Post-print of

Bengtsson J, Hartmann M, Unterseher M, Vaishampayan P, Abarenkov K, Durso L, Bik EM, Garey JR, Eriksson KM, Nilsson RH. 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(6-7):407-412.
DOI: 10.1016/j.resmic.2012.07.001
http://dx.doi.org/10.1016/j.resmic.2012.07.001

Brief note

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

Johan Bengtsson[a,b,*], Martin Hartmann[c,d], Martin Unterseher[e], Parag Vaishampayan[f], Kessy Abarenkov[g], Lisa Durso[h], Elisabeth M. Bik[i], James R. Garey[j], K. Martin Eriksson[b], R. Henrik Nilsson[b,k]

[a] Institute of Neuroscience and Physiology, The Sahlgrenska Academy, University of Gothenburg, Medicinaregatan 11, Box 434, 405 30 Gothenburg, Sweden; johan@microbiology.se

[b] Department of Biological and Environmental Sciences, University of Gothenburg, Box 461, 405 30 Göteborg, Sweden

[c] Molecular Ecology, Agroscope Reckenholz-Tänikon Research Station ART, Reckenholzstrasse 191, 8046 Zurich, Switzerland; martin.hartmann@microbiome.ch

[d] Soil Sciences, Swiss Federal Research Institute WSL, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

[e] Institute of Botany and Landscape Ecology, Ernst-Moritz-Arndt-University Greifswald, Grimmer Strasse 88, D-17487 Greifswald, Germany; martin.unterseher@uni-greifswald.de

[f] Biotechnology and Planetary Protection Group, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA; Parag.A.Vaishampayan@jpl.nasa.gov

[g] Natural History Museum, University of Tartu, 46 Vanemuise St., 51014 Tartu, Estonia; kessy@ut.ee

[h] USDA, ARS, Agroecosystem Management Research Unit, 137 Keim Hall UNL-East Campus, Lincoln, NE 68583, USA; Lisa.Durso@ars.usda.gov

[i] Department of Microbiology & Immunology, Stanford School of Medicine, Fairchild Science Building, 299 Campus Drive, Stanford, CA 94305, USA; eliesbik@stanford.edu

[j] Department of Cell Biology, Microbiology and Molecular Biology, University of South Florida, 4202 E. Fowler Ave. ISA 2015,Tampa, FL 33620, USA; garey@usf.edu

[k] Department of Botany, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[*] Correspondence and reprints: johan@microbiology.se

Abstract

Metagenomic libraries represent subsamples of the total DNA found at a study site and offer unprecedented opportunities to study ecological and functional aspects of microbial communities. To examine the depth of a community sequencing effort, rarefaction analysis of the ribosomal small sub-unit (*SSU/16S/18S*) gene in the metagenome is usually performed. The fragmentary, non-overlapping nature of *SSU* sequences in metagenomic libraries poses a problem to this analysis, however. We introduce a software package – Megraft – that grafts *SSU* fragments onto full-length *SSU* sequences, accounting for observed and unobserved variability, for accurate assessment of species richness and sequencing depth in metagenomics endeavors.

Key words: metagenomics; rarefaction; species richness analysis; pyrosequencing; software; ribosomal DNA; rDNA; *16S*; *18S*

1. Introduction

Metagenomics is the study of the full span of DNA sequences obtained from a given sample, typically environmental substrates such as soil, seawater, or intestinal mucus (Trevors and Masson, 2010). A great number of sequences is needed to obtain a representative sample of the metagenome, such that these efforts typically employ high-throughput next-generation sequencing (NGS) techniques such as 454 pyrosequencing (Margulies et al., 2005). Metagenomic data offer windows on diverse questions relating to the ecology and functional profiles of the underlying organism community, but two pursuits are common to most studies: the taxonomic affiliation of the species recovered and whether the sequencing effort was deep enough to characterize the community with reasonable accuracy. In environmental sequencing studies, both of these are typically addressed through the ribosomal small subunit (*SSU/16S/18S*) sequences in the dataset. These are clustered into operational taxonomic units (OTUs; Blaxter et al., 2005), representative sequences of which are subjected to, e.g., BLAST searches (Altschul et al., 1997) in the International Nucleotide Sequence Databases (Cochrane et al., 2011) or other reference databases for taxonomic annotation. The relative abundance of the OTUs can similarly be used to assess the sampling effort and taxon richness with rarefaction or other mathematically smoothed species accumulation curves (Unterseher et al., 2011).

This clustering-and-rarefaction approach is not ideal for NGS-derived metagenomes, however, since the NGS technologies produce sequences of limited length – 300-400 base-pairs (bp.) in the case of pyrosequencing and the new Ion Torrent release (http://www.iontorrent.com). The *SSU* gene, in contrast, is 1,500-2,000 bp. long, such that metagenomes will contain random *SSU* fragments from different, disjunct parts of the gene. In these cases, non-overlapping *SSU* sequence fragments from the same species will not cluster together but rather form separate OTUs. To inflate datasets with such artificial OTUs is detrimental to subsequent analyses of sequencing depth and species richness. To address this problem, we introduce an open source software package – Megraft – that grafts *SSU* fragments onto full-length *SSU* sequences, accounting for observed as well as unobserved sequence variability. Using published prokaryotic and eukaryotic datasets from environmental sampling efforts as reference, we show that Megraft enables diversity assessments that are very close to the published ones and that are markedly better than the corresponding results from as-is clustering-and-rarefaction approaches.

2. Materials and Methods

Megraft is written in Perl and released under the GNU-GPL open source software license for UNIX-type operating systems, including MacOS X and Linux (Supplementary Item 1; http://microbiology.se/software/megraft/). It expects a query dataset of *SSU* sequences in the FASTA format (Pearson and Lipman, 1988), which can be extracted from metagenomes using, e.g., Metaxa (Bengtsson et al., 2011). The entries are processed sequentially (Fig. 1). Any highly similar query sequences with a significant overlap as established by Megraft through GramCluster (Russel et al., 2010) are treated as a single entry during the analysis and are then expanded back into their original number in the output FASTA file. Drawing from Metaxa (Bengtsson et al., 2011) and V-RevComp (Hartmann et al., 2011), Megraft uses HMMER version 3 (Eddy, 2011) and the hidden Markov models (HMMs) of Hartmann et al. (2010) to assess what part of the *SSU* each query sequence stems from. The *SSU* is composed of ten (semi-)conserved regions and eight (eukaryotes) or nine (prokaryotes) intercalary, hypervariable "V" regions (Hartmann et al., 2010). The HMMs are positioned immediately upstream and downstream of each V region to allow an accurate view of the position and span of each query sequence. A local BLAST search is then performed against a bundled copy of the full-length *SSU* sequences of the non-redundant SILVA reference database release 108 (Pruesse et al., 2007). The query and the closest reference sequence are compared for similarity in all shared conserved and hypervariable regions; the two sequences could, for example, be 99% similar in a conserved region but only 93% similar in the adjacent V region. Three increasingly sophisticated options are available to the Megraft user: 1) *Proxy mode:* The closest BLAST match is used as a proxy for the query sequence; 2) *Insert mode:* The closest BLAST match is used as reference. The region in the reference sequence corresponding to the query sequence is cut out and replaced with the query sequence; and 3) *Insert-differential-introduce mode*: As mode 2, with the addition that variation (mutations) is introduced along the full length of the reference sequence following the Jukes-Cantor model of nucleotide evolution. The degree of variation introduced is proportional to the distance between the query sequence and the reference sequence; the more dissimilar the query sequence is with respect to the closest reference sequence, the more variation introduced into the new sequence. The distribution of conserved and hypervariable regions across the *SSU* is taken into account in this process, such that variation is more likely to be added in hypervariable regions than in conserved regions and at a magnitude proportional to the distance to the reference sequence of each respective region. With each option, the end product is a FASTA file of full-length, partly artificial sequences to serve as basis for clustering, sequencing depth analysis, and assessment of species richness.

We evaluated the software using two different approaches. For the first one we used eight

published studies that featured Sanger-derived, long or near-full-length environmental *SSU* sequences, clustering, and rarefaction/species richness analysis (Table 1). Six of these studied prokaryotic, and two eukaryotic, microbial communities from diverse habitats such as forest soil, human gut, and a spacecraft assembly facility. The number of sequences in the datasets ranged from 198 to 3,834 (average: 1,423). To simulate the *SSU* component of pyrosequencing/Ion Torrent-derived metagenomes, a random continuous 350 bp. stretch of each *SSU* sequence in each published dataset was selected to form eight new datasets. These "daughter datasets" were identical to their published "mother dataset" in all regards except for the sequence length (Supplementary Item 2). The daughter datasets were run through Megraft using all three modes. Each pair of mother dataset and publication was examined for clustering settings, and GramCluster was tweaked to mimic the clustering conditions used by the respective original authors. The same settings were then used to cluster the respective daughter datasets as processed by Megraft in all three modes. The resulting lists of OTUs and their relative abundance were used for richness analysis with mathematically smoothed accumulation curves as described by Unterseher et al. (2011; Supplementary Item 3). For comparison, the raw daughter datasets – not processed by Megraft – were used to mimic the conventional clustering-and-rarefaction approach as applied on raw metagenome sequences. We similarly ran all mother datasets through the rarefaction process to verify that we could reproduce the rarefaction results of the underlying published studies. As a second approach to evaluate the performance of Megraft, we employed a simulated metagenome of 100,000 sequences of 350 bp. random nucleotide data generated through the EMBOSS 6.2.0 suite (Rice et al., 2000). A total of 1,522 full-length archaeal, bacterial, and eukaryote *SSU* sequences, forming 227 OTUs of different sizes, were downloaded from SILVA and subjected to clustering and rarefaction. A random 350 bp. stretch of each sequence was then chosen and added to the simulated metagenome. We used Metaxa 1.1 to identify and extract the SSU fragments from the metagenome. The extracted fragments were then subjected to clustering and rarefaction in four configurations: unprocessed as well as run through the three modes of Megraft.

3. Results and Discussion

The rarefaction results of the six prokaryote studies are shown in Fig. 2, with the results from the two eukaryotic studies given in Supplementary Item 3. Using the unaltered mother datasets, we were able to recapture the chief species richness results of all studies, which were used as baseline for further comparison. We found the consistently least accurate rarefaction approach for metagenomics datasets to be the one based on the unprocessed daughter datasets; in both eukaryote datasets, and in four of the six prokaryote datasets, the corresponding curve was higher by approximately a factor of two (Fig. 2; Supplementary Item 3). At the other end of the spectrum, the most sophisticated mode of Megraft – *Insert-differential-introduce* – was found to be the best approximation of the original authors' analysis in six of the eight datasets. In five of these, it was very close to, or nearly indistinguishable from, the published one (Supplementary Item 3). At its worst, the *Insert-differential-introduce* mode underestimated the diversity by a factor of 1.5 in the eukaryotic dataset of Wu et al. (2009). The two remaining modes of Megraft produced increasingly better results along a progression from the least sophisticated one – *Proxy mode* – to the semi-advanced *Insert mode*. A similar observation was made for the *SSU* fragments extracted from the simulated metagenome: the three Megraft modes gave rise to rarefaction curves hovering very closely around the results from the full-length dataset, whereas the unprocessed 350 bp. dataset gave rise to dramatic overestimations of the underlying diversity (Supplementary Item 4). These observations jointly suggest that it is crucial to account for both the fragmentary nature of metagenomic *SSU* sequences and for the amount of variation *not directly observed* in those sequences for lack of long enough sequence reads. Even so, Megraft should be seen as an improvement over existing approaches rather than as a panacea for pyrosequencing metagenomics rarefaction. Only in one of the eight datasets (Sunagawa et al., 2010) did Megraft produce results that were all but *identical* to the published data. Megraft will perform better the richer the database of reference sequences, such that suboptimal performance can be expected for datasets with a large proportion of previously unsequenced lineages. The eukaryote *SSU* reference datasets available lag behind the prokaryotic ones in terms of taxon and sequence sampling (cf. Pruesse et al., 2007), suggesting that Megraft may perform better on prokaryote datasets. Indeed, this may explain the somewhat larger deviation found for the eukaryotic dataset of Wu et al. (2009).

Megraft relies on HMMs that were tailored over conserved segments to the 5' and 3' ends of each of the hypervariable V regions of the *SSU*. There is little room for additional robust HMMs along the *SSU*, since the remaining non-hypervariable regions are less well conserved and thus not very suitable for robust HMM construction (Hartmann et al., 2010). This means that Megraft handles *SSU* fragments down to about 200 bp. in length; sequences shorter than that could, at least in theory, fit

166 *between* two HMMs, thus escaping detection as *SSU* sequences. For this reason, Megraft is not
167 expected to do a robust job with Illumina sequences, which at present typically are shorter than 200 bp.
168 This also means that Megraft works equally well for all sequencing technologies that produce
169 sequences of 200 bp. or longer, including pyrosequencing, Ion Torrent, and regular Sanger sequencing.
170 Similarly, although metagenomics represents the most obvious field of application for Megraft, it could
171 equally well be used in, e.g., amplicon- and data mining-oriented efforts. Megraft shares some
172 similarities with EMIRGE (Miller et al., 2011), which is a powerful iterative software package for
173 reconstruction of full-length ribosomal genes from paired-end Illumina sequences. However, EMIRGE
174 is not oriented towards rarefaction and does not readily process sequences from sequencing
175 technologies other than Illumina. Megraft, in contrast, has a strong focus on rarefaction, and it is
176 tailored for sequences longer than ~200 bp. These observations place EMIRGE and Megraft alongside
177 one another in that they seek to solve a roughly similar problem, but for different sequencing
178 technologies.
179 Sequence quality control is a crucial element in NGS-based efforts; failure to account for
180 incorrect or substandard sequences may have far-reaching negative ramifications for the downstream
181 results. In particular, the pyrosequencing technology struggles with base-calling in regions rich in
182 homopolymers, often giving rise to a range of artificial sequences differing only in these regions (Huse
183 et al., 2007; Balzer et al., 2011). Programs for denoising NGS datasets are readily available for
184 amplicon-based efforts (Quince et al., 2011) as well as for genome-centered studies (Ilie et al., 2011;
185 Salmela & Schröder, 2011). Such datasets typically contain large or very large numbers of sequences
186 from the same gene or region. Here, sequence templates as well as deviations from those templates can
187 be established, and the latter can be compared and corrected according to the former. Metagenomes
188 tend to be more challenging to denoise due to the potentially very low number of sequences
189 representing each marker, and regular denoising software can typically not be used for metagenomes.
190 Progress in sequence preprocessing, filtering, and read-score trimming has recently been made and
191 released as free software (Schmieder & Edwards, 2011), however, offering some degree of sequence
192 quality control also for metagenomes. Sequence quality control measures should be undertaken before
193 running Megraft. We nevertheless offer a level of sequence quality control in that Megraft features a
194 clustering step to average out minor differences among sequences. We use GramCluster for this step, a
195 program written with NGS datasets in mind to be robust against sequencing errors (Russell et al.,
196 2010). For instance, overlapping sequences that only differ in homopolymer-rich regions are very
197 likely to be considered as representing a single OTU in the clustering step, providing some degree of
198 protecting against overestimation of diversity due to technical artefacts.

About 250 MB of free computer memory is needed to run Megraft. It took 31 minutes to run the 3,843-sequence dataset of Eckburg et al. (2005) on a four-core 2.3 GHz Linux machine through the full Megraft pipeline (including the HMMER and BLAST steps; Figure 1); the same dataset took 280 minutes to process on a legacy one-core 2.0 GHz MacBook Pro. The time of execution of Megraft has a roughly linear relationship with the number of query (*SSU*) sequences, such that a doubling of the number of *SSU* sequences will double the runtime. The clustering step will however lead to shorter runtimes for datasets with two or more overlapping, conspecific sequences in that only one sequence in each cluster produced will be run through the BLAST step. In conclusion, we present an open source software tool to facilitate the analysis of sequencing depth and subsequent assessment of species richness based on the ribosomal *SSU* gene sequences in metagenomic datasets. The software produces artificial sequences that should not be used for other purposes than analyses of the rarefaction type, but using the sequences for precisely that will give rise to far more accurate views of the underlying sequencing depth and species richness than any comparable attempt with the raw *SSU* fragments retrieved directly from the metagenome in question.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Balzer, S., Malde, K., Jonassen, I. 2011. Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 27, i304-i309.

Bengtsson, J., Eriksson, K.M., 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. Anton Leeuw. Int. J. G. 100, 471-475.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., Abebe, E., 2005. Defining operational taxonomic units using DNA barcode data. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1935–1943.

Cochrane, G., Karsch-Mizrachi, I., Nakamura, Y., 2011. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 39(S1), D15-D18.

Durso, L.M., Harhay, G.P., Smith, T.P.L., et al., 2010. Animal-to-animal variation in fecal microbial diversity among beef cattle. Appl. Environ. Microbiol. 76, 4858-4862.

Eckburg, P.D., Bik, E.M., Bernstein, C.N., et al., 2005. Diversity of the human intestinal microbial flora. Science 308, 1635-1638.

Eddy, S.R., 2011. Accelerated profile HMM searches. PLoS Computational Biology 7: e1002195.

Hartmann, M., Howes, C.G., Abarenkov, K., Mohn, W.W., Nilsson, R.H., 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. J. Microbiol. Meth. 83, 250-253.

Hartmann, M., Howes, C.G., Veldre, V., et al., 2011. V-RevComp: Automated high-throughput

247 detection of reverse complementary 16S ribosomal RNA gene sequences in large environmental and
248 taxonomic datasets. FEMS Microbiol. Lett. 319, 140-145.
249
250 Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., Welch, D.M. 2007. Accuracy and quality of
251 massively parallel DNA pyrosequencing. Genome Biol. 8. R143.
252
253 Ilie, L., Fazayeli, F., Ilie, S. 2011. HiTEC: accurate error correction in high-throughput sequencing
254 data. Bioinformatics 27, 295-302.
255
256 Jangid, K., Williams, M.A., Franzluebbers, A.J., Schmidt, T.M., Coleman, D.C., Whitman, W.B.,
257 2011. Land-use history has a stronger impact on soil microbial community composition than
258 aboveground vegetation and soil properties. Soil Biol. Biochem. 43, 2184-2193.
259
260 Margulies, M., Egholm, M., Altman, W.E. et al., 2005. Genome sequencing in microfabricated high-
261 density picolitre reactors. Nature 437, 376-380.
262
263 Miller, C.S., Baker, B.J., Thomas, B.C., Singer, S.W., Banfield, J.F., 2011. EMIRGE: reconstruction of
264 full-length ribosomal genes from microbial community short read sequencing data. Genome Biol. 12,
265 R44.
266
267 Mohamed, D.J., Martiny, J.B.H., 2011. Patterns of fungal diversity and composition along a salinity
268 gradient. ISME J. 5, 379-388.
269
270 Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. P. Natl. Acad.
271 Sci. USA 85, 2444-2448.
272
273 Pruesse, E.C., Quast, C., Knittel, K., Fuchs, B., Ludwig, W., Peplies, J., Glöckner, F.O., 2007. SILVA:
274 a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data
275 compatible with ARB. Nucleic Acids Res. 35, 7188-7196.
276
277 Quince, C., Lanzén, A., Davenport, R. J., Turnbaugh, P. J. 2011. Removing noise from pyrosequenced
278 amplicons. BMC Bioinf. 12, 38.
279

Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276-277.

Russel, D.J., Way, S.F., Benson, A.K., Sayood, K., 2010. A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. BMC Bioinf. 11, 601.

Salmela, L., Schröder, J. 2011. Correcting errors in short reads by multiple alignments. Bioinformatics 27, 1455-1461.

Schmieder, R., Edwards, R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-864.

Shivaji, S., Pratibha, M.S., Sailaja, B., et al., 2011. Bacterial diversity of soil in the vicinity of Pindari glacier, Himalayan mountain ranges, India, using culturable bacteria and soil 16S rRNA gene clones. Extremophiles 15, 1-22.

Sunagawa, S., Woodley, C.W., Medina, M., 2010. Threatened corals provide underexplored microbial habitats. PLoS ONE 5, e9554.

Trevors, J.T., Masson, L., 2010. DNA technologies: what's next applied to microbiology research? Anton Leeuw. Int. J. G. 98, 249-262.

Unterseher, M., Jumpponen, A., Öpik, M., Tedersoo, L., Moora, M., Dormann, C.F., Schnittler, M., 2011. Species abundance distributions and richness estimations in fungal metagenomics – lessons learned from community ecology. Mol. Ecol. 20, 275-285.

Vaishampayan, P., Osman, S., Andersen, G., Venkateswaran, K., 2010. High-density 16S microarray and clone library–based microbial community composition of the Phoenix spacecraft assembly clean room. Astrobiology 10, 499-508.

Wu, T., Ayres, E., Li, G., Bardgett, R.D., Wall, D.H., Garey, J.R., 2009. Molecular profiling of soil animal diversity in natural ecosystems: Incongruence of molecular and morphological results. Soil Biol. Biochem. 41, 849-857.

313

Figure Legends

Figure 1. Schematic overview of the Megraft execution process. Rectangles represent operations done by Megraft. The input file is shown at the top, and the output file is indicated at the bottom. Remaining tilted rectangles represent data items used by Megraft. Dashed lines indicate use of data from a previous analysis step.

Figure 2. Rarefaction results of the six prokaryote studies used to evaluate Megraft. In each subfigure, the black dotted line indicates a 1:1 relationship between the number of sequences and OTUs. The black solid line indicates the results of the rarefaction of the original, full-length *SSU* sequences. The orange line indicates the rarefaction results of clustering of the 350-bp. fragments without any regrafting or other use of Megraft. The green, red, and blue dotted lines indicate the rarefaction performance of the three increasingly sophisticated Megraft modes *Proxy*, *Insert*, and *Insert-differential-introduce*, respectively.

Tables

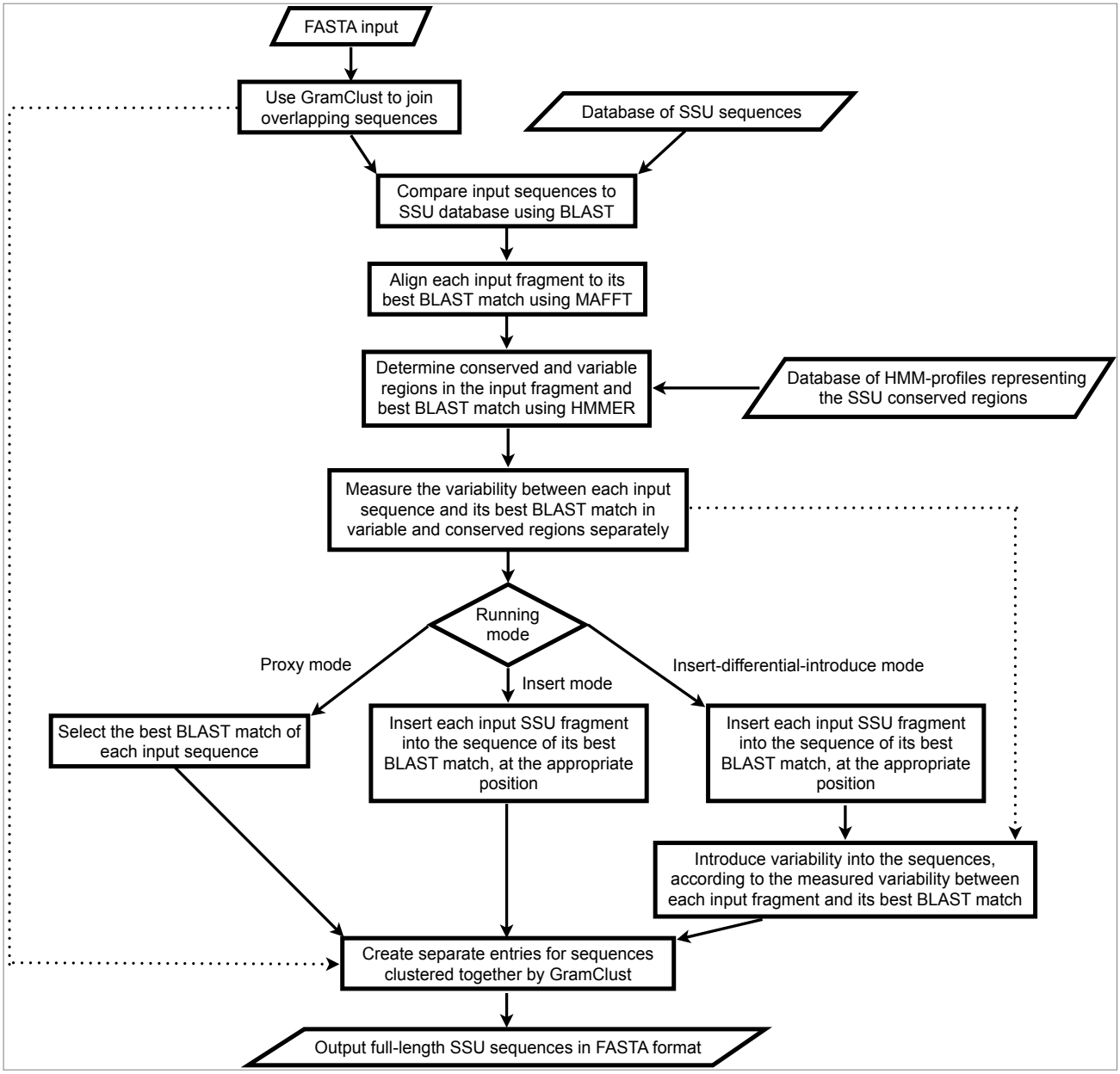
Table 1. Summary of the eight reference studies used to evaluate the performance of Megraft.

Study	Sample name in study	#seq	Sequencing strategy	SSU coverage	#OTUs	Domain	Locality/substrate	Country
Durso et al., 2010	Animal 2	2,084	bidirectional	near-complete 16S	416	Prokaryote	Bovine gut	USA
Eckburg et al., 2005	Subject C	3,834	bidirectional	near-complete 16S	206	Prokaryote	Human distal gut	USA
Jangid et al., 2011	CF	538	single run	~800 bp. 16S	272	Prokaryote	Forest soil	USA
Mohamed et al., 2011	Freshwater	370	single run	~700 bp. 18S	80	Eukaryote	Freshwater marsh	USA
Shivaji et al., 2011	P4S	198	bidirectional	~1000 bp 16S	44	Prokaryote	Himalayan glacier	India
Sunagawa et al., 2010	SGUS	943	bidirectional	near-complete 16S	178	Prokaryote	Caribbean coral	Panama
Vaishampayan et al., 2010	PHX-A	777	bidirectional	near-complete 16S	100	Prokaryote	Spacecraft assembly facility	USA
Wu et al., 2009	TK	2,641	unidirectional	519 bp. 18S	305	Eukaryote	Alaskan tundra	USA

330 **Figures**

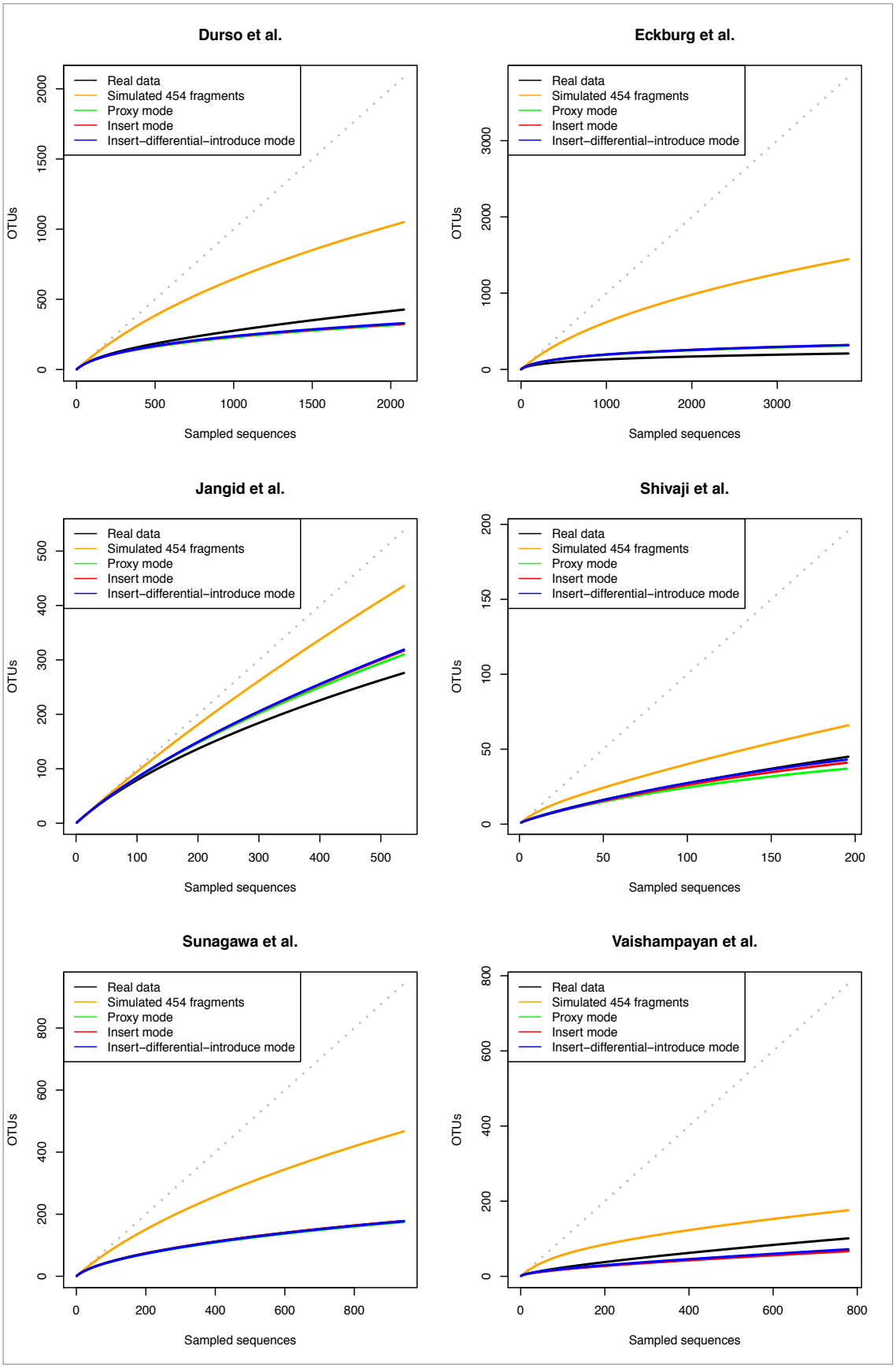
331

332 **Figure 1**



333

334 **Figure 2**



335

Supplementary Items

Supplementary Item 1. The software package together with its documentation and a test dataset. In addition, the user will have to install NCBI-BLAST, HMMER, and GramCluster; detailed installation instructions are provided in the documentation.
[For review purposes, this file is available at
http://microbiology.se/publ/megraft/Supplementary_Item_1.pdf]

Supplementary Item 2. The daughter datasets generated from the original eight published datasets. The sequences are given in the FASTA format. Each sequence is a randomly chosen continuous 350 bp. stretch of the original, published sequence.
[For review purposes, this file is available at
http://microbiology.se/publ/megraft/Supplementary_Item_2.tar.gz]

Supplementary Item 3. Rarefaction results from the six prokaryotic, and two eukaryotic, reference studies. Annotated R code to compute species accumulation curves and estimators of species richness from the Megraft-data with the R 'vegan' package.
[For review purposes, this file is available at
http://microbiology.se/publ/megraft/Supplementary_Item_3.tar.gz]

Supplementary Item 4. Input and output files and data items from the analysis of the simulated metagenome. The *SSU* component of the metagenome is provided, but the sizable metagenome itself is not. It can be found at http://microbiology.se/publ/megraft/simulated_metagenome.tar.gz
[For review purposes, this file is available at
http://microbiology.se/publ/megraft/Supplementary_Item_4.tar.gz]