Diversidades alfa y beta en datos metagenomicos de shotgun de fresa

Camila Silva

2023-03-08

Pre-procesamiento de los datos

DIVERSIDADES CON TODO EL CONJUNTO DE DATOS Y LOS DATOS FILTRADOS POR CALIDAD

Los datos fueron entregados en una carpeta de drive https://drive.google.com/drive/folders/1x0106TYUr54gfqE6uod3g5qN8D
Q5x. Para poderlos usar en el servidor, se descargaron en mi maquina local y luego se pasaron por ssh al servidor.

En el equipo local

```
# $ scp Downloads/kraken_results-20230203T201634Z-001.zip camila@132.248.196.39:/home/camila/GIT/Tesis_
#  camila@132.248.196.39's password:
#  kraken_results-20230203T201634Z-001.zip          100%   12MB   2.2MB/s   00:05
# $ scp Downloads/fastp_results-20230203T175936Z-001.zip camila@132.248.196.39:/home/camila/GIT/Tesis_M
#  camila@132.248.196.39's password:
#  fastp_results-20230203T175936Z-001.zip           100% 2728KB   2.1MB/s   00:01
# $ scp Downloads/bracken_results-20230203T203724Z-001.zip camila@132.248.196.39:/home/camila/GIT/Tesis
#  camila@132.248.196.39's password:
#  bracken_results-20230203T203724Z-001.zip         100% 3798KB   2.4MB/s   00:01
```

En el servidor

```
# $ unzip kraken_results-20230203T201634Z-001.zip
# $ unzip bracken_results-20230203T203724Z-001.zip
# $ unzip fastp_results-20230203T175936Z-001.zip
```

Luego de tener las muestras Kraken y braken en el servidor con RStudio; es necesario generar el archivo **.biom**; para esto, se debe activar el ambiente de conda **metagenomis**, **kraken-biom** es un programa ampliamente utilizado para M a partir de la salida de kraken.

Con esto se obtiene una matriz de abundancia a partir de los archivos de salida de Kraken, que permite el uso de el paquete Phyloseq en R, que nos permiten analizar la diversidad y la abundancia mediante la manipulación de datos de asignación taxonómica.

```
# $ conda activate metagenomics

# $ kraken-biom kraken_results/* --fmt json -o fresa_kraken.biom
```

```
getwd()
```

```
## [1] "/home/camila/GIT/Tesis_Maestria/Analisis_Comparativo/Fresa_Solena"
```

```
# if (!requireNamespace("BiocManager", quietly = TRUE))
#   +     install.packages("BiocManager")
# BiocManager::install("phyloseq")

# install.packages(c("ggplot2", "readr", "patchwork"))
```

Phyloseq es un paquete de Bioconductor (Open Source Software For Bioinformatics) para la manipulación y análisis (herramienta para importar, guardar, analizar y visualizar) de datos metagenómicos generados por metodologías de secuenciación de alto rendimiento.

```
library("phyloseq")
library("ggplot2")
library("igraph")
```

```
##
## Attaching package: 'igraph'

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
library("readr")
library("patchwork")
library("vegan")
```

```
## Loading required package: permute

##
## Attaching package: 'permute'

## The following object is masked from 'package:igraph':
##
##     permute

## Loading required package: lattice

## This is vegan 2.6-4

##
## Attaching package: 'vegan'

## The following object is masked from 'package:igraph':
##
##     diversity
```

```
library("GUniFrac")
library("pbkrtest")
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
#library("BiodiversityR")
library("kableExtra")
```

```
## Registered S3 method overwritten by 'httr':
##   method         from
##   print.response rmutil
```

```
library("RColorBrewer")
```

Datos kraken

Importamos los datos kraken en un archivo biom

```
setwd("/home/camila/GIT/Tesis_Maestria/Data/fresa_solena")
fresa_kraken <- import_biom("fresa_kraken.biom")
class(fresa_kraken) # objeto phyloseq
```

```
## [1] "phyloseq"
## attr(,"package")
## [1] "phyloseq"
```

Queremos acceder a los datos que contiene nuestro objeto phyloseq **fresa_kraken**

Primero la tabla de taxonomia

```
#fresa_kraken@tax_table@.Data)
```

Cambiamos los nombres de las columnas a los niveles taxonomicos

```
colnames(fresa_kraken@tax_table@.Data) <- c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "S
```

Queremos cortar la parte inicial nombres, ya que aparecen,por ejemplo: "B___Bacteria" y queremos que solo se vea "Bacteria" substring(). Este comando ayuda a extraer o reemplazar caracteres en un vector.

```
fresa_kraken@tax_table@.Data <- substr(fresa_kraken@tax_table@.Data,4,100)
#View(fresa_kraken@tax_table@.Data)
```

Queremos ver la tabla de OTUs, esta tabla contiene las abundancias de los otus de cada muestra

```
#fresa_kraken@otu_table@.Data
```

Queremos cortar los nombres de las muestras para que coincida con los metadatos

```
colnames(fresa_kraken@otu_table@.Data) <- substr(colnames(fresa_kraken@otu_table@.Data),1,6)
#View(fresa_kraken@otu_table@.Data)
```

Cargar los metadatos

Ya que hay un desface de dos muestras entre los metadatos y las muestras de la otu_table se deben ver cuales son y quitarlas del archivo de metadatos

```
# $ ls kraken_results |cut -d'.' -f1 > lista_kraken.txt
# $ ls metadata.csv |cut -d',' -f1 > lista_metadata.txt
# $ wc *txt
# $ cat lista_metadata.txt lista_kraken.txt | sort | uniq -c
# $ cat lista_metadata.txt lista_kraken.txt | sort | uniq -c | sort | head
# $  1 MD2145
# $  1 MD2146
# $  2 MD2055
# $  2 MD2056
```

Eliminamos las dos muestras que no estaban en nuestra otu_table y cargamos los metadatos

```
metadata_fresa <- read.csv2("/home/camila/GIT/Tesis_Maestria/Data/fresa_solena/metadata.csv",header =
#metadata_fresa <- sample_data(metadata_fresa)
```

luego hacemos que los metadatos pertenezcan al objeto phyloseq en la seccion de **sam_data**

```
fresa_kraken@sam_data <- sample_data(metadata_fresa)
```

Creamos una columna extra en sam_data por necesidad de funcionamiento de mas adelante

```
fresa_kraken@sam_data$Sample<-row.names(fresa_kraken@sam_data)
colnames(fresa_kraken@sam_data)<-c('Treatment','Samples')
#View(fresa_kraken)
```

Ahora tenemos tambien la tabla de datos de calidad (fastp_kraken_summary),

para ver que muestra podemos eliminar de nuestro dataset, que no cumpla ciertos estandares de calidad + ID de la muestra + Reads_B - Reads_Before -> total de reads crudos + Reads_A - Reads_After -> total de reads despues del analisis de calidad + Reads_diff -> diferencia en tre Reads_B y Reads_A + Q30_B -> porcentaje arriba de 30 (escala fred) antes del analisis de calidad + Q30_A -> porcentaje arriba de 30 (escala fred) despues del analisis de calidad + LowQua -> reads de baja calidad + N_reads -> readas que contienen N y se descartan + too_short -> no pasan el tamaño minimo de calidad + Duplication -> porcentaje de duplicados + LengthR1 -> longitud promedio de los reads + LengthR2 -> longitud promedio de los reads + Classified -> porcentaje de clasificados del total despues del filtrado

ejemplo muestra MD2055 -> contienen 97millones de reads antes del filtrado de calidad, y despues queda con 79millones filtro usado para eliminar muestras, que luego del filtrado de calidad contengan menos de 25millones de reads los que nos da 5 muestras a eliminar (MP2079,MP2080,MP2088,MP2109,MP2137)

Eliminiar las muestras de baja calidad,usando filtro de menos de 25 millones de reads luegodel analisis de calidad se eliminan las muestras por su nombre

```
samples_to_remove <- c("MP2079","MP2080","MP2088","MP2109","MP2137")
fresa_kraken_fil <- prune_samples(!(sample_names(fresa_kraken) %in% samples_to_remove), fresa_kraken)
```

podemos comprobar el número de muestras antes y después del filtrado

```
nsamples(fresa_kraken) # 58
```

```
## [1] 58
```

```
nsamples(fresa_kraken_fil) # 53
```

```
## [1] 53
```

Queremos hacer un analisis de diversidad de nuestras muestras, para esto las dos metricas las usadas son: Diversidad Alfa y Beta