 doi:10.20944/preprints201909.0170.v1

ARSA-16S: a new and conceptually different approach for 16S-based taxonomic profiling

Eduardo Pareja-Tobes, Raquel Tobes

Oh no sequences! research group, Era7 Bioinformatics

Plaza de Campo Verde, 3, 18001 Granada, Spain

September 16, 2019

Abstract

Here we describe ARSA-16S, a tool and accompanying reference database for the analysis of bacterial 16S amplicons. Among other features, ARSA-16S is based on a new model, approach, and algorithm for sequence-level assignment of reads understood as probability distributions, assigns reads individually, and is designed with non-overlapping amplicons covering two non-contiguous regions. A new set of primers for the amplification and sequencing of the V4 and V6 regions is also provided.

Keywords *16S analysis, metagenomics, amplicons, bioinformatics, DNA sequencing.*

Corresponding author Raquel Tobes rtobes@era7.com

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

1 Introduction

Massive sequencing of amplicons spanning variable regions of 16S rRNA is the most used approach for taxonomic profiling. Although the use of long reads provided by third generation sequencing technologies to sequence the whole 16S rRNA gene is augmenting, approaches based on short reads NGS technologies have been and continue being the more used ones. Consequently the majority of the bioinformatics methods for taxonomic profiling are oriented to analyze short reads-based approaches. The review [11] about "Best Practices" for 16S Microbiome studies highlights the methods QIIME [5], MG-RAST [9], UCHIME [6], and mothur [12] as pipelines commonly used. A recent work focused on the benchmarking of the main 16S bioinformatics tools has been published [1] analyzing and evaluating the accuracy of mothur and QIIME, tools that they consider as most widely used taxonomic analysis tools, compared with two recently released alternatives, MAPseq and QIIME 2. They use synthetic simulated datasets with different reference databases testing also different variable sub-regions of the 16S rRNA gene to analyze these four bioinformatics tools. The authors conclude that QIIME 2 obtains the largest proportion of classified sequences with more accurate relative abundances and that MAPseq is a more conservative and precise approach with fewer misassignments (at genus rank). QIIME 2 [2], is a more advanced optimization of QIIME that collect different algorithms for taxonomic classification of marker-gene amplicon sequences. QIIME 2 uses new taxonomy classifiers as those based on the use of BLAST+, or VSEARCH or the naive Bayes scikit-learn classifier. Scikit-learn is a set of Machine Learning tools for data mining and data analysis built on NumPy, SciPy, and matplotlib. Using QIIME 2, the selection of the most appropriate algorithm and the correct setting of the different parameters is very complex and, hence, not easy to do for lab researchers. Algorithm and parameters setting dependence implies that results obtained with different algorithms and/or different parameters are not comparable. MAPseq [8] is based on sequence read mapping against hierarchically clustered and annotated reference sequences that are pre-clustered in hOTUs at different identity thresholds, and pre-classified to taxonomic categories based on the NCBI taxonomy and on the All-species Living Tree Project dataset [14]. Its computational efficiency is due to include improved k-mer counting based on a pre-clustering step, to align only the high scoring segment pairs using Needleman-Wunsch algorithm and to use a sensitive algorithm to evaluate the confidence level of the assigned classification. In [3] the concept of ASVs (Amplicon Sequences variant) was proposed as a substitute for OTUs, commonly used in microbiome taxonomic profiling tools as QIIME and mothur and MG-RAST. Operational Taxonomic Unit (OTU) is an operational definition used to classify groups of closely related individuals while ASVs are inferred by a **de novo** process in which biological sequences are discriminated from errors mainly based on that biological sequences are more likely to be repeatedly observed than are error-containing sequences. This philosophy is applied in the related open-source software package DADA2 [4].

Here we describe ARSA-16S, a new approach for 16S bacterial taxonomic profiling, designed to analyze any Illumina-based amplicon design, be it single or paired read, overlapping or not.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

Figure 1: **V3-V4 standard amplicon**

2 Amplicon Design

2.1 Non-overlapping Amplicons

While almost all existing 16S amplicon designs consist in reads whose sequences are expected to overlap in the target, nothing actually forces this overlapping design. On the contrary, we want to argue that non-overlapping amplicons are conceptually better: overlapping amplicons basically waste sequence. The overlapping region must be of considerable size (in a standard `2x300` MiSeq run as shown in *Figure 1*, around `135bp`, `22%` of the available sequence) and we are sequencing that region *twice* for no apparent reason. It could be argued that this has become a standard only because of sequence assignment/comparison normally used work best with a contiguous sequence, and extracting the amplicon-specific regions from reference databases was deemed too hard/unnconvenient. In [7], the only reference we could find where the analysis of non-overlapping amplicons is considered, we can read "Out of many possible ways of assigning taxonomy for paired-end 16S reads, we have decided to use k-mer-based methods, since, besides their speed and accuracy, they allow us to query reads with gaps in them (marked as unknown nucleotides) without loss of accuracy", thus lending force to our hypothesis for the disregard of non-overlapping amplicons.

Non-overlapping amplicons, on the other hand, let us sample two variable regions of our choice, limited only by the presence of adequate conserved regions for primers, and in the case of Illumina a reasonable total insert size. By not wasting sequence we can, as we will see later, sequence V4-V6 with `2x150bp` reagents, or the standard V3-V4 with `2x250bp` (or even `2x150bp` with a different set of non-degenerate primers).

What ARSA-16S does instead is concatenating the two reads in the orientation they are expected to be in the target, and compare that with an amplicon-specific reference database. As we will see this approach works the same with overlapping and non-overlapping amplicons, only making it obvious

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

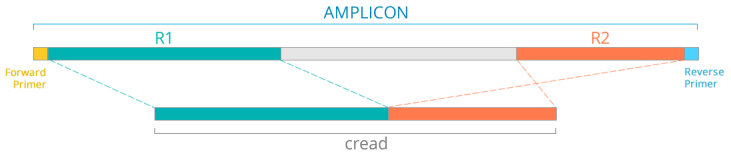

Figure 2: **cread construction** creads are formed by first removing the primer sequences from both reads, and then concatenating the resulting R1 suffix with (the reverse complement of) the R2 one.

that overlapping amplicons waste precious sequence by at best sampling it twice, at worst introducing artifacts in the assembly rendering otherwise good reads unassignable. The target-oriented concatenated reads will be called *creads* from now on. These creads is what needs to be assigned to an amplicon-specific reference database formed by *creferences*, obtained by determining the positions at which the primers could bind and extracting the cread that would be produced thereof.

2.2 Amplicon-Specific Reference Databases

The extraction of creferences starts with a set of 16S sequences which are expected to contain the amplified region. In this set we localize where the primers could match[1], and extract what would be sequenced according to the read length used. These two sequences (for paired reads) are then concatenated. The primers are never included in either creads or creferences: what we sequence in a PCR amplicon is the primer *used* when amplifying a molecule during a particular cycle, not the original target sequence. In short, we simulate which reads could be obtained with a given set of primers from the input reference set. For overlapping amplicons the resulting creferences will of course contain inside the overlapping sequence twice. These database, once constructed, can be used for the assignment of any dataset using the same primers and read lengths.

2.2.1 Amplicon Resolution

Multiple reference sequences can be expected to yield the same creference; the ratio creferences/references provides a coarse measure of the global discrimination power of the corresponding amplicon. Each creference is of course linked with the references containing it, so that for example taxonomic assignments of the references can be combined[2]. The taxonomic resolution of different amplicons can be thus analyzed beforehand, looking for instance at whether relevant species/strains could be distinguished, choosing then the right amplicon for the experiment.

[1] By default looking for perfect matches of a suffix of the primer 3' end.
[2] by computing their LCA (lowest common ancestor), for example.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

Figure 3: **creference database construction** An example showing the construction of a creference for our V4-V6 amplicon, with source the canonical *E. coli* 16S sequence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

3 ARSA-16S Reference Database

The default global 16S reference database (from which amplicon-specific ones are extracted) is a subset of RNAcentral [13], the most comprehensive compilation of non-conding RNA sequences. Sequences are included in our database if they satisfy all of the following

- they are annotated as `rRNA`
- the sequence contains no unambiguous nucleotides (only `ATGC`)
- its header description in some of the source databases contains the string `16S` / `16s`
- it is present in at least one open database[3]

In total, ~6 millions of sequences are included. Code and database itself are accessible under the open AGLPv3 and ODbLv1 licenses, respectively.

3.1 Taxonomy Annotations

Each sequence in the global 16S reference database has a set of linked RNAcentral entries, each with its own NCBI taxonomy annotation; and each creference in an amplicon-specific database is present in potentially several sequences of the global 16S database. These different annotations are merged by computing their LCA (lowest common ancestor). Simply computing their LCAs in the whole NCBI taxonomy tree would lead in most cases to completely uninformative assignments, though[4]; as a workaround, we have defined three subtrees: *classified*, *environmental*, and *unclassified* which together cover the whole NCBI tree. LCAs of taxonomic assignments across each of these trees yield a far better picture regarding what their name imply: the *classified* tree corresponds to assignments to well-defined, standard taxa, *environmental* to sequences linked to particular environments/samples, and *unclassified* to novel/unknown organisms.

Each of these trees is what we call the *covering tree* of a generating set of taxa: the nodes which are ancestors or descendants of some node in the generating set. Generating sets for each are

- *unclassified* its scientific name contains the `unclassified` string
- *environmental* its scientific name contains `environmental samples` as prefix
- *classified* those being neither *unclassified* nor *environmental*

[3] RNAcentral includes all SILVA entries, which does not have an open license. When/if the long announced open release of SILVA happens we will also include the 16S sequences from SILVA that are not present in any other database. In any case, this set of sequences is fairly small.

[4] Nodes for Environmental samples can branch at any level of the Bacteria subtree, same for unclassified sequences. The use of the same tree for this radically different notion of "taxa" is in our opinion a serious mistake.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

4 Sequence Assignment

The definition of cread assignment and its implementation is based on [10]. For completeness, we outline the relevant notions and detail the specifics of how ARSA-16S relies on the approach and results thereof.

4.1 Read Sequence Model

All reads and the resulting creads are treated throughout as (the product of) probability distributions on $\{A, C, G, T\}$. For fastq input, these distributions are determined in the standard way, considering that all error bases are equally likely [10]. The input though can be any product of product of probability distributions, one per position; we can thus analyze consensuses resulting from amplicons tagged with UMIs.

4.2 Sequence Assignment Definition

ARSA-16S computes an exact solution to a precisely defined problem. This problem is defined considering the creferences as possible observations of the cread, a probability distribution. The best possible creference is thus the most likely one under the cread distribution. In short, for each cread `r` , ARSA-16S outputs the most likely creference in the provided creference set [10].

This assignment definition is of course sensitive to read qualities: two creads with the same most likely sequence (the sequence in the fastq file) could perfectly have different assignments in the same creference set.

While we consider OTUs and binning erroneous approaches[5], ARSA-16S could also be used on sets of reads by generating a joint distribution for each cluster.

4.3 Sequence Assignment Algorithm

The assignment algorithm[6] is based on the techniques described in [10].

First the most likely sequences of all creads are compared for exact matches against the set of creferences[7], normally[8] producing a considerable set of cread assignments. The remaining creads are then

[5] even more so when this clustering process is done without considering sequence quality.

[6] Here algorithm is understood as a particular way of computing something well-defined (the assignment in this case), not (as is normally the case in the bioinformatics literature) as a procedure providing an implicit notion of what we want to compute, of which its explicit formulation, if any, is frequently fuzzy.

[7] Obviously, if the creference set contains the most likely sequence of a cread, that is its best assignment.

[8] In our experience, 40-60% of creads have an exact assignment. This number of course will depend on both the particular sample and the global coverage of the reference database used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

compared with the cache of creferences with some read assigned to it, and the best one such is used as starting point for finding the best assignment. The creferences are indexed using an ultrametric index, which can of course be re-used for the assignment of reads of the same amplicon type; it is this index that is key for a fast[9] assignment computation in the case of creads having no exact creference counterpart; we refer the reader to [10] for details on how these ultrametric indexes work. Reads for which the best assignment probability is lower than user-provided threshold are output as unassigned.

4.4 Taxonomic Assignments

Taxonomic assignments of individual reads are computed independently of the sequence assignment we have just described, simply as its assignment creference taxon in each of the *classified*, *unclassified*, *environmental* trees.

5 Asymmetric V4-V6 Amplicon

As a proof of concept we have designed and developed a new V4-V6 amplicon compatible with any Illumina sequencer (such as iSeq) providing `2x150bp` reads. To sequence these two non-contiguous variable regions with the least possible number of sequencing cycles we have designed new primers as close as possible to each variable region. V4 being larger than V6, we use asymmetric reads with `180bp` for R1 and `120bp` for R2, each fully covering V4 and V6 respectively.

5.1 V4-V6 Primers

The new primers are a set of 14 (7 forward, 7 reverse) *non-degenerate* primers. The common use of degenerate positions in the primers restricts valid regions to those where the differences between targets occur in the same position, while normally generating a lot of extraneous undesirable primers which do not match any known target (while possibly matching partially sequences which we don't want to amplify).

[9] ARSA-16S, operating single-threadedly, assigns reads at a speed of around 10^4 reads per second.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 September 2019 doi:10.20944/preprints201909.0170.v1

Table 1: **V4-V6 primers** Primer IDs together with their sequence in the 5' ⟶ 3' orientation. The Primer IDs follow the pattern `Era7-<start_pos>-<end_pos>-<F/R><number>` where positions are those in the standard *E. coli* reference and `F,R` denote forward, reverse primers. Nucleotides shared by all primers appear in bold.

Primer ID	5' -> 3'
Era7-516-528-F1	**TGCCAGCAGCCGC**
Era7-516-528-F2	**TGCCAGCCGCCGC**
Era7-516-528-F3	**TGTCAGCCGCCGC**
Era7-516-528-F4	**TGCCGGCAGCCGC**
Era7-516-528-F5	**TGCCAACAGCCGC**
Era7-516-528-F6	**TGCCAGCGGCCGC**
Era7-516-528-F7	**TGCCATCAGCCGC**
Era7-1063-1079-R1	**CGTCAGCTCGT**GTCG**TG**
Era7-1063-1079-R2	**CGTCAGCTCGT**GCCG**TG**
Era7-1063-1079-R3	**CGTCAGCTCGT**GTTG**TG**
Era7-1063-1079-R4	**CGCCAGCTCGT**GCCG**TG**
Era7-1063-1079-R5	**CGTCAGCTCGT**GCTG**TG**
Era7-1063-1079-R6	**CGTCAGCTCGT**ACCG**TG**
Era7-1063-1079-R7	**CGTCAGCTCGT**GCCT**TG**

These primers have underwent significant laboratory testing, involving sequencing real samples from various origins, and synthetic amplicons covering all possible primer combinations. Using our global reference database, these primers cover with exact matches ~96% of the sequences coming from full 16S genes. Considering that the primers will also amplify references with some mismatch along its sequence, it is probable that this set of primers covers almost if not all the references included in our database.

References

[1] **Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments**
Alexandre Almeida, Alex L Mitchell, Aleksandra Tarkowska, and Robert D Finn
GigaScience **7** (5 2018)
DOI: 10.1093/gigascience/giy054.

[2] **Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin**
Nicholas A. Bokulich, Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso
Microbiome **6** (1 2018)
DOI: 10.1186/s40168-018-0470-z.

[3] **Exact sequence variants should replace operational taxonomic units in marker-gene data analysis**
Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes
The ISME Journal **11** (12 2017), 2639–2643.
DOI: 10.1038/ismej.2017.119.

[4] **DADA2: High-resolution sample inference from Illumina amplicon data**
Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes
Nature Methods **13** (7 2016), 581–583.
DOI: 10.1038/nmeth.3869.

[5] **QIIME allows analysis of high-throughput community sequencing data**
J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld, and Rob Knight
Nature Methods **7** (5 2010), 335–336.
DOI: 10.1038/nmeth.f.303.

[6] **UCHIME improves sensitivity and speed of chimera detection**
Robert C. Edgar, Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight
Bioinformatics **27** (16 2011), 2194–2200.
DOI: 10.1093/bioinformatics/btr381.

[7] **IM-TORNADO: A Tool for Comparison of 16S Reads from Paired-End Libraries**
Patricio Jeraldo, Krishna Kalari, Xianfeng Chen, Jaysheel Bhavsar, Ashutosh Mangalam, Bryan White, Heidi Nelson, Jean-Pierre Kocher, Nicholas Chia, and Paul Jaak Janssen
PLoS ONE **9** (12 2014)
DOI: 10.1371/journal.pone.0114804.

[8] **MAPseq: highly efficient k-mer search with confidence estimates, for rRNA sequence analysis**
João F Matias Rodrigues, Thomas S B Schmidt, Janko Tackmann, Christian von Mering, and Inanc Birol
Bioinformatics **33** (23 2017), 3808–3810.
DOI: 10.1093/bioinformatics/btx517.

References

[9] **The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes**
F Meyer, D Paarmann, M D'Souza, R Olson, EM Glass, M Kubal, T Paczian, A Rodriguez, R Stevens, A Wilke, J Wilkening, and RA Edwards
BMC Bioinformatics **9** (1 2008)
DOI: 10.1186/1471-2105-9-386.

[10] **Amplicon Analysis I**
Eduardo Pareja-Tobes and Raquel Tobes (2019).

[11] **The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies**
Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, Mick Watson, and Shuang-Jiang Liu
Applied and Environmental Microbiology **84** (7 2018)
DOI: 10.1128/AEM.02627-17.

[12] **Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities**
P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber
Applied and Environmental Microbiology **75** (23 2009), 7537–7541.
DOI: 10.1128/AEM.01541-09.

[13] **RNAcentral: a hub of information for non-coding RNA sequences**
Blake A Sweeney, Anton I Petrov, Boris Burkov, Robert D Finn, Alex Bateman, Maciej Szymanski, Wojciech M Karlowski, Jan Gorodkin, Stefan E Seemann, Jamie J Cannone, Robin R Gutell, Petra Fey, Siddhartha Basu, Simon Kay, Guy Cochrane, Kostantinos Billis, David Emmert, Steven J Marygold, Rachael P Huntley, Ruth C Lovering, Adam Frankish, Patricia P Chan, Todd M Lowe, Elspeth Bruford, Ruth Seal, Jo Vandesompele, Pieter-Jan Volders, Maria Paraskevopoulou, Lina Ma, Zhang Zhang, Sam Griffiths-Jones, Janusz M Bujnicki, Pietro Boccaletto, Judith A Blake, Carol J Bult, Runsheng Chen, Yi Zhao, Valerie Wood, Kim Rutherford, Elena Rivas, James Cole, Stanley J F Laulederkind, Mary Shimoyama, Marc E Gillespie, Marija Orlic-Milacic, Ioanna Kalvari, Eric Nawrocki, Stacia R Engel, J Michael Cherry, SILVA Team, Tanya Z Berardini, Artemis Hatzigeorgiou, Dimitra Karagkouni, Kevin Howe, Paul Davis, Marcel Dinger, Shunmin He, Maki Yoshihama, Naoya Kenmochi, Peter F Stadler, and Kelly P Williams
Nucleic Acids Research **47** (D1 2019), D221–D229.
DOI: 10.1093/nar/gky1034.

[14] **The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks**
Pelin Yilmaz, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner
Nucleic Acids Research **42** (D1 2013), D643–D648.
DOI: 10.1093/nar/gkt1209.