Tara_analysis

TARA Oceans Data Analysis

Setup and Package Loading

Importing CSV Files

```
piconano = read.csv("Data/tara_phytoplankton_piconano_forhomework_092821.csv")
sample = read.csv ("Data/tara_sample_information_092821.csv" )
```

Creating Merged, pivoted dataframes

```
#here we are pivoting the columns containing TARA such that they are data fields.
#made sure that column names are the same between the two data frames so they can be joined!
piconano_long <- piconano %>%
  pivot_longer(
    cols = starts_with("TARA"),
    names_to = "sampleID"
  )
#We are merging them based on the piconano arrangement!
sample_merge <- left_join(piconano_long, sample, by="sampleID")
```

Homework Questions

Question 1:

Determine the ten most abundant OTUs, using the column 'totab.piconano', which is the sum of all reads of each OTU across all samples.

Creating top 10 list

```
top_10_piconano <- piconano %>%
  # Remove duplicate taxogroups, keeping the one with highest abundance
  arrange(desc(totab.piconano)) %>%
  distinct(taxogroup, .keep_all = TRUE) %>%
  # Get top 10
  slice_head(n = 10)

#pivoting this data as well for calculations later
pivoted_top_10_piconano <-  pivot_longer(top_10_piconano,
    cols = starts_with("TARA"),
    names_to = "sampleID"
  )
```

BLASTING sequences

So based on our results here we can pull out info from the relevant sequence columns and perform a nucleotide BLAST

Plain text will be below but I am also going to create a dataframe of this data from a csv

```
list_blast_data = read.csv("Data/top_ten_otus.csv")
```

- Sequence 1 (cid = 161224):
  - Most closely related cultured isolate: (Per Ident = 100%) LC009879.1 - Dictyochophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) JQ781966.1 - Stramenopiles
- Sequence 2 (cid = 171085):
  - Most closely related cultured isolate: (Per Ident = 94.62%) MW750340.1 - Bacillariophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KJ757845.1 - uncultured Eukaryote - no info
- Sequence 3 (cid = 83541):
  - Most closely related cultured isolate: (Per Ident = 100%) MT760788.1 - Prymnesiophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KP404791.1c - uncultured Eukaryote
- Sequence 4 (cid = 146146):
  - Most closely related cultured isolate: (Per Ident = 100%) AY254857.1 - Dictyochophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KC583002.1 - uncultured Stramenopile
- Sequence 5 (cid = 73122):
  - Most closely related cultured isolate: (Per Ident = 90.15) HG973006.1 - Trebouxiophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KJ760716.1 - uncultured Eukaryote
- Sequence 6 (cid = 178099):
  - Most closely related cultured isolate: (Per Ident = 100%) MZ611704.1 - Prymnesiophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) HM581637.1 - uncultured marine eukaryote
- Sequence 7 (cid = 231524):
  - Most closely related cultured isolate: (Per Ident = 100%) AB058358.1 - Prymnesiophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KF62097c1.1 - Haptophyta
- Sequence 8 (cid = 78011):
  - Most closely related cultured isolate: (Per Ident = 100%) JF791030.1 - Cryptophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) KF130577.1 - uncultured eukaryote
- Sequence 9 (cid = 211092):
  - Most closely related cultured isolate: (Per Ident = 100%) KY368637.1 - Mamiellophyceae
  - Most closely related uncultured isolate: (Per Ident = 100%) FR874356.1 - uncultured marine picoeukaryote
- Sequence 10 (cid = 40687):
  - Most closely related cultured isolate: (Per Ident = 90.08%) AB081517.1 - Dictyochophyceae

- Most closely related uncultured isolate: (Per Ident = 90.15%) DQ647511.1 - uncultured marine eukaryote

Taxonomic Class Comparisons

Compare this information to the 'lineage' and 'taxogroup' columns in the dataset, which contain the taxonomic annotation performed by the Tara group. How much taxonomic diversity is represented by the top 10 OTUs of piconanoeukaryote phytoplankton (according to these methods)?

```
taxocomparison <- data.frame(list_blast_data$cult.class, list_blast_data$uncult.class, unique(top_10_pi

length(unique(taxocomparison$list_blast_data.cult.class))
```

```
## [1] 6
```

```
length(unique(taxocomparison$list_blast_data.uncult.class))
```

```
## [1] 6
```

```
length(unique(taxocomparison$top_10_piconano.taxogroup))
```

```
## [1] 0
```

```
length(unique(taxocomparison$top_10_piconano.lineage))
```

```
## [1] 0
```

Based on the above comparisons, I would conclude that the TARA group found more taxonomic diversity based on their methods, than by our top ten analysis.

Question 2

Let's look at some variation in distribution among the top ten OTUs. The Tara group took samples at two depths, 'surface' (~3-7 meters, coded SUR) and deep chlorophyll maximum (coded DCM). The column names of the samples in the first file correspond to the Pangaea Accession Number column in the second file, which contains metadata on the samples. Merge the information in these two files in order to classify the OTU samples as surface or DCM. If you do this in Excel, transposing the data and sorting will be helpful. If you do this in R, the function pivot.longer() in the package tidyr, and merge(), may be helpful.

Merging information

Calculating Relative Abundance

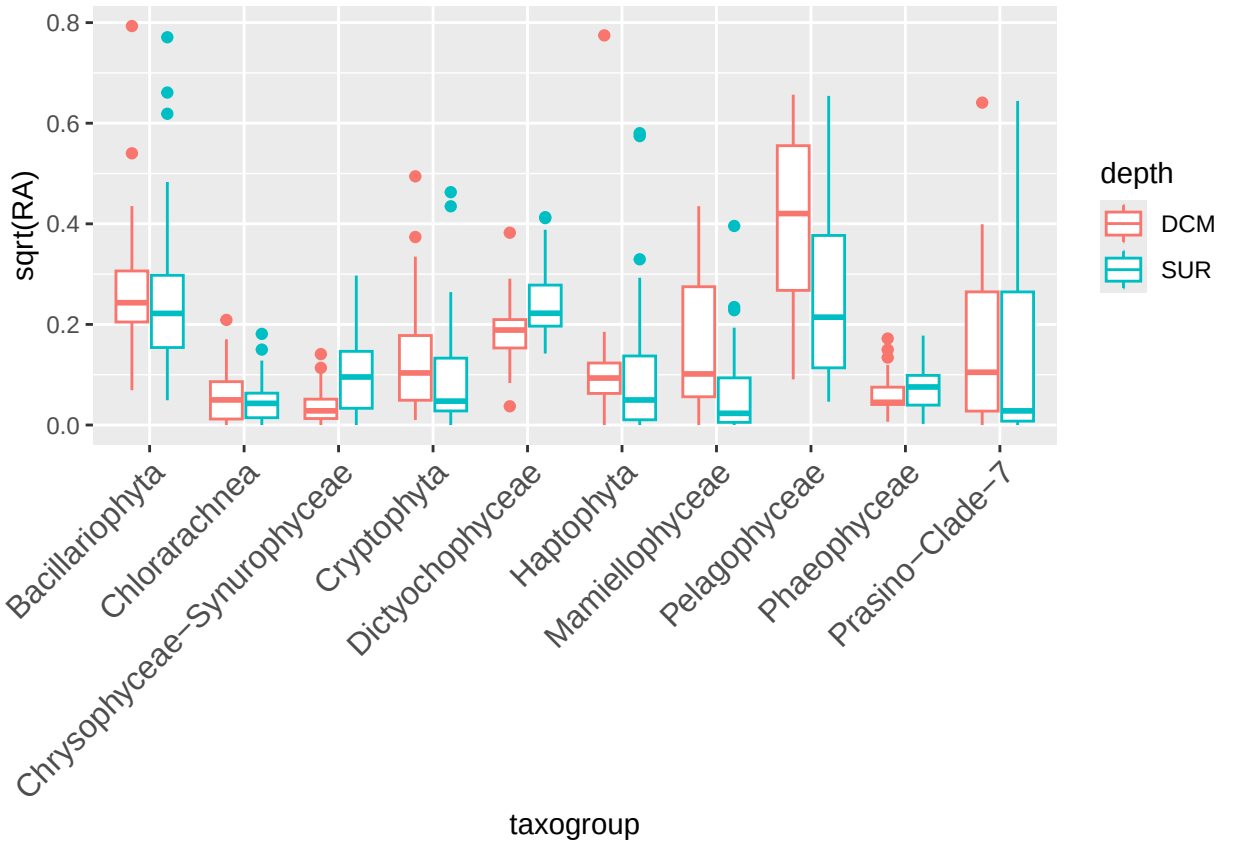
```
top_10_merge <- left_join(pivoted_top_10_piconano, sample, by = "sampleID")

abun_merge <- top_10_merge %>%
  select(value, total.reads, depth, taxogroup) %>%
  mutate(
    RA = value / total.reads,
    rootRA = sqrt(RA)
  )
```

Here I have created a new data frame and added two columns, one with the calculated value and the other is root transformed!

Making Boxplot

```
OTUabund <- ggplot(abun_merge, aes(x=taxogroup, y=sqrt(RA), color=depth))  + geom_boxplot()
OTUabund + theme(
  axis.text.x = element_text(angle = 45, hjust = 1, size = 12)
)
```



Question 3: Some Hypotheses

It's hard for me to be confident in my produced graph, since It is defying my expectations. I would expect that in the DCM there would be an *increased* relative abundance relative to the surface water due to both grazing pressure and population density (it is the deep chlorophyll maximum after all). I think perhaps I have made some errors in my calculations above.

Some hypothesis for how environmental factors could cause shifts in community composition.

- I think one of the most consequential environmental factors is most likely nitrogen availability, which would limit an organisms ability to maintain it's metabolism under limitation. We know from our lectures that the concentration increases with depth as primary production slows down.
  - ex situ perhaps using environmental samples we could alter the physical condition of light (light dark bottles in the lab?) and note how diversity and thus nutrient concentrations change over time?
  - I think a good in situ experimental approach might be some kind of "Nitrate Ocean Timeseries" (NOT?) where we measure nitrogen concentrations and associate them with microbial diversity.
- Adding on to the previous hypothesis, how could we demonstrate that *nitrifying archaea* play a key role in the availability of nitrates in the DCM?

- ex situ this could be demonstrated by using the aformentioned environmental samples and comparing the production of nitrate under low light conditions with control and archaea added experimental groups - who produces more nitrate?
- This could be demonstrated in situ through the use of metagenomics techniques looking for organisms who possess known genes for nitrification, identifying them taxonomically, and comparing nutrient availability - are the archaea there? are they nitrifying? and and is nitrate indeed elevated?