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. I am going to use the raw piconano .csv for this, because I am just interested in determining the highest number of totab - better to do that with the unpivoted data.

Creating top 10 list

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

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, (top_10_piconano

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

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

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

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

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

```
## [1] 7
```

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

```
## [1] 10
```

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.

Calculating Relative Abundance

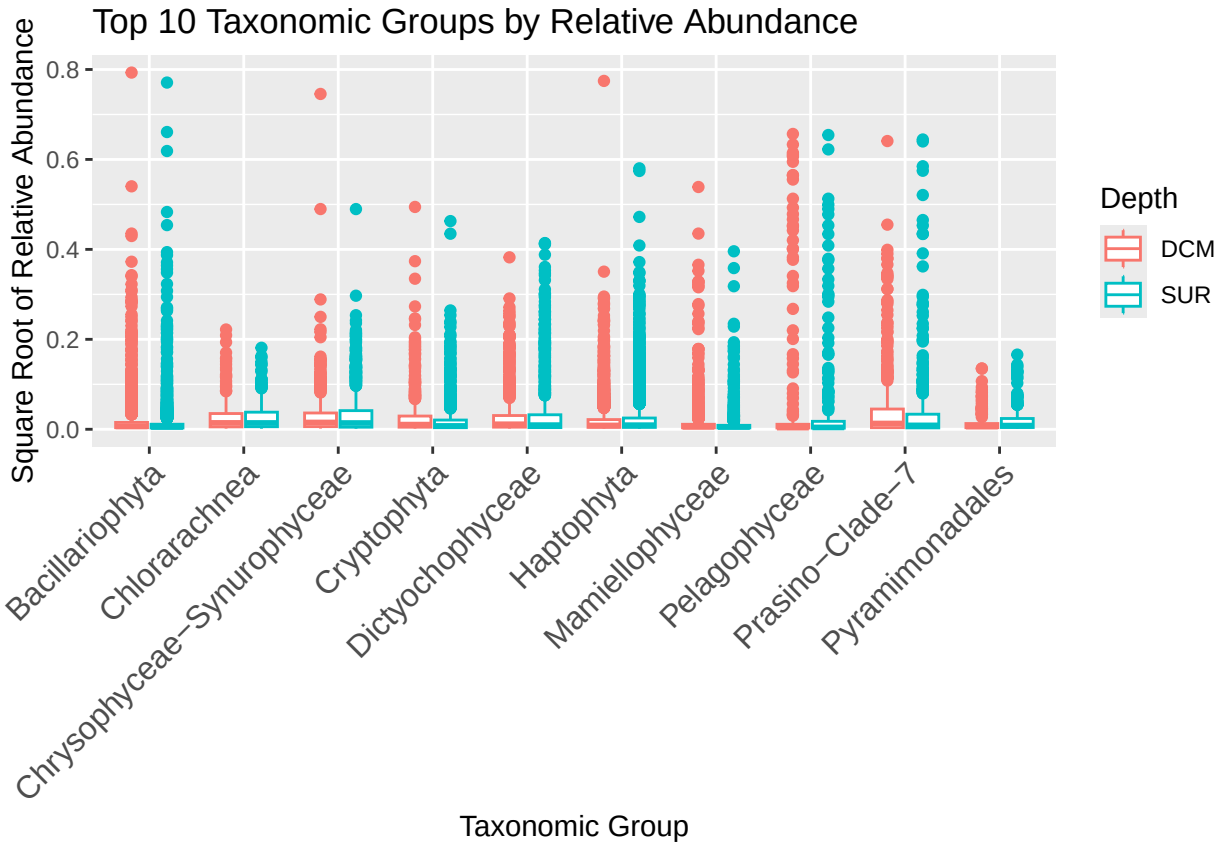
```
# First some prep, we need to throw out the data where the "value" are 0 or it will skew our data like
abun_merge <- sample_merge %>%
  filter(value != 0) %>%
  select(value, total.reads, depth, taxogroup, cid, totab.piconano) %>%
  mutate(
    RA = value / total.reads,
    rootRA = sqrt(RA)
  )
```

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

```
# Identify the top 10 taxogroups by total relative abundance
top_taxogroups <- abun_merge %>%
  group_by(taxogroup) %>%
  summarise(total_RA = sum(RA, na.rm = TRUE)) %>%
  slice_max(total_RA, n = 10, with_ties = FALSE) %>%  # Added with_ties = FALSE for clarity
  pull(taxogroup)
```

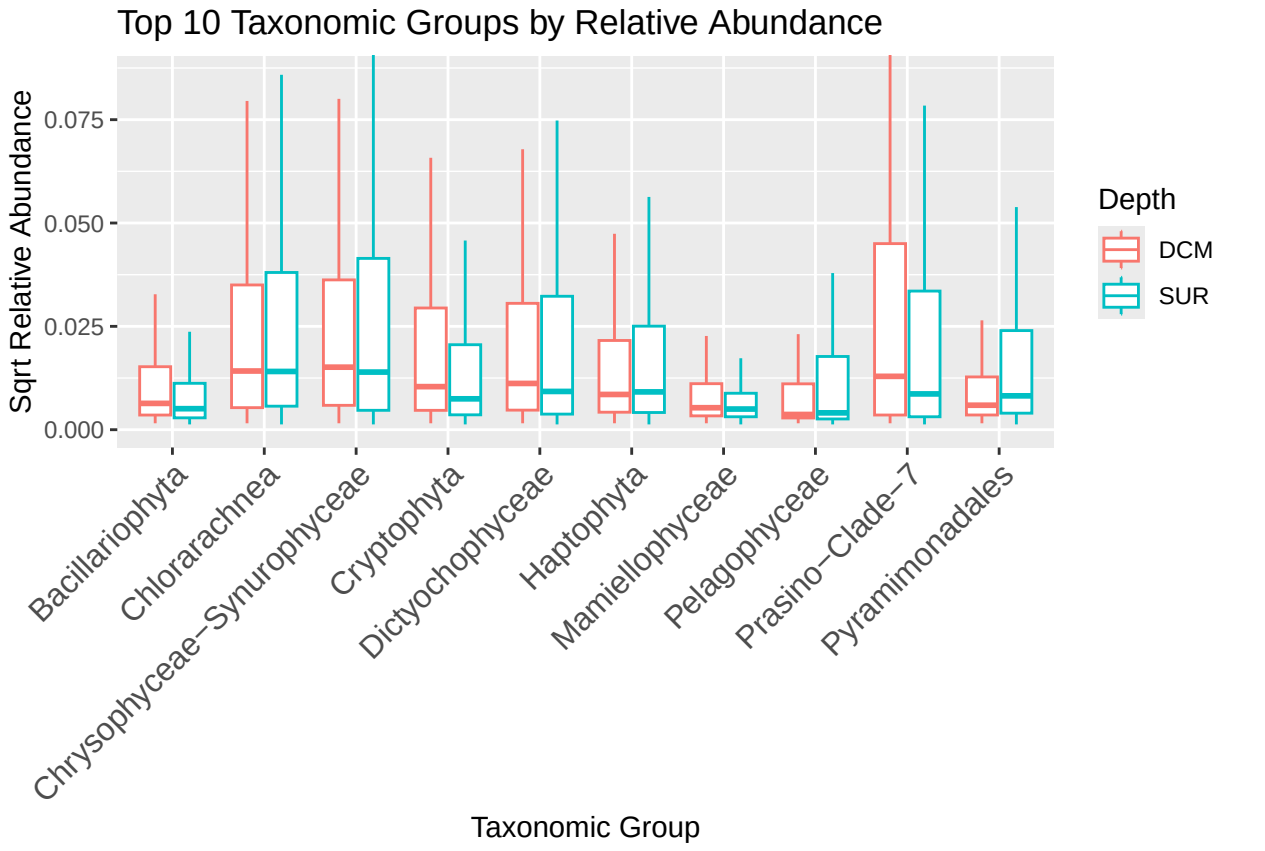
```
# Filter and create the plot
OTUabund <- abun_merge %>%
  filter(taxogroup %in% top_taxogroups) %>%
  ggplot(aes(x = taxogroup, y = sqrt(RA), color = depth)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12)) +
  labs(
    x = "Taxonomic Group",
    y = "Square Root of Relative Abundance",
    color = "Depth",
    title = "Top 10 Taxonomic Groups by Relative Abundance"
  )

# Display the plot
OTUabund
```



This boxplot is almost impossible to interpret, and I believe I am calculating correctly based off of our conversation, so I am going to try to remove some outliers and clean it up. I will also give some better names and make it a bit pretty.

```
OTUabund <- abun_merge %>%
  filter(taxogroup %in% top_taxogroups) %>%
  ggplot(aes(x = taxogroup, y = sqrt(RA), color = depth)) +
  geom_boxplot(outlier.shape = NA) +  # Remove outliers
  coord_cartesian(ylim = c(0, quantile(sqrt(abun_merge$RA), 0.95, na.rm = TRUE))) +  # Focus on 95% of
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12)) +
  labs(
```

```
    x = "Taxonomic Group",
    y = "Sqrt Relative Abundance",
    color = "Depth",
    title = "Top 10 Taxonomic Groups by Relative Abundance"
  )
#Display plot
OTUabund
```



Question 3: Some Hypotheses

About the Graph: After all of this I am still not very confident in what I have generated, mostly due to the meta reason that a homework assignment should probably result in some clear relationships and they are still not visible here.

What is the Cause of This Phenomenon?:

Availability of nutrients and light are going to be the primary driving factors of the phenomena we are observing here, some of the taxa are primary producers and others are grazers or decomposers so this will affect their abundance at certain depths of the water column based on their metabolic strategy..

Some hypothesis for how light iradiance could cause shifts in community composition:

- Primary producers vary in the provided data, with some having more abundance at the surface, and others at the DCM. Particularly the eukaryote Dictyochophyceae seems to prefer the DCM. Some plankton are more adapted to these low light conditions and others like Pyramimonadales appear to prefer the surface.
  - **Ex Situ:** Taking samples from various depths via niskin bottles one could conduct an incubation experiment (maybe first after dilution?) where the samples from each depth are exposed to light

irradiance levels that would reflect the opposite of which they were derived. Does the community start to shift in a direction that reflects it's counterpart?

  - **In Situ:** Measuring responses to light in situ I think is more challenging, as mesocosms or more representative samples are difficult in the open ocean. The work being done by the Hawaii Ocean Timeseries I think reflects one of the only real ways to collect truly "in situ" data from this environment. Perhaps one could contrive some kind of isolation of water from one depth and moving it in the water column to another, measuring community shift over time.

Hypothesis for nutrient concentrations and diversity:

- Nitrogen is frequently limiting in the oligotrophic gyres, and both producers and consumers rely on it to maintain their metabolism. From the surface ocean to the DCM nitrogen should be limiting, due to the presence of photosynthetic microorganisms. As depth increases and photosynthesis becomes less possible, nitrogen should increase as a function of depth.
  - **Ex Situ:** I would pursue a similar method as described above for light ex situ. Altering concentrations of nitrogen would likely cause limitations in primary producers at the low end of the concentration spectrum, and blooms at the higher end.
  - **In Situ:** 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?