SKIOME collection 2.0: coupling a novel framework and manual curation step for enriching skin microbiome datasets

Bulla N.[1], Fumagalli S.[1], Fontana S.[2], Ghisleni G.[1], Bruno A.[1], Casiraghi M.[1], Labra M.[1]

E-mail: n.bulla@campus.unimib.it

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

[2] School of Law, University of Milano-Bicocca, Milan (Italy)

Keywords: skin microbiome, metadata, datasets, collection, ENA, NCBI

Abstract: An exponential amount of data from microbiome-related studies is being stored in public repositories like the European Nucleotide Archive (ENA) and the National Center for Biotechnology Information (NCBI). These datasets constitute a significant resource for conducting meta-analysis and implementing machine-learning methodologies. Despite efforts to establish guidelines, the lack of metadata standardization and missing information poses obstacles to data interpretability and usability.

Starting from SKIOME Project, we have developed a novel framework aimed at maximizing the retrieval of amplicon sequencing and whole-genome sequencing (WGS) based projects related to the human skin microbiome. The framework involves querying and downloading datasets from ENA using MADAME, a metadata retrieval tool, and from NCBI via both the SRAdb database in R and the pysradb library in Python.

Applying the developed framework, we have compiled a comprehensive collection of datasets, encompassing over 550 projects and more than 120000 samples deposited since 2012. We integrated these datasets using in-house Python scripts and curated the data by selectively including samples from human skin microbiome, focusing on amplicon and WGS sequencing. Moreover, we performed an enrichment step through EntrezDirect from NCBI and, utilizing Madame, retrieved publications essential for the curation process.

In the upcoming weeks, the manual curation step will be performed as an engaging and educational activity involving this university's students. Thus, SKIOME collection 2.0, meticulously curated, will soon become accessible online through a dedicated website. The platform will simplify datasets download and provide the necessary resources for seamless secondary analysis. In conclusion, in alignment with the FAIR principles, SKIOME collection 2.0 facilitates the accessibility of metadata, aiming to contribute to the advancement of future human skin microbiome research.