Advances in Biology Laboratory Education
Publication of the Association for Biology Laboratory Education
Volume 43, Article 16, 2023

Snowflake CURE: Isolating microbial DNA from snow for metagenomic analysis to teach undergraduate students molecular and biostatistical tools

Jenean H. O'Brien[1] and Anne E. Kruchten[2]

[1]The College of St. Scholastica, Biology Department, 1200 Kenwood Avenue, Duluth, MN 55811, USA

[2]University of Minnesota Duluth, Swenson College of Science and Engineering, 140 Engineering, 1405 University Drive, Duluth, MN 55812, USA (**jobrien@css.edu**; **jenk0064@d.umn.edu**)

Designed to help undergraduate students gain hands-on experience with molecular biology and to become familiar with bioinformatic analysis of large-scale datasets, our Snowflake Course-based Undergraduate Research Experience (CURE) focuses on identifying how microbial populations isolated from different snow samples differ in the types and amounts of species present. In an eight-week course, upper division biology major students isolate and purify DNA from snow samples, amplify microbial 16S DNA by polymerase chain reaction (PCR), purify microbial DNA for metagenomic sequencing, use Excel and R statistical software to analyze Illumina sequencing data, and specifically explore the gene Ice Nucleation Protein (inaQ) through PCR, agarose gel electrophoresis and online database analyses. Our goal is that students gain an understanding of what bioinformatics analysis means and hopefully recognize that this skillset is within their abilities. Students are first introduced to molecular and bioinformatic skills and these are further reinforced in the duration of the course. Additionally, students gain experience in writing laboratory notebooks and communicating their science textually and visually through figure development. This course could be easily modified to become a module in a larger course, divided into smaller units to fit into pre-existing courses and/or be adapted for any water-based samples.

Keywords: metagenomics, inquiry-based learning, weather, heatmap

Introduction

This Course-based Undergraduate Research Experience (CURE) is designed to mimic working in a research laboratory environment. Upper division biology and biochemistry majors work toward four course objectives focused on developing proficiency in basic molecular biology and biostatistical skills; utilizing research protocols; discerning benefits and limitations of these approaches and communicating results in both written and oral forms (Kruchten, 2020). Our course format is eight weeks, with twice weekly meetings for two hours each. While this course utilizes snow samples, any local water sample with potential microbial populations can be used.

The eight weeks are divided into six content modules. We start with Module 1: Snowflakes and Weather focusing on setting the foundation for our research questions. This module includes watching a video defining microbiomes (Microbiology Society, n.d.), reading a review article that explores airborne bacteria (Smets et al, 2016), perusing the National Oceanic and Atmospheric Administration (NOAA, 2022) and National Weather Service (NWS, 2022) websites to learn about weather and jet streams, and watching several videos explaining how snowflakes form (BeSmart, n.d. and SciShow, n.d.) including the potential role for bacteria in this process. Briefly,

snowflakes start by forming ice crystals around a nucleating particle such as dust, pollen or a microbial cell. Some species of bacteria, such as *Pseudomonas syringae*, contain Ice Nucleation Proteins (INPs) to help with this process (Li et al, 2012). Our overarching research question is, “do snowstorms from different geographic origins carry different bacterial nuclei in their snowflakes?” During the preceding winter, three snow samples were collected from each storm that produced at least three inches of fallen snow. During this first module, students filter the thawed snow samples through an AeroPress coffee press (AeroPress, 2022) containing a 0.2-micron polyethersulfone (PES) membrane (Sterlitech, 2022) to collect the microbes contained in our samples.

In Module 2: DNA Isolation, resource videos focus on DNA purification techniques, specifically comparing precipitation (MrSimpleScience, n.d.) versus silica-gel purification (Choy, n.d.) methods. Students practice micropipetting skills, use QIAamp® Fast DNA Stool Mini kits from Qiagen to purify DNA from the microbes they collected in the previous module, and measure DNA concentration on a Nanodrop™ spectrophotometer.

Module 3 is focused on asking whether our snow samples contain bacteria. To answer this question, students use polymerase chain reaction (PCR) to amplify the 16S ribosomal RNA (rRNA) gene, which should be present in all bacterial species. After setting up and running this PCR, students utilize agarose gel electrophoresis to visualize their results.

At this point in the course, most groups have confirmed the presence of 16S rRNA in the samples, indicating that bacteria are present. Module 4 asks a more pointed question, whether any of our snow samples contain a specific INP protein called inaQ. Students again use PCR and agarose gel electrophoresis to answer this question. Further, we utilize the GenBank Nucleotide database (NCBI GenBank, 2022) to find the *P. syringae inaQ* sequence, BLAST (NCBI BLAST, 2022) to find similar sequences in other bacterial species, and Clustal Omega (EMBL, 2022) to align these multiple sequences. Then students are tasked with designing a new pair of PCR primers to answer a new question of their choice about inaQ. For example, some students want to know if inaQ specifically from *P. syringae* is present in their samples, and design primers to test this. Students then run this PCR and gel electrophoresis analysis as well.

Module 5 starts to explore metagenomic methods that could be used to analyze the diversity and abundance of bacteria in each of the snow samples. Students watch videos explaining Sanger Sequencing (Quick Biochemistry Basics, n.d.) and Illumina Sequencing (Henrik’s lab, n.d. and Illumina, n.d.) methods and complete an electronic worksheet comparing these techniques.

In Module 6, the final module, students are given a large dataset containing the 16S rRNA Illumina Sequencing results from a previous set of snow samples. They use Excel and R software programs to analyze this dataset, learning about the benefits and limitations of using each of these options. Specifically, students create heatmaps to demonstrate the diversity and abundance of each bacterial species present in each of the snow samples analyzed. It becomes very clear how much easier it is to make heatmaps in R than in Excel during these lab exercises.

Here we provide detailed instruction focused on the R analysis component of the final metagenomics module. This includes R dataset files/codes and freely available instructional materials, such as YouTube videos, used to help inform both faculty and students about metagenomics and R techniques. Please note, throughout this article, many of the cited references are links to videos and/or other useful course materials.

Student Outline – R Prep

Please complete this R pre-work before lab.

Each of you will need to prepare for performing metagenomic analyses using R programming.

- By the end of this assignment, you should:
  - Have an overview of R,
  - Have created an R Studio Cloud account on your computer,
  - Uploaded and imported data (using the friends.csv example) into R and used some basic functions to explore it,
  - Installed the tidyverse package, and
  - Used tidyverse to calculate the body mass index (BMI) of Star Wars characters.

Learning to Use R
We're going to use a series of YouTube videos (R Programming 101, n.d.) to learn about what R can do and practice using it. You will also create an R Studio Cloud account (it's free and used all over the world safely) and learn some simple techniques.

- Start and stop the videos to complete the activities as you go - don't wait until the end.
- *To be clear, all we want you to do is exactly replicate what Greg Martin, the host of the videos, does. There is no snowflake data set to analyze until you get to class.*

Here are the video and instruction links. Please watch them and do the activities in this order.

1. Watch Video #1 "R programming for beginners - Why you should use R" (3:56)
https://www.youtube.com/watch?v=9kYUGMg_14s&list=PLtL57Fdbwb_Chn-dNR0qBjH3esKS2MXY3

Activity: Just watch - no activity to do. This first video is an overview to help you see what R can do.

2. ***Get Started in R Studio Cloud***
We will be using R Studio Cloud instead of R or R Studio.

Activity: Use these instructions (Appendix A) to get started for free, sign up for a free account, start a new project, upload and import the friends.csv file and install the tidyverse package.

3. Watch Video #2 How to use R Studio (7:36 total, but START watching at 1:30)
https://www.youtube.com/watch?v=orjLGFmx6l4&list=PLtL57Fdbwb_Chn-dNR0qBjH3esKS2MXY3&index=2

Activity: Skip the first 1:30 of the video, then simply watch the rest of the video to get a further overview of R.

Note: Shortcuts using keys in R Studio Cloud to zoom into a quadrant are the same. The difference in Cloud is that you use the same shortcut to zoom out of a quadrant. For example, "Shift" + "Control" + "1" will zoom you *into* Quadrant 1 and also zoom you *out of* Quadrant 1 in R Studio Cloud.

4. **Watch Video #3 R Studio for beginners (11:53 total, but START watching at 4:32)**
https://www.youtube.com/watch?v=e8B9YU_M5FM&list=PLtL57Fdbwb_Chn-dNR0qBjH3esKS2MXY3&index=3

Activity: You have already uploaded & imported the Friends.csv data file and installed the tidyverse package in R Studio Cloud, so you can skip the first 4:32 of the video (which shows how to do this in R Studio software instead of in the Cloud version). When you start watching the video at 4:32, you should follow along, trying each step in R Studio Cloud.

5. ***Watch Video #4 Manipulating data using tidyverse* (6:55)**
https://www.youtube.com/watch?v=nRtp7wSEtJA&list=PLtL57Fdbwb_Chn-dNR0qBjH3esKS2MXY3&index=5

Activity: Follow along with the video and replicate exactly what Greg is doing. By the end, you should have used R and the tidyverse package to analyze the BMI of Star Wars characters. Basically, you will have written code!

Note: On a PC, you can use “Ctrl” + “Enter” whenever Greg says “Command” + “Enter”.

Cited References

R Programming 101. (n.d.). Playlists: R Programming for Beginners [YouTube channel]. Retrieved May 12, 2022, from https://www.youtube.com/c/RProgramming101/playlists

Materials

A computer with internet access and an R Studio Cloud account (freely available) is required for each student. Most of the cited references are also helpful course resources.

Notes for the Instructors

The pre-work for the R metagenomics module should give students experience with creating an R Studio Cloud account, opening a new project, importing data and installing packages that contain functions needed to perform their practice analysis. Further, students, should have attempted to follow along with the Star Wars BMI analysis performed in the YouTube channel videos. Overall, students should be familiar with what R is and how the interface is organized before they come to lab.

We start lab by having the students open their R Studio Cloud account and then go to the URL for our Snowflake Data project. The project simply contains a folder with all .csv data files and .r code files needed to run the analyses, just make sure to change the access settings to "Everyone (all Cloud users)" under "Who can view this project." The URL format should be as follows: https://rstudio.cloud/content/XXXXXXX where the XXXXXXX is a numeric code specific to your project.

Once students have our project open, we walk them through how to import the .csv data files into the project, as outlined in their lab pre-work assignment (Appendix A, Step 2). Then, students click on the .r code files (Appendices B and C), which will open directly in the upper left quadrant of their screen. For the rest of the lab, we run the R code together, explaining what each step along the way is doing. These explanations are also provided at the beginning of each step of the code, following the hashtag. By the end of lab, each student has created two heatmaps demonstrating which bacterial species are present in each of our snow samples, based on the 16S sequencing data.

We designed this class to be an immersive experience that represents working in a molecular biology laboratory. To that end, all assessments for the course aligned with activities that molecular biologists perform as part of their career. For each module, students designed a data figure with text that explained the procedures they performed and/or the results they collected. For example, at the end of the metagenomics analysis module, students complete an assignment (Appendix D) that requires them to create five data figures from Excel and R, including figure captions. Appendix E contains a student-generated example. At the end of the semester, the student compiles these figures into a research poster that they present at our Scholarly and Creative Activities Symposium.

Additional assessments include keeping a laboratory notebook and recording pipetting technique videos. These two activities are performed twice throughout the course in a pre-feedback-post format. The final assessment students complete is a writing a case statement, similar to the self-evaluative letters that molecular biologists would include in a portfolio for promotion/tenure. The case statement is described as an opportunity for students to state their case, or demonstrate, how they have made progress during this semester on the course objectives. This should include evidence in the form of specific examples of progress. Students often include references to their lab notebooks, pipetting videos and figures/poster to as pieces of evidence. A case statement is a combination of an argument demonstrating progress and a reflection on that progress. Therefore, students are asked to reflect on what grade they think they should earn for the course and to explain why. These assessments have generally been well received, as demonstrated in our student evaluations of the course (Appendix F).

Cited References

AeroPress. Original Coffee Maker. [Internet]. c2022. Palo Alto (CA). [accessed 2022 July 12]. Available from: https://aeropress.com/products/aeropress-coffee-maker

Be Smart. (n.d.). The Science of Snowflakes. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/fUot7XSX8uA

Choy M. (n.d.). Silica-based chromatography spin column. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/vYOKjcmjVrI

European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EBI). Clustal Omega – Multiple sequence alignment. c2022. Cambridgeshire (UK). [accessed 2022 July 12]. Available from: https://www.ebi.ac.uk/Tools/msa/clustalo/

Henrik's Lab. (n.d.). Next generation sequencing (Illumina): An introduction. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/CZeN-IgjYCo
Illumina. (n.d.). Illumina sequencing by synthesis. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/fCd6B5HRaZ8
Kruchten A. 2020. A curricular bioinformatics approach to teaching undergraduates to analyze metagenomic datasets using R. Front Microbiol. 11:578-600.
Li Q, Yan Q, Chen J, He Y, Wang J, Zhang H, Yu Z, Li L. 2012. Molecular characterization of an ice nucleation protein variant (inaQ) from Pseudomonas syringae and the analysis of its transmembrane transport activity in Escherichia coli. Int J Biol Sci. 8(8):1097-1108.
Microbiology Society. (n.d.). What is a microbiome? [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/YqFCXpA7O3k
MrSimpleScience. (n.d.). DNA Isolation: Simple animated tutorial. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/8cYvyYOjzOc
National Center for Biotechnology Information (NCBI). Basic Local Alignment Search Tool (BLAST). [Internet]. c2022. Bethesda (MD). [accessed 2022 July 12]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi
National Center for Biotechnology Information (NCBI). Nucleotide (GenBank). [Internet]. c2022. Bethesda (MD). [accessed 2022 July 12]. Available from: https://www.ncbi.nlm.nih.gov/nucleotide/
National Oceanic and Atmospheric Administration [Internet]. c2022. Silver Spring (MD). [accessed 2022 July 12]. Available from: https://www.nws.noaa.gov/outlook_tab.php
National Weather Service Jetstream. [Internet]. c2022. Silver Spring (MD). [accessed 2022 July 12]. Available from: https://www.weather.gov/education/jetstream
Quick Biochemistry Basics. (n.d.). Sanger sequencing. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/KTstRrDTmWI
SciShow. (n.d.). Bioprecipitation: How bacteria makes snow. [Internet]. YouTube.com; [accessed 2022 July 12]. Available from: https://youtu.be/mFJLFXhycSQ
Smets W, Moretti S, Denys S, Lebeer S. 2016. Airborne bacteria in the atmosphere: Presence, purpose, and potential. Atmos Environ. 139:214-221.
SterliTech. Polyethersulfone (PES) Membrane Filters. [Internet]. c2022. Auburn (WA). [accessed 2022 July 12]. Available from: https://www.sterlitech.com/polyethersulfone-membrane-filter-pes023001.html

Acknowledgments

Thank you very much to all of the BIO 4160 - Molecular Biology students who have participated in this laboratory course and provided their feedback over the last three years.

About the Authors

Jenean O'Brien has been teaching at The College of St. Scholastica since 2019, where she teaches a range of first year and upper level classes to biology and biochemistry majors and non-major students.

Anne Kruchten has taught undergraduate courses in cell biology, biochemistry, microbiology, and climate change as a faculty member. She now works as an Educational Specialist developing innovative cross-college curriculum solutions in the Swenson College of Science and Engineering at the University of Minnesota Duluth.

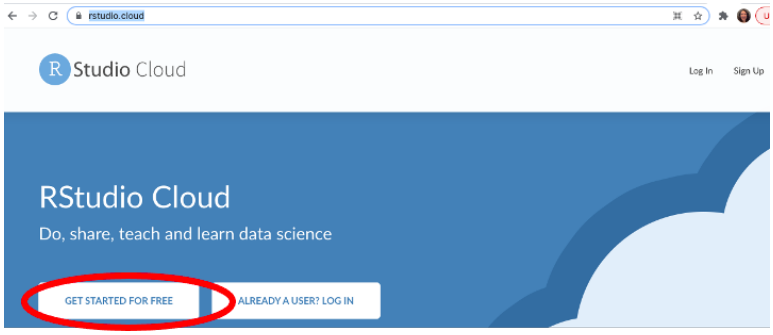
Advances in Biology Laboratory Education
Publication of the Association for Biology Laboratory Education
Volume 43, Article 16, 2023

Appendix A – Student Instructions for How to Create an R Studio Cloud Account - Plus More!

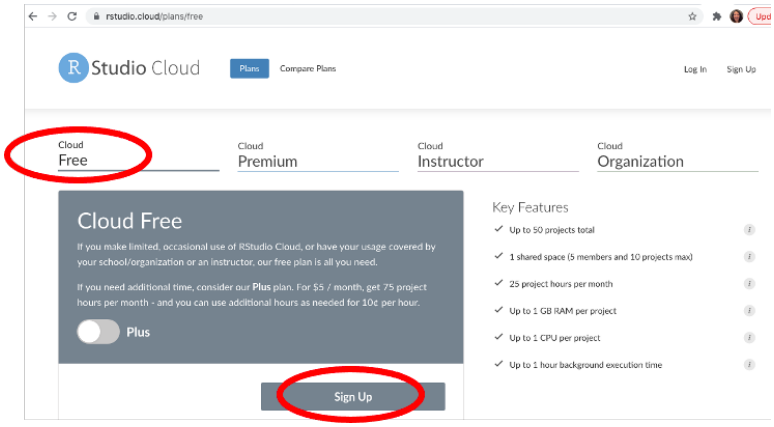
Please follow these step by step instructions to ensure that your account and laptop are ready for class.

1. **Create a free RStudio account that you will use in class.**

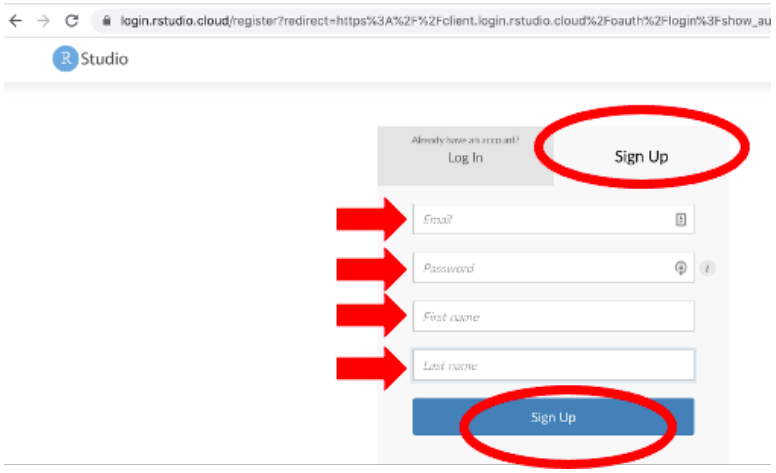
a. Go to the website https://rstudio.cloud/ and you should see the website below and click on "Get Started for Free".



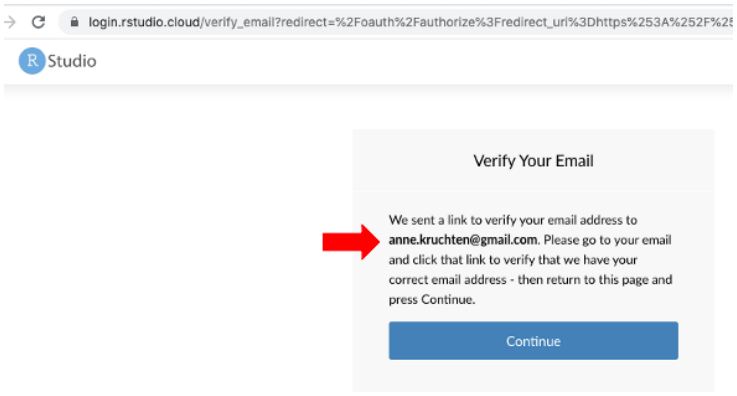
b. On the next screen, click on Cloud Free and then click on Sign Up at the bottom of the screen.



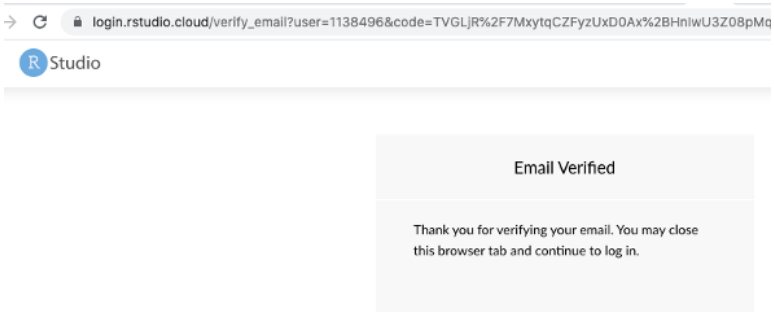
c. Create an account by making sure you're on the Sign Up tab and then filling in the fields with your information. **BE SURE TO NOTE YOUR EMAIL AND PASSWORD SO YOU CAN LOG IN DURING CLASS!** Click "Sign Up" at the bottom when you're finished.



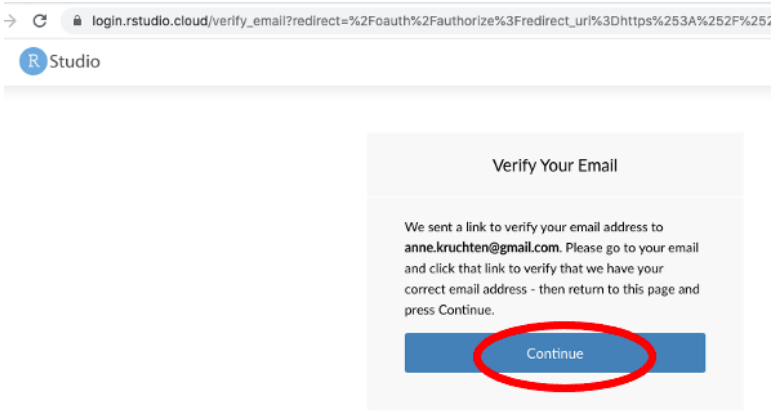
d. Before you can continue, you must verify the email address you used for sign up. Go to that email address and follow the instructions you received. Leave this window open - you'll return to it in a minute.



e. Once you've followed the emailed instructions, you will see this screen. From here return to the window or tab you were on before.

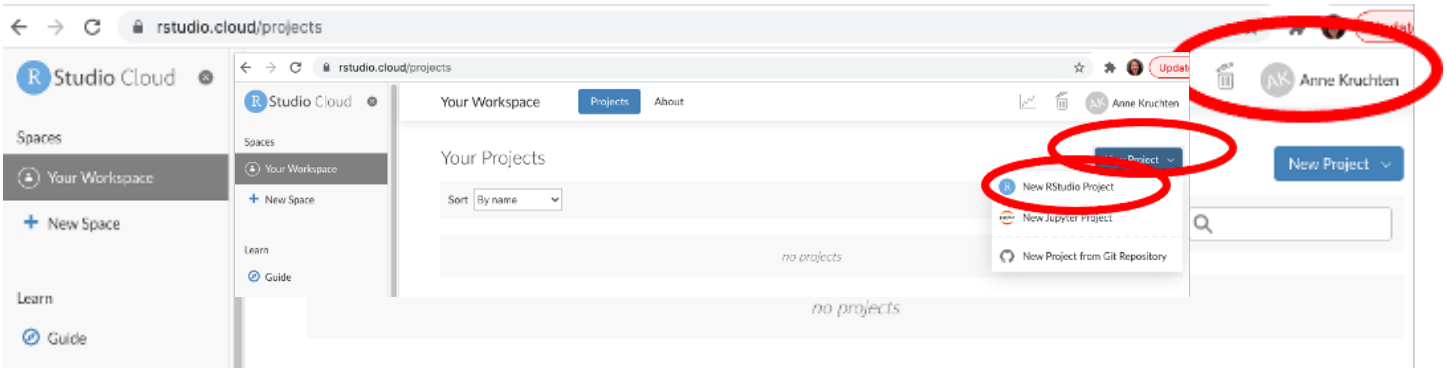


f. When you're back on the previous tab or window, you'll see this screen again. Click Continue.



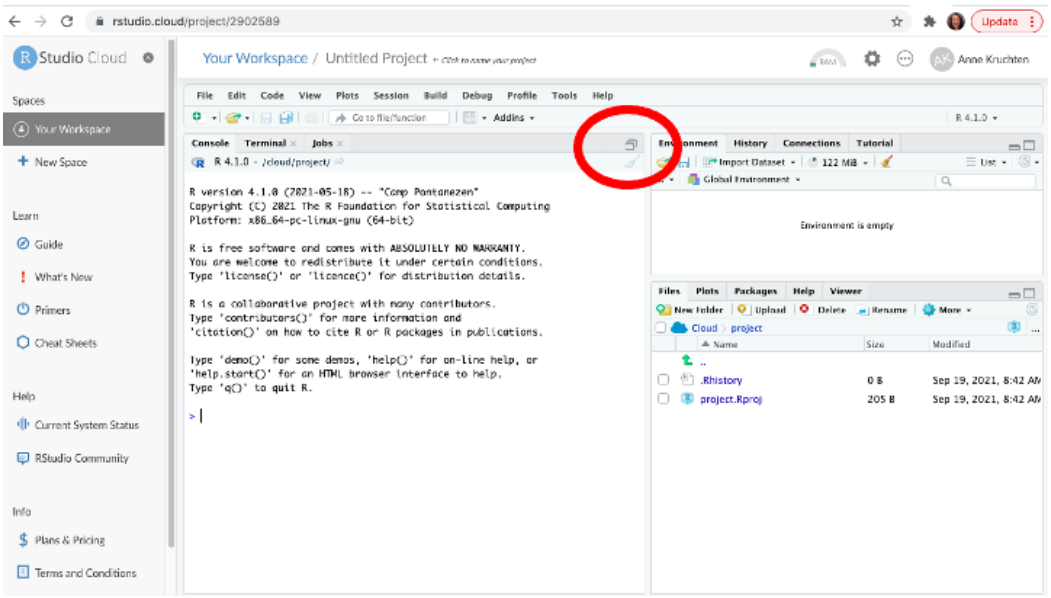
g. You should now be on a screen that looks like the picture below, with your name in the top righthand corner in the red circle. You now officially have an RStudio account.

2. Create a new project in RStudio.

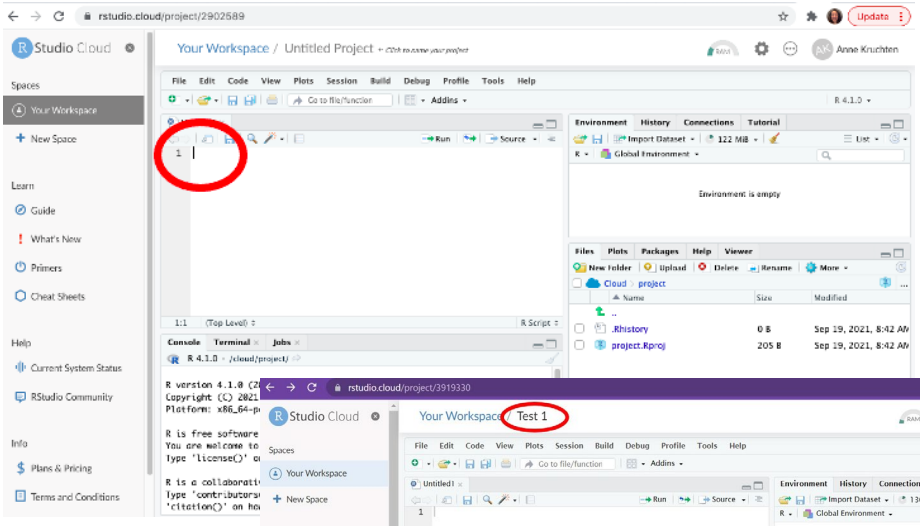


a. Towards the top right of the screen, click on the blue box that says "New Project" and then in the drop down click on "New RStudio Project".

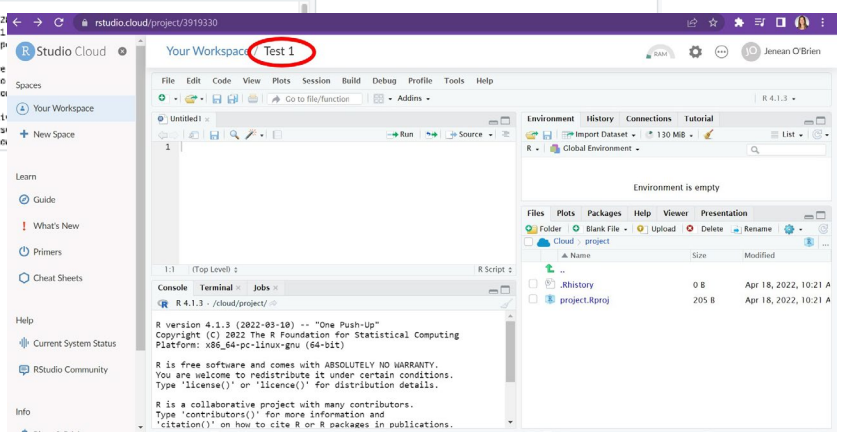
b. A new project screen will pop up that looks like the screen below. On your project screen, click on the double screen icon inside the red circle below.



c. Your screen will now have four quadrants in the working area: upper left, lower left, upper right, and lower right. Place your cursor in the upper left quadrant by clicking in the white box.



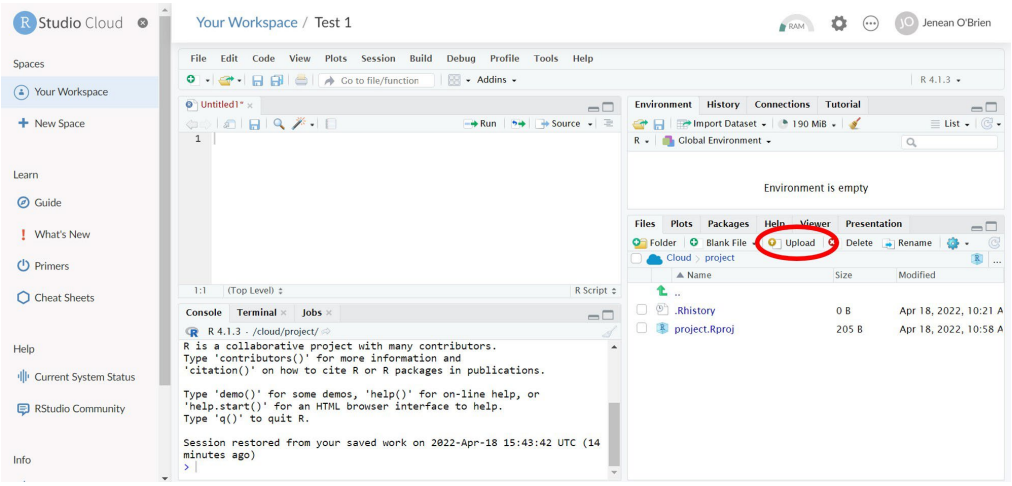
d. Rename your project by clicking on "Untitled Project" at the top and typing in the new name. I've renamed this example "Test 1." You're now ready to continue on and upload data files into your project.



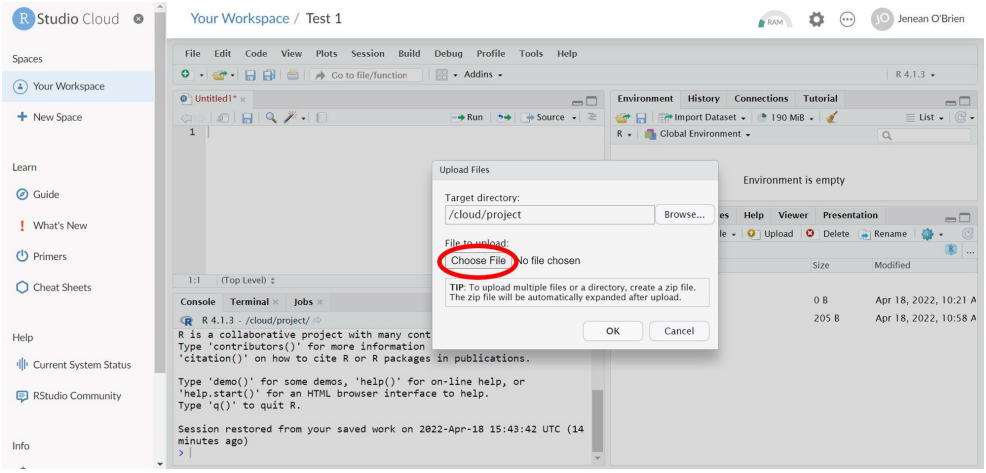
3. Upload data files into your project.

For your first project (Test 1), download the "friends.csv" file. Then go back to R Studio Cloud and continue with the next step below.

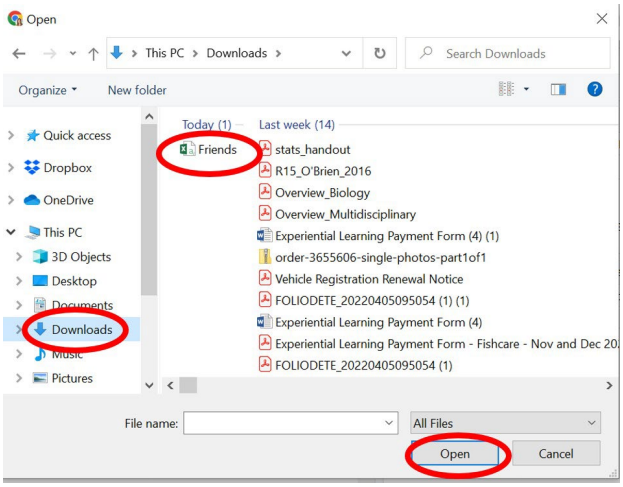
a. In the bottom right quadrant, click on "Upload".



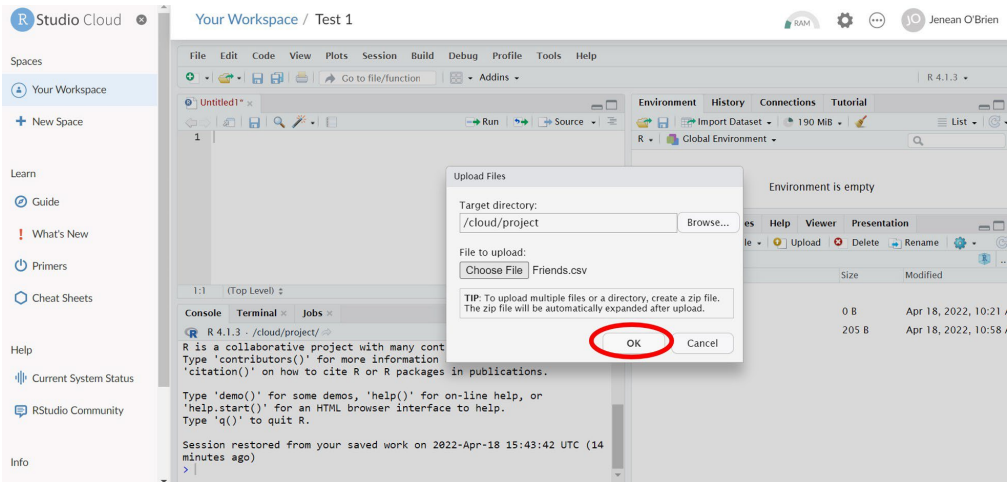
b. In the pop-up window, click on "Choose File".



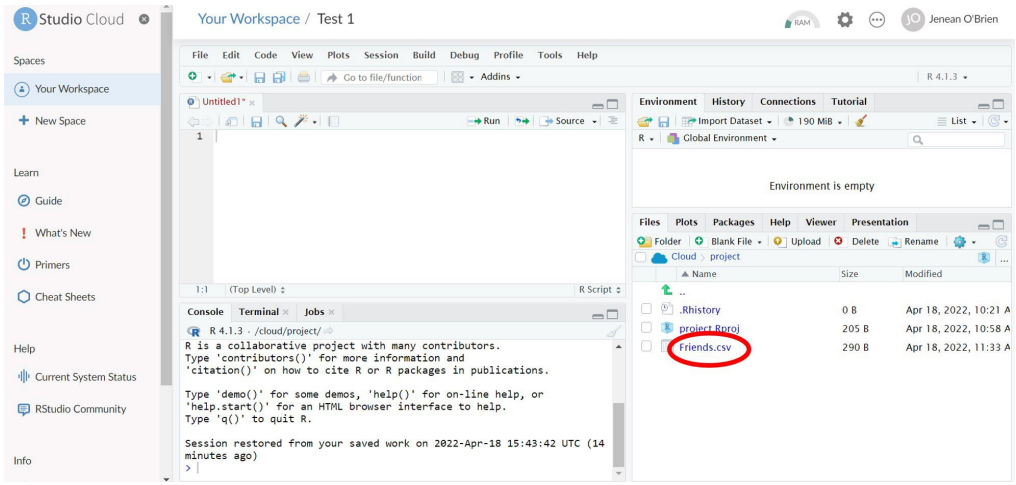
c. In the next pop-up window, select your Downloads, then select the friends.csv file. Click "open" in the bottom right corner.



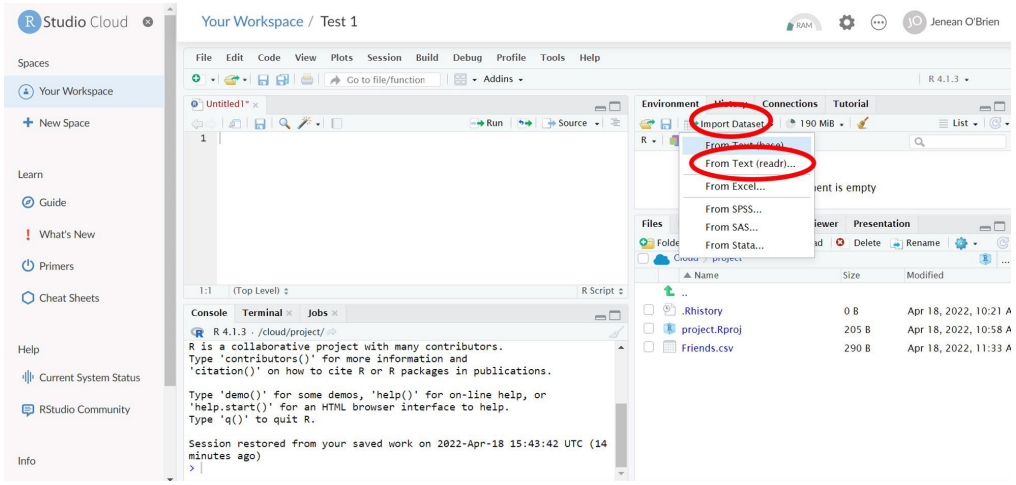
d. The friends.csv file now appears next to Choose File. Click OK at the bottom of the pop up window.



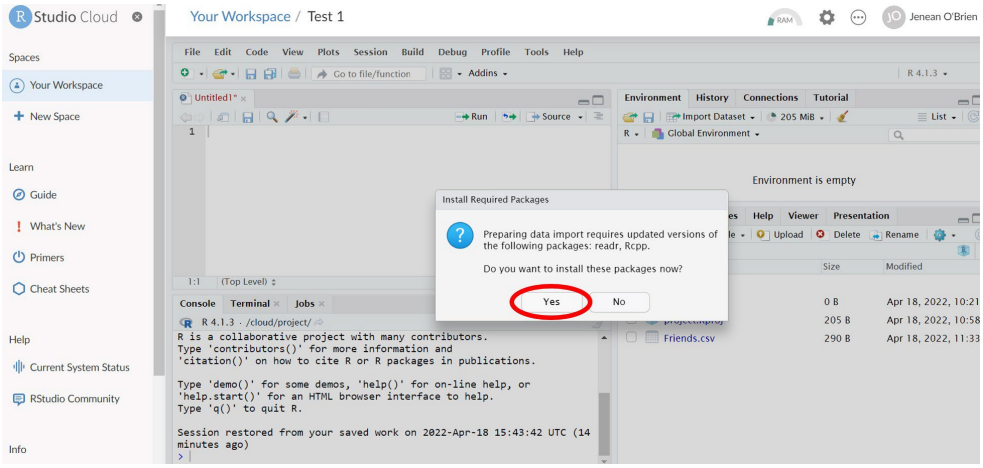
e. The friends.csv file now appears in the file list in the bottom right quadrant.



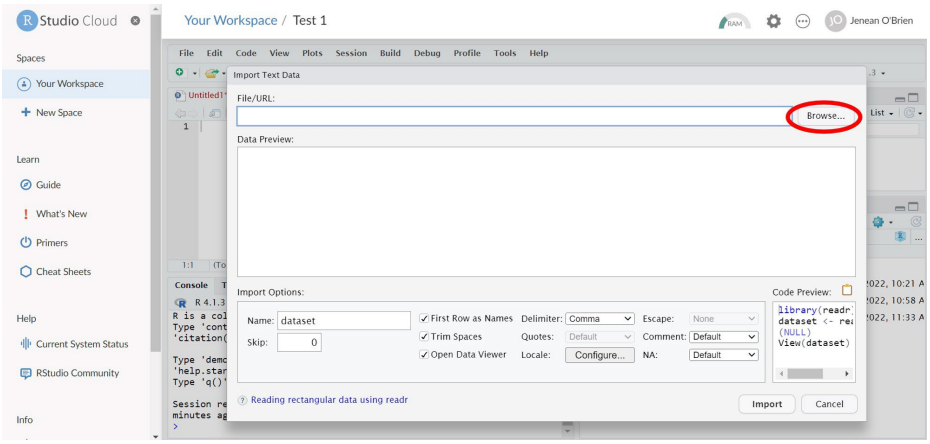
f. The datafile is now uploaded into RStudio, and now we want to import it into our Test 1 project. Start this process by clicking on "Import Dataset" in the top right quadrant. In the dropdown menu, click on "From Text (readr)..."



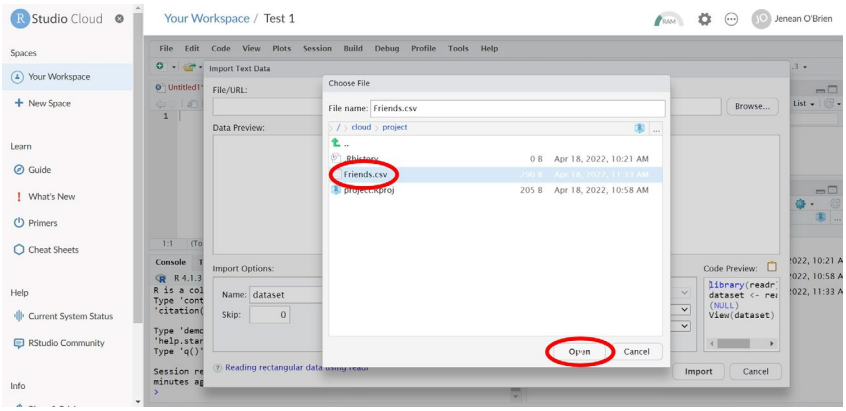
g. A pop-up window will ask if you want to Install Required Packages. Click yes, and then ***wait patiently***. It may take a few minutes before it looks like anything is happening.



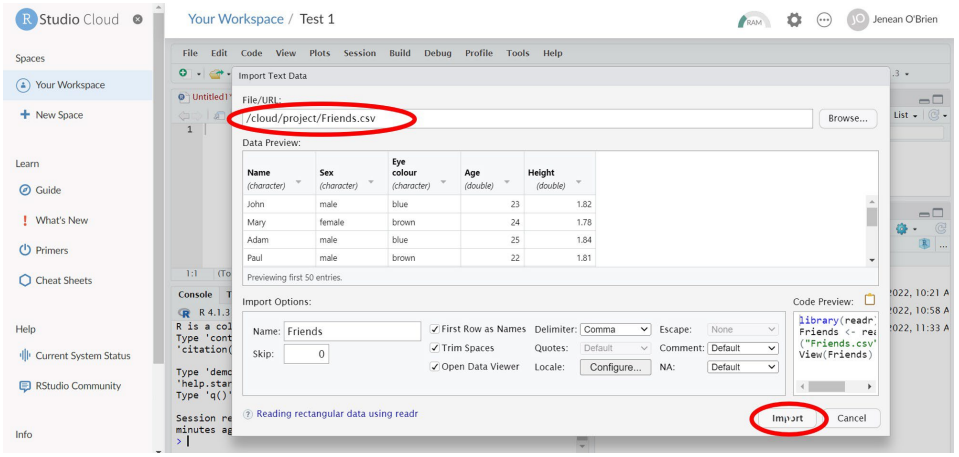
h. Wait. It will take a few minutes and then suddenly you'll start seeing a lot of red text scrolling in the bottom left quadrant followed by this screen popping up. Click on "Browse".



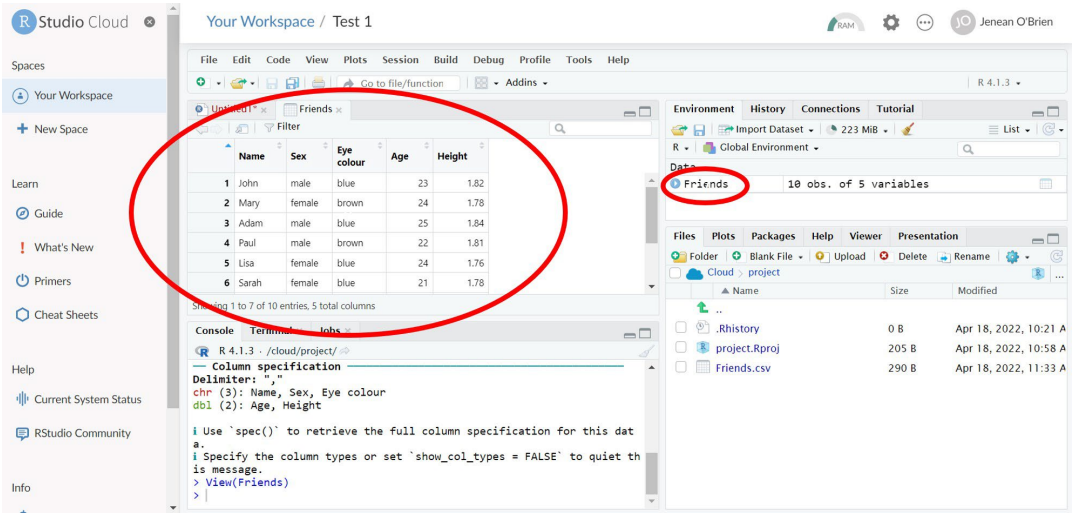
i. The Browse function will bring up a pop-up window that allows you to choose from the files you uploaded into the bottom right quadrant. Choose the friends.csv file and then click Open at the bottom of the window.



j. Now you'll see the friends.csv file name under File/URL. In the center you can see a preview of the data file. This file has five columns of data: Name, Sex, Eye colour, Age and Height. You're just seeing the very tops of the columns in an Excel file. Finally, click "Import" at the bottom right.

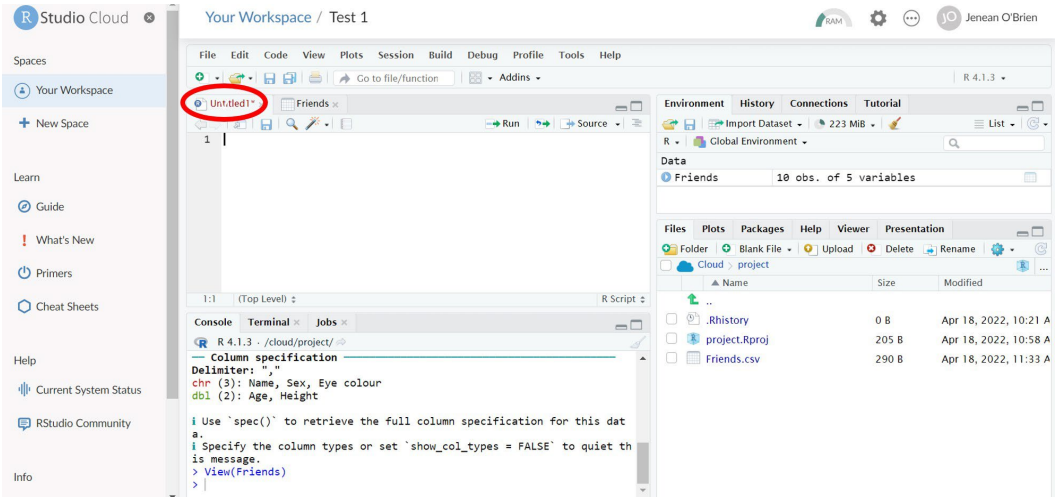


k. The friends.csv file has now been imported into your Test 1 project. You can see it listed in the upper right quadrant, and you can see a preview of it in the upper left quadrant. Continue to the last step.



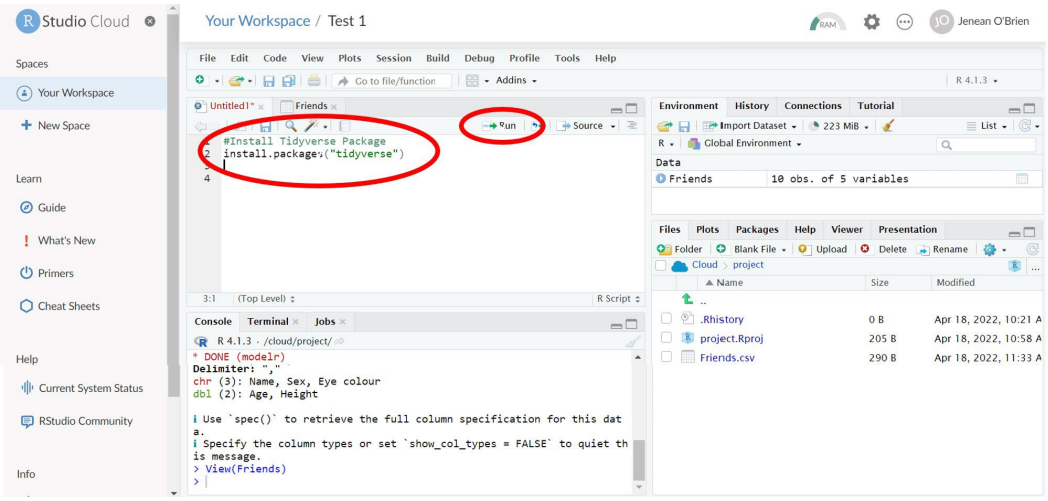
4. Install Packages into RStudio that contain the tools we'll need for data analysis.

a. Click on the tab "Untitled" in the upper left quadrant. This will return you to Quadrant 1. Place your cursor at the beginning of line #1.



b. Type the following into the quadrant.

```
#Install Tidyverse Package
install.packages("tidyverse")
```



c. In the upper left quadrant click on the "Run" button. By doing this, you are actually running "code", just like a computer programmer would. Any line that begins with # is skipped by the computer. The # lines include notes and instructions for you. When you click "Run", the computer will skip all the # lines and run the first command it sees on line 2 "Install packages ("tidyverse")...

The bottom left quadrant is the output window. It shows what the computer is doing. You will briefly see the blue text below that says > install.packages("tidyverse)... which means that the computer is installing a package named "tidyverse" which has a bunch of tools for data presentation. While the computer is working, there is a small red stop sign in the upper right of the lower left quadrant box. You need to wait for this to go away before you can do the next step.

d. As the computer works, a bunch of red text will start to scroll through the lower left quadrant. This is good. When it's finished, the red stop sign will disappear, red text will tell you where the downloaded packages are, and your cursor will appear on the next line (number 3) of text in the upper left quadrant, ready to run the next command.

e. Now that you have installed the tidyverse package, you need to have R call it up. To do this, type the following into Quadrant 1:

   library("tidyverse")

   and click run.

f. You're ready to run R!! Refer back to the "R Prep" assignment and finish watching the videos, trying each step along the way in your R Studio Cloud account.

****Remember to note your email and password that you used to create your RStudio account. For easy access, you can bookmark RStudio in your browser. Bring this laptop with you to class and any charging cords needed.****

Appendix B – R code to make a heatmap with small dataset

```
# To rename the R script tab in the upper left quadrant, click on the floppy disc and rename as "Phyla Heatmap Code."

# Install and load the gplots package which contains various tools for plotting data.
install.packages("gplots")
library(gplots)

# Install and load the BiocManager package for some analysis tools.
install.packages("BiocManager")
BiocManager::install("Heatplus")
library(Heatplus)

# Install and load the vegan package for hierarchical clustering if you want to use distance functions (how closely related objects are in a dendrogram) not specified in the basic command called dist.
install.packages("vegan")
library(vegan)

# Install and load the RColorBrewer package for better color options in heatmaps.
install.packages("RColorBrewer")
library(RColorBrewer)

# Load our data 2020 Winter Storm Phylum Data Transposed and call it Phyla_data -- remember that the .csv file must be imported into your working folder first!
Phyla_data <- (X2020_Winter_Storm_Phylum_Data_Transposed)
head(Phyla_data)

# look at the dimensions of the data (number of rows vs columns) using the dim command
dim(Phyla_data)
# ours should read out [1] 12 23, meaning twelve rows and twenty-three columns.

# Check out the first three rows and first four columns to make sure it looks right
Phyla_data[1:3, 1:4]

# We'll have to strip off the sample IDs and convert them to row names so that the data matrix contains only sequence count data.
row.names(Phyla_data) <- Phyla_data$sample
Phyla_data <- Phyla_data[, -1]

# Look at the dimensions of the data (number of rows vs columns) using the dim command to confirm that the removal of the sample IDs worked.
dim(Phyla_data)
# ours should read out [1] 12 22, meaning twelve rows and twenty-two columns (one less column than before since we removed the sample IDs).

# Check out the first three rows and first four columns to make sure it looks right
Phyla_data[1:3, 1:4]

# colorRampPalette is in the RColorBrewer package.  This creates a color palette that shades from light yellow to red in RGB space with 100 unique colors.
scaleyellowred <- colorRampPalette(c("lightyellow", "red"), space = "rgb")(100)

# Here's a very basic heatmap using our data with the color palette we created above.
heatmap(as.matrix(Phyla_data), Rowv = NA, Colv = NA, col = scaleyellowred)
```

```
# If you look at the heatmap, we only see 10 phyla, but we know there should be 22, so this command will change to margins, effectively changing the ‘width’ of the heatmap we see so we can read everything on that x axis
heatmap(as.matrix(Phyla_data), Rowv = NA, Colv = NA, col = scaleyellowred, margins = c(10, 2))

# Now let’s add a dendrogram for the samples. The heatmap function will do this for you, but we prefer to make our own using the vegan package as it has more options for distance metrics. Also, this means that you can do hierarchical clustering using the full dataset, but only display the more abundant taxa in the heatmap.
library(vegan)

# Perform the statistical analyses needed to organize (or clump) the data by patterns. This statistical analysis is called the Bray-Curtis dissimilarity matrix, which will be calculated on the full dataset.
data.dist <- vegdist(Phyla_data, method = "bray")

# Start to organize (or clump) the data by patterns. We choose to do average linkage hierarchical clustering. Other options are 'complete' or 'single'. You'll need to choose the one that best fits the needs of your situation and your data.
row.clus <- hclust(data.dist, "aver")

# Add the dendrogram to one axis, the rows. This dendrogram clusters the snowstorm samples that have similar phyla patterns.
heatmap(as.matrix(Phyla_data), Rowv = as.dendrogram(row.clus), Colv = NA, col = scaleyellowred, margins = c(10, 3))

# You can also add the dendrogram to the other axis, the columns. This dendrogram clusters the phyla that occur more often together. This demonstrates the phyla that have a similar distribution pattern across snowstorms.

# First you have to transpose the dataset to get the phyla as rows.
data.dist.g <- vegdist(t(Phyla_data), method = "bray")
col.clus <- hclust(data.dist.g, "aver")

# Then you can make the column dendrogram with Rowv = as.dendrogram(row.clus).
heatmap(as.matrix(Phyla_data), Rowv = as.dendrogram(row.clus), Colv = as.dendrogram(col.clus), col = scaleyellowred, margins = c(10, 3))

#To save the plot image, in the lower right quadrant, click “Zoom.” Right click on the image, select “save image as.”

#As a reminder: Samples 1, 2 and 3 are three replicates from Storm #1. Samples 4, 5 and 6 are three replicates from Storm #2. Samples 7, 8 and 9 are three replicates from Storm #3. Samples 10, 11 and 12 are three replicates from Storm #4.
```

Appendix C – R code to make a heatmap with large dataset

```
# To start a new tab for this code in the upper left quadrant, click on the green + icon and select "R script."

# To rename the R script tab in the upper left quadrant, click on the floppy disc and rename as "OTU Heatmap Code."

# Install and load the gplots package which contains various tools for plotting data.
install.packages("gplots")
library(gplots)

# Install and load the BiocManager package for some analysis tools.
install.packages("BiocManager")
BiocManager::install("Heatplus")
library(Heatplus)

# Install and load the vegan package for hierarchical clustering if you want to use distance functions (how closely related objects are in a dendrogram) not specified in the  basic command called dist.
install.packages("vegan")
library(vegan)

# Install and load the RColorBrewer package for better color options in heatmaps.
install.packages("RColorBrewer")
library(RColorBrewer)

# Load our data 2020 Winter Storm OTU Data Transformed and call it OTU_DATA -- remember that the .csv file must be imported into your working folder first!
OTU_DATA <- read_csv("ABLE files/2020 Winter Storm OTU Data Transformed.csv")

# Look at the dimensions of the data (number of rows vs columns) using the dim command.
dim(OTU_DATA)
# ours should read out [1] 12 496, meaning twelve rows and 496 columns.

# Check out the first four rows and first 15 columns to make sure it looks right.
OTU_DATA[1:4, 1:15]

# We'll have to strip off the sample IDs and convert them to row names so that the data matrix contains only sequence count data.
row.names(OTU_DATA) <- OTU_DATA$sample
OTU_DATA <- OTU_DATA[, -1]

# Look at the dimensions of the data (number of rows vs columns) using the dim command.
dim(OTU_DATA)
# ours should read out [1] 12 495, meaning twelve rows and 495 columns, (one less column than before since we removed the sample IDs).

# Check out the first four rows and first 15 columns to make sure it looks right.
OTU_DATA[1:4, 1:15]

# colorRampPalette is in the RColorBrewer package.  This creates a color palette that shades from light yellow to red in RGB space with 100 unique colors.
scaleyellowred <- colorRampPalette(c("lightyellow", "red"), space = "rgb")(100)

# Here's a very basic heatmap using our data with the color palette we created above.
heatmap(as.matrix(OTU_DATA), Rowv = NA, Colv = NA, col = scaleyellowred)
```

```
# It’s pretty clear that this plot is inadequate in many ways. For one, the OTU labels are all squished along the bottom and impossible to read. One solution to this problem is to remove OTUs that are exceedingly rare from this figure. Let’s try removing OTUs whose relative read abundance is less than 1% of at least 1 sample.

# Determine the maximum relative abundance for each column.
maxab <- apply(OTU_DATA, 2, max)
head(maxab)

# Remove the OTUs with less than 1% as their maximum relative abundance.
n1 <- names(which(maxab < 0.01))
OTU_DATA.1 <- OTU_DATA[, -which(names(OTU_DATA) %in% n1)]

# The margins command sets the width of the white space around the plot. The first element is the bottom margin and the second is the right margin.
heatmap(as.matrix(OTU_DATA.1), Rowv = NA, Colv = NA, col = scaleyellowred, margins = c(10, 2))

# This is better, but there are still many different OTU names.  You can check how many with the dim command as we did in the beginning of the code.
dim(OTU_DATA.1)
# Our result is 12 rows, 470 columns.  Let's see what happens if we make the cutoff 2%.

# Determine the maximum relative abundance for each column.
maxab <- apply(OTU_DATA, 2, max)
head(maxab)

# Remove the OTUs with less than 2% as their maximum relative abundance.
n1 <- names(which(maxab < 0.02))
OTU_DATA.2 <- OTU_DATA[, -which(names(OTU_DATA) %in% n1)]

# Clean up the margins again.
heatmap(as.matrix(OTU_DATA.2), Rowv = NA, Colv = NA, col = scaleyellowred, margins = c(10, 2))

# This is better, but there are still many different OTU names.  You can check how many with the dim command as we did in the beginning of the code.
dim(OTU_DATA.2)
# Down to 447 columns. Let's try a 4% cutoff.

# Determine the maximum relative abundance for each column.
maxab <- apply(OTU_DATA, 2, max)
head(maxab)

# Remove the OTUs with less than 4% as their maximum relative abundance
n1 <- names(which(maxab < 0.04))
OTU_DATA.4 <- OTU_DATA[, -which(names(OTU_DATA) %in% n1)]

# Clean up the margins again.
heatmap(as.matrix(OTU_DATA.4), Rowv = NA, Colv = NA, col = scaleyellowred, margins = c(10, 2))

# This is better, but there are still many different OTU names.  You can check how many with the dim command as we did in the beginning of the code.
dim(OTU_DATA.4)
# 418 columns!  Let's go big and try a 8% cutoff!

# Determine the maximum relative abundance for each column.
maxab <- apply(OTU_DATA, 2, max)
head(maxab)
```

```
# Remove the OTUs with less than 8% as their maximum relative abundance.
n1 <- names(which(maxab < 0.08))
OTU_DATA.8 <- OTU_DATA[, -which(names(OTU_DATA) %in% n1)]

# Clean up the margins again.
heatmap(as.matrix(OTU_DATA.8), Rowv = NA, Colv = NA, col = scaleyellowred, margins = c(10, 2))

# Check how many OTU names with the dim command.
dim(OTU_DATA.8)
# 367 columns.  Seems ok for now.

# Now let's add a dendrogram for the samples. The heatmap function will do this for you, but we prefer to make our own using the vegan package as it has more options for distance metrics. Also, this means that you can do hierarchical clustering using the full dataset, but only display the more abundant taxa in the heatmap.
library(vegan)

# Perform the statistical analyses needed to organize (or clump) the data by patterns. This statistical analysis is called the Bray-Curtis dissimilarity matrix, which will be calculated on the full dataset.
OTU_DATA.dist <- vegdist(OTU_DATA, method = "bray")

# Start to organize (or clump) the data by patterns. We choose to do average linkage hierarchical clustering. Other options are 'complete' or 'single'. You'll need to choose the one that best fits the needs of your situation and your data.
row.clus <- hclust(OTU_DATA.dist, "aver")

# Add the dendrogram to one axis, the rows.  This dendrogram clusters the snowstorm samples that have similar OTU patterns.
heatmap(as.matrix(OTU_DATA.8), Rowv = as.dendrogram(row.clus), Colv = NA, col = scaleyellowred, margins = c(10, 3))

# You can also add a column dendrogram to cluster the OTUs that occur more often together. This dendrogram clusters the OTUs that occur more often together. This demonstrates the OTUs that have a similar distribution pattern across snowstorms.
Note that this one must be done on the same dataset that is used in the Heatmap (i.e. reduced number of OTUs due to maximum relative abundance percentage cutoffs).

# First you have to transpose the dataset to get the OTUs as rows.
OTU_DATA.dist.g <- vegdist(t(OTU_DATA.8), method = "bray")
col.clus <- hclust(OTU_DATA.dist.g, "aver")

# Then you can make the column dendrogram with Rowv = as.dendrogram(row.clus).
heatmap(as.matrix(OTU_DATA.8), Rowv = as.dendrogram(row.clus), Colv = as.dendrogram(col.clus), col = scaleyellowred, margins = c(10, 3))

#To save the plot image, in the lower right quadrant, click "Zoom." Right click on the image, select "save image as."

#As a reminder: Samples 1, 2 and 3 are three replicates from Storm #1. Samples 4, 5 and 6 are three replicates from Storm #2. Samples 7, 8 and 9 are three replicates from Storm #3. Samples 10, 11 and 12 are three replicates from Storm #4.
```

Appendix D - Student Assignment Instructions

Metagenomics and Sequencing Module
Five Data Figures from Excel and R

For this module assignment:

1) Think about our work so far and decide on an **overarching question(s)** that you want to ask about the snowflake data that we have. This will help you narrow down which figures from below you want to include and/or create.

2) Submit **five (minimum) Excel/R figures with figure captions**:

    a) Choose at least two figures from Excel that we created. Your options are:

    Averages + standard deviations of Phylum in graph format

    Averages + standard deviations of OTUs in graph format

    Conditionally Formatted (Red/White color scale) of Phylum data

    Conditionally Formatted (Red/White color scale) of OTU data

    b) Include two figures from R that we created:

    Heat map of Phylum data

    Heat map of OTU data

    c) Create one new figure of your choosing, using either R or Excel

Figure captions will include the following information:

- Informative title that summarizes the result demonstrated by the figure
- Brief description of methods used (e.g. “DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit”, “DNA was sequenced using Illumina sequencing”, “R was used to analyze Illumina sequencing data)
- Define the axes, color scale, symbols, data that was deleted or is represented, etc.
- Summarize the results demonstrated in the figure (e.g. “Bar A is larger than Bar B, and therefore A is...”)

No citations are needed.

Appendix E - Student Assignment Example

Questions:
1) What is the difference in count of phylum between snow samples?
2) How does the content of bacteria from certain phyla differ between snow samples?
3) Is the count of phylum between snow samples statistically significant different?

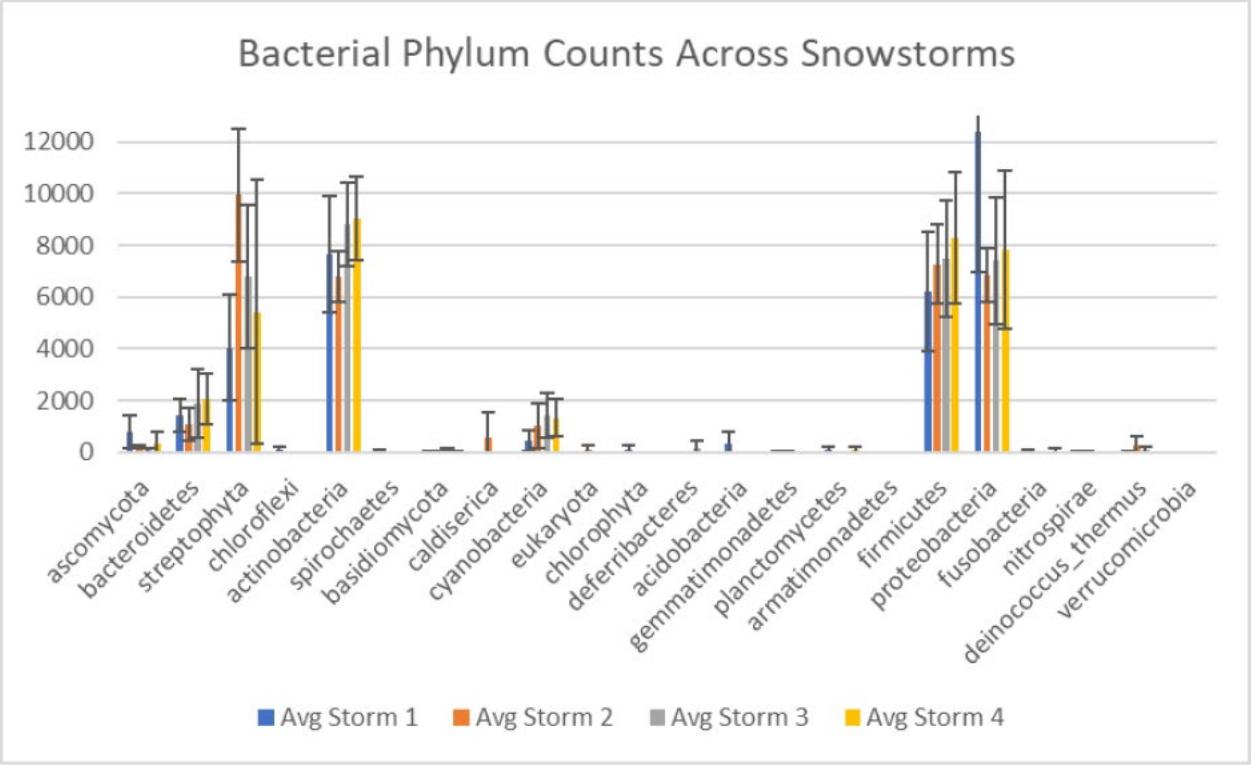


Figure 1: Bacterial phylum count across snowstorms. DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit, DNA was sequenced using Illumina sequencing, Excel was used to analyze Illumina sequencing data. X axis showing all phylum analyzed and bars from all four storms, Y axis showing count in number. The bars are particularly high for streptophyta, actinobacteria, firmicutes, and proteobacteria, which tell us that these phylum are the most abundant in our samples.

phylum	Avg Storm 1	Avg Storm 2	Avg Storm 3	Avg Storm 4
ascomycota	811.33	227.33	91.00	339.33
bacteroidetes	1443.67	1096.00	1891.33	2066.67
streptophyta	4053.33	9939.67	6800.33	5429.00
chloroflexi	104.33	0.00	0.00	0.67
actinobacteria	7667.67	6801.67	8828.33	9048.00
spirochaetes	34.00	0.33	0.00	0.00
basidiomycota	40.33	38.00	59.67	22.33
caldiserica	0.00	563.33	0.00	0.67
cyanobacteria	446.33	1012.00	1445.33	1325.33
eukaryota	0.00	121.33	1.67	0.00
chlorophyta	105.67	5.67	0.33	0.00
deferribacteres	0.67	0.33	170.67	0.33
acidobacteria	327.67	0.00	0.67	0.00
gemmatimonadetes	11.67	27.33	0.00	0.33
planctomycetes	110.33	0.33	0.33	83.67
armatimonadetes	1.33	0.00	0.00	0.00
firmicutes	6229.67	7279.67	7505.33	8283.33
proteobacteria	12409.00	6842.00	7402.00	7819.00
fusobacteria	61.67	0.33	0.00	70.33
nitrospirae	19.00	10.33	0.33	0.00
deinococcus_thermus	8.67	262.67	107.33	0.33
verrucomicrobia	0.00	0.00	0.00	7.00

Figure 2: Average count of phylum across snow storms. DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit, DNA was sequenced using Illumina sequencing, Excel was used to analyze Illumina sequencing data. X axis showing the average count from three snow samples from each snow storm, Y axis showing the count of each phylum. Blue color means low count, red means high count. We can see red color throughout streptophyta, actinobacteria, firmicutes, and proteobacteria, acrss all 4 snow storms. Which indicates high abundance of these phyla.

phylum	T Test storm 1 v 2	T Test storm 2 v 3	T Test Storm 3 v 4	T Test Storm 4 v 1	T Test Storm 1 v 3	T Test Storm 2 v 4
ascomycota	0.18	0.07	0.42	0.35	0.18	0.72
bacteroidetes	0.54	0.40	0.86	0.41	0.63	0.24
streptophyta	0.04	0.22	0.70	0.69	0.25	0.27
chloroflexi	0.27	#DIV/0!	0.37	0.27	0.33	0.42
actinobacteria	0.57	0.14	0.88	0.44	0.51	0.13
spirochaetes	0.38	0.37	#DIV/0!	0.37	0.42	0.42
basidiomycota	0.92	0.72	0.53	0.38	0.76	0.41
caldiserica	0.37	0.37	0.12	0.12	#DIV/0!	0.42
cyanobacteria	0.36	0.57	0.86	0.14	0.18	0.65
eukaryota	0.25	0.26	0.24	#DIV/0!	0.30	0.31
chlorophyta	0.40	0.36	0.37	0.37	0.42	0.39
deferribacteres	0.52	0.37	0.37	0.52	0.42	1.00
acidobacteria	0.28	0.12	0.12	0.28	0.34	#DIV/0!
gemmatimonadetes	0.63	0.37	0.37	0.39	0.42	0.43
planctomycetes	0.14	1.00	0.38	0.81	0.21	0.42
armatimonadetes	0.37	#DIV/0!	#DIV/0!	0.37	0.42	#DIV/0!
firmicutes	0.54	0.89	0.71	0.36	0.53	0.60
proteobacteria	0.16	0.73	0.86	0.27	0.25	0.64
fusobacteria	0.21	0.37	0.37	0.92	0.27	0.42
nitrospirae	0.71	0.39	0.37	0.37	0.43	0.42
deinococcus_thermus	0.32	0.54	0.13	0.37	0.22	0.36
verrucomicrobia	#DIV/0!	#DIV/0!	0.37	0.37	#DIV/0!	0.42

Figure 3: T Test result comparing each snowstorm to the next snowstorm by phylum. DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit, DNA was sequenced using Illumina sequencing, Excel was used to perform T Test calculations. X axis showing comparison between snowstorms, Y axis showing comparison between phylum. Blue means low number (vary similar), red means high number, green indicates significant difference ($p<0.05$). Only two samples had significant differences: ascomycota between storm 2 and 3 as well as streptophyta between storm 1 and 2.

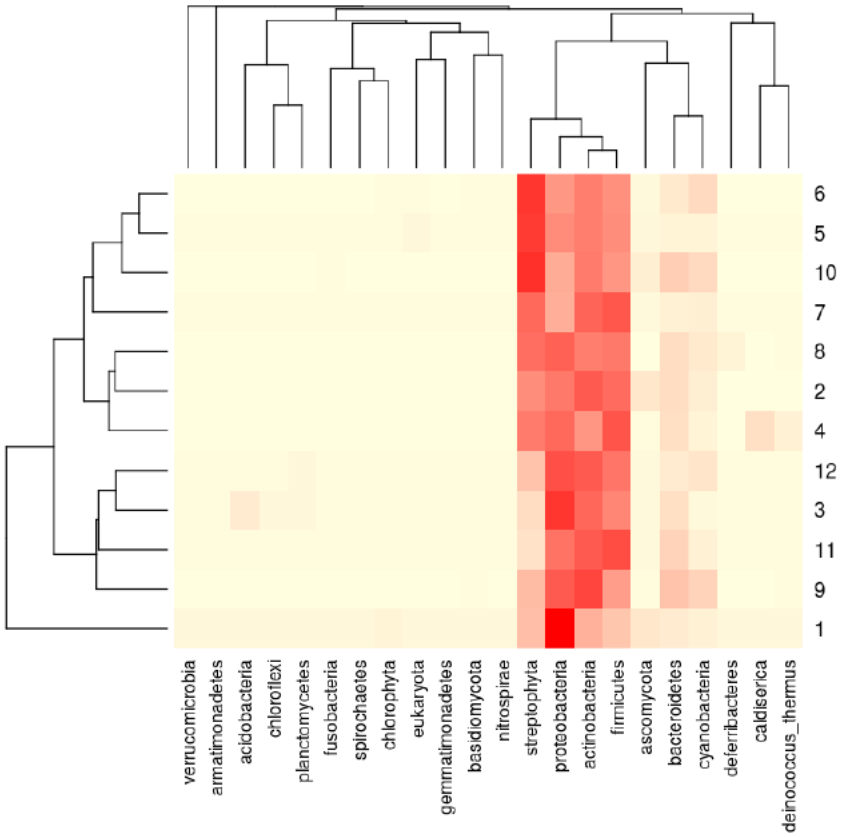


Figure 4: Heatmap and phylogenetic correlation between phylum and snow samples. DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit, DNA was sequenced using Illumina sequencing, R was used to create the figure. X axis showing concentration and correlation between content from each phylum, Y axis showing content from each snow storm. Red color indicates high abundance, lighter color indicates lower abundance. Streptophyta, proteobacteria, actinobacteria, and firmicutes, have high levels throughout all snow storms. No significant difference between the content of each snow storm.

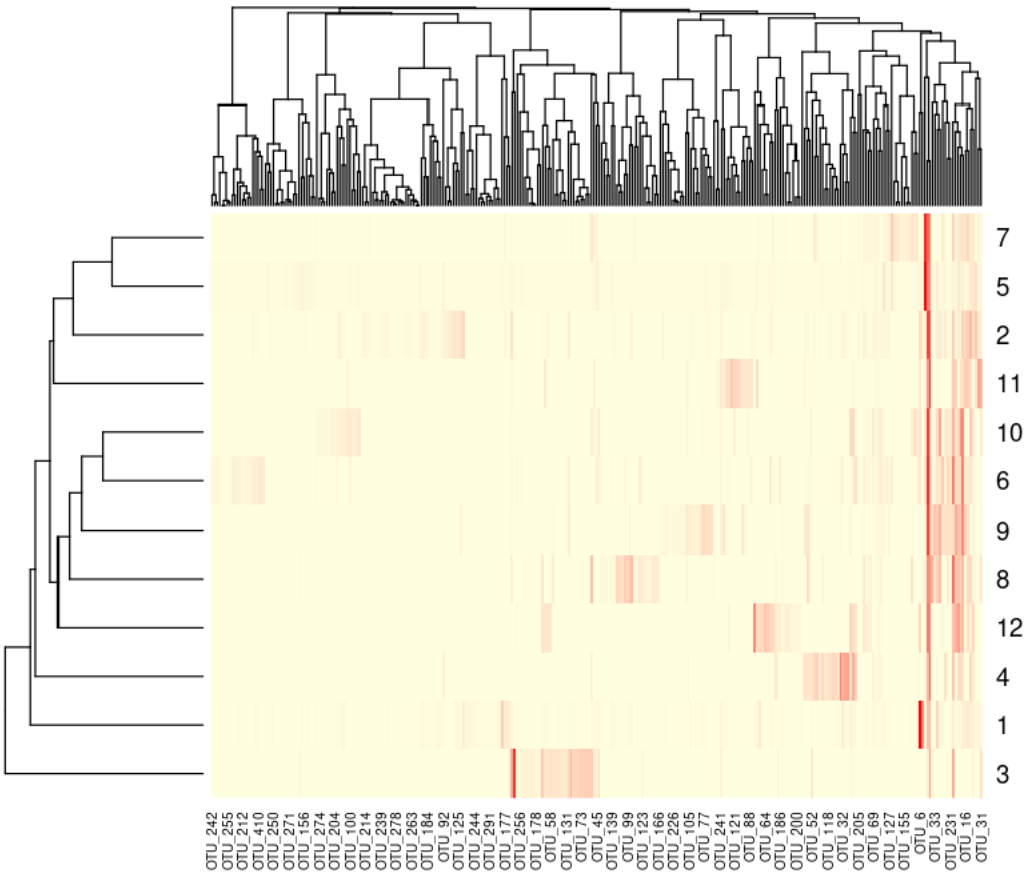


Figure 5: Heatmap and correlation of OTU content between snow samples. DNA was purified using Qiagen QIAamp Fast DNA Stool Mini Kit, DNA was sequenced using Illumina sequencing, R was used to create the figure. X axis showing concentration and correlation between content from each OTU, Y axis showing content from each snow storm. Red color indicates high abundance, lighter color indicates lower abundance. The most significant data is how we can see that there is a high abundance of certain species throughout multiple snow samples, indicating that similar bacteria end up in the same place. Shown by the many red cells on the right side.

Appendix F - Student Evaluations of the Course

Student Feedback – anonymous, collected by instructor
BIO 4160 – Molecular Biology
Spring 2022
n= 6 (out of 8 total students)

Questions #1-4 were scored on the following Likert scale:
1 = Fell way too short
2
3 = Acceptable
4
5 = Great improvement

1) To what extent did the course help you achieve the objective: Read, understand, develop, and practice research protocols in the process of answering a research question.
4.83

2) To what extent did the course help you achieve the objective: Discern the benefits and limitations of several molecular biology approaches.
4.5

3) To what extent did the course help you achieve the objective: Communicate your results in both written and oral form as would occur in a research laboratory environment.
4.5

4) To what extent did the course help you achieve the objective: Demonstrate proficiency in basic molecular biology skills including pipetting and bench organization.
4.83

5) In what ways were Assignments #1-6 [data figure assignments] beneficial to you?

- It helped me prepare for the presentation
- Gaining a better understanding of the goal of the entire research project
- Assignments 1-6 were very beneficial to me because they forced me to explore and research topics that I would not typically look into. Prior to this class, I did not know any of the content that was looked into within any of the assignments which kept the course fun, interesting, and unique.
- There helped me to improve my writing skills and prepare for the poster.
- It was very helpful to divide all the different parts of our experiment into different sections and assignments. Focusing on one aspect at a time was very useful to me.
- I thought that assignments 3, 5, and 6 were beneficial to learning how to make figure captions. Assignments 1, 2, and 4 helped me understand the information that was presented in the class, and why we were performing the techniques to obtain these results.

6) How could the Assignments #1-6 [data figure assignments] be improved?

- Not having word count limitation since we are using these as a reference for the poster putting that into the curriculum first or second week to explain the purpose of these assignments with a poster example
- More direction or background information on where to find sources or on what's to be expected
- The assignments could be improved by doing a walk through of how to make good figures and captions for them.
- Make the formatting more specific to what is needed on the paper.
- I think the content was perfect, needed to provide enough detail that you really needed to understand the content but short enough that it wouldn't take up hours. Perhaps it could be useful to incorporate peer review or a discussion post forum where you have to read what your peers are doing.

- Assignment 4 was interesting to learn about Sanger and Illumina Sequecing, but since it was not a large part of the poster, I feel like the video worksheet was enough to understand the two processes used. However, the feedback on the assignment was helpful, so it was still beneficial. Assignment 6 would be a lot more beneficial if the data was your own, which I understand is a time constraint of the course.

7) In what ways was the Poster and Presentation project beneficial to you?

- It helped put the ideas together
- Putting it all together so we can see all the research we did
- The poster and presentation were beneficial to me because I have not had to do a presentation on a poster in a formal way in quite a long time, so it refreshed all of my skills. It was also beneficial because it gave me one last chance to practice making posters.
- It gave me experience that I will be able to take into the future.
- Making a poster, both content and formatting wise is a good skill to learn. It was also a really good way to summarize our project in a nice visual representation. Another thing is that we had to include only the most important parts, which is difficult.
- I did not make many Posters in my past courses, so making the poster was a learning experience that helped me build some skills I did not have before. The Poster allowed for an overall understanding of everything completed in the course, which was fun to see in one final project.

8) How could the Poster and Presentation project be improved?

- Not sure
- More explanation on why we do this and more time to work on it
- N/A The expectations and guidelines were laid out very clearly and were easy to follow.
- Provide examples of the poster and video
- I'm not quite sure, I thought it was done really well. Particularly nice that we could get peer review and feedback from Dr. O'Brien before turning in the final version.
- I think the project could be improved if a little more time went into making and understanding what goes into the poster. For example, along with completing the assignments, the student could also start adding condensed information and images to a draft poster throughout the semester.

9) In what ways was the Case Statement experience beneficial to you?

- Make an argument as to what I did and why I deserve a certain grade is amazing
- It allowed me to reflect on my skills I've improved on and learned in class
- The case statement experience is very beneficial because it forces you to look back at your own work and judge how much you have improved over the course of the semester. It was a good chance for me to be honest with myself and make notes of how I could continue to improve in the future.
- was a nice way to wrap up the semester and take a look back on what we did.
- I had never done a case statement before so that was a new experience. I believe it was a nice assignment in terms of reflecting on what we have been doing and the outcome of the course.
- The Case Statement was beneficial in looking back at everything you learned through the course, and realizing how much progress you actually made. The case statement showed the benefit of this course, and made for a better understanding of why every assignment mattered.

10) How could the Case Statement experience be improved?

- Not sure
- More time to brainstorm what to put on it
- N/A guidelines were very clear and easy to follow
- Assign It sooner so it isn't put together as last minute.
- It was very open, which I believe is a good thing when reflecting. But it felt a little unclear what you were really asking and just starting was a little difficult.

I thought the case statement experience was very good, and since this was the first time I wrote one, I think a draft to receive feedback on would be beneficial into making the case statement more professional.

Mission, Review Process & Disclaimer

The Association for Biology Laboratory Education (ABLE) was founded in 1979 to promote information exchange among university and college educators actively concerned with teaching biology in a laboratory setting. The focus of ABLE is to improve the undergraduate biology laboratory experience by promoting the development and dissemination of interesting, innovative, and reliable laboratory exercises. For more information about ABLE, please visit http://www.ableweb.org/.

Papers published in *Advances in Biology Laboratory Education: Peer-Reviewed Publication of the Conference of the Association for Biology Laboratory Education* are evaluated and selected by a committee prior to presentation at the conference, peer-reviewed by participants at the conference, and edited by members of the ABLE Editorial Board.

Citing This Article

O'brien JH. And Kruchten AE. 2023. Snowflake CURE: Isolating microbial DNA from snow for metagenomic analysis to assist undergraduate students with learning molecular and biostatistical tool. Article 16 In: Boone E and Thuecks S, eds. *Advances in biology laboratory education*. Volume 43. Publication of the 43rd Conference of the Association for Biology Laboratory Education (ABLE). https://doi.org/10.37590/able.v43.art16