The North Atlantic Aerosols and Marine Ecosystems Study (NAAMES) microbial genetic profiling using 16S rRNA amplicon high-trhoughput sequencing

Luis M. Bolaños

June 2020

Contents

1 Introduction

This document was created to provide a methodological framework of the analysis of four 16S rRNA amplicon data sets generated as part of the The North Atlantic Aerosols and Marine Ecosystems Study (NAAMES). Each of the four data sets represent the collection of genetic profiles targeting specific seasonal events in the annual plankton cycle (https://doi.org/10.3389/fmars.2019.00122 for more information). These data sets will be referred as:

- **NAAMES1**: Campaign 1 Winter Transition (November 2015). 56 microbial biomass samples were collected from 7 stations at 8 different depths (5m, 25m, 50m, 75m, 100, 150, 200, 250m and 300m).
- **NAAMES2**: Campaign 2 Climax Transition (May 2016). 64 microbial biomass samples were collected from 5 stations at 8 different depths (5m, 25m, 50m, 75m, 100, 150, 200, 250m and 300m). Station 4 was sampled daily during a four-day occupation (4 depth profiles).
- **NAAMES3**: Campaign 3 Declining Phase (September 2017). 112 microbial biomass samples were collected from 11 stations at 8 different depths (5m, 25m, 50m, 75m, 100, 150, 200, 250m and 300m). Station 6 was sampled daily during a four-day occupation (4 depth profiles).
- **NAAMES4**: Campaign 4 Accumulation Phase (March-April 2018). 40 microbial biomass samples were collected from 5 stations at 8 different depths (5m, 25m, 50m, 75m, 100, 150, 200, 250m and 300m).

2 Samples, data generation and raw data availability

16S rRNA amplicon sequencing was performed on libraries made of amplicons using 27F (5'-AGAGTTTGATCNTGGCTCAG-3) and 338 RPL (5'- GCWGCCWCCCGTAGGWGT-3') primer set. For more information: https://doi.org/10.1038/s41396-020-0636-0 and https://github.com/lbolanos32/Phyto_NAAMES_2019. Raw 16S rRNA datasets are publicly available at:

- **NCBI SRA.** BioProject: PRJNA627189 (SRA accession numbers SRR11596858 to SRR11597110)
- **NASA SeaWiFS Bio-optical Archive and Storage System (SeaBASS).** Amplicon Sequences download Instructions

1) Go to https://seabass.gsfc.nasa.gov/investigator/Giovannoni,%20Stephen

2) Click the blue [search] button

3) Check the "Include all associated files" box.

4) Click download all

Questions? please contact

- Stephen Giovannoni (Steve.giovannoni@oregonstate.edu) or
- Luis M Bolanos (bolanosl@oregonstate.edu, lbolanos@lcg.unam.mx)

Citation:

- If curated phytoplankton fraction from "https://github.com/lbolanos32/Phyto_NAAMES_2019" is used, plase cite: "Small Phytoplankton Dominate Western North Atlantic Biomass" (https://doi.org/10.1038/s41396-020-0636-0).
- If data sets are used in any other way, you followed this pipeline or used the attached scripts, please cite: "Seasonality of the microbial community composition in the western North Atlantic".

3 Pipeline: From Raw Data to ASV tables

The following scripts were used sequentially to achieve four ASVs tables, one for each sequencing run. Samples from each campaign were sequenced in one Illumina lane. Processed tables, product of this pipeline, can be found along this document.

Task 3.1 Pre-processing sequence files: CUTADAPT

Cutadapt is used to chop the primers from the raw sequences. In this SOP, we are using a fixed number of bp to trim each paired-end. The fixed number match the primer length: 27F (20bp) and 338RPL (18bp) (**trimf.sh** and **trimrev.sh**) to chop the fixed number of bp from the raw sequences

Task 3.2 Run DADA2 version 1.2

We used the script **DadaR.R** to generate an ASV table coupled with taxonomic assignation using SILVA database train version 123 (silva_nr_v123_train_set.fa) and the R version used was R-3.4.1 DadaR.R This script generates an ASV (seqtab-nochimtaxa.txt)table from the trimmed reads.

Task 3.3 Parsing Dada output

"seqtab-nochimtaxa.txt" files use unique sequence as row name (identifier) and the assigned taxa is shown in the last columns (SILVA hierarchical format). We parsed the names with **parseNames.pl** which outputs the file seqtab-par.txt.tmp. Then we added sequential field-specific identifiers using **addcolnm.pl** which outputs seqtab-par.txt.tmp.co

*This can be automatized and run it for the four data sets.

4 Pipeline: From ASV tables to the merged dataset (used in 'Seasonality of the microbial community composition in the western North Atlantic')

We created a single ASV table merging the N1N2, N3 and N4 tables by the ASV sequence. ASV tables for each campaign were generated using dada2 script (DadaR.R, above). It is done individually due to the error frequency estimation, which underlying assumption is that each sequencing run have different errors.

Listing 1: parsing the files for merging

```
Create files without taxa and merge them using the sequence (
    first row)
cut -f 1-112 N1N2/seqtab-par.txt.tmp > seqtab-parN1N2.txt.tmp #
    NAAMES 1 and NAAMES 2 were analyzed together in the previous
    paper and we are keeping it consistent.
cut -f 1-150 N3/seqtab-par.txt.tmp > seqtab-parN3.txt.tmp
cut -f 1-154 N4/seqtab-par.txt.tmp > seqtab-parN4.txt.tmp
```

Listing 2: R processing

```
count_tab1 <- read.table("Path where the file is located/seqtab-
    parN1N2.txt.tmp", header=T, row.names=1, check.names=F)
count_tab2 <- read.table("Path where the file is located/seqtab-
    parN3.txt.tmp", header=T, row.names=1, check.names=F)
count_tab3 <- read.table("Path where the file is located/seqtab-
    parN4.txt.tmp", header=T, row.names=1, check.names=F)

count_tab_phy1 <- otu_table(count_tab1, taxa_are_rows=T)
count_tab_phy2 <- otu_table(count_tab2, taxa_are_rows=T)
count_tab_phy3 <- otu_table(count_tab3, taxa_are_rows=T)

OTU_physeqN1N2 <- phyloseq(count_tab_phy1)
OTU_physeqN3 <- phyloseq(count_tab_phy2)
OTU_physeqN4 <- phyloseq(count_tab_phy3)
```

```
NAAMESphyseq<-merge_phyloseq(OTU_physeqN1N2,OTU_physeqN3,OTU_
    physeqN4)
write.table(otu_table(NAAMESphyseq),file= "/Users/luisbolanos/
    Documents/OSU_postdoc/NAAMES/MergeAll/seqtab-parNAAMES.txt.
    tmp", quote=FALSE, sep = "\t")

seqtab-parNAAMES.txt #is a merged ASV table
seqtab-parNAAMES.txt = total with repeats 412 samples #Repeats
    are N3 samples re-sequenced in the NAAMES 4 line. So we
    removed the low quality repeats below and some N2 which we
    already knew were low quality from the previous paper (
    Bolanos et al., 2020).
```

List of removed samples from seqtab-parNAAMES.txt (NEW FILE seqtab-parNAAMES.tab should have 375)

- N3S1-5_S132
- N3S1-150_S133
- N3S1Ml1_S134
- N31a-150_S135
- N3S2Ml1_S137
- N3S3-75_S140
- N3S3-150_S142
- N3S3d1-8_S144
- N3S4Ml1_S147
- N3S4-5-200_S148
- N3S4-5Ml1_S149
- N3S5-150_S150
- Undetermined_S0
- NAAMES2-20_S45
- NAAMES2-23_S47
- NAAMES2-32_S53
- NAAMES2-62_S58
- N3S1-5_S1
- N3S1-150_S6
- N3S1-Ml1_S9
- N3S1a-150_S16
- N3S1-5-Ml1_S27
- N3S2-Ml1_S36
- N3S3-5_S38
- N3S3-25_S39
- N3S3-75_S41
- N3S3-100_S42
- N3S3-150_S43
- N3S3-Ml1_S46
- N3S3d1-8_S50
- N3S3d2-3_S52
- N3S3-5-100_S60
- N3S4-Ml1_S73
- N3S4-5-200_S81
- N3S4-5-Ml1_S83
- N3S5-150_S89
- N3S5-200_S90
- N3S6C4-25_S103

Task 4.1 Parsing and formatting to obtain the final taxonomy

- Add sequential IDs using the **addcolnm.pl** script to seqtab-parNAAMES.tab and generate seqtab-parNAAMES.txt.tmp.co
- add taxonomy from the dada2 output to create seqtab-par.NAAMES.tax.tab.txt

- removing non 16S (SILVA assign these as ”Eukaryota”)

```
grep -vw "Eukaryota" seqtab-par.NAAMES.tax.tab.txt > seqtab-
    par.NAAMESclean.tax.tab.txt
cut -f 1,2 seqtab-par_on16.txt | sed "s/^N/>N/" | sed "s/\t
    /\n/" > seqtab-par.NAAMESclean.tax.tab.fa
```

- Splitting photosynthetic from heterotrophic

Phytoplankton fraction

Taxonomic assignment was done as in Bolanos et al., 2020. Photosynthetic fractions for each campaign were annotated and collapsed using custom databases. Final files are provided along this document.

After aligning the phytoplankton ASVs, we manually removed sequences which were not correctly aligned (potential misamplifications).

- N36924
- N13977
- N35961
- N20866
- N12585
- N11320
- N30700
- N34104
- N29233
- N13881
- N34339
- N36715
- N33444
- N35157
- N34541
- N36383
- N35218
- N21599
- N22708
- N35953
- N36824
- N35909
- N30104
- N32625
- N35101
- N30022
- N33118
- N35337

```
grep -wf Photlink_final_collapsed.lst seqtab-parNAAMESclean.
    tax.tab.txt > Photlink_final_collapsedVR2.otu #This VR2
    is the file without the above sequences
```

Heterotrophic fraction

```
grep -vwf Photlink_final_collapsed.lst seqtab-parNAAMESclean
    .tax.tab.txt > Hetlink_final_collapsedV.otu

cut -f 1,3-377 Hetlink_final_collapsedV.otu >
    Hetlink_final_collapsedV2.otu #Get the Heterotrophic ASV
    Table

cut -f 1,2 Hetlink_final_collapsedV.otu | sed "s/^N/>N/" |
    sed "s/\t/\n/g" > Hetlink_seq.fasta
tail -n +3 Hetlink_seq.fasta > tmp1_rm
mv tmp1_rm Hetlink_seq.fasta

cut -f 1,378-383 Hetlink_final_collapsedV.otu >
    Hetlink_temptax.txt #Get the Heterotrophic Taxonomic
    Table to add highly defined SAR11 and SAR202
```

- Adding SAR11 and SAR202 highly defined taxonomy

```
grep "SAR11_clade" seqtab-par.NAAMESclean.tax.tab.txt | tail
     -n +1 | cut -f 1,2 | sed "s/^N/>N/" | sed "s/\t/\n/" >
    SAR11_merge.fasta

#SAR202
grep "SAR202" seqtab-par.NAAMESclean.tax.tab.txt | tail -n
    +1 | cut -f 1,2 | sed "s/^N/>N/" | sed "s/\t/\n/" >
    SAR202_merge.fasta

#Run Phyloassigner with a custom SAR11 DB
perl phyloassigner.pl --hmmerdir --pplacerdir -o SAR11_allN
    /datasets/SAR11_full/Sar11CoreV3_aln_inpclean1_10.
    phyloassignerdb/ SAR11_merge.fasta

#NOTE: Sar11CoreV3_aln_inpclean1_10.phyloassignerdb/ and how
     was it created is provided in this repository. Check
    folder SAR11DB

#Run Phyloassigner with a custom SAR202 DB
perl phyloassigner.pl --hmmerdir --pplacerdir -o SAR202_allN
     /datasets/SAR202/ss_aln.phyloassignerdb/ SAR202_merge.
    fasta

cut -f 1,3 SAR202_allN/SAR202_merge.fasta.aln.jplace.tab |
    tail -n +2 > sar202_Nall.twocols

cut -f 1,3 SAR11_allN/SAR11_merge.fasta.aln.jplace.tab |
    tail -n +2 > sar11_Nall.twocols

#Use the perl script substPAon16.pl to update the taxonomy
    table
perl substPAon16.pl Hetlink_temptax.txt sar202_Nall.twocols
    > HetlinkTax_202.txt
perl substPAon16.pl HetlinkTax_202.txt sar11_Nall.twocols >
    HetlinkTax_202_11.txt
```

5 Pipeline: R analysis of the merged dataset

Final datasets are:

- **Photlink_final_collapsedVR2.otu** (Photosynthetic ASV table)
- **Photlink_final_collapsedVR.tax** (Photosynthetic Taxa table)

- **Hetlink_final_collapsedV2.otu** (Heterotrophic ASV table)
- **HetlinkTax_202_11.txt** (Heterotrophic Taxa table)
- **Allcat.otu.txt**(Concatenated Phot+Het ASV table)
- **Allcat.tax** (Concatenated Phot+Het Taxa table)
- **Merge_envFileDNA.txt** (Environmental data)
- **CoordsStV2.txt** (formatted station coordinates)

Task 5.1 FIGURE 1: MAPS

```r
library(ggplot2)
library(ggmap)
library(maps)
library(mapdata)

samps <- read.table("/Users/luisbolanos/Documents/MisDrafts/
    InProgress/NAAMES_annual/CoordsStV2.txt", header=T,sep="\t")

#Get the row number
N1<-which(samps$cruise=="NAAMES1")
N2<-which(samps$cruise=="NAAMES2")
N3<-which(samps$cruise=="NAAMES3")
N4<-which(samps$cruise=="NAAMES4")

#NAAMES1
svg("N1map.svg")
image(x=-75:-20, y = 30:60, z = outer(0, 0), xlab = "lon", ylab =
     "lat")
map("world", add = TRUE, fill=TRUE,bg='light blue')
points(samps[N1,3], samps[N1,2], pch=19, col="blue4", cex=1.5,
    type="p")
dev.off()

#NAAMES2
svg("N2map.svg")
image(x=-75:-20, y = 30:60, z = outer(0, 0), xlab = "lon", ylab =
     "lat")
map("world", add = TRUE, fill=TRUE,bg='light blue')
points(samps[N2,3], samps[N2,2], pch=19, col="darkgreen", cex
    =1.5, type="p")
dev.off()

#NAAMES3
```

```
svg("N3map.svg")
image(x=-75:-20, y = 30:60, z = outer(0, 0), xlab = "lon", ylab =
     "lat")
map("world", add = TRUE, fill=TRUE,bg='light blue')
points(samps[N3,3], samps[N3,2], pch=19, col="firebrick", cex
    =1.3, type="p")
dev.off()

#NAAMES4
svg("N4map.svg")
image(x=-75:-20, y = 30:60, z = outer(0, 0), xlab = "lon", ylab =
     "lat")
map("world", add = TRUE, fill=TRUE,bg='light blue')
points(samps[N4,3], samps[N4,2], pch=19, col="black", cex=1.5,
    type="p")
dev.off()
```

Task 5.2 FIGURE 2: Ordination

R script (OrdinationNAAMES.R) provided

Task 5.3 FIGURE 3: Phytoplankton Community Composition

R script (PhytoCCNAAMES.R) provided

Task 5.4 FIGURE 4: Heterotrophic Bacteria Community Composition

R script (HetCCNAAMES.R) provided

Task 5.5 FIGURE 5: SAR 11 Community Composition

R script (SAR11NAAMES.R) provided

Task 5.6 FIGURE 6: ASVs modularity and relative contributions

Network and modularity analysis is provided as an R script (this includes Fig S3)

Task 5.7 FIGURE S1: Cladogram of SAR11 phylogenetic tree used as database for taxonomic placement

A SAR11 phylogenetic tree was built based on full-length 16S rRNA retrieved from SILVA DB. (find more in lbolanos32/NAAMES_2020/SAR11_Phy_DB/)

A) Data Retrieval

SAR11_1167seqs_SILVA138.fasta is the original file with the sequences downloaded from SILVA with the following parameters:

- length >1400 (only full length, in that way we know we are covering the V1 and V2)
- Qseq >90
- Qaln >90
- Qpintail >90

B) Cleaning, aligning, cropping and QC

- B.1) Sequences were aligned usign Clustalw and misaligned sequences were manually removed.
- B.2) DNA dist was estimated for the cropped sequences, pairs >99.5 identical were de-multiplexed and only one representative was used
- B.3) Final alignment **Sar11CoreV3_aln_inpclean1_10.phy** is provided in the SAR11DB directory

C) Phylogenetic reconstruction using raxmlHPC

```
raxmlHPC -f a -100 -m GTRGAMMA -x 345 -p 678 -s Sar11CoreV3_aln_inpclean1_10.phy -n ar11CoreV3_aln_inpclean1_10.tree -o Gmet_R0046
```

D) setting the phyloassigner DB

```
perl phyloassigner-6.166/setupdb.pl Sar11CoreV3_aln_inpclean1_10.nwk Sar11CoreV3_aln_inpclean1_10.phy 'Gmet_R0046' –nopack –pplacerdir /raid1/home/micro/bolanosl/local/source/phyloassigner-6.166/binaries/
```

Task 5.8 FIGURE S2: Dendogram and Heatmap of phytoplankton communities

R script (Dendro_Phyto.R) provided

Task 5.9 FIGURE S3: Network analysis

6 List of Tables provided

- N1N2_seqtab-par.txt.tmp (original output for N1 and N2 DADA2 run)
- N3_seqtab-par.txt.tmp (original output for N3 DADA2 run)
- N4_seqtab-par.txt.tmp (original output for N4 DADA2 run)

- seqtab-parNAAMES.txt (ASV table merged by sequence from N1N2, N3 and N4,, provided upon request as it is too large for github)

- seqtab-par.NAAMESclean.tax.tab.fa (working sequences)

- seqtab-par.NAAMESclean.tax.tab.txt (working ASV table with New IDs and taxa assignment, provided upon request as it is too large for github)

- Photlink_final_collapsed.lst (list of phytoplankton ASVs)

- Photlink_final_collapsedVR.tax (phytoplankton taxonomic assignment)

- Photlink_final_collapsedVR2.otu (phytoplankton ASV table)

- Merge_envFileDNA.txt (working environmental file)

- Allcat.env (concatenated het+phot environmental file)

- Allcat.tax (concatenated het+phot tax table)

- Allcat.otu (concatenated het+phot ASV table)

- Hetlink_final_collapsedV2.otu (Het Bacteria ASV table)

- HetlinkTax_202_11.txt (Het Bacteria tax table)

- Taxa_custom.tax (Phytoplankton custom taxonomic names)

- barplots_addposition_min1600.txt (table required for fig 3)

- SAR11V2 taxa, custom file to create FIG 5

7 Other objects provided

- EXTRA: netNAAMESall.rds This file contains the co variance associations computed for the network analysis. This file is required to fully reproduce the NAAMESNTWK.R and the manuscript figures. Or it can be computed from scratch running the commented section in R file.