SOFTWARE TOOL ARTICLE

bit: a multipurpose collection of bioinformatics tools [version 1; peer review: 2 not approved]

Michael Lee

Blue Marble Space Institute of Science, Seattle, Washington, USA

V1 First published: 31 Jan 2022, 11:122
<https://doi.org/10.12688/f1000research.79530.1>
Latest published: 31 Jan 2022, 11:122
<https://doi.org/10.12688/f1000research.79530.1>

Abstract

bit is a collection of small scripts and programs that facilitate many common tasks in bioinformatics. It operates in a Unix-like command-line environment and is comprised of bash and python code. *bit* is openly available on GitHub, archived with Zenodo, and is conda installable. The package is useful for users who want to do things such as manipulate fasta files, calculate GC content, quickly summarize nucleotide assemblies, easily download assemblies from NCBI just based on accessions, pull amino-acid sequences from GenBank files, calculate Shannon uncertainty for columns in multiple sequence alignments, and more. The source code is hosted on GitHub: github.com/AstroBioMike/bit

Keywords

bioinformatics, toolkit, command-line

This article is included in the **Bioinformatics** gateway.

This article is included in the **Python** collection.

Open Peer Review

Approval Status

	1	2
version 1 31 Jan 2022	 view	 view
1. Kai Zhang , Ludwig Institute for Cancer Research, La Jolla, USA		
2. Georges Hattab , University of Marburg, Marburg, Germany		
Any reports and responses or comments on the article can be found at the end of the article.		

Corresponding author: Michael Lee (mikelee@bmsis.org)

Author roles: Lee M: Software, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2022 Lee M. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Lee M. **bit: a multipurpose collection of bioinformatics tools [version 1; peer review: 2 not approved]**
F1000Research 2022, 11:122 <https://doi.org/10.12688/f1000research.79530.1>

First published: 31 Jan 2022, 11:122 <https://doi.org/10.12688/f1000research.79530.1>

Introduction

There are of course several great and widely used packages of bioinformatics helper programs already available. Some of these include the likes of seqtk,¹ fastX-toolkit,² and bbtools³ – all of which I use regularly and have facilitated goals I was trying to accomplish. But there are always more tasks that crop up that may not yet have a helper program or script already written to accomplish them. *bit* is a collection of small scripts and programs that were not written for any single piece of research work. Rather it is a collection that has been built (and is still being built) over several years. Anytime I need to write something to perform a task that has more than a one-off ad hoc use, something I end up using frequently, I consider adding it to the *bit* package. Some programs are light wrappers that extend and/or simplify the utility of existing software (like taxonkit⁴ and goatools⁵); many are written in Python leveraging the Biopython⁶ module (e.g. programs to summarize assemblies, calculate gc content, calculate Shannon uncertainty per column in multiple sequence alignments, pulling amino-acid sequences from GenBank files); and many are bash scripts to do things like download any assembly in

```
# looking at help menu
$ bit-dl-ncbi-assemblies -h

----- HELP INFO -----

This program downloads assembly files for NCBI genomes. It takes as input
assembly accessions (either GCA_* or GCF_*) and optionally a specification of
which format to download. For version info, run `bit-version`.

Required input:

  - [-w <file>] single-column file of NCBI assembly accessions

Optional arguments include:

  - [-f <str>] default: genbank
                Specify the desired format. Available options currently
                include: genbank, fasta, protein, gff, feature_tab, report, stats.

  - [-j <int> ] default: 1
                The number of downloads you'd like to run in parallel. Write speeds
                can become problematic generally with around 15 or more, so a max
                of 12 will be used to be safe even if more are requested.

Example usage:

  bit-dl-ncbi-assemblies -w ncbi_accessions.txt -f protein -j 4

# creating an input file holding the wanted accessions
$ printf "GCF_000005845.2\nGCA_000008865.2\n" > wanted-accessions.txt

# viewing file just made
$ cat wanted-accessions.txt
GCF_000005845.2
GCA_000008865.2

# running program
$ bit-dl-ncbi-assemblies -w wanted-accessions.txt -f fasta -j 2

Targeting 2 genomes in fasta format.

DONE!

# viewing new output files
$ ls *.gz
GCA_000008865.2.fa.gz GCF_000005845.2.fa.gz
```

Figure 1. Example accessing the help menu and using the program for downloading genome assemblies from NCBI.

different file formats from NCBI⁷ just by providing a list of wanted accessions. It is a rather random collection, but it is of convenience to many users.

Methods

Implementation

The package is written in Bash and Python (3+), and is built to run in a Unix-like environment.

Operation

bit is packaged in conda,⁸ which serves as its primary means of installation. All dependencies are handled by the conda installation, but they include: python v3+; biopython⁶ v1.7.9+; pybedtools⁹ v0.8.2+; GNU parallel¹⁰ v20211022+; pandas¹¹ v1.3.4+; entrez direct¹² v16.2+; taxonkit⁴ v0.9.0+; goatools⁵ v0.8.12.

Use cases

All commands are prefixed with ‘bit-’ and so can be seen by typing that and hitting tab twice. Each comes with a help menu by running the command with no arguments or with ‘-h’. In [Figure 1](#) is an example with the program for downloading genome assemblies from NCBI by providing accessions.

Software availability

Source code available from: <https://www.github.com/AstrobioMike/bit>.

Archived analysis code as at time of publication: <https://doi.org/10.5281/zenodo.3383647>.

License: GNU GPL v3.0.

References

1. Li H: **seqtk**. [Reference Source](#)
2. Hanon GJ: **FastX-Toolkit**. [Reference Source](#)
3. Bushnell B: **BBtools**. [Reference Source](#)
4. Shen W, Ren H: **TaxonKit: a practical and efficient NCBI taxonomy toolkit**. *J. Genet. Genomics*. 2021; **48**(9): 844–850.
5. Klopfenstein DV, Zhang L, Pedersen BS, *et al.*: **GOATOOLS: A Python library for Gene Ontology analyses**. *Sci. Rep.* 2018; **8**: 1–17.
6. Cock PJA, Antao T, Chang JT, *et al.*: **Biopython: Freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009; **25**: 1422–1423. [Publisher Full Text](#)
7. Leary NAO, Wright MW, Brister JR, *et al.*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. 2016; **44**: 733–745.
8. Anaconda-Team: **Anaconda Software Distribution**. 2016.
9. Dale RK, Pedersen BS, Quinlan AR: **Pybedtools: A flexible Python library for manipulating genomic datasets and annotations**. *Bioinformatics* 2011; **27**: 3423–3424. [PubMed Abstract](#) | [Publisher Full Text](#)
10. Tange O: **GNU Parallel**. 2021.
11. Team T pandas development: **pandas**. *Zenodo* 2020.
12. Agarwala R, Barrett T, Beck J, *et al.*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2018; **46**: D8–D13.

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 23 March 2022

<https://doi.org/10.5256/f1000research.83522.r128277>

© 2022 Hattab G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Georges Hattab

Department of Mathematics and Computer Science, University of Marburg, Marburg, Hesse, 35032, Germany

The author has put some effort to compile a set of useful scripts. By looking at the number of references alone, it is clear the author lacks knowledge in the field of bioinformatics as many works are left out or not even mentioned. Aside from this fact, there is no description of the methodologies that are made available. I would invite the author to see what a normal methods section looks like. Although a practical set of scripts is provided and for the most part well written, there is a fundamental gap in describing and reproducing this work. Moreover, I would invite the author to look up relevant and openly available data sets where use cases can be presented and the usefulness can be demonstrated. In its current state, there is need for more than considerable work to improve this submission.

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Artificial Intelligence, Machine Learning, Data Mining, Data Visualization

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 22 February 2022

<https://doi.org/10.5256/f1000research.83522.r122738>

© 2022 Zhang K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

Kai Zhang

Ludwig Institute for Cancer Research, La Jolla, CA, USA

The author reports a collection of scripts for various bioinformatics tasks. However, the manuscript is too short and lacks important details for me to assess the significance and novelty of this work. The rationale for this study is unclear. The description of the tool is not enough. I suggest the author add more details to the Methods section and perform a thorough comparison with related tools/packages.

Specific comments:

Need to be more clear about the problems or tasks that this toolkit tries to address. How efficient are these scripts? What are the expected input sizes for your scripts? Have you done any comparison with other similar packages?

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

No

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, single-cell genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research