


SOFTWARE

Open Access



getSequenceInfo: a suite of tools allowing to get genome sequence information from public repositories

Vincent Moco¹, Damien Cazenave¹, Maëlle Garnier¹, Matthieu Pot¹, Isabel Marcelino¹, Antoine Talarmin¹, Stéphanie Guyomard-Rabenirina¹, Sébastien Breurec^{1,2,3}, Séverine Ferdinand¹, Alexis Dereeper¹, Yann Reynaud¹ and David Couvin^{1*} 

*Correspondence:
david.couvin@googlemail.com

¹ Unité Transmission, Réservoir et Diversité des Pathogènes, Institut Pasteur de Guadeloupe, Les Abymes, Guadeloupe, France

² Faculté de Médecine Hyacinthe Bastaraud, Université des Antilles, Pointe-à-Pitre, France

³ Centre d'Investigation Clinique Antilles Guyane, Inserm CIC 1424, Pointe-à-Pitre, France

Abstract

Background: Biological sequences are increasing rapidly and exponentially world-wide. Nucleotide sequence databases play an important role in providing meaningful genomic information on a variety of biological organisms.

Results: The *getSequenceInfo* software tool allows to access sequence information from various public repositories (GenBank, RefSeq, and the European Nucleotide Archive), and is compatible with different operating systems (Linux, MacOS, and Microsoft Windows) in a programmatic way (command line) or as a graphical user interface. *getSequenceInfo* or gSeqI v1.0 should help users to get some information on queried sequences that could be useful for specific studies (e.g. the country of origin/isolation or the release date of queried sequences). Queries can be made to retrieve sequence data based on a given kingdom and species, or from a given date. This program allows the separation between chromosomes and plasmids (or other genetic elements/components) by arranging each component in a given folder. Some basic statistics are also performed by the program (such as the calculation of GC content for queried assemblies). An empirically designed nucleotide ratio is calculated using nucleotide information in order to tentatively provide a "NucleScore" for studied genome assemblies. Besides the main gSeqI tool, other additional tools have been developed to perform various tasks related to sequence analysis.

Conclusion: The aim of this study is to democratize the use of public repositories in programmatic ways, and to facilitate sequence data analysis in a pedagogical perspective. Output results are available in FASTA, FASTQ, Excel/TSV or HTML formats. The program is freely available at: <https://github.com/karubiotools/getSequenceInfo>. *getSequenceInfo* and supplementary tools are partly available through the recently released Galaxy KaruBioNet platform (http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html).

Keywords: Genome sequences, Nucleotide diversity, Assembly, DNA, Repository, Metadata



Background

Sequencing technologies are widely used nowadays and sequencing data is increasing at a rapid rate. Whole genome sequencing (WGS) projects can be applied in various settings and scientific studies, fostering the analysis of a large amount of data which are generally deposited in public archives such as those belonging to the International Nucleotide Sequence Database Collaboration (INSDC), which comprises the DNA Data Bank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank [1–4]. Unlike GenBank sequences, RefSeq sequences are not part of the INSDC but are derived from INSDC sequences to provide non-redundant curated data representing our current knowledge of known genes. Differences between GenBank and RefSeq genome assemblies are well described (https://www.ncbi.nlm.nih.gov/books/NBK50679/#RefSeqFAQ.what_is_the_difference_between_1).

Several important and useful initiatives have been conducted to foster data retrieval from these public sequence repositories, and various programs or online tools are available for this task. Among these programs/tools, we can non-exhaustively mention the NCBI Genome Downloading Scripts (<https://github.com/kblin/ncbi-genome-download>), Kraken_db_install_scripts (https://github.com/mw55309/Kraken_db_install_scripts), Entrez Programming Utilities (<https://www.ncbi.nlm.nih.gov/books/NBK25501/>; <https://www.ncbi.nlm.nih.gov/books/NBK179288/>), enaBrowserTools (<https://github.com/enasequence/enaBrowserTools>), NCBI genomes FTP site (<https://ftp.ncbi.nih.gov/genomes/>), and the recently released NCBI Datasets (<https://www.ncbi.nlm.nih.gov/datasets/>). Another interesting software tool named metatools_ncbi (https://github.com/farhat-lab/metatools_ncbi) allows downloading Biosample and SRA run information metadata from the NCBI. SRA tools such as fastq-dump (<https://github.com/ncbi/sra-tools>) or SRAdb [5] are also interesting for downloading FASTQ reads and associated information. Furthermore, a wide range of additional bioinformatics resources are available in the INSDC and elsewhere, adhering to the Findable, Accessible, Interoperable, Reusable (FAIR) data principles, which are fostering better practices in sequencing data analyses [6]. Some research studies are becoming more and more complex due to a wide variety of biological data and formats. Sharing information in a more reproducible, understandable and pedagogical manner is important to make scientific research affordable and accessible to a greater number of scientists and/or students.

Dedicated strategies have been developed to better organize and analyze genomic data such as the notion of genome assembly [7]. A wide variety of software tools have been developed to tentatively respond to specific or more general requests regarding sequence analysis. Nucleotide sequence data could be more or less complex to study, depending on the species of interest or the studied genomic elements/components (e.g. chromosomes, plasmids, etc.). Approaches and methods facilitating comparative genomic studies have been implemented such as the well-known average nucleotide identity (ANI) which allows improving taxonomic assignments [8].

Despite the usefulness of public genomic repositories, metadata related to genomic sequences are not always well annotated or available. As mentioned before [9], the World Health Organization (WHO) has repeatedly advocated open sharing of pathogen genetic sequences as well as the knowledge and benefits resulting from the genetic data (<https://www.who.int/blueprint/meetings-events/meeting-report-pathogen-genet>

[ic-sequence-data-sharing.pdf](#)). We aim to develop intuitive software to obtain genomic sequences and their associated data when available in a relatively simple way through a graphical user interface (GUI) or using the command line. Such an approach could be of relevant significance for the design of specific biological databases or the improvement of existing ones.

Implementation and results

Programming language and modules

The Perl programming language was used to develop the freely available `getSequenceInfo` (`gSeqI`) software. Perl/Tk graphical user interface toolkit has been employed to design the GUI in order to make an intuitive interface. Several other Perl modules are required to run the software, including BioPerl (<http://bioperl.org/>) [10]. The list of needed modules is provided in the software installation file (<https://github.com/dcouverin/getSequenceInfo/tree/master/install>). Further details regarding programming are available on the software GitHub page (<https://github.com/karubiotools/getSequenceInfo>). A user manual is also available from the GitHub page to help using the tool. The software is compatible with Linux and Microsoft Windows operating systems.

How `getSequenceInfo` tool differs from other existing tools

Unlike other existing tools, `getSequenceInfo` allows users to query specific information regarding wanted sequences. For example, users can download NCBI sequences from a given release date, and they can also download specific sequences (from the assemblies) such as plasmids, chromosomes, etc. Users can query both ENA and NCBI GenBank or RefSeq databases. Extraction of metadata associated to genome assemblies such as country, host, or isolation source is.

Furthermore, one may notice the notion of “NucleScore” (further explanations are provided below) which could bring new information for classification/characterization of genome assemblies. This score is a reduction of the nucleotide information which is intended to be discriminating and informative. To calculate the NucleScore, measures such as nucleotide variance, GC content, genome size, and AT/GC ratio were used. It allows different species to be distinguished on the basis of nucleotides alone (the use of a reference genome is not necessary).

Main options

Users can select two methods for querying nucleotide databases managed by the `gSeqI` software: (i) the method based on NCBI databases (GenBank and RefSeq) designed to download FASTA and GenBank files associated to various metadata (when available), or (ii) the method based on ENA designed to download particularly FASTQ files (among others). Both methods provide different functionalities allowing users to download various sequence file formats. The tool provides a parser allowing to extract assembly metadata (such as country, PubMed ID, isolation source, host, etc.) from FASTA and GenBank files. Specific string search and regular expressions were used to retrieve metadata (further details are provided in `Supplementary_tools` scripts “`nucleScore`.”

pl” and “genbank_info.pl”). Some options allow users to determine several features of their search. Users can select the sequence database from which they want to download genomic data. For example, users can choose a NCBI sequence database with the option “-directory” or “-dir” followed by the database of interest (i.e. “genbank” or “ref-seq”). Once the database has been selected, the user can indicate the species with option “-species” followed by the wanted species (e.g. *Escherichia coli*). Users can also query the sequence database by indicating a NCBI Taxonomy ID (Taxid) with the option “-taxid”. Furthermore, users can limit the number of NCBI assemblies they want to download (with the option “-n”). Regarding options dedicated to ENA sequence database, users can choose the option “-ena” to download FASTA sequence records from ENA according to an accession number (<https://ena-docs.readthedocs.io/en/latest/retrieval/general-guide/data-classes.html>) or the option “-fastq” to download compressed FASTQ files obtained from the ENA run accession numbers (starting with “ERR...” or “SRR...”) provided (<https://ena-docs.readthedocs.io/en/latest/submit/general-guide/accessions.html>). Further options are available for both querying methods (i.e. based on NCBI databases or on ENA) and are visible in the provided user manual (https://github.com/karubiotools/getSequenceInfo/blob/master/User_Manual.pdf).

Metrics and NucleScore calculation

Common and customized genome metrics (GC content, AT/GC ratio, nuc%, and variance of nuc%) were used to calculate the NucleScore in order to potentially assess genome assembly quality and completeness.

The GC content is calculated as follows: $GC\ content = \frac{G+C}{A+T+C+G} \times 100$.

AT/GC ratio is calculated as follows: $AT/GC\ ratio = \frac{A+T}{G+C}$.

The percentage for each nucleotide (nuc%) is calculated as follows: $nuc\% = \frac{nuc}{A+T+C+G} \times 100$.

The variance (Var) was calculated taking into account the nuc% value for each nucleotide using the basic formula ([https://fr.wikipedia.org/wiki/Variance_\(math%C3%A9matiques\)](https://fr.wikipedia.org/wiki/Variance_(math%C3%A9matiques))).

The NucleScore was calculated taking into account the variance Var, the GC content, the AT/GC ratio, and the total sequence length (using log2 and square root to tentatively normalize the result) as follows:

$$Nucle\ Score = \log_2 \left(\frac{Var \times GC\ content \times ATG\ Cratio^3}{\sqrt{length}} \right)$$

Length variable represents the total size of a given genome assembly.

A graphical user interface (GUI) has been developed to facilitate the utilization of the program (Fig. 1).

Examples of use

Note that the ‘\$’ symbol indicates a shell prompt (or command terminal).

- (i) Coronavirus genome assemblies (example of 50 genomes available from December 1st 2019)

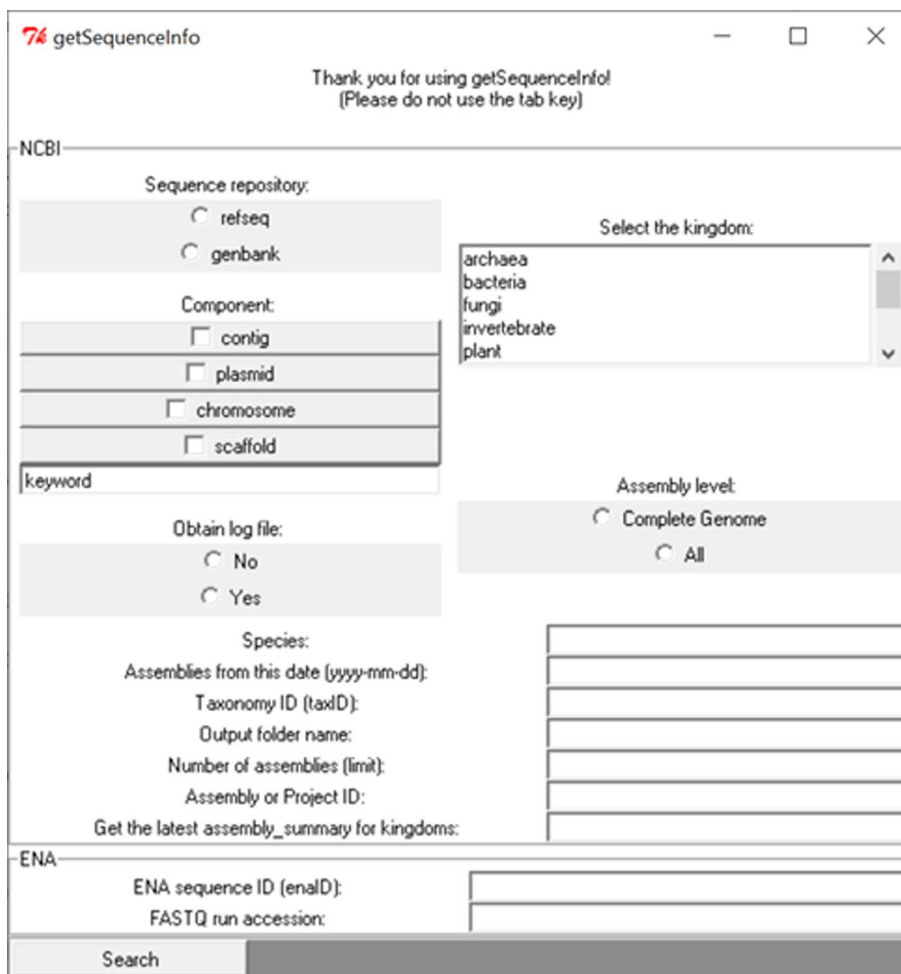


Fig. 1 Overview of the GUI of getSequenceInfo

```
$ perl getSequenceInfo.pl -s coronavirus -k viral -date 2019-12-01 -n 50 -o COVID19.
```

Please note that nowadays, a wide range of specific tools have been developed for querying SARS-CoV-2 data such as this NCBI resources page: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

- (ii) *E. coli* complete genome assemblies from RefSeq and their associated plasmids (example of 50 genomes)

```
$ perl getSequenceInfo.pl -k bacteria -s "Escherichia coli" -d refseq -level "Complete Genome" -c plasmid -n 50.
```

- (iii) 5 sequences belonging to *Naegleria fowleri* species from GenBank

```
$ perl getSequenceInfo.pl -k protozoa -s "Naegleria fowleri" -d genbank -n 5.
```

- (iv) FASTA sequences and associated text reports from ENA

```
$ perl getSequenceInfo.pl -ena GCA_000195955,BN000065.
```

- (v) FASTQ run accessions from ENA (example with four Brazilian *M. bovis* genomes)

```
$ perl getSequenceInfo.pl -fastq SRR7693912,SRR7693877,SRR9850824,SRR9850830.
```

Supplementary tools

A set of additional or complementary tools is also offered to help users analyze their DNA sequences.

These tools are available in a dedicated folder named “Supplementary_tools” on the GitHub repository.

A dedicated program named “nucleScore.pl” was designed to get metrics corresponding to the aforementioned “NucleScore” directly from users’ data. The code can be used as follows (considering a set of FASTA files available in the current repository):

```
$ perl nucleScore.pl *.fasta.
```

The result file will be a tabulated file containing sequence information corresponding to each FASTA file (nucleotide frequencies, GC-content, AT/GC ratio, etc.).

Another program named “countDifferences.pl” allows users to compare sequences from a multi-Fasta alignment file by calculating differences (in bp) and percentage identity. The code is used as follows:

```
$ perl countDifferences.pl multi_alignment.fasta.
```

It generates percentage_identity and distance matrices from the given multi-FASTA alignment input file.

Then, a program named “SRARunInfo.pl” allows users to get information on running accessions using ID (e.g. SRR7693912). The program can be runned as follows:

```
$ perl SRARunInfo.pl SRR7693877,SRR9850824,SRR9850830.
```

A result summary information corresponding to the query is generated (Additional file 1). In addition, XML and CSV result files are generated for each queried accession.

“removeChar.pl” tool allows to remove positions (or columns) from a multi-Fasta alignment file in function of a given character (e.g. ‘N’ or ‘-’). The program can be used as follows:

```
$ perl removeChar.pl alignment.fasta -
```

The result will be a multi-Fasta alignment file without the given character (‘-’) in the sequences.

Availability through the Galaxy KaruBioNet instance

Partial functionalities of the getSequenceInfo and supplementary tools have been made available through the Galaxy KaruBioNet instance (http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html) [11]. Users who are not comfortable with the command line interface (CLI) or the GUI can use Galaxy [12] to perform the analysis in an even more user-friendly way. Furthermore users can easily register and login to the website providing an email address and a password. A screenshot of the welcome page of the Galaxy KaruBioNet highlighting the tool suite is shown in Fig. 2. Another tool named catchSequenceInfo (available in our Galaxy instance) allows to get resistance, virulence, plasmids, and multilocus sequence typing (MLST) [13] information from 2,518 complete genome assemblies (mainly collected from RefSeq). This tool uses: (a) ABRicate (<https://github.com/tseemann/abricate>) with ResFinder [14], PlasmidFinder [15], and VFDB [16] databases to predict resistance, plasmid, and virulence genes; as well as (b) MLST tool (<https://github.com/tseemann/mlst>). A percentage coverage of 90.00% was used to screen resistance and plasmid genes, whereas a percentage coverage of 80.00% was

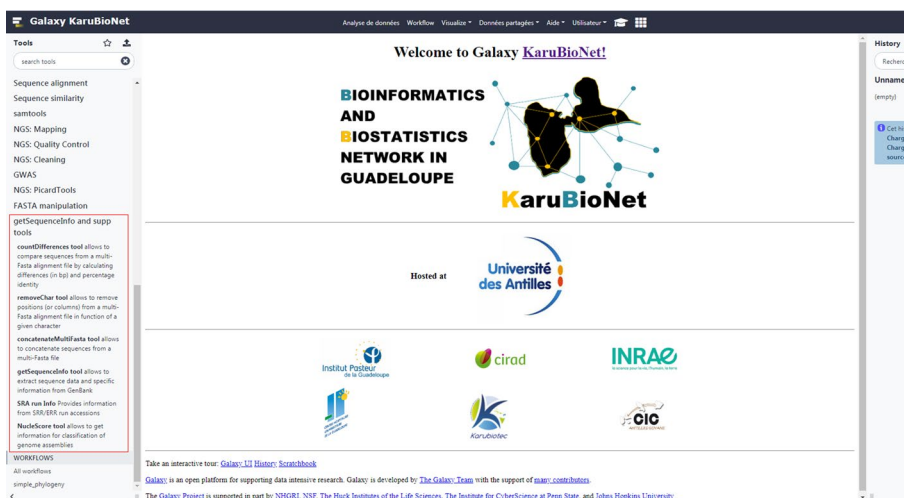


Fig. 2 Galaxy KaruBioNet screenshot with gSeqI and additional tools framed in red

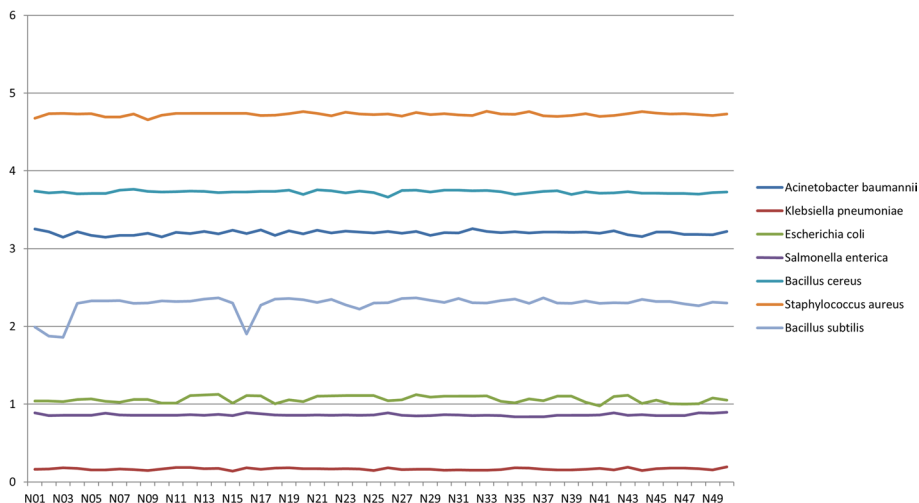


Fig. 3 NucleScore value of 50 genomes belonging to 7 different bacterial species (*Acinetobacter baumannii*, *Bacillus cereus*, *Bacillus subtilis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Salmonella enterica*, *Staphylococcus aureus*)

used to filter virulence genes. getSeqI and catchSequenceInfo tools were used to build a small dataset containing assemblies belonging to *Escherichia coli* (n=1835), *Klebsiella pneumoniae* (n=622) and *Enterobacter cloacae* (n=61) species (results are provided in Additional file 2). Please note that due to our low data storage capacity, the total number of assemblies that can be downloaded has been limited to 50 in Galaxy (users can use the CLI interface to download further data). Note that, after using getSequenceInfo tools from Galaxy, users are invited to delete and purge generated data.

Potential use of NucleScore for bacterial organisms delineation

NucleScore calculation was tentatively used to potentially delineate bacterial organisms.

NucleScore values varied by species as follows (see Fig. 3):

- (i) *Staphylococcus aureus* (min = 4.655; max = 4.768)
- (ii) *Escherichia coli* (min = 0.978; max = 1.125)
- (iii) *Acinetobacter baumannii* (min = 3.146; max = 3.256)
- (iv) *Salmonella enterica* (min = 0.838; max = 0.895)
- (v) *Bacillus cereus* (min = 3.66; max = 3.764)
- (vi) *Bacillus subtilis* (min = 1.858; max = 2.369)
- (vii) *Klebsiella pneumoniae* (min = 0.140; max = 0.195)

The NucleScore metric can potentially be used as a score to assess the quality or completeness of a genome assembly.

We can notice that some species (such as *Staphylococcus aureus* and *Klebsiella pneumoniae*) are easily identifiable thanks to their NucleScore. However, other species may be difficult to identify based on their NucleScore (e.g. scores of *Escherichia coli* and *Salmonella enterica* are very close). This may constitute a limitation to the use of the NucleScore.

Discussion

The getSequenceInfo software tool attempts to allow users to easily download sequences as well as associated metadata from NCBI's GenBank or RefSeq and EBI's ENA without advanced computing knowledge or complex manipulations.

This tool has been designed to work preferably on a small number of sequences of interest. It could be helpful in designing specific genomic databases or strains collection, by facilitating access to sequences using accession numbers and keywords. For example, users can easily download plasmid sequences from available strains belonging to given species (e.g. *Escherichia coli*).

Although existing sequence downloading tools offer great functionality to download sequences in various manners, getSeqI provides a different way to download sequences making association with several metadata (such as country and host), and allowing users to make specific queries.

The main tool has also been made accessible through a Singularity container to facilitate its utilization (the GitHub page of the tool shows how to install the Singularity image) [17]. The tool will also be deposited in a Docker container (<https://docs.docker.com/>). Prospects for improvement of this tool will be performed. The fact that the software tool is for the moment available as a standalone program (although it is also available through a GUI) could represent a limit for the utilization of the tool. Therefore, efforts have been made to make the getSequenceInfo tool partly accessible through a Galaxy platform. Future developments of getSequenceInfo will consist in adding some functions for improving extraction of useful information and in adding other methods to make the tool more accessible. Moreover, additional programming tools (using machine learning and other techniques) will be applied to improve the developed tools and enhance the prediction/calculation of the NucleScore. Several methods exist for the delineation of bacterial genomes [18], and we can draw inspiration from them to improve our methodology. For now, the getSeqI tool can be used to download a small amount of sequence data (<hundreds). The ability to clearly define sequences of interest and the number of sequences that can be downloaded represent the limits of this tool.

However, despite the limits concerning the NucleScore, it nevertheless makes it possible to identify certain species in a fairly simple way. Furthermore, we believe that future investigations and improvements will potentially make this score more relevant.

Some initiatives such as the Nagoya Protocol on Access and Benefit Sharing (https://en.wikipedia.org/wiki/Nagoya_Protocol) and the Convention on Biological Diversity (https://en.wikipedia.org/wiki/Convention_on_Biological_Diversity) could play a role in helping nations for a better managing of biological resources. We promote a fair sharing of bioinformatics resources. Strategies are also needed to make scientific results more accessible and affordable for all interested people (https://en.wikipedia.org/wiki/Open_science). In our opinion, it is important to facilitate the understanding of biological sequence analysis for a large audience by making/sharing less complex, simpler, low-cost, reproducible and accessible methodologies.

Conclusion

To conclude, `getSequenceInfo` offers an accessible way to get genome sequences (in FASTA, FASTQ or GenBank formats) and associated metadata in a programmatic way (command line) or using a graphical user interface (GUI). The software tool allows users to get multi-level information from queried genome assembly such as country of origin/isolation, release date, host, etc. (using NCBI GenBank or RefSeq databases). `getSequenceInfo` could allow users to retrieve sequencing data in the function of a given kingdom or species, and from a given date. The separation between chromosomes and plasmids (or other elements/components) by ranging corresponding sequences in a given folder, is also possible. Some basic statistics are also performed by the program (such as the calculation of GC content for queried assemblies). Furthermore, an empirically designed nucleotide ratio (“NucleScore”) is proposed to tentatively provide a score explaining nucleotide diversity of queried genome assemblies. Supplementary tools are provided to help users with some occasional needs for sequence analysis and data retrieval. `getSequenceInfo` and tools are partly available from the recently released Galaxy KaruBioNet platform (http://calamar.univ-ag.fr/c3i/galaxy_karubionet.html). Finally, `getSequenceInfo` could be used as an educational tool allowing students to better comprehend sequence data.

Availability and requirements

Lists the following:

- Project name: `getSequenceInfo`
- Project home page: <https://github.com/karubiotools/getSequenceInfo>
- Operating system(s): Platform independent
- Programming language: Perl
- Other requirements: e.g. Perl 5.26 or higher, Perl modules
- License: e.g. GNU GPLv3.
- Any restrictions to use by non-academics: licence needed

Abbreviations

gSeqI	GetSequenceInfo tool
GUI	Graphical user interface
CLI	Command-line interface
nuc	Nucleotide
GC-content	Guanine-cytosine content

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-022-04809-5>.

Additional file 1: Result dataset obtained using "SRArunInfo.pl" for 3 accessions (SRR7693877, SRR9850824, and SRR9850830).

Additional file 2: Result datasets corresponding to 2,518 complete genome assemblies belonging to *E. coli*, *K. pneumoniae*, and *E. cloacae* from RefSeq or GenBank repositories. The first sheet shows results obtained from gSeqI, the second sheet provides catchSequenceInfo results, and the following sheets provide Sequence Types (ST) distribution for each species.

Acknowledgements

We are grateful to Nalin Rastogi for helpful discussions and reading of this manuscript. We thank Erick Stattner and Wilfried Segretier (from LAMIA laboratory of the Université des Antilles) for their help regarding the NucleScore. We also thank Vincent Guerlais, Isaure Quérel, Géliza Gamiette, Gaëlle Gruel, Youri Vingataramin and Degrâce Batantou (as well as other members of the TReD-Path unit) for their help.

Author contributions

VM, DCa, MG and DCo conceived the software; MP, SB, SF, IM, AT, SGR, YR and AD tested the software; MP, SB, SF, YR, AD and DCo wrote the manuscript with contributions from all authors. All authors read and approved the final manuscript.

Funding

This study was partly conducted in the framework of the MALIN project (<https://www.projet-malin.fr/>), grant number 2015-FED-186, supported by the European Union in the framework of the European Regional Development Fund (ERDF) and the Regional Council of Guadeloupe. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All data is available on the GitHub repository (<https://github.com/karubiotools/getSequenceInfo>). A folder named "datasets" contains additional files that were used in the manuscript as output examples.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 26 October 2021 Accepted: 23 June 2022

Published online: 08 July 2022

References

1. Karsch-Mizrachi I, Takagi T, Cochrane G. International nucleotide sequence database collaboration. *Int Nucleotide Seq Database Collab Nucleic Acids Res.* 2018;46:D48–51. <https://doi.org/10.1093/nar/gkx1097>.
2. Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T. DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res.* 2020;48:D45–50. <https://doi.org/10.1093/nar/gkz982>.
3. Amid C, Alako BT, BalavenkataramanKadhirvelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martinez-Villacorta J, Milano A, Pakseresht A, Rahman N, Rajan J, Reddy K, Richards E, Smirnov D, Sokolov A, Vijayaraja S, Cochrane G. The European nucleotide archive in 2019. *Nucleic Acids Res.* 2020;48:D70–6. <https://doi.org/10.1093/nar/gkz1063>.
4. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. *Nucleic Acids Res.* 2020;48:D84–6. <https://doi.org/10.1093/nar/gkz956>.
5. Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRADB: query and use public next-generation sequencing data from within R. *BMC Bioinform.* 2013;14:19. <https://doi.org/10.1186/1471-2105-14-19>.
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 'tHoen PA, Hoof R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone

- ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. <https://doi.org/10.1038/sdata.2016.18>.
7. Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A, DiCuccio M, Murphy TD, Pruitt KD, Kimchi A. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res*. 2016;44(D1):D73–80. <https://doi.org/10.1093/nar/gkv1226>.
 8. Ciufu S, Kannan S, Sharma S, Badretin A, Clark K, Turner S, Brover S, Schoch CL, Kimchi A, DiCuccio M. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol*. 2018;68:2386–92. <https://doi.org/10.1099/ijsem.0.002809>.
 9. Amid C, Pakserehsht N, Silvester N, Jayathilaka S, Lund O, Dynovski LD, Pataki B, Visontai D, Xavier BB, Alako BTF, Belka A, Cisneros JLB, Cotten M, Haringhuizen GB, Harrison PW, Höper D, Holt S, Hundahl C, Hussein A, Kaas RS, Liu X, Leinonen R, Malhotra-Kumar S, Nieuwenhuijse DF, Rahman N, Dos Ribeiro SC, Skiby JE, Schmitz D, Stéger J, Szalai-Gindl JM, Thomsen MCF, Cacciò SM, Csabai I, Kroneman A, Koopmans M, Aarestrup F, Cochrane G. The COMPARE data hubs. *Database (Oxford)*. 2019;2019:baz136. <https://doi.org/10.1093/database/baz136>.
 10. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korfi I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E. The bioperl toolkit: perl modules for the life sciences. *Genome Res*. 2002;12(10):1611–8. <https://doi.org/10.1101/gr.361602>.
 11. Couvin D, Dereeper A, Meyer DF, Noroy C, Gaete S, Bhakkan B, Poulet N, Gaspard S, Bezault E, Marcelino I, Pruneau L, Segretier W, Stattner E, Cazenave D, Garnier M, Pot M, Tressières B, Deloumeaux J, Breurec S, Ferdinand S, Gonzalez-Rizzo S, Reynaud Y for the KaruBioNet Team. KaruBioNet: a network and discussion group for a better collaboration and structuring of bioinformatics in Guadeloupe (French West Indies). *Bioinform Adv*. 2022;2(1):vbac010. <https://doi.org/10.1093/bioadv/vbac010>
 12. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*. 2016;44:W3–10. <https://doi.org/10.1093/nar/gkw343>.
 13. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. 2018;3:124. <https://doi.org/10.12688/wellcomeopenres.14826.1>.
 14. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*. 2012;67(11):2640–4. <https://doi.org/10.1093/jac/dks261>.
 15. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, MøllerAarestrup F, Hasman H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. 2014;58(7):3895–903. <https://doi.org/10.1128/AAC.02412-14>.
 16. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res*. 2019;47(D1):D687–92. <https://doi.org/10.1093/nar/gky1080>.
 17. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. *PLoS ONE*. 2017;12(5):e0177459. <https://doi.org/10.1371/journal.pone.0177459>.
 18. Maderankova D, Jugas R, Sedlar K, Vitek M, Skutkova H. Rapid bacterial species delineation based on parameters derived from genome numerical representations. *Comput Struct Biotechnol J*. 2019;17:118–26. <https://doi.org/10.1016/j.csbj.2018.12.006>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

