In-House Tools for Public Meta-Omics Metadata Curation

Samuel Piquer-Esteban[1,2], Wladimiro Diaz-Villanueva[1], Vicente Arnau[1], Andrés Moya[1,2,3]

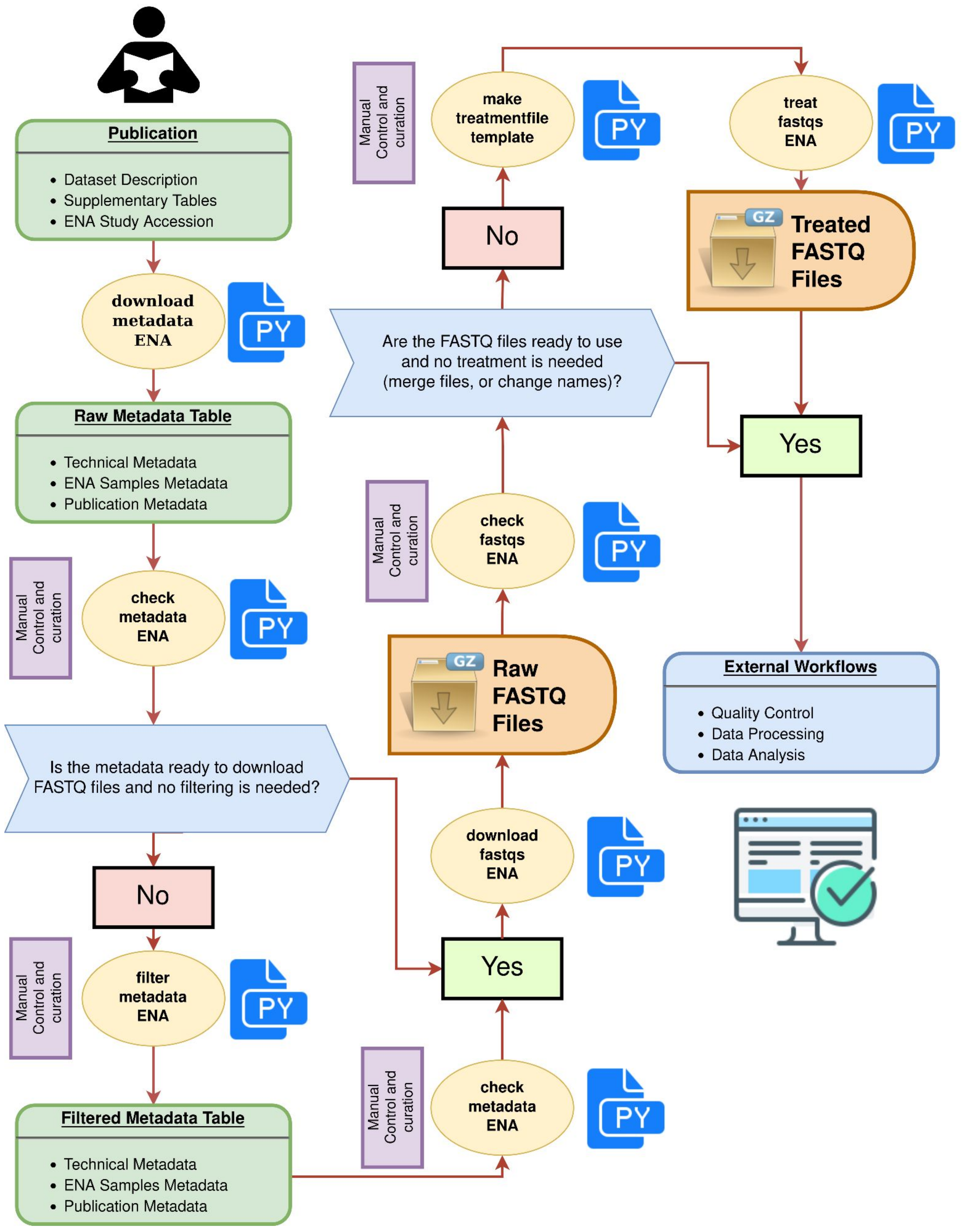
1. Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2. Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
3. Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain

INTRODUCTION

One of the key elements of any meta-omics study is its metadata. However, the metadata curation of public studies remains a costly and challenging endeavor despite its importance. Even though some initiatives support data reuse, most are focused on human metagenomic data and generally do not allow processing with external pipelines (1-6). Furthermore, their metadata is usually incomplete or scattered, which often leads to the need for in-house curation and processing. Currently, to the best of our knowledge, there are no tools designed to facilitate, accompany and guide the researcher during this manual curation process.

To address this need, we present a collection of scripts developed to extract and curate metadata and FASTQ files associated with research projects hosted in the European Nucleotide Archive (ENA).

WORKFLOW



Publication
- Dataset Description
- Supplementary Tables
- ENA Study Accession

download metadata ENA — PY

Raw Metadata Table
- Technical Metadata
- ENA Samples Metadata
- Publication Metadata

Manual Control and curation — check metadata ENA — PY

Is the metadata ready to download FASTQ files and no filtering is needed?

No

Manual Control and curation — filter metadata ENA — PY

Filtered Metadata Table
- Technical Metadata
- ENA Samples Metadata
- Publication Metadata

Manual Control and curation — check metadata ENA — PY

Yes

download fastqs ENA — PY

Raw FASTQ Files

Manual Control and curation — check fastqs ENA — PY

Are the FASTQ files ready to use and no treatment is needed (merge files, or change names)?

No

Manual Control and curation — make treatmentfile template — PY

treat fastqs ENA — PY

Treated FASTQ Files

Yes

External Workflows
- Quality Control
- Data Processing
- Data Analysis

CONCLUSIONS & NEAR FUTURE

- We have reviewed more than 200 ENA projects and applied that acquired knowledge to specially **designed a collection of scripts to accompany and guide the researcher during the in-house curation process of meta-omics data**.
- In the near future, **we intend to convert these scripts into a Python package and make them available to the general public**.
- Our ultimate goal is to **provide a uniform and reproducible workflow to simplify the curation of public sequencing data**.

CONTACT

Samuel.Piquer@uv.es

DESCRIPTION

download metadata ENA

This program allows downloading the metadata associated with a study project by collecting the information available in ENA Browser and using the mg-tooolkit package (https://pypi.org/project/mg-toolkit/). Additional metadata from the publication can also be provided in an alternative mode.

check metadata ENA

This program enables to analyze and carry out some checks of interest on the previously generated metadata tables at different points of the workflow.

filter metadata ENA

This program allows different filtering operations to be performed in a sequential manner on the previously generated metadata table based on the provided filtering information.

download fastqs ENA

This program allows downloading the FASTQ files associated with a study project using the parfive package (https://pypi.org/project/parfive/).

check fastqs ENA

This program enables to analyze and carry out some checks of interest on the downloaded FASTQ files.

make treatmentfile template

This program allows to generate a raw treatmentfile template using the information available in the previously generated metadata table and the downloaded FASTQ files.

treat fastqs ENA

This program allows different treatment operations (merge, change names, and copy) to be performed on the downloaded FASTQ files based on the treatment information provided.

Program Type	Main Programs	Control Check Programs	Extra Treatment Programs

ACKNOWLEDGMENTS

This work was supported by MICINN-PID2019-105969GB-I00, CIPROM/2021/042 and CIBEResp. SPE was funded by an FPU fellowship from the Spanish Ministry of Universities (Reference: FPU20/05756).

REFERENCES

1. Mitchell A. L. et al. (2020). Nucleic acids research, 48(D1), D570-D578.
2. Oliveira F. S. et al. (2018). Nucleic acids research, 46(D1), D684-D691.
3. Pasolli E. et al. (2017). Nature methods, 14(11), 1023-1024.
4. Shao L. et al. (2021). BMC microbiology, 21(1), 1-12.
5. Wu S. et al. (2020). Nucleic acids research, 48(D1), D545-D553.
6. Zhang Q. et al. (2021).Briefings in Bioinformatics, 22(3), bbaa082.