

MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes

Pierre Pericard, Yoann Dufresne, Loïc Couderc, Samuel Blanquart, Hélène Touzet

▶ **To cite this version:**

Pierre Pericard, Yoann Dufresne, Loïc Couderc, Samuel Blanquart, Hélène Touzet. MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes. Bioinformatics, 2017, 10.1093/bioinformatics/btx644 . hal-01646297v1

HAL Id: hal-01646297

https://inria.hal.science/hal-01646297v1

Submitted on 16 Jan 2018 (v1), last revised 1 Oct 2019 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

MATAM: reconstruction of phylogenetic marker genes from short sequencing reads in metagenomes

Pierre Pericard [1,2]*, Yoann Dufresne [1,2], Loïc Couderc [3,1], Samuel Blanquart [2,1] and Hélène Touzet [1,2]*

[1]CRIStAL (UMR CNRS 9189, Université Lille 1), [2]Inria Lille Nord-Europe, and [3]Bilille, 59650 Villeneuve d'Ascq, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Advances in the sequencing of uncultured environmental samples, dubbed metagenomics, raise a growing need for accurate taxonomic assignment. Accurate identification of organisms present within a community is essential to understanding even the most elementary ecosystems. However, current high-throughput sequencing technologies generate short reads which partially cover full-length marker genes and this poses difficult bioinformatic challenges for taxonomy identification at high resolution.
Results: We designed MATAM, a software dedicated to the fast and accurate targeted assembly of short reads sequenced from a genomic marker of interest. The method implements a stepwise process based on construction and analysis of a read overlap graph. It is applied to the assembly of 16S rRNA markers and is validated on simulated, synthetic and genuine metagenomes. We show that MATAM outperforms other available methods in terms of low error rates and recovered fractions and is suitable to provide improved assemblies for precise taxonomic assignments.
Availability: https://github.com/bonsai-team/matam
Contact: pierre.pericard@gmail.com, helene.touzet@univ-lille1.fr
Supplementary information: Supplementary material is available at *Bioinformatics* online.

1 Introduction

Shotgun metagenomic sequencing provides an unprecedented opportunity to study uncultured microbial samples, with multiple applications ranging from the human microbiome to soil or marine samples, for which the vast majority of microbial diversity remains unknown (Locey and Lennon, 2016).

A major goal of metagenomic studies is to characterize microbial diversity and ecological structure. This is often achieved by focusing on one of several phylogenetic marker genes (Liu *et al.*, 2011; Segata *et al.*, 2012), that are ubiquitous in the taxonomic range of interest and exhibit variable discriminative regions. For bacterial communities, the gold standard marker is the 16S ribosomal RNA (rRNA, ∼1500bp avg. length), for which millions of sequences are available in curated reference databases, such as Silva (Quast *et al.*, 2013), RDP (Cole *et al.*, 2014) or GreenGenes (DeSantis *et al.*, 2006). Traditionnal approaches such as amplicon sequencing are limited to the analysis of small portions of the marker sequences. This leads to strong technological limitations for organisms identification at sufficiently precise taxonomic levels, typically beyond genus (Poretsky *et al.*, 2014). To assign marker sequences to species, or even strains, we need to be able to recover full length rRNA with less than a few errors per kilobase. Metagenomic assemblers are not suitable for this task, because they are optimized to deal with whole genomes, and struggle to differentiate between very similar sequences (Sczyrba *et al.*, 2017). To this respect, marker-oriented methods such as EMIRGE (Miller *et al.*, 2011) and REAGO (Yuan *et al.*, 2015) were recently developed in order to assemble metagenomic read subsets into full length 16S rRNA contigs, thus aiming to improve the taxonomic assignment accuracy of environmental samples. EMIRGE uses a Bayesian approach to iteratively reconstruct 16S rRNA full length sequences. REAGO identifies rRNA reads using Infernal (Nawrocki *et al.*, 2009), and then constructs an overlap graph by searching for exact overlaps between reads using a suffix/prefix

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

array. However, such tools still show some limitations in terms of recovery error rates as well as dealing with low abundance species.

In this work, we present MATAM, a new approach based on the construction and exploitation of an overlap graph, carefully designed to minimize the error rate and the risk of chimera formation. MATAM was validated on both simulated and genuine sequencing data and showed excellent results.

2 Methods

2.1 Overview of MATAM

The MATAM (Mapping-Assisted Targeted-Assembly for Metagenomics) pipeline takes as input a set of shotgun metagenomics short reads and a reference database containing the largest possible set of sequences from a given target marker gene. MATAM identifies reads originating from that marker, and assembles nearly full length sequences of it. It is composed of four major steps illustrated in Figure 1. Although this method should work for any conserved and widely surveyed gene, we will focus on the 16S rRNA for the remainder of the article. Additional technical details and parameters are available in the Supplementary Methods.

2.2 Reference database construction

The availability of a reference database for the marker gene is an essential feature of the method, because it allows us to model the target sequences. For applications to 16S rRNA assembly, MATAM utilizes Silva 128 SSU Ref NR database (Quast *et al.*, 2013). From this reference database that we denote as *complete*, we also build a *clustered* reference database, that provides a coarse-grained representation of the taxonomic space. For that task, we use Sumaclust (Mercier *et al.*, 2013; Kopylova *et al.*, 2016) (http://metabarcoding.org/sumaclust) using a 95% identity threshold.

2.3 rRNA reads identification and mapping

In the first step, reads are mapped against the clustered reference database using SortMeRNA (Kopylova *et al.*, 2012, 2014). This step allows to quickly sort out 16S rRNA reads from the whole set of reads, providing high quality alignments. For each read, we keep up to ten best alignments against the reference database. Moreover, this mapping step yields a broad classification of the 16S rRNA reads. Indeed, reads coming from distantly related species are aligned against their respective closest known references, which nest in distant lineages of the taxonomy, while reads from closely related species are aligned against closely related references.

2.4 Construction of the overlap graph

The identified 16S rRNA reads are then organized into an *overlap graph* defined as follows: graph nodes are reads, and an undirected edge connects two nodes if the two reads overlap with a sufficient length and with a sufficient identity to assert that they originated from a common sampled taxon. The standard approach to build such an overlap graph requires comparison of each read with each other, which is time-consuming. Here, we use alignment information to sort through candidate read pairs in a very efficient manner. For each pairing, we consider only reads that share alignments with at least one common reference sequence and for which the alignments are overlapping with a minimal length ℓ and a minimum identity percentage m.

2.5 Extracting contigs from the overlap graph

The overlap graph reveals some general trends. While it exhibits highly connected subgraphs, it also displays disjoint paths (see Figure 1 for an example). We simplify the graph by performing a breadth first traversal starting from a random node to annotate the nodes with their depth. All nodes with equal depth that are connected in a single connected component are collapsed into a single *compressed node* and outgoing edges are merged into a *compressed edge*. Low support compressed nodes containing a single read, and compressed edges representing a single overlap are removed. The resulting graph, called the *compressed graph*, is several order of magnitude smaller than the initial overlap graph. We partition this graph in three categories of subgraphs: *hubs*, that are nodes with an degree strictly greater than two, *specific paths* that are sequences of nodes of degree two or one, and *singletons* that are non-connected nodes. Intuitively, hubs correspond to the highly connected subgraphs in the overlap graph, and are likely to contain mainly reads coming from conserved regions shared in many species, thus overlaping without error even for distantly related taxa. Specific paths tend to contain reads originating from variable regions of the 16S gene, that are specific to one or few closely related species. For each subgraph in the compressed graph (hubs, specific paths, singletons), we extract the underlying sets of reads and build an individual assembly using the genomic assembler SGA (Simpson and Durbin, 2012). As a result, we obtain one or more contigs for each subgraph.

2.6 Contigs scaffolding

We use a greedy algorithm to scaffold the contigs obtained in the previous step. The idea of this algorithm is first to align all contigs against the complete reference database, and then to cluster contigs according to the matching reference sequences to build a consensus scaffold. When doing so, a long contig with a unique alignment will be selected for scaffolding before a short contig exhibiting a large number of alignments. Such long contig can be assigned non-ambiguously to a single species, while the short contig with multiple matches rather corresponds to a conserved region of the marker and is used to fill in the blanks between the specific contigs. Finally, only scaffolds larger than 500bp are retained. The full details of this algorithm are given in Supplementary Methods, section 2.6.

2.7 Abundance estimation and taxonomic composition

The last step consists in estimating abundances by remapping the rRNA reads onto the scaffolds (see Supplementary Methods section 2.7), and assigning those scaffold to a taxon using the RDP classifier (Wang *et al.*, 2007). The estimated abundances and the taxonomic assignations are summarized in a Krona file (Ondov *et al.*, 2011) that allows users to easily visualize the estimated sample taxonomic composition.

3 Implementation

MATAM was implemented in Python 3, except for the overlap graph building and compression steps that were written in C++11 using the SeqAn library (Döring *et al.*, 2008), and is available *via* Docker and Conda. MATAM is distributed under the GNU Affero GPL v3.0 licence and the source code is freely available at the following URL: https://github.com/bonsai-team/matam. All MATAM runs presented in this article were performed using MATAM v0.9.9.

4 Results

MATAM performance was compared with those of two general-purpose metagenomic assemblers, SPAdes (Bankevich *et al.*, 2012; Nurk *et al.*, 2016) and MEGAHIT (Li *et al.*, 2015), as well as with two methods specialized in 16S rRNA assembly, EMIRGE (Miller *et al.*, 2011) and REAGO (Yuan *et al.*, 2015). The five tools were run on three different datasets, chosen for their complementarity and the possibility to validate the reconstructed candidate 16S rRNA

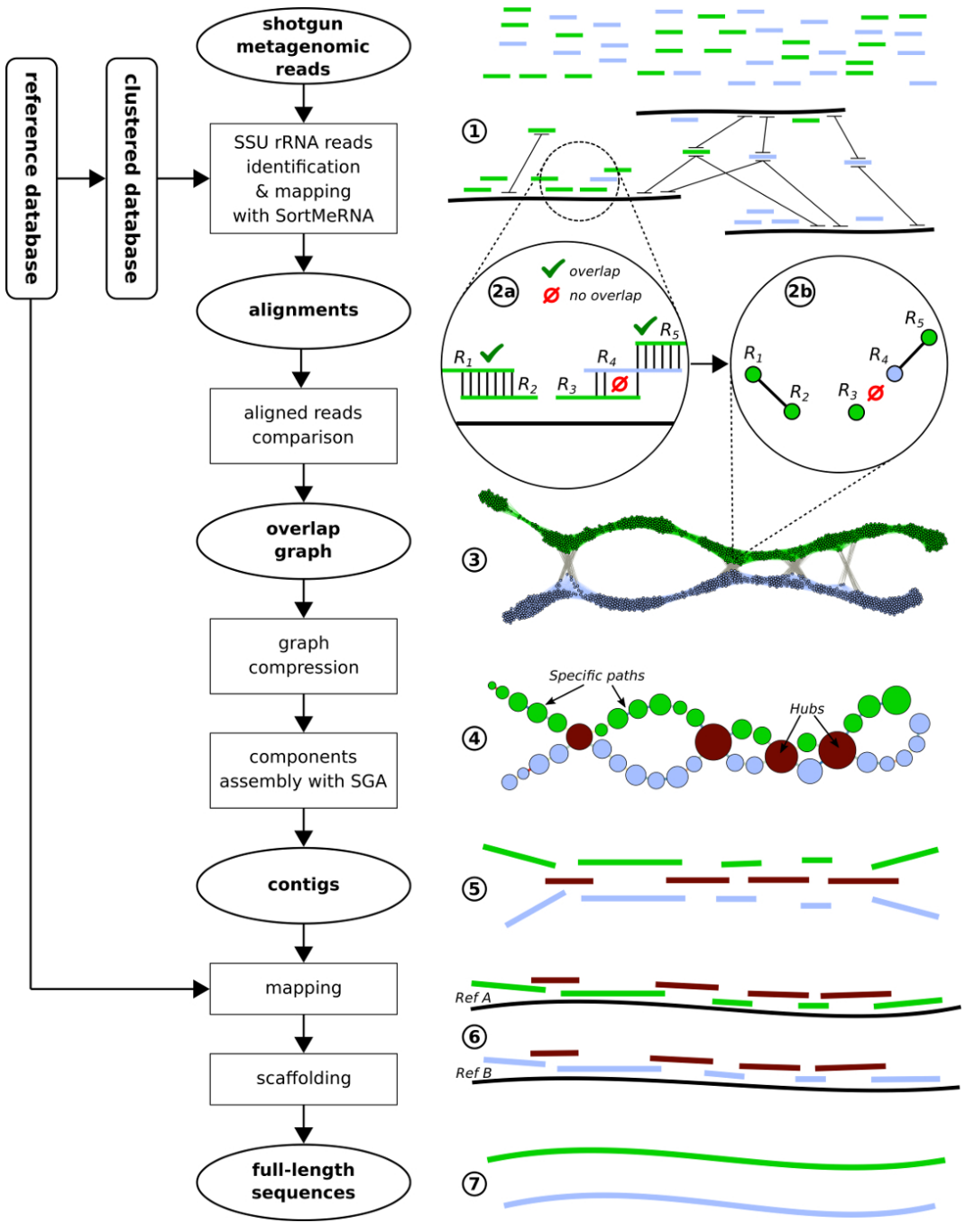


Fig. 1. MATAM overview. On the left, we describe the main steps of the pipeline. On the right, we illustrate those steps when the sample contains two species (the first one in blue, the latter one in green). Starting from shotgun metagenomic reads, (1) we first identify SSU rRNA reads and align them on up to 10 sequences from a clustered reference database. (2a) Reads alignments are compared between them to compute reads pairwise alignments. (2b) An overlap with 100% identity between two reads corresponds to an edge in the (3) read overlap graph. (4) Using a breadth-first search, the overlap graph is then simplified into a compressed graph and subgraphs (hubs, specific paths, singletons) are identified. (5) Reads from each subgraph are assembled into contigs with SGA. (6) Contigs are aligned on the complete reference database and alignments are selected using a greedy algorithm. (7) Contigs aligned on the same reference are then scaffolded into full-length sequences.

sequences: a simulated dataset (Mavromatis *et al.*, 2007), a synthetic microbial community (Shakya *et al.*, 2013), and two environmental samples from human gut and mouth providing amplicon based taxonomic assignments (The Human Microbiome Project Consortium, 2012). On all those Illumina datasets, MATAM overlap graph parameters were set to: minimal overlap length $\ell = 50$ nt and minimum identity $m = 100$ %. By doing so, we discard read pairs containing sequencing errors in their overlap, which is reasonnable when working with low-error quality-cleaned Illumina datasets.

SortMeRNA was used to extract 16S rRNA reads from these datasets before assembling them with SPAdes and MEGAHIT. Complete command-lines and parameters are available in the Supplementary Results.

In order to compare the five methods on a common ground, the same validation procedure was applied for all experiments. Only reconstructed

Table 1. Abundance estimation and taxonomic composition statistics for the simulated datasets. G is the total number of genera, TP (true positives) the number of correct genera, FP (false positives) the number of falsely identified genera, FN (false negatives) the number of missed genera.

		G	TP (% est. abd)	Pearson corr.	FP (% est. abd)	FN (% theo. abd)
50x	MATAM	87	81 (78.58)	0.999	6 (21.41)	2 (1.19)
	EMIRGE	79	60 (82.02)	0.874	19 (17.98)	23 (13.35)
20x	MATAM	87	82 (78.48)	0.999	5 (21.52)	1 (1.00)
	EMIRGE	74	67 (81.54)	0.981	7 (18.45)	16 (8.17)
10x	MATAM	85	81 (78.12)	0.999	4 (21.88)	2 (1.19)
	EMIRGE	61	59 (89.31)	0.854	2 (10.69)	24 (16.33)

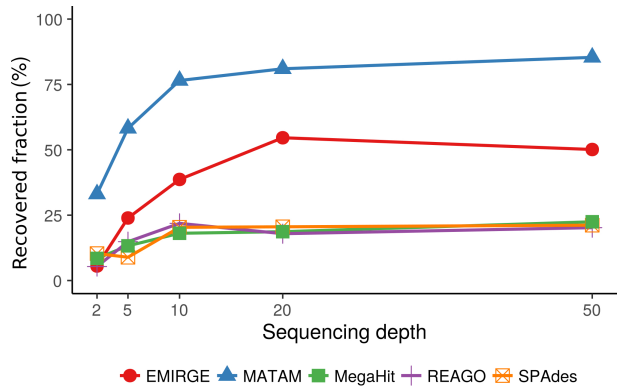


Fig. 2. Effect of sequencing depth on the assemblies recovered fractions for the simulated datasets.

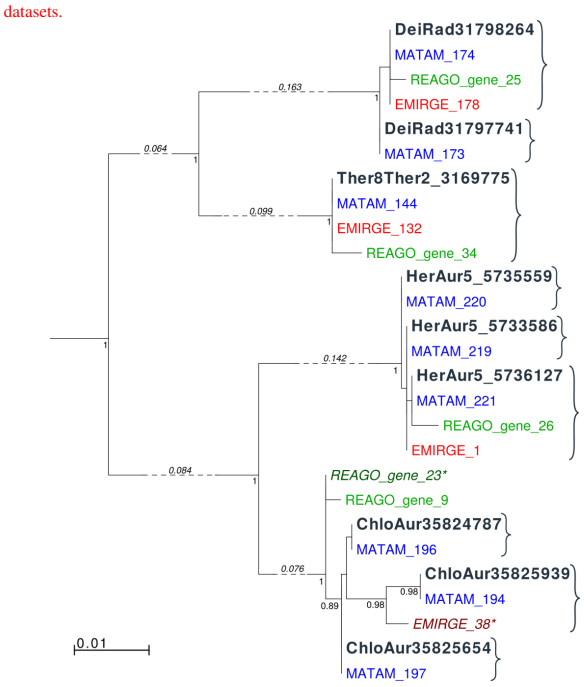


Fig. 3. Alignment of the reference sequences with the assembled contigs from the synthetic community shows MATAM ability to differentiate between very close sequences. MATAM, EMIRGE and REAGO contigs are shown respectively in blue, red and green. In a ideal setting, each software should produce contigs that cluster closely to each reference (black) sequence. Contigs followed by a star, and drawn in a darker color, were considered as chimeric by VSEARCH.

sequences with lengths exceeding 500bp were considered, and chimeric sequences were filtered out by the UCHIME algorithm (Edgar *et al.*, 2011) implemented in VSEARCH (Rognes *et al.*, 2016) and querying the Silva 128 SSU Ref Nr99 database. For each experiment, we indicate the proportion of chimeric contigs (*% chimeras*, which is the total size of all chimeric contigs divided by the assembly total size). All the following measures were then computed on the remaining assemblies. When the sequences present in the sample are actually known (see Sections 4.1 and 4.2), the assembly quality assessment was performed with MetaQuast (Mikheenko *et al.*, 2016) by aligning the contigs against the original sample sequences, and considering the following metrics: the *number of contigs* (#contigs), which is the total number of contigs of lengths greater than 500bp; the *total length* (TL), which is the total number of bases in the contigs; the *total aligned length* (TAL), which is the total number of aligned nucleotides in the contigs; the *recovered fraction* (RF), which stands for the proportion of nucleotides from the original sample sequences covered with contigs; the *error rate* (ER), which consists in the percentage of observed mismatches and indels with respect to the closest matched sequence in the original sample. Finaly, taxonomic assignments were carried out with the RDP Classifier (Wang *et al.*, 2007). The assemblies evaluation protocol, command-lines and parameters can be found in the Supplementary Results, section 4.2.

4.1 Simulated metagenomic datasets with varying sequencing depth

In the first experiment, we evaluated the ability of methods to correctly reconstruct the 16S rRNA sequences in the context of low sequencing depth. For that, we used a selection of 122 genomes from 83 genera providing a realistic taxomical diversity (Mavromatis *et al.*, 2007; Pignatelli and Moya, 2011), and that contains 287 distinct 16S rRNA copies. We generated five datasets with varying sequencing depths: 50x, 20x, 10x, 5x and 2x per genome. Illumina reads were simulated with the ART simulator (Huang *et al.*, 2012), using the HiSeq2500 built-in error profile, 101bp read length, and 250bp fragment length with a 30bp standard-deviation (SD). In this simulation, all species are equally distributed, which corresponds to the *high complexity community* introduced in Mavromatis *et al.* (2007). Simulation command-line and parameters can be found in the Supplementary Results, section 4.3.1.

Table 2 shows the results averaged over the five datasets (*mean* metrics and their respective standard deviation, SD). The complete results for all five datasets can be found in Supplementary Table 1. More than 99% of the MATAM sequences were aligned by MetaQuast to one of the 287 16S rRNA sequences from the initial sample (mean TAL/TL), while among other methods, this proportion reached at best 91%, with REAGO. Congruently, MATAM sequences obtained the lowest average error rate (ER=0.03%), which represents more than a ten-fold accuracy gain compared to the other assemblers, and a twenty-fold improvement over EMIRGE. Furthermore, EMIRGE sequences contained 0.5% of unknown nucleotides (Ns), bringing its effective ER above 1%. Additionally, MATAM recovered about thirty times less chimeras than REAGO and EMIRGE did.

For each of the five tools, we reported the recovered fraction (RF) with respect to increasing sequencing depth (Figure 2). MATAM recovered from 76% to 85% of the reference sequences for sequencing depths greater than 10x, while EMIRGE recovered less than 55% of the reference sequence, and the RF for other methods is lower than 22%. MATAM also achieved the best performance facing a low sequencing depth of 2x, reaching a RF of 33%, while RFs ranged between 5% and 10% with all other assemblers.

We also evaluated the abundance estimation and taxonomic composition accuracy for MATAM and EMIRGE (Table 1). Compared to the theoretical taxonomic composition of the community, MATAM reconstructed genera with a better accuracy than EMIRGE, missing only one or two genera representing about 1% of the total theoretical abundance.

Table 2. Results for the simulated dataset with varying sequencing depth. We provide averaged metrics for the five sequencing depths. ACL is the average contig length.

	Chimera (%)		TAL/TL (%)		ER (%)		Ns (%)		ACL	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
MATAM	1.28	0.55	99.3	0.2	0.03	0.02	0.00	0.00	1252	116.9
EMIRGE	36.89	9.42	79.9	11.6	0.62	0.16	0.55	0.36	1436	15.4
REAGO	42.11	10.36	91.5	0.8	0.31	0.13	0.00	0.00	1333	298.9
SPAdes	21.23	9.05	73.5	15.9	0.60	0.49	0.02	0.04	966	47.4
MEGAHIT	23.81	2.85	80.3	4.9	0.36	0.18	0.00	0.00	962	87.6

Table 3. Results for the synthetic community.

	Chimera (%)	#contigs	TL	TAL	RF (%)	ER (%)	Ns (%)	#recov. genera
MATAM	3.2	101	139220	130654	83.1	0.05	0	47/48
EMIRGE	17.4	82	117138	102856	50.7	0.17	1.12	44/48
REAGO	15.5	59	90269	81297	42.8	0.06	0	44/48
SPAdes	5.5	59	70229	59988	39.9	0.11	0.05	43/48
MEGAHIT	3.0	61	77251	68904	44.3	0.18	0	45/48

Table 4. Results for the gut and mouth HMP datasets. The column #classes indicates the total number of taxonomic classes found with RDP from the assemblies, with the number of these classes validated with the QIIME OTUs (in parentheses). The column #genera gives the same information at the genus level.

		Chimera (%)	#contigs	TL	#classes	#genera
SRS011405	MATAM	3.37%	218	187710	5 (4)	21 (17)
	EMIRGE	43.04%	273	393152	2 (2)	12 (8)
SRS016002	MATAM	4.92%	353	320748	13 (13)	31 (28)
	EMIRGE	46.01%	282	394087	12 (12)	25 (23)

Moreover the estimated abundance distribution for the correct genera was closer to the expected abundances (according to a Pearson correlation coefficient) than the one from EMIRGE. For both methods, and for all tested sequencing depths, about 10 to 20% of the total estimated abundance corresponds to genera not present in the community.

4.2 Synthetic archaeal and bacterial community

Inching toward more realistic applications, a second dataset provides Illumina reads extracted from a synthetic microbial community composed of 16 archaeal species from 12 genera, as well as 48 bacterial species from 36 genera (accession SRR606249; Shakya *et al.* (2013)). As emphasized by the authors, the selected organisms cover a wide range of environmental conditions and adaptation strategies. In contrast to the previous simulated dataset (Section 4.1), the proportion of each species in the sample is not uniform, which results in individual genome average sequencing depth varying from 6x to 318x. The number of 16S rRNA paralogs per genome appears also highly diverse, ranging from 1 to 10 copies per genome. Altogether, this dataset represents a total amount of 106 distinct 16S rRNA sequences with pairwise sequence identities ranging from 59.64% to 99.93%.

The organisms were sequenced on Illumina HighSeq2000, providing 109 million 101bp paired-end reads with an average fragment size of 250bp. We quality cleaned the reads using Prinseq Lite (Schmieder and Edwards, 2011), removed adapter sequences using Cutadapt (Martin, 2011), filtered out short reads (< 50bp), and obtained a total number of 67.6 million reads, which were analyzed with MATAM and EMIRGE. The uncleaned raw dataset was provided to REAGO, considering that the method could not handle reads with varying lengths. Finally, for SPAdes and MEGAHIT, the 16S rRNA reads were extracted from the cleaned dataset using SortMeRNA, which provided 108,560 16S rRNA reads to assemble. Cleaning and pre-processing command-lines and parameters for the synthetic community can be found in the Supplementary Results, section 4.3.2.

Results are shown in Table 3. Confirming the trends observed on the simulated dataset, MATAM is able to recover the highest number of sequences together with the highest RF (83%). Most importantly, with lower ER than achieved by the other tested methods, the MATAM assembly appears highly accurate. While EMIRGE is the second best approach in terms of RF, it also yields the greatest ER and Ns over all the compared tools. Moreover, a RDP classification of MATAM and EMIRGE sequences indicates that while MATAM missed one expected genus only, EMIRGE missed 4 genera out of 48.

Inspection of the MetaQuast alignments of the assemblies against the original 16S rRNAs revealed that all methods accurately assembled the genes sharing less than 90% sequence identity with their closest relatives within the sample. However, performances significantly dropped when attempting to assemble the closely related genes in the dataset. This especially concerned the paralogous 16S rRNA copies sharing around 99% sequence identity. Supplementary Table 2 (Supplementary Results, section 4.3.2) provides pairwise distances between sequences from a representative subset of four related species possessing one to

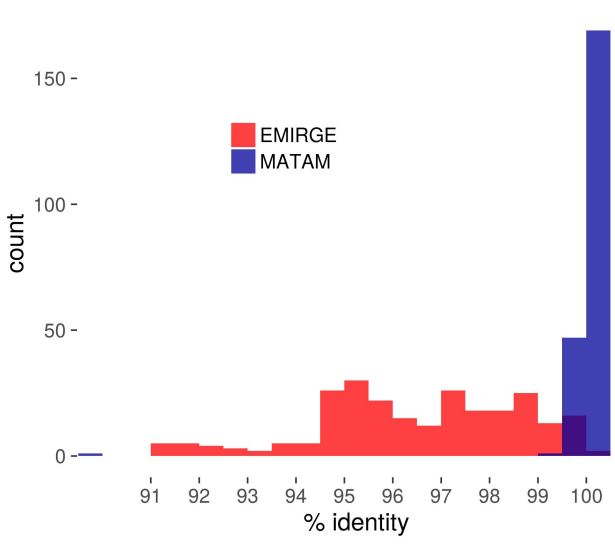


Fig. 4. Human gut sample SRS011405. % identity distribution of best matches against Silva 128 SSU Ref NR

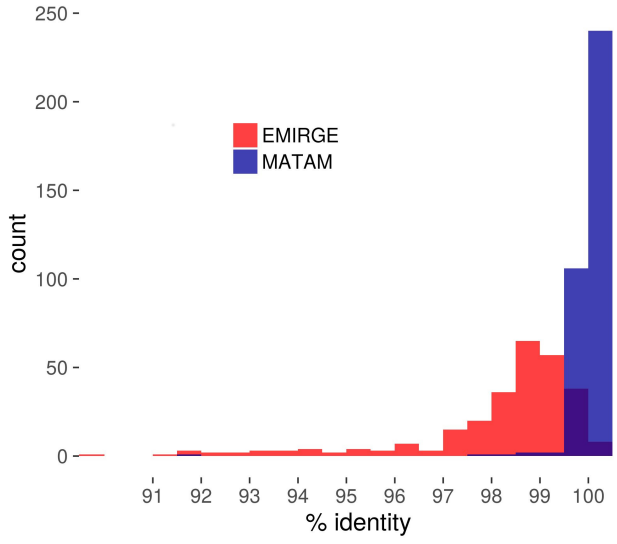


Fig. 5. Human mouth sample SRS016002. % identity distribution of best matches against Silva 128 SSU Ref NR

three such paralogous copies. Those 16S rRNAs and their corresponding assembled candidate sequences were selected for a phylogenetic tree reconstruction. The obtained tree (Figure 3) demonstrates that MATAM correctly assembled all the different paralogs with nearly no error, while EMIRGE and REAGO only managed to recover one candidate sequence per species. Thus, EMIRGE and REAGO merged into a single candidate sequence the reads issued from distinct paralogs, resulting in erroneous assemblies with high ER and underestimated RF. Indeed, each of the sequences assembled with REAGO, as well as one EMIRGE sequence over four, appear to cluster at a slight distance from their respective targeted paralogs. Those distances simply account for the methods reconstruction errors. Consistently, in two cases, the candidates assembled by EMIRGE and REAGO were identified as chimeras by VSEARCH.

4.3 Sensitivity to reference database completeness

A natural question is to evaluate MATAM sensitivity to the reference database completeness in the experiments of Sections 4.1 and 4.2. To evaluate this concern, we constructed two depleted databases by removing from Silva the closest relatives to the targeted sequences, sharing 99% identity over 90% alignment length (6 243 and 1 900 sequences removed for the simulated and synthetic community datasets, respectively). The results, presented in Supplementary Results, section 4.3.3, show that MATAM recovery performances are degraded, but still improves over those obtained with other tools.

4.4 Human Microbiome Project

Finally, we used two metagenomic samples from the Human Microbiome Project (gut: SRS011405, and mouth: SRS016002, The Human Microbiome Project Consortium (2012)) in order to validate MATAM on real metagenomic datasets sequenced from genuine environments. The reads were already quality cleaned and trimmed, and no additional filtering was performed. Hence, reads having different lengths, we were not able to run REAGO on these datasets. Results obtained with SPAdes and MEGAHIT using the following protocol appeared highly inaccurate and therefore, they are not further commented in this work. Thus, we only present the results obtained with EMIRGE and MATAM. Datasets availability, and additional details on the evaluation protocol can be found in Supplementary Results, section 4.3.4.

For these two datasets, the exact ground truth is unknown. Thus we could not perform the same validation procedure as in the two previous examples and we had to resort to alternative strategies. First, we took advantage of the availability of OTU sequences inferred through a QIIME analysis of the V1-V3 hypervariable regions for the same biological samples (available from the SRS accession numbers). We compared the assignments obtained from assemblies, calculated with RDP, with these of amplicon OTUs (Table 4). For both samples, MATAM identified more classes and genera than EMIRGE did, and most of these taxa were validated by the amplicon OTUs. Interestingly, we observed that in the two samples, three genera were recovered both by MATAM and EMIRGE, but not by the amplicon approach: *Odoribacter*, *Peptococcus*, and *Bergeyella*. Since some species from these genera are known to be adapted to the human gut and mouth environments, it is plausible that they were missed by the amplicon approach while being accurately recovered by MATAM and EMIRGE from the metagenomic samples.

Moreover, we evaluated assembly quality by aligning MATAM and EMIRGE sequences against the complete Silva 128 SSU Ref NR database, using BLAST. The rationale for this experiment is that most of the species in these human gut and mouth samples are possibly already known, and therefore should be found in Silva. We observed that nearly all MATAM sequences matched with a known 16S rRNA in Silva with more than 99% identity, among which a majority matched with 100% identity (Figures 4 and 5), which suggests that MATAM sequences could possibly be assigned at the species or even the strain level. On the other hand, EMIRGE sequences provided a discordant picture. In the case of the human mouth sample, most of the EMIRGE sequences obtained a match above 97% identity, but only a slight proportion of them matched with 100% identity against a known 16S rRNA (Figure 5). The observation is even more pronounced with the human gut sample, where only 43% of the EMIRGE sequences obtained a match above 97% identity against a Silva 16S rRNA sequence (Figure 4). Thus, conversely to MATAM, EMIRGE sequences would suggest that only a slight proportion of the human gut and mouth diversity has a known isolate registered in Silva. However, considering our previous conclusions on controlled datasets, we assume that part of this diversity inferred with EMIRGE might in fact corresponds to reconstruction artifacts.

5 Discussion

Taxonomic assignments of environmental samples is a strikingly difficult task which suffers from inherent limitations of high-throughput sequencing technologies. In this respect, we designed MATAM as an alternative to existing software helping to better understand the taxonomic structures of shotgun metagenomic samples. Our experimental results show that MATAM outperforms other available tools providing phylogenetic marker assemblies. Reconstructing full length 16S rRNAs allows to reach a higher precision of taxonomic assignments than individual read analysis or amplicon sequencing do, because the reconstructed sequences effectively contain stronger phylogenetic signal. Moreover, metagenomic shotgun sequencing is naturally immune against the primer and amplification biases attached to the amplicon sequencing technology, and therefore is more adequate to sequence unknown species.

Our approach opens up several new perspectives. Although we have focused this work on the assembly of 16S rRNA genes, MATAM was designed to deal with any marker of taxonomic interest. Indeed, there is currently an emerging trend to consider a combination of universal (single-copy) marker families, such as provided in the recently published database proGenomes (Mende *et al.*, 2017). Sequences from this database, or from any other customized one, could be used with MATAM to target a variety of markers, and thus provide improving taxonomic assignments. MATAM could also be used in combination with other types of sequencing data. Long read sequencing is able to produce fragments that cover large regions of the DNA molecules, up to several thousands of bases. When long reads are available, they could serve as a guide in the scaffolding step of MATAM and concomitantly, MATAM low-error contigs could be used to correct them. Finally, targeted gene capture, that allows to sequence at high depth captured DNA regions of interest from an environmental sample (Gasc *et al.*, 2016), could also prove to be an exciting application field for MATAM.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, **19**(5), 455–477.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, **42**(D1), D633–D642.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16s rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, **72**(7), 5069–5072.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). Seqan an efficient, generic c++ library for sequence analysis. *BMC Bioinformatics*, **9**(1), 1–9.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**(16), 2194–2200.

Gasc, C., Peyretaillade, E., and Peyret, P. (2016). Sequence capture by hybridization to explore modern and ancient genomic diversity in model and nonmodel organisms. *Nucleic Acids Research*, **44**(10), 4504–4518.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**(4), 593–594.

Kopylova, E., Noé, L., and Touzet, H. (2012). Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, **28**(24), 3211–3217.

Kopylova, E., Noé, L., Pericard, P., Salson, M., and Touzet, H. (2014). Sortmerna 2: ribosomal rna classification for taxonomic assignation. In *Workshop on Recent Computational Advances in Metagenomics, ECCB 2014*.

Kopylova, E., Navas-Molina, J. A., Mercier, C., Xu, Z. Z., Mahé, F., He, Y., Zhou, H.-W., Rognes, T., Caporaso, J. G., and Knight, R. (2016). Open-source sequence clustering methods improve the state of the art. *mSystems*, **1**(1).

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**(10), 1674–1676.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*, **12**(Suppl 2), S4.

Locey, K. J. and Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, **113**(21), 5970–5975.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), pp. 10–12.

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P., and Kyrpides, N. C. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods*, **4**(6), 495–500.

Mende, D. R., Letunic, I., Huerta-Cepas, J., Li, S. S., Forslund, K., Sunagawa, S., and Bork, P. (2017). proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Research*, **45**(D1), D529–D534.

Mercier, C., Boyer, F., Bonin, A., and Eric, C. (2013). Sumatra and sumaclust: fast and exact comparison and clustering of sequences.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, **32**(7), 1088–1090.

Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., and Banfield, J. F. (2011). Emirge: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, **12**(5), R44.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**(10), 1335–1337.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. (2016). metaSPAdes: a new versatile de novo metagenomics assembler. *arXiv:1604.03071 [q-bio]*. arXiv: 1604.03071.

Ondov, B. D., Bergman, N. H., and Phillippy, A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.

Pignatelli, M. and Moya, A. (2011). Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data. *PLOS ONE*, **6**(5), e19984.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., and Konstantinidis, K. T. (2014). Strengths and Limitations of 16s rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLoS ONE*, **9**(4), e93827.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**(D1), D590–D596.

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**(6), 863–864.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jorgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Hansen, L. H., Sorensen, S. J., Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Kang, D. D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., Goeker, M., Kyrpides, N., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - a benchmark of computational metagenomics software. *bioRxiv*, page 099127.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**(8), 811–814.

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using Archaeal and Bacterial synthetic communities. *Environmental microbiology*, **15**(6), 1882–1899.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, **22**(3), 549–556.

The Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). NaÃ¯ve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, **73**(16), 5261–5267.

Yuan, C., Lei, J., Cole, J., and Sun, Y. (2015). Reconstructing 16s rRNA genes in metagenomic data. *Bioinformatics*, **31**(12), i35–i43.