant corpus, as described in the text.

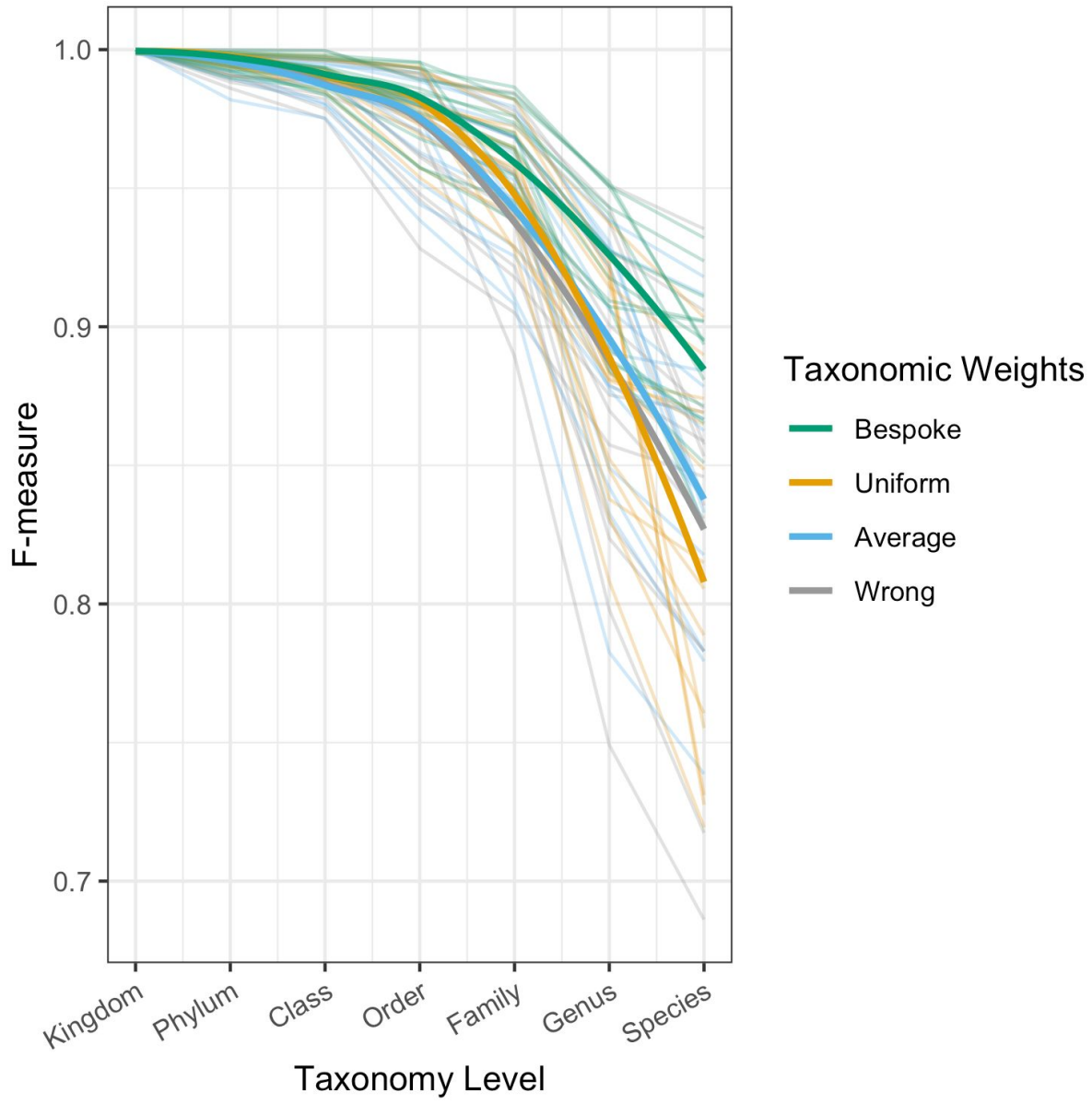

Figure 3. Classification accuracy drops with increasing taxonomic specificity. Using the same cross-validation strategy as for Figure 1, classification F-measure drops as finer levels of classification are required, but is much more consistent across levels for classifiers with bespoke (habitat-specific) weights. Faint lines show results for 14 EMPO 3 habitat types. Bold lines show LOESS plots to demonstrate trends.

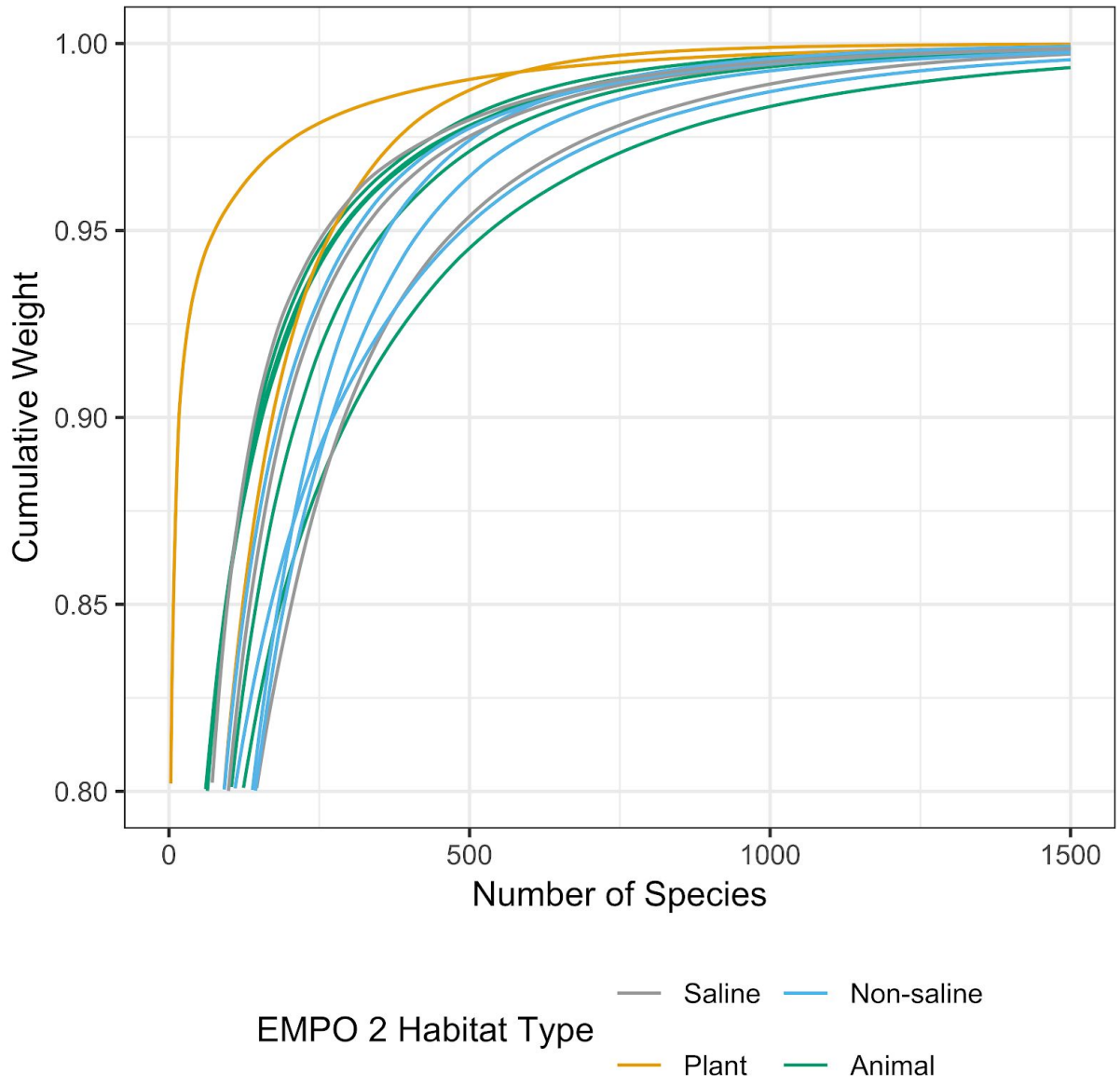

Figure 4. Cumulative weights for different habitat types display similar diversity trends. Lines show the cumulative taxonomic weights for 14 EMPO 3 habitat types as a function of species count, coloured by EMPO 2 habitat type. Taxa are ordered separately for each habitat type from most to least abundant. The most peaked distribution is Plant corpus, where ~73% of reads were mapped to a single taxonomy in the Cyanobacteria phylum, Chloroplast class.

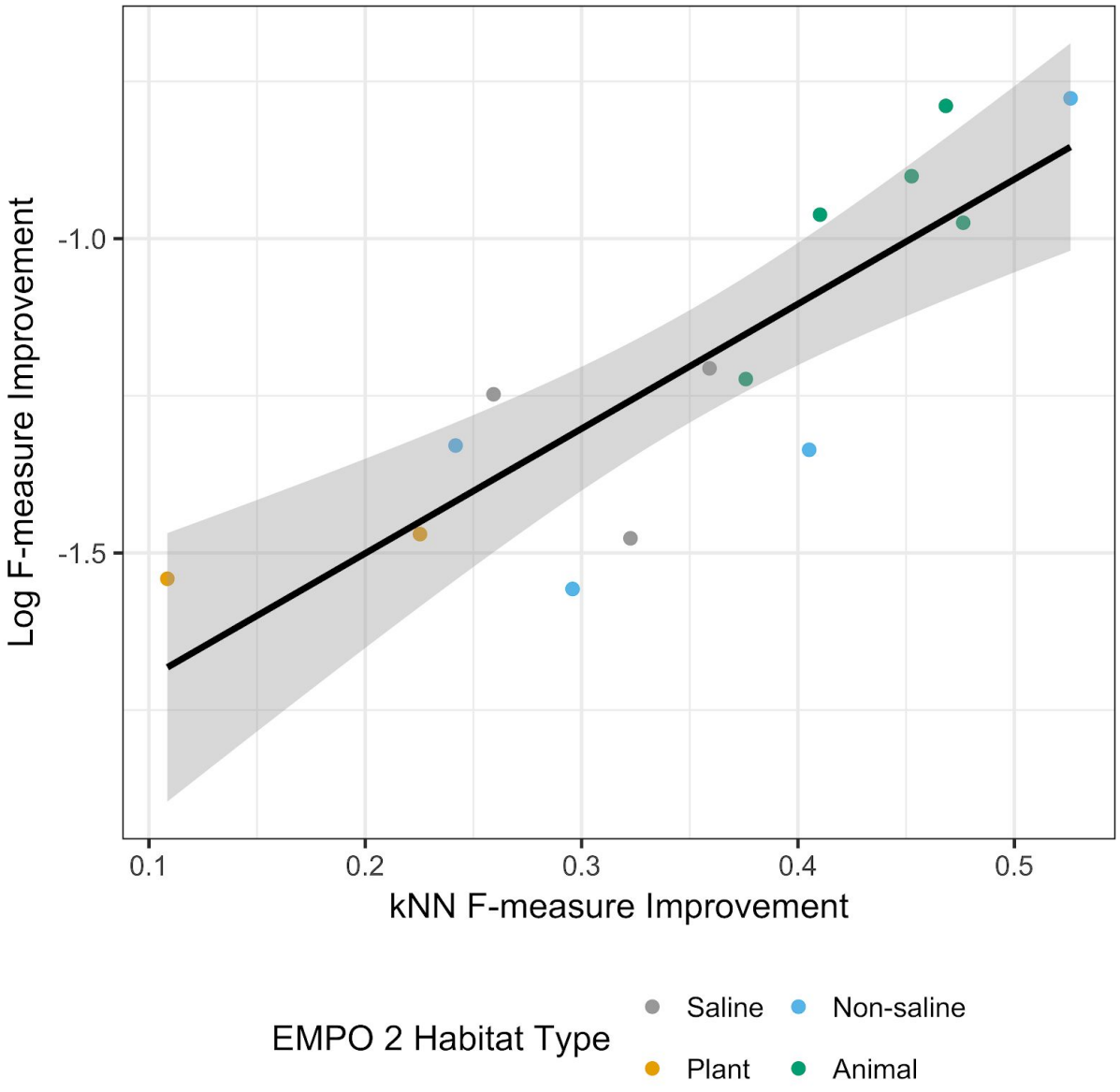

Figure 5. Improvement for classification with habitat-specific weights is largely dependent on weight distributions. There is a strong relationship between the performance improvements observed using full cross-validation experiments reported in Figure 1 and a much simpler k nearest neighbours, leave-one-out cross-validation experiment (Pearson r^2 ~ 0.68, p-value ~ $3x10^{-4}$). The simpler experiment did not include cross validation on the taxonomic abundances observed in empirical abundances, indicating that much of the variance in performance improvement over standard methods comes from the relationship between the bespoke (habitat-specific) weights and the reference data geometry.