Journal of Biotechnology **18**(4): 733-743, 2020

BIOINFORMATIC APPROACHES FOR ANALYSIS OF CORAL-ASSOCIATED BACTERIA USING R PROGRAMMING LANGUAGE

Doan Thi Nhung[1], Bui Van Ngoc[1, 2,✉]

[1]*Institute of Biotechnology, Vietnam Academy of Science and Technology*
[2]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology*

✉To whom correspondence should be addressed. E-mail: bui@ibt.ac.vn

Received: 28.7.2020
Accepted: 25.9.2020

SUMMARY

Recent advances in metagenomics and bioinformatics allow the robust analysis of the composition and abundance of microbial communities, functional genes, and their metabolic pathways. So far, there has been a variety of computational/statistical tools or software for analyzing microbiome, the common problems that occurred in its implementation are, however, the lack of synchronization and compatibility of output/input data formats between such software. To overcome these challenges, in this study context, we aim to apply the DADA2 pipeline (written in R programming language) instead of using a set of different bioinformatics tools to create our own workflow for microbial community analysis in a continuous and synchronous manner. For the first effort, we tried to investigate the composition and abundance of coral-associated bacteria using their *16S rRNA* gene amplicon sequences. The workflow or framework includes the following steps: data processing, sequence clustering, taxonomic assignment, and data visualization. Moreover, we also like to catch readers' attention to the information about bacterial communities living in the ocean as most marine microorganisms are unculturable, especially residing in coral reefs, namely, bacteria are associated with the coral *Acropora tenuis* in this case. The outcomes obtained in this study suggest that the DADA2 pipeline written in R programming language is one of the potential bioinformatics approaches in the context of microbiome analysis other than using various software. Besides, our modifications for the workflow execution help researchers to illustrate metagenomic data more easily and systematically, elucidate the composition, abundance, diversity, and relationship between microorganism communities as well as to develop other bioinformatic tools more effectively.

Keywords: *16S rRNA, Acropora tenuis, bioinformatics, coral-associated bacteria, R programming language*

INTRODUCTION

Less than 2% of the approximately 10^{30} microorganisms on earth can be cultured using traditional methods (Wade, 2002). Therefore, only a very small part of the benefits from microorganisms are exploited, the remaining potential is still a challenge. Healthy corals, together with various microorganisms such as bacteria, algae, fungi, viruses, form a community living in close relationships in the coral mucus layer, skeleton, and tissues (Rosenberg *et al.*, 2007). The most important role of organisms living on corals is to provide coral nutrition and play a protective role by synthesizing antibiotics to fight pathogens (Kvennefors *et al*., 2012). The dynamic equilibrium between organisms in coral holobiont is the result of interactions between symbiotic microorganisms under certain environmental conditions (Reshef *et al*., 2006). Bacteria have been known to be an important part of ecosystem function, participating in

geochemical and ecosystem processes (Balser *et al*., 2006) and being able to perform necessary functions in the coral reef environment. Several studies have demonstrated that there are differences in the species composition of bacteria between some corals (Bourne, Munn, 2005). Besides, bacteria are potential subjects that play an important role in coral health, the explorations and clarifications of the interactions of whole bacterial communities in the coral mucus layer would be contributed by the examination of coral bacteria, their composition and abundance. Metagenomic studies have proposed that there are a wide distribution and high diversity of bacteria, phages in the mucus of organisms, such as the intestinal tract, human respiratory tract, and coral (Breitbart *et al*., 2002). Therefore, in the direction of coral microbiological research, bacterial studies are very necessary and meaningful, contributing to establish fundamental science to identify the coral pathogens and disease-forming mechanisms. Thereby, we can have appropriate approaches for the protection of coral health and toward sustainable development of coral reef ecosystems.

Recently, one of the most significant events in the field of microbial ecology has been the emergence and development of metagenomics (Thomas *et al.*, 2012). Metagenomics is a new approach of dry and wet laboratory technique combination, aimed for determination and nomenclature of microorganisms, especially marine bacteria based on their DNA/RNA sequences. It provides a tool to assess the phylogenetic and functional aspects of associated-coral bacterial diversity without the need for lab cultivation and isolation of individual species. It also helps determine the species of community, metabolism, and functional roles of the microbes in the environmental sample (Langille *et al*., 2013). Metagenomics provides genetic information on evolutionary profiles of community, novel enzymes or biocatalysts, genomic linkages between function, and phylogeny for unculturable organisms. It is also a powerful tool for initiating new hypotheses of microbial function. Metagenomics and the next-generation sequencing (NGS) technology allow creating huge sequencing data sets from a variety of environments such as soil, the human body or ocean water samples. The approach of metagenomics based on direct extraction of nucleic acids from environmental samples has been shown to be highly effective for ecological comparison and analysis, metabolic data collection of microorganisms in complex environments as well as identifying new molecules using a library built from previously isolated nucleic acids.

Besides, the support of bioinformatics tools, especially the R programming language (Callahan *et al*., 2016), is also a great help for analyzing the metagenomic results. R is a programming language and a developed environment used in statistical computing and graphics. The R language has many advantages, such as helping users to store data and process data effectively, render results in the form of graphs that can be easily used by the users. The R programming language can be used for free, open-source, and installs on all operating systems. Therefore, R is very suitable for use in the field of biology that requires the processing of huge amounts of data and the ability to efficiently exploit data, especially in the field of microbiome analysis. Besides, R has the strength of the ability to display results in a graphical form, allowing users to make visual assessments. On the other hand, currently, the popular bioinformatics software (FastQC, Mothur, QIIME, Blast, SILVA) when used to analyze thousands of sequences simultaneously, will consume resources and slow down the computer. Moreover, these softwares are designed and developed by different companies, consequently, the analysis will not be continuous due to different input/output formats. Therefore, in order to reduce software installation costs, save time, increase compatibility and synchronization for NGS data, the R programming language will be used to develop into packages and pipelines to perform all steps of analyzing and processing

NGS data continuously and synchronously instead of using each processing software for each separate analysis step. Another advantage when developing tools in R language is the ability to use server resources to analyze massive data sets according to the PCA – Principal component analysis or other dimensions as Big Data. It is also the inevitable trend for bioinformatics in particular in the era of the 4.0 revolution.

Therefore, the study "Bioinformatic approaches for analysis of coral-associated bacteria using R programming language*"* was carried out to take full advantage of metagenomics and R programming language to analyze the bacterial communities living in coral *Acropora tenuis*.

MATERIALS AND METHODS

The dataset used in this study is highly-overlapping Illumina 2x250 amplicon sequences of the *16S RNA* gene of bacteria living in the coral *Acropora tenuis*, sequenced by the Illumina next-generation sequencing technology. These paired-end sequences were downloaded from the public NCBI Genbank database (Accession number: PRJNA517286) with the format of fastqfiles.

At first, the metagenomic sequence data was processed by *DADA2* package in R programming language (Callahan et al., 2016) through the following steps: filter and trim low-quality reads (QC<30), dereplicate the filtered fastq files, infer the sequence variants in each sample, merge the denoised sequences, remove chimeric sequences, mapping short sequence reads to a reference genome and assign taxonomy, respectively. Then, the composition and abundance of the bacterial community were analyzed in *Phyloseq* package to assess the phylogenetic and bacterial diversity, and the results were visualized with *ggplot2* package subsequently.

The process of analyzing 16S metagenome data using R programming language will be carried out according to the following procedure:

Packages loading with function library()

```
library("Rcpp")
library("dada2"); packageVersion("dada2")
library("phyloseq"); packageVersion("phyloseq")
library("ggplot2"); packageVersion("ggplot2")
```

Metagenomic data preprocessing

```
path <- "/Users/pwd/16srRNA/" # Create a path for writing and downloading data
```

The names of the fastq files must be checked using some of the string-changing functions to create two lists of F (forward) and R (reverse) read file. The sequence files have the name of Name_1.fastq or Name_2.fastq where 1 are the forward file, and 2 stand for the reverse. The above two types of sequences were grouped into two different groups: fnFs and fnRs.

```
fnFs <- sort(list.files(path, pattern="_1.fastq", full.names = TRUE))
fnRs <- sort(list.files(path, pattern="_2.fastq", full.names = TRUE))
plotQualityProfile (fnFs [1: 2]) # Visualize the quality of the first two reads of the forward sequence file
plotQualityProfile (fnRs [1: 2])
```

Assigning folder containing the filtered sequences from the fastq.gz file

```
filt_path <- file.path (path, "filtered") # Put filtered items in the filtered subdirectory
out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(220,220), maxN=0, maxEE=c(2,2), truncQ=2, trimLeft=c(25,20) , rm.phix=TRUE, compress=TRUE, multithread=TRUE) # Filter and trim low-quality sequences
```

After filtering and trimming low-quality reads (QC<30), DADA2 pipeline inferred amplicon sequence variants (ASVs) – the unique sequences from forward and reverse reads with their abundances.

```
derepFs <- derepFastq(filtFs, verbose=TRUE) # Dereplicate the filtered forward reads
derepRs <- derepFastq(filtRs, verbose=TRUE)
errF <- learnErrors(filtFs, multithread=TRUE) # Calculate the sequence error rates of the forward reads
errR <- learnErrors(filtRs, multithread=TRUE)
```

dadaFs <- dada(derepFs, err=errF, multithread=TRUE) # Infer sequence variants from forward reads

dadaRs <- dada(derepRs, err=errR, multithread=TRUE)

Subsequently, DADA2 merged together the inferred forward and reverse sequences and eliminated the residual errors (non-overlap paired sequences) to construct the ASVs table (ASVs and their abundances).

mergers <- mergePairs(dadaFs, derepFs, dadaRs, derepRs, verbose=TRUE) # Merge paired-end sequences

seqtab <- makeSequenceTable(mergers) # Construct ASVs table

seqtab.nochim <- removeBimeraDenovo(seqtab, method = "consensus", multithread = TRUE, verbose = TRUE) # Remove chimeric sequences from ASVs table

dim(seqtab.nochim)

sum(seqtab.nochim)/sum(seqtab) # Calculate the percentage of remained sequences after removing the chimera

If this value equals to 1, 0% of chimeric sequences is removed from the ASVs table, and if 0.9 then there are 10% of chimera are removed.

Taxonomy assignment

Download "silva_nr_v132_train_set.fa" and "silva_species_assignment_v132.fa" from https://benjjneb.github.io/dada2/training.html

taxaRC <- assignTaxonomy(seqtab.nochim, "/Users/pwd/Trainset/silva_nr_v132_train_set.fa", tryRC=TRUE) # Assign taxonomy to genus level

taxaSp <- addSpecies(taxaRC, "/Users/pwd/Trainset/silva_species_assignment_v132.fa") # Assign taxonomy to species level

Microbiome analysis

The package phyloseq synthesized and combined data into a phyloseq object containing the ASVs table, taxonomic table, and sample data table, which can be easily manipulated to analyze microbiome.

ps <- phyloseq(otu_table(seqtab.nochim, taxa_are_rows=FALSE), tax_table(taxaSp), sample_data(sampledata)) # Combine the ASVs table, taxonomic table and sample data table into phyloseq object

Analyze the fluctuations of microorganisms at different taxonomic ranks

Filtering phyloseq objects at the "Phylum" level with options

Bacphy<- ps %>%

+ tax_glom(taxrank = "Phylum") %>% # Combined data at the phylum level

+ transform_sample_counts(function(x) {x/sum(x)}) %>% # Convert to a diversity analysis object

+ psmelt() %>% # Combine format

+ filter(Abundance > 0.02) %>% # Filter out low diversity taxa

+ arrange(Phylum) # Arrange data frame alphabetically by phylum

Illustration of the bacterial community composition and abundance at phylum level

ggplot(Bacphy, aes(x = Samples, y = Abundance, fill = Phylum)) # Assign values to axes

+ geom_bar(stat = "identity") + scale_fill_manual(values = phylum_color) # set palette

+ facet_wrap(~feature, nrow=1, scales = "free_x") # Separate by sample features

+ ylab("Relative Abundance (Phylum >0.02%)") # Assign y axis name

+ scale_y_continuous(expand = c(0,0)) # Clear the space below the 0 of the y-axis in the chart

+ ggtitle("Phylum Composition of microbiota") # Assign chart title

Microbiological component representation at two consecutive classification levels (kingdom and phylum)

treemap(Bacphy, index=c("Kingdom", "Phylum"), vSize="Abundance", type="index",

fontsize.labels=c(15,12), # Size of the label. Give size for each aggregation level: size for group, size for subgroup

fontcolor.labels=c("white","black"), # The color format of the labels

fontface.labels=c(2,1), # Label font: 1,2,3,4 for lowercase, bold, italic, bold italics...

bg.labels=c("transparent"), # Background color of label

align.labels=list(c("center", "center"),

c("left", "bottom")), # Where to put labels in rectangles?

overlap.labels=0.5, # Determine the tolerance for overlap between labels

inflate.labels=F,#If true, the label is larger when the rectangle is larger

palette = "Set1", # Choose a palette from the RColorBrewer presets or make your own

fontsize.title=12)

All of these above works were executed and performed in R studio by R packages and command lines to reduce software installation costs, increase compatibility and synchronization for metagenomic data. This is the advantage of R programming language which make it preeminent comparing to other bioinformatic tools.

RESULTS AND DISCUSSION

Table 1. Filterandtrim statistic results of forward reads.

	Reads.in	Reads.out	Frequency (%)
SRR8491659	23013	20738	90.11
SRR8491657	37786	31047	89.12
SRR8491660	28360	26007	91.70
SRR8491663	32697	30103	92.07
SRR8491658	37158	33595	90.41
SRR8491656	39483	35678	90.36

Initially, the low-quality sequencing reads (QC<30) were cleaned from the dataset by the *filterandtrim* function. The results are shown in table 1: 8-10% of the number of reads were removed from the forward sequence data for all samples. After filtering and trimming, the cleaned database was verified again and processed in further downstream analyses, which were all executed by R packages and command lines, as opposed to using bioinformatics tools and software.

Subsequently, in the DADA2 pipeline, the filtered data was dereplicated by inferring unique sequences - amplicon sequence variants (ASVs) with their abundances (the number of reads with that ASV) – from sequencing reads. At the same time, the DADA2 sequence inference step also estimated the error rates of those filtered sequences and removed all substitution and errors from the data. After that, the forward and reverse reads from the inferred unique sequences were merged to construct the sequence table containing the ASVs and the number of times those sequences appear in each sample. Then, about 35% of inferred sequences, which was identified as chimeric sequences by DADA2, was removed from the sequence table. By using training sets from Silva reference database, we obtained the taxonomic assignments for bacteria living in the coral *Acropora tenuis* (Table 2). These bacteria have been identified from previous studies as regular occupants of coral reefs. From this taxonomic table, the composition and abundance of the bacterial associations were accessed to analyze and visualize the fluctuations of bacterial communities at different taxonomic levels by the *phyloseq* and *ggplot2* packages of R.

The results in panel A, Figure 1 represent the composition and abundance of the bacterial association at the phylum level. The phylum Proteobacteria dominates the community in all six samples with a percentage of higher than 70% of total identified bacteria, while the second phylum in terms of quantity is Bacteroidetes. Notably, the phylum Epsilonbacteraeota accounts for a notable percentage in sample 2 and the phylum Dadabacteria in sample 3, contributing to the community diversity. At class level (Panel B), the bacterial classes with the greatest contributions are Gammaproteobacteria (phylum Proteobacteria), Alphaproteobacteria (phylum Proteobacteria) and Bacteroidia (phylum Bacteroidetes). In addition, a noticeably greater proportion of the class Campylobacteria (phylum Epsilonbacteraeota) and Dadabacteriia (phylum Dadabacteria) bacteria was contained in samples 2 and 3, at 8 and 3% respectively. In summary, the results in Figure 1 demonstrate that the bacterial associations are diverse; and the differences between the bacterial composition in this study and other previously conducted metagenomic characterizations are presumably due to the dissimilar environmental conditions and areas of sampling.

Table 2. The first 20 bacterial species identified in the coral *Acropora tenuis.*

Kingdom	Phylum	Class	Order	Family	Genus	Species
Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Nitrincolaceae	Neptuniibacter	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Nitrincolaceae	Neptuniibacter	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Methylobacterium	*NA*
Bacteria	Bacteroidetes	Bacteroidia	Cytophagales	Cyclobacteriaceae	Reichenbachiella	*NA*
Bacteria	Bacteroidetes	Gammaproteobacteria	Oceanospirillales	Alcanivoracaceae	Alcanivorax	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Marivita	cryptomonadis
Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Cutibacterium	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Idiomarinaceae	Idiomarina	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Aestuariibacter	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomonadaceae	Maricaulis	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Thalassobius	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Alteromonadales	Colwelliaceae	Thalassotalea	sediminis
Bacteria	Proteobacteria	Alphaproteobacteria	Caulobacterales	Hyphomonadaceae	Maricaulis	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Ralstonia	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Burkholderia-Caballeronia-Paraburkholderia	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseobacter	*NA*
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Erythrobacter	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Saccharospirillaceae	Thalassolituus	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Cellvibrionales	Spongiibacteraceae	Spongiibacter	*NA*
Bacteria	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Saccharospirillaceae	Oleibacter	marinus

After using the Silva training set and species assignment, the coral-associated bacterial data was processed to analyze the microbiome and the diversity of the bacterial community at the genus level (Figure 2). Nine genera *Alteromonas* (phylum Proteobacteria, class Gammaproteobacteria), *Endozoicomona*s (phylum Proteobacteria, class Gammaproteobacteria), *Neptuniibacter* (phylum Proteobacteria, class Gammaproteobacteria), *Thalassolituus* (phylum Proteobacteria, class Gammaproteobacteria), *Oleibacter* (phylum Proteobacteria, class Gammaproteobacteria), *Erythrobacter* (phylum Proteobacteria, class Alphaproteobacteria), *Roseobacter* (phylum Proteobacteria, class Alphaproteobacteria), *Ekhidna* (phylum Bateroidetes, class Bacteroidia), *Ruegeria* (phylum Proteobacteria, class Alphaproteobacteria) were quite abundant in all six samples.

In summary, the six analyzed bacterial associations living in coral *Acropora tenuis* are almost all similar to each other, the number of common bacteria is quite large with great diversity. However, there is still a degree of

variation in composition, such as the genus *Vibrio* making up a notable amount in samples 2 and 5, which has been shown to be an etiological factor of coral bleaching in previous studies.

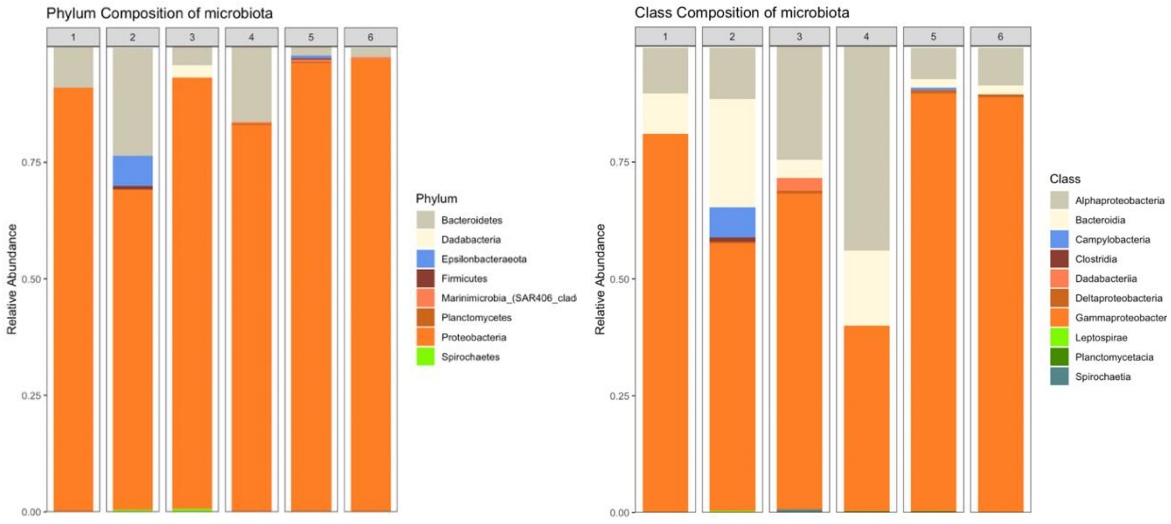

Figure 1. Composition and abundance of bacterial communities at phylum (A) and class (B) level. The bar charts show taxonomic classification of bacterial reads from pooled DNA amplicon from different locations into phylum and class level using DADA2 (classification applied 50% confidence thresholds). The two right corners next to each chart indicates bacterial phyla and classes with the relative abundances bigger than 0.02% of the total bacteria.

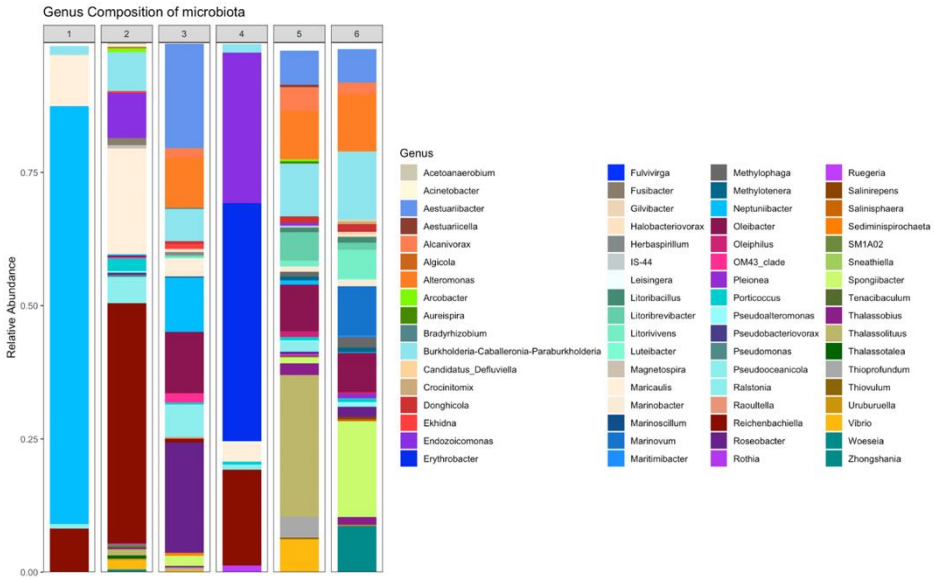

Figure 2. Composition and abundance of viral communities at genus level. The bar charts show taxonomic classification of bacterial reads from pooled DNA amplicon from different locations into genus level using DADA2 (classification applied 50% confidence thresholds). The right corners next to bar chart indicates bacterial genera with the relative abundances bigger than 0.02% of the total bacteria.

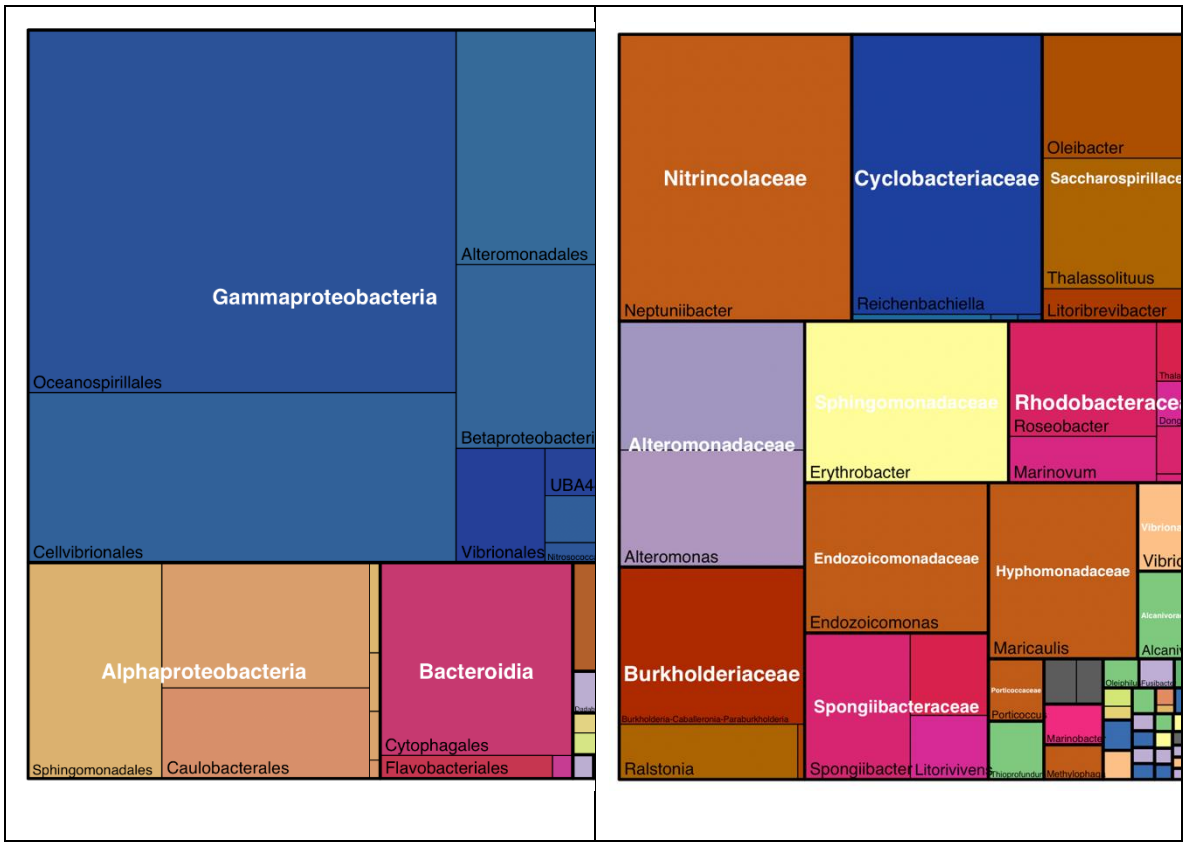

Figure 3. Microbiological component representation at class and order level (A) along with family and genus level (B). The treemap shows bacterial proportion in two consecutive taxonomic ranks (class and order - A, family and genus - B); the white letters indicate bacteria with higher taxonomic ranks and the black are the lower consecutive ranks.

The result of the treemaps in Figure 3 indicates the composition of the bacterial community and also highlights the dominance of certain bacteria at each taxonomic level. For example, in panel A, it shows the bacterial composition at the order level in each class and which order accounts for the highest amounts of total bacteria, similar to family and genus level in panel B. From the results shown, the "treemap" function seems to have remarkable advantages in data visualization, indeed, it provides not only an overview but also highlights the constituents and proportion of bacterial communities in two successive taxonomic levels.

Overall, the composition and abundance of marine bacterial communities were illustrated in details. Being the largest in terms of composition in our study, the phylum Proteobacteria also constitutes the most abundant and diverse microbial group on Earth (Bradley, Pollard, 2017; Spain *et al*., 2009), including pathogenic species such as *Salmonella* (Cortese *et al.*, 2016; Mithal *et al*., 2017), *Campylobacter* (Mithal *et al*., 2017), *Helicobacter*, *Vibrio*, and *Escherichia* (Cortese *et al*., 2016). Moreover, the phylum Epsilonbacteraeota which makes up for a remarkable percentage in sample 2 has recently been reclassified, previously classified as phylum Proteobacteria. Based on the assessment of nearly 300 phylogenetic tree topologies, it was reported that Epsilonproteobacteria within the Proteobacteria is not warranted, and this group was reassigned to a novel phylum under the proposed name Epsilonbacteraeota (phyl. nov.) (Waite *et al.*, 2017). This phylum is crucial

chemolithotrophic primary producers in deep-sea hydrothermal vent systems, in which they are often the dominant bacterial lineage in vent plumes and deposits (Flores *et al.*, 2011; Huber *et al.*, 2010). Dadabacteria is only observed in sample 3. Appealingly, the Dadabacteria have the potential to degrade microbial particulate organic matter, particularly peptidoglycan and phospholipids. The marine Dadabacteria were divided into two clades with distinctive ecological niches in worldwide metagenomic data: a shallow clade with the potential for photoheterotrophy through the utilization of proteorhodopsin, exist principally in surface waters up to 100m depth; and a deep clade that is more varied in the deep photic zone without the photoheterotrophic potential (Graham, Tully, 2020).

In the amplicon workflow, DADA2 discovered more real variants and output fewer spurious sequences and more reference strains than other methods (Callahan *et al*., 2016). This could be the reason why the taxonomic assignment witnessed the most efficiency in DADA2 pipeline, compared to other commonly used algorithms. There was still a variety of unclassified taxonomy which was called "NA" or "Unknown". However, the limitation of assigned microbial species is one of the drawbacks of using sequencing technology for practical analyses of 16 S rRNA metagenomic (Whon *et al*., 2018). This is not specific to DADA2: other popular bioinformatic also algorithms achieved low accuracy in taxa prediction using SILVA database. High within-sample accuracies are rarely achieved at the species level (Escobar-Zepeda *et al*., 2018).

These results have shown how microbial identification at the genus level has can provide useful fundamental information for further in-depth studies. In the future, phylogenetic analysis, functional prediction and phage-host interaction should be performed following analyses similar to ours to broaden understanding about the functions of microbial communities and their interactions in the context of specific habitats.

CONCLUSION

In this study, the taxonomic assignment, composition and diversity of the coral microbiome were successfully accessed and analyzed by the means of bioinformatics, in this case the R language. The metagenomic dataset was processed through the DADA2 pipeline of R in a continuous manner, which is a more optimal approach compared to the use of a typical bioinformatics pipeline. R studio interface grants users more control over the operation, which includes steps such as executing the command line and visualizing the results in graphical forms. All the steps are linked together to ensure the synchronization and compatibility of input and output format of the database via the R object and environment. Therefore, this study has shown the potential of the R programming language in analyzing metagenomic data. Further development and optimization of R and existing bioinformatic tools for data analysis and visualization will aid researchers in understanding the diversity, function, and relationships between microbial communities in different environmental conditions.

Acknowledgement: *This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 106.04-2018.309 and Dalida project under grant number NĐT.37.FRA/18.*

REFERENCES

Balser TC, McMahon KD, BartD, Bronson D, Coyle DR, Craig N, Flores-Mangual ML, Forshay K, Jones SE, Kent AE, Shade AL (2006) Bridging the gap between micro - and macro-scale perspectives on the role of microbial communities in global change ecology. *Plant and Soil* 289(1): 59–70.

Bourne DG, Munn CB (2005) Diversity of bacteria associated with the coral Pocillopora damicornis from the Great Barrier Reef. *Environmental Microbiology* 7(8): 1162–1174.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral

communities. *Proceedings of the National Academy of Sciences* 99(22): 14250 LP – 14255.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13(7): 581–583.

Cortese F, Scicchitano P, Gesualdo M, Filaninno A, De Giorgi E, Schettini F, Laforgia N, Ciccone MM (2016) Early and Late Infections in Newborns: Where Do We Stand? A Review. *Pediatrics and Neonatology* 57(4): 265–273.

Escobar-Zepeda A, Godoy-Lozano EE, Raggi L, Segovia L, Merino E, Gutiérrez-Rios RM, Juarez K, Licea-Navarro AF, Pardo-Lopez L, Sanchez-Flores A (2018) Analysis of sequencing strategies and tools for taxonomic annotation: Defining standards for progressive metagenomics. *Scientific Reports* 8(1): 12034.

Flores GE, Campbell JH, Kirshtein JD, Meneghin J, Podar M, Steinberg JI, Seewald JS, Tivey MK, Voytek MA, Yang ZK, Reysenbach AL (2011) Microbial community structure of hydrothermal deposits from geochemically different vent fields along the Mid-Atlantic Ridge. *Environmental Microbiology* 13(8): 2158–2171.

Graham ED, Tully BJ (2020) Marine Dadabacteria exhibit genome streamlining and phototrophy-driven niche partitioning. *BioRxiv* 2020.06.22.165886.

Huber JA, Cantin HV, HuseSM, Welch DBM, Sogin ML, Butterfield DA (2010) Isolated communities of Epsilonproteobacteria in hydrothermal vent fluids of the Mariana Arc seamounts. *FEMS Microbiology Ecology* 73(3): 538–549.

Kvennefors ECE, Sampayo E, Kerr C, Vieira G, Roff G, Barnes AC (2012) Regulation of bacterial communities through antimicrobial activity by the coral holobiont. *Microbial Ecology* 63(3): 605–618.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology* 31(9): 814–821.

Mithal LB, Palac HL, Yogev R, Ernst LM, Mestan KK (2017) Cord Blood Acute Phase Reactants Predict Early Onset Neonatal Sepsis in Preterm Infants. *PloS One* 12(1): e0168677.

Reshef L, Koren O, Loya Y, Zilber-Rosenberg I, Rosenberg E (2006) The coral probiotic hypothesis. *Environmental Microbiology* 8(12): 2068–2073.

Rosenberg E, Kellogg CA, Rohwer F (2007) Coral Microbiology. *A Sea of Microbe* 20: 146–154.

Thomas T, Gilbert J, Meyer F (2012) Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* 2(1): 3.

Wade W (2002) Unculturable bacteria--the uncharacterized organisms that cause oral infections. *Journal of the Royal Society of Medicine* 95(2): 81–83.

Waite DW, Vanwonterghem I, Rinke C, Parks DH, Zhang Y, Takai K, Sievert SM, Simon J, Campbell BJ, Hanson TE, Woyke T, Klotz MG, Hugenholtz P (2017) Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). *Frontiers in Microbiology* 8: 682.

Whon TW, Chung WH, Lim MY, Song EJ, Kim PS, Hyun DW, Shin NR, Bae JW, Nam YD (2018) The effects of sequencing platforms on phylogenetic resolution in 16 S rRNA gene profiling of human feces. *Scientific Data* 5(1): 180068.

Journal of Biotechnology **18**(4): 733-743, 2020

CÁCH TIẾP CẬN TIN SINH HỌC TRONG VIỆC PHÂN TÍCH CÁC QUẦN XÃ VI KHUẨN SỐNG TRÊN SAN HÔ BẰNG NGÔN NGỮ LẬP TRÌNH R

Đoàn Thị Nhung[1], Bùi Văn Ngọc[1,2]

[1]Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
[2]Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

TÓM TẮT

Những tiến bộ gần đây trong nghiên cứu đa hệ gen và tin sinh học đã cho phép các nhà khoa học phân tích tổng thể về đa dạng sinh học, thành phần và số lượng của các quần xã vi sinh vật, cũng như các gen chức năng và các con đường trao đổi chất của chúng. Cho đến nay đã có rất nhiều công cụ và phần mềm tính toán/thống kê để phân tích vi sinh vật. Tuy nhiên, vấn đề gặp phải khá phổ biến là sự thiếu đồng bộ và tính tương thích về các loại định dạng dữ liệu đầu ra/đầu vào giữa các phần mềm với nhau. Để khắc phục trở ngại này, chúng tôi sử dụng DADA2 pipeline được viết trên ngôn ngữ lập trình R để sửa đổi và cải tiến chúng thay vì sử dụng các phần mềm tin sinh học khác thành công cụ phân tích các quần xã vi khuẩn một cách liên tục và đồng bộ. Trong bước đầu nghiên cứu, chúng tôi thử nghiệm phân tích thành phần và số lượng vi khuẩn sống trên san hô dựa trên trình tự gen 16S rRNA của chúng. Quy trình làm việc bao gồm các bước sau: xử lý dữ liệu, phân cụm trình tự, gán đơn vị phân loại, trực quan kết quả. Hơn nữa, chúng tôi kỳ vọng hướng độc giả chú ý đến thông tin rằng các quần xã vi khuẩn sống trong đại dương hầu hết là các vi sinh vật không nuôi cấy được, trong đó có vi sinh vật sống trên san hô nói chung và san hô *Acropora tenuis* nói riêng. Kết quả thu được trong nghiên cứu này cho thấy DADA2 pipeline viết trên ngôn ngữ lập trình R là một trong những công cụ tin sinh học ứng dụng tiềm năng trong lĩnh việc phân tích các quần xã vi sinh vật thay vì sử dụng các phần mềm riêng rẽ khác nhau. Bên cạnh đó, các sửa đổi trong quy trình làm việc của chúng tôi cũng giúp các nhà nghiên cứu dễ dàng minh họa một cách có hệ thống các dữ liệu metagenomic, làm sáng tỏ thành phần, sự phong phú, sự đa dạng và mối quan hệ giữa các quần xã vi sinh vật, cũng như để phát triển các công cụ tinh sinh học khác một cách hiệu quả.

Từ khoá: *16S rRNA, Acropora tenuis, tin sinh học, vi khuẩn sống trên san hô, ngôn ngữ lập trình R.*