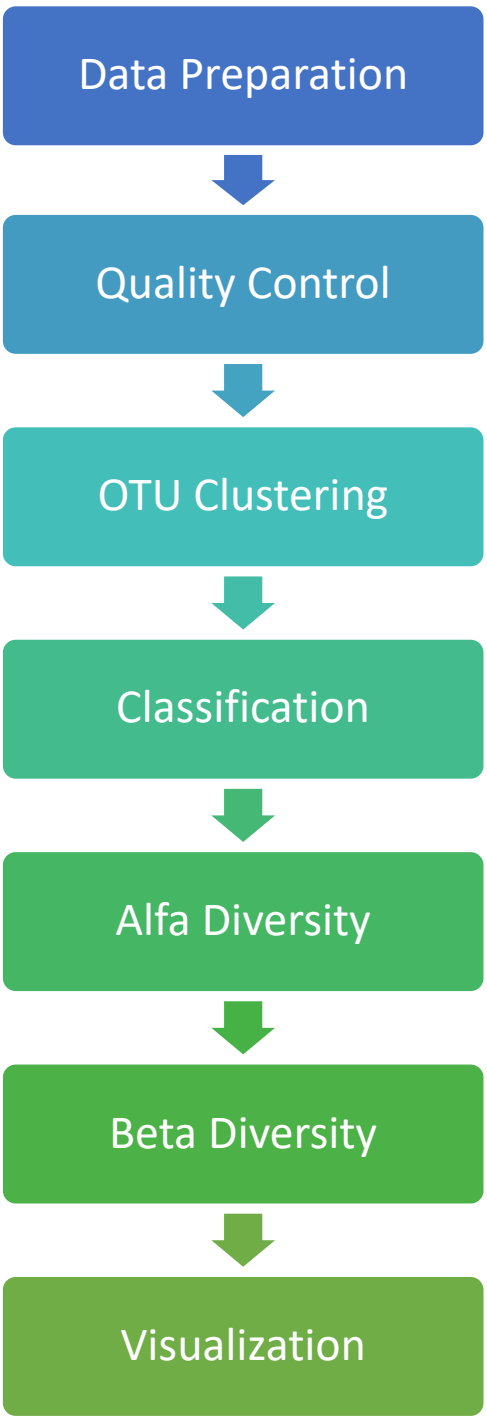
Student: Martí Díez Macià

PRACTICAL – EXIT TICKET

We applied this workflow in the practical. Fill it with the format file and software used in each step:



GALAXY TOOLS: .fastq, List of Dataset Pairs

GALAXY TOOLS: .fastq, Make.contigs

GALAXY TOOLS: .fasta, Cluster.split

GALAXY TOOLS: .list, Classify.otu

GALAXY TOOLS: .shared, Rarefaction.single

GALAXY TOOLS: .shared, Dist.shared

GALAXY TOOLS: .taxonomy, Taxonomy-to-Krona

Describe the experimental design of the data set used in the practical session

The Schloss lab collected daily fecal samples from mice for 365 days to study gut microbiome variation. We use a subset of this data from a single mouse at 20 different time points. In order to evaluate the error rates, a mock community with a known composition was sequenced alongside. The data set is organized into 20 pairs of files (40 .fastq files into pairs), with each pair representing a sample with its forward and reverse reads. We will use Galaxy to analyze this data and follow this process

How does the duplicate sequences removal works?

The process of removing duplicate sequences involves identifying unique reads, recording their occurrences, and optimizing files for computation. This process is key when we handle microbiome samples were many identical sequences are expected. It also helps speeding up computation and reduce disk space usage

Why is the OTU clustering one of the main steps on the analysis? What would you expect (in terms of diversity) if we apply an identity threshold < 97%?

OTU clustering groups similar 16S rRNA sequence reads into clusters, which reduces the data volume. This process helps manage the large amount of data generated. It is also an essential tool for taxonomic profiling, which is one of the main goals of this analysis. It allows for the identification and classification of organisms. We also group our sequences with real sequences, and it improves the accuracy of the analysis. If we apply an identity threshold lower than 97% we would expect larger diversity since a low identity threshold means more clusters being formed (each can represent a different species).

Explain the importance of rarefaction curves. If you have a sample which show a rarefaction curve which has not started to level off, how would you solve this problem? Give at least two possible solutions

Rarefaction curves help estimate species richness and assess sampling suitableness. If a curve hasn't leveled off, it indicates incomplete sampling of biodiversity. To address this, we could increase sampling effort by collecting more samples/sequencing more deeply or use subsampling techniques to reduce the impact of abundant species. With a rarefaction curve, the goal is to get a plateauing curve that indicates that the diversity of most species has been captured.