A Pipeline to Create OTU Table

Lei Yu

Outline

Introduction

Bioinformatics Pipeline

Statistical Results

Data Description

- Sequences Data (22.8 GB):

  18 Plots and 4 subplots in each plot + negative control ->73 samples
  Forward and complement reverse for each sample

- Environmental Data:
  Soil Chemistry, Trees, Location

PC or Cluster?

HPCC

High-Performance Computing Center (Linux System)

ssh -x [account name]@cluster.hpcc.ucr.edu

Login through terminal (Mac and Linux)

Login through Cygwin (Windows)

Login through FTP (FileZilla)

GitHub and GitLab

https://github.com/YULEITSINGTAO/

https://yuleitsingtao.github.io/TNC_project/

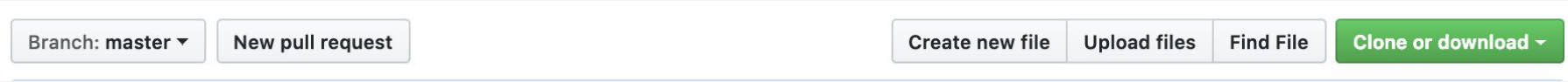
Repository management services are vital aspects of successfully developing software, either individually or collaboratively.

GitLab

vs.

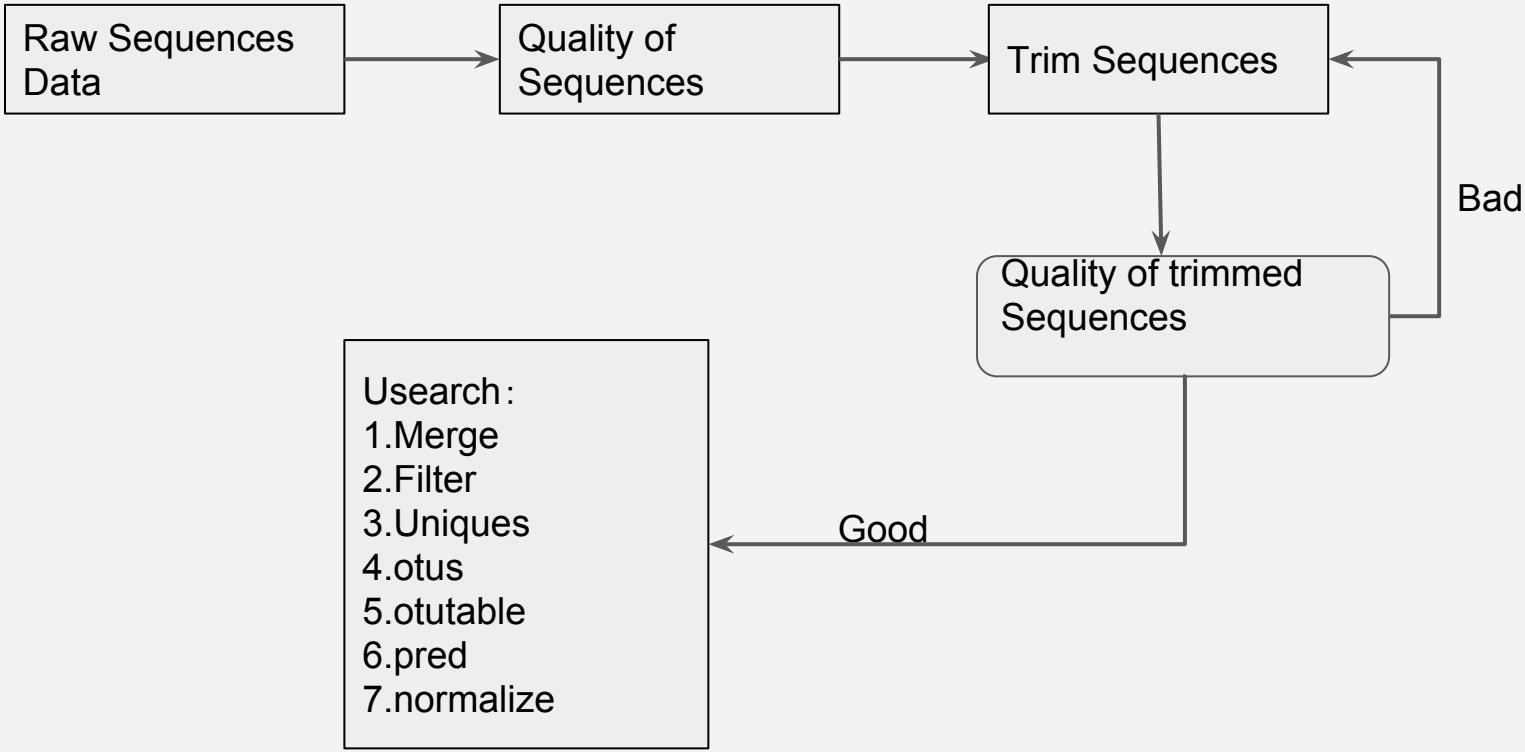
GitHub

Use the code source on GitHub




```
$ git clone [url]
```

Bioinformatics Pipeline



Run Programmes

1. Copy the pipeline

```
$ ssh -x [account name]@cluster.hpcc.ucr.edu
$ cd ~/shared
$ cp -a Lei_pipeline [your local directory]
```

2. Go to your local directory

```
$ cd [your local directory/Lei_pipeline]
```

3. Check files

```
$ ls
FastQC  cutadapt  raw_sequences_report  Qiime2
data_processing  trimmed_sequences  Usearch_project
input_raw_sequences  trimmed_sequences_report
```

4. Upload your sequences into [raw_sequences_report]

I like using FileZilla

Notes: Blue color -> Folders Green color-> Programmes

5. Data processing

```
$ cd data_processing
$ ls
trim  test_qc_input_raw_sequences
test_qc_trimmed_sequences
$ ./test_qc_input_raw_sequences
$ ./code_trimme
$ ./test_qc_trimmed_sequences
```

6. Go to check trimmed sequences

I like using FileZilla

Notes: If the quality of trimmed sequences are not good, you want to change the parameters in trim

7. Run Usearch

```
$ cd Usearch_project
$ ls
combine_otu_reference.shell  otus       summary
filter                       otutable   uniques.sh
make_database                out        unitITS.udb
merge                        pred       usearch
normalize.sh                 run
utax_reference_dataset_01.12.2017.fasta
$ ./run
```

Notes: You may want to use more CPUs, change the parameters in *uniques.sh*
Some parameters also need to change depend on trimmed sequences

8. See results in out

```
$ cd out
merged.fq  otutable.txt  reads_survey.txt  filtered.fa
otus.fa  otutable_normalized.txt  uniques.fa
```

9. Merge reads_survey.txt and otutable.txt
(This step can be finished in bash or any other languages you like)

```
$ cd ..
$ ./combine_otu_reference.shell
```

10. Merge otutable will be in out as well

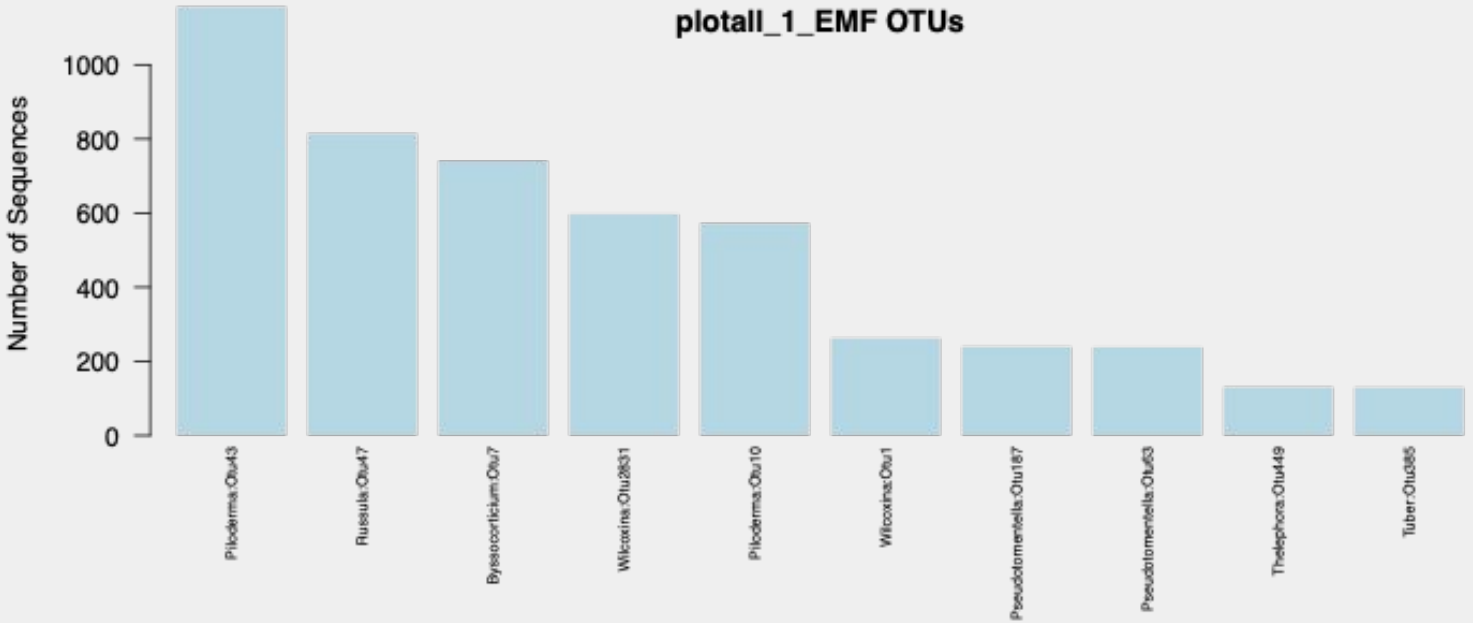
```
$ cd out
$ combine  merged.fq  otutable.txt
reads_survey.txt filtered.fa  otus.fa
otutable_normalized.txt  uniques.fa
```

Finally, we get the otu table

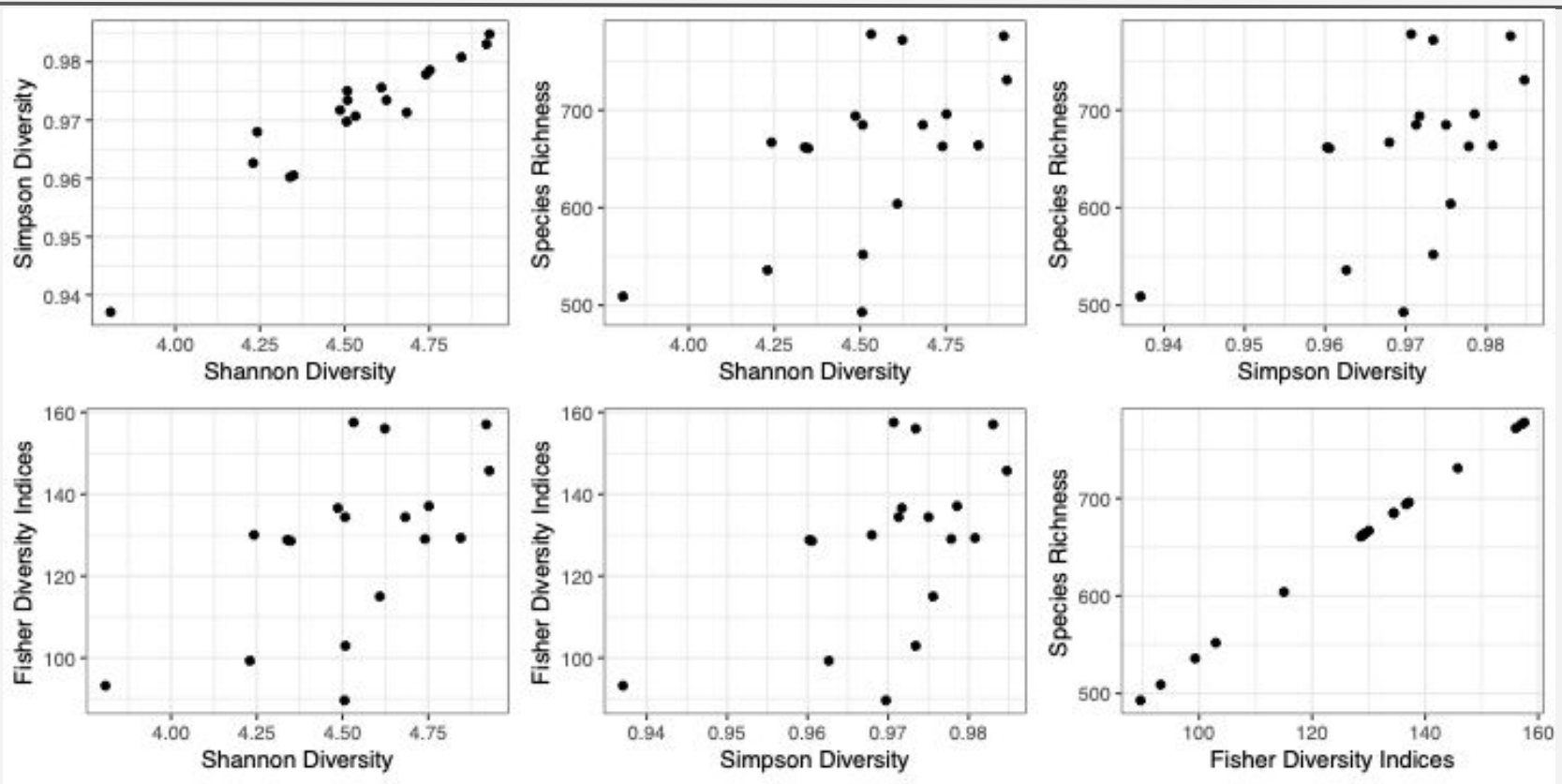
Statistics Work

- Language: R
- There are 4235 outs
- Abundance Curve
- Alpha diversity and Beta diversity

Abundance Curve



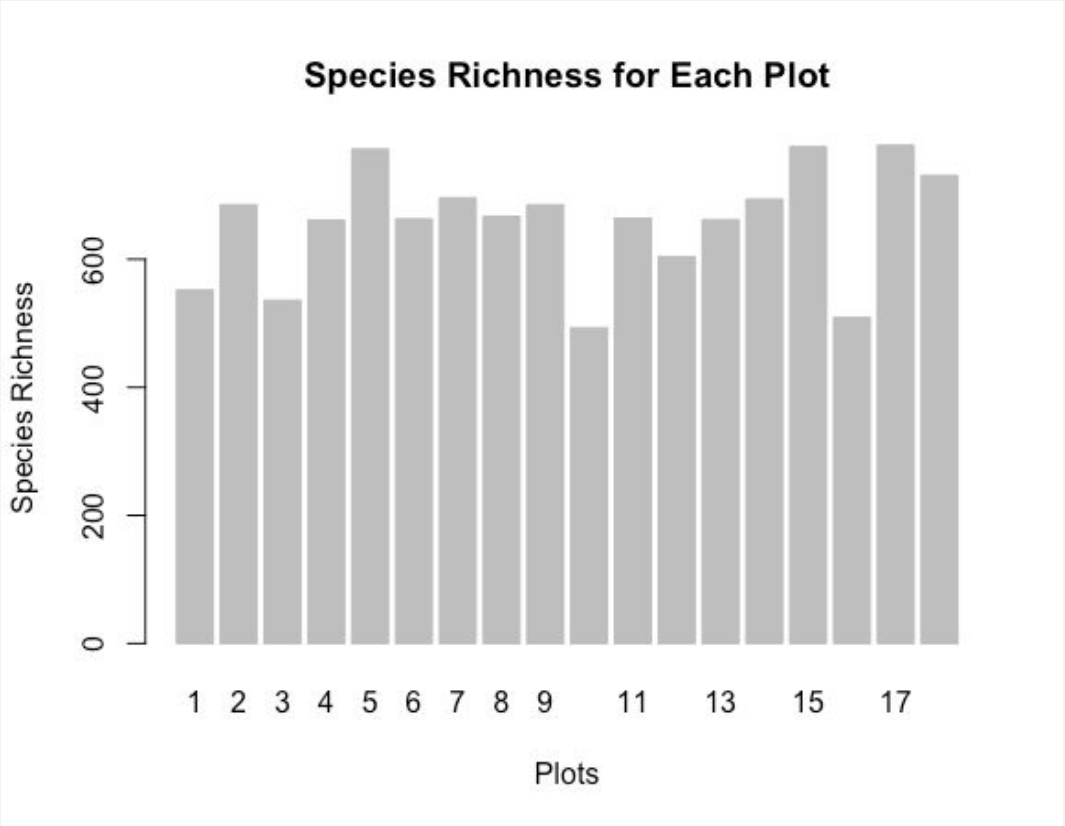
Alpha diversity



Species Richness

Plot 1, 3 10 and 16 show less species richness than others.

What factors influence the alpha diversity?



Environmental Data set

There are 226 features in the raw environmental data.

We clean the environmental data deleting the NA value and manual delete some useless features.

Then analysis the linear relationship between each features and delete the features which can be linear represented by others.

General Linear Regression

Based on general linear regression model, we find the 12 key features:

MG, CA, H, CEC, NO3_N, S, CC_Mult_FR, FRI, CEI, Precip_Yr, Water_Bal, CC_LIDAR.

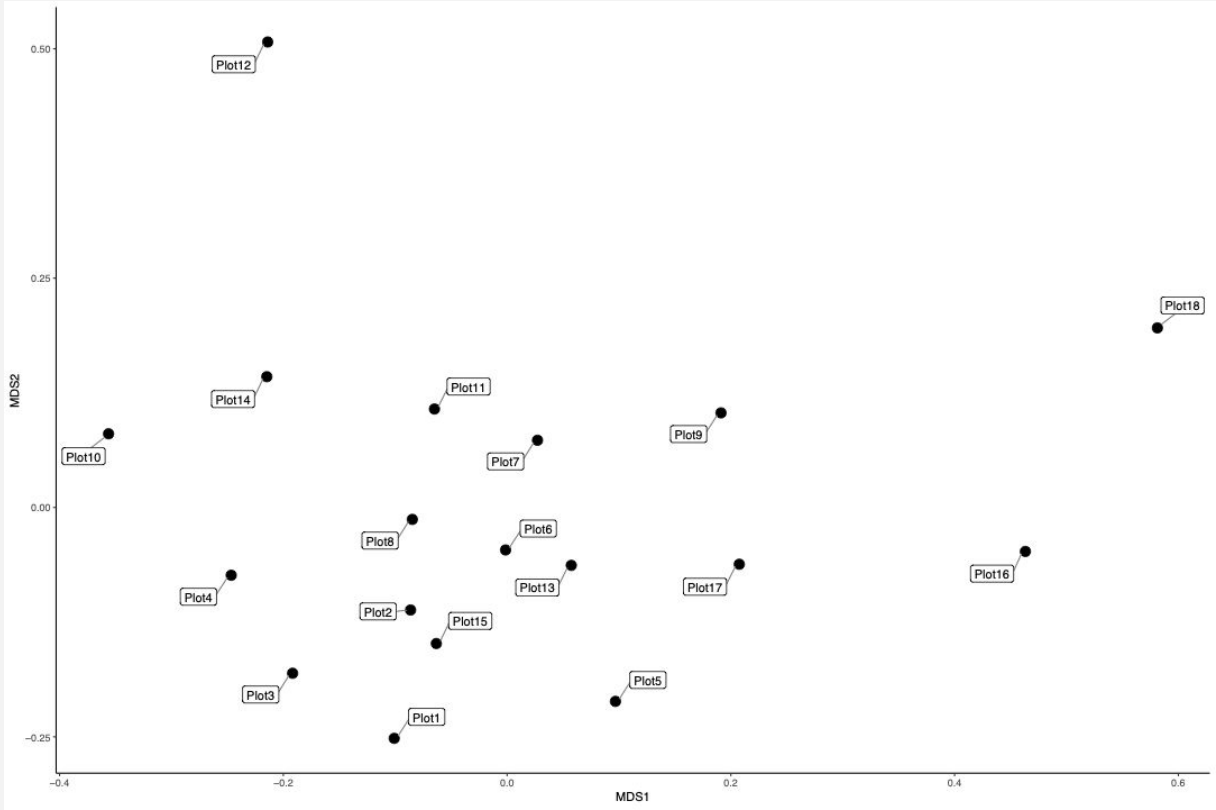
Soil chemistry > Trees Environment.

Beta Diversity

We see each plot as a community.

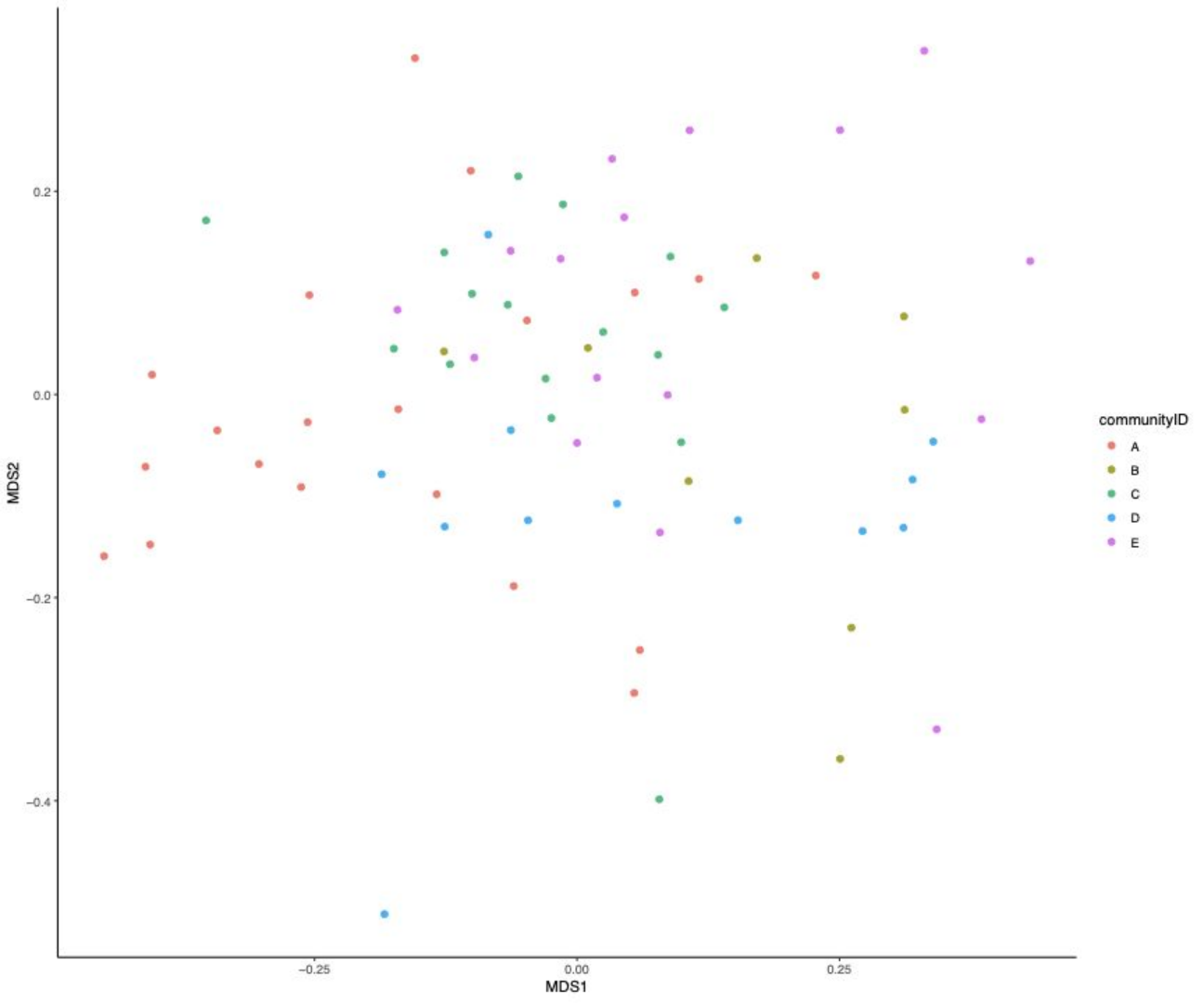
Plot 12, 16, 18

What features are influencing beta diversity?



We determine the community according to physical distance.

What features are influencing beta diversity?



Thanks