Metadata retrieval from sequence databases with *ffq*

Ángel Gálvez-Merchán[1], Kyung Hoi (Joseph) Min[2], Lior Pachter[1,3]* and A. Sina Booeshaghi[4]*

[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA
[2]Department of Computer Science and Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts
[3]Department of Computing and Mathematical Sciences, Pasadena, CA
[4]Department of Mechanical Engineering, California Institute of Technology, Pasadena, CA
*Address correspondence to: lpachter@caltech.edu and abooesha@caltech.edu

Abstract

We present a command-line tool, called *ffq*, for querying user-generated data and metadata from sequence databases. The code can be found here: https://github.com/pachterlab/ffq.

Introduction

The extraordinary large volume of user-generated sequencing data available in public databases is increasingly being utilized in research projects alongside novel experiments (Simon *et al.*, 2018; Razmara *et al.*, 2019; Lung *et al.*, 2020; Rajesh *et al.*, 2021; Hippen and Greene, 2021; Wartmann *et al.*, 2021; Kasmanas *et al.*, 2021; Huang *et al.*, 2021; Klie *et al.*, 2021; Booeshaghi *et al.*, 2022). Collation of metadata is crucial for such reuse of publicly available data since it can provide information about the samples assayed and can facilitate the acquisition of raw data. For example, *sra-tools* enables users to query and download data from the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA), which currently hosts 13.67 PB of data. An alternative to *sra-tools* is the *pysradb* tool (Choudhary, 2019). *pysradb* was developed to access metadata from the Sequence Read Archive (SRA), using metadata obtained from the regularly updated SRAdb SQLite database (Zhu *et al.*, 2013). MetaSRA adds additional standardized metadata on top of the SRAdb SQLite database (Bernstein *et al.*, 2017) and also provides an API for accessing them. While these and other tools (Mahi *et al.*, 2019; Li *et al.*, 2018; Eaton, 2020; Bernstein *et al.*, 2020) have proven to be very useful, they are limited in terms of the scope of databases they provide access to. We developed *ffq* to facilitate metadata retrieval from a diverse set of databases, including

1. National Center for Biotechnology Information Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO),
2. European Molecular Biology Lab-European Bioinformatics Institute European Nucleotide Archive (EMBL-EBI ENA),
3. DNA Data Bank of Japan Gene Expression Archive (DDBJ GEA), and
4. Encyclopedia of DNA Elements (ENCODE) database (Davis *et al.*, 2018; ENCODE Project Consortium, 2012).

In order to facilitate a modular architecture for *ffq*, we first studied the structure of these databases in detail to identify commonalities and relationships between them (Figure 1).

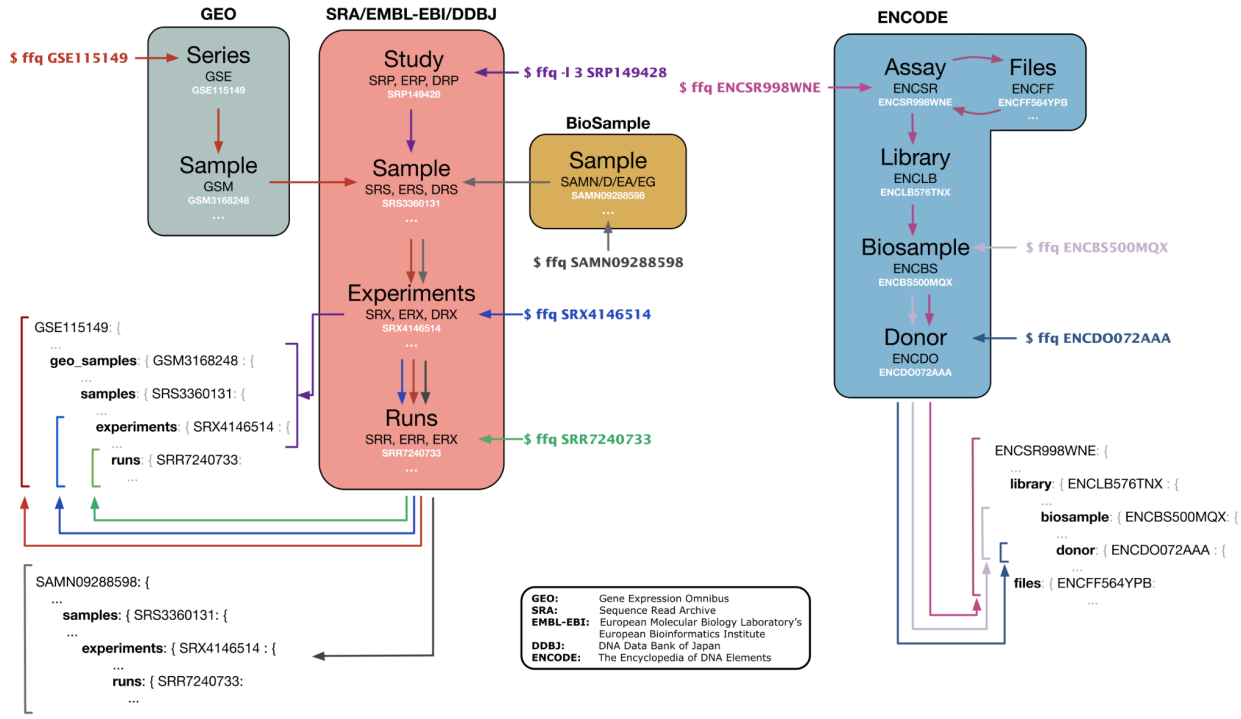

Figure 1: Metadata retrieval. *ffq* fetches and returns metadata as a JSON object by traversing the database hierarchy. Subsets of the database hierarchy can be returned by specifying *-l [level]*.

The SRA, ENA, and DDBJ databases all follow a similar hierarchical structure where studies are grouped into samples, experiments, and runs, a shared architecture that is useful and likely the result of the longstanding International Nucleotide Sequence Database Collaboration (INSDC) between the ENA, NCBI, and DDBJ. We note that the Genome Sequence Archive (GSA) (Chen *et al.*, 2021; CNCB-NGDC Members and Partners, 2022) is not a member of the INSDC. However it also uses a similar hierarchical structure for its database, and regularly ingests data from the SRA, but does not expose its publicly available data for programmatic access.

The consistent database schemas used by members of the INSDC greatly simplifies metadata retrieval for *ffq*. For example, GEO accession codes are grouped hierarchically through Series and Samples and have external relations to SRA accession codes for raw sequencing data submitted to the SRA. This enables *ffq* to fetch metadata and processed data from GEO that submitters have associated with raw sequencing data stored in the SRA.

Description

Based on the database architectures, we created *ffq* to fetch metadata using database accessions or paper DOIs as input. Importantly, *ffq* only fetches metadata and links to data files and does not offer data downloading. This deliberate design decision was motivated by the UNIX philosophy “Make each program do one thing well” (McIlroy *et al.*, 1978).

The *ffq* options are summarized below:

- *ffq* [accession(s)]

    - Where [accession] can be any of the following: SR(R/X/S/P), ER(R/X/S/P), DR(R/X/S/P), GS(E/M), ENC(SR/BS/DO), CXR, SAM(N/D/EA/EG), DOI.
- *ffq* [-l level] [accession(s)]
    - Where [level] defines the hierarchy in the database to which data is subset data.
- *ffq* [--ftp] [--aws] [--gcp] [--ncbi] [accession(s)]
    - Where the flags correspond to the types of data-storage links for the raw data.
- *ffq* [-o out] [--split] [accession(s)]
    - Where [out] corresponds to a path on disk to save the JSON file and [--split] splits the metadata from multiple accessions into their own file.

Accession-based *ffq* metadata retrieval uses the NCBI's Entrez programming utilities, ENA's API, GEO's FTP, and ENCODE's API to programmatically access metadata with HTTP requests. DOI-based metadata retrieval first converts the DOI to the manuscript tile via the CrossRef API (Hendricks *et al.*, 2020) and then retrieves all study accessions associated with the manuscript title with the ENA search API. Metadata is returned as a Javascript Object Notation (JSON) object. Run times for metadata retrieval vary depending on database up-time, server connection speed, and database rate-limiting, but generally we find that *ffq* can download metadata at a rate of 10s per sample. This rate includes short and deliberate delays we have added between HTTP requests to prevent a perceived Denial-of-Service.

Usage and Documentation

The *ffq* tool is written in Python and can be installed with pip and conda. It has four dependencies and undergoes quality control via an automated testing framework that validates behavior against three Python versions (3.6, 3.7, and 3.8) covering 88% of the code. The JSON return objects make *ffq* interoperable with other tools such as *jq* for easy command-line parsing. Additionally, *ffq*'s modularity and simplicity make it extensible to other genomic databases. By leveraging existing APIs, *ffq* offers a lightweight solution for querying data that is guaranteed to be more up-to-date than tools that rely on regular database builds.

Discussion

While *ffq* facilitates downloading of data from numerous genomic databases, the results retrieved are only useful to the extent that the metadata uploaded is meaningful and complete. Meaningful and complete user-generated data underlies the curation of genomic references essential for comparative genomic data analysis (Luebbert and Pachter, 2022). Unfortunately, there is little to no standardization of user-uploaded sequencing metadata (Wang *et al.*, 2019; Rajesh *et al.*, 2021), and metadata descriptions can become exceedingly complex for current multiplexed experiments where different assays with distinct data types are combined. Improvement of metadata uploading in machine-readable standard formats is essential if publicly available genomic data are to be usable by scientists in the future.

Acknowledgments

This work was motivated by the need to obtain metadata for (Booeshaghi and Pachter, 2020). We thank Ali Mortazavi for his suggestion to include *ffq* querying of the ENCODE database.

Funding

This work was funded in part by NIH U19MH114830.

References

Bernstein,M.N. *et al.* (2020) Jupyter notebook-based tools for building structured datasets from the Sequence Read Archive. *F1000Res.*, **9**, 376.

Bernstein,M.N. *et al.* (2017) MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, **33**, 2914–2923.

Chen,T. *et al.* (2021) The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics Proteomics Bioinformatics*, **19**, 578–583.

Choudhary,S. (2019) pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. *F1000Res.*, **8**, 532.

CNCB-NGDC Members and Partners (2022) Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2022. *Nucleic Acids Res.*, **50**, D27–D38.

Davis,C.A. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.

Eaton,K. (2020) NCBImeta: efficient and comprehensive metadata retrieval from NCBI databases. *J. Open Source Softw.*, **5**, 1990.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Hendricks,G. *et al.* (2020) Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, **1**, 414–427.

Hippen,A.A. and Greene,C.S. (2021) Expanding and Remixing the Metadata Landscape. *Trends Cancer Res.*, **7**, 276–278.

Huang,Y.-N. *et al.* (2021) The systematic assessment of completeness of public metadata accompanying omics studies. *bioRxiv*, 2021.11.22.469640.

Kasmanas,J.C. *et al.* (2021) HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. *Nucleic Acids Res.*, **49**, D743–D750.

Klie,A. *et al.* (2021) Increasing metadata coverage of SRA BioSample entries using deep learning-based named entity recognition. *Database* , **2021**.

Li,Z. *et al.* (2018) GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata. *Database* , **2018**.

Luebbert,L. and Pachter,L. (2022) Efficient querying of genomic databases for single-cell RNA-seq with gget. *bioRxiv*, 2022.05.17.492392.

Lung,P.-Y. *et al.* (2020) Maximizing the reusability of gene expression data by predicting missing metadata. *PLoS Comput. Biol.*, **16**, e1007450.

Mahi,N.A. *et al.* (2019) GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. *Sci. Rep.*, **9**, 7580.

McIlroy,M. *et al.* (1978) UNIX time-sharing system. *The Bell system technical journal*, **57**, 1899–1904.

Rajesh,A. *et al.* (2021) Improving the completeness of public metadata accompanying omics studies. *Genome Biol.*, **22**, 106.

Razmara,A. *et al.* (2019) recount-brain: a curated repository of human brain RNA-seq datasets metadata. *bioRxiv*, 618025.

Simon,L.M. *et al.* (2018) MetaMap, an interactive webtool for the exploration of metatranscriptomic reads in human disease-related RNA-seq data. *bioRxiv*, 425439.

Booeshaghi,A. *et al.* (2022) Depth normalization for single-cell genomics count data. *bioRxiv*, 2022.05.06.490859.

Booeshaghi,A. and Pachter,L. (2020) Decrease in ACE2 mRNA expression in aged mouse lung. *bioRxiv*, 2020.04.02.021451.

Wang,Z. *et al.* (2019) Mining data and metadata from the gene expression omnibus. *Biophys. Rev.*, **11**, 103–110.

Wartmann,H. *et al.* (2021) Bias-invariant RNA-sequencing metadata annotation. *Gigascience*, **10**.

Zhu,Y. *et al.* (2013) SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, **14**, 19.