ciimar

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

# Meta_Microbial Workshop

## Exploring Online Resources and Repositories **I**

- Practical example (Silva NGS)

**Catarina Magalhães**

Funding Sponsorship

fct Fundação para a Ciência e a Tecnologia

BIOPORTUGAL S.A. Químico, Farmacêutica

SPECO

# PLAN FOR THE NEXT 45’

- Silva NGS | Short Introduction
- Start a Case Study
  - Samples/Data Description
  - Silva NGS Registration
  - Execute a Project
  - Analyze the Outputs in Working Groups by Responding to a Challenge

    Present Your Results to the Class

# SILVA NGS | SHORT INTRODUCTION

- SILVAngs is a data analysis service for ribosomal RNA gene (rDNA) amplicon reads generated by next-generation sequencing (NGS), based on an automatic software pipeline.
- It uses the SILVA databases to classify rDNA reads and provides several outputs, like taxonomy tables and multiple graphs for download.
- SILVAngs uses the de.NBI Cloud (German Network for Bioinformátics Infrastructure) to process all user projects. The de.NBI cloud provide free compute resources for academic users.

*Quast et al. 2013*
Developed by Frank Oliver Group
Max Planck Institute for Marine Microbiology

*Contact: ngs-contact@arb-silva.de*

# SILVA NGS | SHORT INTRODUCTION

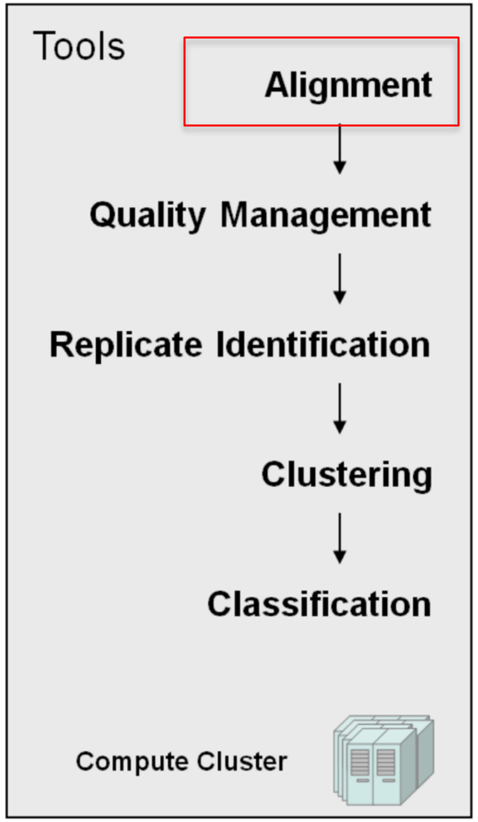
SILVAngs User Guide:

https://www.google.com/search?q=silva+ngs+user+guide&rlz=1C5CHFA_enPT857PT857&oq=silva+ngs+user+guide&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIGCAEQRRg80gEINDk2NWowajSoAgCwAgE&sourceid=chrome&ie=UTF-8

The basic workflow of the pipeline can be divided into the following steps

- Alignment | Initial quality control
- Quality management | further quality filtering
- De-replication (identical sequences)
- Clustering (OTU definition)
- Classification of the OTUs/reads

# 1º Alignment



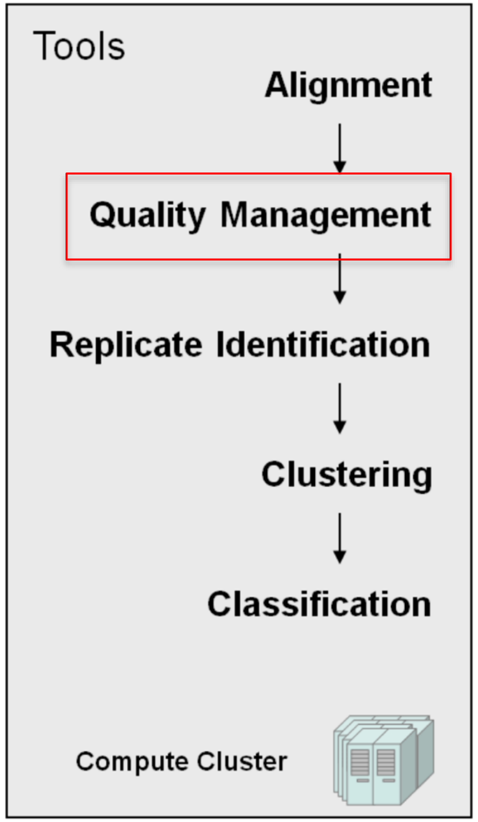
## Alignment | Initial quality control

All input reads are aligned by SILVA Incremental Aligner (SINA - http://www.arb-silva.de/aligner/sina-download/)

According to a number of parameters measured by SINA, problematic reads (such as PCR artefacts) or even contamination of the dataset with non-rRNA gene sequences are identified and not considered for further processing

*Pruesse et al. 2012*
*https://academic.oup.com/bioinformatics/article/28/14/1823/218226*

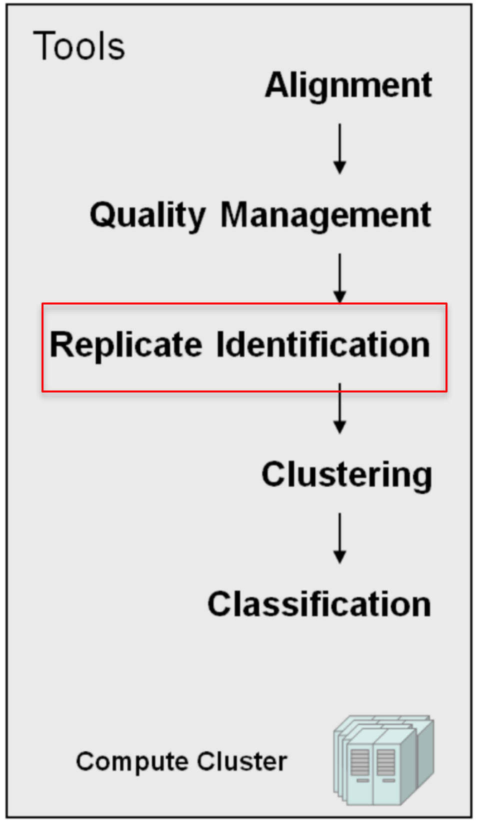
# 2º Quality management



## Further quality filtering

All reads which have not been rejected by the previous alignment step undergo further quality filtering, including **length**, **ambiguity** (ambiguous bases) and **homopolymer** (series of consecutive identical bases) checks. The sequence length cut-off can be defined by the user, whereas for ambiguities and homopolymers thresholds of max. 2% are used.

# 2º Replication Identification

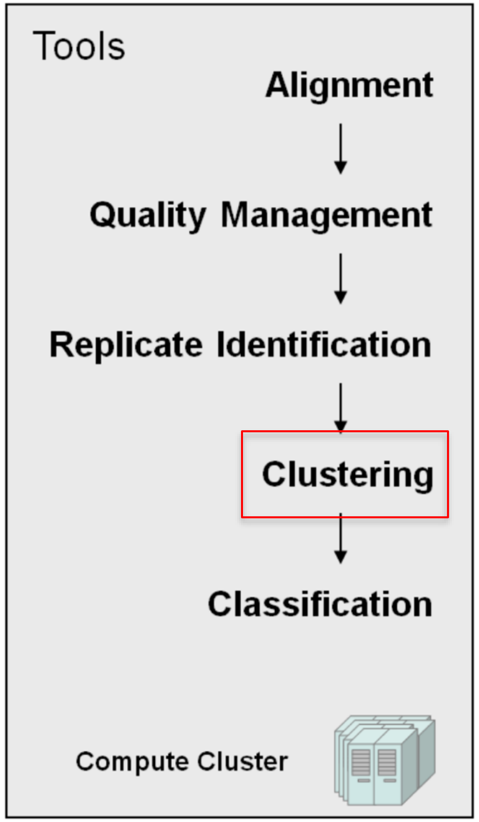


**De-replication | identification of identical sequences**

All remaining reads enter the de-replication stage of the pipeline. 100% identical reads, are identified and only one read is retained for further processing.

This is to reduce calculation time, since processing redundant reads is a waste of computing power.

# 4º Clustering

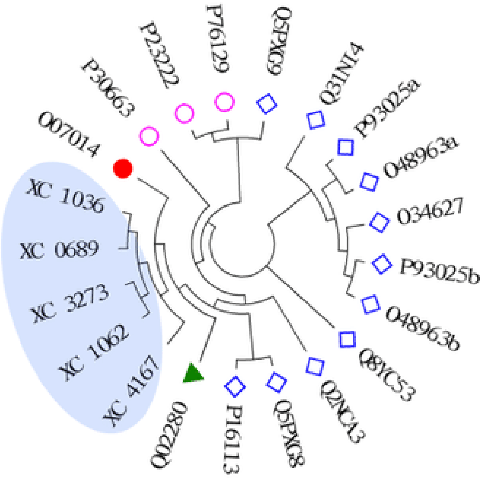


## Clustering (OTU definition)

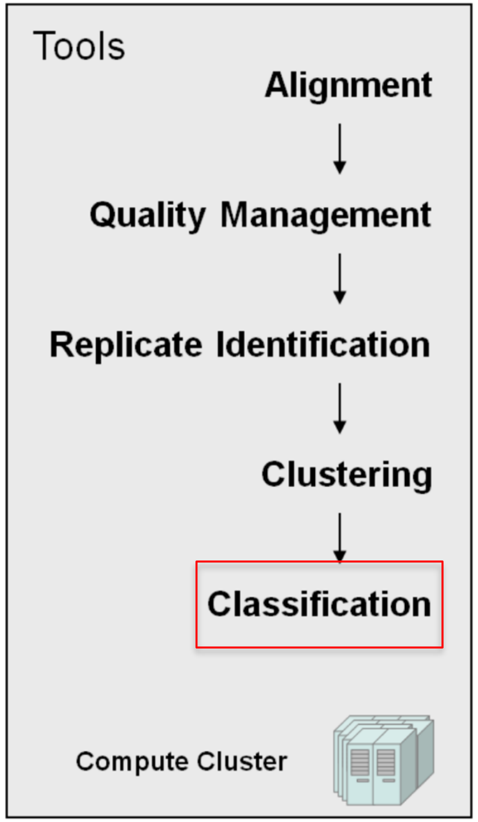
Clustering is done on a **97-99% identity level**. This is motivated by the fact that PCR and sequencing errors can easily introduce 3% artificial divergence in the sequences.

Technically, this is just another de-replication step to further reduce the number of reads that needs to be classified. Compared to previous de-replication,

97-99 % Identity
(can be adjusted)



# 4º Classification



Tools

Alignment

Quality Management

Replicate Identification

Clustering

Classification

Compute Cluster

## - Classification of the OTUs/reads

In the classification step the representative reads (the longest read of each OTU) are compared to the SILVA reference datasets of the small- (16S/18S) and large (23S/28S) subunit rDNA with its corresponding SILVA taxonomy (Quast et al. 2013).

Only significant hits are considered; everything else is assigned to a class called ‘No Relative’.

The classification result of each representative read is mapped back to all other reads of the OTU cluster and also to the corresponding identical reads from de-replication step.

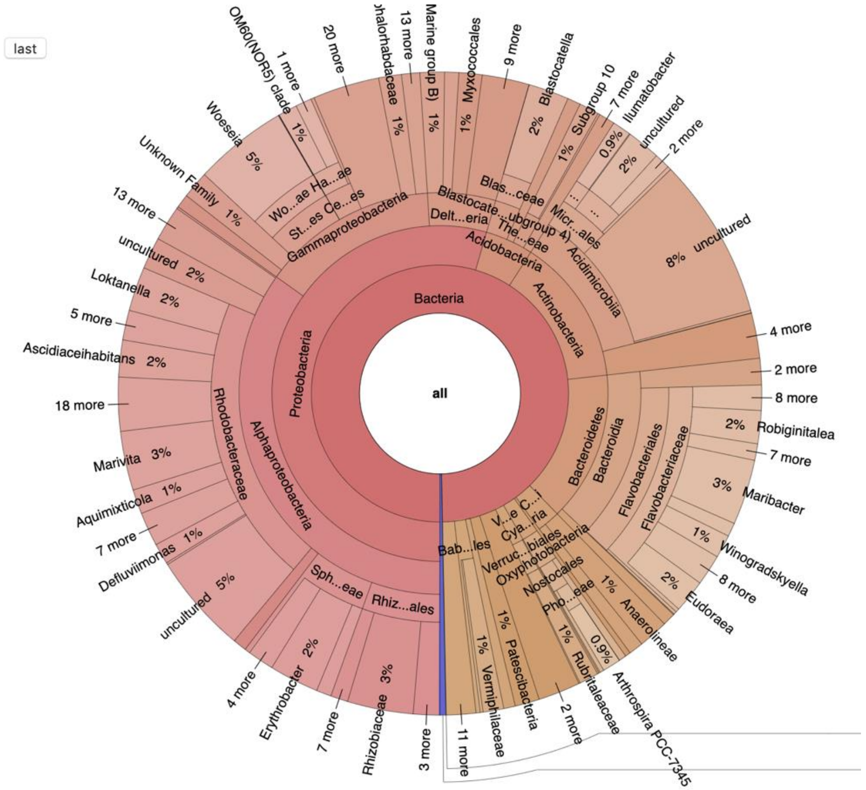
Is recommended to avoid overinterpretation of the results especially for reads below 1200 bases. For amplicon illimina data SILVA’s phylogeny-based taxonomy has a reliable resolution down to the genus level.

Meta_Microbial
Workshop

silva ngs

# 5º RESULTS

- Chart Gallery (e.g. Sequence lengh distribution; Rarefaction curves)
- Fingerprints (It offers the option to specify the taxonomic depth, filter by taxonomy,
- include/remove samples)
- Taxplotkrona (Interactive visualization)
- Archive Zip (with all the autputs)
- Report (HTML and PDF)



Meta_Microbial
Workshop

# **SILVA** NGS Pipeline | Let’s Make It Happen!

silva ngs

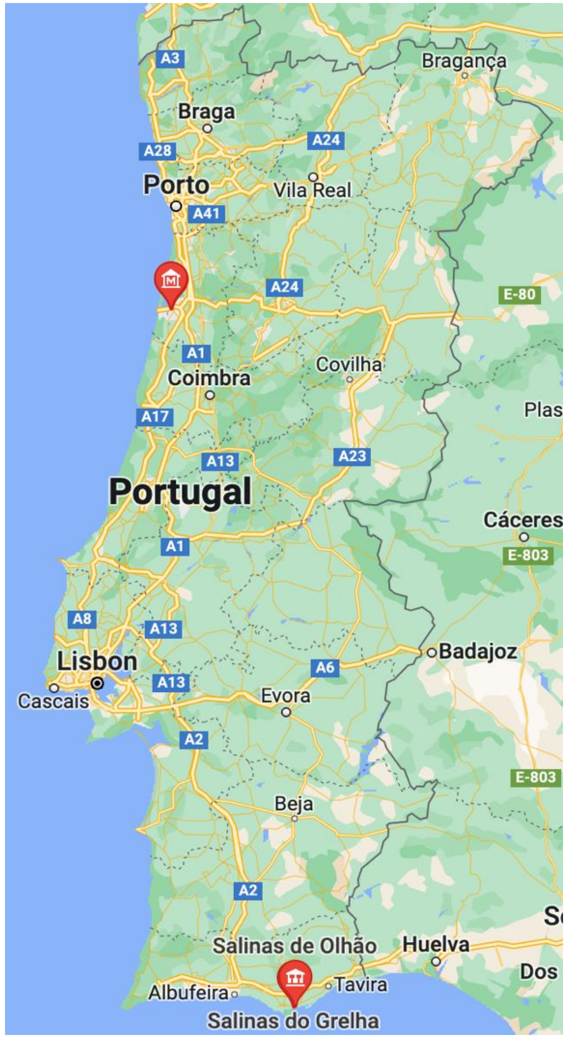
1. Go to SILVA NGS Web Page at https://www.arb-silva.de/ngs/
2. Register at SILVA NGS platform and log in.
3. Start a project
   3.1 Go to ”My Projects”
   3.2 Select ”Create Project”
   3.3 Add a Name in “Project Name” – *No spaces or special characters*
   3.4 In Sequence Type chose 16S/18S, SSU
   3.5 In Sequence Technology select “Illumina (MiSeq/HiSeq)”
   3.6 In Expected Sequence Quantity write “300000”
   3.7 In Expected Read Length write “400”
   3.8.In project description write for example “Samples from Salt Pans”
4. In Upload Files “Upload your sequences”. The FASTA files E1, E2, E3 and E4 available in: https://drive.google.com/drive/folders/1dqUnQHbp1VJHUSSDCVdCoZIA3FLep9LS?usp=sharing
5. Select “Execute project”.

*SILVAngs - rDNA-based microbial community analysis using next-generation sequencing (NGS) data - User Guide_2020. https://ngs.arb-silva.de/silvangs/#*

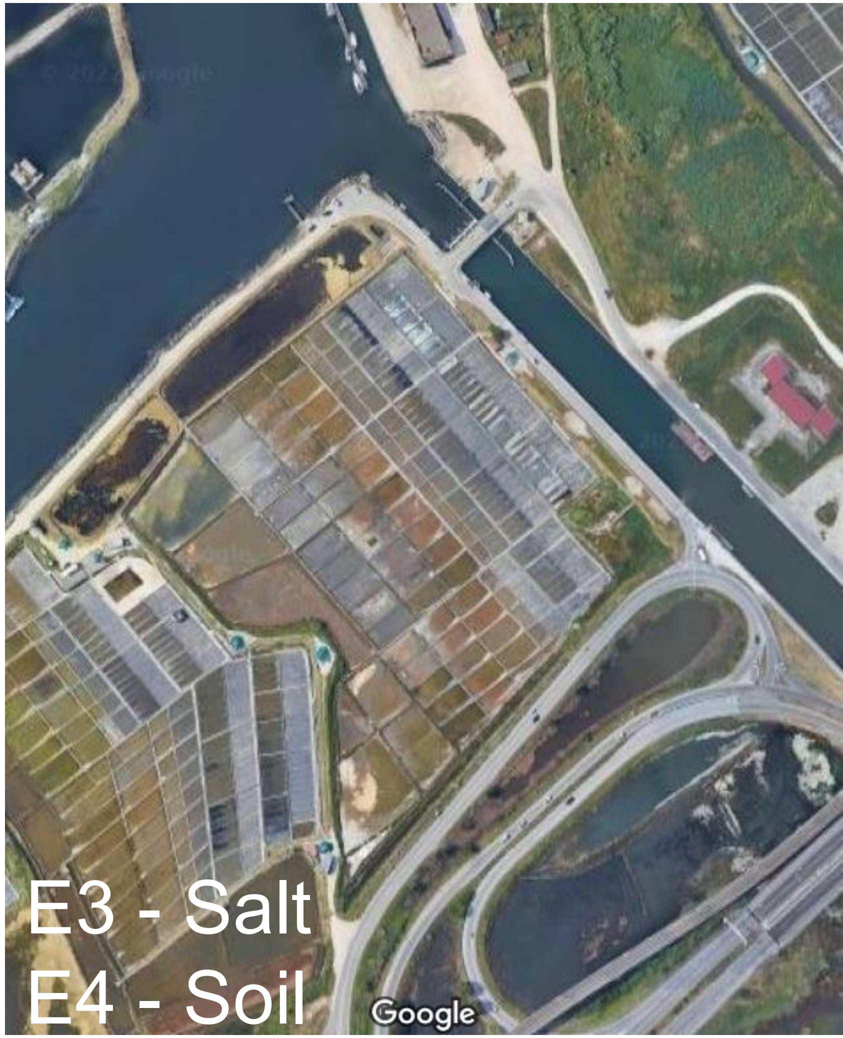
# SAMPLES WE WILL ANALYZE

## Olhão Salt Pans

E1 - Sal
E2 - Soil

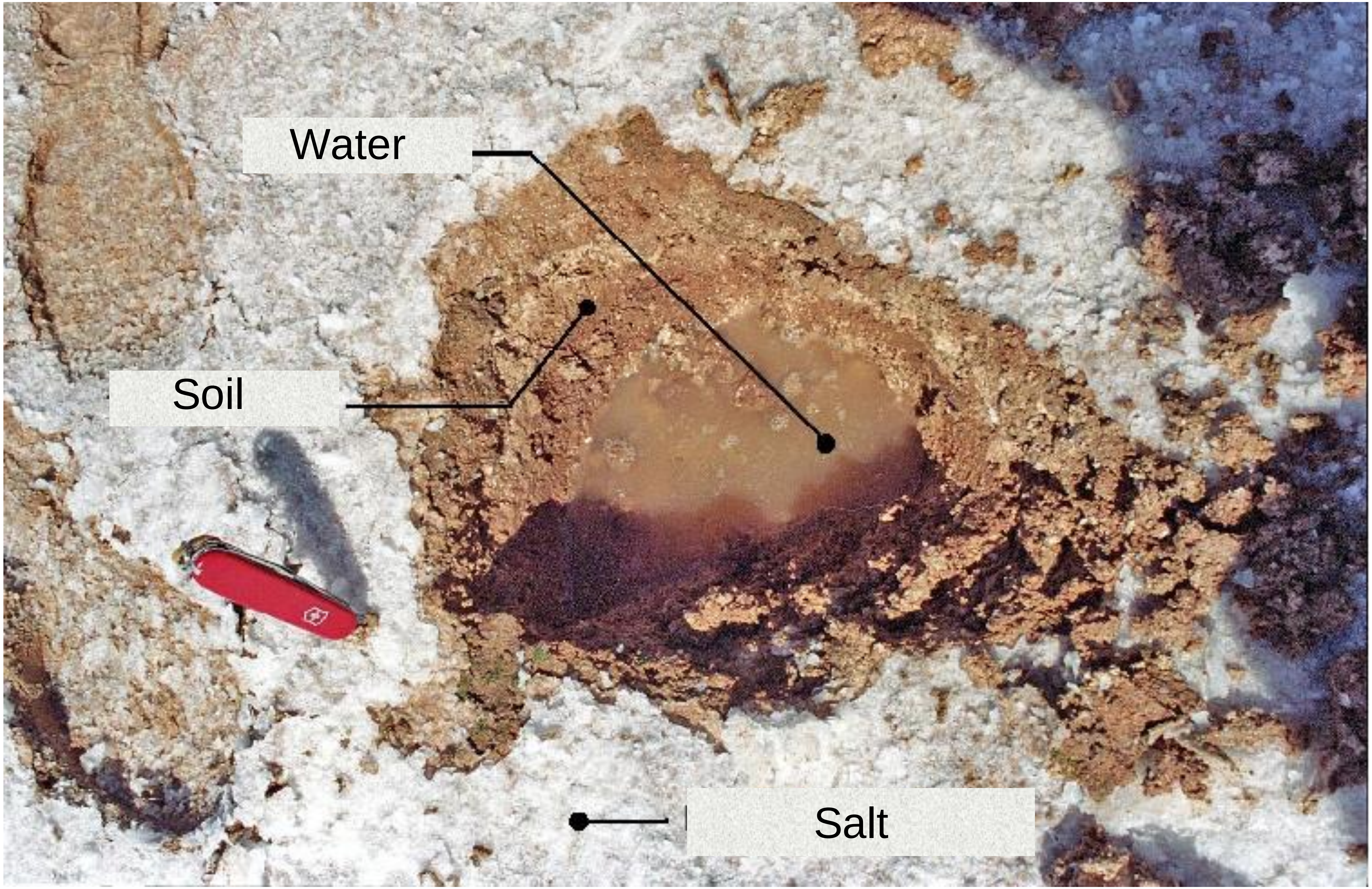


## Aveiro Salt Pans



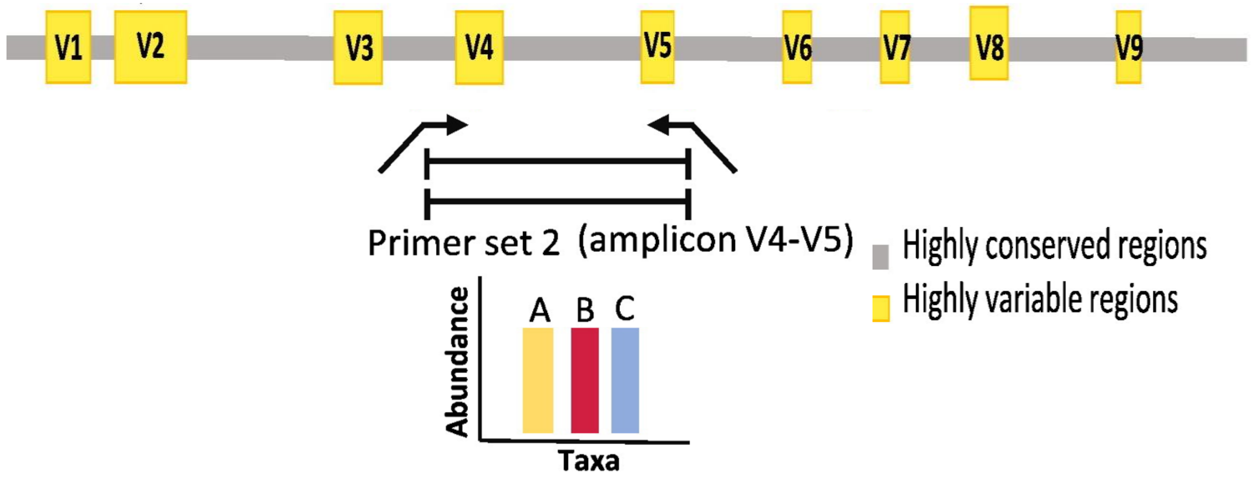
E3 - Salt
E4 - Soil

# SAMPLES WE WILL ANALYZE



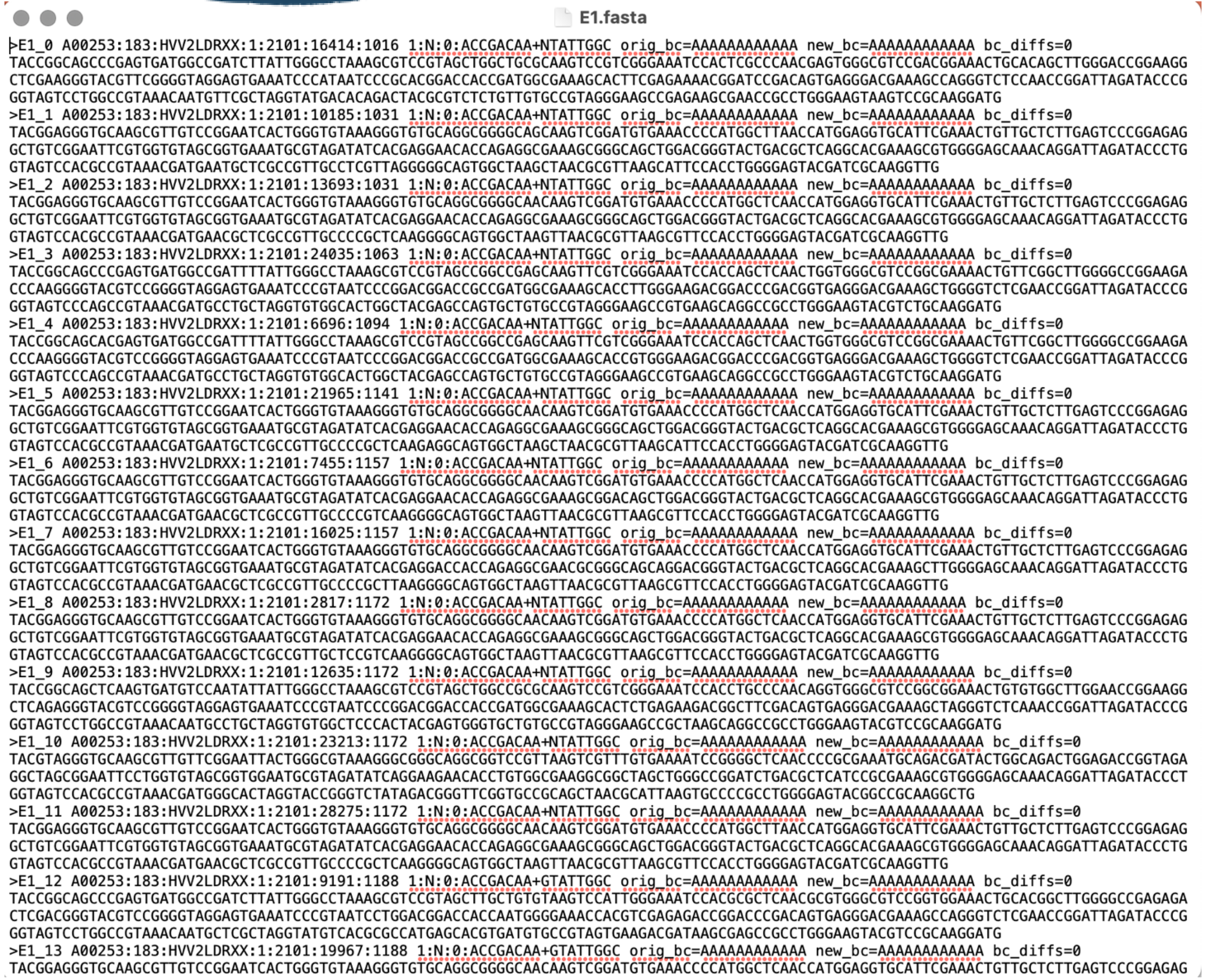
# METABARCODING
# GENETIC MARKER | 16S rRNA GENE

## 515F-Y/926R - primers
## Hypervariable Regions V4 and V5 of the 16S rRNA Gene



earth
microbiome project

*Parada et al. 2016*
https://pubmed.ncbi.nlm.nih.gov/26271760/

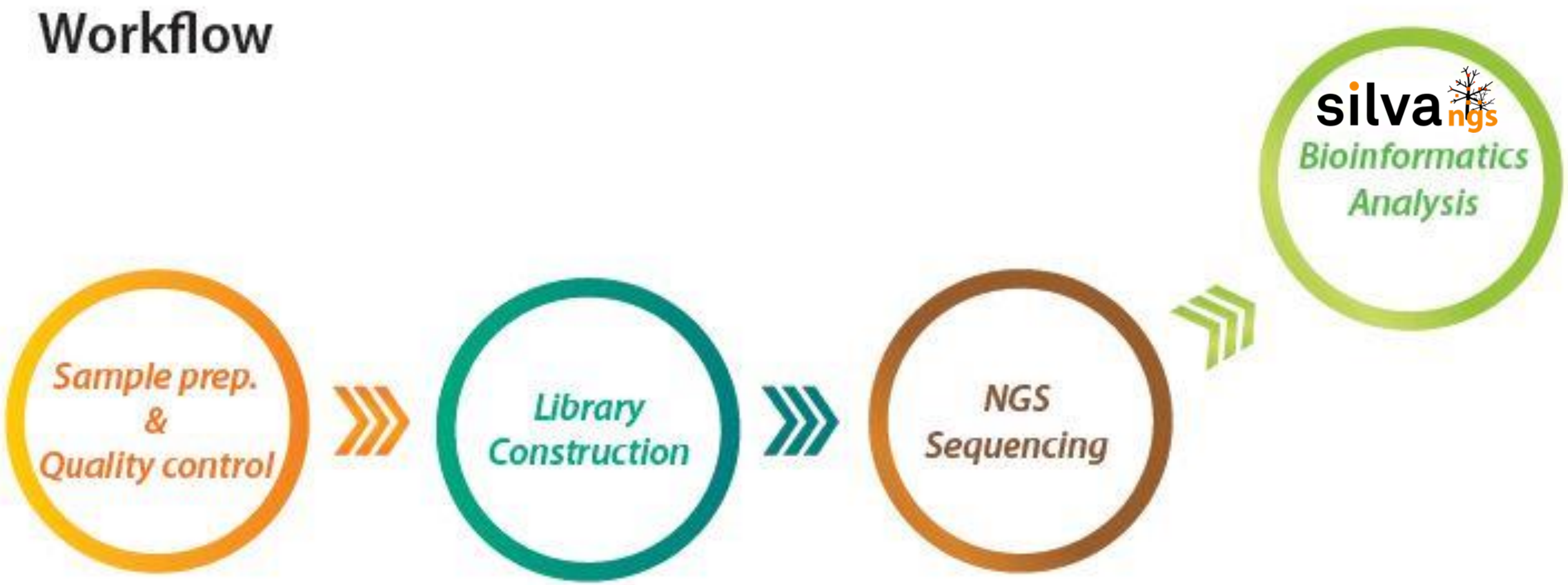


```
>E1_0 A00253:183:HVV2LDRXX:1:2101:16414:1016 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACCGGCAGCCCGAGTGATGGCCGATCTTATTGGGCCTAAAGCGTCCGTAGCTGGCTGCGCAAGTCCGTCGGGAAATCCACTCGCCCAACGAGTGGGCGTCCGACGGAAACTGCACAGCTTGGGACCGGAAGG
CTCGAAGGGTACGTTCGGGGTAGGAGTGAAATCCCATAATCCCGCACGGACCACCGATGGCGAAAGCACTTCGAGAAAACGGATCCGACAGTGAGGGACGAAAGCCAGGGTCTCCAACCGGATTAGATACCCG
GGTAGTCCTGGCCGTAAACAATGTTCGCTAGGTATGACACAGACTACGCGTCTCTGTTGTGCCGTAGGGAAGCCGAGAAGCGAACCGCCTGGGAAGTAAGTCCGCAAGGATG
>E1_1 A00253:183:HVV2LDRXX:1:2101:10185:1031 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAGCAAGTCGGATGTGAAACCCCATGGCTTAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGGCAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAATGCTCGCCGTTGCCTCGTTAGGGGGCAGTGGCTAAGCTAACGCGTTAAGCATTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_2 A00253:183:HVV2LDRXX:1:2101:13693:1031 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGGCAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAACGCTCGCCGTTGCCCCGCTCAAGGGGCAGTGGCTAAGTTAACGCGTTAAGCGTTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_3 A00253:183:HVV2LDRXX:1:2101:24035:1063 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACCGGCAGCCCGAGTGATGGCCGATTTTATTGGGCCTAAAGCGTCCGTAGCCGGCCGAGCAAGTTCGTCGGGAAATCCACCAGCTCAACTGGTGGGCGTCCGGCGAAAACTGTTCGGCTTGGGGCCGGAAGA
CCCAAGGGGTACGTCCGGGGTAGGAGTGAAATCCCGTAATCCCGGACGGACCGCCGATGGCGAAAGCACCTTGGGAAGACGGACCCGACGGTGAGGGACGAAAGCTGGGGTCTCGAACCGGATTAGATACCCG
GGTAGTCCCAGCCGTAAACGATGCCTGCTAGGTGTGGCACTGGCTACGAGCCAGTGCTGTGCCGTAGGGAAGCCGTGAAGCAGGCCGCCTGGGAAGTACGTCTGCAAGGATG
>E1_4 A00253:183:HVV2LDRXX:1:2101:6696:1094 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACCGGCAGCACGAGTGATGGCCGATTTTATTGGGCCTAAAGCGTCCGTAGCCGGCCGAGCAAGTTCGTCGGGAAATCCACCAGCTCAACTGGTGGGCGTCCGGCGAAAACTGTTCGGCTTGGGGCCGGAAGA
CCCAAGGGGTACGTCCGGGGTAGGAGTGAAATCCCGTAATCCCGGACGGACCGCCGATGGCGAAAGCACCGTGGGAAGACGGACCCGACGGTGAGGGACGAAAGCTGGGGTCTCGAACCGGATTAGATACCCG
GGTAGTCCCAGCCGTAAACGATGCCTGCTAGGTGTGGCACTGGCTACGAGCCAGTGCTGTGCCGTAGGGAAGCCGTGAAGCAGGCCGCCTGGGAAGTACGTCTGCAAGGATG
>E1_5 A00253:183:HVV2LDRXX:1:2101:21965:1141 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGGCAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAATGCTCGCCGTTGCCCCGCTCAAGAGGCAGTGGCTAAGCTAACGCGTTAAGCATTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_6 A00253:183:HVV2LDRXX:1:2101:7455:1157 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGACAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAACGCTCGCCGTTGCCCCGTCAAGGGGCAGTGGCTAAGTTAACGCGTTAAGCGTTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_7 A00253:183:HVV2LDRXX:1:2101:16025:1157 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGACCACCAGAGGCGAACGCGGGCAGCAGGACGGGTACTGACGCTCAGGCACGAAAGCTTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAACGCTCGCCGTTGCCCCGCTTAAGGGGCAGTGGCTAAGTTAACGCGTTAAGCGTTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_8 A00253:183:HVV2LDRXX:1:2101:2817:1172 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGGCAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAACGCTCGCCGTTGCTCCGTCAAGGGGCAGTGGCTAAGTTAACGCGTTAAGCGTTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_9 A00253:183:HVV2LDRXX:1:2101:12635:1172 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACCGGCAGCTCAAGTGATGTCCAATATTATTGGGCCTAAAGCGTCCGTAGCTGGCCGCGCAAGTCCGTCGGGAAATCCACCTGCCCAACAGGTGGGCGTCCGGCGGAAACTGTGTGGCTTGGAACCGGAAGG
CTCAGAGGGTACGTCCGGGGTAGGAGTGAAATCCCGTAATCCCGGACGGACCACCGATGGCGAAAGCACTCTGAGAAGACGGCTTCGACAGTGAGGGACGAAAGCTAGGGTCTCAAACCGGATTAGATACCCG
GGTAGTCCTGGCCGTAAACAATGCCTGCTAGGTGTGGCTCCCACTACGAGTGGGTGCTGTGCCGTAGGGAAGCCGCTAAGCAGGCCGCCTGGGAAGTACGTCCGCAAGGATG
>E1_10 A00253:183:HVV2LDRXX:1:2101:23213:1172 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGTAGGGTGCAAGCGTTGTTCGGAATTACTGGGCGTAAAGGGCGGGCAGGCGGTCCGTTAAGTCGTTTGTGAAAATCCGGGGCTCAACCCCGCGAAATGCAGACGATACTGGCAGACTGGAGACCGGTAGA
GGCTAGCGGAATTCCTGGTGTAGCGGTGGAATGCGTAGATATCAGGAAGAACACCTGTGGCGAAGGCGGCTAGCTGGGCCGGATCTGACGCTCATCCGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCT
GGTAGTCCACGCCGTAAACGATGGGCACTAGGTACCGGGTCTATAGACGGGTTCGGTGCCGCAGCTAACGCATTAAGTGCCCCGCCTGGGGAGTACGGCCGCAAGGCTG
>E1_11 A00253:183:HVV2LDRXX:1:2101:28275:1172 1:N:0:ACCGACAA+NTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTTAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
GCTGTCGGAATTCGTGGTGTAGCGGTGAAATGCGTAGATATCACGAGGAACACCAGAGGCGAAAGCGGGCAGCTGGACGGGTACTGACGCTCAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTG
GTAGTCCACGCCGTAAACGATGAACGCTCGCCGTTGCCCCGCTCAAGGGGCAGTGGCTAAGTTAACGCGTTAAGCGTTCCACCTGGGGAGTACGATCGCAAGGTTG
>E1_12 A00253:183:HVV2LDRXX:1:2101:9191:1188 1:N:0:ACCGACAA+GTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACCGGCAGCCCGAGTGATGGCCGATCTTATTGGGCCTAAAGCGTCCGTAGCTTGCTGTGTAAGTCCATTGGGAAATCCACGCGCTCAACGCGTGGGCGTCCGGTGGAAACTGCACGGCTTGGGGCCGAGAGA
CTCGACGGGTACGTCCGGGGTAGGAGTGAAATCCCGTAATCCTGGACGGACCACCAATGGGGAAACCACGTCGAGAGACCGGACCCGACAGTGAGGGACGAAAGCCAGGGTCTCGAACCGGATTAGATACCCG
GGTAGTCCTGGCCGTAAACAATGCTCGCTAGGTATGTCACGCGCCATGAGCACGTGATGTGCCGTAGTGAAGACGATAAGCGAGCCGCCTGGGAAGTACGTCCGCAAGGATG
>E1_13 A00253:183:HVV2LDRXX:1:2101:19967:1188 1:N:0:ACCGACAA+GTATTGGC orig_bc=AAAAAAAAAAAA new_bc=AAAAAAAAAAAA bc_diffs=0
TACGGAGGGTGCAAGCGTTGTCCGGAATCACTGGGTGTAAAGGGTGTGCAGGCGGGGCAACAAGTCGGATGTGAAACCCCATGGCTCAACCATGGAGGTGCATTCGAAACTGTTGCTCTTGAGTCCCGGAGAG
```

Number of Sequences
63,244

*Ficheiros* ***E1****; E2; E3; E4*

# Microbial Diversity
## NGS Data Analysis



Workflow

Sample prep.
&
Quality control

Library
Construction

NGS
Sequencing

silva ngs
Bioinformatics
Analysis

# CHALLENGE – Extract Insights from Your Data

## Form Groups of 4-5 Elements

In this challenge, we ask to conduct a visual and interpretative analysis of the SilvaNGS outputs to study the diversity of microbiomes in Olhão and Aveiro Salt Pans.

This platform will generate graphical visualizations and taxonomic tables are available at
https://drive.google.com/drive/folders/1d9I1ThGaiH-wB29PLyyCMNGCOVelPOG0?usp=sharing

Meta_Microbial
Workshop

silva ngs

# CHALLENGE – Extract Insights from Your Data

## Form Groups of 4-5 Elements

**Analyze the results obtained in groups in order to answer one of the following questions**:

1. Which samples showed the greatest diversity of prokaryotes?
2. Which prokaryotic phyla were present in all samples?
3. Which prokaryotic phyla differed between the two studied salt pans (Olhão and Aveiro)?
4. Among the prokaryotes, which domain is better represented in the studied salt pan samples: Archaea or Bacteria? What differences were observed in terms of relative abundance  of Archaea or Bacteria phyla identified in the different samples?
5. At a lower taxonomic level (genus), analyze a group of prokaryotes of your choice and compare this taxonomic group between the different analyzed samples.
6. Identify differences in the structure of microbial communities inhabiting the Salt and Sediment in the Aveiro Salt Pan.
7. Characterize the distribution of cyanobacteria in the different sampling stations.
8. Answer your own question...