Bioinformatics, 2024, **40(4)**, btae152
https://doi.org/10.1093/bioinformatics/btae152
Advance Access Publication Date: 19 March 2024
Applications Note

Data and text mining

GINSA: an accumulator for paired locality and next-generation small ribosomal subunit sequence data

Eric Odle [1,*], Samuel Kahng [2,3], Siratee Riewluang[1], Kyoko Kurihara[1], Kevin C. Wakeman[3,4,*]

[1]Department of Natural History Sciences, Graduate School of Science, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan
[2]Department of Oceanography, University of Hawaii at Manoa, Honolulu, HI 96822, United States
[3]Institute for the Advancement of Higher Education, Hokkaido University, Sapporo, Hokkaido 060-0817, Japan
[4]Graduate School of Science, Hokkaido University, Sapporo, Hokkaido 060-0810, Japan

*Corresponding authors. Institute for the Advancement of Higher Education, Hokkaido University, Kita-17 Nishi-8, Kita Ward, Sapporo, Hokkaido 060-0817, Japan. E-mail: wakeman.k@oia.hokudai.ac.jp (K.C.W.); Department of Natural History Sciences, Graduate School of Science, Sapporo, Kita-10 Nishi-8, Kita Ward, Sapporo, Hokkaido 060-0810, Japan. E-mail: ericmichael.odle.q5@elms.hokudai.ac.jp (E.O.)

Associate Editor: Jonathan Wren

Abstract

Motivation: Motivated by the challenges of decentralized genetic data spread across multiple international organizations, GINSA leverages the Global Biodiversity Information Facility infrastructure to automatically retrieve and link small ribosomal subunit sequences with locality information.

Results: Testing on taxa from major organism groups demonstrates broad applicability across taxonomic levels and dataset sizes.

Availability and implementation: GINSA is a freely accessible Python program under the MIT License and can be installed from PyPI via pip.

1 Introduction

Advances in nucleic acid sequencing technologies have led to a rapid increase in the amount of available genetic data (Keen *et al.* 1996, Warburton and Sebra 2023). To better organize and share this emergent abundance of sequence data between researchers, public databases such as GenBank (Benson *et al.* 1993, Strasser 2011), the European Nucleotide Archive (ENA) (Burgin *et al.* 2023), and the DNA Data Bank of Japan (DDBJ) (Tanizawa *et al.* 2023) were established in the 1980s. For example, evolutionary biologists often rely on small ribosomal subunit rRNA gene (SSU) sequences archived in these databases to study new species. However, sequence databases do not require a complete set of metadata (e.g. site of collection, date of collection, species-level identification, or link to publication) when uploading sequences. Absence of a complete set of metadata can lead to the omission of locality data, forcing biologists to manually seek associated location information elsewhere. To address this disconnect among archived data, we developed GINSA (GbIf Next-gen Sequence Accumulator): a biodiversity research tool that fetches SSU sequences and their associated localities. This tool takes advantage of the Global Biodiversity Information Facility (GBIF) [www.gbif.org], which links taxa (scientific names), localities (sites of occurrence), and SSU sequences.

Manually pairing high-volume SSU sequence and locality data is prohibitively slow. Although GBIF provides links to specific sequences used for identification, researchers must currently follow a convoluted chain of websites to FASTA files stored in off-site repositories (typically ENA for next-generation sequencing). Upon downloading FASTA/FASTQ files, researchers must then manually search massive lists (often hundreds of thousands) of SSU sequences. This time-consuming step is required for each species occurrence on GBIF, for which there can be thousands. Finally, researchers must manually trace sequences back to their occurrence locality from GBIF. We developed GINSA to automate this process.

Previous attempts to address the inaccessibility of sequence metadata include pysrabd (Choudhary 2019), grabseqs (Taylor *et al.* 2020), and *ffq* (Gálvez-Merchán *et al.* 2023). While helpful for specific applications, these tools address use cases that differ from those of GINSA. Python package pysradb provides convenient access to next-generation sequences stored on the National Center for Biotechnology Information (NCBI) Sequence Read Archive but does not focus on data from ENA. Another tool, grabseqs, automates next-generation sequence acquisition for multiple repositories, but requires users to have prior knowledge of the specific accession numbers associated with their organism of interest. Similarly, *ffq* addresses the difficulty in acquiring sequence metadata from ENA. However, *ffq* requires database accession or article DOI numbers as input. In contrast, GINSA leverages the structure provided by GBIF to link specific taxa with their known localities and SSU sequences. Users simply enter the name of their target organism (taxon), and then wait for the collection process to finish automatically.

2 Applications and implementation

The GINSA tool provides researchers with large datasets in line with the Big Data (De Mauro *et al.* 2016, Miralles *et al.* 2020)

Received: 27 December 2023; **Revised:** 15 February 2024; **Editorial Decision:** 11 March 2024; **Accepted:** 16 March 2024

nature of current molecular taxonomy. For instance, evolutionary biologists rely on large molecular sequence datasets to study speciation trends (Schlegel 1991, Adl *et al.* 2019). Large datasets help resolve cryptic diversity, which is a challenge seen across the tree of life—animals (Marchán *et al.* 2018, Li and Wiens 2023), plants (Vieu *et al.* 2023, Windham *et al.* 2023), fungi (Koufopanou *et al.* 1997, Pringle *et al.* 2005), bacteria (Meyer *et al.* 2023), protists (Wakeman and Leander 2013, Krienitz *et al.* 2015, Martin *et al.* 2016), archaea (Câmara *et al.* 2023), and viruses (Roux *et al.* 2019). Specialists across a range of taxa can therefore use GINSA to collect more data for their phylogenetic (SSU sequence) and biogeographic (locality) analyses.

There are currently over 2.6 billion occurrences on GBIF representing 1.3 million confirmed species. Occurrence coverage is uneven across taxa; animals account for 79.8%, followed by plants at 16.8%, and other taxa at <1.5% each (Table 1). When considering only next-generation SSU sequences archived by ENA (via the publisher MGnify), coverage favors bacteria and protists (Supplementary Fig. S1). This ENA/MGnify subset includes approximately 23.7 million GBIF occurrences, which together comprise the pool of data accessible by GINSA. Moreover, this data pool is expected to grow. Since 2008, 25–50 thousand new species from each major global region have been added to GBIF every two years (Waller 2020).

The GINSA tool offers an efficient method for accessing ENA next-generation sequence repositories linked to GBIF taxon occurrences (Fig. 1). Users enter a search taxon, then GINSA queries GBIF for all recorded occurrences of that taxon. Next, the program extracts respective ENA links from GBIF occurrence records, downloading and processing FASTA/FASTQ files into a curated list of SSU sequences. The following details outline how GINSA automates this task.

2.1 User prompt

Upon running GINSA, users are prompted for two inputs: the project folder path and the target taxon. All subsequent sub-folders and output files are saved inside the project folder. Search taxa are parsed in Python as a string, and users may enter either one-word (e.g. genus name *Lecudina*) or two-word (e.g. species name *Lecudina longissima*) queries.

2.2 GBIF taxon search

An API call searches GBIF for all instances of the queried taxon, and a list is generated by the function search_species_occurrences() for all matching GBIF occurrences.

Table 1. Summary of GBIF biodiversity coverage across major groups.[a]

Taxon	Total occurrences	ENA/MGnify
Animals	2 097 448 406	33 065
Plants	442 531 533	376 547
Fungi	38 914 204	955 943
Bacteria	22 722 639	18 355 383
Protists	15 895 213	3 098 014
Archaea	442 031	335 722
Viruses	910 025	0
Incertae sedis	8 014 894	630 700

[a] Coverage on GBIF for each major group is quantified by number of occurrences. Column 1 (Taxon) lists the major groups of life recognized by GBIF. Groups Chromista and Protozoa are combined as Protists. Column 2 (Total Occurrences) shows the current total number of GBIF occurrences. Column 3 (ENA/MGnify) shows the number of occurrences based on material samples archived by MGnify.

2.3 FASTA and MAPseq download

Sequential API calls are made to ENA/MGnify for each occurrence linked from GBIF. Next, the function ssu_fasta_grab() downloads FASTA/FASTQ files containing SSU sequences belonging to the search taxon. This process is then repeated by mapseq_grab() for the associated MAPseq files. These MAPseq files are necessary because next-generation sequencing read assembly generates long lists of sequences with complex names.

2.4 SSU contig decode

For each occurrence, a text search is run within the MAPseq file to locate all sequences associated with the search taxon. Next, sequence labels are tracked to specific sequences in the corresponding FASTA/FASTQ file.

2.5 Generate FASTA master file

Extracted SSU sequences are gathered into a file named seq_master.FASTA alongside a corresponding metadata table named occurrences.csv. Users may annotate seq_master.FASTA with additional GBIF metadata (e.g. latitude, longitude, or country of origin). A script (misc/suffix_annotator.py) demonstrating annotation with occurrences.csv is provided on the project GitHub repository.

3 Availability and testing

The GINSA project was written using Python (Van Rossum and Drake 1995) version 3.12 and is free to use under the MIT License. Code was structured into two scripts: a command line interface (CLI) implementation named GINSA_cli.py and a graphical user interface (GUI) implementation named GINSA_gui.py. Following installation via pip,

```
pip3 install GINSA
```

the CLI can be run with a single line of text:

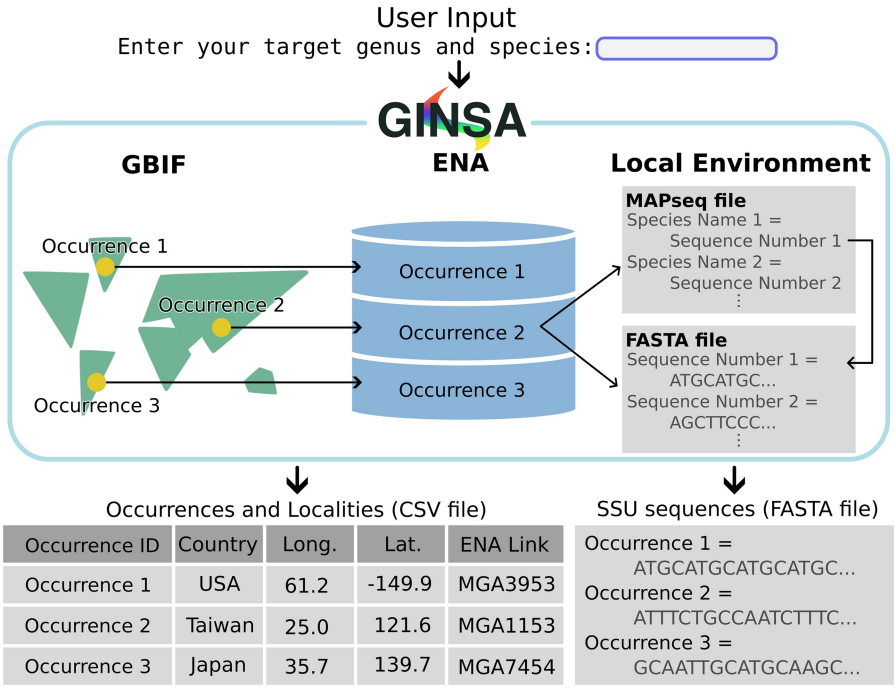
```
GINSA_cli <path/to/project/directory>
<"search taxon">
```

Moreover, the GUI can be run by simply entering:

```
GINSA_gui
```

A broad range of taxonomic groups were represented when testing GINSA (Table 2). These groups include animals (arthropod genus *Lambia*), plants (*Aneura mirabilis*, *Chrysymenia brownii*), fungi (*Malassezia globosa*), bacteria (*Altibacter lentus*), protists (*Lecudina longissima*, *Tetraselmis marina*, *Lecudina tuzetae*, and *Labyrinthula* spp.), and archaea (*Nanohaloarchaea*). This set of taxa allowed us to evaluate the speed and utility of GINSA across multiple taxa and occurrence sizes. Testing was performed on an Intel Xeon W-2235 CPU 3.80 GHz system with 31.0 GiB of available memory running Linux kernel 5.15.0–87. Network download speed during testing was stable, ranging from 443 to 540 Mbp.

Following testing, GINSA exhibited applicability across a spectrum of taxa and occurrence sizes. Taxa with smaller datasets (*Lambia* spp., *Lecudina longissima*) took less time to analyze than taxa with larger datasets (*Malassezia globosa*, *Labyrinthula* spp.) (Table 2). All tests completed without interruption, although the larger taxa required significantly more storage (97.8–178.5 GB). Network speed and local storage capacity were the only observed bottlenecks to performance. With sufficient storage and internet connectivity, taxa with an even greater number of GBIF occurrences could theoretically be analyzed using GINSA.



Occurrence ID	Country	Long.	Lat.	ENA Link
Occurrence 1	USA	61.2	-149.9	MGA3953
Occurrence 2	Taiwan	25.0	121.6	MGA1153
Occurrence 3	Japan	35.7	139.7	MGA7454

Figure 1. Chart visualizing the GINSA workflow. User input is taken as a GBIF search taxon. Occurrences are then linked with their source sequences archived on ENA. Output CSV and FASTA files link GBIF occurrence IDs, localities, and sequences.

Table 2. Summary of test taxa occurrences, process runtimes, and output directory sizes.[a]

Taxon	OC (*n*)	RT (min)	Size (GB)
Lambia spp.	11	3.4	0.8
Lecudina longissima	26	10.1	6.80
Tetraselmis marina	190	36.0	5.26
Nanohaloarchaea	253	152.2	58.3
Lecudina tuzetae	309	86.8	31.9
Aneura mirabilis	549	65.3	0.118
Altibacter lentus	628	336.0	79.3
Chrysymenia brownii	655	95.1	2.21
Malassezia globosa	1379	327.5	97.8
Labyrinthula spp.	2602	593.0	178.5

[a] Occurrences (OC) reflect the number (n) of Global Biodiversity Information Facility (GBIF) entries corresponding to a particular search taxon. Processing runtimes (RT) are measured in minutes (min). Resulting output directory sizes (Size) are measured in gigabytes (GB).

Neither runtime nor output directory size were linearly associated with occurrence count. These observations are attributed to the presence of GBIF occurrences identified through means (e.g. human identification, museum specimens, and Sanger sequencing) other than next-generation sequencing. For example, although *Aneura mirabilis* had 549 occurrences on GBIF, only two of those occurrences linked back to next-generation SSU sequences. For this reason, GINSA generates an output plot summarizing the proportion of searched occurrences containing next-generation SSU sequence data. Examples of these plots are provided on the project GitHub page (https://github.com/ericodle/GINSA).

4 Conclusion

This article introduced GINSA (GbIf Next-gen Sequence Accumulator), a novel tool designed to bridge the gap between genetic sequence data and locality metadata. Rapid growth in the amount of data from next-generation sequencing technologies has generated increasing demand for more efficient methods to pair sequence information with biogeographic context. The GINSA tool addresses this challenge by automating the collection of SSU sequences and locality metadata for a given taxon through integration with the Global Biodiversity Information Facility (GBIF). By streamlining the process of accessing and pairing these crucial data, GINSA enables researchers to work more efficiently. This tool has a beginner-friendly design, open-source code base, and is applicable across major organism groups. As such, GINSA is offered as a free resource for evolutionary biologists navigating the complexities of cryptic speciation and Big Data research.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions. We would also like to thank Shaun Cunningham for taking the time to test GINSA and proofread our manuscript. Lastly, we would like to extend a special thanks to Professor Kazuhiro Kogame at Hokkaido University for his academic support.

Author Contributions

Eric Odle (Conceptualization [lead], Data curation [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [lead], Supervision [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Samual Kahng (Investigation [supporting], Validation [supporting], Writing—original draft [equal], Writing—review & editing [equal]), Siratee Riewluang (Investigation [supporting], Methodology [supporting], Software [supporting],

Validation [supporting], Visualization [supporting], Writing—review & editing [equal]), Kyoko Kurihara (Software [supporting], Validation [supporting], Visualization [supporting], Writing—review & editing [supporting]), and Kevin C. Wakeman (Conceptualization [supporting], Funding acquisition [supporting], Project administration [supporting], Supervision [supporting], Validation [supporting], Writing—review & editing [supporting])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by DX Fellowship Code PH8D230001.

Data availability

All code and associated data for GINSA are freely available on the project repository at https://github.com/ericodle/GINSA.

References

Adl SM, Bass D, Lane CE *et al.* Revisions to the classification, nomenclature, and diversity of eukaryotes. *J Eukaryot Microbiol* 2019;**66**:4–119.

Benson D, Lipman DJ, Ostell J. Genbank. *Nucleic Acids Res* 1993; **21**:2963–5.

Burgin J, Ahamed A, Cummins C *et al.* The European nucleotide archive in 2022. *Nucleic Acids Res* 2023;**51**:D121–5.

Câmara PE, Bones FLV, Lopes FAC *et al.* DNA metabarcoding reveals cryptic diversity in Forest soils on the isolated Brazilian Trindade Island, South Atlantic. *Microb Ecol* 2023;**85**:1056–71.

Choudhary S. pysradb: a python package to query next-generation sequencing metadata and data from NCBI sequence read archive. *F1000Res* 2019;**8**:532.

De Mauro A, Greco M, Grimaldi M. A formal definition of big data based on its essential features. *Library Rev* 2016;**65**:122–35.

Gálvez-Merchán Á, Min KH, Pachter L *et al.* Metadata retrieval from sequence databases with FFQ. *Bioinformatics* 2023;**39**:btac667.

Keen G, Burton J, Crowley D *et al.* The genome sequence database (GSDB): meeting the challenge of genomic sequencing. *Nucleic Acids Res* 1996;**24**:13–6.

Koufopanou V, Burt A, Taylor JW. Concordance of gene genealogies reveals reproductive isolation in the pathogenic fungus *Coccidioides immitis*. *Proc Natl Acad Sci USA* 1997;**94**:5478–82.

Krienitz L, Huss VA, Bock C. *Chlorella*: 125 years of the green survivalist. *Trends Plant Sci* 2015;**20**:67–9.

Li X, Wiens JJ. Estimating global biodiversity: the role of cryptic insect species. *Syst Biol* 2023;**72**:391–403.

Marchán DF, Cosín DJD, Novo M. Why are we blind to cryptic species? lessons from the eyeless. *Eur J Soil Biol* 2018;**86**:49–51.

Martin DL, Chiari Y, Boone E *et al.* Functional, phylogenetic and host-geographic signatures of *Labyrinthula* spp. provide for putative species delimitation and a global-scale view of seagrass wasting disease. *Estuaries and Coasts* 2016;**39**:1403–21.

Meyer JL, Gunasekera SP, Brown AL *et al.* Cryptic diversity of black band disease cyanobacteria in *Siderastrea siderea* corals revealed by chemical ecology and comparative genome-resolved metagenomics. *Mar Drugs* 2023;**21**:76.

Miralles A, Bruy T, Wolcott K *et al.* Repositories for taxonomic data: where we are and what is missing. *Syst Biol* 2020;**69**:1231–53.

Pringle A, Baker D, Platt J *et al.* Cryptic speciation in the cosmopolitan and clonal human pathogenic fungus *Aspergillus fumigatus*. *Evol* 2005;**59**:1886–99.

Roux S, Krupovic M, Daly RA *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across earth's biomes. *Nat Microbiol* 2019;**4**:1895–906.

Schlegel M. Protist evolution and phylogeny as discerned from small subunit ribosomal RNA sequence comparisons. *Eur J Protistol* 1991;**27**:207–19.

Strasser BJ. The experimenter's museum: Genbank, natural history, and the moral economies of biomedicine. *Isis* 2011;**102**:60–96.

Tanizawa Y, Fujisawa T, Kodama Y *et al.* DNA data bank of Japan (DDBJ) update report 2022. *Nucleic Acids Res* 2023; **51**:D101–5.

Taylor LJ, Abbas A, Bushman FD. grabseqs: simple downloading of reads and metadata from multiple next-generation sequencing data repositories. *Bioinformatics* 2020;**36**:3607–9.

Van Rossum G, Drake FLJ. *Python tutorial*, Vol. 620. Amsterdam, The Netherlands: Centrum voor Wiskunde en Informatica, 1995.

Vieu JC, Koubínová D, Grant JR. Population genetic structure and diversity of cryptic species of the plant genus *Macrocarpaea* (gentianaceae) from the tropical Andes. *Plants* 2023;**12**:1710.

Wakeman KC, Leander BS. Molecular phylogeny of marine gregarine parasites (apicomplexa) from tube-forming polychaetes (sabellariidae, cirratulidae, and serpulidae), including descriptions of two new species of *Selenidium*. *J Eukaryot Microbiol* 2013;**60**:514–25.

Waller J. *Gbif Regional Statistics – 2020*, 2020. https://data-blog.gbif.org/post/gbif-regional-statistics-2020 (26 December 2023, date last accessed).

Warburton PE, Sebra RP. Long-read DNA sequencing: recent advances and remaining challenges. *Annu Rev Genomics Hum Genet* 2023; **24**:109–32.

Windham MD, Picard KT, Pryer KM. An in-depth investigation of cryptic taxonomic diversity in the rare endemic mustard *Draba maguirei*. *Am J Bot* 2023;**110**:e16138–22.

Bioinformatics, 2024, **40(4)**, btae152
https://doi.org/10.1093/bioinformatics/btae152
Advance Access Publication Date: 19 March 2024
Applications Note