Microbiome data and related metadata at user’s fingertips: aiming to an automated tool for effortless retrievement and reconstruction

Soletta G.[1], Agostinetto G.[1], Fumagalli S.[1], Bruno A.[1]

E-mail: g.soletta@campus.unimib.it

[1] Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan (Italy)

Keywords: microbiome data, metadata, ENA, bioinformatics, omics data

Abstract: Publicly available repositories, as the European Nucleotide Archive (ENA), provide huge amounts of microbiome-related data derived from high-throughput DNA sequencing studies. These datasets are a crucial resource for meta-analyses and secondary analyses, leading to novel discoveries, data reuse approaches and data mining applications.

In this work, we want to present a currently in development tool for downloading sequencing data and related metadata to enhance microbiome re-analyses strategies. The aim of the work is to automate the process of data retrieval, creating an easy-to-use framework to download data in a standardized format, according to FAIR principles, and reconstruct both submission and studies metadata. The tool is mainly written in Python and it implements a set of scripts for simplified programmatic access to ENA repository. Overall, three main step were developed: 1) Download of reads from ENA in gzipped FASTQ format, given a list of microbiome projects; 2) Check of the downloaded files - through associated metadata, the tool performs a double-check step, in order to completely guarantee the further analysis; 3) Metadata retrieval for each project, considering both submission and sample metadata. Metadata reconstruction is now under development, as it is a crucial step to easily allow the secondary analysis.
Currently, the SKIOME Project was used as a case study to test the tool, taking the manually curated list of projects as a starting point.

Looking ahead, our tool may be included into larger workflows for processing, analyzing, and visualizing microbiome data, also implementing an automatic integration into the QIIME2 framework for 16S rRNA amplicon studies.