

iSeq: an integrated tool to fetch public sequencing data

Chao H.Y.¹, Li Z.J.², Chen D.J.², Chen M.^{1*}

¹ Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China

² Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China

* mchen@zju.edu.cn

Key words: iSeq; NGS; metadata; SRA; FASTQ

Motivation and Aim: High-throughput DNA sequencing technologies have revolutionized biological research [1]. Despite the significant volume of publicly available sequencing data, accessing it programmatically remains challenging. In this context, we introduce iSeq, a tool designed to streamline the retrieval of metadata and NGS data from various databases. iSeq aims to facilitate efficient data access, supporting multiple accession formats and offering functionalities such as FASTQ file merging and multi-threaded processes.

Methods and Algorithms: iSeq is a Bash script that supports the retrieval of sequencing data and metadata from GSA [2], SRA [3], ENA [4], and DDBJ [5] databases. It offers a range of features including support for multiple accession formats (Project, Study, Sample, Experiment, or Run), metadata download, direct retrieval of gzip-formatted FASTQ files, multi-threaded processes, and parallel downloads. These features enhance the efficiency and flexibility of data retrieval, catering to diverse research needs.

Results: Utilizing iSeq, we conducted a series of analyses to showcase its capabilities in retrieving public sequencing data and metadata. In one instance, we aimed to investigate the transcriptomic profiles of different tissues in a model organism. Leveraging iSeq, we effortlessly retrieved RNA-seq data from the SRA database using the corresponding accession IDs. The tool's ability to handle multiple accession formats streamlined the retrieval process, enabling us to obtain comprehensive datasets encompassing diverse tissue types. Furthermore, we explored the microbial diversity in various environmental samples by retrieving metagenomic data from the ENA database. iSeq facilitated the retrieval of metagenomic datasets associated with specific environmental conditions, allowing for in-depth analyses of microbial community structures and functional potentials. In another scenario, we sought to investigate the genetic basis of a complex trait by analyzing whole-genome sequencing data. iSeq proved instrumental in retrieving large-scale genomic datasets from the GSA database, providing access to a wealth of genetic information for downstream analyses. Moreover, iSeq's support for direct download of gzip-formatted FASTQ files expedited data retrieval, particularly for projects hosted on the GSA database. This feature proved invaluable for projects requiring rapid access to sequencing data in a compressed format, minimizing download times and storage requirements. Overall, our experiences demonstrate the versatility and efficacy of iSeq in accessing public sequencing data across multiple databases. The tool's intuitive interface, coupled with its robust functionality, empowers researchers to efficiently retrieve and utilize diverse genomic datasets for various biological inquiries.

Conclusion: iSeq presents a valuable resource for researchers seeking to access and utilize public sequencing data efficiently. By providing seamless access to multiple

databases and offering advanced features for data retrieval, iSeq empowers researchers to explore and analyze genomic datasets effectively, thereby accelerating biological research.

Funding: This study was supported by the National Natural Sciences Foundation of China (No. 32070677, 32270709, 32261133526, 32300532).

References

1. Karouia F., Peyvan K., Pohorille A. Toward biotechnology in space: High-throughput instruments for in situ biological research beyond Earth. *Biotechnol Adv.* 2017;35(7):905-932. doi 10.1016/j.biotechadv
2. Chen T., Chen X., Zhang S. et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics Proteomics Bioinf.* 2021;19(4):578-583. doi 10.1016/j.gpb.2021.08.001
3. Katz K., Shutov O., Lapoint R., Kimelman M., Brister J.R., O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* 2022;50(D1):D387-D390. doi 10.1093/nar/gkab1053
4. Cummins C., Ahamed A., Aslam R. et al. The European Nucleotide Archive in 2021. *Nucleic Acids Res.* 2022;50(D1):D106-D110. doi 10.1093/nar/gkab1051
5. Fukuda A., Kodama Y., Mashima J., Fujisawa T., Ogasawara O. DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res.* 2021;49(D1):D71-D75. doi 10.1093/nar/gkaa982