iSeq: An integrated tool to fetch public sequencing data

Haoyu Chao [1], Zhuojin Li [2], Dijun Chen [2,*], Ming Chen [1,*]

[1] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China;

[2] Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing 210023, China;

Abstract

High-throughput sequencing technologies (Next Generation Sequencing; NGS) are increasingly utilized by researchers to tackle a diverse array of biological inquiries. Leveraging the remarkable scale and efficiency of modern sequencing, significant advancements are made across various fields, spanning from genome analysis to the intricate dynamics of protein-nucleic acid interactions. Recognizing that NGS data harbors rich biological information, the International Nucleotide Sequence Database Collaboration (INSDC) was established nearly 40 years ago to collect and disseminate public nucleotide sequence data and associated metadata. The National Genomics Data Center (NGDC) has also provided open access to vast amounts of raw sequence data. These databases have greatly enhanced the capacity for reanalyzing NGS data. In recent years, amid the rise of large language models, biological sequences and data have emerged as inputs for training models to address biological challenges. However, methods for programmatically accessing this public sequencing data remain limited. To address this gap, we have developed iSeq, an integrated tool that allows for quick and straightforward retrieval of metadata and NGS data via the command-line interface. iSeq is currently the only tool that supports simultaneous retrieval from multiple databases (GSA, SRA, ENA, DDBJ, and GEO). Additionally, iSeq supports a wide range of accession formats as input and features parallel downloads, multi-threaded processes, and FASTQ file merging. It is freely available on Bioconda (https://anaconda.org/bioconda/iseq) and GitHub (https://github.com/BioOmics/iSeq).

Keywords: iSeq, NGS, Metadata, SRA, FASTQ

Highlights:

- iSeq supports multiple databases for accessing a wide range of raw sequencing data and metadata.
- iSeq supports at least 25 different accession formats as input.
- iSeq supports parallel downloads, multi-threaded processes, FASTQ file merging, and integrity verification.

INTRODUCTION

Next Generation Sequencing (NGS) represents a significant advancement in molecular biology, allowing for high-throughput sequencing of DNA and RNA [1]. Alongside the development of NGS technology, numerous omics applications have emerged, including bulk RNA-Seq [2], ATAC-Seq [3], ChIP-Seq [4] Hi-C [5], and WGS [6]. Exciting new applications are also being explored, such as single-cell and spatial omics in epigenomics [7], transcriptomics [8], proteomics [9], and metabolomics [10]. These technologies have revolutionized various fields. For instance, in oncology, NGS is extensively used to identify genetic mutations in tumors, aiding in the development of targeted therapies [11]. Moreover, during the COVID-19 pandemic, NGS played a crucial role in identifying and tracking the SARS-CoV-2 virus [12]. Beyond medical applications, NGS is utilized in agriculture to improve yields, quality and resistance by studying plant genomes [13].

With the widespread adoption of NGS technology, the volume of NGS data has dramatically increased. Large-scale mining and analysis of NGS data can provide a more comprehensive understanding of biological mysteries. For example, Gavish *et al*. utilized 1,456 public single-cell RNA-seq (scRNA-seq) samples to uncover hallmarks of transcriptional intratumour heterogeneity across a thousand tumors [14]. Zuntini *et al.* built the tree of life for almost 8,000 angiosperm genera using the publicly available multi-omics data of 2,054 species and their internal NGS data [15]. Nowadays, in the era of flourishing large language models, large-scale NGS data can significantly enhance model performance. For instance, Genoformer used 561 publicly available scRNA-seq datasets (about 30 million cells) to train a model, accelerating the discovery of key network regulators and candidate therapeutic targets [16]. Similarly, scGPT used 65 scRNA-seq datasets (over 33 million cells) to train a model optimized for superior performance across diverse downstream applications, such as cell type annotation and multi-omic integration [17].

New NGS data are increasingly generated from economically growing countries, and data reuse and reanalysis are becoming common [18]. To support NGS data storage and sharing for the global scientific community, the International Nucleotide Sequence Database Collaboration (INSDC) was established nearly 40 years ago to collect and disseminate public nucleotide sequence data and associated metadata [18]. INSDC includes three partner organizations: the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp) in Japan [19], the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) in the UK [20], and GenBank (https://www.ncbi.nlm.nih.gov/genbank) in the USA [21]. The Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) was established as part of INSDC in 2009 [22]. In addition to INSDC, the Genome Sequence Archive (GSA; https://ngdc.cncb.ac.cn/gsa) at The National Genomics Data Center (NGDC) in China also provides vast amounts of raw sequence reads [23]. These organizations collaboratively maintain NGS data for the benefit of science and the wider community.

While high-throughput DNA sequencing technologies have revolutionized biological research, the continuously growing volume of data presents challenges for efficient and convenient access and deep mining. In this context, we introduce iSeq, a tool designed to streamline the retrieval of metadata and NGS data from GSA, SRA, ENA, and DDBJ databases. iSeq aims to facilitate efficient data access, supporting multiple accession formats and offering functionalities such as parallel downloads and multi-threaded processes.

METHODS

iSeq is a comprehensive Bash script designed to efficiently fetch public sequencing data from multiple bioinformatics databases, including GSA, SRA, ENA, and DDBJ. The workflow begins by providing an accession number **(Figure 1)**, which can be in various formats such as Project, Study, Sample, Experiment, or Run. iSeq automatically detects the accession format and fetches metadata from the appropriate source, prioritizing ENA among the partner organizations of INSDC or GSA due to their extensive data availability.

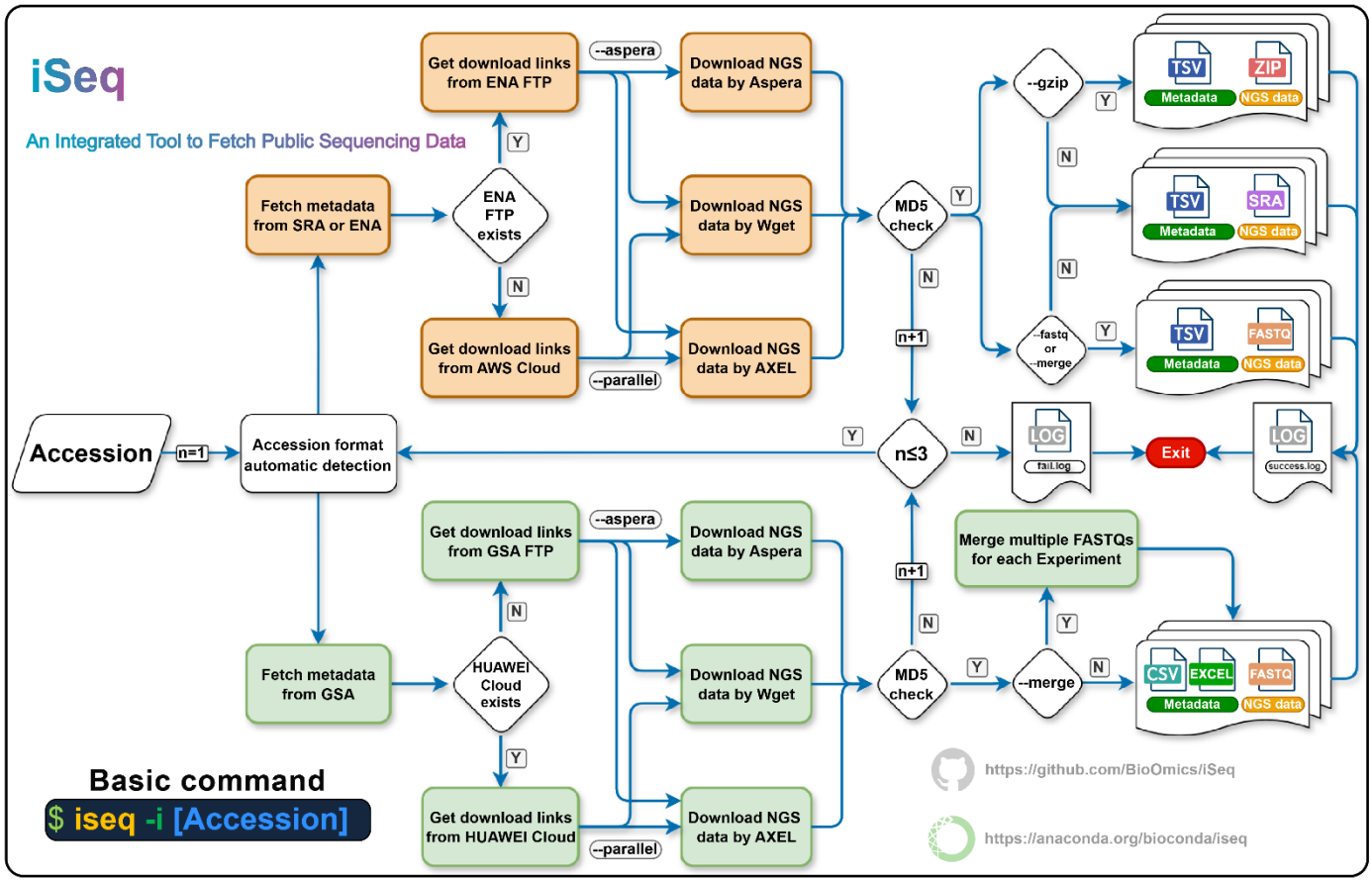


Figure 1. iSeq workflow pipeline. Once iSeq receives the correct accession, it will execute from left to right in sequence.

Once the metadata is fetched, iSeq determines the best source for downloading the sequencing data. If the fetched accession is in the INSDC databases and the FTP channel is available in ENA, iSeq will proceed to download the data using either Aspera (https://www.ibm.com/products/aspera) for fast transfer, Wget (https://www.gnu.org/software/wget), or AXEL (https://github.com/axel-download-accelerator/axel) for parallel downloading. In cases where the FTP channel

is not available in ENA, iSeq will fetch download links from AWS Cloud channel, again utilizing Wget or AXEL based on user preferences. If data is stored in GSA and the HUAWEI Cloud channel is accessible, iSeq prioritizes accessing the HUAWEI Cloud channel for its speed and reliability. Alternatively, iSeq will use the FTP channel with Aspera, Wget, or AXEL for downloading. After the initial download, iSeq will conduct an MD5 checksum to ensure the integrity of the downloaded files. If the checksum does not match the expected value in public databases, iSeq will attempt to redownload the files up to three times. Files that pass the MD5 check are logged as successfully downloaded, while those that fail after three attempts are recorded in a failure log.

For users requiring compressed data, iSeq can directly download FASTQ files in gzip format if available. Otherwise, iSeq will download SRA files and converts them to FASTQ format using fasterq-dump (https://github.com/ncbi/sra-tools), followed by compression with pigz (https://zlib.net/pigz), which is only considered when fetching from INSDC-related databases. For the GSA database, iSeq always directly fetches compressed or BAM format data. Additionally, iSeq can merge multiple FASTQ files from the same experiment into a single file for single-end (SE) sequencing data, or maintain the order and consistency of read names in two files for paired-end (PE) sequencing data. If the experiment consists of only one run, the run accession will be renamed in the format of the experiment accession. Finally, upon successful execution, iSeq will generate the corresponding metadata and NGS data.

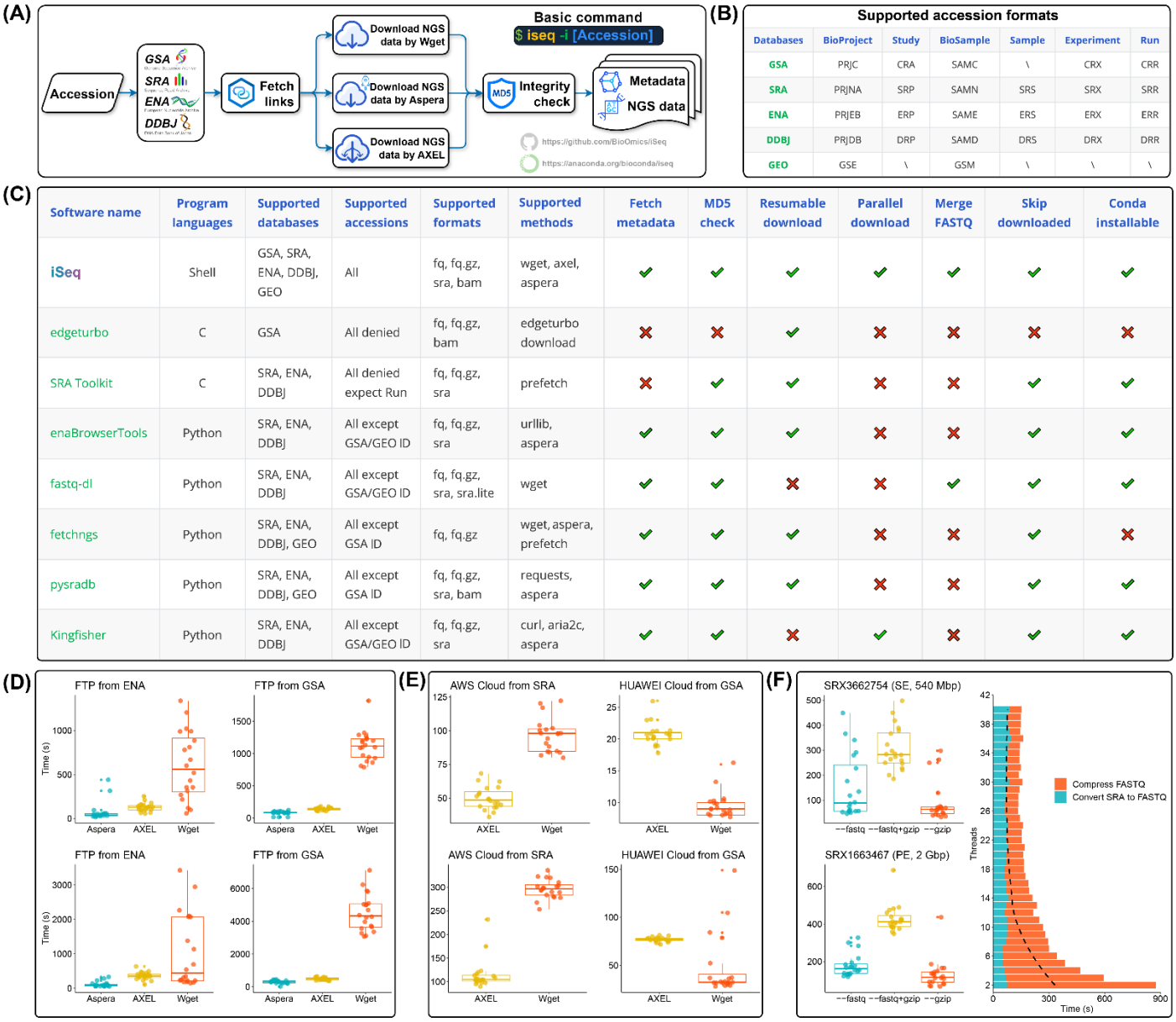
RESULTS AND DISCUSSIONS

Design of iSeq and comparative analysis

Given the exponential growth of publicly stored NGS data and its widespread usage, we developed iSeq using Bash scripting **(Figure 2A)**. iSeq efficiently retrieves sequencing data from diverse bioinformatics repositories like GSA, SRA, ENA, and DDBJ. To simplify user retrieval processes, iSeq currently supports 25 accession formats as inputs **(Figure 2B)**, then fetches relevant metadata and sequentially downloads the included Run accessions. In fact, iSeq was inspired by existing tools like fastq-dl (https://github.com/rpetit3/fastq-dl), fetchngs (https://github.com/nf-core/fetchngs) [24], pysradb (https://github.com/saketkc/pysradb) [25], and Kingfisher (https://github.com/wwood/kingfisher-download). However, given the incomplete functionalities and database support of these tools, iSeq was devised.

Specifically, iSeq stands as the shell command-line tool for fetching NGS data and metadata. Compared to other tools **(Figure 2C)**, it introduces support for the GSA database and its accession formats. Additionally, in terms of supported NGS data formats, iSeq allows retrieval of FASTQ, gzip-formatted FASTQ, SRA, and BAM file formats. For SRA formats, iSeq allows retrieval of SRA Normalized, the original format with full base quality scores. However, iSeq currently does not support SRA Lite, which reduces file size by simplifying quality scores to a uniform Q30. This decision is based on the potential impact of base quality scores on downstream omics analyses, where inaccurate quality scores could

introduce uncontrollable biases during read quality control [26]. Therefore, iSeq refrains from incorporating SRA Lite as an output format. Moreover, recognizing the time-consuming process of single-threaded retrieval for large-scale NGS data via Wget, in addition to integrating Aspera, iSeq also supports multi-threaded downloading using AXEL for individual files. However, multi-threaded downloading may result in data format inconsistencies. Therefore, iSeq conducts integrity checks on downloaded NGS data to ensure the accuracy of downstream analysis. Additionally, in the event of unforeseen interruptions during the iSeq program process, it includes breakpoint reconnection functionality, maximizing the integrity of NGS data retrieval processes.



Databases	BioProject	Study	BioSample	Sample	Experiment	Run
GSA	PRJC	CRA	SAMC	\	CRX	CRR
SRA	PRJNA	SRP	SAMN	SRS	SRX	SRR
ENA	PRJEB	ERP	SAME	ERS	ERX	ERR
DDBJ	PRJDB	DRP	SAMD	DRS	DRX	DRR
GEO	GSE	\	GSM	\	\	\

Software name	Program languages	Supported databases	Supported accessions	Supported formats	Supported methods	Fetch metadata	MD5 check	Resumable download	Parallel download	Merge FASTQ	Skip downloaded	Conda installable
iSeq	Shell	GSA, SRA, ENA, DDBJ, GEO	All	fq, fq.gz, sra, bam	wget, axel, aspera	✔	✔	✔	✔	✔	✔	✔
edgeturbo	C	GSA	All denied	fq, fq.gz, bam	edgeturbo download	✘	✘	✔	✘	✘	✘	✘
SRA Toolkit	C	SRA, ENA, DDBJ	All denied expect Run	fq, fq.gz, sra	prefetch	✘	✔	✔	✘	✘	✔	✔
enaBrowserTools	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra	urllib, aspera	✔	✔	✔	✘	✘	✔	✔
fastq-dl	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra, sra.lite	wget	✔	✔	✘	✘	✔	✔	✔
fetchngs	Python	SRA, ENA, DDBJ, GEO	All except GSA ID	fq, fq.gz	wget, aspera, prefetch	✔	✔	✔	✘	✘	✔	✘
pysradb	Python	SRA, ENA, DDBJ, GEO	All except GSA ID	fq, fq.gz, sra, bam	requests, aspera	✔	✔	✔	✘	✘	✔	✔
Kingfisher	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra	curl, aria2c, aspera	✔	✔	✘	✔	✘	✔	✔

Figure 2. Comparison of iSeq with existing tools and performance benchmarks. (A) Overview of iSeq design. **(B)** Supported accession formats by iSeq. **(C)** Comparative analysis of iSeq and existing tools. **(D)** Download speed comparison using different methods. **(E)** Cloud storage download speed comparison. **(F)** Efficiency of downloading gzip-formatted FASTQ Files.

126 **Benchmarks**

127 Firstly, to assess the stability of NGS data downloads using iSeq, we attempted to fetch 3,000 gzip-formatted FASTQ
128 files (~7 Terabyte of base pairs, Tbp) generated from RNA-Seq studies from the GSA database. The results showed
129 that after the initial download, 2,998 files were successfully downloaded and passed integrity checks **(Table S1)**.
130 Subsequently, we tried to obtain 3,000 SRA files (~5 Tbp) used in multi-omics studies from INSDC-related databases.
131 Ultimately, 2,999 files were successfully downloaded and passed integrity checks **(Table S1)**. Upon rerunning the same
132 command with iSeq, it automatically skipped the already completed downloads and successfully fetched all NGS data
133 and metadata.

134

135 Next, we evaluated the efficiency of iSeq in fetching NGS data through different channels and methods. For accessing
136 NGS data in INSDC-related databases, iSeq prioritizes the FTP channel in ENA. We performed 20 repeated downloads
137 of SRX3662754 (SE, 540 Mbp) and SRX1663467 (PE, 2 Gbp) using the three methods supported by iSeq. The results
138 showed that using Aspera (Command: iseq -i [accession] -a) to fetch NGS data provided the fastest and most stable
139 download speeds **(Figure 2D)**. Considering that Aspera requires specific firewall ports and an additional private key file
140 (which iSeq is pre-configured with automatically when installed using conda), iSeq also supports parallel downloads
141 using AXEL (Command: iseq -i [accession] -p [number of parallel]), achieving speeds close to Aspera **(Figure 2D)**. This
142 conclusion holds true for the GSA FTP channel as well **(Figure 2D)**.

143

144 With the development of cloud technology, SRA and GSA have migrated data to cloud storage. Since Aspera relies on
145 the FASP protocol, it cannot access data over the HTTPS protocol. Therefore, for cloud-stored data, iSeq exclusively
146 uses AXEL and Wget. The results showed that AXEL, when accessing data via AWS Cloud channel from SRA
147 (Command: iseq -i [accession] -p [number of parallel] -d sra), achieved faster download speeds **(Figure 2E)**, whereas
148 Wget was slightly faster when accessing data via HUAWEI Cloud channel from GSA (Command: iseq -i [accession])
149 **(Figure 2E)**.

150

151 Lastly, we evaluated efficiency of iSeq in downloading gzip-formatted FASTQ files through different methods. The results
152 indicated that directly fetching gzip-formatted FASTQ files (Command: iseq -i [accession] -g) was the fastest and most
153 computationally efficient method **(Figure 2F)**. However, the "--fastq" parameter is particularly useful for downloading
154 single-cell data, especially for scATAC-Seq data, as it can effectively decompose the files into four parts: Index, Barcode,
155 Read 1, and Read 2. If FASTQ files are directly downloaded using the "--gzip" parameter, only Read 1 and Read 2 files
156 are obtained (e.g., SRR13450125), which may cause issues during subsequent data analysis. Therefore, iSeq can
157 simultaneously use the "--fastq" and "--gzip" parameters (Command: iseq -i [accession] -q -g -t [number of threads]),
158 first fetching SRA files, then decomposing and compressing them. Notably, increasing the number of threads does not

linearly reduce time. The results indicated that increasing the number of threads did not significantly enhance the speed of decomposing SRA files, but it did notably improve compression speed **(Figure 2F)**. However, when the number of threads exceeded 15, further increases in threads did not result in a substantial reduction in time **(Figure 2F)**.

In summary, users can achieve faster and more stable download speeds for FTP channel data by setting the "--aspera" parameter in iSeq, regardless of whether it's for SRA format data or gzip-formatted files. For cloud-stored data, the same objective can be attained by setting the "--parallel [number of parallel]" parameter in iSeq. It's important to note that actual download speeds may still vary depending on the user's network environment and the real-time throughput of public databases.

CONCLUSION

iSeq stands as a command-line tool, adept at accommodating diverse input formats and providing multifaceted support. In an era characterized by the continuous expansion of NGS data and the popularity of large language models, iSeq empowers researchers with the convenient and swift ability to fetch NGS data and metadata. We believe that through iSeq, the mysteries hidden within NGS data can be brought to light once more.

AUTHOR CONTRIBUTIONS

Haoyu Chao conducted the sample collection, conceptualization, methodology, data analysis, curation, visualization, and the initial drafting and revision of the manuscript. **Zhuojin Li** debugged the iSeq program. Notably, **Ming Chen** conceived the study. **Ming Chen** and **Dijun Chen** jointly provided overall conceptualization and supervision. All authors reviewed and approved the submitted version.

ACKNOWLEDGMENTS

The authors are grateful to the members of Ming Chen’s laboratory for helpful discussions and valuable comments.

DATA AVAILABILITY STATEMENT

NGS data utilized in this study are publicly available from various repositories. Specifically, experiments SRX3662754 and SRX1663467 were retrieved from ENA and SRA, respectively. Additionally, experiments CRX917377 and CRX095512 were obtained from GSA. Furthermore, iSeq is accessible to all researchers via GitHub (https://github.com/BioOmics/iSeq) or Bioconda (https://anaconda.org/bioconda/iseq) [27], and users can install it on any Linux system. The code and data used for performance benchmarks can also be accessed on GitHub (https://github.com/BioOmics/iSeq/tree/main/docs/benchmark).

FUNDING STATEMENT

This work was supported by the National Natural Sciences Foundation of China (31771477, 32070677, 32070656), the National Key Research and Development Program of China (SQ2022YFE012895, 2023YFE0112300).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest associated with this study.

ORCID

Haoyu Chao: https://orcid.org/0000-0002-5475-0443
Zhuojin Li: https://orcid.org/0000-0002-4309-9798
Dijun Chen: http://orcid.org/0000-0002-7456-2511
Ming Chen: https://orcid.org/0000-0002-9677-1699

REFERENCES

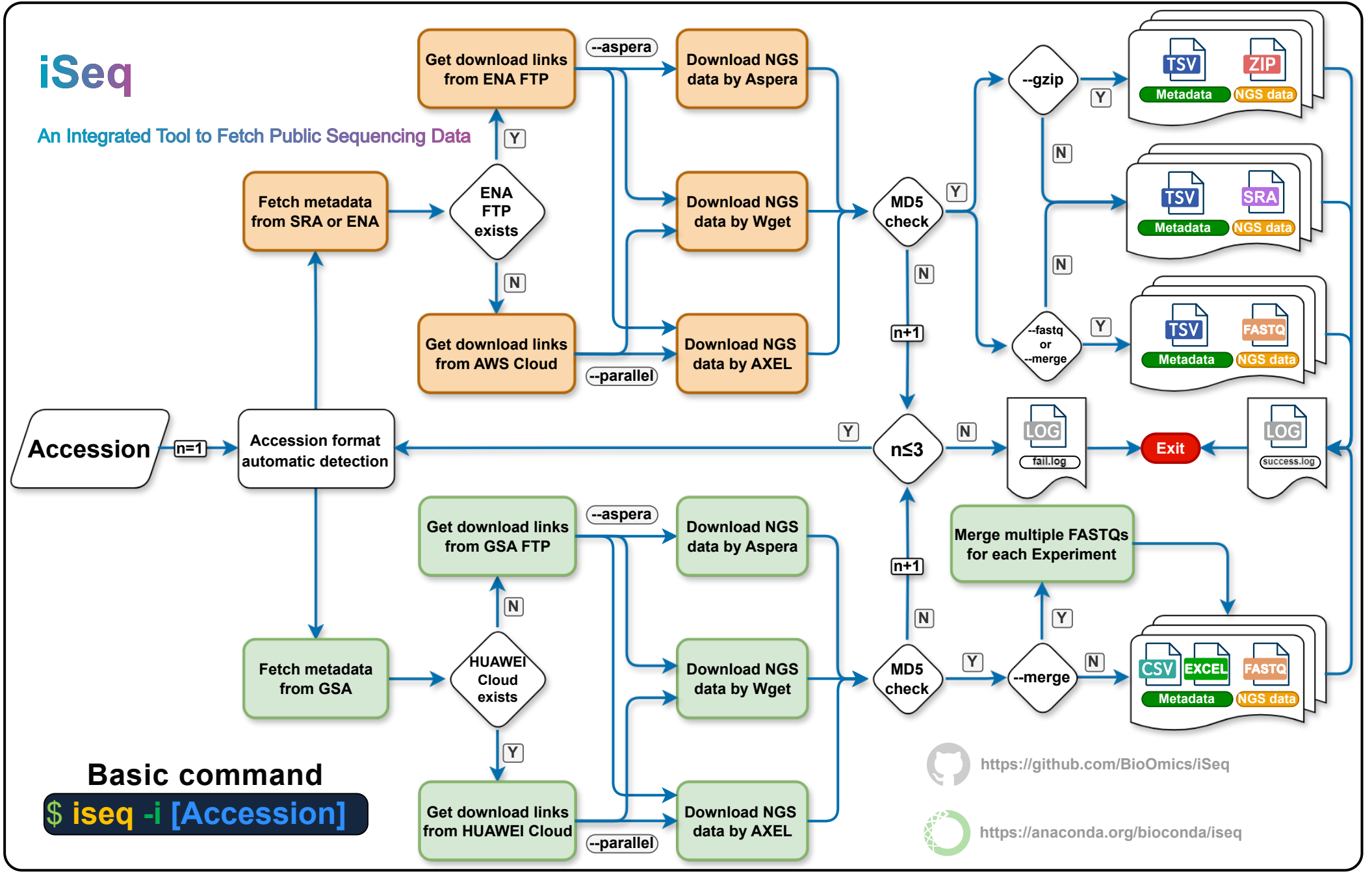
1. Goodwin, S., J. D. McPherson, W. R. McCombie. 2016. "Coming of age: ten years of next-generation sequencing technologies." *Nat Rev Genet* 17: 333-351. https://doi.org/10.1038/nrg.2016.49
2. Stark, R., M. Grzelak, J. Hadfield. 2019. "RNA sequencing: the teenage years." *Nat Rev Genet* 20: 631-656. https://doi.org/10.1038/s41576-019-0150-2
3. Buenrostro, J. D., B. Wu, H. Y. Chang, W. J. Greenleaf. 2015. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Curr Protoc Mol Biol* 109: 21.29.21-21.29.29. https://doi.org/10.1002/0471142727.mb2129s109
4. Park, P. J. 2009. "ChIP-seq: advantages and challenges of a maturing technology." *Nat Rev Genet* 10: 669-680. https://doi.org/10.1038/nrg2641
5. Belton, J. M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, J. Dekker. 2012. "Hi-C: a comprehensive technique to capture the conformation of genomes." *Methods* 58: 268-276. https://doi.org/10.1016/j.ymeth.2012.05.001
6. Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, P. Fox, et al. 2008. "Whole-genome sequencing and variant discovery in C. elegans." *Nat Methods* 5: 183-188. https://doi.org/10.1038/nmeth.1179
7. Kelsey, G., O. Stegle, W. Reik. 2017. "Single-cell epigenomics: Recording the past and predicting the future." *Science* 358: 69-75. https://doi.org/10.1126/science.aan6826
8. Longo, S. K., M. G. Guo, A. L. Ji, P. A. Khavari. 2021. "Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics." *Nat Rev Genet* 22: 627-644. https://doi.org/10.1038/s41576-021-00370-8
9. Mund, A., A. D. Brunner, M. Mann. 2022. "Unbiased spatial proteomics with single-cell resolution in tissues." *Mol Cell* 82: 2335-2349. https://doi.org/10.1016/j.molcel.2022.05.022
10. Saunders, K. D. G., H. M. Lewis, D. J. Beste, O. Cexus, M. J. Bailey. 2023. "Spatial single cell metabolomics: Current challenges and future developments." *Curr Opin Chem Biol* 75: 102327. https://doi.org/10.1016/j.cbpa.2023.102327
11. Horak, P., S. Fröhling, H. Glimm. 2016. "Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls." *ESMO Open* 1: e000094. https://doi.org/10.1136/esmoopen-2016-000094
12. Chiara, M., A. M. D'Erchia, C. Gissi, C. Manzari, A. Parisi, N. Resta, F. Zambelli, et al. 2021. "Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities." *Brief Bioinform* 22: 616-630. https://doi.org/10.1093/bib/bbaa297
13. Sun, Y., L. Shang, Q. H. Zhu, L. Fan, L. Guo. 2022. "Twenty years of plant genome sequencing: achievements and challenges." *Trends Plant Sci* 27: 391-401. https://doi.org/10.1016/j.tplants.2021.10.006
14. Gavish, A., M. Tyler, A. C. Greenwald, R. Hoefflin, D. Simkin, R. Tschernichovsky, N. Galili Darnell, et al. 2023. "Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours." *Nature* 618: 598-606. https://doi.org/10.1038/s41586-023-06130-4
15. Zuntini, A. R., T. Carruthers, O. Maurin, P. C. Bailey, K. Leempoel, G. E. Brewer, N. Epitawalage, et al. 2024.

"Phylogenomics and the rise of the angiosperms." *Nature* https://doi.org/10.1038/s41586-024-07324-0

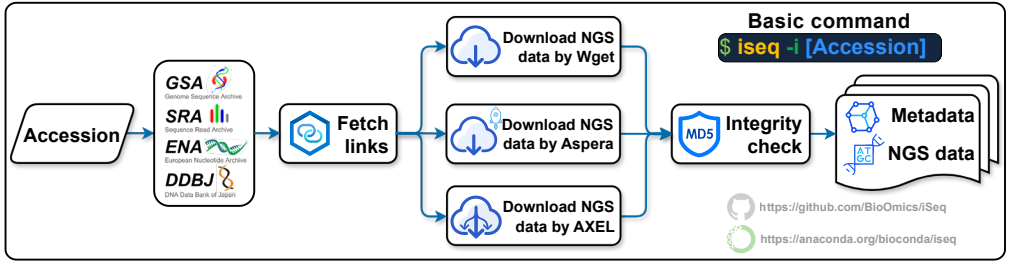
16. Theodoris, C. V., L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, et al. 2023. "Transfer learning enables predictions in network biology." *Nature* 618: 616-624. https://doi.org/10.1038/s41586-023-06139-9
17. Cui, H., C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, B. Wang. 2024. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI." *Nat Methods* https://doi.org/10.1038/s41592-024-02201-0
18. Arita, M., I. Karsch-Mizrachi, G. Cochrane. 2021. "The international nucleotide sequence database collaboration." *Nucleic Acids Res* 49: D121-d124. https://doi.org/10.1093/nar/gkaa967
19. Ara, T., Y. Kodama, T. Tokimatsu, A. Fukuda, T. Kosuge, J. Mashima, Y. Tanizawa, et al. 2024. "DDBJ update in 2023: the MetaboBank for metabolomics data and associated metadata." *Nucleic Acids Res* 52: D67-d71. https://doi.org/10.1093/nar/gkad1046
20. Yuan, D., A. Ahamed, J. Burgin, C. Cummins, R. Devraj, K. Gueye, D. Gupta, et al. 2024. "The European Nucleotide Archive in 2023." *Nucleic Acids Res* 52: D92-d97. https://doi.org/10.1093/nar/gkad1067
21. Sayers, E. W., M. Cavanaugh, K. Clark, K. D. Pruitt, S. T. Sherry, L. Yankie, I. Karsch-Mizrachi. 2024. "GenBank 2024 Update." *Nucleic Acids Res* 52: D134-d137. https://doi.org/10.1093/nar/gkad903
22. Katz, K., O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, C. O'Sullivan. 2022. "The Sequence Read Archive: a decade more of explosive growth." *Nucleic Acids Res* 50: D387-d390. https://doi.org/10.1093/nar/gkab1053
23. Chen, T., X. Chen, S. Zhang, J. Zhu, B. Tang, A. Wang, L. Dong, et al. 2021. "The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types." *Genomics Proteomics Bioinformatics* 19: 578-583. https://doi.org/10.1016/j.gpb.2021.08.001
24. Ewels, P. A., A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, S. Nahnsen. 2020. "The nf-core framework for community-curated bioinformatics pipelines." *Nat Biotechnol* 38: 276-278. https://doi.org/10.1038/s41587-020-0439-x
25. Choudhary, S. 2019. "pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive." *F1000Res* 8: 532. https://doi.org/10.12688/f1000research.18676.1
26. Gargis, A. S., L. Kalman, M. W. Berry, D. P. Bick, D. P. Dimmock, T. Hambuch, F. Lu, et al. 2012. "Assuring the quality of next-generation sequencing in clinical laboratory practice." *Nat Biotechnol* 30: 1033-1036. https://doi.org/10.1038/nbt.2403
27. Grüning, B., R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, J. Köster. 2018. "Bioconda: sustainable and comprehensive software distribution for the life sciences." *Nat Methods* 15: 475-476. https://doi.org/10.1038/s41592-018-0046-7

SUPPORTING INFORMATION

Table S1： Summary of NGS data download and integrity check results.


iSeq
An Integrated Tool to Fetch Public Sequencing Data
Accession
n=1
Accession format automatic detection
Fetch metadata from SRA or ENA
ENA FTP exists
Y
N
Get download links from ENA FTP
Get download links from AWS Cloud
--aspera
--parallel
Download NGS data by Aspera
Download NGS data by Wget
Download NGS data by AXEL
MD5 check
Y
N
n+1
n≤3
Y
N
LOG
fail.log
Exit
LOG
success.log
--gzip
--fastq or --merge
TSV
ZIP
SRA
FASTQ
Metadata
NGS data
Fetch metadata from GSA
HUAWEI Cloud exists
Get download links from GSA FTP
Get download links from HUAWEI Cloud
Merge multiple FASTQs for each Experiment
--merge
CSV
EXCEL
Basic command
$ iseq -i [Accession]
https://github.com/BioOmics/iSeq
https://anaconda.org/bioconda/iseq

A



B

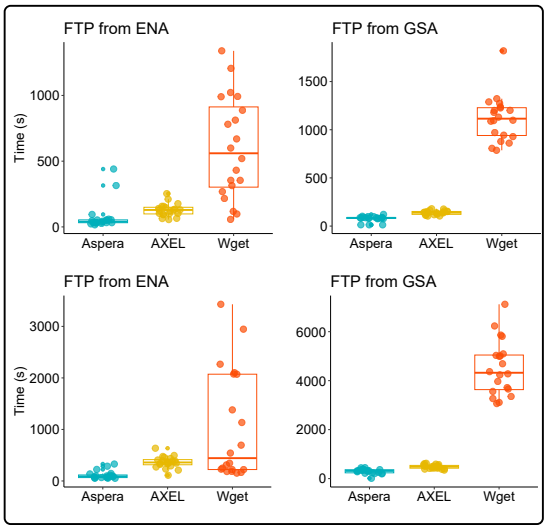
Supported accession formats

Databases	BioProject	Study	BioSample	Sample	Experiment	Run
GSA	PRJC	CRA	SAMC	\	CRX	CRR
SRA	PRJNA	SRP	SAMN	SRS	SRX	SRR
ENA	PRJEB	ERP	SAME	ERS	ERX	ERR
DDBJ	PRJDB	DRP	SAMD	DRS	DRX	DRR
GEO	GSE	\	GSM	\	\	\

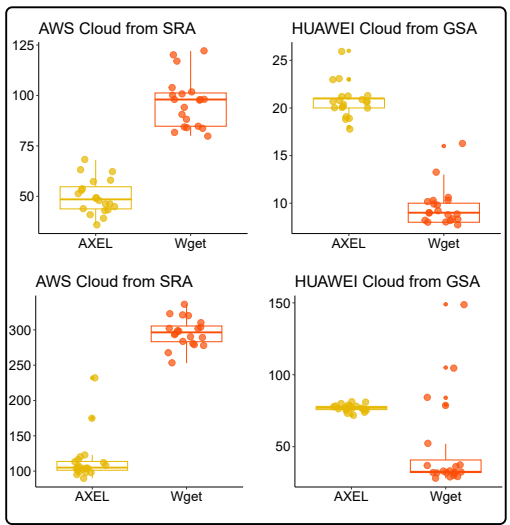
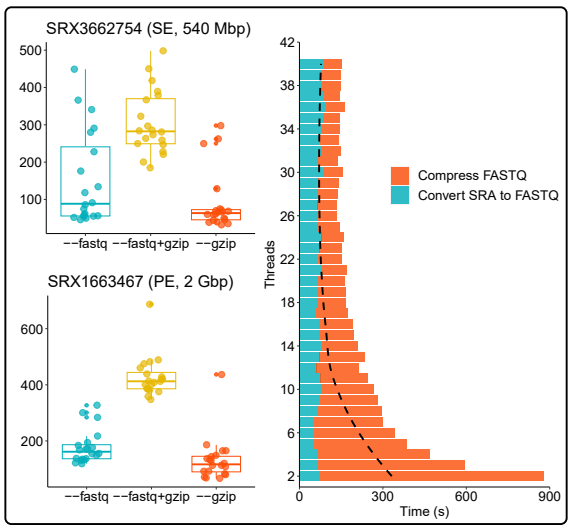
C

Software name	Program languages	Supported databases	Supported accessions	Supported formats	Supported methods	Fetch metadata	MD5 check	Resumable download	Parallel download	Merge FASTQ	Skip downloaded	Conda installable
iSeq	Shell	GSA, SRA, ENA, DDBJ, GEO	All	fq, fq.gz, sra, bam	wget, axel, aspera	✔	✔	✔	✔	✔	✔	✔
edgeturbo	C	GSA	All denied	fq, fq.gz, bam	edgeturbo download	✘	✘	✔	✘	✘	✘	✘
SRA Toolkit	C	SRA, ENA, DDBJ	All denied expect Run	fq, fq.gz, sra	prefetch	✘	✔	✔	✘	✘	✔	✔
enaBrowserTools	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra	urllib, aspera	✔	✔	✔	✘	✘	✔	✔
fastq-dl	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra, sra.lite	wget	✔	✔	✘	✘	✔	✔	✔
fetchngs	Python	SRA, ENA, DDBJ, GEO	All except GSA ID	fq, fq.gz	wget, aspera, prefetch	✔	✔	✔	✘	✘	✔	✘
pysradb	Python	SRA, ENA, DDBJ, GEO	All except GSA ID	fq, fq.gz, sra, bam	requests, aspera	✔	✔	✔	✘	✘	✔	✔
Kingfisher	Python	SRA, ENA, DDBJ	All except GSA/GEO ID	fq, fq.gz, sra	curl, aria2c, aspera	✔	✔	✘	✔	✘	✔	✔

D



E



F