Data collection

TFM. PEC2 - Deliverable 1

Nuria Fernández González

10/06/2023

Index

The scripts and code of this project has been developed in different computers and servers that has been synced. Although the project folder structure is the same in the different machines, the working directory path and the way some functions are loaded can change through the deliverables depending where the code was run.

Versions: * Linux version: Ubuntu 20.04.4 LTS

1 Sequences

1.1 Envision project

ENVISION sequence data can be found in the The European Nucleotide Archive (ENA) or The Sequence Read Archive (SRA) databases under bioproject ID PRJEB36188.

Using Aspera and SRA-explorer (https://sra-explorer.info/#) to download fastq-files.

```
# Install aspera
wget https://ak-delivery04-mul.dhe.ibm.com/sar/CMA/OSA/0adrj/0/ibm-aspera-connect_4.1.3.93_linux.tar.gz
tar zxvf ibm-aspera-connect_4.1.3.93_linux.tar.gz
bash ibm-aspera-connect_4.1.3.93_linux.sh

# Temporary add the aspera executable to $path
PATH="$HOME/.aspera/connect/bin:$PATH"
```

Generate Aspera commands for downloading all fastq files of the project with SRA-explorer that are saved in a script: `sra_explorer_download.sh`

```
$ head sra_explorer_download.sh
#!/usr/bin/env bash
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
ascp -QT -l 300m -P33001 -i $HOME/.aspera/connect/etc/asperaweb_id_dsa.openssh era-fasp@fasp.sra.ebi.ac
```

The data table with the bioproject information was directly downloaded from the SRA website.

For that, go to SRA and search for the project or go to https://www.ncbi.nlm.nih.gov/sra/?term=PRJEB36188

Then, click on **Send results to Run selector** and in the **Select box** click on **Download Total-Metadata** (we will need the full table, not only the accession list). Save table as csv file: `SRARunTable_metadata.csv`.

After visualizing the metadata table in Excel to check all variables included, remove those not needed keeping: study_accession, sample_accession, experiment_accession, run_accession, tax_id, scientific_name, library_name, read_coun, fastq_ftp, submitted_ftp, sra_ft, sample_alias, sample_titleand. Save the file as a tab-delimited file: `filereport_read_run_PRJEB36188.tsv`.

This bioproject contains samples from both field and incubation experiments. In addition, it contains samples of the 16S and 18S rRNA genes targeting **Prokaryotes** and **Eukaryotes** respectively, but this information is lacking in the bioproject metadata. To distinguish between them and being able to select the samples of interest for this project, my collaborators provided the sequencing form used when submitting the samples to the sequencing company.

Dimension project

Dimension project raw fastq sequences and sample correspondence table were directly obtained from collaborators.

2 Environmental data

Tables with the environmental metadata were obtained from collaborators.