The Journal of Open Source Software

NCBImeta: efficient and comprehensive metadata retrieval from NCBI databases

Katherine Eaton[1, 2]

1 McMaster Ancient DNA Centre, McMaster University **2** Department of Anthropology, McMaster University

DOI: 10.21105/joss.01990

Software

- Review
- Repository
- Archive

Editor: Lorena Pantano

Reviewers:

- @coughls
- @druvus

Submitted: 21 December 2019
Published: 03 February 2020

Summary

`NCBImeta` is a command-line application that downloads and organizes biological metadata from the National Centre for Biotechnology Information (NCBI). While the NCBI web portal provides an interface for searching and filtering molecular data, the output offers limited options for record retrieval and comparison on a much larger and broader scale. `NCBImeta` tackles this problem by creating a reformatted local database of NCBI metadata based on user search queries and customizable fields. The output of `NCBImeta`, optionally a SQLite database or text file(s), can then be used by computational biologists for applications such as record filtering, project discovery, sample interpretation, and meta-analyses of published work.

Background

Recent technological advances in DNA sequencing have propelled biological research into the realm of big data. Due to the tremendous output of Next Generation Sequencing (NGS) platforms, numerous fields have transformed to explore this novel high-throughput data. Projects that quickly adapted to incorporate these innovative techniques included monitoring the emergence of antibiotic resistance genes (Zankari et al., 2012), epidemic source tracking in human rights cases (Eppinger et al., 2014), and global surveillance of uncharacterized organisms (Connor et al., 2015). However, the startling rate at which sequence data are being deposited online have presented significant hurdles to the efficient reuse of published data. In response, there is growing recognition within the computational community that effective data mining techniques are a dire necessity (Mackenzie, McNally, Mills, & Sharples, 2016; Nakazato, Ohta, & Bono, 2013).

An essential step in the data mining process is the efficient retrieval of comprehensive metadata. These metadata fields are diverse in nature, but often include the characteristics of the biological source material, the composition of the raw data, the objectives of the research initiative, and the structure of the post-processed data. Several software applications have been developed to facilitate bulk metadata retrieval from online repositories. Of the available tools, `SRAdb` (Zhu, Stephens, Meltzer, & Davis, 2013), the Pathogen Metadata Platform (Chang, Peterson, Garay, & Korves, 2016), `MetaSRA` (Bernstein, Doan, & Dewey, 2017), and `pysradb` (Choudhary, 2019) are among the most widely utilised and actively maintained. While these software extensions offer substantial improvements over the NCBI web browser experience, there remain several outstanding issues.

1. Existing tools assume external programming language proficiency (ex. R, Python, SQL), thus reducing tool accessibility.

2. Available software focuses on implementing access to singular NCBI databases in isolation, for example, the raw data repository the Sequence Read Archive (SRA). This does not empower researchers to incorporate evidence from multiple databases, as it fails to fully leverage the power of interconnected information within the relational database scheme of NCBI.
3. Existing software provides only intermittent database updates, where users are dependent on developers releasing new snapshots to gain access to the latest information. This gives researchers less autonomy over what data they may incorporate as newer records are inaccessible, and may even introduce sampling bias depending on when the snapshots are generated.

In response, `NCBImeta` aims to provide a more user-inclusive experience to metadata retrieval, that emphasizes real-time access and provides generalized frameworks for a wide variety of NCBI's databases.

NCBImeta

`NBCImeta` is a command-line application that executes user queries and metadata retrieval from the NCBI suite of databases. The software is written in Python 3, using the `BioPython` module (Cock et al., 2009) to connect to, search, and download XML records with NCBI's E-Utilities (Kans, 2019). The `lxml` package is utilised to perform XPath queries to retrieve nodes containing biological metadata of interest. `SQLite` is employed as the database management system for storing fetched records, as implemented with the `sqlite3` python module. Accessory scripts are provided to supply external annotation files, to join tables within the local database so as to re-create the relational database structure, and finally to export the database as tabular text for downstream analyses. `NCBImeta` currently interfaces with the molecular and literature databases described in Table 1 (*Entrez Help*, 2016).

Table 1: NCBI databases supported in NCBImeta.

Database	Description
Assembly	Descriptions of the names and structure of genomic assemblies, statistical reports, and sequence data links.
BioSample	Characteristics of the biological source materials used in experiments.
BioProject	Goals and progress of the experimental initiatives, originating from an individual organization or a consortium.
Nucleotide	Sequences collected from a variety of sources, including GenBank, RefSeq, TPA and PDB.
PubMed	Bibliographic information and citations for biomedical literature from MEDLINE, life science journals, and other online publications.
SRA	Composition of raw sequencing data and post-processed alignments generated via high-throughput sequencing platforms.

The typical workflow of `NCBImeta` follows four major steps as outlined in Figure 1. Users first configure the program with their desired search terms. `NCBImeta` is then executed on the command-line to fetch relevant records and organize them into a local database. Next, the user optionally edits the database to, for example, add their own custom metadata. Finally, the resulting database, kept in SQLite format or exported to text, delivers 100+ biologically-relevant metadata fields to researcher's fingertips. This process not only saves significant time compared to manual record retrieval through the NCBI web portal, but additionally unlocks attributes for comparison that were not easily accessible via the web-browser interface.

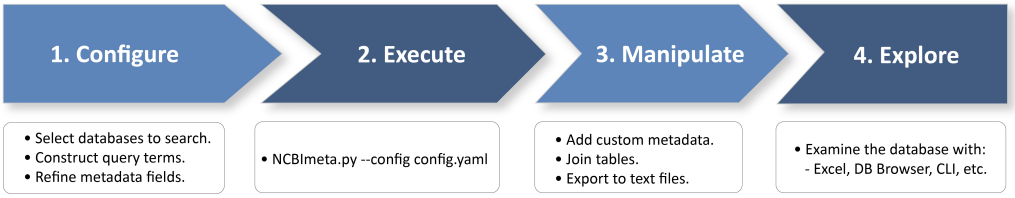

Figure 1: NCBImeta user workflow.

`NCBImeta`'s implementation offers a novel approach to metadata management and presentation that improves upon the prevously described limitations of existing software in a number of ways. First, `NCBImeta` is run on the command-line, and the final database can be exported to a text file, thus no knowledge of an external programming language is required to generate or explore the output. Second, a general parsing framework for tables and metadata fields was developed which can be extended to work with diverse database types contained within NCBI's infrastructure. Finally, a query system was implemented for record retrieval that allows users to access records in real-time, as opposed to working with intermittent or out-dated database snapshots.

Use Case

The following section demonstrates how `NCBImeta` can be used to obtain current and comprehensive metadata for a pathogenic bacteria, *Pseudomonas aeruginosa*, from various sequencing projects across the globe. *P. aeruginosa* is an opportunistic pathogen associated with the disease cystic fibrosis (CF) and is highly adaptable to diverse ecological niches (Stewart et al., 2014). As such, it is a target of great interest for comparative genomics and there are currently over 15,000 genomic sequence records available which are spread across two or more databases. In cases such as this, it is critical to leverage the tremendous power of these existing datasets while being conscious of the labor typically required to retrieve and contextualize this information. `NCBImeta` renders the problem of acquiring and sifting through this metadata trivial and facilitates the integration of information from multiple sources.

To identify publicly available *P. aeruginosa* genomes, `NCBImeta` is configured to search through the tables *Assembly* (assembled genomes) and *SRA* (raw data). For additional context, `NCBImeta` is used to retrieve metadata from the *Nucleotide* table for descriptive statistics of the genomic content, from the *BioProject* table to examine the research methodology of the initiative, from *Pubmed* to identify existing publications, and finally from the *Biosample* table to explore characteristics of the biological material. A small subset of the 100+ retrieved columns is shown in Figure 2, to provide a visual example of the output format and the metadata that is retrieved.

	Organism	Strain	Date	Location	HostDisease	Source	LatLon	Status	Contig	Length	LibrarySelection	Platform
1	Pseudomonas aeruginosa	BK4	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	90	6409337	PCR	ILLUMINA
2	Pseudomonas aeruginosa	CLJ1	05-May-2010	France: Grenoble	Chronic obstructive p...	lungs (tracheal as...	45.199444 N 5....	Scaffold	78	6514464	unspecified	PACBIO_SMRT;ILLUN
3	Pseudomonas aeruginosa	BK2	2010	India: Madurai	Keratitis	cornea	9.93 N 78.12 E	Scaffold	63	6386147	PCR	ILLUMINA
4	Pseudomonas aeruginosa	PA121617	04-Jun-2012	China: Guangzhou	Respiratory disease	sputum	23.0538554170...	Complete ...	2	6853510	RANDOM	PACBIO_SMRT;ILLUN
5	Pseudomonas aeruginosa	TUEPA7472	2015	Germany:Tuebingen	Pseudomonas aerugi...	blood	48.532072 N 9....	Scaffold	19	6806824	PCR	PACBIO_SMRT;ILLUN
6	Pseudomonas aeruginosa	BK6	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	172	7056854	PCR	ILLUMINA
7	Pseudomonas aeruginosa	BK3	2013	India: Madurai	Keratitis	cornea of keratitis...	9.93 N 78.12 E	Scaffold	143	7194702	PCR	ILLUMINA
8	Pseudomonas aeruginosa	CLJ3	17-May-2010	France: Grenoble	Chronic obstructive p...	lungs (tracheal as...	45.199444 N 5....	Contig	135	6353571	unspecified	ILLUMINA
9	Pseudomonas aeruginosa	PAL0.1	2016	France: Lille	Pneumonia	lung	50.38 N 3.03 E	Contig	131	7040354	Hybrid Selection	ILLUMINA
10	Pseudomonas aeruginosa	BK5	2013	India: Madurai	Keratitis	cornea from kerat...	9.93 N 78.12 E	Scaffold	104	6364667	PCR	ILLUMINA
11	Pseudomonas aeruginosa	24Pae112	2015-03-05	Colombia	Sepsis	blood	4.814278 N 75....	Complete ...	1	7097241	size fractionation	PACBIO_SMRT
12	Pseudomonas aeruginosa	PA_D22	21-Mar-2014	China: Nanning	Ventilator associated ...	Sputum; Late isol...	22.817 N 108.3...	Complete ...	1	6681981	size fractionatio...	PACBIO_SMRT;ILLUN
13	Pseudomonas aeruginosa	PA_D21	10-Mar-2014	China: Guangxi	Ventilator associated ...	Sputum; Late isol...	22.8167 N 108....	Complete ...	1	6639108	size fractionatio...	PACBIO_SMRT;ILLUN
14	Pseudomonas aeruginosa	PA_D16	06-Mar-2014	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6681975	size fractionatio...	PACBIO_SMRT;ILLUN
15	Pseudomonas aeruginosa	PA_D9	21-Jan-2014	China: Nanning	Ventilator associated ...	Sputum; Late isol...	22.817 N 108.3...	Complete ...	1	6645477	size fractionatio...	PACBIO_SMRT;ILLUN
16	Pseudomonas aeruginosa	PA_D5	13-Jan-2014	China: Guangxi	Ventilator associated ...	Sputum; Early iso...	22.8167 N 108....	Complete ...	1	6681992	size fractionatio...	PACBIO_SMRT;ILLUN
17	Pseudomonas aeruginosa	PA_D2	24-Dec-2013	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6642996	size fractionatio...	PACBIO_SMRT;ILLUN
18	Pseudomonas aeruginosa	PA_D1	14-Dec-2013	China: Nanning	Ventilator associated ...	Sputum; Early iso...	22.817 N 108.3...	Complete ...	1	6643823	size fractionatio...	PACBIO_SMRT;ILLUN

Figure 2: A subset of the 100+ metadata columns retrieved for *P. aeruginosa* sequencing projects. Viewed with DB Browser for SQLite (https://sqlitebrowser.org/)

Subsequently, the output of NCBImeta can be used for exploratory data visualization and analysis. The text file export function of NCBImeta ensures downstream compatibility with both user-friendly online tools (ex. Google Sheets Charts) as well as more advanced visualization packages (Wickham, 2016). In Figure 3, the geospatial distribution of *P. aeruginosa* isolates are plotted alongside key aspects of genomic composition (ex. number of genes).

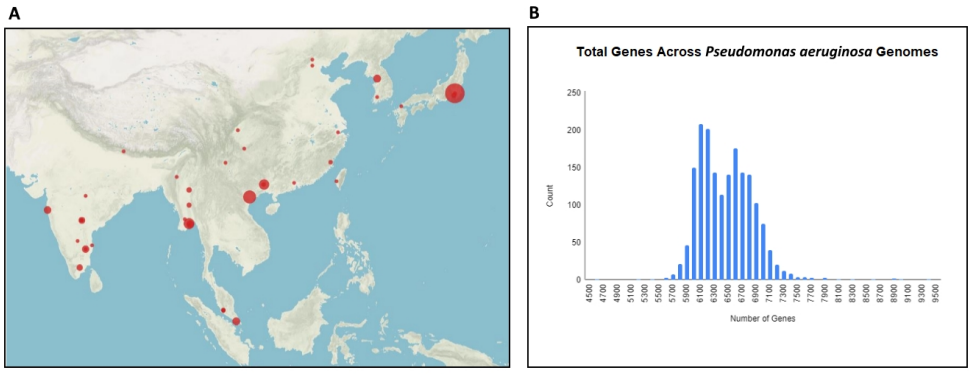

Figure 3: Metadata visualization of *P. aeruginosa* sequencing projects. A) The geographic distribution of samples in this region highlights a large number originating in Japan. Visualized with Palladio (https://hdlab.stanford.edu/palladio/). B) The number of genes per organism reveals a multi-modal distribution within the species.

Finally, NCBImeta can be used to streamline the process of primary data acquisition following careful filtration. FTP links are provided as metadata fields for databases attached to an FTP server (ex. Assembly, SRA) which can be used to download biological data for downstream analysis.

Future Work

The development of `NCBImeta` has primarily focused on a target audience of researchers whose analytical focus is prokaryotic genomics and the samples of interest are the organisms themselves. Chief among those include individuals pursuing questions concerning epidemiology, phylogeography, and comparative genomics. Future releases of `NCBImeta` will seek

to broaden database representation to include gene-centric and transcriptomics research (ex. NCBI's Gene and GEO databases).

Availability

NCBImeta is a command-line application written in Python 3 that is supported on Linux and macOS systems. It is distributed for use under the OSD-compliant MIT license (https://opensource.org/licenses/MIT). Source code, documentation, and example files are available on the GitHub repository (https://github.com/ktmeaton/NCBImeta).

Acknowledgements

I would like to thank Dr. Hendrik Poinar and Dr. Brian Golding for their guidance and support on this project, as well as for insightful conversations regarding biological metadata, the architecture of NCBI, and software deployment. Thank you to Dr. Andrea Zeffiro, Dr. John Fink, Dr. Matthew Davis, and Dr. Amanda Montague for valuable discussions regarding APIs, digital project management, and software publishing. Thank you to all past and present members of the McMaster Ancient DNA Centre and the Golding Lab. This work was supported by the MacDATA Institute (McMaster University, Canada) and the Social Sciences and Humanities Research Council of Canada (#20008499).

References

Bernstein, M. N., Doan, A., & Dewey, C. N. (2017). MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, *33*(18), 2914–2923. doi:10.1093/bioinformatics/btx334

Chang, W. E., Peterson, M. W., Garay, C. D., & Korves, T. (2016). Pathogen metadata platform: Software for accessing and analyzing pathogen strain information. *BMC Bioinformatics*, *17*(1). doi:10.1186/s12859-016-1231-2

Choudhary, S. (2019). Pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive. *F1000Research*, *8*, 532. doi:10.12688/f1000research.18676.1

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. doi:10.1093/bioinformatics/btp163

Connor, T. R., Barker, C. R., Baker, K. S., Weill, F.-X., Talukder, K. A., Smith, A. M., Baker, S., et al. (2015). Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *eLife*, *4*. doi:10.7554/eLife.07335

Entrez Help. (2016). Bethesda, MD: National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK3837/

Eppinger, M., Pearson, T., Koenig, S. S. K., Pearson, O., Hicks, N., Agrawal, S., Sanjar, F., et al. (2014). Genomic epidemiology of the Haitian cholera outbreak: A single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. *mBio*, *5*(6). doi:10.1128/mBio.01721-14

Kans, J. (2019). Entrez Direct: E-utilities on the UNIX Command Line. In *Entrez Programming Utilities Help*. National Center for Biotechnology Information. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK179288/

Mackenzie, A., McNally, R., Mills, R., & Sharples, S. (2016). Post-archival genomics and the bulk logistics of DNA sequences. *BioSocieties*, *11*(1), 82–105. doi:10.1057/biosoc.2015.22

Nakazato, T., Ohta, T., & Bono, H. (2013). Experimental design-based functional mining and characterization of high-throughput sequencing data in the Sequence Read Archive. *PLoS ONE*, *8*(10), e77910. doi:10.1371/journal.pone.0077910

Stewart, L., Ford, A., Sangal, V., Jeukens, J., Boyle, B., Kukavica-Ibrulj, I., Caim, S., et al. (2014). Draft genomes of 12 host-adapted and environmental isolates of *Pseudomonas aeruginosa* and their positions in the core genome phylogeny. *Pathogens and Disease*, *71*(1), 20–25. doi:10.1111/2049-632X.12107

Wickham, H. (2016). Ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. Retrieved from http://ggplot2.org

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., et al. (2012). Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, *67*(11), 2640–2644. doi:10.1093/jac/dks261

Zhu, Y., Stephens, R. M., Meltzer, P. S., & Davis, S. R. (2013). SRAdb: Query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, *14*(1), 19. doi:10.1186/1471-2105-14-19