16S rDNA amplicon sequencing data analysis in Galaxy

Loïc Bourgeois, Amalia Soenens, Nuria Lozano, Juan Imperial
Centro de Biotecnología y Genómica de Plantas (CBGP), Universidad Politécnica de Madrid, Campus de Montegancedo, 28223 Pozuelo de Alarcón, Madrid, Spain

Background :

Most biologists can easily access NGS technologies and data in order to characterize the microbial diversity of a sample with 16S rDNA amplicon sequencing. However, the output of this kind of experiment can be challenging to handle. We assessed the different options to address 16S rDNA amplicon data analysis in Galaxy, and will highlight the benefits and drawbacks of the existing solutions. Indeed, even if the bioinformatics community now provides numerous tools allowing treatment of this sort of data, determining which software best fits the user's needs is not trivial. The choice of the software and the algorithms one should use is important, as it will impact the output of the experiment and relies on the characteristics of the data and the user experience. Here we present the different existing solutions to analyze 16S rDNA amplicon sequencing data inside Galaxy.

Tools implemented separately

There is not any pipeline in 16S rDNA amplicon sequencing data analysis accepted as the golden standard. However, there are several algorithms which have been developped separately, and that can be used together. They sometimes need some modifications of the inputs and/or outputs to be compatible. Some of them have been implemented directly in Galaxy, to be used together as a workflow and are available on the Galaxy Tool Shed. This workflow implemented on Galaxy by qfab, provides some of the most popular algorithms for this sort of analysis (uchime, PyNast, FastTree, RDP classifier, etc.). This is an easy way to install rapidly an efficient and easy-to-use pipeline for 16S rDNA analysis in Galaxy.
However, if one wants to try more options, there are toolsuites that have been developped specially for that, which implements the major part of the available tools for 16S rDNA amplicon sequencing data.

Mothur

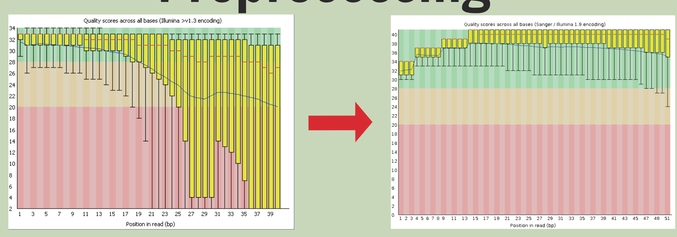
Mothur is the most cited tool suite for 16S rDNA amplicon sequencing data analysis and aims to reimplement the major part of the available tools for these analyses as a standalone in C. There are wrappers provided on the Galaxy Tool Shed which are up to date. However, Mothur implements a lot of different algorithms and all of them are not perfectly integrated.
For instance, the make.contigs tool in its command line version has the option for loading a file of file names (instead of passing manually each file one by one), which is not available in the Galaxy Tool Shed wrapper. Also, there are several tools missing inputs or outputs that are available on the command line version, such as count_table which is lacking in several tools.
Therefore, Mothur is a good solution for scientists looking for a more complete solution for 16S rDNA amplicon sequencing data analysis in Galaxy, but is still not completely integrated.

QIIME

QIIME is the second most cited tool (closely behind Mothur) for 16S rDNA amplicon sequencing data analysis. It aims at the same goal, but with a different philosophy. It is written in Python, and instead of re-implementing the different algorithms, it uses them directly as they are and modify the outputs to make them compatible between each of them. As a result, it requires a lot of dependencies and can be tedious to install.
There is no wrappers that are up to date on the Galaxy Tool Shed for QIIME.
However, a project in collaboration with the QIIME development team succeeded to develop a script that generates wrappers automatically for the QIIME tools. This is based on metadata of the tools about the inputs and output provided by the QIIME developpers. This is a really interesting way of implementing tools in Galaxy, as it can possibly reduce the wrapper creation effort significatively.

Analysis core steps

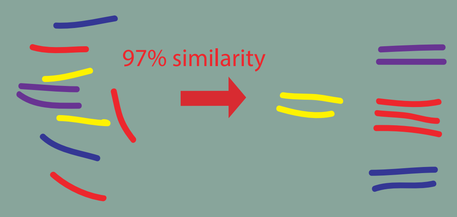
Preprocessing



During this step, the data is cleaned and sequencing errors are reduced

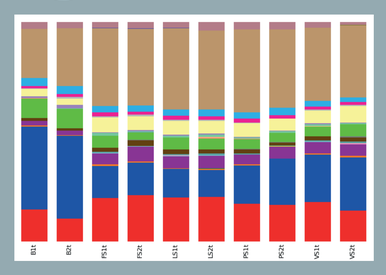
OTUs clustering

Sequences are clustered according to their similarity to estimate the number of different strains in the sample, the consensus is 97% similarity.



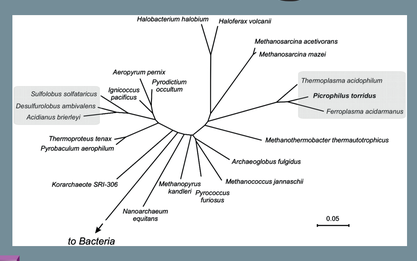
Taxonomy assignment

The previously clustered sequences are represented by a unique sequence. This representative sequences will be compared to a reference database to assign a taxonomy to each of them.



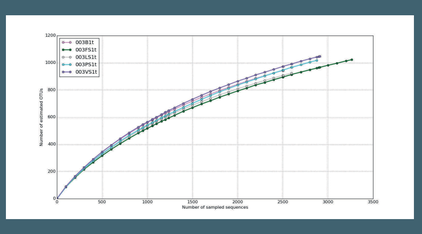
Phylogenetic tree building

Based on the sequences, a phylogenetic tree is built. This tree can be used in downstream analysis to calculate beta diversity by providing distance metrics between the sequences.



Diversity analysis

Diversity analyses allow to estimate the species richness and evenness in the studied sample.



Conclusion and discussions :

When using QIIME and/or Mothur with the command line interface, the tools are directly fully operational. However, when implementing tools in Galaxy, there is a development effort which is required and that can be important. Mothur and QIIME are implemented differently, and have some flaws in different tools. They can be both complementary, (e.g. for visualization) but it is not always the case due to some specificities of the algorithms, such as information storage in the read names. As a result, 16S rDNA amplicon sequencing data analysis on Galaxy works nicely for the standard operating procedures, but is still lacking the implementation of some tools and options.

The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Supercomputing and Visualization Center of Madrid (CeSViMa).