

Genome Analysis

RibDif: can individual species be differentiated by 16S sequencing?

Mikael Lenz Strube  *

Department of Biotechnology and Biomedicine, Technical University of Denmark, Kongens Lyngby DK 2800, Denmark

*To whom correspondence should be addressed.

Associate Editor: Alexandros Stamatakis

Received on June 3, 2021; editorial decision on August 26, 2021; revised on August 12, 2021; accepted on September 1, 2021

Abstract

Motivation: Metataxonomic analysis is now routinely used to profile the microbiome of virtually every ecological niche on planet Earth. The use of amplicon sequence variants (ASVs), proposing to be the exact biological 16S rRNA amplicon sequences of a given biological system, is now considered the gold standard. However, the 16S rRNA genes, and in particular the amplicons derived from it, are not unique for most species nor are they necessarily unique within individual genomes. Despite these restrictions, individual ASVs are often used to make inferences on the state of a given ecosystem, which may cause erroneous conclusions on the effects of a given species on a specific host phenotype or ecosystem.

Results: To support researchers working with metataxonomics, we have developed RibDif, which easily and rapidly evaluates the feasibility of using metataxonomics to profile individual species. We use RibDif to demonstrate that the genus *Pseudoalteromonas* contains species that are impossible to distinguish with 16S amplicons and that this is a common motif in bacterial genera. We propose that researchers consult RibDif when making conclusions on individual species from metataxonomic data.

Availability and implementation: RibDif is freely available along with source code and detailed documentation at <https://github.com/mikaells/RibDif>.

Contact: milst@dtu.dk

1 Introduction

The analysis of microbiomes has been vastly improved by high-throughput sequencing technologies and concurrent improvements in computational biology. Lately, the adaptation of the amplicon sequence variant (ASV) rather than the operational taxonomic unit (OTU) theoretically allows researchers to resolve bacteria at the species level. Since the OTU is an artificially constructed consensus sequence (often 97% similarity), much information is lost compared to the ASV, which instead aims to represent the original biological sequence. The advantage here is evident, e.g. individual species or strains of a genus may have unique properties and ASVs are easier to follow across studies than OTUs (Callahan *et al.*, 2017). This, however, requires that a single sequence, even devoid of technical error, is unique for a particular species.

Unfortunately, many bacterial species have multiple different alleles (Vetrovský and Baldrian, 2013) and the alleles within a genus may not be unique enough to distinguish species (Johnson *et al.*, 2019). The first issue may lead to multiple ASVs from single genomes and the second will not allow one to distinguish species at all. It has become commonplace to associate individual ASVs with ecosystem or host phenotype, but for the reasons given above,

this may not be possible. Moreover, inferring the genome, and hence the metabolic capacity, from amplicons has had limited success (Sun *et al.*, 2020), assumedly for the same reasons. Lastly, diversity estimates become increasingly meaningless if some genomes have multiple alleles or if some species share the same unique allele.

To investigate this, we have developed RibDif, which allows users to easily and quickly answer if species within a given genus can be resolved using standard metataxonomic methods, hence letting users evaluate if ASVs are useful for their particular research question or if alternative methods are to be employed. As described previously, one solution is to develop primers not targeting the 16S region (Lauritsen *et al.*, 2021; Strube *et al.*, 2018).

The software provides a detailed analysis of individual genomes and graphical representations of overall 16S amplicon diversity as well as a report on exactly which species overlap.

2 Methods and implementation

The code for analysis is available at <https://github.com/mikaells/RibDif>. Given a genus or a species, the software downloads all available complete genomes from NCBI using `ncbi-genome-download`

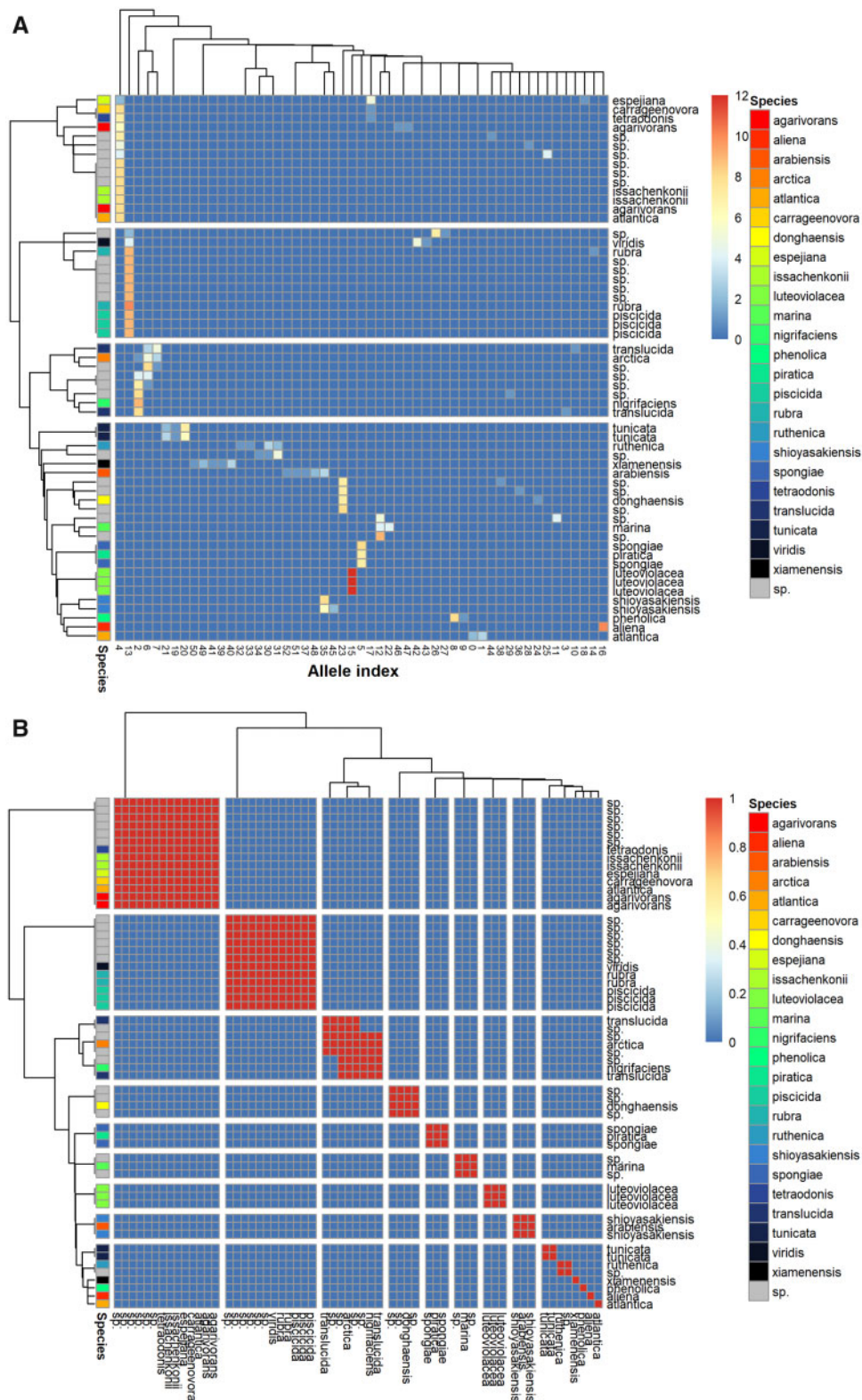


Fig. 1. Heatmaps of 16S rRNA V3V4 alleles in *Pseudoalteromonas* as generated by RibDif. V3V4 amplicons were generated from the 16S rRNA gene of *Pseudoalteromonas* and clustered at 100% identity. (A) Alleles are across columns and individual genomes on the rows. The count of members in a given cluster is given as color intensity. Rows and columns are additionally clustered by similarity in cluster abundance using Wards method. (B) Confusion matrix of each genome showing which will overlap in V3V4-based metataxonomy, clustered using Wards method. The figure is generated using 'RibDif -g Pseudoalteromonas'. sp., genomes only classified as *Pseudoalteromonas* species.

(<https://github.com/kblin/ncbi-genome-download>). All 16S rRNA genes at least 90% of full length are then found by Barrnap (<https://github.com/tseemann/barrnap>) and these sequences are next

subjected to *in silico* polymerase chain reaction (PCR) using *in silico*_pcr.pl (https://github.com/egonozet/in_silico_pcr), by default using primers targeting the V3V4-region (CCTACGGGNGGCNGCAG and

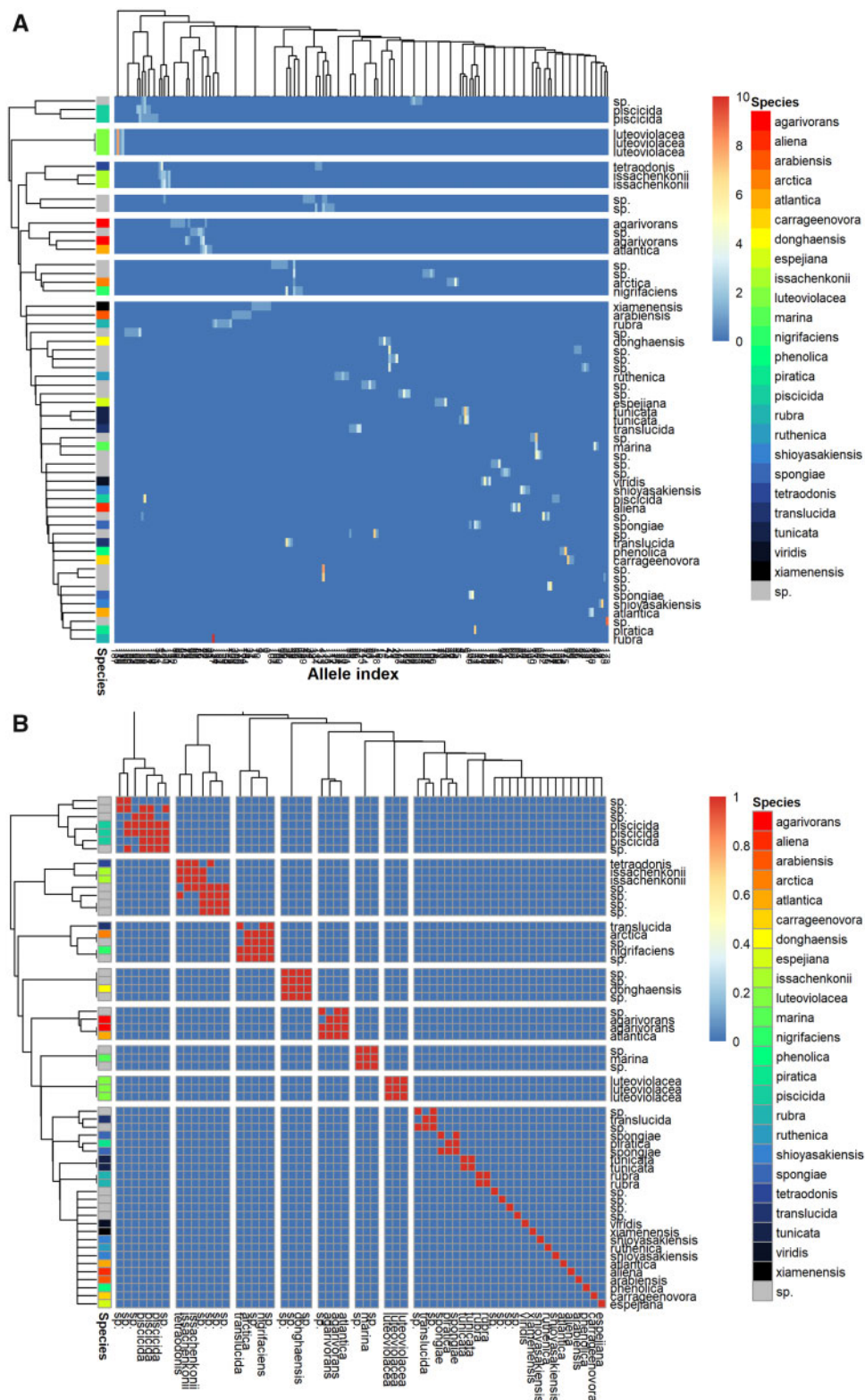


Fig. 2. Heatmaps of 16S rRNA V1V9 alleles in *Pseudoalteromonas* as generated by RibDif. V1V9 amplicons were generated from the 16S rRNA gene of *Pseudoalteromonas* and clustered at 100% identity. (A) Alleles are across columns and individual genomes on the rows. The count of members in a given cluster is given as color intensity. Rows and columns are additionally clustered by similarity in cluster abundance using Wards method. (B) Confusion matrix of each genome showing which will overlap in V1V9-based metataxonomy, clustered using Wards method. The figure is generated using 'RibDif-g *Pseudoalteromonas*'. sp., genomes only classified as *Pseudoalteromonas* species.

GACTACNNGGTATCTAATCC) and full-length V1V9-region (AGRGTTYGATYMTGGCTCAG and RGYTACCTTGTTACGA CTT), allowing one mismatch and one insertion. User-defined

primers are allowed using the `-p/-primer` argument. Optionally, the resulting amplicons of all genomes are aligned by use of MAFFT (Katoh *et al.*, 2002) using the `—adjustdirection` option and a

Table 1. Medically and biotechnologically relevant genera and their resolvability using 16S amplicon sequencing of the V3V4 and V1V9 region

Genus	Complete genomes	Species	Allele multiplicity ^a (%)		Overlaps ^b (%)	
			V3V4	V1V9	V3V4	V1V9
<i>Lactobacillus</i>	123	19	30.08	62.60	21.05	10.53
<i>Vibrio</i>	336	52	86.61	98.51	57.69	17.31
<i>Bacillus</i>	856	47	63.90	93.93	74.47	55.32
<i>Pseudomonas</i>	860	92	24.53	59.42	80.43	35.87
<i>Staphylococcus</i>	889	38	34.76	87.49	65.79	13.16
<i>Escherichia</i>	1700	4	32.94	91.47	100.00	50.00

^aThe percentage of genomes having more than one unique allele.

^bThe percentage of species having at least one allele overlapping another species.

corresponding tree is generated with fasttree (Price *et al.*, 2010). Optional detailed intragenomic analysis is performed by pyani (<https://github.com/widowquinn/pyani>).

Next, VSEARCH (Rognes *et al.*, 2016) is used to cluster the amplicons at 100% identity (as is the theoretical cutoff for ASVs). These clusters are then tabulated across genomes, after which a summary and graphics of allele multiplicity and textual information of species overlap are provided.

The software is implemented in bash and R and is fully parallelized using GNU parallel (Tange, 2020). RibDif with default settings will analyze the *Escherichia* genus, consisting of 1700 genomes, in 1481 s, of which genome download (108 s) and rRNA prediction (860 s) are the major bottlenecks on a 10× dual-core Intel(R) Xeon(R) W-2155 CPU @ 3.30 GHz. Analysis of the *Pseudoalteromonas* genus, consisting of 58 genomes, is accomplished in 62 s.

3 Results

The analysis relies on the RibDif software described in Section 2. Briefly, the software downloads complete genomes of a genus/species, extracts the 16S rRNA genes, amplifies them *in silico* and clusters the amplicons at 100% identity to mimic the concept of an ASV.

Three problematic scenarios are observed. A genome may have 16S rRNA amplicons (i) that are the same in different species, (ii) have multiple variants and (iii) may have multiple variants that are individually shared with other species.

3.1 *Pseudoalteromonas* V3V4 amplicons

As an example, consider the genus *Pseudoalteromonas*, a marine bacterium with 48 recognized taxa of which 24 have published complete whole genomes (36 in total). In addition, 22 complete genomes of unspecified species are available. When clustering the 499 V3V4 amplicons derived from these genomes at 100% identity into 53 unique alleles, it is observed that several of these are identical across species (Fig. 1A). Specifically, *Pseudoalteromonas rubra*, *Pseudoalteromonas viridis* and *Pseudoalteromonas piscicida* are indistinguishable as they share allele 13. In addition, a genome of *P. viridis* contains three unique alleles, one of which is also found in allele 13. Allele 4 is represented in genomes of *Pseudoalteromonas espejiana*, *Pseudoalteromonas issachenkonii*, *Pseudoalteromonas atlantica*, *Pseudoalteromonas carragenenovora*, *Pseudoalteromonas agarivorans* and *Pseudoalteromonas tetradonis*, several of which contain multiple other alleles. As seen in Figure 1A and B, a genome of *Pseudoalteromonas arctica* (along with several unclassified species) contains three alleles, each of which overlap separately with genomes of *Pseudoalteromonas translucida* and *Pseudoalteromonas nigrifaciens*. Genomes such as the one representing *P. espejiana*, which have two shared but one unique allele, are still problematic as all alleles are amplified by PCR and the shared alleles (alleles 4 and 17) will incorrectly suggest the presence of 5 and 2 additional species, respectively.

3.2 Full-length V1V9 amplicons

When analyzing full-length V1V9 amplicons rather than the V3V4 amplicons, substantially more (202) unique alleles are generated owing to the sequences being approximately 3 times longer (Fig. 2). The resolving power of these amplicons is substantially improved, e.g. *P. rubra*, *P. viridis* and *P. piscicida* are now distinguishable and all alleles of *P. viridis* are unique. However, the following pairs of species are still indistinguishable: *tetraodonis/lissachenkonii*, *translucida/arctica/nigrifaciens*, *agarivorans/atlantica* and *spongia/piratica*.

3.3 Medically and biotechnologically important genera

Other genera, such as *Lactobacillus*, are easier to distinguish (Table 1). Of the 19 species having a completed whole genome, the only overlaps in V3V4-amplicons are in the known closely related species of *gasserii/paragasserii* and *amylovorus/lultunensis*, although the latter are distinguishable with V1V9-amplicons. On the other hand, other genera investigated show the same pattern as in *Pseudoalteromonas*, e.g. high allele multiplicity and extensive species overlap, which, however, is somewhat mitigated by using full-length V1V9 amplicons. The *Pseudomonas* genus is particularly difficult to separate into species, being both unusually large (92 species) as well as having 80.43% species overlap. It is likely that most genera are not possible to resolve into species level using V3V4 amplicons, although the use of V1V9-amplicons decreases overlap substantially in some genera, such as *Vibrio* and *Staphylococcus*.

4 Discussion

The advantages of ASVs are evident (Callahan *et al.*, 2017). The use of original biological sequences of a sample rather than artificial consensus sequences theoretically allows one to describe the species of a microbiome with higher resolution. Unfortunately, it does not appear that all species can be described as such.

In this analysis, we have observed some issues with the ASV logic as previously described (Johnson *et al.*, 2019; Větrovský and Baldrian, 2013). It is observed that individual bacteria may have multiple 16S genes and, importantly, may have multiple amplicon variants within the V3V4 region. More problematically, several of these are shared between individual species in several major bacterial genera.

As seen from the analysis of *Pseudoalteromonas*, several practical issues can appear. A researcher analyzing a community containing *P. viridis* using V3V4 metaxonomics will—assuming complete absence of technical error—observe three ASVs, which will overinflate the *Pseudoalteromonas* richness and incorrectly suggest the presence of *P. rubra* and *P. piscicida*. In a natural system with potentially 1000s of species across multiple genera, these issues will eventually disturb conclusions made on both diversity and species presence/absence. Condensing ASVs into species or genera may be a more robust option than relying on ASVs as individually meaningful units.

Researchers should consider these issues when evaluation meta-taxonomic data and we propose they consult RibDif before making conclusions on individual ASVs. Researchers being particularly interested in species differentiation should consider full-length V1V9 amplicons or primers tailored to individual genera as previously done for *Pseudomonas* (Lauritsen *et al.*, 2021) and *Staphylococcus* (Strube *et al.*, 2018).

Acknowledgements

We thank the Bacterial Ecophysiology and Biotechnology group at DTU for helpful discussion and naming suggestions and Mikkel Lindegaard for coding suggestions. Henrik Lenz is gratefully acknowledged for language editing.

Funding

This work has been supported by the Center of Excellence for Microbial Secondary Metabolites (CeMiSt) funded by the Danish National Research Foundation [DNRF137].

Conflict of Interest: none declared.

References

- Callahan, B.J. *et al.* (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, **11**, 2639–2643.
- Johnson, J.S. *et al.* (2019) Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.*, **10**, 5029.
- Katoh, K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
- Lauritsen, J.G. *et al.* (2021) Identification and differentiation of pseudomonas species in field samples using an rpoD amplicon sequencing methodology. *mSystems*, **6**, e00704.
- Price, M.N. *et al.* (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
- Rognes, T. *et al.* (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
- Strube, M.L. *et al.* (2018) A detailed investigation of the porcine skin and nose microbiome using universal and *Staphylococcus* specific primers. *Sci. Rep.*, **8**, 12751.
- Sun, S. *et al.* (2020) Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome*, **8**, 46–49.
- Tange, O. (2011) GNU Parallel – The Command-Line Power Tool. *login: The USENIX Magazine*, **2011**, 42–47.
- Větrovský, T. and Baldrian, P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One*, **8**, e57923.