SoftwareX 14 (2021) 100702

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

omicR: A tool to facilitate BLASTn alignments for sequence data

Berenice Talamantes-Becerra [a,b,*], Jason Carling [b], Arthur Georges [a]

[a] Institute for Applied Ecology, University of Canberra, ACT 2601, Australia
[b] Diversity Arrays Technology Pty Ltd, Canberra ACT 2617, Australia

ARTICLE INFO

Article history:
Received 28 December 2020
Received in revised form 19 April 2021
Accepted 26 April 2021

Keywords:
BLASTn
Sequencing
Genotyping-by-sequencing
Software
Discontiguous megaBLAST

ABSTRACT

Bioinformatics tools for the analysis of sequencing data, are becoming accessible for most scientists. Beginners who are unfamiliar to these tools can be overwhelmed when learning to handle large sequencing datasets. We announce omicR for Windows, which is a user-friendly tool with a graphical user interface that creates fastA files from sequencing data in tabular format such as genotyping-by-sequencing data. OmicR downloads genomes or other sequence sets from the NCBI web server and creates a genome database from the selected references. Subsequently, the user query sequences are aligned to the references and the alignment results are filtered, selecting the best match per sequence.

Code metadata

Current code version	*V1*
Permanent link to code/repository used of this code version	*Executable files for Windows:* *OmicR in GitHub:* https://github.com/ElsevierSoftwareX/SOFTX-D-21-00001 *or* *OmicR in FigShare:* https://doi.org/10.6084/m9.figshare.14431469.v1
Code Ocean compute capsule	*Code Ocean not available for GUI.*
Legal Code License	*Apache-2.0 License*
Code versioning system used	*none*
Software code languages, tools, and services used	*Python*
Compilation requirements, operating environments & dependencies	*ncbi-blast-2.7 (makeblastdb.exe. blastn.exe)*
If available Link to developer documentation/manual	*OmicR in GitHub:* https://github.com/BTalamantesBecerra/omicR_for_Windows *or* *OmicR in FigShare:* https://doi.org/10.6084/m9.figshare.14431469.v1
Support email for questions	Berenice.TalamantesBecerra@canberra.edu.au

Software metadata

Current software version	*V1*
Permanent link to executables of this version	*OmicR in GitHub:* https://github.com/BTalamantesBecerra/omicR_for_Windows *or* *OmicR in FigShare:* https://doi.org/10.6084/m9.figshare.14431469.v1
Legal Software License	*Apache-2.0 License*
Computing platforms/Operating Systems	*Microsoft Windows 7 or above.*
Installation requirements & dependencies	*BLAST+ The BLAST+ latest version can be downloaded here:* https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ *The user manual of BLAST+ can be found here:* https://www.ncbi.nlm.nih.gov/books/NBK279684/
If available, link to user manual - if formally published include a reference to the publication in the reference list	*OmicR in GitHub:* https://github.com/BTalamantesBecerra/omicR_for_Windows *or* *OmicR in FigShare:* https://doi.org/10.6084/m9.figshare.14431469.v1
Support email for questions	Berenice.TalamantesBecerra@canberra.edu.au

* Corresponding author at: Institute for Applied Ecology, University of Canberra, ACT 2601, Australia.
E-mail address: Berenice.TalamantesBecerra@canberra.edu.au (Berenice Talamantes-Becerra).

https://doi.org/10.1016/j.softx.2021.100702

1. Motivation and significance

The access to high-throughput sequencing technologies has enabled the generation of large amounts of sequence data, including for use in molecular marker technologies such as genotyping-by-sequencing techniques [1].

An increasing number of software tools are available for alignment of nucleotide sequences, which often require familiarity with Linux or Windows command line tools to be able to run an analysis [2,3]. Some user-friendly tools already exist for this purpose [4–6]. One of the main distinctions of omicR is that it offers the ability to perform BLASTn alignments directly from sequence data, present within tabular formatted datasets as are commonly encountered in molecular marker technologies. OmicR allows the user to input the tabular dataset after which the sequences will be extracted and formatted for BLASTn alignment. BLASTn alignment is then performed and the alignment outputs are filtered, then inserted into the original tabular data file. Additionally, the filtering provided in omicR differs from that which is available from other GUI and it is particularly suited for sorting and filtering the alignments to provide the best alignment according to the user selection criteria and needs, and return this alignment result to the user within their data table. The tool creates fasta files, downloads genomes and has the option for running discontiguous megaBLAST.

Here we announce omicR, a user-friendly tool to perform BLASTn alignment of sequences against a public or private database built from, contigs, scaffolds, genomes assemblies, or any other nucleotide sequence to be used as reference. OmicR was written in Python [7], it is made for Windows and requires only BLAST+ [8] to be installed. The Python software is presented as a Windows executable, allowing windows users to run this software without installing Python or configuring any additional Python libraries. The BLAST+ executable is available in a Windows executable package from NCBI and can be installed without specialist IT capabilities. The BLASTn results obtained allow for similar filtering criteria as those used in the bioinformatics pipeline Currito3.1 [9], which is designed to select the best matching candidate genome for sequences derived from bacterial isolates [10] and for discovery and identification of novel bacteria [11]. The software selects the best match for each query sequence based on percentage overlap, bitscore and percentage identity.

The outline of the typical steps followed in an analysis is provided below. This software will facilitate the process of BLASTn analysis for students, professionals and researchers who are non-experts in bioinformatics. The software includes a user manual and video tutorials suitable for beginners.

2. Software description

The software provides an intuitive user interface to the NCBI+ software tools, specifically allowing users to build BLASTn databases and perform queries against them without requiring knowledge of command lines processes. Additionally, the software performs filtering of BLASTn alignment to provide results meaningful to their alignment task. The software is aimed at users with limited or no experience in running command line tools, it also facilitates downloading datasets from the NCBI website for input to the BLASTn library building function. Many users are familiar with working with sequences presented in a data table as opposed to a fastA file. The software is designed to accept sequences present in data table format such as molecular marker genotype tables for use in BLASTn queries. The filtered outputs of the BLASTn alignment results are then returned to the original tabular format. The size of the dataset and genomes which can be aligned with this tool is limited only by the amount of disk and memory space available on the hardware on which it is run according to the performance principles of BLAST+ software. The details of BLAST+ software can be found in the NCBI website. The performance of the reference download component is determined by the characteristics of the Entrez [12] query server, it has been noted that downloading of large sequence sets using the Entrez query can be slower than directly downloading these sequence sets via the NCBI website. For this reason, users may wish to choose the available option of directly downloading large datasets.

2.1. Software architecture

The software is designed to perform series of steps (Fig. 1) beginning with creating a fastA file from sequences present in a data table, downloading target sequences for BLASTn analyses, building BLASTn databases from the target sequences, performing BLASTn analyses, filtering alignment output results and returning filtered alignment results to a tabular data format. The user can however perform any of these steps independently by running portions of the software separately by selecting the appropriate module from the user interface.

2.2. Software functionalities

2.2.1. Data format and query sequence

The software accepts query sequences formatted as a tabular comma-separated file (csv). Each row in the file (after the headers) should represent one query sequence. Additional values relating to each sequence may be present as columns delimited by commas. This flexible format allows sequences to be read in most genotyping-by-sequencing data formats.

omicR ensures that each sequence is uniquely identified by adding an identifier number (uniqueID) to each row and then creates 2 files for use in the BLASTn analysis: a fastA file containing the uniqueID and the query sequence, and secondly, a copy of the original csv input file, with an extra column containing the uniqueID.

Alternatively, if the query sequence data is not in tabular format, omicR can be run directly from a fastA file provided by the user. When query sequence data is presented in tabular format, the BLASTn results will be appended as additional columns to a copy of the input file containing the uniqueID. For further information of examples of input files, refer to the user guide manual available on GitHub and FigShare [13].

2.2.2. Downloading genomes

Genomes or other sequence sets can be downloaded from the NCBI website (https://www.ncbi.nlm.nih.gov/) using omicR. This option uses the Biopython module Entrez [12]. This module works by fetching data from NCBI and returning results as a handle. To download the data for building a BLASTn database, enter the required RefSeq numbers. All RefSeq accessions entered will then be fetched into a single fastA file and used to build the BLASTn database.

If the desired reference sequence is large, it is recommended to download the RefSeq assembly as a fna file (formal 3 letter extension for fastA files) using your web browser. This step can be omitted if the reference sequence fastA file is available locally.

2.2.3. Database creation

Creating the NCBI database for BLASTn is a mandatory step to run this software. This script uses the makeblastdb program from NCBI BLAST+ to create a database via the graphical user interface, therefore BLAST+ must be installed prior running this software. If the database has been created previously, omicR allows the user to select those files for running BLASTn.

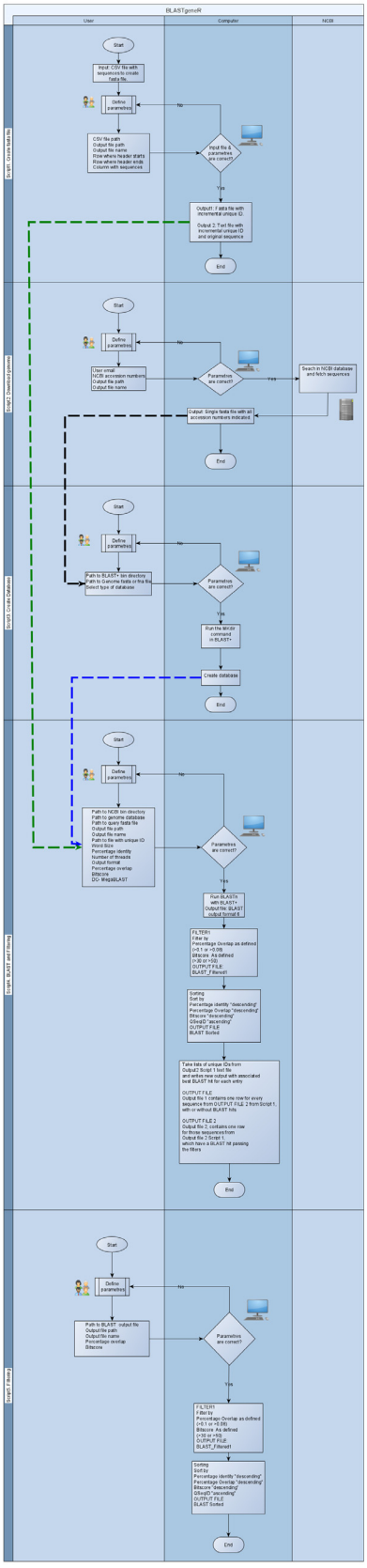

Fig. 1. Diagram of omicR process workflow.

2.2.4. BLASTn alignment and filtering

Running the BLASTn analysis is facilitated by providing the user a graphical user interface (GUI) that includes suggested default parameters to align sequences and filter results to obtain the best match per sequence.

The output of the BLASTn alignment is produced in tabular format. The following columns are output in this format: qseqid (query sequence id), sacc (subject accession number), stitle (subject title), qseq (aligned part of query sequence), sseq (aligned part of subject sequence), nident (number of identical matches), mismatch (number of mismatches), pident (percentage of identical matches), length (alignment length or sequence overlap), evalue (expect value), bitscore (bit score), qstart (start of alignment in query), qend (end of alignment in query), sstart (start of alignment in subject), send (end of alignment in subject), gapopen (number of gap openings), gaps (total number of gaps), qlen (query sequence length), slen (subject sequence length). Note that the calculation of the percentage overlap is done by omicR. The PercentageOverlap column, included in the output file, is calculated as the ratio of the alignment length divided by the query length or subject length, whichever is shortest of these two lengths. During filtering, hits with percentage overlap values lower than the chosen threshold are removed.

The recommended filtering parameters for BLASTn analysis of sequences from similar species are: word size 11, percentage identity 70%, percentage overlap 80% and bitscore 50. If BLASTn is applied for alignment of highly dissimilar sequences, it is recommended to select discontiguous megaBLAST, reducing the percentage overlap to 1% and bitscore to 30 for a less stringent analysis. Other BLAST+ parameters remain at default settings.

The BLASTn alignment script generates 5 output files: The first file has the raw BLASTn output, without headers and without filtering. The second file has a header included to provide the identity for each column, and an additional column which includes the calculated percentage overlap for each alignment. These files can potentially contain multiple alignments per query sequence. In this case, each alignment is represented by a single row. This second file is filtered according to the thresholds selected at run time. Only alignments exceeding these thresholds would be present in this file. The third file contains only the selected best match for each query sequence. This file contains one row for every query sequence including those with no BLASTn hit found or no BLASTn hit greater than the selected threshold. If the input data was provided only in fastA format, this is the final result. The fourth file is identical to the third, with the exception that it contains only sequences with a BLASTn hit with the selected thresholds. The fifth file contains all sequences with BLASTn hits from the third and fourth files, appended to a copy of the original input csv file.

Additional filtering can be done without repeating BLASTn analysis. This step requires the BLASTn output format in a tabular format with the columns ordered in the same order for this software. The filtering allows for the selection of more stringent parameters of percentage overlap and bitscore.

3. Illustrative examples

The omicR tool has a graphical user interface that facilitates BLASTn analyses for people who are unfamiliar to using the terminal command line. Fig. 2, shows the graphical user interface with a simple and intuitive set of buttons corresponding to the set of functions available. The order of the buttons presented in the user interface, follows the sequential order of the data processing scheme, and users can process their data from beginning to end by simply following the steps presented in the interface.

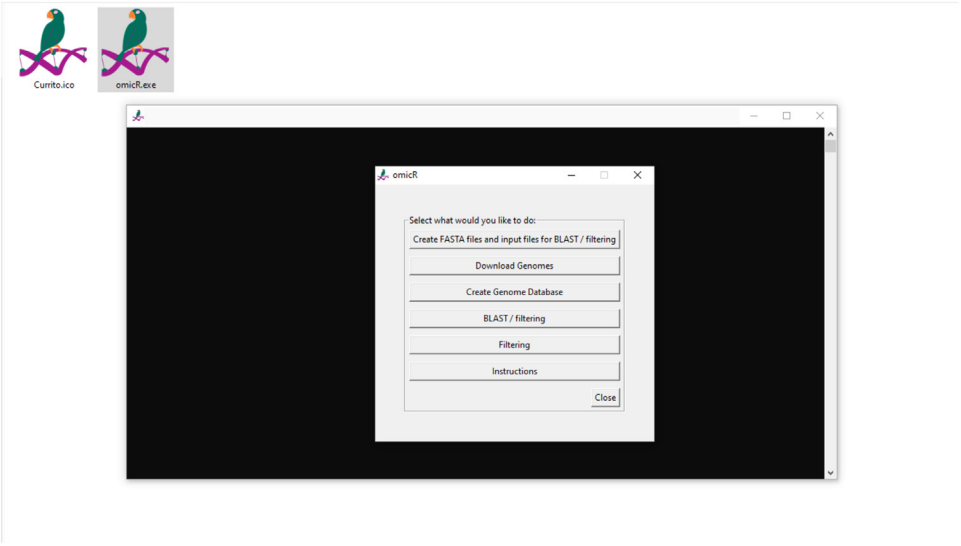

Fig. 2. omicR graphical user interface.

4. Impact

One of the most common analyses required amongst students and researchers who utilize DNA sequencing in their research, is the need to perform alignment of DNA sequence analyses. The majority of postgraduate students studying in the biological sciences do not have the experience and skills needed to perform their own bioinformatic analyses. University departments sometimes employ bioinformaticians but they are unable to meet the demand or sometimes are not available at all. These problems which affect undergraduate students, also affect seasoned researchers who also lack the skills needed to run bioinformatic analyses. Previous alternatives available such as performing BLASTn analyses on the NCBI website, usually limit to a small number of queries using manual copy and paste operations from data tables into web pages, building very unsatisfactory results.

The omicR tool provides a solution for performing BLASTn analyses with complete end to end support, using only simple functions presented in a graphical user interface. Additionally, it performs effective filtering of the BLASTn alignments results, something which is often not available even from existing command line tools. The students who have been given access to omicR have found the freedom to perform all of their BLASTn analyses when required without needed to seek the limited available help.

5. Conclusions

In developing the omicR software, our main aim was to simplify BLASTn analyses for inexperienced users. Although there are options which offer similar alternatives, OmicR allows users to input a tabular dataset from which the sequences are extracted and formatted for BLASTn alignment. After BLASTn alignment is performed, the alignment outputs are filtered and then inserted into the original tabular data file. Inexperienced users may not recognize the importance of filtering of BLASTn results using criteria appropriate for the alignment task at hand.

The software can be used to align sequence data such as genotyping marker sequences to locate SNP markers within the genome. The purpose of this software is to provide an opensource, user-friendly bioinformatic tool for users who need to perform BLASTn alignments of nucleotide sequences against one or more references. The omicR package for Windows users is available for download via GitHub and FigShare [13].

Availability of tools

The BLAST+ latest version can be downloaded here: https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.

The latest version of omicR for Windows can be downloaded here: https://github.com/BTalamantesBecerra/omicR_for_Windows.

An additional description of omicR for Windows, following the MIABi-analysis guidelines has been included in FigShare and can be downloaded here: https://doi.org/10.6084/m9.figshare.14431469.v1.

The BLAST+ user manual created by NCBI fur users to use as reference: https://www.ncbi.nlm.nih.gov/books/NBK279684/.

CRediT authorship contribution statement

Berenice Talamantes-Becerra: Conceptualization, Project administration, Methodology, Software, Visualization, Writing - original draft preparation and editing. **Jason Carling:** Conceptualization, Methodology, Software, Writing - original draft preparation and editing. **Arthur Georges:** Conceptualization, Resources, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to thank our testers Duminda Dissanayake and Sarah Whiteley for testing and providing feedback on this package. The author B. Talamantes-Becerra, would like to acknowledge Consejo Nacional de Ciencia y Tecnología (CONACYT) for providing a scholarship “Becas CONACYT al extranjero 2015” to pursue postgraduate studies.

References

[1] Gruber B, Unmack PJ, Berry OF, Georges A. dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. Mol Ecol Resour 2018;18:691–9. https://doi.org/10.1111/1755-0998.12745.

[2] Kumar S, Dudley J. Bioinformatics software for biologists in the genomics era. Bioinformatics 2007;23:1713–7. https://doi.org/10.1093/bioinformatics/btm239.
[3] Attwood TK, Blackford S, Brazas MD, Davies A, Schneider MV. A global perspective on evolving bioinformatics and data science training needs. Brief Bioinform 2019;20:398–404. https://doi.org/10.1093/bib/bbx100.
[4] Santiago-Sotelo P, Ramirez-Prado JH. prfectBLAST: a platform-independent portable front end for the command terminal BLAST+ stand-alone suite. Biotechniques 2012;53:299–300. https://doi.org/10.2144/000113953.
[5] Du Z, Wu Q, Wang T, Chen D, Huang X, Yang W, et al. BlastGUI: A python-based cross-platform local BLAST visualization software. Mol Inform 2020;39:1900120. https://doi.org/10.1002/minf.201900120.
[6] Priyam A, Woodcroft BJ, Rai V, Moghul I, Munagala A, Ter F, et al. Sequenceserver: A modern graphical user interface for custom BLAST databases. Mol Biol Evol 2019;36:2922–4. https://doi.org/10.1093/molbev/msz185.
[7] van Rossum G, Drake FL. The Python language reference manual. Network Theory Ltd.; 2011.
[8] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
[9] Talamantes-Becerra B, Carling J. Currito3.1 DNA fragment analysis software. 2020, https://doi.org/10.5281/zenodo.3748447.
[10] Talamantes-Becerra B, Carling J, Kennedy K, Gahan M, Georges A. Identification of bacterial isolates from a public hospital in Australia using complexity-reduced genotyping. J Microbiol Methods 2019;160:11–9. https://doi.org/10.1016/j.mimet.2019.03.016.
[11] Talamantes-Becerra B, Carling J, Kilian A, Georges A. Discovery of thermophilic bacillales using reduced-representation genotyping for identification. BMC Microbiol 2020;20:114. https://doi.org/10.1186/s12866-020-01800-z.
[12] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics 2009;25:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
[13] Talamantes-Becerra B, Carling J, Georges A. DATA FROM: omicR: a tool to facilitate BLASTn alignments for sequence data. FigShare 2021. https://doi.org/10.6084/m9.figshare.14431469.v1.