Identifying accurate metagenome and amplicon software via a meta-analysis of benchmarking studies

Paul P. Gardner[1,2], Renee J. Watson[1], Xochitl C. Morgan[3], Jenny L. Draper[4], Robert D. Finn[5], Sergio E. Morales[3], Matthew B. Stott[1]

1. Biomolecular Interactions Centre, School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.
2. Department of Biochemistry, University of Otago, Dunedin, New Zealand.
3. Department of Microbiology and Immunology, University of Otago, Dunedin, New Zealand.
4. Institute of Environmental Science and Research, Porirua, New Zealand.
5. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Abstract

Environmental DNA sequencing has rapidly become a widely-used technique for investigating a range of questions, particularly related to health and environmental monitoring. There has also been a proliferation of bioinformatic methods for analysing metagenomic and amplicon datasets, which makes selecting adequate methods a significant challenge. A number of benchmark studies have been undertaken; however, these often present conflicting results. We have applied a network meta-analysis method to identify software methods that are generally accurate for mapping DNA sequences to taxonomic hierarchies. Based upon these results we have identified some methods and computational strategies that produce robust predictions.

Introduction

Metagenomics, meta-barcoding and related high-throughput environmental DNA (eDNA) sequencing approaches have accelerated the discovery of small and large scale interactions between ecosystems and their biota. The application of these methods has advanced our understanding of microbiomes, disease, ecosystem function, security and food safety [1–3]. A number of strategies have been explored for interpreting eDNA sequencing results (Fig. 1). These can be broadly divided into amplicon (barcoding) and genome-wide (metagenome) based approaches. Amplicon, or barcoding, -based approaches typically target genomic marker sequences such as ribosomal RNA genes[4–7] (16S, 18S, mitochondrial 12S), RNase P RNA[8], or internal transcribed spacers (ITS) between ribosomal RNA genes[9]. These regions are amplified from extracted DNA by PCR, and the resulting DNA libraries are sequenced. In contrast, genome-wide, or metagenome, -based approaches sequence the entire pool of DNA extracted from a sample with no preferential targeting for particular markers or taxonomic clades. Both approaches have limitations that influence downstream analyses. For example, amplicon target regions may have unusual DNA features (e.g. large insertions or diverged primer annealing sites), and consequently these DNA markers may fail to be amplified by PCR[10]. While the metagenome-based methods are not vulnerable to primer bias, they may fail to detect genetic signal from low- abundance taxa if the sequencing does not have sufficient depth, or may under-detect sequences with a high G+C bias [11,12].

The resulting reads from high-throughput sequencing (HTS) can be analysed using a number of different strategies (outlined in Fig. 1)[13–15]. The fundamental goal of many of these studies is to assign taxonomy to reads as specifically as possible, and in some cases to cluster highly-similar reads into "operational taxonomic units" (OTUs)[16]. For greater accuracy in taxonomic assignment, reads can be assembled into longer "contigs" using any of the large number of sequence assembly tools[17–19], that operate in a *de novo* or reference-based fashion. The reference-based methods (also called "targeted gene assembly") make use of conserved sequences to constrain the sequence assembly problem. These have a number of reported advantages including reducing chimeric sequences, and improving the speed and accuracy of assembly relative to *de novo* methods[20,21].

For amplicon-based studies, contigs or reads may be clustered into closely related groupings; again, these can be clustered with either *de novo* or reference-based methods, forming "operational taxonomic units" (OTUs). Metagenomic reads are generally mapped to a reference database of sequences labelled with a hierarchical taxonomic classification. The mapped sequences will range from identical to divergent matches, and the level of divergence and distribution of taxonomic assignments for the matches allows an estimate to be made of where the sequence belongs in the established taxonomy with high probability. This is commonly performed using the lowest common ancestor approach (LCA)[22]. Some tools, however, avoid this computationally-intensive sequence similarity estimation, and instead use alignment-free approaches based upon sequence composition statistics (e.g. nucleotide and k-mers frequencies) to estimate taxonomic relationships[23].

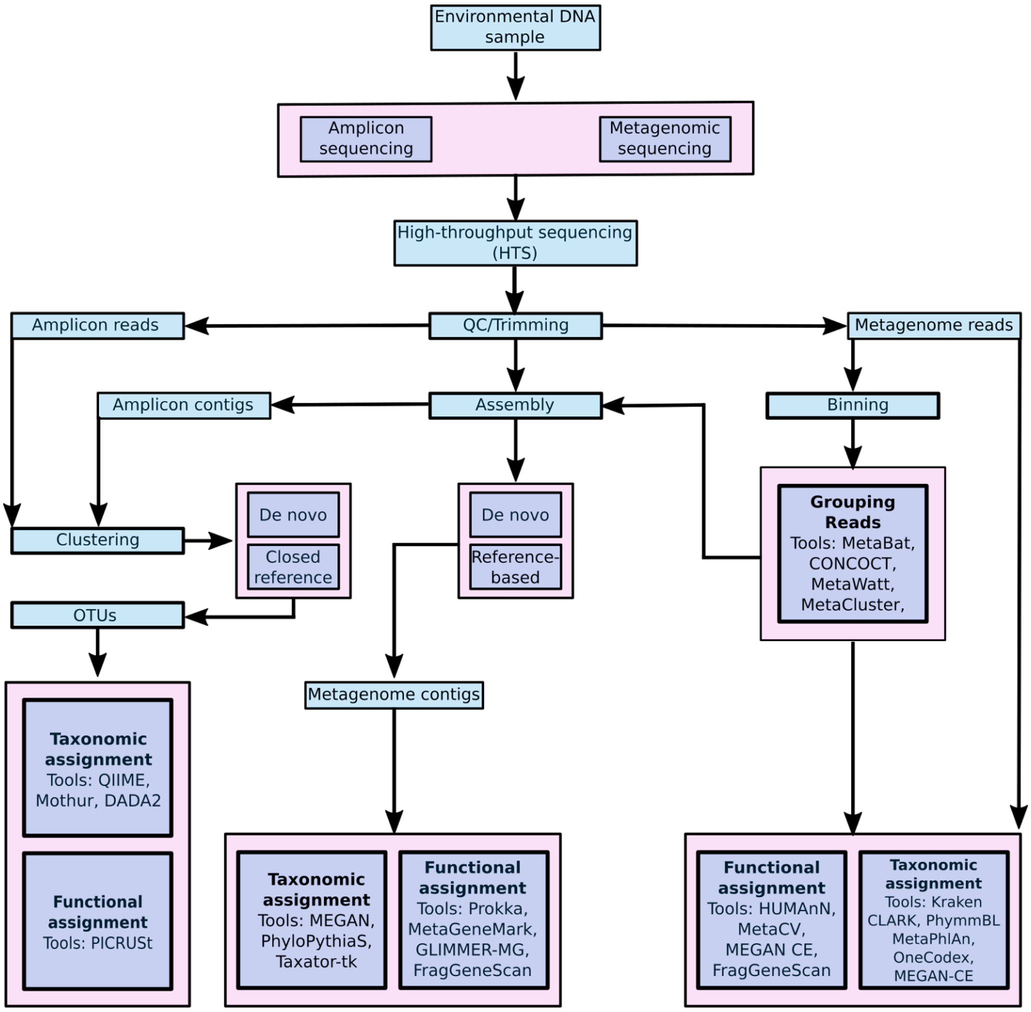


Figure 1: Different eDNA data production and analysis pipelines. The main split is between amplicon or marker-gene based approaches and the shotgun metagenomics strategies.

As eDNA methods become more widely adopted, the potential for false discoveries also increases -- with the apparent false discovery of high-risk bacterial pathogens in subway systems and a global outbreak of an Australian monotreme as two notable examples [24]. These issues are exacerbated by biased selection of species for genome sequencing, as pathogens and iconic species have been prioritised over other organisms, resulting in an over-abundance of sequences corresponding to these in nucleotide databases. While projects such as the "genomic encyclopedia of bacteria and archaea" have attempted to partially address this, the bias remains an issue [25].

In this study we review and compare six independent evaluations [26–31] of 39 eDNA analysis tools. The comparison is restricted to evaluations of the mapping from DNA sequence to taxonomic origin accuracy, the corresponding software tools all perform this (and potentially additional functions) using a range of different strategies (see Figure 1). We have used network meta-analysis techniques to resolve the considerable amount of variation and conflicting reports between the different studies, resulting in a short list of methods that have been consistently reported to produce accurate interpretations of metagenomics results.

Overview of eDNA analysis evaluations

Independent **benchmarking of bioinformatic software** provides a valuable resource for determining the relative performance of software tools, particularly for problems with an overabundance of tools. Some established principles for reliable benchmarking are: 1. The main focus of the study should be the evaluation and not the introduction of a new method; 2. The authors should be reasonably neutral (i.e. not involved in the development of methods included in an evaluation); and 3. The test data, evaluation and methods should be selected in a rational way [32]. The first two criteria are straightforward to discern, but the latter criteria is the more difficult to evaluate as it includes avoiding the over-optimistic reporting of accuracy due to overtraining, and the cherry-picking of data and metrics for accuracy reporting, which can result in inflated accuracy valuations [33–35]. Based upon literature reviews and citation analyses, we have identified six published evaluations of eDNA analysis methods that meet our criteria for inclusion (in our assessment)[26–31]. These evaluations are summarised in Table 1. In the following section we will discuss issues around collecting trusted datasets, including selection of positive and negative control data that avoid any datasets upon which methods may have been over-trained. We describe measures of accuracy for predictions and describe the characteristics of ideal benchmarks, with examples of published benchmarks that meet these criteria.

Paper	**Positive Control**	**Negative Controls**	**Reference exclusion method**	**Metrics**	**Number of methods**
Bazinet *et al.* (2012) [26]	4 published *in silico* mock communities from **742** taxa[36–39]	-	-	Read level	10
Lindgreen *et al.* (2016) [28]	6 *in silico* mock communities from **417** different genera.	Shuffled sequences	Simulated evolution	Read level	14
McIntyre *et al.* (2017) [29,40]	14 *in vitro* and 21 *in silico* mock communities from **846** species.	(1) blanks with human DNA spiked in, (2) "nullomers" -- 17-mers not found in any reference (3) environmental samples known to **not** contain Bacillus	-	Read & Taxonomy level	11
Peabody *et al.* (2015) [27]	1 published *in silico* mock community,[41] 1 mixed *in silico*+*in vitro* mock community from **11** different species.	-	Clade exclusion	Read level	10
Sczyrba *et al.* (2017) [30]	3 *in silico* mock communities from **689** newly sequenced bacteria & archaea plus 598 plasmid, viral & other circular elements.	-	New genome sequences	Taxonomy level	14
Siegwald *et al.* (2017) [31]	36 *in silico* mock communities from **125** bacterial genomes.	-	-	Read level	6

Table 1: A summary of the main features of the software evaluations selected for this study.

Positive and negative control dataset selection

The selection of datasets for evaluating software can be a significant challenge due to the need for these to be reliable, well-curated, robust and representative of the population of possible datasets[42]. The published **positive control** datasets can be divided into two different strategies (summarised in Table 1). These "*in vitro"* and "*in silico"* approaches generate mock communities for evaluation purposes. The *in vitro* methods involve generating microbial consortia with predetermined ratios of microbial strains then extracting the consortium DNA and sequencing and analysing the mix using standard metagenomic pipelines[43,44]. The accuracy of the genome assembly, genome partitioning (binning) and read depth proportional to consortium makeup can then be used to confirm software accuracy. The *in silico* method takes the same approach except that publicly-available genome sequences are selected, and reads simulating metagenome sequencing are derived from these [41,45–47]. It is important to note that ideally-simulated reads are derived from species that are **not present** in established reference databases, as this is a more realistic simulation of most eDNA surveys. A number of different strategies have been used to control for this [26,27,29] (Fig. 2). Peabody *et al.* used "clade exclusion", in which sequences used for an evaluation are removed from reference databases for each software tool[27]. Lindgreen *et al.* used "simulated evolution" to generate simulated reads of varying evolutionary distances from reference sequences[28]. Sczyrba *et al.* restricted their analysis to reads sampled from recently-deposited genomes, increasing the chance that these are not included in any reference database[30].

Another important consideration is the use of **negative controls**. These can be reads derived from randomised sequences[28], or from sequences not expected to be found in reference databases[29]. The resultant reads from these sets can be used to determine false-positive rates for different tools. We have summarised the positive and negative control datasets from published software evaluations in Table 1, along with other features of different evaluations of eDNA analysis software.

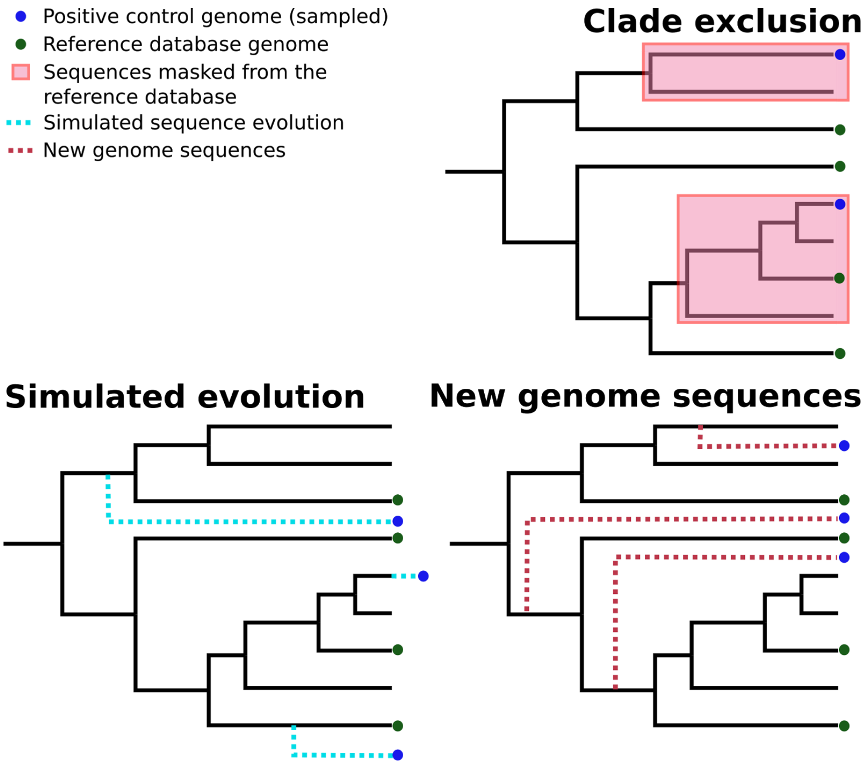

● Positive control genome (sampled)
● Reference database genome
Sequences masked from the reference database
Simulated sequence evolution
New genome sequences
Clade exclusion
Simulated evolution
New genome sequences

Figure 2: Three different strategies for generating positive control sequencing datasets, i.e. genome/barcoding datasets of known taxonomic placement that are absent from existing reference databases.

Metrics used for software benchmarking.

The metrics used to evaluate software play an important role in determining the fit for different tasks. For example, if a study is particularly interested in identifying rare species in samples, then a method with a high true-positive rate (also called **sensitivity** or **recall**) may be preferable. Conversely, for some studies, false positive findings may be particularly detrimental, in which case rates of true positive calling may be sacrificed in exchange for a lower false positive rate. Some commonly used measures of accuracy, including *sensitivity* (recall/true positive accuracy), *specificity* (true negative accuracy) and *F-measure* (the trade-off between recall and precision) are summarised in Table 2.

The definitions of “true positive”, “false positive”, “true negative” and “false negative” (TP, FP, TN and FN respectively) are also an important consideration. Many metagenome analysis methods report a taxonomic distribution as the default output (e.g. 15% Acidobacteria, 11% Actinobacteria, 4% Bacteroidetes, …). This means that the expected taxonomic distributions and the predicted distributions can be compared. Some studies, for selected taxonomic ranks, use the presence/absence of taxa in the expected and predicted taxonomic profiles to determine values for TP, FP, TN and FN. For example if the phylum Bacteroidetes was found in both, then the TP value is incremented by 1 [30].

This approach has some drawbacks as errors have the potential to cancel each other out. For example, if reads from taxon 1 are wrongly mapped to taxon 2, this error may be masked by reads from taxon 2 mapping to taxon 1 (larger, less direct error-cycles are also possible). This results in an incorrect true-positive assignments for both errors. Therefore, a more robust estimation of TP can be made by determining whether individual sequence reads were correctly assigned, at a particular taxonomic rank [27,28,31]. However, this per-read information can be challenging to access from some software tools.

Successfully recapturing the frequencies of different taxonomic groups as a **measure of community diversity** is a major aim for eDNA analysis projects. There have been a variety of approaches for quantifying the accuracy of this information. Pearson’s correlation coefficients [26], L1-norm [30], the sum of absolute log-ratios [28], the log-modulus [29] and the Chao 1 error [31] have each been used. This lack of consensus has made comparing these results a challenge.

The amount of variation between the published benchmarks, including varying taxonomies, taxonomic thresholds and whether reads or taxa were used for evaluations can also impede comparisons between methods and the computation of accuracy metrics. To illustrate this we have summarised the variation of F-measures (a measure of accuracy) between the six benchmarks we are considering in this work (Figure 3).

$Sensitivity = \frac{TP}{TP+FN}$ (a.k.a. recall, true positive rate)	$Specificity = \frac{TN}{TN+FP}$ (a.k.a. true negative rate)	$PPV = \frac{TP}{TP+FP}$ (a.k.a. positive predictive value, precision, sometimes mis-labelled "specificity")
$F\ measure = \frac{2*Sensitivity*PPV}{Sensitivity+PPV} = \frac{2TP}{2TP+FP+FN}$ (a.k.a. F1 score)	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	$FPR = \frac{FP}{FP+TN}$ (a.k.a false positive rate)

Table 2: Some commonly used measures of "accuracy" for software predictions. These are dependant upon counts of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) which can be computed from comparisons between predictions and ground-truths [48].

Review of Results

We have mined independent estimates of sensitivity, positive predictive values (PPV) and F-measures for 39 eDNA software tools, from six published software evaluations. A matrix showing presence-or-absence of software tools in each publication is illustrated in Figure 4A. Comparing the list of 39 eDNA software tools to a publicly available list of eDNA software tools based upon literature mining and crowd-sourcing, we found that 45% (39/87) of all published tools have been evaluated [49]. The unevaluated methods may generally fall into the very recently published (and therefore have not been evaluated yet) or may no longer be available, functional, or provide results in a suitable format for evaluation (see Figure 3A). Several software tools have been very widely cited (Figure 3B), yet caution must be used when considering citation statistics, as the number of citations is not a reliable proxy for accuracy [50]. For example, the tools that are published early are more likely to be widely adopted. Furthermore, some articles are not necessarily cited for the software. For example, the MEGAN1 manuscript is often cited for one of the first implementations of the LCA algorithm for assigning read-similarities to taxonomy [22].

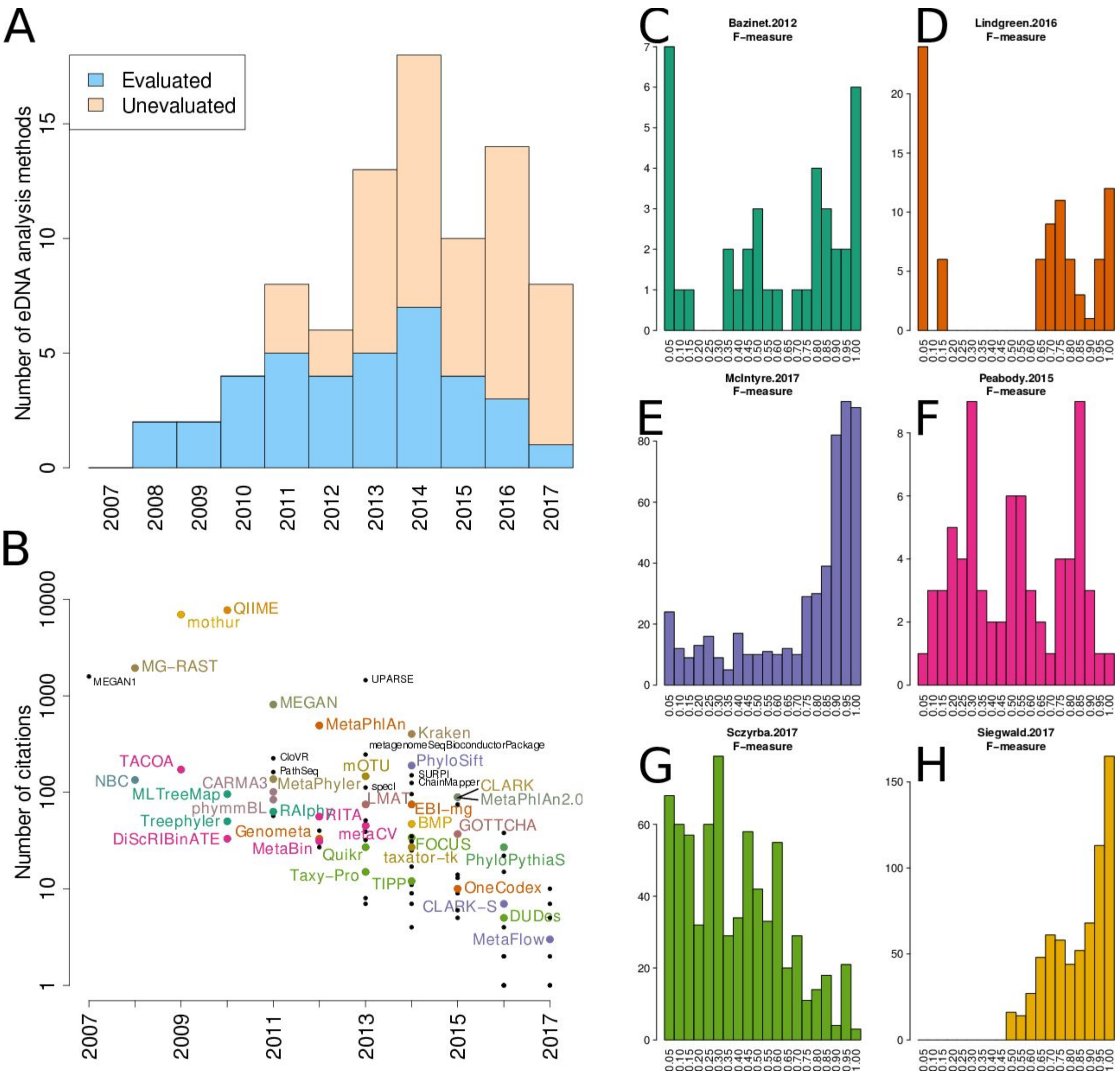


Figure 3: A: more than 80 eDNA analysis tools have been published in the last 10 years. A fraction of these (46%) have been independently evaluated. **B:** The number of citations for each software tool versus the year it was published. Methods that have been evaluated are coloured and labelled (using colour combinations consistent with evaluation paper(s), see right). Those that have not been evaluated, yet have been cited >100 times are labelled in black. **C-H:** The distributions of accuracy estimates based upon reported F-measures using values from 6 different evaluation manuscripts [26–31].

After manually extracting sensitivity, PPV and F-measures (or computing these) from the tables and/or supplementary materials for each publication,[26–31] we have considered the within-publication distribution of accuracy measures (see Figures 3C-H). These figures clearly show that for each publication there are major differences in the accuracy distributions (skewed and multimodal), and measures of centrality and variance. For example, the F-measures from the McIntyre and Siegwald studies are comparatively high (median values of 0.91 and 0.86 respectively), whereas the F-measures reported in the

Sczyrba study are comparatively low (median value 0.31) -- implying that the latter was a much more challenging benchmark.

A further confounding factor is the conflicting rankings between evaluations. For example, by considering median F-measures we found the following: Bazinet *et al.* report that MEGAN (F=0.8) outperforms MetaPhyler (F=0.01); Lindgreen *et al.* report that CLARK (F=0.98) outperforms MEGAN (F=0.70) which outperforms MetaPhyler (F=0.01); and McIntyre *et al.* report that CLARK (F=0.93) outperforms MEGAN (F=0.87); all of these are consistent measurements. However, Peabody *et al.* report that MetaPhyler (F=0.44) outperforms CLARK (F=0.28) and Sczyrba *et al.* report that MetaPhyler (F=0.33) outperforms CLARK (F=0.20) which in-turn outperforms MEGAN (F=0.06) (see Figure 5B). The different software versions do not explain this discrepancy. Therefore our between-study comparison needs to correct for within-study characteristics, in order to fairly compare accuracy metrics for software tools. We have used two different approaches to address this.

Firstly, we use a non-parametric approach for comparing corrected accuracy measures. We converted each F-measure to a “robust Z-score” (see Methods). A median Z-score was then computed for each software tool, and used to rank all the different methods. A 95% confidence interval was also computed for each median Z-score, using a bootstrapping procedure. The results are illustrated in Figure 4B.

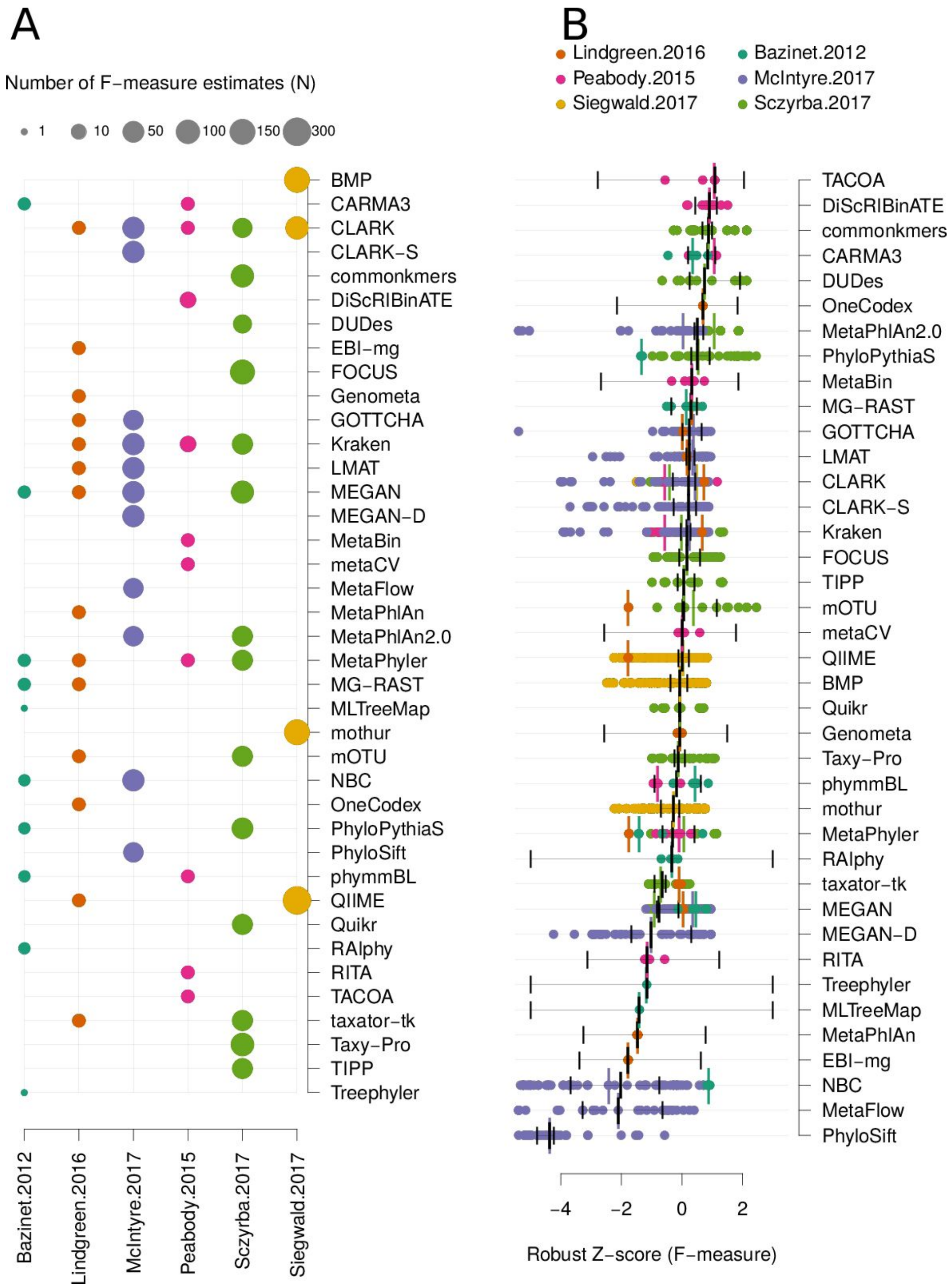


Figure 4: A: a matrix indicating metagenome analysis tools in alphabetical order (named on the right axis) versus a published benchmark on the bottom axis. The circle size is proportional to the median F-measure value for each benchmark. **B:** a ranked list of eDNA analysis tools. The median F-measure for each tool is indicated with a thick black vertical line. Bootstrapping each distribution (seeded with the extremes from the interval) 1000 times, was used to determine a 95% confidence interval for each median. These are indicated with thin vertical black lines. Each F-measure for each tool is indicated with a coloured point, colour indicates the manuscript where the value came from. Coloured vertical lines indicate the median F-measure for each benchmark for each tool.

The second approach we have used is a network meta-analysis to compare the different results. This approach is becoming widely used in the medical literature, predominantly as a means to compare estimates of drug efficacy across multiple studies, different cohorts, sample sizes and experimental designs [51–55]. This approach can incorporate both direct and indirect effects, stemming from a diverse intersecting sets of evidence. This means that indirect comparisons can be used to rank treatments (or software accuracies) even when a direct comparison has not been made.

We have used the “netmeta” software utility (implemented in R)[56] to investigate the relative performance, of each of the 39 software tools for which we have data, using the F-measure as a proxy for accuracy. A random- effects model and a rank-based approach were used for assessing the relative accuracy of different software tools. The resulting forest plot is shown in Figure 5A.

The two distinct approaches for comparing the accuracies from diverse software evaluation datasets resulted in remarkably consistent software rankings. Methods such as CARMA3, DiScRIBinATE, MetaPhlan2.0, OneCodex and TACOA are consistently ranked highly with both approaches. While this result may prove to be due to the small numbers of direct comparisons between these methods, it is suggestive that they may generally perform well across assessment conditions.

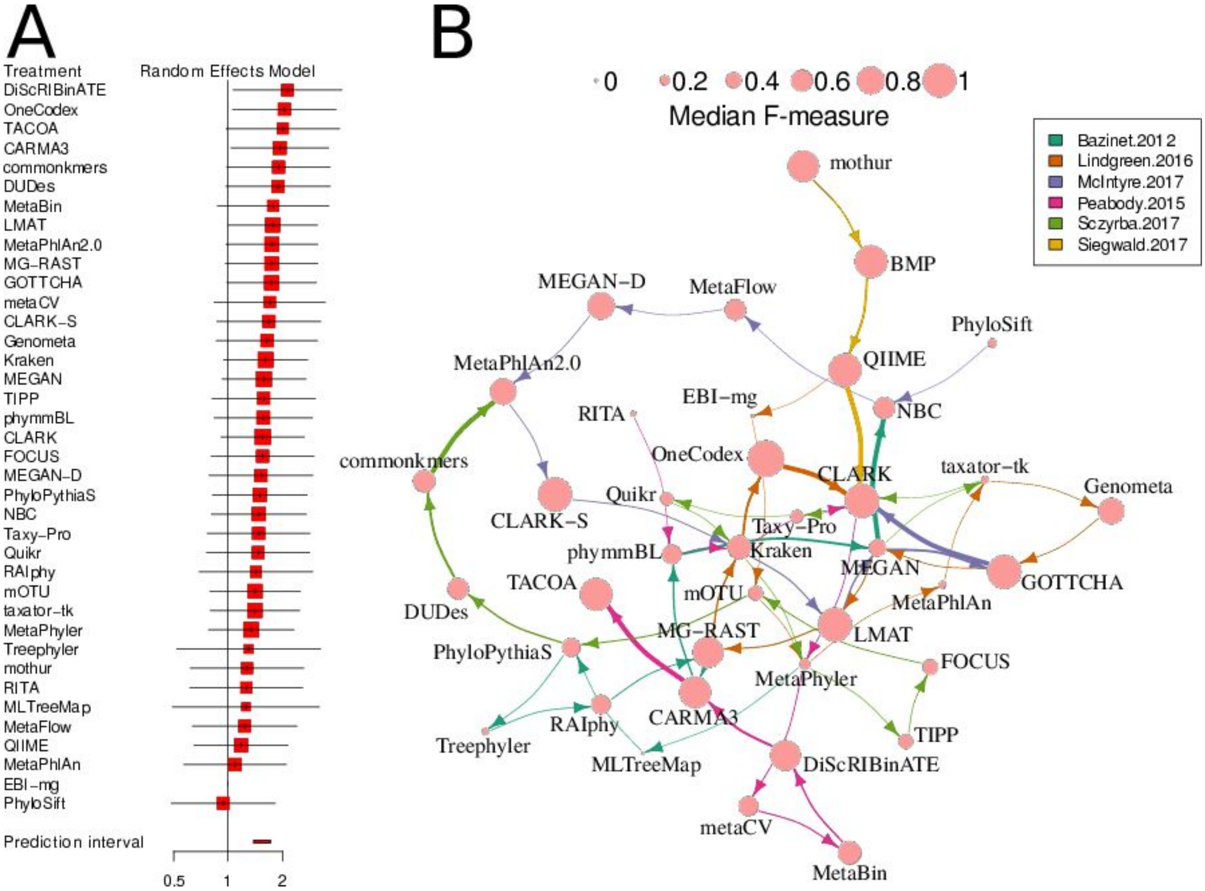


Figure 5: A. A forest plot of a network analysis, indicating the estimated accuracy range for each method. The plot shows the relative F-measure with a 95% confidence interval for each software tool. The tools are sorted based upon relative performance, from high to low. **B.** A network representation of the software tools, published evaluations, ranks for each tool and the median F-measure. The edge-widths indicate the rank of a method within a publication (based upon median, within-publication, rank). The edge-colours indicate the different publications, and node-sizes indicate median F-measure (based upon all publications). An edge is drawn between methods that are ranked consecutively within a publication. For example, CLARK and GOTTCHA were the top two methods from the McIntyre *et al.* (2017) evaluation, therefore a thick purple arrow connects these tools. Each has a high median F-measure and is therefore represented by a large node. CLARK has been evaluated multiple times and therefore has many incoming edges, whereas CLARK-S has only been evaluated once and therefore has a single incoming edge.

Remarkably, these five methods do not fall into a single class of algorithm outlined in Figure 1. CARMA3 [38] and DiScRIBinATE [57] are both BLASTX-based methods which use a lowest-common ancestor (LCA) methodology to determine the likely taxonomic placement of reads. CARMA3 uses additional reciprocal BLASTP searches to ensure orthology assignment, and has additional options for using HMMER3 searches. Both use the NCBI nr database for sourcing sequences with “known” taxonomy assignments. In contrast, MetaPhlan2.0[58] assigns taxonomy using a curated database of clade-specific marker gene sequences. Sequence reads are mapped to the database using the fast read mapper, Bowtie 2 [59]. OneCodex [60] finds sequence“K-mers” (sub-sequences of length *K*) that are unique to specific taxa, then combines these K-mers in a weighted taxonomic root-to-leaf algorithm to identify the most specific taxon for a given read; this approach is similar to the approaches used by Kraken [61] and CLARK [62]. Finally, TACOA [63] uses a composition-based methodology to identify likely taxa. This approach considers all subsequences, and calculates the probability of each using the nucleotide frequencies.The ratio between observed subsequence frequencies and expected frequencies is used to train a method for binning reads into likely taxonomic groups.

Conclusions

The analysis of environmental sequencing data remains a challenging task despite many years of research and many software tools for assisting with this task. In order to identify accurate methods for addressing this problem a number of benchmarking studies have been published. However, these studies have not shown a consistent or clearly optimal approach. We have reviewed and evaluated the existing published benchmarks using a network meta-analysis approach and have identified a small number of methods that are consistently predicted to perform well. Our aim here is to make non-arbitrary software recommendations that are based upon robust criteria rather than how widely-adopted a method is or the reputation of software developers, which are common proxies for how accurate a software tool is for eDNA analyses.

The high-performing methods do not fall into a consistent methodological group (e.g. alignment-based, K-mer, composition or hybrid methods). They include BLASTP-based methods (CARMA3 and DiScRIBinATE), a K-mer method (One Codex), specialised search database with a fast read mapper (MetaPhlan2.0) and a sequence composition-based approach (TACOA).

These results can by no means be considered the definitive answer to how to analyse eDNA datasets since methods will continue to be refined and results are based on broad averages over multiple conditions. Therefore, other methods may be more suited for specific problems. Furthermore, we have not addressed the issue of scale -- i.e. do these tools have sufficient speed to operate on the increasingly large-scale datasets that new sequencing methods are capable of producing.

Our analysis has not identified an underlying cause for the inconsistency between benchmarks. We found a core set of software tools that have been evaluated in most benchmarks. These are CLARK, Kraken, MEGAN, and MetaPhyler, but the relative ranking of these methods is quite different between benchmarks (Figures 4&5). We have not found an obvious mechanism for this, however, due to potential confounding factors such as benchmarking methodology, metrics, or software version. Some follow-up analysis of the Peabody *et al*. (2015) study revealed discrepancies between taxonomies (*Anabaena variabilis* and *Trichormus variabilis* are synonymous) [64], but this does not explain the performance estimates at higher taxonomic levels, which should be identical from family to phylum. Nevertheless, this suggests that robust evaluations should avoid using small numbers of taxa (the Peabody *et al.* evaluation only included 11 species), should evaluate at higher taxonomic levels and should be wary of synonym usage between taxonomies.

"Spike-ins" are a method of quality control widely used for gene expression analysis[65]. In this method, known amounts of DNA from known sources that are unlikely to also be in the environmental sample are added to a eDNA sample. Spike-ins can also include "negative" samples (i.e. sequences that are unlikely to be found in sequence databases), the nullomers from the McIntyre *et al.* study could be very useful for this. When the analysis pipeline does not return a result that is concordant with the spike-in proportions, this indicates a probable error. Spike-in controls can be used both during sample preparation and during sequence analysis. We believe that widespread inclusion of these types of controls by the metagenomics community will facilitate both improvement and increased confidence in the resulting taxonomic assignments.

Methods

Literature search: In order to identify benchmarks of metagenomic and amplicon software methods, an initial list of publications was curated. Further literature searches and trawling of citation databases (chiefly GoogleScholar) identified a relatively comprehensive list of six evaluations (Table 1), in which "F-measures" were either directly reported, or could be computed from supplementary results.

A list of published eDNA software analysis tools was curated manually. This made use of a community-driven project led by Jonathan Jacobs. The citation statistics for each publication were manually collected from GoogleScholar (on 7th July 2017). These values were used to generate Figure 3.

Data extraction: Accuracy metrics were collected from published datasets using a mixture of manual collection from supplementary materials and automated harvesting of data from online repositories. The data, scripts and results are available from: https://github.com/UCanCompBio/meta-analysis-eDNA-software

Data analysis: Each benchmark manuscript reports one or more F-measures for each software method. Due to the high variance of F-measures between studies (see Figure 3 and Supplementary Figure 2 for a comparison), we renormalised the F-measures using the following formula:

$$Robust\ Z\ score\ =\ \frac{x_i - median(X)}{mad(X)}$$

Where the "*mad*" function is the median absolute deviation, "*X*" is a vector containing all the F-measures for a publication and "x_i" is an individual F-measure for a particular software tool. Robust Z-scores can then be combined to provide an overall ranking of methods that is independent of methodological differences between studies (Figure 4). The 95% confidence intervals for median robust Z-scores illustrated in Figure 4 were generated using 1,000 bootstrap re-samplings from the distribution of values for each method.

Network metanalysis was used to provide a second evaluation method that accounts for differences between studies. We used the "netmeta" and "meta" software packages to perform the analysis. As outlined in Chapter 8 of the textbook "Meta-Analysis with R",[66] the ‘metacont’ function with Hedges' G was used to standardise mean differences and estimate fixed and random effects for each method within each benchmark. The ‘netmeta’ function was then used to conduct a pairwise meta-analysis of treatments (methods) across studies. This is based on a graph-theoretical analysis that has been shown to be equivalent to a frequentists network meta-analysis [67]. The ‘forest’ function was used on the resulting values to generate Figure 5A.

Acknowledgements

We are grateful to Jonathan Jacobs (@bioinformer) for maintaining a freely accessible list of eDNA analysis methods. We thank the authors of the benchmark studies used for this work, in particular Adam Bazinet, Christopher E. Mason, Alexa McIntyre, Fiona Brinkman, Alice McHardy, Alexander Sczyrba, David Koslicki and Léa Siegwald for discussions and feedback on some early results. We are grateful to Anne-Laure Boulesteix for valuable discussions.

This work was the outcome of a workshop funded as part of the Biological Heritage National Science Challenge, New Zealand. PPG is funded by the Bioprotection Research Centre, the National Science Challenge “NZ’s Biological Heritage,” and a Rutherford Discovery Fellowship.

References

1. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13,** 260–270 (2012).
2. Baird, D. J. & Hajibabaei, M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol. Ecol.* **21,** 2039–2044 (2012).
3. Bohan, D. A. *et al.* Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. *Trends Ecol. Evol.* **32,** 477–487 (2017).
4. Woese, C. R. Bacterial evolution. *Microbiol. Rev.* **51,** 221–271 (1987).
5. Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4576–4579 (1990).
6. Hugenholtz, P. & Pace, N. R. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.* **14,** 190–197 (1996).
7. Tringe, S. G. & Hugenholtz, P. A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* **11,** 442–446 (2008).
8. Brown, J. W. *et al.* Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 3001–3006 (1996).
9. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 6241–6246 (2012).
10. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523,** 208–211 (2015).

11. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308,** 554–557 (2005).

12. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14,** R51 (2013).

13. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* **2,** 3 (2012).

14. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **5,** 209 (2014).

15. Oulas, A. *et al.* Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights* **9,** 75–88 (2015).

16. Sneath, A. & Sokal, R. R. Principles of numerical taxonomy. *San Francisco and London I* **963,** (1963).

17. Wommack, K. E., Bhavsar, J. & Ravel, J. Metagenomics: read length matters. *Appl. Environ. Microbiol.* **74,** 1453–1463 (2008).

18. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14,** 157–167 (2013).

19. Magoc, T. *et al.* GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* **29,** 1718–1725 (2013).

20. Zhang, Y., Sun, Y. & Cole, J. R. A scalable and accurate targeted gene assembly tool (SAT-Assembler) for next-generation sequencing data. *PLoS Comput. Biol.* **10,** e1003737 (2014).

21. Wang, Q. *et al.* Xander: employing a novel method for efficient gene-targeted metagenomic assembly. *Microbiome* **3,** 32 (2015).

22. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).

23. Gregor, I., Dröge, J., Schirmer, M., Quince, C. & McHardy, A. C. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* **4,** e1603 (2016).

24. Gonzalez, A. *et al.* Avoiding Pandemic Fears in the Subway and Conquering the Platypus. *mSystems* **1,** (2016).

25. Wu, D. *et al.* A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462,** 1056–1060 (2009).

26. Bazinet, A. L. & Cummings, M. P. A comparative evaluation of sequence classification programs. *BMC Bioinformatics* **13,** 92 (2012).

27. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. L. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* **16,** 363 (2015).

28. Lindgreen, S., Adair, K. L. & Gardner, P. P. An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* **6,** 19233 (2016).

29. McIntyre, A., Ounit, R., Afshinnekoo, E. & Prill, R. Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers. *bioRxiv* (2017).

30. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation- a benchmark of computational metagenomics software. bioRxiv: 099127. *BioRxiv* (2017). doi:https://doi.org/10.1101/099127

31. Siegwald, L. *et al.* Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. *PLoS One* **12,** e0169563 (2017).

32. Boulesteix, A.-L., Lauer, S. & Eugster, M. J. A. A plea for neutral comparison studies in computational sciences. *PLoS One* **8,** e61562 (2013).

33. Boulesteix, A.-L. Over-optimism in bioinformatics research. *Bioinformatics* **26,** 437–439 (2010).

34. Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K. & Boulesteix, A.-L.

Over-optimism in bioinformatics: an illustration. *Bioinformatics* **26,** 1990–1998 (2010).

35. Norel, R., Rice, J. J. & Stolovitzky, G. The self-assessment trap: can we all be better than average? *Mol. Syst. Biol.* **7,** 537 (2011).

36. Stranneheim, H. *et al.* Classification of DNA sequences using Bloom filters. *Bioinformatics* **26,** 1595–1600 (2010).

37. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 95–100 (2010).

38. Gerlach, W. & Stoye, J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.* **39,** e91 (2011).

39. Patil, K. R. *et al.* Taxonomic metagenome sequence assignment with structured output models. *Nat. Methods* **8,** 191–192 (2011).

40. McIntyre, A. B. R. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18,** 182 (2017).

41. Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson, D. H. MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One* **3,** e3373 (2008).

42. Boulesteix, A.-L., Wilson, R. & Hapfelmeier, A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Med. Res. Methodol.* **17,** 138 (2017).

43. Singer, E. *et al.* High-resolution phylogenetic microbial community profiling. *ISME J.* **10,** 2020–2032 (2016).

44. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* **7,** e39315 (2012).

45. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* **40,** e94 (2012).

46. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).

47. Caboche, S., Audebert, C., Lemoine, Y. & Hot, D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics* **15,** 264 (2014).

48. Lever, J., Krzywinski, M. & Altman, N. Points of Significance: Classification evaluation. *Nat. Methods* **13,** 603–604 (2016).

49. Metagenomics - Tools, Methods and Madness. *Google Docs* Available at: https://docs.google.com/document/d/1qLczhk4MAKjkOtz-PnXhmgGEWnWQJHLf0Mjd1B9U2RU/edit. (Accessed: 21st August 2017)

50. Gardner, P. P. *et al.* A meta-analysis of bioinformatics software benchmarks reveals that publication-bias unduly influences software accuracy. *bioRxiv* 092205 (2017). doi:10.1101/092205

51. Lumley, T. Network meta-analysis for indirect treatment comparisons. *Stat. Med.* **21,** 2313–2324 (2002).

52. Lu, G. & Ades, A. E. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat. Med.* **23,** 3105–3124 (2004).

53. Salanti, G., Higgins, J. P. T., Ades, A. E. & Ioannidis, J. P. A. Evaluation of networks of randomized trials. *Stat. Methods Med. Res.* **17,** 279–301 (2008).

54. Greco, T. *et al.* The attractiveness of network meta-analysis: a comprehensive systematic and narrative review. *Heart Lung Vessel* **7,** 133–142 (2015).

55. Higgins, J. P. T. *et al.* Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Res Synth Methods* **3,** 98–110 (2012).

56. Rücker, G., Schwarzer, G., Krahn, U. & König, J. netmeta: Network meta-analysis using frequentist methods. *R package version 0. 8-0. Available at)(Accessed December 1, 2016)* (2015).

57. Ghosh, T. S., Monzoorul Haque, M. & Mande, S. S. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics* **11 Suppl 7,** S14 (2010).

58. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (2015).

59. Liu, Y. & Schmidt, B. Long read alignment based on maximal exact match seeds. *Bioinformatics* **28,** i318–i324 (2012).

60. Minot, S. S., Krumm, N. & Greenfield, N. B. One codex: a sensitive and accurate data platform for genomic microbial identification. *bioRxiv* (2015).

61. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).

62. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16,** 236 (2015).

63. Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K. & Nattkemper, T. W. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* **10,** 56 (2009).

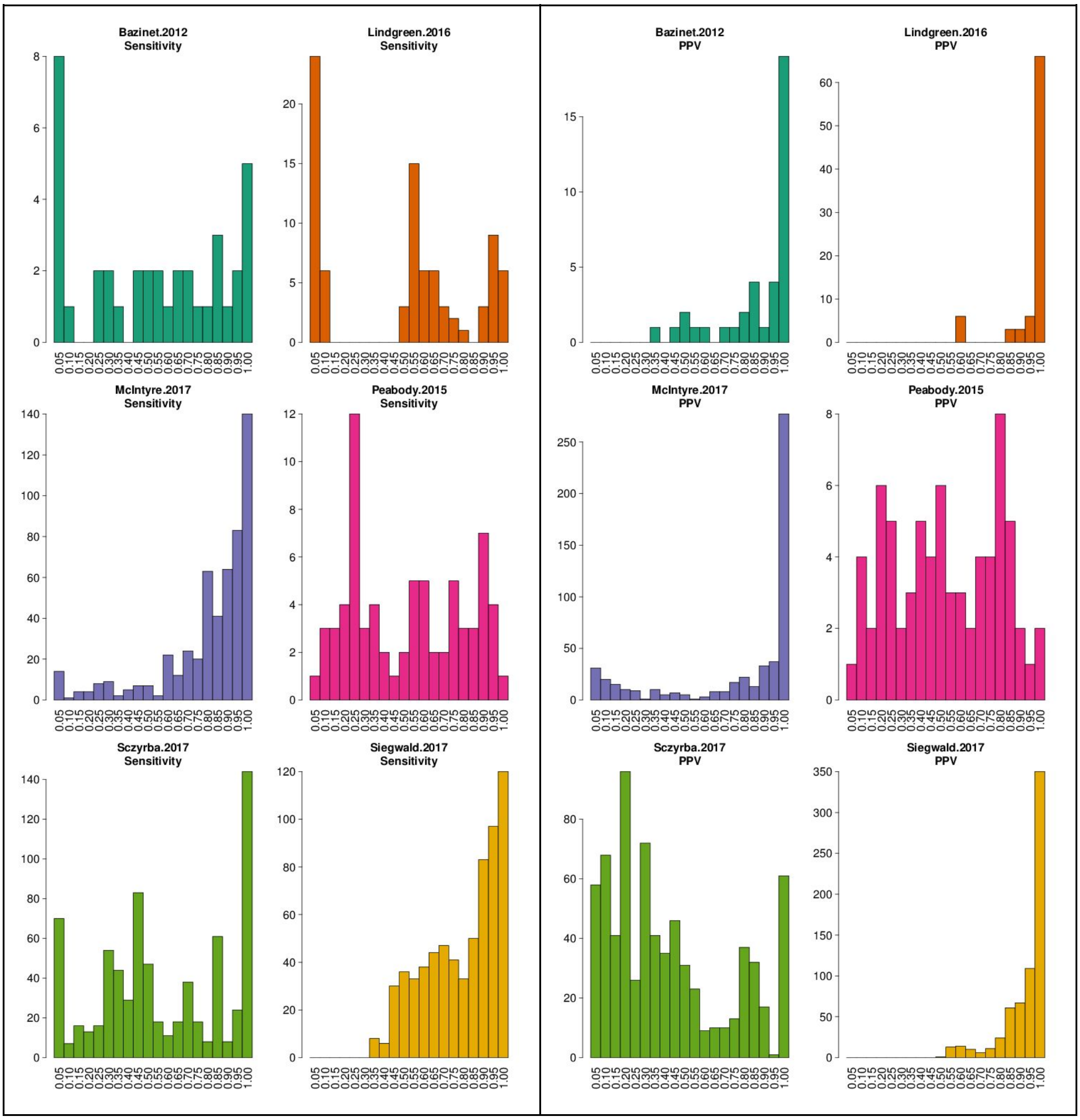
64. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3,** e104 (2017).

65. Yang, I. V. [4] Use of External Controls in Microarray Experiments. *Methods Enzymol.* **411,** 50–63 (2006).

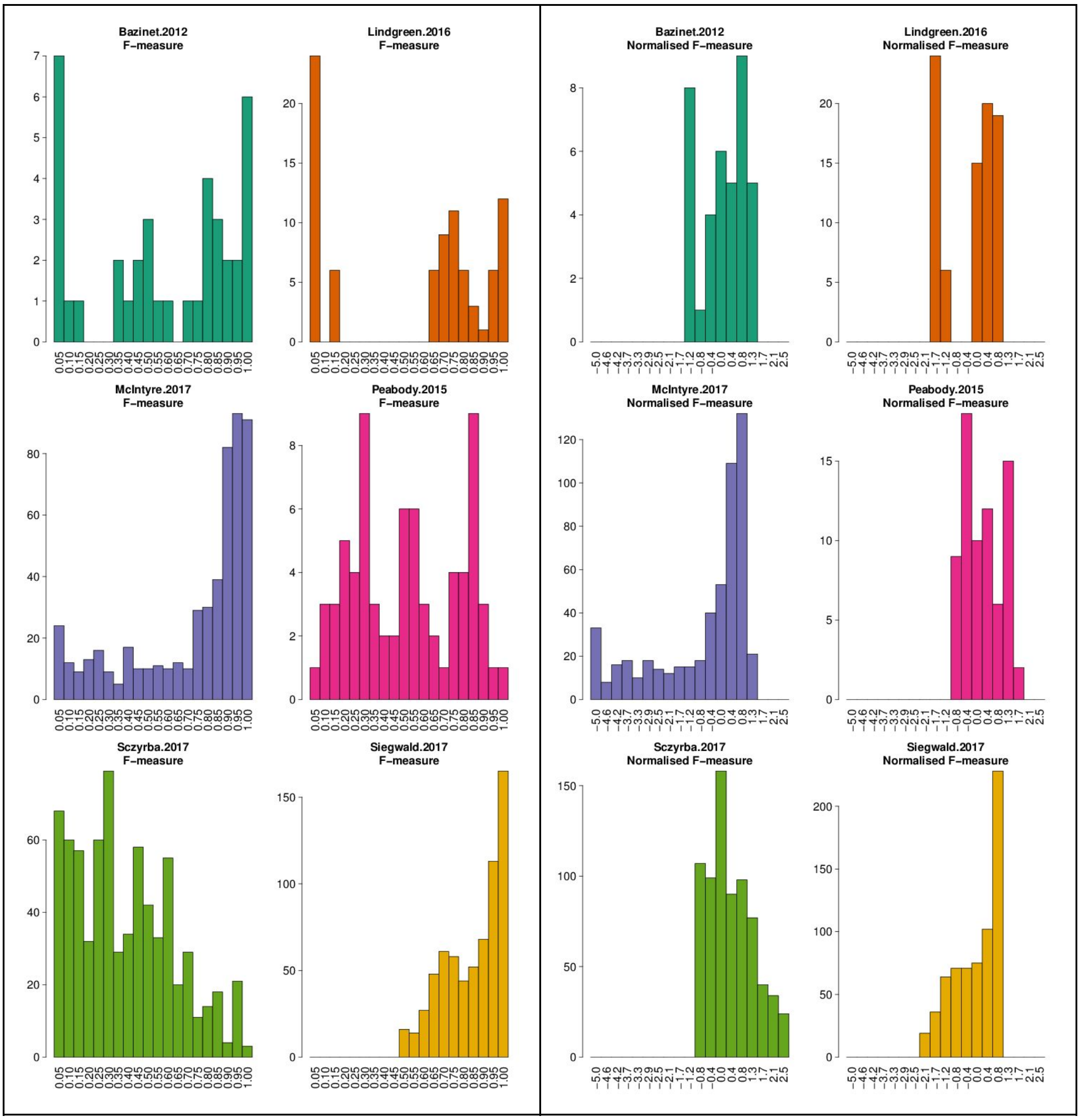
66. Schwarzer, G., Carpenter, J. R. & Rücker, G. *Meta-Analysis with R*. (Springer, 2015).

67. Rücker, G. Network meta-analysis, electrical networks and graph theory. *Res Synth Methods* **3,** 312–324 (2012).

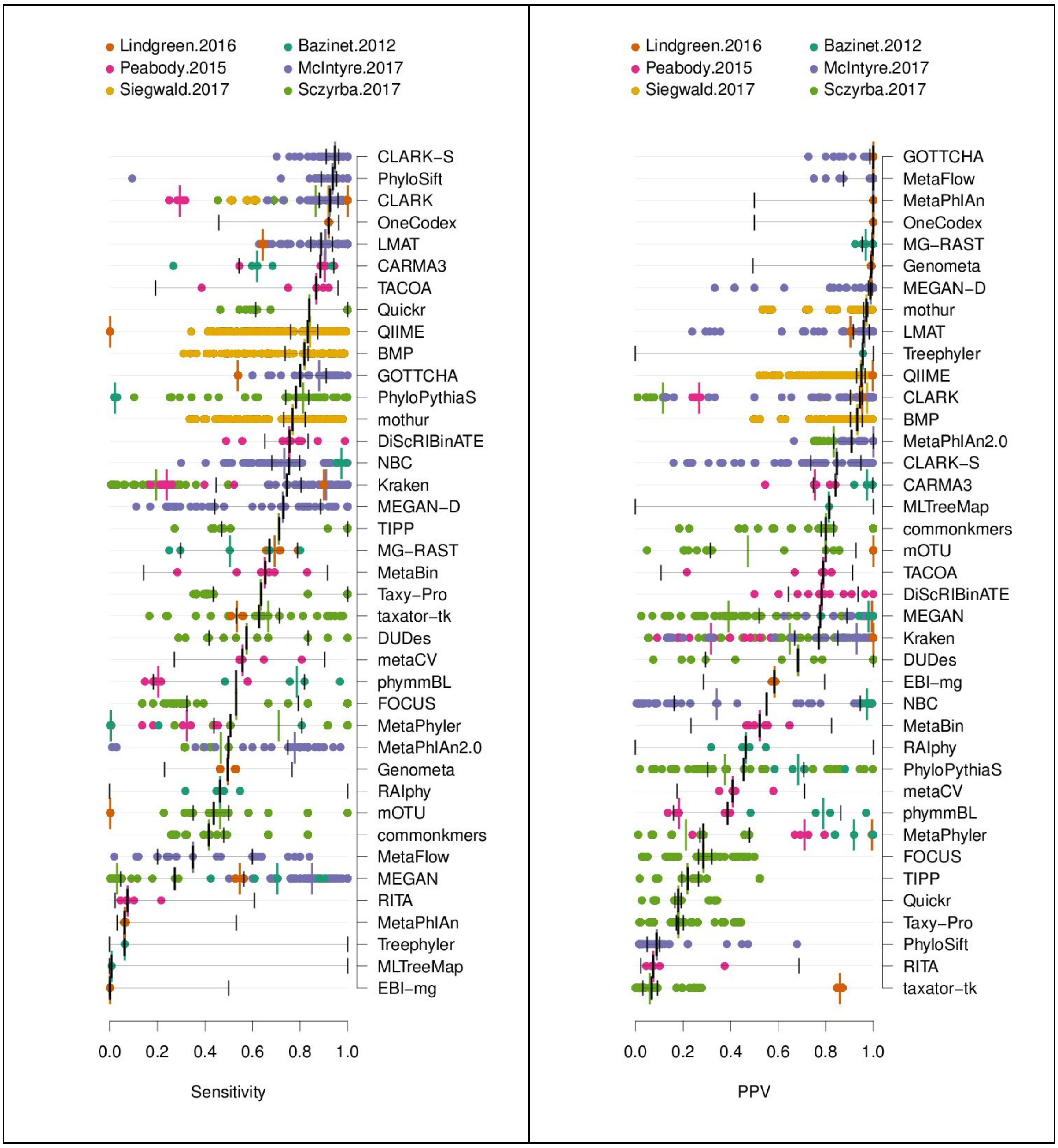
Supplementary Results



Supplementary Figure 1: The distribution of Sensitivity and PPV estimates for each of the six benchmark publications.



Supplementary Figure 2: A. The distributions of F-measure estimates for each of the six benchmark publications. B. The distributions of robust Z-scores for F-measure estimates for each of the six benchmark publications.



Supplementary Figure 3: Ranked lists of eDNA analysis tools, based upon median Sensitivity and PPV measures. Coloured points indicate and estimated accuracy measure from one of six benchmark publications. Median values are indicated by a vertical bar (black for the overall median value, coloured bars for the median value from a publication). Bootstrap derived 95% confidence intervals for the Sensitivity or PPV are indicated with a thin black lines for each method.