Next-Generation Teaching: A Template For Bringing Genomic and Bioinformatic Tools into the Classroom

Supplementary Materials

Supporting Information

Bioinformatics

As a first step in exploring our NGS data and to get students introduced to using a UNIX command-line computing environment, we merged our forward and reverse reads (the two directions of sequencing reads generated by the paired-end approach described earlier) using the program FLASH (Magoč and Salzberg 2011) with settings: -m 20 -M 250, which specified the minimum overlap for merging (-m) as 20 basepairs (bp) and the maximum overlap (-M) as 250 bp. The purpose of this step was to improve the quality and length of our sequence data by merging forward and reverse reads together into one combined sequence. Next, we filtered our data to remove low-quality (i.e., those reads that were unreliable due to failing the Illumina "chastity filter" or having low-quality sequence scores) reads using the FASTX-Toolkit package (Hannon, 2010). The FASTX-Toolkit package is a collection of command line tools, designed to work with processing 'short-read' sequence data, the type of data generated by Illumina NGS platforms. Specifically, we used the fastq_quality_filter script with settings: -q 25 -p 80, which meant that 80% of all bases must have a Q33 quality score of 25 or greater.

After read merging and quality filtering, we moved on to the analysis of microbial diversity in our data. For all analyses, we utilized the QIIME (pronounced 'chime') pipeline (Caporaso et al. 2010) due to its popularity in the field, integrated 'pipeline' (meaning many modules act on the output of previous components of the program), and the ease of generating output plots,

community statistics, and taxonomic identifications. Another benefit of the QIIME pipeline is a well-documented online manual (http://qiime.org/) that provides overviews of specific tools, data analysis tutorials, and example files included with the program source code. As a pre-packaged, well-documented but still completely command-line driven framework, QIIME provides a relatively easy-to-use introduction to command-line programming while also providing the tools to properly analyze 16S rRNA microbial data.

In QIIME, we first converted our sequence data from FASTQ to FASTA format using the convert_fastaqual_fastq.py script, as we no longer need quality scores for our data after initial filtering (see an example script for this in the *.zip file included with this study's Supplementary Materials). Next, we built a "mapping file" per the QIIME documentation instructions. This file specified which libraries corresponded to which sampling treatments (an example mapping file is also included in the *.zip file). We checked this mapping file for errors using the check_id_map.py script. Note: This step almost always generates at least one error. If you don't see more than a single error, your mapping file is likely fine.

Next, we added QIIME specific labels to each of our sequence libraries and combined them all into one large FASTA file using add_qiime_labels.py. We picked operational taxonomic units (OTUs) using the script pick_de_novo_otus.py, which compares the sequence data contained in the combined FASTA file with publicly available 16S rRNA sequence databases (e.g., Greengenes) to assign taxonomy to the full dataset (and each sampling treatment). On a 2015 MacBook Pro, this step took approximately one hour to execute. The OTU table is generated as a Biological Observation Matrix (BIOM) and we generated basic summary stats BIOM table using the summarize-table script. In 2014 and 2015, we removed libraries from all analyses when sequencing yield was less than 30K sequences for a given library. In 2016, we reduced this

number because we ran a "Nano" MiSeq run instead of the full, more expensive run option. We elected to make this change to save money as we were no longer pursuing publication-quality results, and rather, focusing on student learning objectives solely. Next, we summarized communities by taxonomic composition with the summarize_taxa_through_plots.py script. Depending on number of treatments (e.g., for 2014 our treatments were: Treatment 1 = gender, Treatment 2 = location, Treatment 3 = restroom surface) this script can be run for the full dataset then for as many individual treatments as necessary by invoking the -c flag. We computed alpha and beta diversity using the scripts alpha_rarefaction.py and beta_diversity_through_plots.py. The last three scripts produce a multitude of visual output in the form of bar and line graphs as well as principal coordinate analyses. This output can be opened as an HTML file in a web browser (e.g., Google Chrome), allowing students to interact with the data by selecting various treatments, diversity metrics, or other choices. It should be noted that the outline described here is only a subset of the QIIME pipeline, many additional tools for curating, analyzing, and exploring the data are available.

Removing contaminant OTUs

One particularly useful component of the QIIME pipeline for education is the investigation of possible contaminants OTUs. By collecting (and sequencing) the necessary negative controls (e.g., field, classroom, and PCR), it is possible to identify OTUs at each stage, consider them as a class (e.g., the bacterial genus *Pseudomonas* is associated with human acne and a common contaminant in 16S experiments), and remove the ones that appear likely to be a product of contamination. This process is outlined in Step 6 of the accompanying "rRNA Instructor Analysis Guide.sh" although it should be noted that we have not incorporated into our

course materials. Briefly, the process begins by looking in the taxa summary data for each negative control (whether PCR, classroom, or field) and identifying if particularly dominant OTUs are contaminants or not. This isn't always clear and a combination of literature and internet searching can typically help identify obvious ones. Next, the contaminant OTUs need to be identified and copied into a list for removal. Following this step, the script filter_otus_from_otu_table.py along with the list for removal can be combined to create a new data table the contaminant OTUs removed. A follow-up summary of the output table using the command 'biom summarize-table' is useful for confirming the contaminant OTUs have been removed (e.g., there will be a noticeable before and after difference when comparing pre- and post-contaminant filtering summaries).

Results

For UK bathrooms, when microbial diversity was compared across genders, no clear pattern emerged (Figure S1A). However, when the same samples were grouped by location, some loose, location-specific clustering did emerge, specifically for William T. Young Library and the Student Center (Figure S1B). At the UK climbing wall, no clear patterns emerged for microbial diversity across difficulty or type of holds (Figure S1C-D). One outlier sample, footholds of 'Hard' routes was disconnected from the rest of the samples (Figure S1C-D). Across years, a much wider variability in microbial diversity was observed for 2014 samples as the top two principal components (PCs) explained 30.8% and 8.2% of the variation whereas in 2015, the top two PCs explained 9.9% and 7.3% of the variability among samples. Comparisons of between sample (alpha) diversity further show the difference between projects in variability among treatments. In general, between treatment diversity was much lower in 2015 versus 2014

(Figure S2). When biodiversity metrics were compared across treatments and iterations of the experience, no major patterns emerged. For Shannon diversity (SD), the highest diversity observed in 2014 was for sink samples from the Main building in male bathrooms (SD = 6.61; Table S3). For 2015, the highest diversity in a single treatment was observed for a foothold high on an easy route (SD = 6.41; Table S4) and the five highest diversities observed were all on easy routes (Table S4).

Discussion

Additional analytical directions

In this CURE, many analytical stones were left unturned. This was largely a product of the course objective to be a component in a larger course that was not solely focused on genomic education. With more time, or in a post-course independent research setting, a number of complementary analyses could be explored. Chief among these is the question of contamination. Built into the QIIME pipeline are tools for bioinformatically removing OTUs called from negative or control samples from the broader OTU database. For instance, in 2015, it would be possible to remove those OTUs associated with 'clean' climbing hold from all of the OTU libraries linked to treatment holds and thereby clarify diversity with this common library of control OTUs removed.

Furthermore, QIIME provides a mechanism for exploring specific taxonomic diversity at any level of organization from kingdom to species. Taking a taxonomic-focused perspective on the data, perhaps with a human health slant, would provide students with another outlet for understanding where microorganisms occur and the frequency with which putative pathogens, for instance, are present in the environment.

References

1. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., *et al.* (2010). QIIME allows analysis of high-throughput sequencing data. Nature Methods, *7*, 335-336.
2. Hannon, G. (2010) FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/
3. Magoč, T., Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics, *27*, 2957-63.

Supplementary Figures

FIGURE S1. Three-dimensional principal coordinate analysis plots of prokaryotic beta diversity for two years of the course-based undergraduate research experience. While plots appear static in this representation, they are interactive for students and can be rotated, scaled, or modified in a number of ways. (A) 2014: A comparison of prokaryotic diversity across genders in University of Kentucky (UK) bathrooms. (B) 2014: UK bathroom prokaryotic diversity by location. (C) 2015: Prokaryotic diversity of the climbing wall at the UK Johnson Center with samples colored by route difficulty. (D) 2015: Prokaryotic diversity of the climbing wall at the UK Johnson Center with samples colored by hold type.

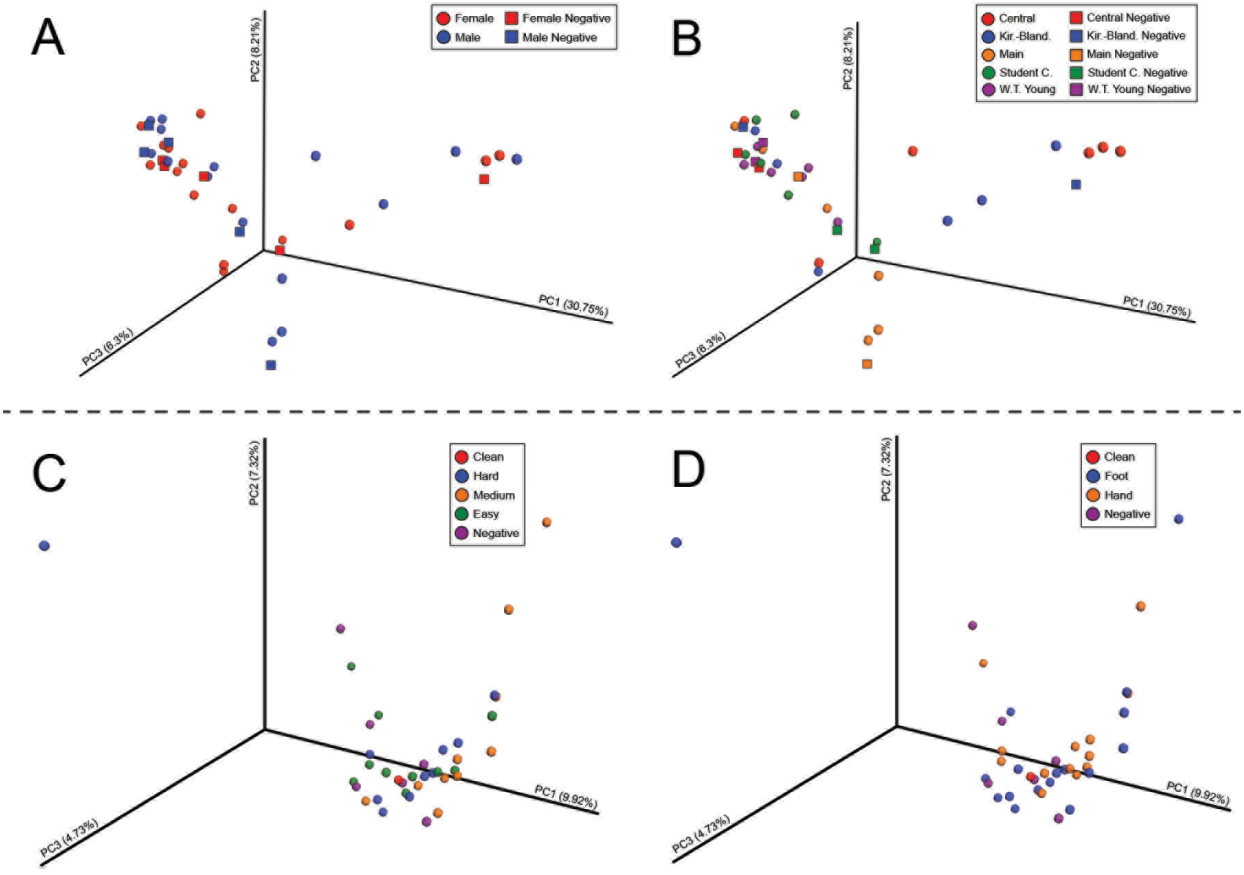
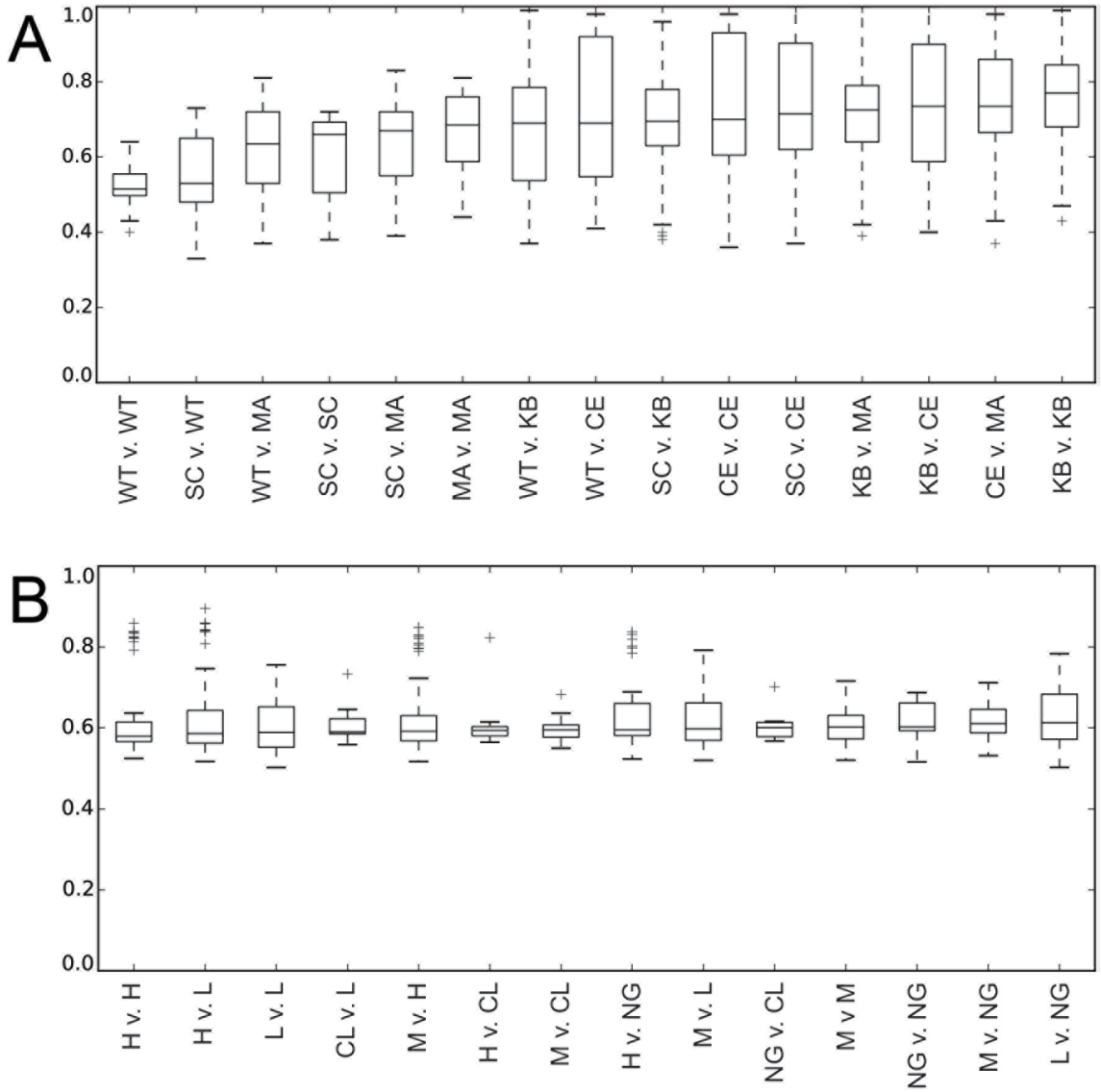


FIGURE S2. Within and between treatment prokaryotic beta diversity comparisons for two years of the course-based undergraduate research experience. Hash marks represent individual samples. (A) 2014: location comparisons (WT = William T. Young Library, SC = Student Center, MA = Main Building, KB = Kirwan and Blanding Dormitories, CE = Central Dormitory. (B) Route height comparisons (H = high, M = middle, L = low, CL = clean hold, NG = negative).



Supplementary Tables

TABLE S1. Weekly schedule of laboratory equipment we used for a section of the course with ~20 students.

Item	Quantity	Notes
Week 2 – Sampling and DNA extraction		
Sterilizing tubes	1 per group	Benchtop heated tubes for sterilizing scissors
Scissors	5+ pairs	Metal with no coating (so they can be sterilized)
Week 3 – PCR amplification		
Thermocycler	1 per lab	
PCR primers	1 pair per library	Earth microbiome primers. See Bates et al. (2011) *ISME* and Gilbert et al. (2014) *BMC Biology*. http://press.igsb.anl.gov/earthmicrobiome/emp-standard-protocols/16s/
Week 4 – Gel electrophoresis		
Centrifuge	1 per group	
Gel electrophoresis rig and power supply	1 per group	
Weeks 5-7 – Bioinformatics		
Computer	1 per student	We used MacBook Pros
Miscellaneous equipment		
Containers for ice	1 per group	For reagents and extracted DNA
Micropipettes	1 set per 2 groups	10 μL, 100 μL, 1000 μL
Kimwipes	1 box per group	
Tube racks	1 per group	

TABLE S2. Budget and item descriptions for resources needed for implementing this course-based undergraduate research experience in prokaryotic diversity. All costs are in USD.

Item	Quantity	Cost	Catalog number or link
Illumina MiSeq	1 lane	$1371[a]	
Sterile culture swabs	100-pack	$36.74/each	B4320115
Extract N Amp PCR Ready mix	1.2 mL bottle	$116.50/each	E3004
Extraction buffer	24 mL bottle	$86.10/each	E7526-24mL
Dilution buffer	12 mL bottle	$33.60/each	D5688-12mL
EZ Vision DNA dye	1 kit	$139.12/kit	N313-KIT
515f/806r primers	1 pair per library	$43.50 per forward and reverse pair	Earth microbiome primers. See Bates et al. (2011) *ISME* and Gilbert et al. (2014) *BMC Biology*. http://press.igsb.anl.gov/earthmicrobiome/emp-standard-protocols/16s/ https://press.igsb.anl.gov/earthmicrobiome/emp-standard-protocols/primer-ordering-and-resuspension/
2 mL screw cap tubes	1 bag (500)		
PCR strip tubes and lids			
1 kb ladder			
1.75% agarose gels			
Various software packages	n/a	Free	All software used is freely available
MacBook Pro	30	$1,499	Provided by the UK Department of Biology for use in a wide array of courses

[a] in-house rate at the University of Kentucky as of Fall 2015

TABLE S3. Prokaryotic alpha diversity across treatments and metrics for the 2014 iteration of the course-based undergraduate research experience. Treatments are sorted by Shannon diversity in descending order. Observed = total observed OTUs.

Location	Surface	Gender	Shannon	Observed	chao1
Main	Sink	M	6.61	923.8	1740.2
Student Center	Negative	F	6.59	732.6	1555.1
Student Center	Sink	F	6.24	722.7	1132.8
Student Center	Negative	M	6.19	618	1059.1
W.T. Young	Faucet	M	6.1	872.3	1819.8
Main	Negative	M	6.01	622.9	1224.4
W.T. Young	Sink	M	6	633.4	987.6
Central	Door	F	5.97	653.9	1197.9
Kirwan-Blanding	Door	F	5.97	963.3	2842.6
Main	Door	M	5.95	599.8	1045.6
Central	Faucet	M	5.91	658.4	1082.1
Kirwan-Blanding	Sink	M	5.85	688.7	1239.4
Kirwan-Blanding	Sink	F	5.77	624.3	1204.9
Main	Door	F	5.71	642.5	975.9
Main	Sink	F	5.56	675.4	1032.6
W.T. Young	Door	M	5.56	633.2	1008.2
Central	Negative	F	5.46	590.8	900.0
Student Center	Door	F	5.45	691.9	1088.3
Student Center	Sink	M	5.4	618	953.2
Main	Faucet	F	5.39	683.6	1029.9
W.T. Young	Sink	F	5.37	625.8	1036.0
Central	Door	M	5.35	696.7	1129.3
Main	Negative	F	5.31	495.1	999.3
Central	Sink	M	5.3	577.6	1764.3
W.T. Young	Door	F	5.3	635.8	974.3
Kirwan-Blanding	Faucet	F	5.23	554.4	894.9
Kirwan-Blanding	Faucet	M	5.17	633.9	1501.7
Kirwan-Blanding	Negative	M	5.17	517.8	874.3
Student Center	Door	M	5.13	655.1	994.8
Kirwan-Blanding	Door	M	5.09	707.7	1099.7
W.T. Young	Faucet	F	5.07	608.4	972.4
W.T. Young	Negative	F	5.05	500	1052.3
Central	Negative	M	5.04	423.4	798.3
Kirwan-Blanding	Negative	F	5.04	692	1607.8
Student Center	Faucet	F	5.04	632.2	1043.9
Central	Sink	F	5.02	573.9	1830.0
Central	Faucet	F	5	547.7	1280.0
W.T. Young	Negative	M	4.98	435.1	1115.5
Student Center	Faucet	M	4.92	544.4	853.2
Main	Faucet	M	n/a	n/a	n/a

TABLE S4. Prokaryotic alpha diversity across treatments and metrics for the 2015 iteration of the course-based undergraduate research experience. Treatments are sorted by Shannon diversity in descending order. Observed = total observed OTUs.

Difficulty	Height	Hold type	Shannon	Observed	chao1
Easy	High	Foot	6.41	950.7	1706.9
Easy	Low	Hand	5.61	756.9	1238.2
Easy	Middle	Hand	4.93	556.7	819.9
Easy	Middle	Foot	4.83	651.3	939.0
Easy	Middle	Foot	4.71	551.8	804.2
Medium	Low	Foot	4.67	620.7	933.5
Easy	Middle	Negative	4.67	556.9	798.9
Clean	Clean	Clean	4.65	411.3	593.3
Easy	Low	Negative	4.65	491.1	696.4
Easy	Low	Hand	4.50	537.6	759.8
Hard	Low	Foot	4.49	543.8	804.1
Hard	Low	Foot	4.48	162.5	415.9
Medium	Middle	Hand	4.45	466.2	678.9
Hard	Low	Hand	4.39	348.2	564.0
Hard	Middle	Hand	4.35	429.6	657.2
Easy	Low	Negative	4.31	365.7	584.5
Hard	Middle	Foot	4.30	311.3	477.9
Hard	High	Foot	4.26	427.4	631.0
Medium	Low	Hand	4.25	497.4	748.2
Hard	Low	Hand	4.24	467.5	642.8
Medium	Middle	Foot	4.17	268	404.7
Easy	High	Hand	4.11	453.1	697.0
Medium	High	Foot	4.10	378.3	568.7
Easy	High	Negative	4.06	340.9	599.0
Hard	Middle	Hand	4.06	475	734.9
Hard	Middle	Foot	4.03	340.8	499.2
Easy	Middle	Negative	3.98	310.8	595.4
Medium	High	Hand	3.96	386.2	548.5
Hard	High	Hand	3.92	401.3	586.6
Easy	Low	Foot	3.92	537.3	822.7
Easy	Low	Foot	3.73	302.2	425.3
Easy	High	Negative	3.68	304.1	672.2
Easy	High	Foot	3.65	357.6	562.9
Medium	High	Hand	3.60	341.7	579.9
Medium	Middle	Foot	2.75	360.7	578.7
Medium	Low	Foot	1.71	311.3	676.2
Medium	Middle	Hand	1.68	294	641.3

Supplementary Materials – Instructor Guide and Example Scripts.zip:

For a step-by-step run through of the analyses performed with more computational detail, see the provided document 'rRNA Instructor Analysis Guide.sh'. We recommend working in TextWrangler (http://www.barebones.com/products/textwrangler/) or an equivalent text editor (not Microsoft Word).

We have also included several template scripts.

Supplementary Materials – Handouts and Protocols:

The following are all of the handouts and protocols used in 2016[1,2].

[1] There is no Week 1 handout because the first was a discussion of the project and experimental design. Students were tasked with developing hypotheses in small groups for the class to vote on.

[2] Though we did not require students to turn in answers to handout questions, we have provided general thoughts for what we were steering them towards in red.

Week 2 – Sampling and DNA Extraction.

It is the third BIO 198 16S rRNA prokaryotic next-generation sequencing project and this year we conducting a project to address hypotheses of prokaryotic species diversity in elevators in buildings on the UK campus.

We will be able to address a number of hypotheses with this study. Some of these include:

1. The originally proposed hypotheses is that taller buildings will facilitate greater elevator use, relative to shorter buildings, and that the "up" elevator buttons in the taller buildings will have greater prokaryotic diversity and abundance compared to shorter buildings.

2. In addition, buildings on campus vary in their age, with some very new buildings having less human traffic and use, relative to older buildings of the same height. We may expect elevators in the older buildings to have greater prokaryotic diversity and abundance.

3. And there are a number of additional ways in which we may be able to compare samples drawn from within and between elevators on campus. These will not be enumerated here, but I bet everyone can identify the possibilities from the sampling design.

To test these hypotheses, we will perform environmental sampling from elevators in three different buildings on campus: The Jacobs Science Building (JSB), Patterson Office Tower POT, and the Thomas Hunt Morgan (THM) Building. The elevators across each of these buildings vary in their controls, and so we won't sample exactly the same surfaces, but there will be overlap.

Each lab section will be divided up into five groups. Individual group sampling breaks down like this:

Group 1
JSB north elevator
Built in 2015

Sample 1 Ground floor outside up button
Sample 2 Inside 1st floor button
Sample 3 Inside 2nd floor button
Sample 4 Inside 3rd floor button
Sample 5 Inside 1st floor push bar
Sample 6 Inside 2nd floor push bar
Sample 7 Inside 3rd floor push bar
Sample 8 Alarm button
Sample 9 Clean swab (negative)

Group 2
JSB service elevator
Built in 2015

Sample 1 Ground floor outside up button
Sample 2 Inside 1st floor button
Sample 3 Inside 2nd floor button
Sample 4 Inside 3rd floor button
Sample 5 Inside Basement button
Sample 6 Inside Penthouse button
Sample 7 Alarm button
Sample 8 Clean swab (negative)

Group 3
POT Elevator A
Built in 1968

Sample 1 Ground floor outside up button
Sample 2 Left panel, 1st floor button
Sample 3 Left panel, 6th floor button
Sample 4 Left panel, 12th floor button
Sample 5 Left panel, 18th floor button
Sample 6 Mezzanine button
Sample 7 Alarm button
Sample 8 Clean swab (negative)

Group 4
POT Elevator D
Built in 1968

Sample 1 Ground floor outside up button
Sample 2 Left panel, 1st floor button
Sample 3 Left panel, 6th floor button
Sample 4 Left panel, 12th floor button
Sample 5 Left panel, 18th floor button
Sample 6 Mezzanine button
Sample 7 Alarm button
Sample 8 Clean swab (negative)

Group 5
THM south elevator
Built in 1974

Sample 1 Ground floor outside up button
Sample 2 1st floor button
Sample 3 2nd floor button
Sample 4 3rd floor button
Sample 5 Alarm button
Sample 6 Clean swab (negative)

This results in a total of **39 samples** that will be collected during each lab section. Each lab section will collect the same set of samples, and perform DNA extractions and PCR for these samples. This will result in three replicate samples of each surface environment, which will be pooled prior to sequencing.

Surface Sampling Protocol

Throughout this work, handle materials and perform work while always aiming to minimize the potential for contaminating samples. Wear clean gloves when handling materials and collecting samples. Avoid touching non-target surfaces (and yourself) once you have put on your gloves.

Preparation

1) All students will perform sampling in each section, with each student working within their group to help sample the different buttons or bars associated with their elevator. Before heading over to the different buildings and elevators, we will assemble the following materials in the lab. We will wear gloves when collecting these items. Each group needs enough of these for each student.
   - Clean gloves to be used when you have arrived at your sampling location. Place these in a Ziploc bag.
   - A set of sterile swabs (contained in sterile tubes). Collect enough of these for the number of samples to be collected from your elevator.
   - Sharpie for labeling tubes.

2) Descriptively label sterile swabs to account for all of the individual surface samples to be collected. Include section number (i.e., 001, 002, or 003), elevator in abbreviated format (e.g., JSB north elevator = JSBN), and sample number (1, 2, 3, …..).

   For example: 001, JSBN, 1

3) Note that one tube in each group will be used as a negative control. This swab will be opened when you get to the elevator, and then simply placed back in its tube without actually swabbing anything.

4) Assemble all of the materials that will be used for each elevator sampling. Place into a bag.

5) Head to your elevator with Professor X or a TA!

Note: Remember that sample collection steps will involve both targeted samples and a negative control!

Sampling

1) When you have reached your elevator, carefully put on a clean set of gloves.
   - You should have a plastic bag with you containing multiple sets of clean gloves.
   - Identify one person in your group to serve as the button pusher (to open the door) and door holder.

2) Next, collect one negative control sample. To do so, open one sterile swab and remove it from its tube, then replace it back into the tube.
   - Carefully remove sampling swab from its container to maintain sterility.

3) Next, sample from the outside "up" button. Use the appropriate swab to gently brush the button. You do not need to press super hard to pick up cells.
   - Carefully remove sampling swabs from their container to maintain their sterility.
   - Carefully place each individual swab in its proper sample tube and apply cap.

4) Hail the elevator and step inside to sample the buttons inside the elevator. Repeat Step 3 for a the remaining button or push bar surfaces.
   - Each student can be responsible for 1-2 samples within their elevator.

5) Once you have finished sampling, head back to the lab. If you're one of the POT groups, hustle!

DNA Extraction

1) When everyone has returned to the lab, glove up!

2) Each group will obtain:
   - 1-9 sterile 2.0 ml screw-cap tubes, depending on how many samples you collected
   - Scissors for cutting swab tips.
   - 2.0 ml tubes containing Extract-N-Amp Plant **Extraction** Solution. These tubes will be labeled specifically for each group and will contain enough solution to use in your extractions. Be sure to obtain the correct tube.
   - 2.0 ml tubes tube containing Extract-N-Amp Plant **Dilution** Solution.These tubes will be labeled specifically for each group and will contain enough solution to use in your extractions. Be sure to obtain the correct tube.
   - Sharpie for labeling tubes.

3) Label the 2.0 ml screw-cap tubes to correspond with each surface sample that your group is working with.
   - Use the same labeling style that you used for your sterile swabs.
   - Also note on the tube that it contains DNA.
   - It is often a good idea to label both the cap and the tube, in case one of them is accidentally removed (e.g., with EtOH).

3) Prepare your work surface by cleaning with a light application of 95% Ethanol (EtOH). Use care when working to limit contamination of surface and supplies

4) Sterilize your scissors using a heat sterilizer.

5) Carefully remove one of your sterile swabs and place it into its corresponding 2.0 ml tube. Then using your scissors, cut off most of the stick so that the end with the cotton tip fits in the tube with the cap closed.
   - Before you cut the stick, pull the swab up slightly so that it is not touching the bottom of the tube, which will allow it to drop a bit when you make your cut.
   - Discard the other end of the stick in your waste container.
   - Use care to make sure you are transferring swabs to their appropriately labeled tubes.

6) Repeat steps 4 and 5 for all of your sample swabs, including the negative controls.
   - Pay attention to using good aseptic (sterile) technique. Always be thinking about how you can avoid cross-contaminating tools, reagents, samples, etc.

When you have transferred all of your sample swabs to their 2.0 ml screw-cap tubes:

7) Add 250 µL of Extract-N-Amp Plant **Extraction** Solution to each tube and re-cap the tubes.
   - Be sure to use a new tip for every tube! Avoid cross contamination.
   - We suggest you do each tube individually, recapping before moving on to the next one.

8) When all groups have completed step 7, place them in a floating tube rack and heat in a 95^{o}C water bath for 10 minutes. Use a timer to monitor this time.

9) Centrifuge your tubes at 2500 g for 5 minutes. Ask for assistance if you have never loaded a centrifuge before. Everything should be properly balanced.
   - Use care when removing your tubes from the centrifuge so that you do not splash liquid around inside the tube. We want everything to remain at the bottom so that you can remove the lid without anything coming out of the tube.

10) Add 250 µL of Extract-N-Amp Plant **Dilution** Solution to each tube. When adding the solution, keep your pipette tip in the liquid at the bottom of your tube and slowly pipette up and down 2-3 times. Do this delicately so as to not splash liquid up out of the tube.
   - Again, you should do this one tube at a time, closing each tube before you move on to the next one.
   - Use a new tip each time.

11) Collect all your tubes in your tube rack, label the rack with a piece of tape/sharpie with your group initials, location, and date. Give your samples to an instructor to be stored at 4^{o}C until next week.

Week 3 - PCR Amplification of the v4 region of prokaryotic 16S rRNA

This week you will work in the same groups as in Week 4 when you collected and prepared your DNA samples.

Our goal this week is to amplify copies of an ~250 bp segment of the 16S rRNA gene from the various prokaryotic genomes in your DNA sample. This segment is referred to as the v4 region because it is one of nine variable regions of the 16S rRNA gene in prokaryotes. We will use the variation in sequence data from the v4 region to discriminate among the different prokaryotic species that we have sampled in our bathroom surfaces.

From each elevator, you collected a range of surface samples. We will perform PCR on each sample using uniquely barcoded (also referred to "indexed") PCR primers. So, for example, PCR of the north elevator in the Thomas Hunt Morgan Building will have the following reactions:

001 Group 5 samples	# of PCRs
Sample 1, ground floor up button	1
Sample 2, 1st floor button	1
Sample 3, 2nd floor button	1
Sample 4, 3rd floor button	1
Sample 5, Alarm button	1
Sample 6, Clean swab (negative)	1
PCR negative	1
Total	7

Table 1. Example of the PCR work to be performed in today's lab, using the THM north elevator samples.

Let's pay attention to a number of aspects of our sampling design here and our PCR work on these samples.

First, each section will be performing a replicate of our overall elevator sampling design, meaning that each surface sample will be PCRd three times. **Can you think of a good reason for us to perform these PCR replicates across sections?**

SUGGESTED ANSWER: We performed PCR replicates across sections incase individual PCR replicates failed.

Second, note that we are performing PCR on our clean swab samples (the ones you opened and then closed at your particular elevator. **Why do we perform PCR for these samples?**

SUGGESTED ANSWER: We perform PCR for our negative swabs to amplify any possible contaminants so that we can sequence them (and see what they are) later in the project.

Third, also note that we will perform a negative control for your PCR. This is separate from the PCR we perform on your clean swab. **What is a PCR negative control? Why do we do this?**

SUGGESTED ANSWER: A PCR negative control is an identical reaction with water substituted for DNA. This provides a control for in case reagents (and not our DNA sample) are contaminated. Amplification in the negative control would indicate contamination.

Fourth, note that a different number of PCRs will be performed across groups within a section. Make sure you pay attention to setting up PCR in a way that accounts for your group's samples.

Finally, remember that for each surface, we will want to generate PCR amplicons (amplified stretches of DNA for our target gene region) for our different surface samples that can be distinguished from each other in our Illumina sequence data. This will require us to use PCR primers that have different barcode (indexed) sequences for each set of surface samples. We will use the same index sequences for our PCR replicates of a particular sample across sections. You can reference Table 2 (see below) to see which sets of PCR primers you need to use for each of your surface samples.

Table 2. Information for 16S rRNA indexed primers. The Primer Mix name is the label given to the tube of combined forward and reverse indexed primers. This tube will be used in preparing the PCR cocktails. The individual forward and reverse primer names are listed for each primer mixture and refer to the primer names provided in Flores et al. (2013).

Group/elevator/Treatment	Primer Mix Name	Index Primer Forward	Index Primer Reverse
G1 / JSB north / Sample 1	Primer Index 1	v4.SA501	v4.SA701
G1 / JSB north / Sample 2	Primer Index 2	v4.SA502	v4.SA701
G1 / JSB north / Sample 3	Primer Index 3	v4.SA503	v4.SA701
G1 / JSB north / Sample 4	Primer Index 4	v4.SA504	v4.SA701
G1 / JSB north / Sample 5	Primer Index 5	v4.SA505	v4.SA701
G1 / JSB north / Sample 6	Primer Index 6	v4.SA506	v4.SA701
G1 / JSB north / Sample 7	Primer Index 7	v4.SA507	v4.SA701
G1 / JSB north / Sample 8	Primer Index 8	v4.SA508	v4.SA701
G1 / JSB north / Sample 9	Primer Index 9	v4.SA501	v4.SA702
G1 / JSB north / PCR Neg.	Primer Index 10	v4.SA502	v4.SA702
G2 / JSB service / Sample 1	Primer Index 11	v4.SA503	v4.SA702

G2 / JSB service / Sample 2	Primer Index 12	v4.SA504	v4.SA702
G2 / JSB service / Sample 3	Primer Index 13	v4.SA505	v4.SA702
G2 / JSB service / Sample 4	Primer Index 14	v4.SA506	v4.SA702
G2 / JSB service / Sample 5	Primer Index 15	v4.SA507	v4.SA702
G2 / JSB service / Sample 6	Primer Index 16	v4.SA508	v4.SA702
G2 / JSB service / Sample 7	Primer Index 17	v4.SA501	v4.SA703
G2 / JSB service / Sample 8	Primer Index 18	v4.SA502	v4.SA703
G2 / JSB service / PCR Neg.	Primer Index 19	v4.SA503	v4.SA703
G3 / POT A / Sample 1	Primer Index 20	v4.SA504	v4.SA703
G3 / POT A / Sample 2	Primer Index 21	v4.SA505	v4.SA703
G3 / POT A / Sample 3	Primer Index 22	v4.SA506	v4.SA703
G3 / POT A / Sample 4	Primer Index 23	v4.SA507	v4.SA703
G3 / POT A / Sample 5	Primer Index 24	v4.SA508	v4.SA703
G3 / POT A / Sample 6	Primer Index 25	v4.SA501	v4.SA704
G3 / POT A / Sample 7	Primer Index 26	v4.SA502	v4.SA704
G3 / POT A / Sample 8	Primer Index 27	v4.SA503	v4.SA704
G3 / POT A / PCR Neg.	Primer Index 28	v4.SA504	v4.SA704
G4 / POT D / Sample 1	Primer Index 29	v4.SA505	v4.SA704
G4 / POT D / Sample 2	Primer Index 30	v4.SA506	v4.SA704
G4 / POT D / Sample 3	Primer Index 31	v4.SA507	v4.SA704
G4 / POT D / Sample 4	Primer Index 32	v4.SA508	v4.SA704
G4 / POT D / Sample 5	Primer Index 33	v4.SA501	v4.SA705
G4 / POT D / Sample 6	Primer Index 34	v4.SA502	v4.SA705
G4 / POT D / Sample 7	Primer Index 35	v4.SA503	v4.SA705
G4 / POT D / Sample 8	Primer Index 36	v4.SA504	v4.SA705
G4 / POT D / PCR Neg.	Primer Index 37	v4.SA505	v4.SA705
G5 / THM north / Sample 1	Primer Index 38	v4.SA506	v4.SA705
G5 / THM north / Sample 2	Primer Index 39	v4.SA507	v4.SA705
G5 / THM north / Sample 3	Primer Index 40	v4.SA508	v4.SA705
G5 / THM north / Sample 4	Primer Index 41	v4.SA501	v4.SA706
G5 / THM north / Sample 5	Primer Index 42	v4.SA502	v4.SA706
G5 / THM north / Sample 6	Primer Index 43	v4.SA503	v4.SA706
G5 / THM north / PCR Neg.	Primer Index 44	v4.SA504	v4.SA706

428

Now let's look at what goes into each PCR. They will each be conducted in a 20 µL reaction that includes:

- 10 µL of Extract-N-Amp Ready Mix (Sigma-Aldrich)
- 1 µL of a mixture containing both the forward and reverse primers, each at 5 µM concentration. See Table 2 for guidance on which primer mixtures to use.
- 5 µL of PCR-grade dH_2O
- 4 µL of your prepared DNA sample (from Week 4).

Let's get started!

1) Prepare to take lab notes for all information you think is important while working through this lab protocol and exercise.

2) Obtain one strip of twelve 200 µL PCR tubes. You will use these tubes to cover:
   - the PCRs of your actual surface samples (including your clean swab).
   - one negative control for your PCR.
   - Each group will have left-over tubes that will not be used. No worries.

3) Label your tubes using a fine Sharpie industrial marker. Make sure you provide enough information on each tube so that they are clearly distinguished from each other, and that your tubes can be identified from other groups.
   - These tubes are small and have limited space for writing. It will be best if you develop a reduced labeling scheme that will allow you to distinguish different tubes (e.g., numbers) and record the details of each tube in your lab notes. Make sure to include information for section number and group!
   - It is also best if you put a label on both the side of the tube AND the top, just in case one of them rubs off.
   - Because your tubes will be connected to each other, you don't need to fully label every tube with all of this information.
   - Finally, place an X on the empty tubes at the end.

4) Collect ice using your ice bucket, then obtain the following reagents:
   - A microcentrifuge tube containing Extract-N-Amp PCR Ready Mix. Note that your group will obtain a specific tube of this Ready Mix with enough volume to perform your PCRs. Be sure to obtain the correct tube.
   - 0.5 mL microcentrifuge tube containing the primer mixes needed for the different surface samples that your group is working with. Be sure to retrieve the proper primer mix for each surface sample!
   - A 1.5 mL microcentrifuge tube containing 60 µL of PCR-grade dH_2O.

   Keep all of your reagents on ice. Some of these reagents are temperature sensitive. Furthermore, the Taq polymerase within your Extract-N-Amp PCR Ready Mix should be

kept inactive to prevent unwanted polymerase activity, which is done by keeping it nice and cold.

5) Obtain your DNA samples from last week. Using a centrifuge, spin them at high speed for a short amount of time (a few seconds) to make sure all of the liquid is at the bottom of the tube.
   - This is always a good idea before you open any of your tubes, including reagents. It is particularly important for your DNA tubes.
   - Keep your DNA samples on ice.

Preparing PCR: An overview.

To set up your PCRs, you will not pipette all reagents for each reaction individually. **Why not?**

SUGGESTED ANSWER: This answer is two-fold. First, we will be pipetting some very small volumes. It's easy to make an error with small volumes so it's best to avoid pipetting them when possible. Second, by combing reagents into a master cocktail or mix, we save significant time by reducing the number of individual pipettes needed.

Instead, a more appropriate way to prepare your PCR mixture is to create a cocktail of reagents that can be dispensed as aliquots to your individual PCR tubes. This cocktail will be prepared for all PCRs using reagents that are exactly the same, and omitting the things that make them different. In this case, each PCR will differ in the primers being used, and the surface-sample DNA. We will be adding the PCR primers and DNA separately to each tube. **How many different cocktails will you need for today?**

SUGGESTED ANSWER: If the PCR primers and DNA will be added separately, then everything else will be the same. Therefore, only one cocktail needs to be made for this approach.

Example: Let's say your intstructor is performing PCR to amplify the same gene region from a set of **five** DNA samples and he isn't concerned with separately barcoding these samples. He would prepare a cocktail that will have enough reagents for:

- a PCR for each of his five DNAs.
- a PCR to serve as a negative control (to test for contamination of PCR reagents).
- one extra PCR that allows for pipetting error. **Why have this extra?**

SUGGESTED ANSWER: We add an extra sample (or two) to make sure that we have enough for every reaction. Some volume is lost to normal pipetting (e.g., some liquid sticks to the pipette tip) so it's important to have a buffer when preparing the cocktail.

Therefore, he would prepare a cocktail for **seven** PCRs using the following reagents and volumes.

10 µL x 7	= 70 µL Extract-N-Amp Ready Mix
1 µL x 7	= 7 µL Primer Mix
5 µL x 7	= 35 µL PCR-grade dH_2O
	= **112 µL total volume of cocktail**

Using this 112 µL cocktail, he would dispense 16 µL to each of his six PCR tubes (five DNA tubes and on negative control). There will probably then be a little cocktail left over in the tube. No big deal.

He would then add 4 µL of the appropriate DNA to each tube to complete the preparation of a 20 µL PCR volume.

For the PCR negative control, the addition of 4 µL of PCR-grade dH_2O will be used to bring the PCR volume up to 20 µL.

What if Professor X needs to do a similar PCR with PCR primers that separately barcoded each sample?

SUGGESTED ANSWER: In this case, Dr. X would make a single cocktail that would not include PCR primers in it. The primers would be added separately to each reaction.

Week 4 - Gel electrophoresis of PCR amplicons

This week you will continue to work in the same groups that you were in to perform surface sampling and PCR.

We'll be using gel electrophoresis to determine if your PCR reactions successfully amplified the expected v4 segment of the 16S rRNA gene.

- In addition, this gel work will be used to determine if there is any potential contamination in your surface sampling (based on PCR of your swab negative controls).
- We'll also be able to check if there is any contamination in your PCR, either in the reagents, or in the preparation of the reactions (based on your PCR negative controls).

A little background on gel electrophoresis….

Electrophoresis of DNA is typically performed using an agarose gel. Agarose is a polysaccharide extracted from seaweed. In a powdered form, agarose will dissolve into a liquid buffer using heat (typically using a microwave). The melted agarose-buffer mixture can be poured into a mold to form a gel. As it cools, the gel will solidify; kinda like jello.

- The liquid buffer is usually a mixture of dH_2O, Tris, Acetic Acid or Boric Acid, and Ethylenediaminetetraacetic acid (EDTA). These buffers are usually referred to by the acronyms TAE and TBE, depending on their composition.
- A comb is used to create wells in the gel.
- DNA (PCR product in your case) will be pipette loaded into these wells.

The solidified gel is submerged in the same buffer (TAE or TBE) that was used to make the gel. This is done in a gel rig that contains a positive electrode (cathode) at one end and a negative electrode (anode) at the other end. DNA has an electric charge to it. When an electric current is applied to the gel rig (filled with TAE or TBE), the DNA will migrate towards the positive electrode.

- The TAE/TBE buffer has ions in it that facilitate the transmission of electrons from negative to positive end. Don't ever put dH_2O in the gel rig. Without ions, no electric charge will move through the gel.

The percentage of agarose in your gel will determine how fast DNA molecules will migrate through the gel matrix. In general, shorter DNA molecules will migrate faster than larger DNA fragments. If you have small DNA fragments, you may want the agarose percentage to be a bit higher so that they don't run too fast.

- Our gels will be made at 1.5% agarose to account for the expected 350 bp size of our PCR fragment.

To load DNA/PCR product into a well, it is first mixed with a loading buffer. This usually has some kind of dense molecule (e.g., sucrose) that helps everything fall to the bottom of the well.

A dye is also included that will give the loading buffer some color, and which has a molecule that will migrate with the DNA. This will allow you to track the movement of your DNA on the gel.

In addition, a DNA ladder is also loaded into a well on the gel, typically to one side or the other of the samples in the run. This contains a series of DNA fragments of known size. It allows you to compare your PCR results to known DNA sizes to assess whether your PCR was successful.

Finally, a molecule is added to your DNA/PCR material prior to loading that interacts with your DNA and allows it to be visualized under UV light.

- Historically, this has been Ethidium Bromide (or commonly written as 'EtBr'), which is also a known mutagen and carcinogen.
- In today's lab, we'll be using a non-toxic DNA dye provided in the EZ-Vision loading buffer.

Let's get started!

1) As usual, take notes throughout your lab activity today.

2) Obtain ice for your ice tray/bucket and the following items:
   - Your PCR tubes from last week. Remember, you should have three sets of strip tubes, one for each of your surface samples.
   - A tube containing 1kb DNA size ladder.
   - A tube of EZ-Vision Three DNA loading dye.
   - A tube of gel-loading buffer.
   - Two new strips of PCR tubes.

3) Centrifuge your PCR tubes to make sure everything is at the bottom of the tube. Remember, you just need to provide a quick spin. Make sure to balance your tubes!

First, we're going to make sure everyone gets a little practice with pipetting into the wells of agarose gels.

4) Obtain an agarose gel and place it into your gel rig.
   - Note that this gel has two rows of 12 wells.

5) Add 1X TAE buffer to the gel rig until it just covers the top of your gel.
   - This doesn't have to be exact, just make sure your gel is completely submerged.
   - If any air bubbles are present in the wells, gently push them out with a yellow pipette tip. Be careful not to puncture the gel.

6) In one of your empty strip-tubes, add 10 µL of gel loading buffer to each of the 12 tubes.
   - Be sure to use the gel loading buffer, and NOT the EZ-Vision DNA dye.

7) Next, pipette 9 µL of the loading buffer from the first tube and practice loading it into the first well on bottom of the gel.
   - This requires precision and a steady hand. And it definitely requires practice.
   - An instructor can help you with technique.
   - Be very careful not to puncture the gel with your tip. This is especially true for the bottom of the well. A punctured well bottom will result in most of your loading material leaking out.
   - Avoid having air in the end of your pipette tip before you submerge it into the buffer. If you do have a small amount of air, slightly press on your pipette plunger to expel the air.

8) Each student will practice loading two to three wells in the gel.
   - After you perform the practice loadings we'll move on to the next step. We can leave the loading buffer in the wells for now.

Now we'll move on to loading and running your PCR product on a gel.

9) Obtain a clean strip-tube and number the tubes according to your PCR numbers moving left to right. Label one additional tube with an "L" for ladder.
   - We will use these tubes to mix the EZ-Vision dye with a sample of our PCR.

10) Pipette 2 µL of EZ-Vision Three DNA dye into each of the newly labeled tubes.

11) Next, using the PCRs you performed last week, pipette 8 µL of each into the tubes containing the EZ-Vision Three DNA dye. There now should be a total of 10 µL (8 of PCR product and 2 of dye) in each of the labeled tubes.
   - Do this pipetting moving left to right, and match each PCR tube to the corresponding tubes containing your EZ-Vision Three DNA dye.

14) Pipette 5 µL of the DNA ladder into the corresponding tube containing EZ-Vision Three DNA dye.

15) Now, pipette 9.5 µL of each of your PCR-dye mixtures and load them into the wells of the top row of your agarose gel. It's OK to leave a tiny amount left in the tube.
   - Use the top row! Not the bottom row that you used for practice loadings.
   - Again, avoid having air in the end of your pipette tip before placing it into the TAE buffer when loading.

16) Load 6.5 µL of your ladder in the next available well of your gel.

17) Place the lid on your gel rig. Check to make sure the connectors are properly secured to the power source. And check to make sure the bottom of your agarose gel is closest to the positive node of the gel rig.

18) Turn on your power source and set the voltage to 75.

19) ***Keep an eye on the tracking dye on your gel.*** The EZ-Vision Three DNA dye contains three tracking dyes that migrate at different rates.
- The light blue dye (Xylene cyanol) runs at about 4000 bp. Big.
- The darker blue dye runs at about 400 bp. Just a bit bigger than our expected fragment size.
- The reddish band runs at about 10 bp. Smaller than one of our primers.

19) When your gel has finished running, turn off the power supply, remove the lid to your gel rig, and bring your gel to the gel visualization equipment.
- This consists of a UV light box, which will be used to see DNA within the gel.

20) With the help of an instructor, check your gels using the UV system to assess the success of your PCR and the presence/absence of amplification in your negative controls.
- If your PCR was successful, you should see a band, or a series of bands, that are ~350 bp in size.
- Hopefully, you will not detect amplification in your negative controls. But, it might be there. We can discuss more.

21) After lab, an instructor will take your gels and photograph them under UV and produce a digital photo of your gel

We will summarize the results across the entire class and (fingers crossed) prepare to send our PCR product to the Advanced Genetic Technology Center (AGTC) on campus for sequencing on the Illumina MiSeq.

Week 5 - Intro to sequence data and command line analysis (Bioinformatics, Part 1)

1) We'll start with an introduction to the data. Look for the raw sequencing data in the **16SrRNA_data** folder on the **Desktop**. This is the entire set of sequence data generated from our sequencing run. Identify the set of sequences corresponding to your particular group.

2) You will see that each set of sequences is coded as a **.gz** file. This is a compressed format. We will need to unpack the sequence file from its compressed format. You could just double click on the file and have your computer do this in an automated way, but we're going to learn a better way to do this using your terminal and command lines.

3) Now, go to the dock on your computer and locate the **Terminal** app. Open it up. It will open a window that has a command prompt. This typically ends in a **$** sign and has a flashing cursor. The command prompt is a powerful thing, and gives you the ability to do way more with a computer than what you normally do with a point and click mouse and a GUI (graphical user interface) interface.

4) We will start with a simple step to check our software. In your terminal prompt type **gcc --version** and hit **enter**. This checks to see if we have command line tools installed. This should return approximately four lines of information. If it doesn't, let us know.

5) Next, we will start learning how to navigate the computer via Terminal. To start, type the command **ls** and press **enter**. This is a list command and will show you all of the files in your current directory. When you open your terminal, the default starting point is your home directory, so you should be seeing the files in your home directory, including the **Desktop** folder (also called a directory).

6) We want to move into the **Desktop** directory. To do this type the command **cd Desktop**. This is a change directory command (hence, **cd**).

7) Use the **ls** command again, and you should see everything on the **Desktop**.

8) Change directories again to move into the **16SrRNA_data** directory (**cd 16SrRNA_data**). Then use another **ls** command to list everything in the directory.

9) You should see a list of compressed data files, each in a **.gz** format. Identify the files associated with your group. You'll see that the each file begins with a group number designation along with a abbreviation of the building and sample number (e.g., G1_JSBN_S1).

10) You can also identify an **Index#** in each file, which refers to a index combination that were in the PCR primers for a particular sample.

Index	Sample description
Index1	JSB north elevator, sample 1
Index2	JSB north elevator, sample 2
Index3	JSB north elevator, sample 3
Index4	JSB north elevator, sample 4
Index5	JSB north elevator, sample 5
Index6	JSB north elevator, sample 6
Index7	JSB north elevator, sample 7
Index8	JSB north elevator, sample 8
Index9	JSB north elevator, sample 9
Index10	JSB north elevator, PCR negative
Index11	JSB service elevator, sample 1
Index12	JSB service elevator, sample 2
Index13	JSB service elevator, sample 3
Index14	JSB service elevator, sample 4
Index15	JSB service elevator, sample 5
Index16	JSB service elevator, sample 6
Index17	JSB service elevator, sample 7
Index18	JSB service elevator, sample 8
Index19	JSB service elevator, PCR negative
Index20	POT elevator A, sample 1
Index21	POT elevator A, sample 2
Index22	POT elevator A, sample 3
Index23	POT elevator A, sample 4
Index24	POT elevator A, sample 5
Index25	POT elevator A, sample 6
Index26	POT elevator A, sample 7
Index27	POT elevator A, sample 8
Index28	POT elevator A, PCR negative
Index29	POT elevator D, sample 1
Index30	POT elevator D, sample 2
Index31	POT elevator D, sample 3
Index32	POT elevator D, sample 4
Index33	POT elevator D, sample 5
Index34	POT elevator D, sample 6
Index35	POT elevator D, sample 7
Index36	POT elevator D, sample 8
Index37	POT elevator D, PCR negative
Index38	THM north elevator D, sample 1
Index39	THM north elevator D, sample 2
Index40	THM north elevator D, sample 3
Index41	THM north elevator D, sample 4
Index42	THM north elevator D, sample 5
Index43	THM north elevator D, sample 6
Index44	THM north elevator D, PCR negative
Index45	JSB service elevator, sample 1, PCR redo
Index46	THM north elevator D, sample 3, PCR redo
Index47	THM north elevator D, sample 5, PCR redo
Index48	Mystery sample

11) Now we will unpack some of the **.gz** files. Each group should be able to identify the set of samples that they worked on. From your group samples, choose to work on a particular sample. Note that some samples did not yield much sequence data. Choose to work with one of the larger files, and not one with a file size <10 KB. To see the file sizes in your terminal, use the command **ls -lh** to provide a detailed list of what is inside your current directory.

12) Note that each surface sample has two different **.gz** files, one with an **R1** in the file name, the other with an **R2** in the file name. These correspond to our paired-end sequences. Read 1 is sequence that starts from one end of our PCR fragment and Read 2 is sequence that starts from the other end. Each 250 bp sequence was generated on an Illumina MiSeq sequencer.

13) When you have identified your particular sequence set to work with, give the following command in your terminal: **gunzip yoursamplename_R1.fastq.gz -k** This will unpack the Read 1 sequences from your particular sequences and produce a text file with a similar filename (without the **.gz** extension). The **-k** flag tells the program to keep a copy of the original gzip file.

14) Use the **ls** command to see the unpacked sequence. It should be similar to the .gz file, except it won't have that extension. Rather, its extension will now just be .**fastq**.

15) Now do the same for your Read 2 sequences: **gunzip yoursamplename_R2.fastq.gz -k**

16) Using your Finder (not your terminal), open the **16SrRNA_data** folder on your Desktop. From your dock (at the bottom of your screen), open TextWrangler. Now, drag your recently created R1 fastq file into the TextWrangler window. This will open your file.

17) Let's take some time to look at one of our raw data files. It may look like a bunch of cryptic information, but we can make sense of it. Sequences are grouped in blocks of 4 lines. The first line of each sequence looks something like this:

- **@** - Each sequence identifier line starts with @.
- **InstrumentID** - unique identifier of the sequencer (M00329)
- **RunNumber** - Run number on instrument (129).
- **Flowcell_ID** - ID of flowcell (000000000-ABTD7).
- **Lane Number** - positive integer, currently 1-8 (1)
- **Tile Number** - positive integer (1)
- **X** - x coordinate of the spot. Integer which can be negative (14930)
- **Y** - y coordinate of the spot. Integer which can be negative (1700)
- **Read Number** - 1 for single reads; 1 or 2 for paired ends (1)
- **Whether it is filtered** - NB: Y if the read is filtered out, not in the fastq file, N otherwise (N)
- **Control number** - 0 when none of the control bits are on, otherwise it is an even number (0)
- **Index Combination** - The barcode sequences from that particular sequence.

18) The second line contains the actual sequence data generated for that read.

19) The third line contains a + sign, indicating an upcoming quality score line.

20) The fourth line contains the quality score for each corresponding base call. You'll notice that these are not actually numbers, but are instead letters and characters. This is a space issue here, as it's easier for the Sequencer to note and store a single ASCII character, than it is a larger number. Here's what the translation might look like:

char	value	Q	Error Prob.
!	33	0	1.000000e+00
"	34	1	7.943282e-01
#	35	2	6.309573e-01
$	36	3	5.011872e-01
%	37	4	3.981072e-01
&	38	5	3.162278e-01
'	39	6	2.511886e-01
(	40	7	1.995262e-01
)	41	8	1.584893e-01
*	42	9	1.258925e-01
+	43	10	1.000000e-01
,	44	11	7.943282e-02
-	45	12	6.309573e-02
.	46	13	5.011872e-02
/	47	14	3.981072e-02
0	48	15	3.162278e-02
1	49	16	2.511886e-02
2	50	17	1.995262e-02
3	51	18	1.584893e-02
4	52	19	1.258925e-02
5	53	20	1.000000e-02
6	54	21	7.943282e-03
7	55	22	6.309573e-03
8	56	23	5.011872e-03
9	57	24	3.981072e-03
:	58	25	3.162278e-03
;	59	26	2.511886e-03
<	60	27	1.995262e-03
=	61	28	1.584893e-03
>	62	29	1.258925e-03
?	63	30	1.000000e-03
@	64	31	7.943282e-04
A	65	32	6.309573e-04
B	66	33	5.011872e-04
C	67	34	3.981072e-04
D	68	35	3.162278e-04

E	69	36	2.511886e-04
F	70	37	1.995262e-04
G	71	38	1.584893e-04
H	72	39	1.258925e-04
I	73	40	1.000000e-04

21) Next let's consider our paired-end sequence data. Read 1 and Read 2 are both about 250 bp in length, and our 16S rRNA PCR fragment is only expected to be somewhere in the 250-300 bp size range. This means that our R1 and R2 sequence reads probably substantially overlap. We can use some software to put them together and create a consensus sequence. We'll do this with the program FLASH (Fast Length Adjustment of SHort Reads).

22) Do a Google search for **FLASH read merger**. Your top hit should be the program website run through Johns Hopkins University. Go to the **FLASH** website and click the **sourceforge** link near the bottom of the page. Next click on the **Download FLASH-1.2.11.tar.gz** link to download the current version of FLASH.

23) Locate the downloaded **FLASH-1.2.11.tar.gz** file in your **Downloads** folder and move it to your **Desktop**.

24) Open your terminal window and move to the Desktop directory. If you are still in the **16SrRNA_data** directory use the command **cd ..** to move back one level to the **Desktop**. If you are unsure of your current directory, you can always use the command **pwd** (print working directory) to identify your current location in directory space.

25) Use an **ls** command to verify the **FLASH-1.2.11.tar.gz** file is in your Desktop directory.

26) Now we'll unpack the **FLASH-1.2.11.tar.gz** file. This one is a little more complex than the .gz files we unpacked earlier. It's referred to as a Tar file, and we use the following command: **tar -zxvf FLASH-1.2.11.tar.gz**

27) This creates a new **FLASH-1.2.11** directory in your Desktop. You can migrate into this directory using the **cd FLASH-1.2.11** command.

28) Use the **ls** command to take a look inside the directory you just unpacked. You'll see a number of different files. Many of these have a **.c** or **.h** extension. These are the actual source code for FLASH, but in their current format they don't function as an actual executable program (we often refer to software as an executable). Other files give us some information about using FLASH.

29) Use the command **cat README** to look inside the README file. The cat command is a quick way to look at the text in a file. The README file gives us a bunch of basic information about FLASH. Notice that it tells us to use the command **make** to compile FLASH. Many programs distributed as source code contain a make file that tells the computer how to compile the software.

30) Use the command **make** to compile FLASH.

31) In your terminal, use the **ls -lh** command to see the file list in the FLASH directory, along with the dates when files were created (among other things). You should see a file simply labeled **flash**. This is the compiled executable. Use the command **./flash --help** to see a list of options to use with FLASH. We'll take a look at a few of the options here.

- First, take a look at the **-m** flag. When we use different settings in an executable, we refer to these as flags, which are typically preceded by a dash. The **-m** flag sets the minimum overlap for our R1 and R2 reads. The default is 10 bp, meaning that if the 2 reads do not overlap by at least 10 bp, they will not be merged. We'll leave this at the default setting.

- Next, look at the **-M** flag. This set the maximum overlap between reads. Any reads that have greater overlap than this setting will not be merged. The default is 65 bp, which is too low for us given that we have 250bp reads covering an ~250 bp fragment. We'll use **-M 250**.

- And now look at the **-o** flag. This sets the prefix for your output files.

- There are a number of other flags to consider, but we'll be good using the default settings for the rest of the parameters of the analysis.

32) Before we run FLASH on your data, let's first move your specific data files (the ones you unpacked that ends in **.fastq**) to your FLASH-1.2.11 directory. To do this, you need to be in your **16SrRNA_data** folder. Navigate there by typing **cd /Users/username/Desktop/16SrRNA_data** and hit enter. Note that this uses a path to guide you to the **16SrRNA_data** folder.

33) Now let's copy the files you will use over to the FLASH directory. This is a good demonstration on how to use the copy (**cp**) command. It will be something like this...

cp yourfilename_R1.fastq yourfilename_R2.fastq /Users/username/Desktop/FLASH.1.2.11

34) Next, let's run FLASH on your data. First, we need to move back to our FLASH-1.2.11 directory. To get there, type **cd /Users/username/Desktop/FLASH-1.2.11**. Now, use the following command. Note: This command is all one line separated by spaces and 'yourfilename' is only being used as a template!

./flash -M 250 yourfilename_R1.fastq yourfilename_R2.fastq -o yourfilename

35) Assuming that flash worked for you (you should see screen output indicating it worked), use an **ls** command to look at the new files that have been created.

36) Use **cat Your_prefix.hist** to see a summary of the merged reads by length.

37) Next, look at the plot for these values using **cat Your_prefix.histogram**.

38) Now that we have merged reads for your data, let's take a little time to explore how we might learn about their prokaryotic source. In TextWrangler, open your file that ends in ***.extendedFrags.fastq**.

39) Open a web browser and go to the following website: **http://www.ncbi.nlm.nih.gov/** This is the NCBI homepage, a clearinghouse for genetic information and resources.

40) Now click on the **DNA & RNA** link. This will list a variety of genetic databases.

41) Next click on the **Genbank** link. This is a database containing all public DNA sequence data. Most journals require you to publish your DNA sequence data prior to publication. It usually is submitted here.

42) Now click on the **Search Genbank** link. We're going to explore some of our sequence data.

43) Click on the **BLAST** link. Then, click on the **nucleotide blast** link.

44) OK, here we are, ready to go. Let's BLAST!

Week 6 – Bioinformatics, Part 2: QIIME.

Installing MacQIIME

1) Locate the file labeled **MacQIIME_1.9.1-20150604_OS10.7.tgz** on the desktop.

2) Open your terminal and change directory (**cd**) to the Desktop. Use an **ls** command to make sure the MacQIIME tar file is available on the desktop.

3) Use the command **tar -xvf MacQIIME_1.9.1-20150604_OS10.7.tgz** to unpack the tar file. This may take a few minutes as it unpacks all of the files. When your command prompt returns, use an **ls** command to make sure there is a new **MacQIIME_1.9.1-20150604_OS10.7** folder on your desktop.

4) In your terminal, change directory into the **MacQIIME_1.9.1-20150604_OS10.7** directory. Use an **ls** command to make sure you have an **install.s** file in this folder. This is similar to a make file and will perform the compiling of MacQIIME.

5) Next, use the command **./install.s** to install a copy of MacQiime on your computer. This installs MacQIIME in the root directory of the computer, meaning we can use it from within any folder that the terminal is in.

6) Use the command **macqiime** to source it, making it functional and ready to be used. If it installed properly, a number of lines should output to the screen and your user prompt will now start with the term "MacQIIME."

7) As a further test, use the command **print_qiime_config.py -t** to see if a bunch of information is output that tells you about the various programs and configurations that make up QIIME.

Converting fastq to fasta

Our fastq files contain lots of extra stuff in them that we no longer need, so we are going to convert them to a simpler file format termed fasta. We will use QIIME to do this for us with a **convert_fastaqual_fastq.py** command. Because we are doing this for multiple files, we can use a shell script to repeat this operation.

8) Using TextWrangler, open the **fastq_fasta_198.sh** file in the **files_for_week_9** folder. Notice that there are two lines in this file corresponding to two different fastq files. We will modify this .sh file to account for the five fastq data files that we'll process today. These are:

G1_JSBN_S1.final.fastq
G2_JSBS_S1.final.fastq
G3_POTA_S1.final.fastq
G4_POTD_S1.final.fastq
G5_THM_S1.final.fastq

- These are the sequence files that have resulted from the FLASH read merging work we did last week. In addition, we used the software package FASTX-Toolkit to filter sequences by quality. In each set of sequences, this removed approximately 25-30 sequences that were of generally low quality.

9) Place a copy of your newly modified **fastq_fasta_198.sh** file into the **week_9_data** folder located on the Desktop where the five fastq files are located.

10) Using your terminal, use a change directory to move into the **week_9_data** folder.

11) Check to make sure that you have sourced **macqiime**. Your command prompt should begin with the term **MacQIIME**.

12) Use the command **chmod 777 fastq_fasta_198.sh**.
    - This is necessary to provide the correct set of permissions to this newly created file. Without doing this your computer will likely not have permission to run the script.

13) Now use the command **./fastq_fasta_198.sh** to execute the script that you have made.
    - This will create a new folder labeled fasta_files on your Desktop that you specified in your script.
    - It'll probably take about 5-6 minutes to complete this job.

Building a QIIME mapping file

14) Next, find the **rRNA_2016_Mapping.txt** file in the **files_for_week_9** folder on the Desktop and open this in TextWrangler.
    - This file links the individual fasta files to their particular treatment variables.
    - It can also contain information for the barcode and Primer sequences, but ours have been removed from the sequences, so these fields are left empty.

15) Note that the underscores in the SampleID names for our fasta files have been replaced with periods. This is because QIIME doesn't recognize the underscore character. Change the file names of the newly created fasta files in the **fasta_files** folder to account for this.

16) Move the **rRNA_2016_Mapping.txt** file to the **Desktop**.

17) In your terminal, change directory to the **fasta_files** folder.

18) In your terminal, use the following long command:

    validate_mapping_file.py -m /Users/homedirectory/Desktop/rRNA_2016_Mapping.txt -o /Users/homedirectory/Desktop/check_id_output -p -b

    - This will check your mapping file to make sure that it matches your data files and is ready to be used in further analysis.

19) A new folder will be created labeled **check_id_output**. Open this in your finder and open the **rRNA_2016_Mapping.log** file in TextWrangler or click on the **.html** file and open in a web browser.
    - QIIME probably said there were errors in the mapping file when used the validation command, but your file should not list any. If you think there are errors, let us know.

20) Return to the terminal and give the following long command:

    add_qiime_labels.py -m /Users/homedirectory/Desktop/rRNA_2016_Mapping.txt -i /Users/homedirectory/Desktop/week_9_data/fasta_files/ -c SampleID -o /Users/homedirectory/Desktop/QIIME_Labeled

    - This adds a new folder to your **Desktop** labeled **QIIME_Labeled** and then creates a new sequence file labeled **combined_seqs.fna** that combines all of your individual fasta files, and adds QIIME-specific labeling to each sequence to keep track of which

sequences belong to which sample. If you would like, open this in TextWrangler to look at the full data set.

Assigning your sequences to Operational Taxonomic Units (OTUs)

21) Next, we will use the following command to go through the process of OTU assignment:

pick_de_novo_otus.py -i /Users/homedirectory/Desktop/QIIME_Labeled/combined_seqs.fna -o /Users/homedirectory/Desktop/QIIME_OTUs

- This may take a few minutes. We've got lots of data.
- This command makes QIIME go through the entire sequence file, which contains over 60,000 sequences, and identify which sequences have high sequence similarity. The default setting here is to lump all sequences with at least 97% sequence similarity into a single representative sequence.
- Here, an OTU is a proxy for species, assuming that all species are divergent in their sequence by at least 3%. It's not a perfect metric, but it'll work for us here.

Summarize OTUs

22) Use this long command to summarize all of our OTU information into a simplified table:

biom summarize-table -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -o /Users/homedirectory/Desktop/QIIME_OTUs/table_summary.txt

- This doesn't do a whole heck of a lot, but it's one of the steps we need to take in the process of using QIIME to summarize our data.

Generate heatmap of OTUs for each sample

23) Here, we'll have QIIME generate a heatmap illustrating the diversity and abundance of different prokaryotic OTUs across our give samples. Use the following command:

make_otu_heatmap.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -o /Users/homedirectory/Desktop/QIIME_OTUs/heatmap.pdf

- This will make a PDF file of the heatmap inside the **QIIME_OTUs** folder. Open it up.

Build summary plots of taxa (OTUs)

24) Use this command to generate plots of taxonomic diversity for our samples at different hierarchical levels (e.g., class, order, family, etc.):

summarize_taxa_through_plots.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -o /Users/homedirectory/Desktop/QIIME_OTUs/taxa_summary -m /Users/homedirectory/Desktop/rRNA_2016_Mapping.txt

- Look inside the **taxa_summary** folder and the the **taxa_summary_plots** folder to find two interactive files labeled **area_charts.html** and **bar_charts.html**. You can open these in Firefox to browse a summary of results across samples.
- Do any of the samples look like they are different from the others?

Alpha diversity and rarefaction

25) Here, we'll calculate measures of prokaryotic diversity for our samples and treatments. In addition, we'll perform some analyses to ask if we have sampled enough to accurately capture the diversity within our samples. Use the following command:

alpha_rarefaction.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -m /Users/homedirectory/Desktop/rRNA_2016_Mapping.txt -o /Users/homedirectory/Desktop/QIIME_OTUs/arare/ -t /Users/homedirectory/Desktop/QIIME_OTUs/rep_set.tre

- This operation will take a little time to complete. Maybe five minutes or so.
- When completed, open the arare folder and we'll take a look at some of the results.
- **What is Alpha diversity? What is a rarefaction plot?**

SUGGESTED ANSWER: Alpha diversity is the taxonomic diversity of each site (or sample). A rarefaction plot in this context refers to a subsampling plot where a number of sequences (e.g., 1000) is subsampled and biodiversity is calculated. This is done iteratively (at least 10x or so) and for different levels of sequencing (1000, 2000, 3000, etc.). The final plot shows how much diversity is discovered at each level of sequencing and provides a way to assess how appropriate the amount of sequencing was for each sample.

Beta diversity and differentiation among samples

26) Now we'll assess a different measure of diversity, beta diversity. **What is this?** Use the following command:

SUGGESTED ANSWER: Beta diversity is the level of dissimilarity among samples or treatments. Essentially, how similar is sample A to samples B, C, etc.

beta_diversity_through_plots.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -m /Users/homedirectory/Desktop/rRNA_2016_Mapping.txt -o /Users/homedirectory/Desktop/QIIME_OTUs/betadiv_even/ -t /Users/homedirectory/Desktop/QIIME_OTUs/rep_set.tre -e 8000 -f

We'll take a look at the results from these analyses and discuss.

Week 7 - Bioinformatics, Part 3: Quantitative analysis and revisiting our prokaryotic diversity hypotheses

This week we will use our larger sequence data set to measure prokaryotic diversity and statistically test hypotheses about how this diversity varies across our sampled environments. We will do this in QIIME. This will be a tractable exercise in our two-hour lab because the very time-consuming steps in the analysis (e.g., the pick_de_novo_otus.py step) have already been performed for you.

One thing to note going into these analyses: some samples have been removed from the analysis due to low sequence output. This is particularly true for a handful of samples (8) for which we recovered < 40 sequences. To provide somewhat reasonable estimates of prokaryotic diversity, we removed any sample with < 5000 sequences (merged reads). Some of these were PCR negative controls. In total, we will still analyze a collection of 32 different sequenced samples. A list of all samples, the number of merged sequences per sample, and which ones were left out of analysis can be found in the Google Drive link that has been emailed to everyone.

1) On your Desktop, check for the **16SrRNA_data_week_10 folder** and look inside to make sure it contains a **fasta_files** folder with 48 (**.fna**) files.

2) Now check for the **QIIME_Labeled** folder on your Desktop. We have performed this step in the analysis in preparation for today's lab, which is saving us computation time.
   - This contains the combined sequences in fasta format from all of the samples to be analyzed.

3) And, check for the **QIIME_OTUs** folder on your Desktop. Here, we have already performed the following steps:
      - pick_de_novo_OTUs.py
      - biom summaize-table
      - make_otu_heatmap.py

   - Remember, sequences that were < 3% divergent from each other were binned together into the same "species" or OTU category. This step also made a record of the number of sequences identified for each OTU, which provides a measure of abundance, or frequency, of that particular OTU.
   - You may find the heatmap interesting, or not.

4) Next, locate on your Desktop the **Full_data_processing_files** folder, which contains a number of scripts and analysis files used in the analysis of the complete data set.
   - Note that some of these have already been run on the data (e.g., read merging and quality filtering scripts). We provide them here so you have access to everything that has been done with the data.

5) Move the **rRNA_2016_48_Mapping.txt** file to your Desktop.

6) Open the **rRNA_2016_48_Mapping.txt** in Textwrangler. You should see lines for the 32 input files that we are including in our analysis. Note that we currently have only one **Treatment** category defined here, specifying if the building sampled was a tall (**high**) or short (**low**) building.

7) First, lets generate some histograms of prokaryotic diversity for the different samples using the following command:

summarize_taxa_through_plots.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -o /Users/homedirectory/Desktop/QIIME_OTUs/taxa_summary_sample -m /Users/homedirectory/Desktop/rRNA_2016_48_Mapping.txt

- Note that a copy of this command is located in the **Long_commands_week10.txt** file found inside the **Full_data_processing_files** folder.

8) Locate these results in the **QIIME_OTUs/taxa_summary_sample/taxa_summary_plots** folder and browse your results in the **bar_charts.html** file. This should open in your web browser.

9) Now lets generate the same histograms summarized by our first treatment using the following command:

summarize_taxa_through_plots.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -o /Users/homedirectory/Desktop/QIIME_OTUs/taxa_summary_treatment -m /Users/homedirectory/Desktop/rRNA_2016_48_Mapping.txt -c Treat1 -f

10) Find your results and see if there appear to be any differences.

11) Next, lets calculate measures of alpha diversity for samples and treatment, and perform rarefaction analysis of the data using the following command:

alpha_rarefaction.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -m /Users/ homedirectory /Desktop/rRNA_2016_48_Mapping.txt -o /Users/ homedirectory /Desktop/QIIME_OTUs/arare/ -t /Users/ homedirectory /Desktop/QIIME_OTUs/rep_set.tre -f

12) Locate the results of this analysis in the **QIIME_OTUs/arare/alpha_rarefaction_plots** folder.

13) This analysis will provide three measures of alpha diversity. What are they and how do they differ in how they calculate alpha diversity? You may need to do some research on these to understand what each measure is telling you.

SUGGESTED ANSWER: This answer will depend on the metric being calculated. Some common defaults are. Shannon = Shannon's diversity index, which is a quantitative measure that reflects how many different types (species or otherwise) are in a sample. Observed OTUs is simply the total number of OTUs observed, purely from a count perspective. Documentation of the various metrics that can be used can be seen here: http://scikit-bio.org/docs/latest/generated/skbio.diversity.alpha.html

14) What can you say about alpha diversity for the various samples? What about diversity when comparing tall vs. short buildings?

15) Based on the rarefaction plots for various measures of alpha diversity, what can you say about the sufficiency of our data to estimate alpha diversity for our samples or treatments?
    - Note that you can save a copy of the **arare** folder and its contents for later reference. A flash drive may work best for this step. You'll need to copy the entire folder to be able to access the results on another computer.

16) OK, now calculate measures of beta diversity and a principal component plot for your samples and treatment using the following command:

    beta_diversity_through_plots.py -i /Users/homedirectory/Desktop/QIIME_OTUs/otu_table.biom -m /Users/homedirectory/Desktop/rRNA_2016_48_Mapping.txt -o /Users/homedirectory/Desktop/QIIME_OTUs/betadiv_even/ -t /Users/homedirectory/Desktop/QIIME_OTUs/rep_set.tre -f

    - You will probably receive a minor warning about a negative eigenvalue, which is a statistic used in the principal component analysis. It's OK.

17) Open the **unweighted_unifrac_dm.txt** file located in the **QIIME_OTUs/betadiv_even** folder. You'll want to use Textwrangler for this.
    - Here you are seeing a matrix of distance measures between samples (i.e., beta diversity), measured using a calculation termed UniFrac, which takes into account phylogenetic distance between samples. What is the difference between a weighted and unweighted measure? Is there a good reason to use, or not to use, the weighted measure of beta diversity for our study?

SUGGESTED ANSWER: Weighted versus unweighted distance measures refer to whether OTU abundances are accounted for (weighted) or ignored (unweighted) when comparing distances among samples. Therefore, if the goal is simply to assess what taxa are present, unweighted distance should be used.

18) It's not easy to interpret all of those pairwise comparisons of beta diversity between samples. There are just too many combinations to identify a clear pattern. So, instead, you've also performed a principal component analysis (PCA) on the data. Open the **index.html** file in the

QIIME_OTUs/betadiv_even folder. You'll need to open this in a web browser to visualize the PCA plot.

- PCA is a way of taking complex data (like the beta diversity matrix) and condensing it down to a smaller number of dimensions (PC axes) that explain all of that variation. It provides a much more visually easy way to look for patterns in your results.
- Here's a cool interactive website that provides an explanation of PCA:
  http://setosa.io/ev/principal-component-analysis/
- Do you see any beta diversity patterns in that would suggest differences among samples or treatment?

19) Now lets ask if we have a statistical difference between our two treatment groups. We'll do this by testing for a significant difference in beta diversity. Use the following command:

compare_categories.py --method anosim -i /Users/homedirectory/Desktop/QIIME_OTUs/betadiv_even/unweighted_unifrac_dm.txt -m /Users/homedirectory/Desktop/rRNA_2016_48_Mapping.txt -c Treat1 -o /Users/homedirectory/Desktop/QIIME_OTUs/anosim_out -n 99

20) Locate the results file in the **QIIME_OTUs/anosim_out** folder. Do the results support the hypothesis that taller buildings on campus have greater prokaryotic diversity compared to shorter buildings?

21) OK, so you've run through the analysis of the full data set and tested the main hypothesis that was proposed for this project. What other hypotheses can be tested with our sampling design and data? Other differences exist among the buildings and elevators we sampled, and we collected a range of samples within elevators.

For the remainder of the lab, you will work in your groups to develop additional hypotheses to be tested. You will then address these hypotheses using your knowledge of how to calculate measures of diversity and perform a statistical test.

This will require you to modify the **rRNA_2016_48_Mapping.txt** in Textwrangler to add in the new treatments. We will be available to help you with this and deal with any errors that may arise.

Supplementary Materials – Survey Questions.

Course survey – A pre-course survey was distributed to gauge student interest, understanding, and prior knowledge of various topics. We performed a follow-up, post-course survey at the end of the semester in order to gauge how the topics discussed in the course addressed knowledge gaps and peaked student interest.

The survey was developed and administered using the online tool Survey Monkey (www.surveymonkey.com) and was completely anonymous. Questions marked with an asterisk were additions for the 2016 CURE.

Pre-course Survey: Questions from the 2016 iteration of the course.

Please use the rating scale below to answer the following questions:

Rating scale: (a) None, (b) little experience/comfort, (c) intermediate experience/Some comfort, or (d) highly experienced/comfort

1. Please rate your previous experience in biology classes with the following subjects:
    a. Scientific method (e.g., deductive logic)
    b. Hypothesis generation
    c. Experimental design
    d. Use of molecular lab equipment
    e. Data analysis
    f. Problem solving
    g. Communicating science (via oral or written reports)
    h. Reading peer review literature

2. For any topics that you rated either intermediate experience or highly experienced please provide a specific example.*

3. Please rate your general knowledge of the following topics
    a. PCR
    b. Genetic analysis
    c. Statistics
    d. Primary literature
    e. Bioinformatic analysis

4. Please rate your comfort level with the following areas (e.g., how comfortable would you be with conducting the following on your own)
    a. Generating a hypothesis from a given observation
    b. Designing an experiment to test a hypothesis
    c. Analyzing data for patterns
    d. Laboratory techniques
    e. Figuring out the next step in an experiment
    f. Ability to work independently
    g. Ability to work with others
    h. Communicating my work
    i. Time management
    j. Understanding peer review literature

5. What about biology excites you?*

6. Please describe what a biologist is to you. What type of skills would you expect them to have? *

7. Please briefly describe your long-term career goals

8. Please briefly describe what you hope to gain from this course

Post-course Survey:

Please use the rating scale below to answer the following questions:

Rating scale: (a) No improvement, (b) improved slightly, (c) moderately improved, or (d) greatly improved

1. Please rate how the overall content and experience in the course has influenced your knowledge of the following topics:
    a. Scientific method (e.g., deductive logic)
    b. Hypothesis generation
    c. Experimental design
    d. Use of molecular lab equipment
    e. Data analysis
    f. Problem solving
    g. Communicating science (via oral or written reports)
    h. Reading peer review literature

2. Please rate how the 16s rRNA project contributed in your knowledge in the following topics:
    a. Scientific method (e.g., deductive logic)
    b. Hypothesis generation
    c. Experimental design
    d. Use of molecular lab equipment
    e. Data analysis
    f. Problem solving
    g. Communicating science (via oral or written reports)
    h. Reading peer review literature

3. Please provide a specific example of a topic that you moderately or greatly improved on*

4. Please rate how this course has influenced your general knowledge of the following topics:
    a. PCR
    b. Genetic analysis
    c. Statistics
    d. Primary literature
    e. Bioinformatic analysis

5. Please rate how the overall content and experience in this course has changed your comfort level (e.g., how comfortable would you be conducting the following on your way) with the following topics:
    a. Generating a hypothesis from a given observation
    b. Designing an experiment to test a hypothesis
    c. Analyzing data for patterns
    d. Laboratory techniques

e. Figuring out the next steps in an experiment
f. Ability to work independently
g. Ability to work with others
h. Communicating my work
i. Time management
j. Understanding peer-review literature

6. Please rate how the 16s rRNA project influenced your comfort level with the following areas:
    a. Generating a hypothesis from a given observation
    b. Designing an experiment to test a hypothesis
    c. Analyzing data for patterns
    d. Laboratory techniques
    e. Figuring out the next steps in an experiment
    f. Ability to work independently
    g. Ability to work with others
    h. Communicating my work
    i. Time management
    j. Understanding peer-review literature

7. Regarding question 6, please provide a specific example of a topic that you moderately or greatly improved on*

8. Please briefly describe how this course has or hasn't influenced your view of what a biologist does*

9. Please provide any feedback about the 16s RNA project