MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12399

METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data

JOHAN BENGTSSON-PALME,* MARTIN HARTMANN,†‡ KARL MARTIN ERIKSSON,§ CHANDAN PAL,* KAISA THORELL,¶**†† DAN GÖRAN JOAKIM LARSSON* and ROLF HENRIK NILSSON‡‡
*Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, 413 46, Gothenburg, Sweden, †Forest Soils and Biogeochemistry, Swiss Federal Research Institute WSL, CH-8903 Birmensdorf, Switzerland, ‡Molecular Ecology, Institute for Sustainability Sciences, Agroscope, CH-8046 Zurich, Switzerland, §Department of Shipping and Marine Technology, Chalmers University of Technology, 412 96 Gothenburg, Sweden, ¶Department of Microbiology and Immunology, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Box 435, 40530 Gothenburg, Sweden, **Department of Chemical and Biological Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden, ††Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Nobels Väg 16, 171 77 Stockholm, Sweden, ‡‡Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Gothenburg, Sweden

Abstract

The ribosomal rRNA genes are widely used as genetic markers for taxonomic identification of microbes. Particularly the small subunit (SSU; 16S/18S) rRNA gene is frequently used for species- or genus-level identification, but also the large subunit (LSU; 23S/28S) rRNA gene is employed in taxonomic assignment. The METAXA software tool is a popular utility for extracting partial rRNA sequences from large sequencing data sets and assigning them to an archaeal, bacterial, nuclear eukaryote, mitochondrial or chloroplast origin. This study describes a comprehensive update to METAXA – METAXA2 – that extends the capabilities of the tool, introducing support for the LSU rRNA gene, a greatly improved classifier allowing classification down to genus or species level, as well as enhanced support for short-read (100 bp) and paired-end sequences, among other changes. The performance of METAXA2 was compared to other commonly used taxonomic classifiers, showing that METAXA2 often outperforms previous methods in terms of making correct predictions while maintaining a low misclassification rate. METAXA2 is freely available from http://microbiology.se/software/metaxa2/.

Keywords: 16S, 18S, metagenomics, microbial communities, rRNA libraries, taxonomic assignment

Received 26 March 2014; revision received 25 February 2015; accepted 26 February 2015

Introduction

Detailed studies of microbial communities are largely dependent on DNA sequencing efforts. Community structure is frequently assessed using taxonomic markers, notably the ribosomal small subunit (SSU; 16S/18S) and/or large subunit (LSU; 23S/28S) rRNA genes. These genes are present not only in archaea, bacteria and the nucleus of eukaryotes, but also in the eukaryotic mitochondria and chloroplasts. It is thus of great importance to separate these categories from each other before assessing, for example, bacterial diversity in a sample, or bias and noise will follow (Taberlet *et al.* 2012). Furthermore, in large shotgun metagenomes, identifying and extracting such marker genes is not trivial and – with steady growth of the size of sequence data sets – increasingly time-consuming. The METAXA software (Bengtsson *et al.* 2011) offers a remedy to these problems for the SSU gene and for sequence data sets with read lengths down to about 200 base pairs (bp). In contrast to tools such as QIIME (Caporaso *et al.* 2010), Mothur (Schloss *et al.* 2009), the RDP classifier (Wang *et al.* 2007) and Rtax (Soergel *et al.* 2012), which primarily seek to discern the taxonomic identity of sequences that are already known to derive from the rRNA genes, such as amplicon libraries, METAXA is principally concerned with identification of fragments from such genes in metagenomic data sets and sorting them by organism or organelle type. This also sets METAXA apart from, for example, MG-RAST (Meyer *et al.* 2008) and MEGAN (Huson & Mitra 2012), which function more as complete pipelines for annotation and analysis of metagenomic sequence data. METAXA has found various uses, not only for the envisaged task

Correspondence: Johan Bengtsson-Palme, Fax: +46 31 786 2560; E-mail: johan.bengtsson-palme@microbiology.se

of identifying SSU sequences in metagenomes, but also in quality control of SSU sequences (e.g. Ghai *et al.* 2012), exclusion of unwanted SSU sequences from environmental data (e.g. Duguma *et al.* 2013) and reconstruction of full-length SSU sequences from fragmentary metagenomic sequence data (Fan *et al.* 2012). Additionally, its software design has carried over into specialized packages for other barcoding regions, such as the ITS region in eukaryotes (Bengtsson-Palme *et al.* 2013b). However, increasing use of very short-read sequencing techniques, frequently using paired-end reads, for metagenomic studies requires tools that can reliably detect rRNA sequence fragments in reads as short as 100 bp (Yoccoz 2012). Although the read lengths obtained by next-generation sequencing technologies are steadily increasing, the need for trustworthy detection of rRNA in sequences as short as 100 bp has not diminished, both because of the need to analyse legacy data, but also to support processing of the many shotgun metagenomics projects that sacrifice read length for sequencing depth. In this study, we introduce a thorough update of METAXA – METAXA2 – which is rewritten to account for such short-read lengths and to support paired-end libraries. METAXA2 also incorporates the ability to process LSU (23S/28S) rRNA sequences. Although LSU sequences are less commonly used for DNA metabarcoding applications, it is in some organismal groups employed as a complement to, or preferred over, the SSU gene (Sonnenberg *et al.* 2007; Schoch *et al.* 2012; Kerekes *et al.* 2013; Wurzbacher *et al.* 2014). In addition, METAXA2 adds support for BLAST+ and HMMER 3.1 and can directly handle input in the FASTQ format in addition to FASTA. Finally, METAXA2 comes with a completely redesigned taxonomic classifier, making it able to accurately deduce the taxonomic information of rRNA sequences down to the genus level and often further.

Materials and methods

Implementation

The fundamental analysis procedure of METAXA has been described previously (Bengtsson *et al.* 2011; Bengtsson-Palme *et al.* 2013a). In brief, input sequences are screened for the presence of the conserved regions flanking the hypervariable (V/D) regions of the SSU/LSU gene using hidden Markov models (HMMs) using HMMER (Eddy 2011). Those HMMs are based on alignments of sequences from all phyla with full-length rRNA sequences deposited in the SILVA database (Quast *et al.* 2013), and due to the high sensitivity of HMMs to sequence variation (Freyhult *et al.* 2007), they should be able to detect the target sequences also from previously nonsequenced organisms. In the case of paired-end input sequences from large-insert libraries, these are merged assuming an insert size that can be specified by the user; the reverse read is reverse-complemented and added after the insert. Sequences found by HMMER to contain rRNA genes are then further subjected to a BLAST search (Altschul *et al.* 1997) against a specialized database for classification. Cases that cannot unambiguously be assigned to archaea, bacteria, eukaryota, mitochondria or chloroplast origin are aligned to their five best BLAST matches using MAFFT (Katoh & Toh 2008) to facilitate further manual examination. For the present update, METAXA has been completely redesigned to be able to handle the enormous data sets generated by modern sequencing techniques, making it much faster and more memory efficient than its predecessor. Moreover, additional HMMs for LSU rRNA detection were constructed in accordance with proposed guidelines (Hartmann *et al.* 2010) to complement previously published LSU HMMs (Kerekes *et al.* 2013).

To enable taxonomic classification to the genus and species level, we have dramatically improved the classification engine in METAXA2. To begin with, the classifier now reports matches with a reliability score. This score is based on the sequence identity to the closest BLAST matches, as well as the taxonomic resolution with which the sequence can be classified given the region matched in the rRNA sequences from the database. For each rRNA entry, METAXA2 compares the taxonomic affiliation of the top five BLAST matches to each other. In each comparison, the percentage identity between the query and the database sequence, weighted by the length of the overlap of the two sequences, is taken into account. If the BLAST matches point to the same taxonomic origin, the query sequence is given a taxonomic affiliation with a high reliability score (close to 100). BLAST matches below a certain identity threshold, specified specifically for each taxonomic level, are not considered in this comparison. If the reliability score is below 80, the comparison is repeated at the taxonomic level above (e.g. genus if the previous comparison was made on species level), until the reliability score is above 80. In this way, all detected rRNA sequences get a taxonomic classification at a taxonomic level that is very likely to be correct, although in some instances not very specific. As a result, cases where a species classification can be uniquely derived from the top five BLAST matches will be reported with taxonomic classification down to the species level. If there are ambiguities at the species level, the software proceeds to the genus level, tries to find a nonconflicting classification and continues in this manner until a unique, definite classification can be given for the input sequence at some taxonomic level.

To allow classification to the genus and/or species level, we have updated the classification databases in

METAXA2, building on manually curated entries from SILVA (release 111; Quast *et al.* 2013) and MITOZOA (version 2.0; release 10; D'Onorio de Meo *et al.* 2012), cross-checked with data from Greengenes (DeSantis *et al.* 2006; McDonald *et al.* 2012), CRW (Cannone *et al.* 2002) and GenBank (Benson *et al.* 2014). From these databases, all full-length SSU and LSU sequences were retrieved, and their taxonomic information assessed. For the curated database, rRNA sequences without complete taxonomic affiliations to the species level were excluded, along with sequences known to be chimeric. This also means that sequences from uncultured organisms, or from metagenomics projects, are not included in the classification database. Finally, sequences with contradictory taxonomic information in the curated database were also excluded from the data set. In this way, we have ensured that there should not be sequences with incomplete or insufficient taxonomic information that might perturb the classification process of METAXA2. Due to incomplete coverage of chloroplast and mitochondrial taxonomic information, classification beyond organelle type is unfortunately generally not possible using the new classifier. The new classification engine is described in detail in the METAXA2 manual (Item S1, Supporting information), which also describes additional implementation details.

Software evaluation

The performance of METAXA2 was evaluated in a similar way as the original METAXA version (Bengtsson *et al.* 2011). All tests were carried out using METAXA2 version 2.0.1 (METAXA2-se is used to indicate single-end reads in this text, and METAXA2-pe is used to indicate paired-end reads). To simulate rRNA fragments derived from shotgun metagenomes, nine data sets were generated by randomly selecting a stretch of base pairs from each sequence in a set of handpicked high-quality SSU and LSU sequences of known taxonomic origin, with the length of that stretch ranging from 50 bp to full-length rRNA sequences. Each of those sequence sets contained 1000 sequences, deriving from 200 archaeal, 200 bacterial, 200 nuclear eukaryote, 200 chloroplast and 200 mitochondrial rRNA sequences, of which 100 from each taxonomic group were SSU sequences and 100 were LSU sequences. The annotations of the sequences that these data sets were derived from were all of high quality, and thus the same sequences, or close variants of them, are in most cases also present in the classification database of METAXA2. This means that at longer read lengths, the classification accuracy should be optimal given the taxonomic divergence within each taxonomic group. In addition, we generated four simulated paired-end data sets of 1000 pairs of sequences each, in the same way as above, but with a 150-bp gap between the two stretches representing the nonsequenced region of the DNA fragment, with length of each 'read' in the pair being 50, 100, 200 or 300 bp, respectively (Data Set S1, Supporting information). We then let METAXA2 classify these simulated read data sets and calculated its performance in terms of sensitivity (proportion of detected simulated reads) and accuracy (proportion of correctly classified simulated reads). In addition, we ran METAXA2 with the options '–taxlevel 6' and '–taxlevel 7' on the same data sets, to force classifications to the genus and species level, respectively. Each read (or read pair) that was detected by METAXA2 as part of an rRNA sequence, regardless of domain or organelle, is reported as 'detected' in the following text, while we will use the word 'assigned' for those sequences correctly assigned to their correct domain or organelle. A simulated read (or read pair) was regarded correctly classified if the taxonomy reported by METAXA2 corresponded to the known taxonomy of the sequence that the read was derived from, at all taxonomic levels reported by METAXA2. If the METAXA2 classification was found to correspond to the known taxonomic affiliation all the way to the species or genus level, the read (or read pair) was regarded perfectly classified to species or genus, respectively. If METAXA2 reported any taxonomic affiliation that was incorrect – at any taxonomic level – the read was regarded as misclassified.

To test the impact of replacing the native METAXA2 reference data set with other databases, we created two separate METAXA2-compatible databases, constructed from the nonredundant SILVA release 111 reference database and from the Greengenes 13_8 reference database, respectively. We then classified the simulated SSU rRNA fragments created above using these two reference databases and compared the result to the result obtained when using the native METAXA2 database.

To assess the impact of which part of the rRNA gene a sequence fragment was derived from, we extracted the individual V/D regions from the full-length data sets with known taxonomic affiliations using V-Xtractor (Hartmann *et al.* 2010) with the additional option '-i long' to retain the HMM region required for detection by METAXA2. Chloroplast V regions were extracted using the bacterial mode, eukaryote V/D regions using the fungal mode, and mitochondrial sequences were excluded, as V-Xtractor does not offer a suitable mode for those. We then randomly introduced errors at rates of 0%, 1%, 3%, 5%, 10%, 25% and 50%, with 90% of errors being substitutions and 10% being insertions or deletions (Data Set S2, Supporting information). Each of the resulting data sets was then classified using METAXA2.

To compare the classification accuracy of METAXA2 to other commonly used taxonomic classifiers, we used the

same simulated sets of reads derived from full-length SSU sequences with known taxonomic information (Data Set S1, Supporting information) and classified those using the Mothur naïve Bayesian classifier (QIIME-Mothur-NBC; Schloss *et al.* 2009), the RDP Classifier (QIIME-RDP-NBC; Wang *et al.* 2007), Rtax (Soergel *et al.* 2012) and Uclust (QIIME-Uclust; Edgar 2010), as implemented in QIIME (Caporaso *et al.* 2010), based on the Greengenes 13_8 database as reference (McDonald *et al.* 2012). Rtax was run on both simulated single-end (QIIME-Rtax-se) and paired-end reads (QIIME-Rtax-pe), while all other software tools were only tested on single-end reads. As Mothur produced surprisingly poor results in the QIIME implementation, we also ran the software separately using default Mothur settings, a bootstrap cut-off value of 80 and the SILVA release 119 reference database, as recommended in the Mothur documentation (Native-Mothur-NBC). As not all of those tools have been designed to handle all rRNA types supported by METAXA2, we only compared the classification accuracy for bacterial 16S rRNA sequences. Classification accuracy was investigated at the genus level, and a classification was counted as correct if the true genus name was predicted using default settings. Classification was regarded to be incorrect if the wrong genus name was predicted. If no classification was given at the genus level for a particular sequence, that entry was not counted towards either the correct or incorrect totals.

As METAXA2 does not provide species diversity estimates and similar metrics, we used the output of METAXA2 to investigate whether that could be employed to recapture species richness. For this purpose, we used the identified SSU sequence fragments produced by METAXA2 in the above evaluation as input to MEGRAFT version 1.0.2 (Bengtsson *et al.* 2012), producing a set of sequences that could be clustered to generate OTUs. These sequences were clustered in USEARCH (Edgar 2010), with the setting '–id 0.93'. Rarefaction curves were produced using the R statistical program (R Development Core Team 2011) and VEGAN (Oksanen *et al.* 2011). To further assess the ability of METAXA2 to reproduce estimates of taxonomic composition from metagenomes compared to 16S rRNA amplicon data, we ran METAXA2 on one of the samples (MG-RAST IDs 4481963.3 and 4481964.3) generated in a study by Gibbons *et al.* (2014), in which the same microbial communities were sequenced using both shotgun metagenomic (100-bp paired-end Illumina HiSeq-2000 data) and amplicon sequencing (16S rRNA V4 region data generated by the Illumina MiSeq platform, read length 151 bp). Here, only the bacterial part of the METAXA2 output was investigated, as the amplicon sequencing targeted the bacterial 16S rRNA gene. We also compared the METAXA2 classifications to the classifications made by MG-RAST (based on SILVA SSU; Meyer *et al.* 2008). Furthermore, we compared the classifications on phylum and genus level of METAXA2-pe, Native-Mothur-NBC, QIIME-RDP-NBC, QIIME-Rtax-pe and QIIME-Uclust on the rRNA fragments extracted from the Illumina shotgun metagenome by METAXA2.

To test the capabilities of METAXA2 on larger data sets, we ran the entire nonredundant SILVA 111 release (Quast *et al.* 2013), including all SSU and LSU sequences, through the software and counted the number of sequences that were classified according to their annotated origin by METAXA2. Finally, to test the susceptibility to false-positive detections, nine data sets of lengths 50, 100, 200, 300, 500, 750, 1000, 1250 and 1500 bp, each containing five-million random DNA sequences, were generated using the EMBOSS 6.2.0 suite (Rice *et al.* 2000) and run through the software in both SSU and LSU detection modes (Data Set S3, Supporting information). We also downloaded the human genome assembly GRCh38 from NCBI GenBank and divided each chromosome into 1000-bp stretches. These 1000-bp sequences were then used as input to METAXA2, the QIIME implementations of RDP (QIIME-RDP-NBC), Rtax (QIIME-Rtax-se) and Uclust (QIIME-Uclust), as well as Mothur with its recommended default options (Native-Mothur-NBC). All entries in these data sets producing a taxonomic prediction (as opposed to 'unknown' or 'unclassified') were regarded as false-positive detections, unless they, in the human genome case, were derived from taxonomic groups within the Craniata.

A complete list of software versions and commands used for this evaluation can be found in Item S2 (Supporting information).

Results

Detection efficiency and classification accuracy

To evaluate the performance of METAXA2, we ran the software on simulated data sets derived from rRNA sequences with known taxonomic classification, 13 for each of the SSU and the LSU genes (Data Set S1, Supporting information). METAXA2 successfully detected over 99% of bacterial and archaeal SSU sequences from 100-bp paired-end reads and longer (Fig. 1; Fig. S1, Supporting information), with <0.5% misassignment to the wrong kingdom or organelle (Fig. S2, Supporting information). Even with paired-end 50-bp reads, METAXA2 could extract 74% of the test SSU sequences (55% for single-end reads). However, the misassignment rate was slightly higher using single-end reads (Fig. S2, Supporting information). For the LSU gene, the number of uncertain assignments increased slightly from 50 bp to 100 and 200 bp, due to the fact that more sequences were detected as rRNA, although many of them were still too short for reliable

assignment. The extraction ability of METAXA2 is thus better or on par with that of the previous version; for paired-end reads, METAXA2 outperforms its predecessor for all SSU types at 100 bp. The detection efficiency for the LSU region was slightly lower than that of the SSU region, especially at shorter read lengths (Fig. 1). This is likely due to the substantially longer variable regions in the LSU compared to the SSU, especially in nuclear eukaryote sequences.

We also examined whether METAXA2 could make reliable taxonomic classifications of fragmentary sequences using the new classifier (Fig. 2). At 100-bp single-end reads, METAXA2 correctly reported a classification at some taxonomic level for 98% of the detected SSUs; the corresponding value for the LSU sequences was 95%. The classifications obtained from paired-end data were as accurate as those for single-end data, although a substantially larger proportion of SSU sequences could be detected (Fig. 2). By forcing METAXA2 to report genus- or species-level classifications (option –taxlevel 6 or 7), we could obtain a higher number of sequences classified to the genus and species levels (Fig. 2; Figs S3 and S4, Supporting information), albeit with considerably higher proportions of misclassified entries, particularly in the forced species classification case (Fig. 2). Finally, we performed OTU richness estimates on the simulated archaeal, bacterial and eukaryote SSU data sets. The output from METAXA2 was used as input to the MEGRAFT software, and we found that METAXA2 can be used to accurately recapture OTU richness estimates of environmental communities (Fig. S5, Supporting information). However, using paired-end sequences derived from METAXA2 as input for MEGRAFT to produce estimates of OTU richness consistently showed higher estimated diversity than did single-end sequences, indicating that single-end reads might be preferable for such applications. This might, however, also reflect a weakness in the way MEGRAFT processes paired-end data – a sequence class for which it was not primarily designed.

The manually curated METAXA2 database has a substantial role in ensuring a high rate of correct classifications. When this database was substituted for the nonredundant SILVA 111 release or the Greengenes 13_8 reference database used by QIIME, classification accuracy decreased substantially, primarily for rRNA sequences from eukaryota, mitochondria and chloroplasts but even for bacteria the accuracy declined slightly (Fig. S6, Supporting information).

Effects of sequencing errors and V-region bias

Testing the performance of METAXA2 across all V/D regions in the SSU and LSU genes showed that the software correctly detected all sequences as SSU genes for the V2 to V9 regions, and more than 99% of the V1 region sequences (Table S1, Supporting information). For the LSU gene, the detection efficiency was 100%, or close to 100%, for all regions except V3/D2 and V10/D8, where efficiency was substantially lower (78.4% and 79.7%, respectively; Table S2, Supporting information). Furthermore, across all regions in the SSU gene, the proportion of classifications being correct at the level reported by METAXA2 among the detected sequences was higher than 98% (Fig. 3a), while for the LSU gene, this proportion was on average 95.7% (Fig. 3b). METAXA2 was found to generally be able to detect all rRNA genes in sequences with up to 5% error rate. Even at a 10% error rate, the detection efficiency was, on average, above 95%. In addition, at up to 10% sequence error rate, METAXA2 was able to classify sequences to their correct taxonomic origin, at some level, in more than 96% of the detected rRNA sequences (Tables S3 and S4, Supporting information). However, it should be noted that METAXA2 seldom classified the query sequences to the species or genus level at the 5% and 10% sequence error rates.

Performance on real data sets

Similar to its predecessor, METAXA2 assigned 99.913% of the 286 858 SSU sequences in the nonredundant SILVA 111 release to their annotated domain/organelle, while 99.782% of the 29 306 sequences in the corresponding LSU release were assigned according to domain/organelle annotation (Table 1). When evaluated on true metagenomic data, we found that METAXA2 results were comparable to results derived from amplicon sequencing of the 16S rRNA gene. On the phylum level, METAXA2 results for amplicons and metagenomic data were generally similar, except for the Actinobacteria, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Gemmatimonadetes and Tenericutes phyla (Fig. 4a). In most cases, the METAXA2 results also mirrored results derived from MG-RAST (Fig. 4a). The majority of the cases with large discrepancies between methods occurred in phyla constituting small proportions of the total community, indicating that the differences may largely be due to noise related to difficulties in detecting taxonomic groups close to the detection limit of either amplicon or shotgun metagenomic sequencing. Indeed, sufficient sampling depth is very important for recovering community structure (Pinto & Raskin 2012), and the number of classifications in each taxonomic group needs to be interpreted in the light of the total number of extracted rRNA sequences from a metagenome. This could, however, not explain why the proportion of reads reported as Actinobacteria was dramatically higher in the metagenomic classifications, both by METAXA2 and MG-RAST,

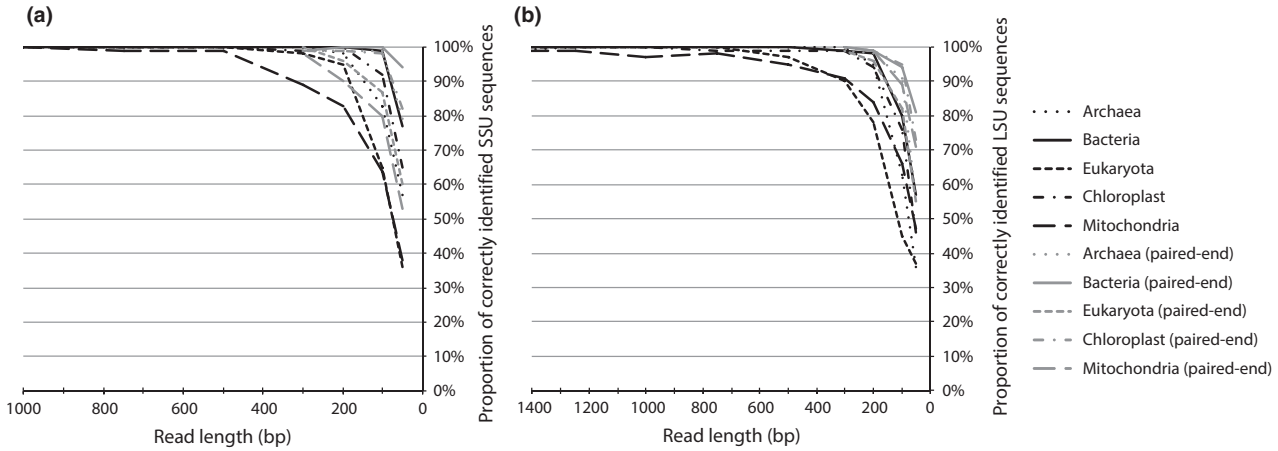

Fig. 1 Proportion of successfully detected and correctly assigned SSU (a) and LSU (b) rRNA gene fragments at decreasing read lengths.

than the amplicon-derived estimate. Also at the genus level, the abundance estimates made by METAXA2 from the metagenome compared well to the amplicon-derived approximations, particularly for large genera (Fig. 4b).

Comparison to other classification tools

There are a number of tools for classifying 16S rRNA sequences, both adapted for full-length or long-read data, and specialized on short reads. We compared the ability to make accurate taxonomic classifications for bacteria at the genus level of Metaxa2 and some commonly used tools implemented in QIIME: the Mothur naïve Bayesian classifier (QIIME-Mothur-NBC), the RDP Classifier (QIIME-RDP-NBC), Rtax (QIIME-Rtax-se/pe) and Uclust (QIIME-Uclust), all based on the Greengenes reference database. The comparison showed that METAXA2 performs better than all of those tools at commonly used read lengths for metagenomic studies (100–200 bp), which is what METAXA2 is primarily designed for (Fig. 5a). In addition, METAXA2 achieves this with a very small number of incorrect classifications (Fig. 5b). At 50-bp read lengths, METAXA2 performed on par with Rtax on single-end sequences, although Rtax performed better using paired-end input (Table S5, Supporting information). However, Rtax achieved this classification power while also displaying a higher misclassification rate than METAXA2 (Fig. 5b and Table S6, Supporting information). On longer pyrosequencing- and Sanger-like read lengths, METAXA2 performs better than all other tested tools, having above 85% correct genus classifications and no incorrect classifications above 300-bp reads. Interestingly, paired-end reads did not improve the classification accuracy of METAXA2 for bacterial 16S rRNA, while it did for Rtax. However, this generally results from METAXA2 showing higher detection efficiency but lower classification certainty using paired reads, emphasizing that using paired-end data will most often increase the number of SSU sequences detected in a metagenome, but not necessarily aid in classifying these sequence fragments to the genus level. Notably, the QIIME implementation of Mothur (QIIME-Mothur-NBC) performed poorly at all read lengths, resulting in very few classifications, of which more were incorrect than correct. When Mothur was run separately with the SILVA release 119 as reference database (Native-Mothur-NBC), it classified a much larger proportion of sequences to the genus level, with low numbers of incorrect classifications, especially at long-read lengths. This indicates that the default Mothur settings in QIIME are not ideal for fragmentary metagenomic 16S rRNA and that the choice of reference database is crucial for the accuracy of taxonomic predictions. We also compared the classifications of the rRNA fragments from the Gibbons *et al*. (2014) real metagenomic data made by these tools to those made by METAXA2 on the phylum and genus levels. Overall, classifications were similar, although some discrepancies were apparent for the Chloroflexi, Cyanobacteria, Firmicutes and Gemmatimonadetes phyla (Fig. S7, Supporting information). The difference between METAXA2 and the other tools in estimated abundance of Cyanobacteria is to a large extent caused by a (likely correct) classification of cyanobacterial rRNA sequences as Chloroplasts instead of Cyanobacteria by METAXA2. The results on genus level were more variable, but generally corresponded between the different tools. However, the similarity seemed to be larger between tools using the same reference database (QIIME-RDP-NBC, QIIME-Rtax-pe and QIIME-Uclust), although the METAXA2 classifications were quite similar to all other tested methods (Pearson correlation coefficient just below 0.9; Table S7, Supporting information).

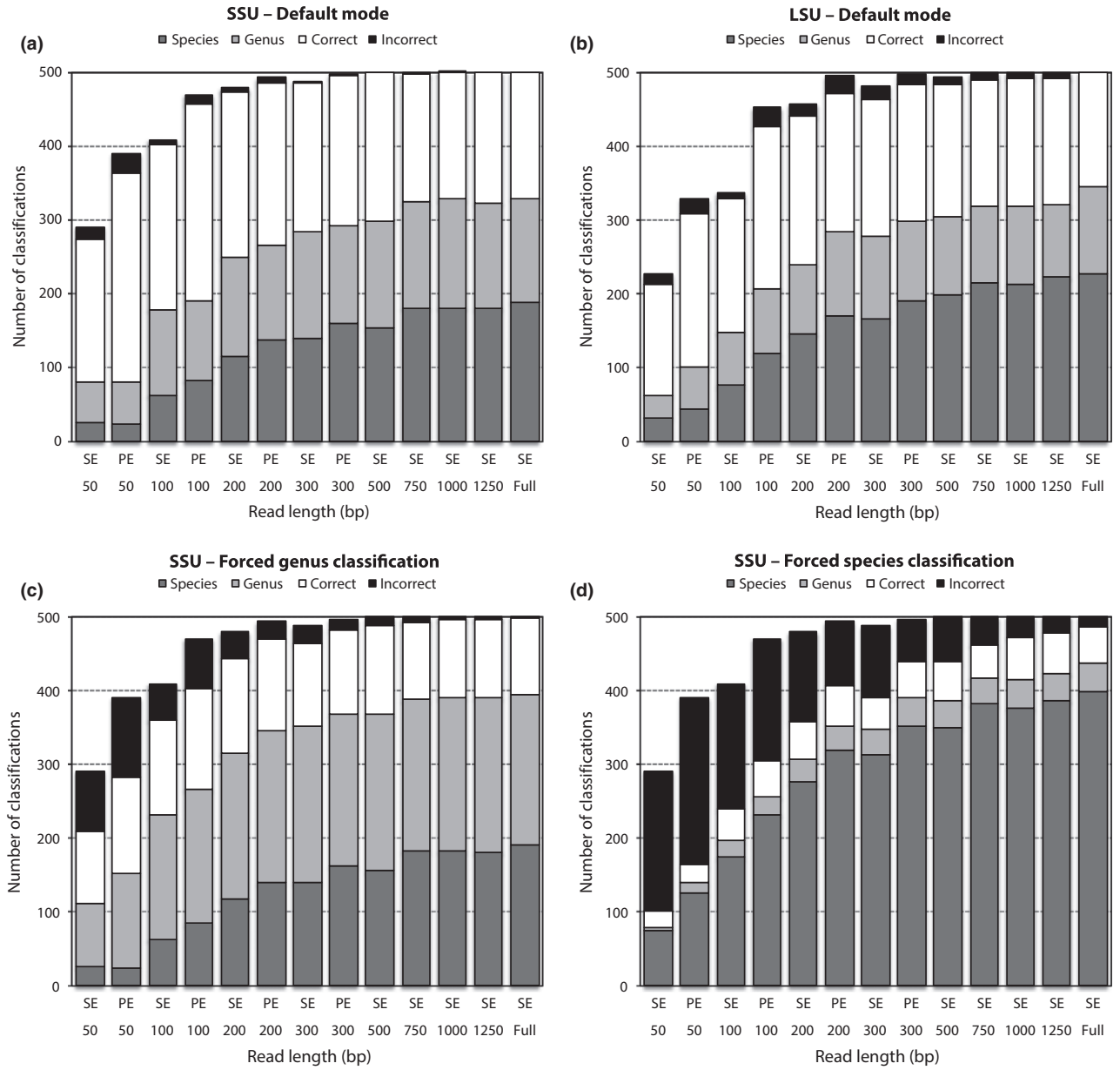

Fig. 2 Number of successful classifications in a manually quality-controlled data set of SSU (a) and LSU (b) sequences. METAXA2 was also forced to issue genus-level (c) and species-level (d) predictions of taxonomic origin. 'Species' denotes classifications where METAXA2 (using default settings) issued a full prediction down to the species level, in accordance with the taxonomic classification of the input sequence. 'Genus' denotes correct classifications at the genus level, and 'Correct' denotes classifications where METAXA2 classified a sequence consistently with the classification of the input sequence, but above the genus level. 'Incorrect' classifications were, at some taxonomic level, misclassified by METAXA2. SE denotes single-end and PE denotes paired-end input sequences. All cases included 500 sequences, and sequences not classified were not detected as rRNA.

False-positive rate

To test the susceptibility to false-positive detections of METAXA2, we ran nine data sets with random DNA sequences through the software in both SSU and LSU detection modes (Data Set S3, Supporting information). We also ran the software on the human genome, counting the number of non-Craniata rRNA detections. Both these approaches generated zero matches, indicating a false-positive rate even lower than the 0.00012% of the previous METAXA version, and also shows that METAXA2 is substantially more robust against false-positive detections than many of the other tested tools primarily designed to work with amplicon data (Table S8,

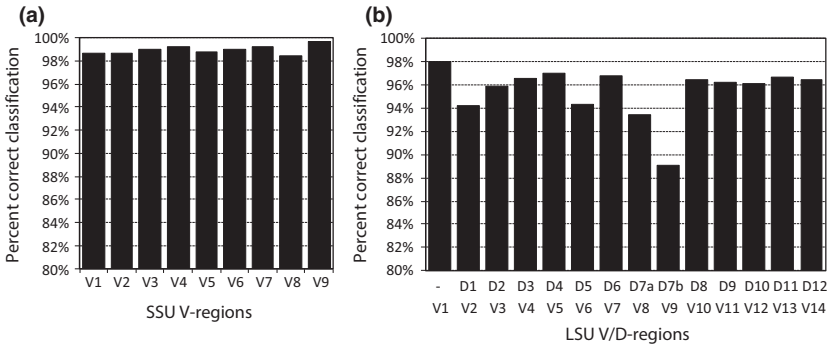

Fig. 3 Classification accuracy of the different V regions of the SSU gene (a) and the V/D regions of the LSU (b) gene (V regions in archaea, bacteria and chloroplasts, D regions in eukaryota). Accuracy was measured as classification consistent with that of the input sequence, but did not require a full species name.

Table 1 Assignment statistics for the SSU and LSU sequences in the SILVA 111 release generated from the METAXA2 output. Note that in most cases, the numbers in the 'As annotated' and 'Different' columns do not add up to the number of 'Annotations in SILVA', because some sequences in SILVA were not annotated even to the domain level

	SILVA 111 SSU			SILVA 111 LSU		
Origin	As annotated	Different	Annotations in SILVA	As annotated	Different	Annotations in SILVA
Archaea	10 908	1	10 919	405	0	405
Bacteria	241 634	63	241 805	17 763	4	17 765
Eukaryota	31 862	2	31 862	9268	53	9268
Chloroplast	1824	61	1843	1528	2	1531
Mitochondria	379	0	429	278	0	284
Uncertain		123			5	
Total	286 607	250	286 858	29 242	64	29 306

Supporting information). Running the 5-million, 300-bp sequence set took 24 min for the SSU rRNA gene on an eight-core Linux computer (2.4 GHz Intel Xeon CPUs), while searching the human genome took 5 min and 25 s.

Discussion

METAXA2 constitutes a substantial update to the original METAXA software, adding support for LSU rRNA detection, large-scale paired-end libraries with short-read lengths, a completely revised taxonomic classifier, along with a range of other important revisions pertaining to speed, memory efficiency, and ease of use. In addition, METAXA2 outperforms its predecessor in terms of sensitivity and accuracy, especially at short-read lengths. Furthermore, as can be seen when comparing simulated single-end and paired-end reads, covering a larger portion of the rRNA gene systematically improves detection efficiency, and in some cases also classification accuracy. This effect is entirely due to pairs of reads having a larger probability to cover the rRNA gene both in conserved and variable regions. Covering a larger portion of the gene also aids in detecting rRNA fragments in the first place. As METAXA2 relies on detection of the conserved regions of the rRNA genes, the use of short-read lengths reduces the ability of the software to extract SSU and LSU sequences from a larger sequence set. However, as can be seen in Fig. 1, already 100-bp reads are sufficient to detect nearly all tested bacterial rRNAs, and using 100-bp paired-end reads recaptures over 90% of SSU sequences regardless of organelle or taxonomic domain. It should be noted that the SSU rRNA gene is far more used in studies on microbial diversity than the LSU gene, which translates into a richer set of SSU reference sequences compared to the LSU. This, in turn, influences the performance of METAXA2 on LSU data (Figs 1 and 2). A further complication with the LSU gene compared to the SSU gene is its longer variable regions (Fig. 3). The variable regions are substantially harder to account for using HMMs, impairing the ability of METAXA2 to reliably detect LSU fragments in short-read data (100 bp and shorter; Fig. 2). Despite these limitations, METAXA2 could detect more than 90% of LSU sequences at 100 bp using paired-end reads, of which 93% could be correctly classified.

METAXA2 refines and streamlines species- and genus-level classification, which required additional manual steps in the former version of METAXA. This classification can be forced to be reported at a certain taxonomic level (using the –taxlevel option). Using this option to force

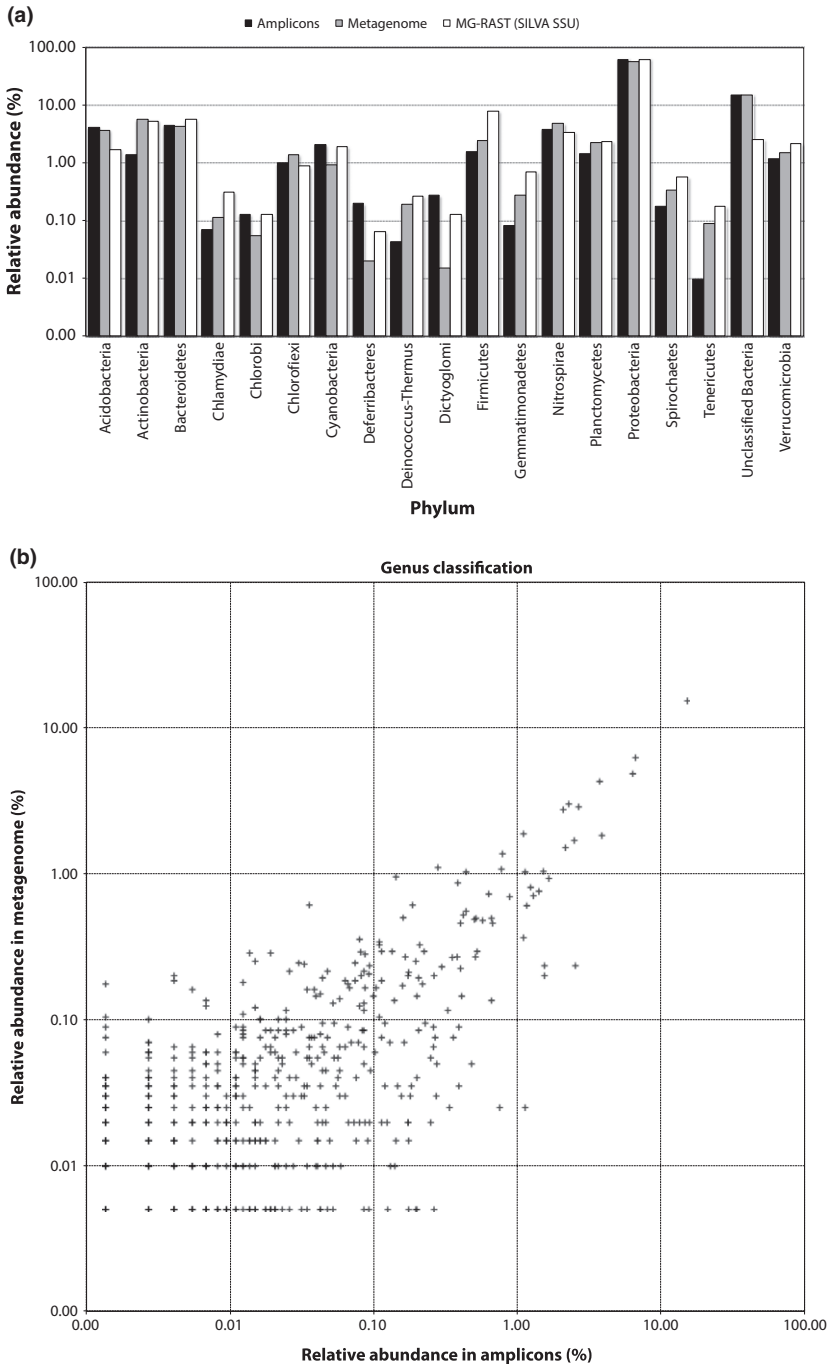

Fig. 4 Evaluation of METAXA2 performance on real metagenomic data. (a) Comparison of phylum relative abundance estimated from amplicon sequencing of the 16S rRNA gene, extraction of SSU sequences from the metagenome using METAXA2 and classification of SSU sequences using MG-RAST. (b) Comparison of the relative abundance of genera estimated using amplicon sequencing (*x*-axis) and extraction of SSU reads from the metagenome using METAXA2 (*y*-axis). Note that the scale in both figure (a) and (b) is logarithmic.

species classifications gives rise to higher numbers of full species classifications, but at the expense of substantially higher misclassification rates (Fig. 2), particularly when read lengths are short. This option should therefore be used with caution on short-read data sets, and knowingly of the fact that simply choosing the taxonomic classification of the best scoring BLAST match as annotation for your input sequence can be highly unreliable for read fragments derived from certain regions of the rRNA sequence, particularly in the conserved parts of the SSU and LSU genes. It is apparent that this effect is affecting METAXA2 far more for the LSU gene than the SSU gene (Fig. 3). For the SSU, previous studies have suggested that the V3 and V4 regions would give more accurate results compared to the V5 or V9 region for short-read data (Mizrahi-Man *et al.* 2013). However, for METAXA2, this problem seems to be minor. The classifier of METAXA2 assigns nearly all sequences in the SILVA database to

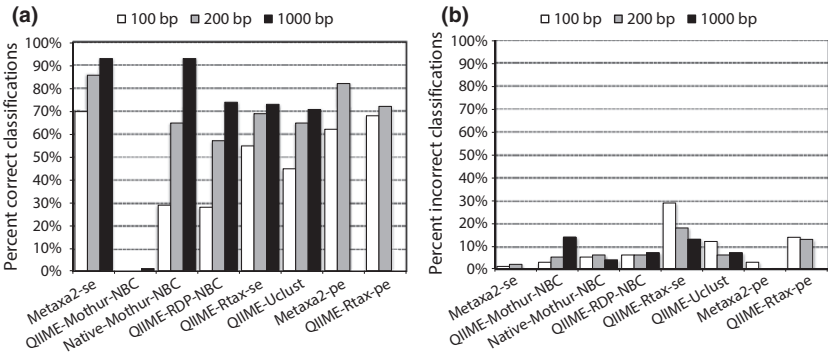

Fig. 5 Comparison of classification accuracy between METAXA2 and other commonly used tools for taxonomic classification, available through QIIME. The percentage correct classifications (a) was measured as classifications where the software predicted the same genus as that of the input sequence, while incorrect classifications (b) was measured as prediction to the wrong genus compared to that of the input sequence. METAXA2 and Rtax were run in both single-end and paired-end modes. Mothur was run both as implemented in QIIME (QIIME-Mothur-NBC) and standalone using settings recommended in the Mothur documentation (Native-Mothur-NBC).

their annotated domain or organelle. Unfortunately, investigating the accuracy of every single classification in the SILVA database more specifically would not be meaningful, as not all sequence entries in the database have complete taxonomy, and we do not know whether the reported taxonomic information is correct for any particular given entry. The minute percentage of sequences (0.087% for the SSU gene and 0.218% for the LSU gene) not classified identically in SILVA and by METAXA2 might reflect misassignment by METAXA2, or mistakes in the database annotation. Such database misannotations might result from, for example, symbionts or parasites in eukaryotes, or contamination in rRNA libraries.

All sequencing platforms currently in use are prone to sequencing errors of different types and at varying rates (Loman *et al.* 2012). Such errors can have an impact on both estimates of diversity and species delimitation, as sequencing errors can be hard to distinguish from actual mutations or variants within the population (Eren *et al.* 2013). As METAXA2 classifies rRNA sequence fragments based on a set of best database matches, the software is reasonably robust against this type of errors. The program is able to identify almost all rRNA sequences with up to 10% errors, but its classifications will be less specific at higher error rates. This allows METAXA2 to have almost as high accuracy for its predictions at 10% error rate as with no errors introduced. It should, however, be noted that at 5% error rate and above, METAXA2 only rarely claims to be able to determine taxonomy at the genus level. The solid performance of METAXA2 also on rRNA sequences with 10% errors indicates that the software would be able to detect rRNA sequences also from species with fairly low sequence similarity to those found in the current databases, which is desirable in the case of DNA sequencing of samples from less well-studied environments. In such cases, METAXA2 would retain classification accuracy for sequences with high degrees of divergence from the reference sequences by only providing a taxonomic prediction at higher levels, such as phylum, order or class, thereby reducing the potential for novel sequences to be misclassified.

The new classifier also fares well when compared with other frequently used classification software for bacterial 16S rRNA sequences implemented in QIIME. In particular, METAXA2 maintains very low misclassification rates even at very short-read lengths, such as 50 bp, choosing not to classify sequences at a given taxonomic level rather than to return an unreliable classification. In contrast, Rtax (QIIME-Rtax-se/pe), which had roughly the same proportion of correctly classified entries as METAXA2 – or even higher at 50-bp reads – often outputs an incorrect genus prediction at short-read lengths. At longer read lengths, the QIIME RDP Classifier (QIIME-RDP-NBC) and Uclust (QIIME-Uclust) implementations perform roughly on par with Rtax, with similar misclassification rates, while Rtax excels when using short reads. Still, METAXA2 has a higher proportion of correctly classified sequences even at longer read lengths. Thus, METAXA2 can be used for classification of both short and long rRNA reads with high accuracy, although the benefit compared to other tools will be smaller the longer the read length. It should be noted that all these comparisons have been made using the default settings and the database supplied with QIIME, which might be suboptimal for some software and read length combinations. On the other hand, we suspect that many users of QIIME (and other software, including METAXA2) will not go beyond the default settings, making our comparison relevant for common usage cases. Finally, we found that Mothur is very conservative about making predictions at the genus level using the default QIIME settings

(QIIME-Mothur-NBC), making it nearly unusable for short-read data unless the settings are tweaked. However, the Mothur naïve Bayesian classifier performs substantially better using the recommended reference set for Mothur – SILVA release 119 – constituting a much larger reference set of SSU rRNA genes (Native-Mothur-NBC). This emphasizes the importance of maintaining a high-quality reference database that is suitable for the purpose of each individual analysis. We strongly believe that the extensive effort put into curating the METAXA2 reference database contributes to the convincing performance of the new classifier, as accuracy dropped when the entire SILVA or Greengenes reference databases were used in place of the native METAXA2 database.

In conclusion, METAXA2 provides increased efficiency and accuracy of rRNA detection and classification in short-read data sets, with negligible proportions of false positives and misclassifications. In addition, METAXA2 combines detection of SSU and LSU rRNA in one package and includes a new classifier capable of species- or genus-level classification. METAXA2 is freely available under the GNU GPLv3 software licence at http://microbiology.se/software/metaxa2/. We believe that METAXA2 will further increase the accuracy of annotation and analysis of metagenomes and similar data sets, even at short-read lengths.

Acknowledgements

DGJL acknowledges financial support from the Swedish Research Council (VR), the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and the Swedish Foundation for Strategic Environmental Research (MISTRA). KME and RHN acknowledge financial support from FORMAS. Five anonymous reviewers are acknowledged for constructive comments on earlier versions of the manuscript.

References

Altschul SF, Madden TL, Schäffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bengtsson J, Eriksson KM, Hartmann M *et al.* (2011) Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, **100**, 471–475.

Bengtsson J, Hartmann M, Unterseher M *et al.* (2012) Megraft: a software package to graft ribosomal small subunit (16S/18S) fragments onto full-length sequences for accurate species richness and sequencing depth analysis in pyrosequencing-length metagenomes and similar environmental datasets. *Research in Microbiology*, **163**, 407–412.

Bengtsson-Palme J, Hartmann M, Eriksson KM, Nilsson RH (2013a) Metaxa, overview. In: *Encyclopedia of Metagenomics* (ed. Nelson KE), pp. 1–5. Springer, New York.

Bengtsson-Palme J, Ryberg M, Hartmann M *et al.* (2013b) Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution*, **4**, 914–919.

Benson DA, Clark K, Karsch-Mizrachi I *et al.* (2014) GenBank. *Nucleic Acids Research*, **42**, D32–D37.

Cannone JJ, Subramanian S, Schnare MN *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.

Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

DeSantis TZ, Hugenholtz P, Larsen N *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**, 5069–5072.

D'Onorio de Meo P, D'Antonio M, Griggio F *et al.* (2012) MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. *Nucleic Acids Research*, **40**, D1168–D1172.

Duguma D, Rugman-Jones P, Kaufman MG *et al.* (2013) Bacterial communities associated with culex mosquito larvae and two emergent aquatic plants of bioremediation importance. *PLoS ONE*, **8**, e72522.

Eddy SR (2011) Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Eren AM, Maignien L, Sul WJ *et al.* (2013) Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, **4**, 1111–1119.

Fan L, McElroy K, Thomas T (2012) Reconstruction of ribosomal RNA genes from metagenomic data. *PLoS ONE*, **7**, e39948.

Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research*, **17**, 117–125.

Ghai R, Hernandez CM, Picazo A *et al.* (2012) Metagenomes of mediterranean coastal lagoons. *Scientific Reports*, **2**, 490.

Gibbons SM, Jones E, Bearquiver A *et al.* (2014) Human and environmental impacts on river sediment microbial communities. *PLoS ONE*, **9**, e97435.

Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH (2010) V-Xtractor: an open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, **83**, 250–253.

Huson DH, Mitra S (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Methods in Molecular Biology*, **856**, 415–429.

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, **9**, 286–298.

Kerekes J, Kaspari M, Stevenson B *et al.* (2013) Nutrient enrichment increased species richness of leaf litter fungal assemblages in a tropical forest. *Molecular Ecology*, **22**, 2827–2838.

Loman NJ, Misra RV, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

McDonald D, Price MN, Goodrich J *et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, **6**, 610–618.

Meyer F, Paarmann D, Dsouza M *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mizrahi-Man O, Davenport ER, Gilad Y (2013) Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS ONE*, **8**, e53608.

Oksanen J, Blanchet FG, Kindt R *et al.* (2011) vegan: Community Ecology Package.

Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS ONE*, **7**, e43093.

Quast C, Pruesse E, Yilmaz P *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. *Trends in Genetics: TIG*, **16**, 276–277.

Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Schoch CL, Seifert KA, Huhndorf S *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6241–6246.

Soergel DAW, Dey N, Knight R, Brenner SE (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *The ISME Journal*, **6**, 1440–1444.

Sonnenberg R, Nolte AW, Tautz D (2007) An evaluation of LSU rDNA D1-D2 sequences for their use in species identification. *Frontiers in Zoology*, **4**, 6.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.

Wurzbacher C, Rösel S, Rychła A, Grossart H-P (2014) Importance of saprotrophic freshwater fungi for pollen degradation. *PLoS ONE*, **9**, e94643.

Yoccoz NG (2012) The future of environmental DNA in ecology. *Molecular Ecology*, **21**, 2031–2038.

J.B.P., K.M.E., K.T., D.G.J.L. and R.H.N. designed the software. J.B.P., M.H. and R.H.N. implemented the software. J.B.P., M.H., K.M.E., C.P. and K.T. evaluated the software. J.B.P. and R.H.N. drafted the manuscript. All authors have contributed to and approved the final manuscript.

Data accessibility

The program, user manual, and example data set are freely available at http://microbiology.se/software/metaxa2/

The manually curated test sets, and their accompanying randomly excerpted subsequences (Data Set S1, Supporting information), the V/D region subsequences (Data Set S2, Supporting information) and the METAXA2 output from the false-positive tests (Data Set S3, Supporting information) are available as supplementary material to this publication.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Proportion of successfully detected and correctly assigned SSU (a) and LSU (b) rRNA fragments at decreasing read lengths.

Fig. S2 Proportion of misassigned entries and sequences marked as 'uncertain' by METAXA2.

Fig. S3 Number of successful classifications at different taxonomic levels in the manually quality-controlled data set of SSU sequences using default options (a), using the –taxlevel 7 option to force species classifications of SSU sequences (b), and using the –taxlevel 6 option to force genus-level classifications (c).

Fig. S4 Number of successful classifications at different taxonomic levels, made as in Fig. S3 (Supporting information), but on the manually quality-controlled dataset of LSU sequences.

Fig. S5 OTU accumulation curves generated using VEGAN, based on rRNA sequence output from the MEGRAFT software generated from SSU sequence sets of different read lengths extracted by METAXA2.

Fig. S6 Correct classification rate of METAXA2 using the native classification database and the SILVA and GreenGenes databases for (a) only bacterial SSU rRNA, (b) archaeal, bacterial, eukaryote 18S, chloroplast and mitochondrial SSU rRNA sequences.

Fig. S7 Comparison of different classification software on phylum level on real metagenomic data.

Table S1 METAXA2 detection efficiency of the different V regions of the SSU sequence, at increasing error rates.

Table S2 METAXA2 detection efficiency of the different V/D regions of the LSU sequence, at increasing error rates.

Table S3 METAXA2 classification accuracy of the sequence fragments detected as rRNA genes from the different V regions of the SSU sequence, at increasing error rates.

Table S4 METAXA2 classification accuracy of the sequence fragments detected as rRNA genes from the different V/D regions of the LSU sequence, at increasing error rates.

Table S5 Percent correct classifications at different read lengths of tools commonly used for taxonomic classification.

Table S6 Percent incorrect classifications at different read lengths of tools commonly used for taxonomic classification.

Table S7 Pearson correlation coefficients between different classification methods for genera on real metagenomic data.

Table S8 False positive detection rate of METAXA2 compared to other commonly used classification tools.

Item S1 The manual for METAXA2.

Item S2 Complete listing of software versions and commands used for the software evaluation.

Data Set S1 The manually curated test sets, and randomly excerpted subsequences of those used for benchmarking accuracy of the performance of METAXA2 on SSU and LSU data sets.

Data Set S2 Excerpted subsequences corresponding to the V/D regions of the rRNA gene extracted by V-Xtractor.

Data Set S3 The output from METAXA2 on the data sets of random DNA sequences used for false-positive detection of METAXA2.