Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy 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 tools for analysing metagenomic and amplicon datasets, which makes selecting adequate tools a significant challenge. A number of benchmark studies have been undertaken; however, these can present conflicting results. We have applied a robust Z-score ranking procedure and a network meta-analysis method to identify software tools that are generally accurate for mapping DNA sequences to taxonomic hierarchies. Based upon these results we have identified some tools and computational strategies that produce robust predictions.

Introduction

Metagenomics, meta-barcoding and related high-throughput environmental DNA (eDNA) or microbiome 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]. The classification of DNA sequences can be broadly divided into amplicon (barcoding) and genome-wide (metagenome) approaches. The amplicon, or barcoding, -based approaches 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].

High-throughput sequencing (HTS) results can be analysed using a number of different strategies (Supplementary Figure 1) [13–16]. The fundamental goal of many of these studies is to assign taxonomy to sequences as specifically as possible, and in some cases to cluster highly-similar sequences into “operational taxonomic units” (OTUs) [17]. For greater accuracy in taxonomic assignment, metagenome and amplicon sequences may be assembled into longer “contigs” using any of the available sequence assembly tools [18,19]. The reference-based methods (also called “targeted gene assembly”) make use of conserved sequences to constrain sequence assemblies. These have a number of reported advantages including reducing chimeric sequences, and improving the speed and accuracy of assembly relative to *de novo* methods [20–23].

Environmental DNA sequences are generally mapped to a reference database of sequences labelled with a hierarchical taxonomic classification. The level of divergence, distribution and coverage of mapped taxonomic assignments allows an estimate to be made of where the sequence belongs in the established taxonomy . This is commonly performed using the lowest common ancestor approach (LCA) [24]. Some tools, however, avoid this computationally-intensive sequence similarity estimation, and instead use alignment-free approaches based upon sequence composition statistics (e.g. nucleotide or k-mer frequencies) to estimate taxonomic relationships [25].

In this study we identified seven published evaluations, of tools that estimate taxonomic origin from DNA sequences [26–32]. Of these, four evaluations met our criteria for a neutral comparison study [33] (see Supplementary Table 1). These are summarised in Table 1 [26,28,30,32] and include accuracy estimates for 25 environmental DNA classification tools. We have used network meta-analysis techniques and non-parametric tests to variable and sometimes conflicting reports from the different evaluation studies, resulting in a short list of methods that have been consistently reported to produce accurate interpretations of metagenomics results. This study reports one of the first meta-analyses of neutral comparison studies, fulfilling the requirement for an apex study in the evidence pyramid for benchmarking [34].

Overview of environmental DNA classification 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 benchmarks 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 [33]. The criteria 1 and 2 are straightforward to determine, but criteria 3 is the more difficult to evaluate as it includes identifying challenging datasets, and appropriate metrics for accurate accuracy reporting [35–37]. Based upon literature reviews and citation analyses, we have identified seven published evaluations of environmental DNA analysis, we have assessed these against the above three principles and four of these studies meet the inclusion criteria (assessed in Supplementary Table 1) [26,28,30,32]. These studies are summarised in Table 1.

In the following sections we discuss issues with collecting trusted datasets, including the selection of positive and negative control data that avoid 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 controls**	**Negative controls**	**Reference exclusion method**	**Metrics**
Almeida *et al.* (2018) [32]	12 *in silico* mock communities from 208 different genera.	-	2% of positions “randomly mutated”	Sequence Level (Sen., F-measure)
Bazinet *et al.* (2012) [26]	Four published *in silico* mock communities from 742 taxa [38–41]	-	-	Sequence Level (Sen., PPV)
Lindgreen *et al.* (2016) [28]	Six *in silico* mock communities from 417 different genera.	Shuffled sequences	Simulated evolution	Sequence Level (Sen., Spec., PPV, NPV, MCC)
Siegwald *et al.* (2017) [30]	36 *in silico* mock communities from 125 bacterial genomes.	-	-	Sequence Level (Sen, PPV, F-measure)

Table 1: A summary of the main features of the four software evaluations used for this study, including the positive controls employed (the sources of sequences from organisms with known taxonomic placements, whether negative control sequences were used, the approaches for excluding reference sequences from the positive control sequences, and the metrics that were collected for tool evaluation. The accuracy measures are defined in Table 2, the abbreviations used above are Matthews Correlation Coefficient (MCC), Negative Predictive Value (NPV), Positive Predictive Value (PPV), Sensitivity (Sen), Specificity (Spec).

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 independent of past training datasets, reliable, well-curated, robust and representative of the large population of all possible datasets [34]. **Positive control** datasets can be divided into two different strategies, namely the *in vitro* and *in silico* approaches for generating mock communities.

In vitro methods involve generating microbial consortia in predetermined ratios of microbial strains, extracting the consortium DNA, sequencing and analysing these using standard metagenomics pipelines [42,43]. Non-reference sequences can also be included to this mix as a form of negative control. The accuracy of the genome assembly, genome partitioning (binning) and read depth proportional to consortium makeup can then be used to confirm software accuracy. In principle, every metagenome experiment could employ *in vitro* positive and negative controls by “spiking” known amounts of DNA from known sources, as has been widely used for gene expression analysis [44] and increasingly for metagenomics [45–47].

In silico methods use selected publicly-available genome sequences. Simulated metagenome sequences are can be derived from these [48–51]. It is important to note that ideally-simulated sequences are derived from species that are **not present** in established reference databases, as this is a more realistic simulation of most environmental DNA surveys. A number of different strategies have been used to control for this [27,28,31]. 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 sequences of varying evolutionary distances from reference sequences [28], similarly Almeida *et al* simulated random mutations for 2% of nucleotides in each sequence [32]. Sczyrba *et al.* restricted their analysis to sequences sampled from recently-deposited genomes, increasing the chance that these are not included in any reference databases [31]. These strategies are illustrated in Figure 1.

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

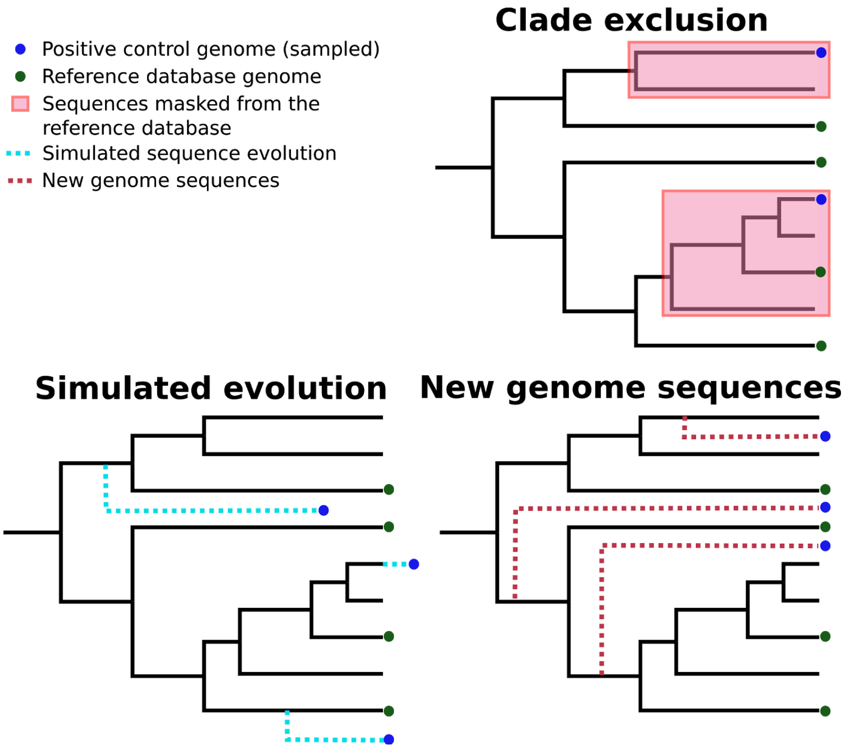

Clade exclusion
● Positive control genome (sampled)
● Reference database genome
▨ Sequences masked from the reference database
Simulated sequence evolution
New genome sequences
Simulated evolution
New genome sequences

Figure 1: 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. These are: “clade exclusion”, where positive control sequences are selectively removed from reference databases [27]; “simulated evolution”, where models of sequence evolution are used to generate sequences of defined divergence times from any ancestral sequence or location on a phylogenetic tree e.g. [52,53]; and “new genome

sequences” are genome sequences that have been deposited in sequence archives prior to the generation of any reference sequence database used by analysis tools [31].

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 a good true positive rate may be sacrificed in exchange for a lowering the 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. There are two main ways this has been approached, namely per-sequence assignment and per-taxon assignment. Estimates of per-sequence accuracy values can be made by determining whether individual sequences were correctly assigned to a particular taxonomic rank [27,28,30]. Alternatively, per-taxon accuracies can be determined by comparing reference and predicted taxonomic distributions [31]. The per-taxon approach may lead to erroneous accuracy estimates as sequences may be incorrectly assigned to included taxa. Cyclic-errors can then cancel, leading to inflated accuracy estimates. However, per-sequence information can be problematic to extract from tools that only report profiles.

Successfully recapturing the frequencies of different taxonomic groups as a **measure of community diversity** is a major aim for environmental DNA analysis projects. There have been a variety of approaches for quantifying the accuracy of this information. Pearson’s correlation coefficient [26], L1-norm [31], the sum of absolute log-ratios [28], the log-modulus [29] and the Chao 1 error [30] 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 levels and whether sequences 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 four benchmarks we are considering in this work (Figure 2).

$Sensitivity = \frac{TP}{TP+FN}$ (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 dependent 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 [54].

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 Google Scholar) identified a comprehensive list of seven evaluations (Table 1), in which “F-measures” were either directly reported, or could be computed from supplementary materials. These seven studies were then evaluated against the three principles of benchmarking [33], with four studies meeting all three principles and were included in the subsequent analyses (see Supplementary Table 1 for details).

A list of published environmental DNA classification software was curated manually. This made use of a community-driven project led by Jonathan Jacobs [55]. The citation statistics for each software publication were manually collected from Google Scholar (in July 2017). These values were used to generate Figure 2.

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. For a number of the benchmarks, a number of non-independent accuracy estimates were taken, for example different parameters, reference databases or taxonomic levels were used for the evaluations. We have combined all non-independent accuracy measurements using a median value. Leaving only single accuracy measures for each tool and benchmark dataset combination. The data, scripts and results are available from: https://github.com/Gardner-BinfLab/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 2C and Supplementary Figure 3 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 each 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 the methodological and data differences between studies (Figure 3). The 95% confidence intervals for median robust Z-scores shown in Figure 3 were generated using 1,000 bootstrap resamplings from the distribution of values for each method, extreme (F={0,1}) values seeded into each *X* in order to capture the full range of potential F-measures.

Network meta-analysis was used to provide a second 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", [56] 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 (tools) across studies. This is based on a graph-theoretical analysis that has been shown to be equivalent to a frequentists network meta-analysis [57]. The 'forest' function was used on the resulting values to generate Figure 4A.

Review of Results

We have mined independent estimates of sensitivity, positive predictive values (PPV) and F-measures for 25 environmental DNA classification tools, from three published software evaluations. A matrix showing presence-or-absence of software tools in each publication is illustrated in Figure 3A. Comparing the list of 25 environmental DNA classification tools to a publicly available list of environmental DNA classification tools based upon literature mining and crowd-sourcing, we found that 29% (25/88) of all published tools have been evaluated in the four of seven studies we have identified as neutral comparison studies (details in Supplementary Table 1) [33]. The unevaluated methods 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 2A). Several software tools have been very widely cited (Figure 2B), yet caution must be used when considering citation statistics, as the number of citations is not correlated with accuracy (Figure 2D) [28,58]. For example, the tools that are published early are more likely to be widely cited, or it may be that some articles are not necessarily cited for the software tool. For example, the MEGAN1 manuscript is often cited for one of the first implementations of the lowest-common-ancestor (LCA) algorithm for assigning read-similarities to taxonomy [24].

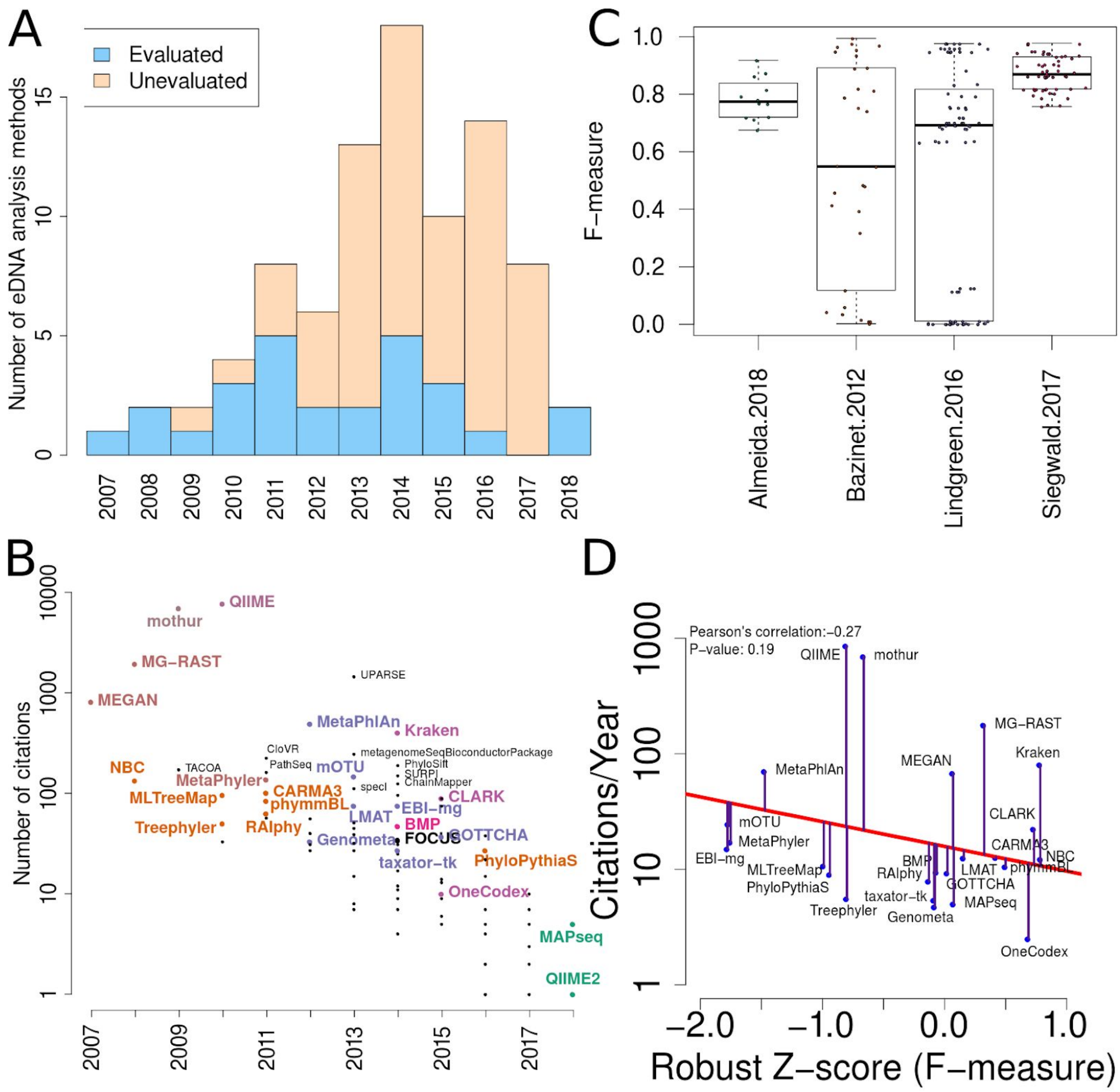


Figure 2: A: More than 80 environmental DNA classification tools have been published in the last 10 years [55]. A fraction of these (29%) have been independently evaluated. **B:** The number of citations for each software tool versus the year it was published. Software tools that have been evaluated 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:** Box-whisker plots illustrating the distributions of accuracy estimates based upon reported F-measures using values from 4 different evaluation manuscripts [26,28,30,32]. **D:** The relationship between publication citation counts and the corresponding tool accuracy estimate, as measured by a normalised F-measure (see Methods for details).

After manually extracting sensitivity, PPV and F-measures (or computing these) from the tables and/or supplementary materials for each publication [26,28,30,32], we have considered the within-publication distribution of accuracy measures (see Figure 2C, Supplementary Figure 2 & Supplementary Figure 3). These figures indicate that each publication has differences in F-measure distributions. These can be skewed and multimodal, and different measures of centrality and variance. Therefore, a correction needs to be used to account for between-benchmark variation.

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 computed for each software tool, and used to rank tools. A 95% confidence interval was also computed for each median Z-score using a bootstrapping procedure. The results are presented in Figure 3B (within-benchmark distributions are shown in Supplementary Figure 3A and B).

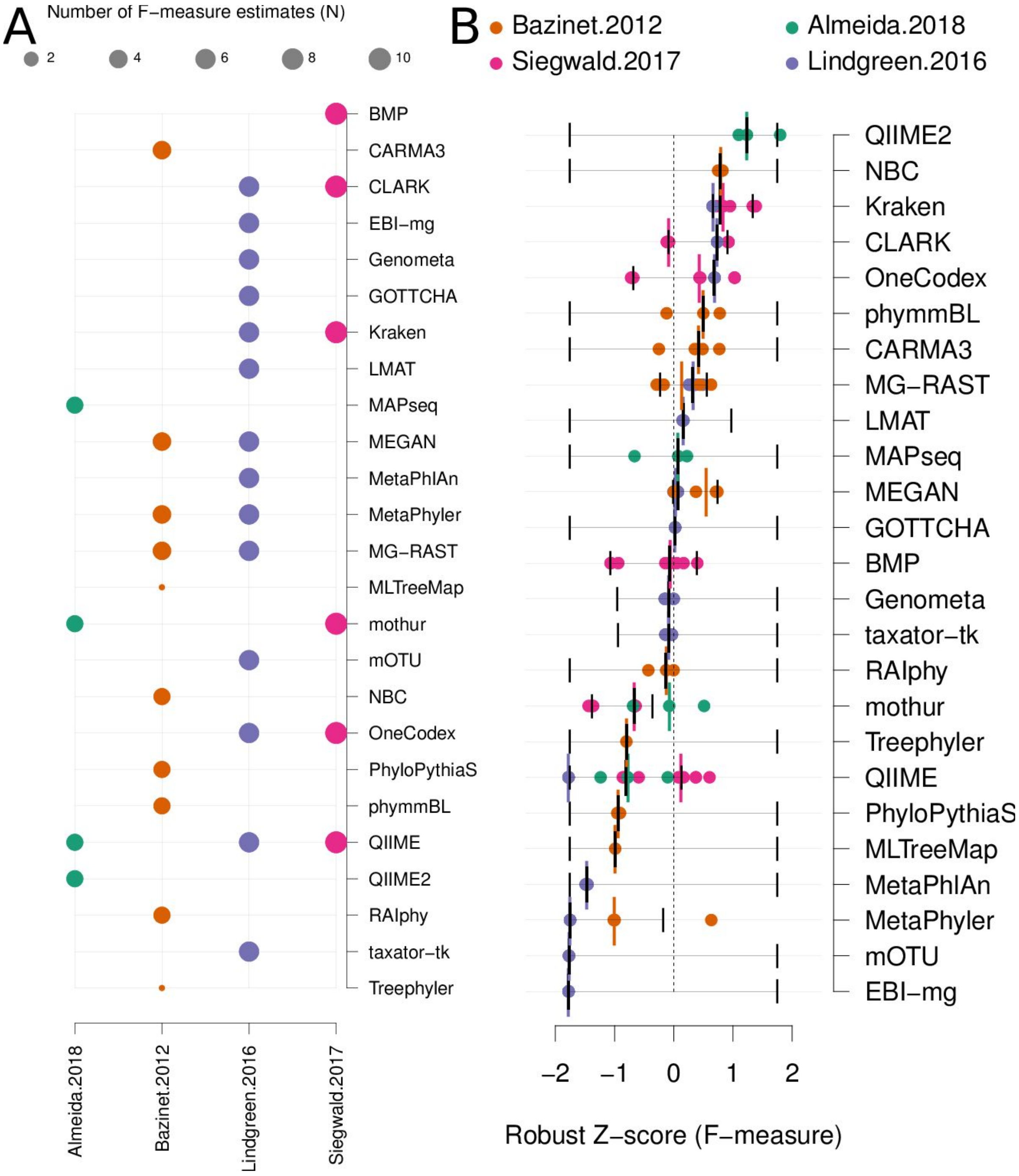


Figure 3: A: a matrix indicating metagenome analysis tools in alphabetical order (named on the right axis) versus a published benchmark on the bottom axis. The circle size is proportional to the number of F-measure estimates from each benchmark. **B:** a ranked list of environmental DNA classification 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 was sourced. 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 from multiple studies that include different cohorts, sample sizes and experimental designs [59–63]. This approach can incorporate both direct and indirect effects, and incorporates diverse intersecting sets of evidence. This means that indirect comparisons can be used to rank treatments (or software tool accuracy) even when a direct comparison has not been made.

We have used the “netmeta” software utility (implemented in R) [64] to investigate the relative performance of each of the 25 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 4A.

The two distinct approaches for comparing the accuracies from diverse software evaluation datasets resulted in remarkably consistent software rankings in this set of results. The Pearson’s correlation coefficient between robust Z-scores and network meta-analysis odds-ratios is 0.95 (P-value=$1.4\text{x}10^{-11}$), see Supplementary Figure 6.

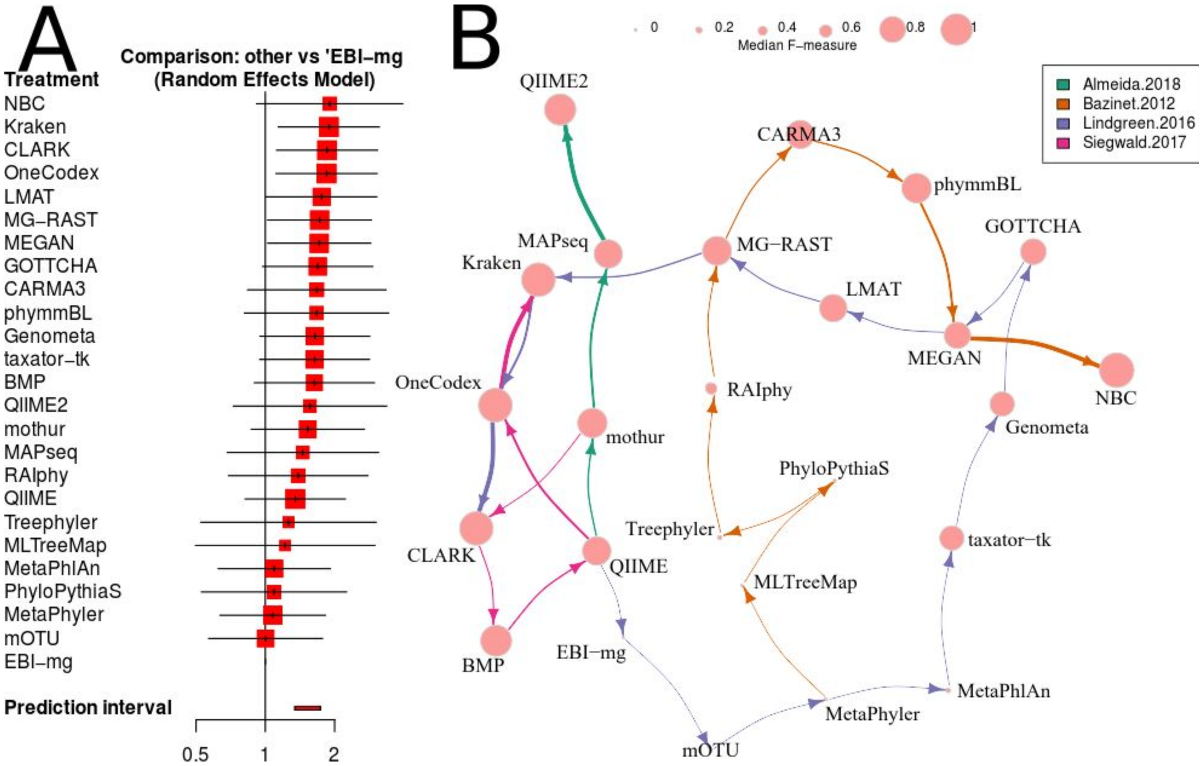


Figure 4: A. A forest plot of a network analysis, indicating the estimated accuracy range for each tool. The plot shows the relative F-measure with a 95% confidence interval for each software tool. The tools are sorted based upon relative performance, from high to low. Tools with significantly higher F-statistics have a 95% confidence interval that does not cover the null odds-ratio of 1.

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 tool 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 tools that are ranked consecutively within a publication.

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 [26–32]. 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 and a non-parametric approach. These methods have identified a small number of tools 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 tool is or the reputation of software developers, which are common proxies for how accurate a software tool is for environmental DNA analyses [58].

Based upon this meta-analysis, the k-mer based approaches, CLARK [65], Kraken [66] and One Codex [67] consistently rank well in both the non-parametric, robust Z-score evaluation and the network meta-analysis. The confidence intervals for both evaluations were comparatively small, so these estimates are likely to be reliable. In particular, the network meta-analysis analysis showed that these tools are significantly more accurate than the alternatives (i.e. the 95% confidence intervals exclude the the odds-ratio of 1).

There were also a number of widely-used tools, MG-RAST [68], MEGAN [69] and QIIME 2 [70] that are both comparatively user-friendly and have respectable accuracy (Z>0 and narrow confidence intervals, see Figure 3B and Supplementary Figure 5). However, the new QIIME 2 tool has only been evaluated in one benchmark [32], and so this result should be viewed with caution until further independent evaluations are undertaken. Therefore has a large confidence interval on the accuracy estimate based upon robust Z-scores (Figure 3) or ranked below high-performing methods with the network meta-analysis (Figure 4). The tools Genometa [71], GOTTCHA [72], LMAT [73], mothur [74] and taxator−tk [75], while not meeting the stringent accuracy thresholds we have used above were also consistently ranked well by both approaches.

The NBC tool [76] ranked highly in both the robust Z-score and network analysis, however the confidence intervals on both accuracy estimates were comparably large. Presumably, this was due to its inclusion in a single, early benchmark study [26] and exclusion from all subsequent benchmarks. To investigate this further, the authors of this study attempted to run NBC themselves, but found that it failed to run (core dump) on test input data. It is possible that with some debugging, this method could compare favourably with modern approaches.

These results can by no means be considered the definitive answer to how to analyse environmental DNA datasets since tools will continue to be refined and results are based on broad averages over multiple conditions. Therefore, some tools may be more suited for more specific problems than those assessed in these results (e.g. human gut microbiome). 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 inconsistencies 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 tools differed greatly between some benchmarks. We did find that restricting the included benchmarks to those that satisfy the criteria for a “neutral comparison study” [33], improved the consistency of evaluations considerably. This may point to differences in the results obtained by “expert users” (e.g. tool developers) compared and those of “amateur users” (e.g. bioinformaticians or microbiologists).

Finally, the results presented in Supplementary Figure 4 indicate that most metagenome analysis tools have a high positive-predictive value (PPV). This implies that false-positive matches between environmental DNA sequences and reference databases are not the main source of error for these analyses. However, sensitivity estimates can be low and generally cover a broad range of values. This implies that false-negatives are the main source of error for environmental analysis. This shows that matching divergent environmental DNA and reference database nucleotide sequences remains a significant research challenge in need of further development.

Acknowledgements

We are grateful to Jonathan Jacobs (@bioinformer) for maintaining a freely accessible list of environmental DNA analysis methods. We thank the authors of the benchmark studies used for this work, all of whom were offered an opportunity to comment on this work before it was released. 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 our early results. We are grateful to Anne-Laure Boulesteix for valuable suggestions on how to conduct the analyses.

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 MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13: 260–270.

2. Baird DJ, Hajibabaei M. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Mol Ecol. 2012;21: 2039–2044.

3. Bohan DA, Vacher C, Tamaddoni-Nezhad A, Raybould A, Dumbrell AJ, Woodward G. Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. Trends Ecol Evol. 2017;32: 477–487.

4. Woese CR. Bacterial evolution. Microbiol Rev. 1987;51: 221–271.

5. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87: 4576–4579.

6. Hugenholtz P, Pace NR. Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. Trends Biotechnol. 1996;14: 190–197.

7. Tringe SG, Hugenholtz P. A renaissance for the pioneering 16S rRNA gene. Curr Opin Microbiol. 2008;11: 442–446.

8. Brown JW, Nolan JM, Haas ES, Rubio MA, Major F, Pace NR. Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. Proc Natl Acad Sci U S A. 1996;93: 3001–3006.

9. Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A. 2012;109: 6241–6246.

10. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523: 208–211.

11. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. Science. 2005;308: 554–557.

12. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14: R51.

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

14. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci. 2014;5: 209.

15. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights. 2015;9: 75–88.

16. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35: 833–844.

17. Sneath A, Sokal RR. Principles of numerical taxonomy. San Francisco and London I. 1963;963.

18. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2017; doi:10.1093/bib/bbx120

19. Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform. 2017; doi:10.1093/bib/bbx098

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

21. Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, et al. Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome. 2015;3: 32.

22. Huson DH, Tappu R, Bazinet AL, Xie C, Cummings MP, Nieselt K, et al. Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. Microbiome. 2017;5: 11.

23. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27: 824–834.

24. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17: 377–386.

25. Gregor I, Dröge J, Schirmer M, Quince C, McHardy AC. PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. PeerJ. 2016;4: e1603.

26. Bazinet AL, Cummings MP. A comparative evaluation of sequence classification programs. BMC Bioinformatics. 2012;13: 92.

27. Peabody MA, Van Rossum T, Lo R, Brinkman FSL. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics. 2015;16: 363.

28. Lindgreen S, Adair KL, Gardner PP. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6: 19233.

29. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. biorxiv.org; 2017;18: 182.

30. Siegwald L, Touzet H, Lemoine Y, Hot D, Audebert C, Caboche S. Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics. PLoS One. 2017;12: e0169563.

31. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods. 2017;14: 1063–1071.

32. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience. 2018;7. doi:10.1093/gigascience/giy054

33. Boulesteix A-L, Lauer S, Eugster MJA. A plea for neutral comparison studies in computational sciences. PLoS One. 2013;8: e61562.

34. 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. 2017;17: 138.

35. Boulesteix A-L. Over-optimism in bioinformatics research. Bioinformatics. 2010;26: 437–439.

36. Jelizarow M, Guillemot V, Tenenhaus A, Strimmer K, Boulesteix A-L. Over-optimism in bioinformatics: an illustration. Bioinformatics. 2010;26: 1990–1998.

37. Norel R, Rice JJ, Stolovitzky G. The self-assessment trap: can we all be better than average? Mol Syst Biol. 2011;7: 537.

38. Stranneheim H, Käller M, Allander T, Andersson B, Arvestad L, Lundeberg J. Classification of DNA sequences using Bloom filters. Bioinformatics. 2010;26: 1595–1600.

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

40. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. 2011;39: e91.

41. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, et al. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 2011;8: 191–192.

42. Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, et al. High-resolution phylogenetic microbial community profiling. ISME J. 2016;10: 2020–2032.

43. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. PLoS One. 2012;7: e39315.

44. Yang IV. [4] Use of External Controls in Microarray Experiments. Methods Enzymol. 2006;411: 50–63.

45. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics. 2015;16: 856.

46. Singer E, Andreopoulos B, Bowers RM, Lee J, Deshpande S, Chiniquy J, et al. Next

generation sequencing data of a defined microbial mock community. Sci Data. 2016;3: 160081.

47. Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. Nat Commun. 2018;9: 3096.

48. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics and metagenomics. PLoS One. 2008;3: e3373.

49. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 2012;40: e94.

50. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28: 593–594.

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

52. Stoye J, Evers D, Meyer F. Rose: generating sequence families. Bioinformatics. 1998;14: 157–163.

53. Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C. ALF—A Simulation Framework for Genome Evolution. Mol Biol Evol. Oxford University Press; 2012;29: 1115–1123.

54. Lever J, Krzywinski M, Altman N. Points of Significance: Classification evaluation. Nat Methods. Nature Research; 2016;13: 603–604.

55. Jacobs J. Metagenomics - Tools, Methods and Madness. In: Google Docs [Internet]. [cited 21 Aug 2017]. Available: https://goo.gl/2gyNxK

56. Schwarzer G, Carpenter JR, Rücker G. Meta-Analysis with R. Springer; 2015.

57. Rücker G. Network meta-analysis, electrical networks and graph theory. Res Synth Methods. 2012;3: 312–324.

58. Gardner PP, Paterson JM, Ghomi FA, Umu SUU, McGimpsey S, Pawlik A. A meta-analysis of bioinformatics software benchmarks reveals that publication-bias unduly influences software accuracy [Internet]. bioRxiv. 2017. p. 092205. doi:10.1101/092205

59. Lumley T. Network meta-analysis for indirect treatment comparisons. Stat Med. 2002;21: 2313–2324.

60. Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med. 2004;23: 3105–3124.

61. Salanti G, Higgins JPT, Ades AE, Ioannidis JPA. Evaluation of networks of randomized trials. Stat Methods Med Res. 2008;17: 279–301.

62. Greco T, Biondi-Zoccai G, Saleh O, Pasin L, Cabrini L, Zangrillo A, et al. The attractiveness of network meta-analysis: a comprehensive systematic and narrative review. Heart Lung Vessel. 2015;7: 133–142.

63. Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Res Synth Methods. 2012;3: 98–110.

64. 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;

65. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16: 236.

66. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15: R46.

67. Minot SS, Krumm N, Greenfield NB. One codex: a sensitive and accurate data platform for genomic microbial identification. bioRxiv. biorxiv.org; 2015; Available: http://biorxiv.org/content/early/2015/09/28/027607.abstract

68. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, Keegan KP, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016;44: D590–4.

69. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016;12: e1004957.

70. 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;6: 90.

71. Davenport CF, Neugebauer J, Beckmann N, Friedrich B, Kameri B, Kokott S, et al. Genometa--a fast and accurate classifier for short metagenomic shotgun reads. PLoS One. 2012;7: e41224.

72. Freitas TAK, Li P-E, Scholz MB, Chain PSG. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. 2015;43: e69.

73. Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics. 2013;29: 2253–2260.

74. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75: 7537–7541.

75. Dröge J, Gregor I, McHardy AC. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics. 2015;31: 817–824.

76. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 2011;27:

127–129.

77. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7: 335–336.

78. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13: 581–583.

79. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol. 2013;31: 814–821.

80. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014; doi:10.1093/bioinformatics/btu153

81. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38: e132.

82. Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res. 2012;40: e9.

83. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38: e191.

84. Brady A, Salzberg S. PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat Methods. 2011;8: 367.

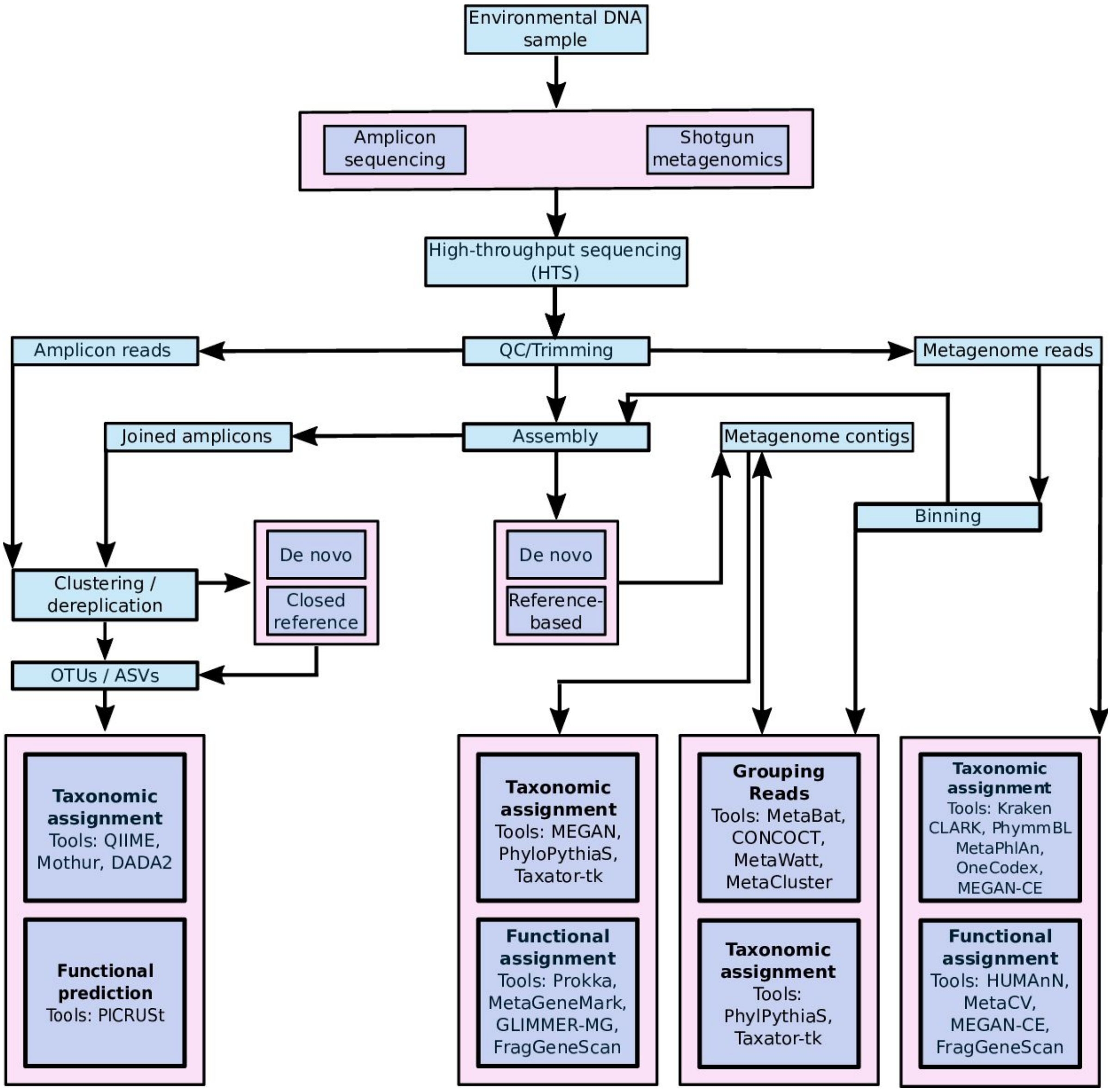
85. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12: 902–903.

86. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol. 2012;8: e1002358.

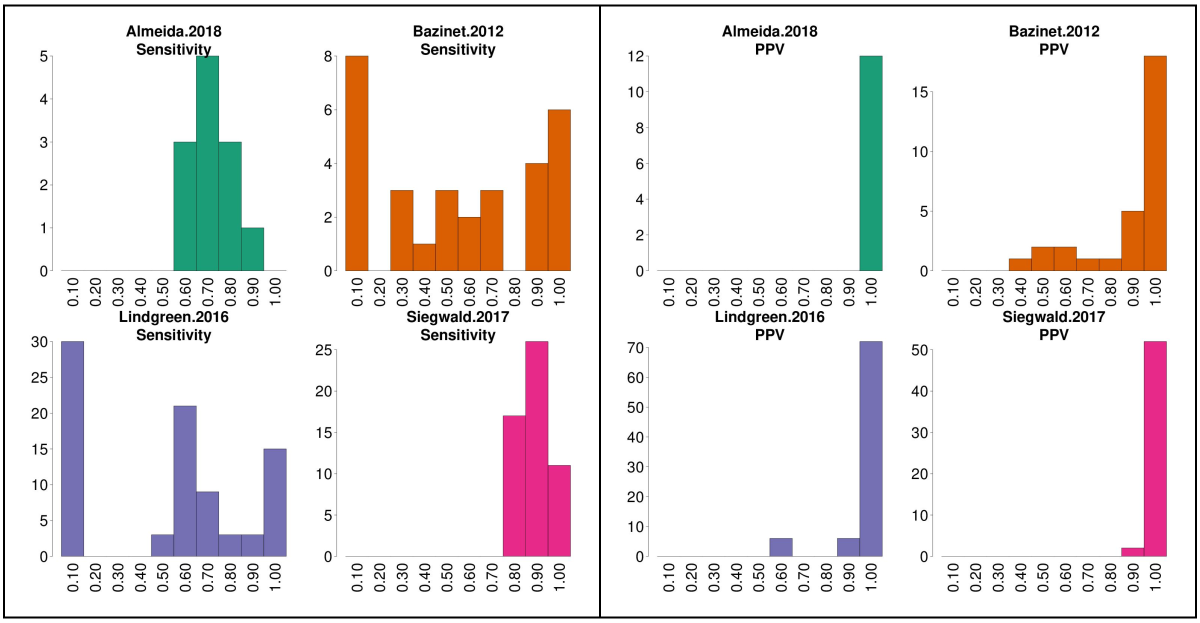
87. Liu J, Wang H, Yang H, Zhang Y, Wang J, Zhao F, et al. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic Acids Res. 2013;41: e3.

88. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. PeerJ Inc.; 2017;3: e104.

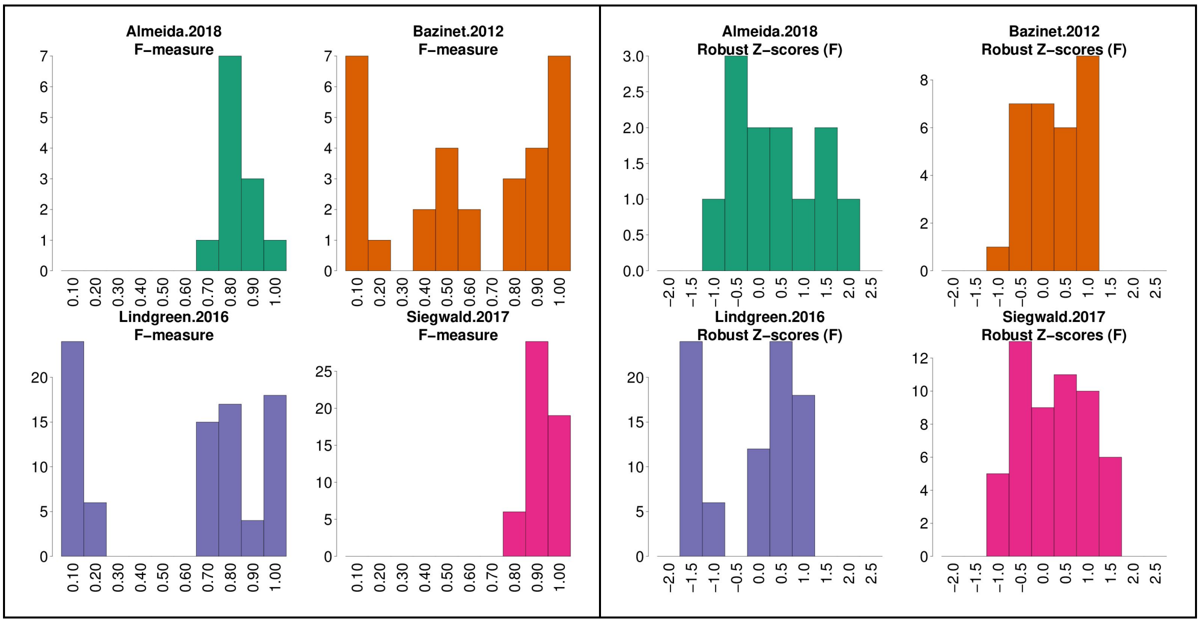
Supplementary Figures and Results


Environmental DNA sample
Amplicon sequencing
Shotgun metagenomics
High-throughput sequencing (HTS)
Amplicon reads
QC/Trimming
Metagenome reads
Joined amplicons
Assembly
Metagenome contigs
Binning
De novo
Closed reference
De novo
Reference-based
Clustering / dereplication
OTUs / ASVs
Taxonomic assignment Tools: QIIME, Mothur, DADA2
Functional prediction Tools: PICRUSt
Taxonomic assignment Tools: MEGAN, PhyloPythiaS, Taxator-tk
Functional assignment Tools: Prokka, MetaGeneMark, GLIMMER-MG, FragGeneScan
Grouping Reads Tools: MetaBat, CONCOCT, MetaWatt, MetaCluster
Taxonomic assignment Tools: PhylPythiaS, Taxator-tk
Taxonomic assignment Tools: Kraken CLARK, PhymmBL MetaPhlAn, OneCodex, MEGAN-CE
Functional assignment Tools: HUMAnN, MetaCV, MEGAN-CE, FragGeneScan

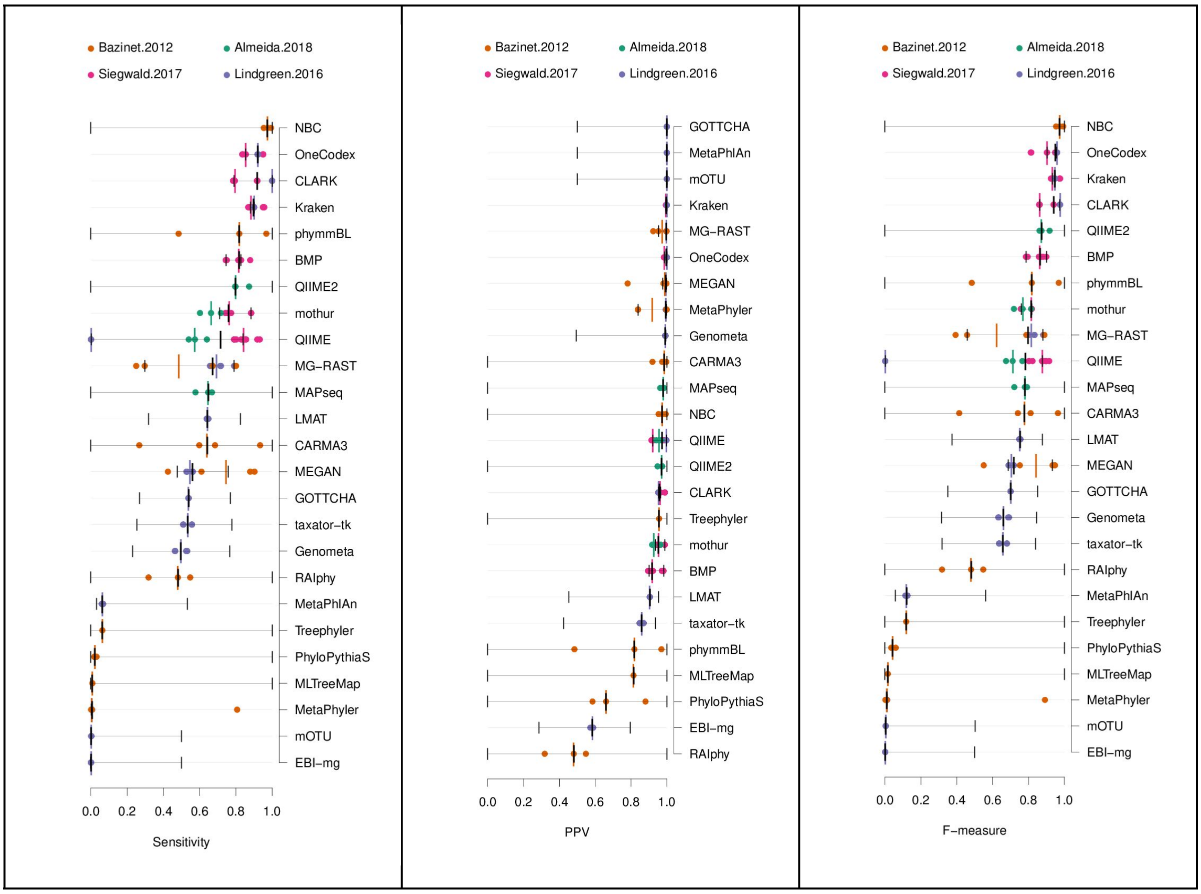
Supplementary Figure 1: A high-level summary of the main eDNA data production and analysis pathways. The main split is between amplicon or marker-gene based approaches and the shotgun metagenomics strategies. These sequences can be further processed and used to generate Operational Taxonomic Units (OTUs), Amplicon Sequence Variants (ASVs), and/or mapped onto reference databases for taxonomic and/or functional assignments. The tools referenced in the figure include QIIME [77], Mothur [74], DADA2 [78], PICRUSt [79], MEGAN [24], PhyloPythiaS [25], Taxator-tk [75], Prokka [80], MetaGeneMark [81], GLIMMER-MG [82], FragGeneScan [83], Kraken [66], CLARK [65], PhymmBL [84], MetaPhlAn [85], One Codex [67], MEGAN-CE [69], HUMAnN [86] and MEtaCV [87].



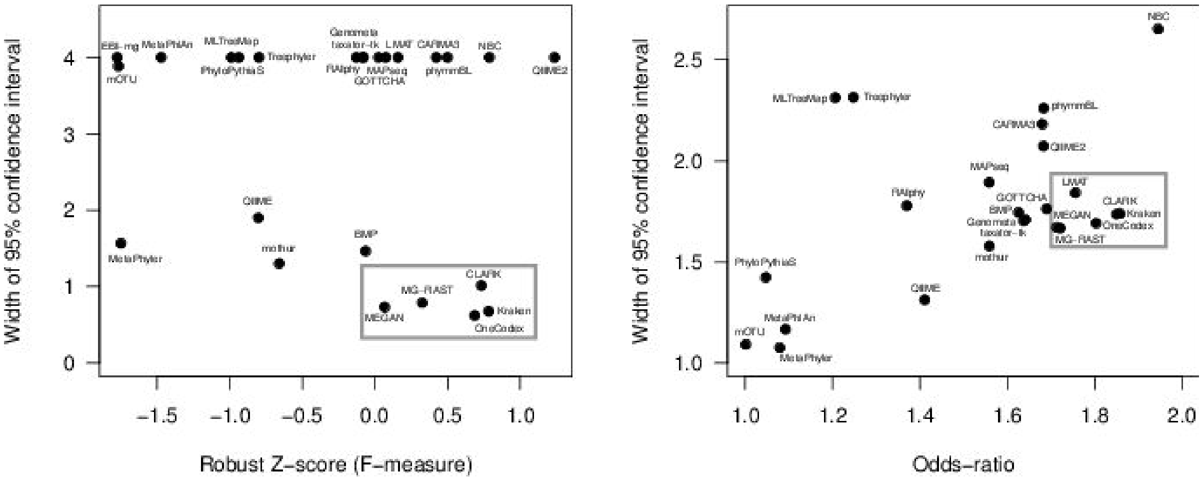
Supplementary Figure 2: The distribution of Sensitivity and PPV estimates for each of the six benchmark publications.



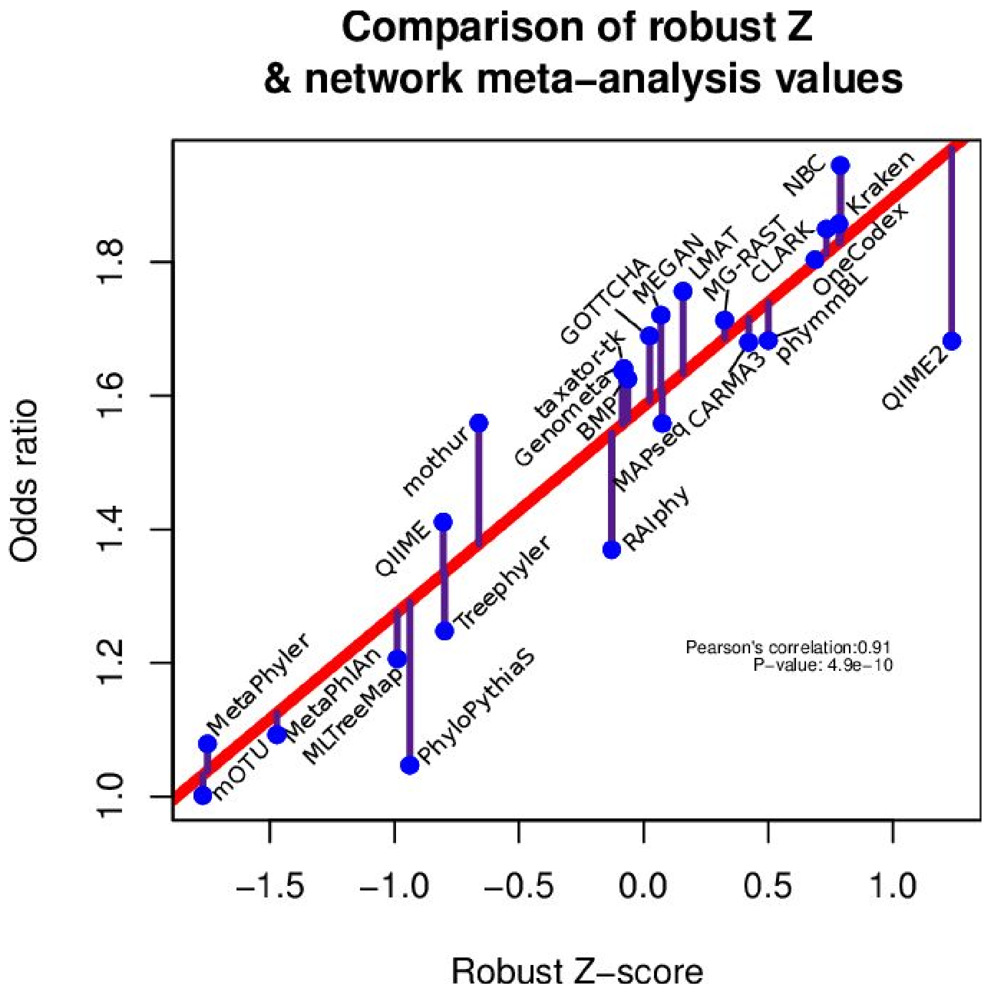
Supplementary Figure 3: A. The distributions of F-measure estimates for each of the six benchmark publications. B. The distributions of robust Z-scores for F-measure estimates for each of the six benchmark publications.



Supplementary Figure 4: Ranked lists of eDNA analysis tools, based upon median Sensitivity, PPV and F 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, PPV or F-measure are indicated with a thin black lines for each method.



Supplementary Figure 5: Estimated effect size (Robust Z-scores or Odds ratios) versus the confidence intervals. These plots show an alternative view of the forest-plots from Figure 3B & Figure 4A. The small sets of tools with comparatively high estimated accuracy and small confidence intervals have been indicated with grey boxes.

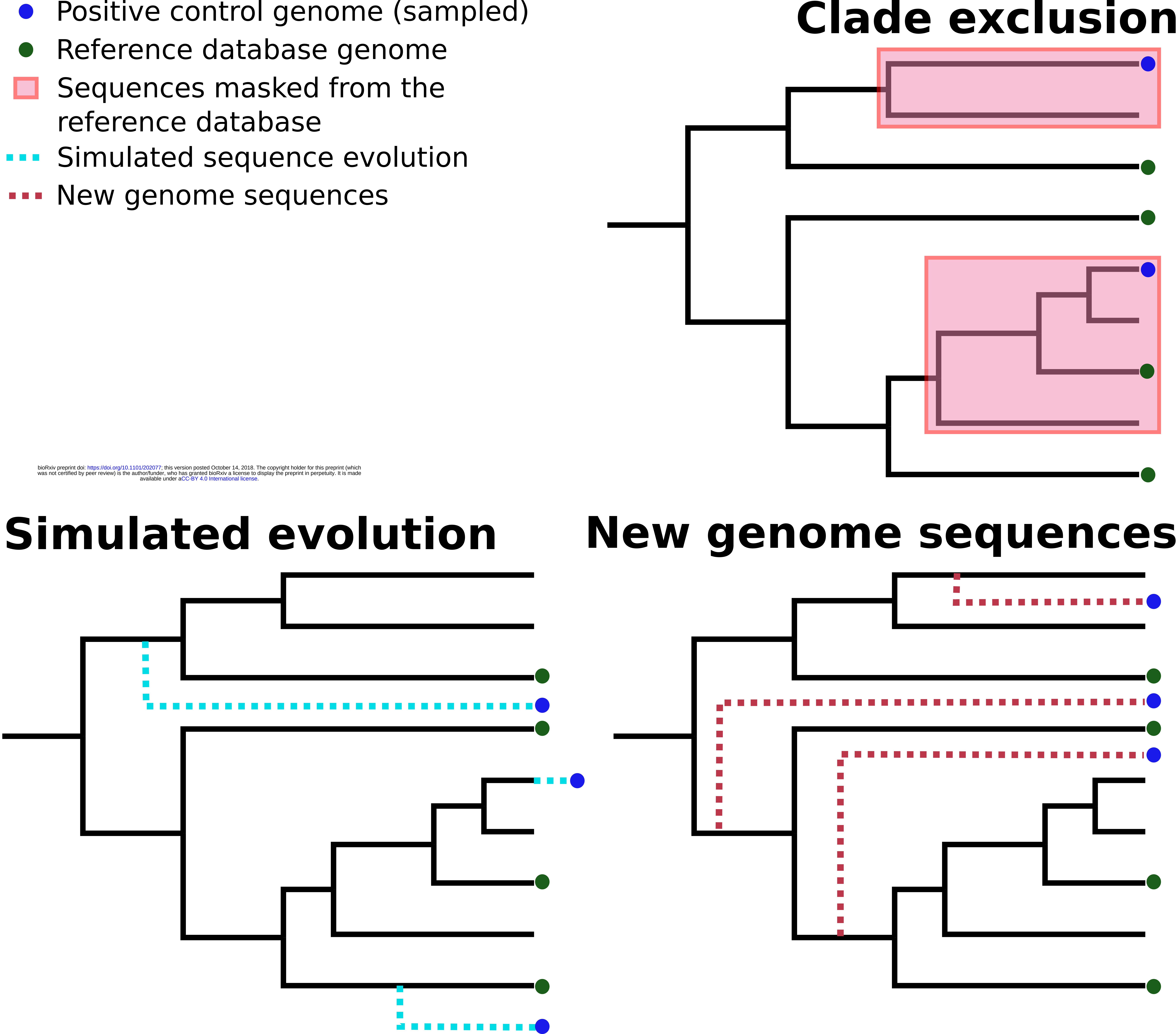


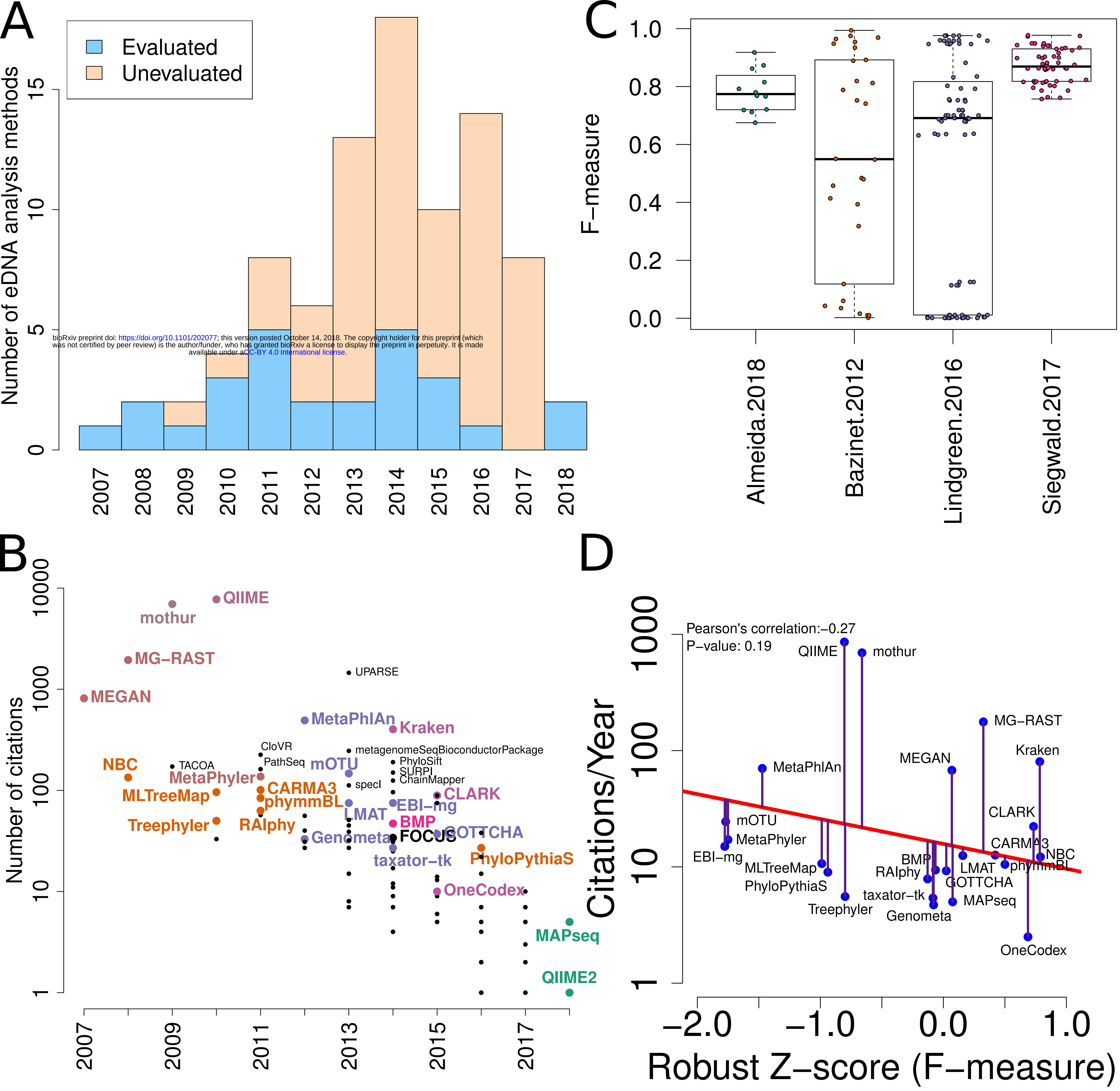
Supplementary Figure 6: Comparison of Robust Z-scores and odds ratios from the network meta-analysis. The Pearson's correlation coefficient between the two approaches for ranking software tools is 0.91 (P-value=4.9x10^{-10}).

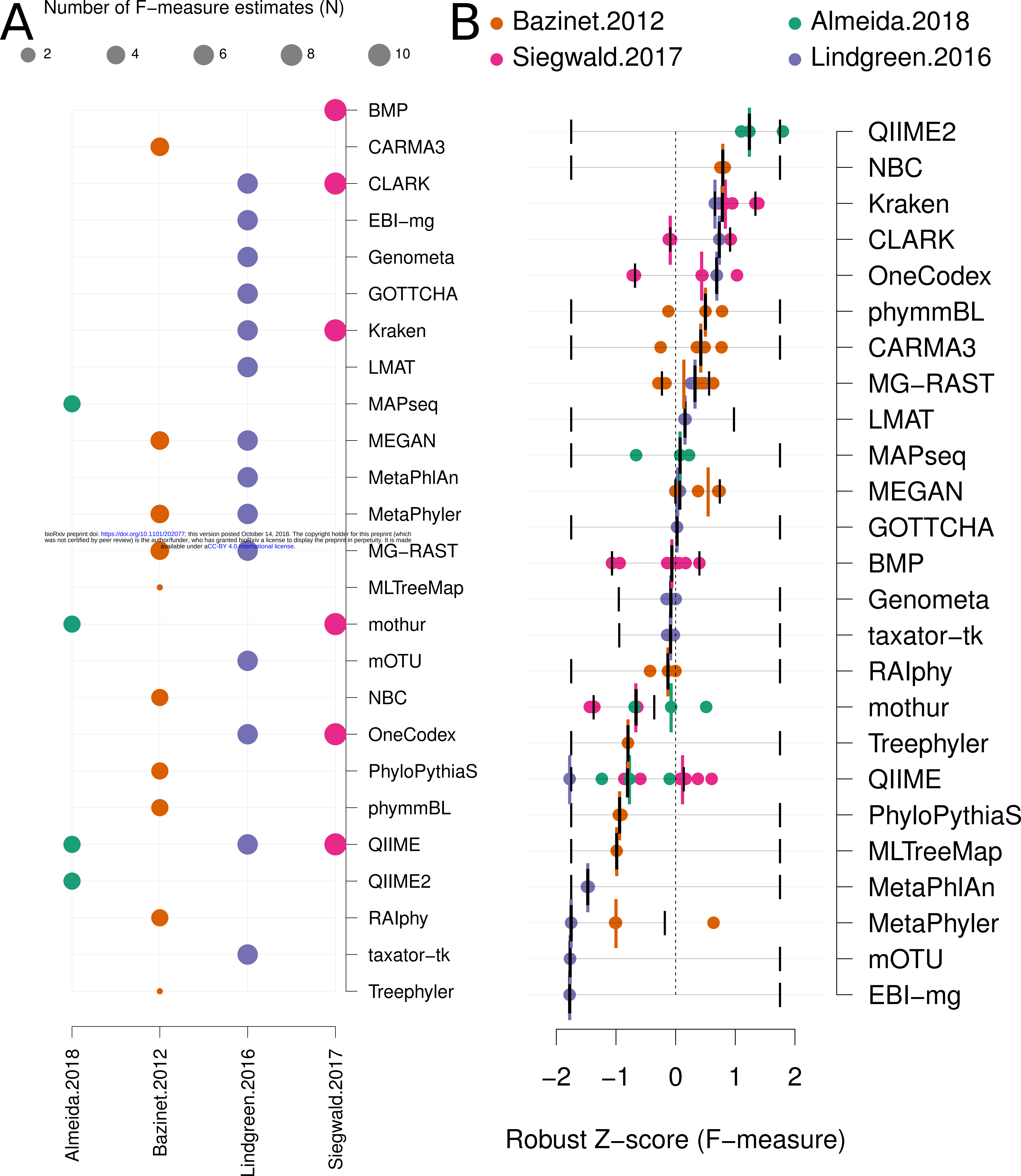
Paper	**Principle 1:** study focus is an evaluation	**Principle 2:** authors should be reasonably neutral	**Principle 3:** test data, evaluation and metrics should be rational
Almeida *et al.* (2018)	Yes	Yes	Yes
Bazinet *et al.* (2012)	Yes	Yes	Yes
Lindgreen *et al.* (2016)	Yes	Yes	Yes
McIntyre *et al.* (2017)	Yes	No*	Yes
Peabody *et al.* (2015)	Yes	Yes	No***
Sczyrba *et al.* (2017)	Yes	No**	Yes
Siegwald *et al.* (2017)	Yes	Yes	Yes

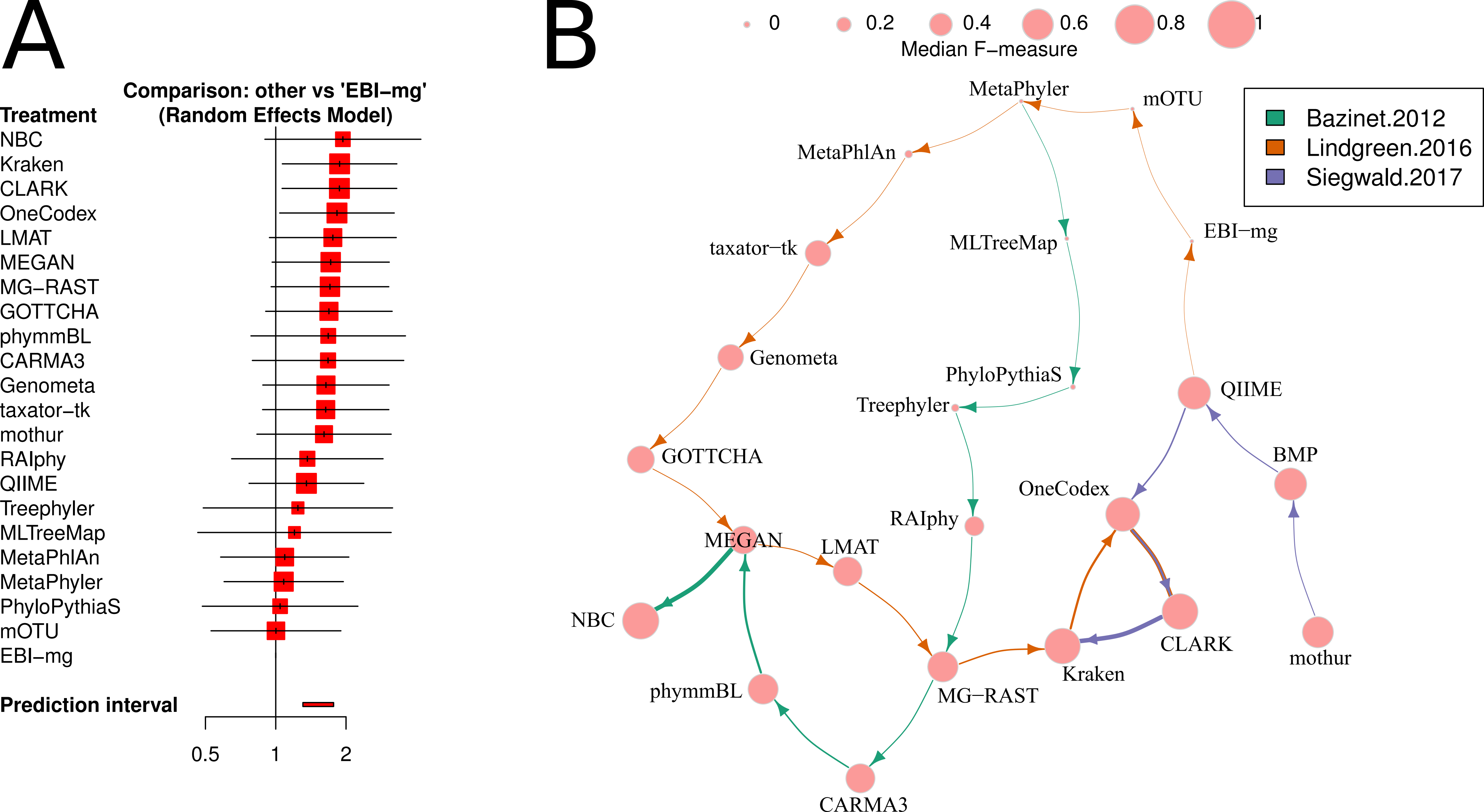
Supplementary Table 1: There are 3 main principles for benchmarking that authors should try to adhere to as suggested by [33]. **Principle 1**, the main focus of the study should be an evaluation. This criteria was evaluated manually be the authors of this study. **Principle 2**, benchmark authors should be reasonably neutral i.e. not involved in the development of methods included in the evaluation. This was evaluated below, by collecting method references provided by benchmark authors, these were tabulated and evaluated manually for overlap between authorship lists for the benchmarks and methods. **Principle 3**, the test data, evaluation and methods should be selected in a rational way. We assessed the number of taxa used and the evaluation metrics reported for each study. If either of these were too low or likely to be biased (e.g. only reporting sensitivity), then principle 3 was not met.

*The benchmark co-authors S Lonardi and R Ounit are co-authors of the methods CLARK and CLARK-S, GL Rosen is a co-author of the method NBC. All three were benchmarked in this study.
**12 or the 67 CAMI benchmark authors are also co-authors for 7 of the 14 methods that were benchmarked in this study.
***Subsequent analysis of the results from this manuscript highlight that the 11 taxa used in this evaluation is too few for robust accuracy estimates. Furthermore, 1 of the taxa has been renamed in subsequent taxonomies, making some of the accuracy estimates lower than these are in practise [88].


● 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


A
Evaluated
Unevaluated
Number of eDNA analysis methods
0
5
10
15
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
C
F−measure
0.0
0.2
0.4
0.6
0.8
1.0
Almeida.2018
Bazinet.2012
Lindgreen.2016
Siegwald.2017
B
Number of citations
1
10
100
1000
10000
QIIME
mothur
MG−RAST
MEGAN
UPARSE
MetaPhlAn
Kraken
CloVR
PathSeq
metagenomeSeqBioconductorPackage
TACOA
mOTU
PhyloSift
SURPI
ChainMapper
NBC
MetaPhyler
CARMA3
specI
MLTreeMap
phymmBL
CLARK
EBI−mg
LMAT
BMP
RAlphy
Treephyler
Genometa
FOCUS
GOTTCHA
taxator−tk
PhyloPythiaS
OneCodex
MAPseq
QIIME2
2007
2009
2011
2013
2015
2017
D
Pearson's correlation:−0.27
P−value: 0.19
Citations/Year
1
10
100
1000
QIIME
mothur
MG−RAST
MetaPhlAn
MEGAN
Kraken
mOTU
MetaPhyler
EBI−mg
CLARK
CARMA3
NBC
phymmBL
MLTreeMap
PhyloPythiaS
BMP
RAlphy
LMAT
GOTTCHA
taxator−tk
Treephyler
Genometa
MAPseq
OneCodex
−2.0
−1.0
0.0
1.0
Robust Z−score (F−measure)


A
Number of F−measure estimates (N)
2
4
6
8
10
BMP
CARMA3
CLARK
EBI−mg
Genometa
GOTTCHA
Kraken
LMAT
MAPseq
MEGAN
MetaPhlAn
MetaPhyler
MG−RAST
MLTreeMap
mothur
mOTU
NBC
OneCodex
PhyloPythiaS
phymmBL
QIIME
QIIME2
RAIphy
taxator−tk
Treephyler
Almeida.2018
Bazinet.2012
Lindgreen.2016
Siegwald.2017
B
Bazinet.2012
Siegwald.2017
Almeida.2018
Lindgreen.2016
QIIME2
NBC
Kraken
CLARK
OneCodex
phymmBL
CARMA3
MG−RAST
LMAT
MAPseq
MEGAN
GOTTCHA
BMP
Genometa
taxator−tk
RAIphy
mothur
Treephyler
QIIME
PhyloPythiaS
MLTreeMap
MetaPhlAn
MetaPhyler
mOTU
EBI−mg
−2
−1
0
1
2
Robust Z−score (F−measure)


A
Comparison: other vs 'EBI–mg'
(Random Effects Model)
Treatment
NBC
Kraken
CLARK
OneCodex
LMAT
MEGAN
MG–RAST
GOTTCHA
phymmBL
CARMA3
Genometa
taxator–tk
mothur
RAIphy
QIIME
Treephyler
MLTreeMap
MetaPhlAn
MetaPhyler
PhyloPythiaS
mOTU
EBI–mg
Prediction interval
0.5
1
2
B
0
0.2
0.4
0.6
0.8
1
Median F–measure
🟩 Bazinet.2012
🟧 Lindgreen.2016
🟪 Siegwald.2017
MetaPhyler
mOTU
MetaPhlAn
taxator–tk
MLTreeMap
EBI–mg
Genometa
PhyloPythiaS
Treephyler
QIIME
GOTTCHA
BMP
OneCodex
RAIphy
MEGAN
LMAT
NBC
CLARK
Kraken
mothur
phymmBL
MG–RAST
CARMA3