Chapter 1: Theme1 GRDI Ecobiomics, Setting Up Packages and Working Directory structures

Emily Giroux

4/02/2019

Dada2: Divisive Amplicon Denoising Algorithm
This script follows closely the tutorial provided by Benjamin J Callahan et al, 2017. *Workflow for Microbiome Data Analysis: from raw reads to community analyses.*
https://bioconductor.org/help/course-materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.html#assign_taxonomy

Using package `BiocManager` to install required packages:

```
# Installing required packages
r <- getOption("repos")
r["CRAN"] <- "http://cran.us.r-project.org"
options(repos = r)

if (!requireNamespace("BiocManager")) install.packages("BiocManager")
BiocManager::install()

library("BiocManager")
.cran_packages <- c("data.table", "kableExtra", "knitr", "rprojroot")
.inst <- .cran_packages %in% installed.packages()
if (any(!.inst)) {
    install.packages(.cran_packages[!.inst])
}
```

Load packages into session, and print package versions:

```
sapply(c(.cran_packages), require, character.only = TRUE)
## data.table kableExtra      knitr  rprojroot
##       TRUE       TRUE       TRUE       TRUE
```

Source our custom R scripts:
For this we will use the rprojroot package to set the directory structures. This will help us when finding our files to source functions. We specify ours is an RStudio project. The root object contains a function that will help us locate our package R files regarless of our current working directory.

```
library("rprojroot")
root        <- rprojroot::is_rstudio_project
scriptsPath <- root$make_fix_file(".")("R")
scripts     <- dir(root$find_file("R", path = root$find_file()))
scriptsl    <- paste(scriptsPath, scripts, sep = "/")
lapply(scriptsl, source)
```

The DADA2 tutorial website contains formatted training fastas for the RDP training set, GreenGenes clustered at 97% identity, and the Silva reference database available. For fungal taxonomy, the General Fasta release files from the UNITE ITS database can be used as is. Download the database files and place them in a designated databases directory.

Obtain databases:
Complete the following on the biocluster. Create a designated directory for DADA2 databases and download the SILVA, UNITE and RDP databases into it.

```
mkdir ~/Databases/dada2DBs
cd ~/Databases/dada2DBs
```

SILVA: https://zenodo.org/record/1172783#.XJOpWiJKhhE

```
wget https://zenodo.org/record/1172783/files/SILVA_LICENSE
wget https://zenodo.org/record/1172783/files/silva_nr_v132_train_set.fa.gz
wget https://zenodo.org/record/1172783/files/silva_species_assignment_v132.fa.gz
```

RDP taxonomic classifier: https://zenodo.org/record/801828#.XJOrWiJKhhE

```
wget https://zenodo.org/record/801828/files/rdp_species_assignment_16.fa.gz
wget https://zenodo.org/record/801828/files/rdp_train_set_16.fa.gz
```

Greengenes: https://zenodo.org/record/158955#.XJOr8CJKhhE

```
wget https://zenodo.org/record/158955/files/gg_13_8_train_set_97.fa.gz
wget https://zenodo.org/record/158955/files/rdp_species_assignment_14.fa.gz
wget https://zenodo.org/record/158955/files/rdp_train_set_14.fa.gz
```

UNITE: General Fasta releases (DADA2 package version 1.3.3 or later) https://unite.ut.ee/repository.php

```
wget https://files.plutof.ut.ee/public/orig/EB/0C/
    EB0CCB3A871B77EA75E472D13926271076904A588D2E1C1EA5AFCF7397D48378.zip
mv EB0CCB3A871B77EA75E472D13926271076904A588D2E1C1EA5AFCF7397D48378.zip unite.zip
unzip unite.zip
```

Extensions: The DADA2 package also implements a method to make species level assignments based on exact matching between ASVs and sequenced reference strains. Recent analysis suggests that exact matching (or 100% identity) is the only appropriate way to assign species to 16S gene fragments. Currently, species-assignment training fastas are available for the Silva and RDP 16S databases. To follow the optional species addition step, download the silva_species_assignment_v128.fa.gz file into the databases directory as well.

Record paths to dadabases:
Note: The COI database was obtained from Terry Porter.

```
dbsPath   <- "/home/CFIA-ACIA/girouxeml/Databases"
dada2DBs  <- paste(dbsPath, "dada2DBs", sep = "/")

rdp16Sset <- "rdp_train_set_16.fa"
rdp16Sspp <- "rdp_species_assignment_16.fa"
silvaSet  <- "silva_nr_v132_train_set.fa"
silvaSpp  <- "silva_species_assignment_v132.fa"
```

```
gGenesSet <- "gg_13_8_train_set_97.fa"
uniteSet  <- "unite"
coiSetDir <- "CO1_v3.2"

rdp16SsetPath <- paste(dada2DBs, rdp16Sset, sep = "/")
rdp16SsppPath <- paste(dada2DBs, rdp16Sspp, sep = "/")
silvaSetPath  <- paste(dada2DBs, silvaSet, sep = "/")
silvaSppPath  <- paste(dada2DBs, silvaSpp, sep = "/")
gGenesSetPath <- paste(dada2DBs, gGenesSet, sep = "/")
uniteSetPath  <- paste(dada2DBs, uniteSet, sep = "/")
coiSetDirPath <- paste(dbsPath, coiSetDir, sep = "/")

# Note: unzip and then paste the path to the sh file:
system2(paste("unzip ", uniteSetPath, ".zip", sep = ""))
uniteSetSh <- paste(dada2DBs, "sh_general_release_dynamic_02.02.2019.fasta", sep = "/")
```

Note for ITS sequences:
In general you should not truncate the reads in ITS analysis, because there is usually no effective single truncation length due to the biological length variation in the ITS region. https://github.com/benjjneb/dada2/issues/609

Checking for Adapter Sequences
It may be a good idea to see if there are adapter sequences still on the raw reads. Below are the read structures for Illumina paired-end reads showing the portions that are adapter sequences:

5' AATGATACGGCGACCACCGAGATCTACAC TCTTTCCCTACACGACGCTCTTCCGATCT
(N)
AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC <- region to select as forward adapter
XXXXXX
ATCTCGTATGCCGTCTTCTGCTTG 3'

3' TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGA
(N)
TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG <- region to select as reverse adapter
XXXXXX
TAGAGCATACGGCAGAAGACGAAC 5'

Where each string of 'X' is the unique 4-, 6, or 8-base barcode from the L2 adaptor mix of the library construction system (where applicable) and (N) is the library insert.
Record the adapter sequences specific to the sequencing run type:

```
fwdAdapMiSeq <- "AGATCGGAAGAGCACAC"
revAdapMiSeq <- "AGATCGGAAGAGCGTCGT"
fwdAdap      <- fwdAdapMiSeq
revAdap      <- revAdapMiSeq
```

User:
Define the path to the shared folder where the main working directory will be.

```
sharedPath   <- "/isilon/cfia-ottawa-fallowfield/users/girouxeml/PIRL_working_directory"
analysis     <- "ecobiomics"
sharedPathAn <- paste(sharedPath, analysis, sep = "/")
```

Read in the most recently updated metadata table:
This table has the read pairs already collapsed to one row each.

```
library("data.table")
metadataName <- "ecobiomics_metadata_edited_Emily19Aug2019.csv"
metadataPath <- paste(sharedPathAn, metadataName, sep = "/")
metadata     <- data.table::fread(metadataPath, sep = "auto", header = TRUE)
```

Note:
I placed all the raw fastq files from the sequencing runs in a directory called "raw/illumina/NRC/" to keep the original saved location format on the GPSC. Also, I renamed the fastq.gz files so that all "-" were underscores instead "_", using perl rename, multiple times. I prefer to organise the processed data by amplicon region, so I created a directory in for each unique amplicon region recorded in the metadata table, and then moved the processed reads belonging to a region to its matching region directory in the shared analysis directory.

Split amplicon regions to separate metadata tables:
Here I use the data.table packages to split the table using the binary keys approach:
https://cran.r-project.org/web/packages/data.table/vignettes/datatable-keys-fast-subset.html

```
library("data.table")
data.table::setkey(metadata, Region)
unique(metadata$Region)
metadata16S <- metadata["16S"]
metadata18S <- metadata["18S"]
metadataCOI <- metadata["COI"]
metadataITS <- metadata["ITS"]
## [1] "16S" "18S" "COI" "ITS"
```

Save the image and load this at the beginning of each amplicon region processing and analysis workflow:

```
imageDirPath <- "/home/CFIA-ACIA/girouxeml/GitHub_Repos/r_environments/ecobiomics/"
if (!dir.exists(imageDirPath)) dir.create(imageDirPath)

# Specify an image name for this chapter:
startUpImage <- "ecobiomics_StartUp.RData"

# Save this chapter's image:
save.image(paste(imageDirPath, startUpImage, sep = ""))
```

To begin, load the image from this script prior to running the analysis steps. When re-starting a session, you can quickly load up the image by running the chunk below:

```
sharedPath <- "/isilon/cfia-ottawa-fallowfield/users/girouxeml/PIRL_working_directory"
analysis <- "ecobiomics"
sharedPathAn <- paste(sharedPath, analysis, sep = "/")
imageDirPath <- "/home/CFIA-ACIA/girouxeml/GitHub_Repos/r_environments/ecobiomics/"
startUpImage <- "ecobiomics_StartUp.RData"
load(paste(imageDirPath, startUpImage, sep = ""))
```

Analyses for each amplicon region will follow these steps:

1. Run the **ggPlotRaw** chunks.
   Check out the quality profile of the raw reads. Most Illumina sequencing data shows a trend of decreasing average quality towards the end of sequencing reads.

2. Run the **SeqPrep** or the **cutadapt** chunks to investigate possible adapter contamination issues.
   For SeqPrep when testing for adapters and performing adapter removal with optional merging:
   To test if the choice of adapters is good using the first fastq read 1 sequence. Ignore broken pipe error. This happens because when the stdin of "cat" is small it may finish writing *before* the exit of the reader, in our case "grep".

3. Run the **filterAndTrimming** chunk.
   Filtering, based on quality profile per region:
   Outside of filtering and trimming, there should be no major loss of reads. If any parameter needs time optimizing - it should be the filtering and trimming step.

4. Run the **ggPlotsProcessed** chunk. Look at the effects of trimming.

5. Run the **keepTrimmed** and the **keepTrimmed2** chunk. Update metadata so that samples that did not have any reads that passed filtering are removed from further analysis, to avoid downstream processing errors.
   Really helpful data.table page on keys and fast binary searches, especially for subsetting rows - here rows are subset based on a list.
   https://cran.r-project.org/web/packages/data.table/vignettes/datatable-keys-fast-subset.html
   Update metadata table so that rows that had read pairs where a direction no longer had reads after filtering are removed from the metadata table.

6. Run the **splitRunsMetadata** chunk. Split the samples by sequencing run, so that error can be calculated properly.

7. Run **errorLearningPool1** chunk:
   Run the error learning on the libraries. Split up sets run on different runs, then merge them back together later from the rdp files.
   The DADA2 method relies on a parameterized model of substitution errors to distinguish sequencing errors from real biological variation. Because error rates can (and often do) vary substantially between sequencing runs and PCR protocols, the model parameters can be discovered from the data itself using a form of unsupervised learning in which sample inference is alternated with parameter estimation until both are jointly consistent.
   Parameter learning is computationally intensive, as it requires multiple iterations of the sequence inference algorithm, and therefore it is often useful to estimate the error rates from a (sufficiently large) subset of the data. Be aware that error rates are being learned from a subset of the data. As a rule of thumb, a million 100nt reads (or 100M total bases) is more than adequate to learn the error rates.

8. Run the **drepDadaMergePool1** chunk. Run the sample inference and merger of paired-end reads. This runs 3 steps:
   i. *Depreplication*
   ii. *DADA2*
   iii. *Merging*

i. Dereplication:
Dereplication combines all identical sequencing reads into into "unique sequences" with a corresponding "abundance": the number of reads with that unique sequence. Dereplication substantially reduces computation time by eliminating redundant comparisons. The sequence data is imported into R from demultiplexed fastq files (i.e. one fastq for each sample) and simultaneously dereplicated to remove redundancy. We name the resulting derep-class objects by their sample name.

ii. Run the DADA2 algorithm to infer sequence variants:
After filtering, the typical amplicon bioinformatics workflow clusters sequencing reads into operational taxonomic units (OTUs): groups of sequencing reads that differ by less than a fixed

dissimilarity threshhold. Here we instead use the high-resolution DADA2 method to to infer *A*mplicon *S*equence *V*ariants (ASVs) exactly, without imposing any arbitrary threshhold, and thereby resolving variants that differ by as little as one nucleotide (Benjamin J Callahan et al. 2016).

The crucial difference between this workflow and the introductory workflow is that the samples are read in and processed in a streaming fashion (within a for-loop) during sample inference, so only one sample is fully loaded into memory at a time. This keeps memory requirements quite low: A Hiseq lane can be processed on 8GB of memory (although more is nice!).

The DADA2 sequence inference method can run in two different modes: Independent inference by sample (pool=FALSE), and inference from the pooled sequencing reads from all samples (pool=TRUE). Independent inference has the advantage that computation time is linear in the number of samples, and memory requirements are flat with the number of samples. This allows scaling out to datasets of almost unlimited size. Pooled inference is more computationally taxing, and can become intractable for datasets of tens of millions of reads. However, pooling improves the detection of rare variants that were seen just once or twice in an individual sample but many times across all samples. As of version 1.2, multithreading can now be activated with the arguments multithread = TRUE, which substantially speeds this step.

The DADA2 sequence inference step should remove (nearly) all substitution and indel errors from the data (Benjamin J Callahan et al. 2016).

iii. We now merge together the inferred forward and reverse sequences. We're also removing paired sequences that do not perfectly overlap as a final control against residual errors.

9. Run the **mergeSplitRuns** chunk. Merge split count matrixes back into R - if samples were split across different runs.

10. Run the **remChimeric** chunk. Remove chimeric sequences from the sequence table.
    The DADA2 method produces a sequence table that is a higher-resolution analogue of the common "OTU table", i.e. a sample by sequence feature table valued by the number of times each sequence was observed in each sample.
    Although exact numbers vary substantially by experimental condition, it is typical that chimeras comprise a substantial fraction of inferred sequence variants, but only a small fraction of all reads.

11. Run the **assignTax** chunk to assign taxonomy.
    One of the benefits of using well-classified marker loci like the 16S rRNA gene is the ability to taxonomically classify the sequence variants. The dada2 package implements the naive Bayesian classifier method for this purpose (Wang et al. 2007). This classifier compares sequence variants to a training set of classified sequences.

Important:
If any parameters are repeated and changed for trimming, error-learning, dereplication, DADA2 or merging, all steps after the altered chunk must be repeated and updated. Do not load a saved image that extends beyond these processes if an earlier has step has been altered. Any changes to trimming, or the steps just mentioned can dramatically alter all the results that follow.

Table 1: 16S Samples

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
16S_NA_GB_1A_1a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1A_1b_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1A_2a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1A_NS	16S	benthic	Gagne	Experimental	NucleoSpin
16S_NA_GB_1A_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_1B_1a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1B_1b_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1B_2a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1B_NS	16S	benthic	Gagne	Experimental	NucleoSpin
16S_NA_GB_1B_PS	16S	benthic	Gagne	Experimental	PowerSoil
16S_NA_GB_1B_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_1C_1a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1C_1b_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_1C_2a_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_2A_PS	16S	benthic	Gagne	Experimental	PowerSoil
16S_NA_GB_2A_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_2B_PS	16S	benthic	Gagne	Experimental	PowerSoil
16S_NA_GB_2B_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_3A_NS	16S	benthic	Gagne	Experimental	NucleoSpin
16S_NA_GB_3A_PS	16S	benthic	Gagne	Experimental	PowerSoil
16S_NA_GB_3A_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_3B_NS	16S	benthic	Gagne	Experimental	NucleoSpin
16S_NA_GB_3B_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_4A_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_4B_NS	16S	benthic	Gagne	Experimental	NucleoSpin
16S_NA_GB_4B_PW	16S	benthic	Gagne	Experimental	PowerWater
16S_NA_GB_5A_PS	16S	benthic	Gagne	Experimental	PowerSoil
16S_NA_GB_K1A_1_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1A_1_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1A_2_FD	16S	soil	Seguin	Experimental	FastDNA
16S_NA_GB_K1A_2_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1A_2_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1A_3_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1A_3_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1B_1_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1B_1_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1B_2_FD	16S	soil	Seguin	Experimental	FastDNA
16S_NA_GB_K1B_2_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1B_2_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1B_3_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1B_3_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1C_1_FD	16S	soil	Seguin	Experimental	FastDNA
16S_NA_GB_K1C_1_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1C_1_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1C_2_FD	16S	soil	Seguin	Experimental	FastDNA
16S_NA_GB_K1C_2_PS	16S	soil	Seguin	Experimental	PowerSoil

Table 1: 16S Samples *(continued)*

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
16S_NA_GB_K1C_2_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_K1C_3_PS	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_K1C_3_PW	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_M0R_PS_1	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PS_2	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PS_3	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PS_4	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PS_5	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PS_6	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M0R_PW_1	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M0R_PW_2	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M0R_PW_3	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M0R_PW_4	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M0R_PW_5	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M0R_PW_6	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PS_1	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M7R_PS_2	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M7R_PS_3	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M7R_PS_4	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M7R_PS_6	16S	water	Lawrence	Experimental	PowerSoil
16S_NA_GB_M7R_PW_1	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PW_2	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PW_3	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PW_4	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PW_5	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_M7R_PW_6	16S	water	Lawrence	Experimental	PowerWater
16S_NA_GB_PS_1A1	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_1A2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_1B1	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_1B2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_1C1	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_1C2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_3A2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_3B1	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_3B2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PS_3C2	16S	soil	Seguin	Experimental	PowerSoil
16S_NA_GB_PW_1A1	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_1A2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_1B1	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_1B2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_1C1	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_1C2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_3A1	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_3A2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_3B1	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_3B2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_PW_3C1	16S	soil	Seguin	Experimental	PowerWater

Table 1: 16S Samples *(continued)*

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
16S_NA_GB_PW_3C2	16S	soil	Seguin	Experimental	PowerWater
16S_NA_GB_SW_PS_1	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_10	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_11	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_13	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_14	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_15	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_16	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_17	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_2	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_3	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_4	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_5	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_6	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_7	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_8	16S	water	Watson	Spike	PowerSoil
16S_NA_GB_SW_PS_9	16S	water	Watson	Spike	PowerSoil

Table 2: 18S Samples

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
18S_NA_GB_B_1A_NS	18S	benthic	Gagne	Experimental	NucleoSpin
18S_NA_GB_B_2A_NS_A	18S	benthic	Gagne	Experimental	NucleoSpin
18S_NA_GB_B_6B_NS_A	18S	benthic	Gagne	Experimental	NucleoSpin
18S_NA_GB_S_1_FD_B	18S	soil	Seguin	Experimental	FastDNA
18S_NA_GB_S_1_PS_A	18S	soil	Seguin	Experimental	PowerSoil
18S_NA_GB_S_2_FD_B	18S	soil	Seguin	Experimental	FastDNA
18S_NA_GB_S_2_PS_A	18S	soil	Seguin	Experimental	PowerSoil
18S_NA_GB_S_2_PW	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_3_FD_B	18S	soil	Seguin	Experimental	FastDNA
18S_NA_GB_S_3_PS_B	18S	soil	Seguin	Experimental	PowerSoil
18S_NA_GB_S_3_PW_A	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_4_PW	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_5_PS_B	18S	soil	Seguin	Experimental	PowerSoil
18S_NA_GB_S_5_PW	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_6_FD_B	18S	soil	Seguin	Experimental	FastDNA
18S_NA_GB_S_6_PS_B	18S	soil	Seguin	Experimental	PowerSoil
18S_NA_GB_S_6_PW	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_HB1_PW_B	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_HB2_PW	18S	soil	Seguin	Experimental	PowerWater
18S_NA_GB_S_NT2_PW_B	18S	soil	Seguin	Experimental	PowerWater

Table 3: COI Samples

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
COI_NA_GB_B_1A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_1A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_1A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_1B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_1B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_1B_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_2A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_2A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_2A_PW_2	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_2A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_2B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_2B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_3A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_3A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_3A_PW_2	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_3A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_3B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_3B_PS_2	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_3B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_3B_PW_2	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_3B_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_4A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_4A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_4A_PW_2	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_4A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_4B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_4B_PS_2	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_4B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_4B_PW_2	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_4B_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_5A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_5A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_5A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_5B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_5B_PS_2	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_5B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_5B_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_6A_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_6A_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_6A_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_B_6B_PS	COI	benthic	Gagne	Experimental	PowerSoil
COI_NA_GB_B_6B_PW	COI	benthic	Gagne	Experimental	PowerWater
COI_NA_GB_B_6B_QPNS	COI	benthic	Gagne	Experimental	QuickPick_NucleoSpin
COI_NA_GB_I_Aaegypti_NS_1	COI	invertebrate	Ogden	Experimental	NucleoSpin
COI_NA_GB_I_Aaegypti_PS_1	COI	invertebrate	Ogden	Experimental	PowerSoil
COI_NA_GB_I_Aaegypti_PW_1	COI	invertebrate	Ogden	Experimental	PowerWater

Table 3: COI Samples *(continued)*

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
COI_NA_GB_I_Ctarsalis_NS_2	COI	invertebrate	Ogden	Experimental	NucleoSpin
COI_NA_GB_I_Ctarsalis_PS_2	COI	invertebrate	Ogden	Experimental	PowerSoil
COI_NA_GB_I_Ctarsalis_PW_2	COI	invertebrate	Ogden	Experimental	PowerWater
COI_NA_GB_S_1PS	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_1PW	COI	soil	Seguin	Experimental	PowerWater
COI_NA_GB_S_2PS	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_3PW	COI	soil	Seguin	Experimental	PowerWater
COI_NA_GB_S_4PW	COI	soil	Seguin	Experimental	PowerWater
COI_NA_GB_S_Holiday_Beach_1	COI	soil	Philips	Experimental	PowerSoil
COI_NA_GB_S_Holiday_Beach_2	COI	soil	Philips	Experimental	PowerSoil
COI_NA_GB_S_NB_15_1_1	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_1_2	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_2_1	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_2_2	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_3_1	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_3_2	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_4_1	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_4_2	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_5_1	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_NB_15_5_2	COI	soil	Philips	Experimental	PowerWater
COI_NA_GB_S_N_trial_1	COI	soil	Philips	Experimental	PowerSoil
COI_NA_GB_S_N_trial_2	COI	soil	Philips	Experimental	PowerSoil
COI_NA_GB_S_PS3_1	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_PS3_2	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_PS3_3	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_PS3_4	COI	soil	Seguin	Experimental	PowerSoil
COI_NA_GB_S_QPFD_1	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QPFD_2	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QPFD_3	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QPFD_4	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QPFD_5	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QPFD_6	COI	soil	Seguin	Experimental	QuickPick_FastDNA
COI_NA_GB_S_QP_Holiday_Beach	COI	soil	Philips	Experimental	QuickPick_PowerSoil
COI_NA_GB_S_QP_N_trial_1	COI	soil	Philips	Experimental	QuickPick_PowerSoil

Table 4: ITS Samples

LibraryName	Region	Sample	Provider	Experiment	ExtractionKit
ITS_NA_GB_B_1A_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_1A_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_1B_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_1B_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_1B_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_2A_NS_B	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_2A_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_2A_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_2B_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_2B_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_2B_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_3A_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_3A_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_3A_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_4A_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_4A_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_4B_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_4B_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_5A_NS_A	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_5A_NS_B	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_5A_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_5B_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_5B_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_5B_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_6A_NS	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_6A_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_B_6A_PW	ITS	benthic	Gagne	Experimental	PowerWater
ITS_NA_GB_B_6B_NS_B	ITS	benthic	Gagne	Experimental	NucleoSpin
ITS_NA_GB_B_6B_PS	ITS	benthic	Gagne	Experimental	PowerSoil
ITS_NA_GB_S_1_FD_A	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_1_PS_B	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_1_PW	ITS	soil	Seguin	Experimental	PowerWater
ITS_NA_GB_S_2_FD_A	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_2_PS_B	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_3_FD_A	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_3_PS_A	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_3_PW_B	ITS	soil	Seguin	Experimental	PowerWater
ITS_NA_GB_S_4_FD	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_4_PS	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_5_FD_A	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_5_FD_B	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_5_PS_A	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_6_FD_A	ITS	soil	Seguin	Experimental	FastDNA
ITS_NA_GB_S_6_PS_A	ITS	soil	Seguin	Experimental	PowerSoil
ITS_NA_GB_S_HB1_PW_A	ITS	soil	Seguin	Experimental	PowerWater
ITS_NA_GB_S_NT2_PW_A	ITS	soil	Seguin	Experimental	PowerWater