SOFTWARE TOOL ARTICLE

Viewing RNA-seq data on the entire human genome [version 1; referees: 3 approved]

Eric M. Weitz[1], Lorena Pantano[2], Jingzhi Zhu[3], Bennett Upton[4], Ben Busby [1]

[1]National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD, 20894, USA
[2]Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA
[3]Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[4]LSU Shreveport Laboratory for Advanced Biomedical Informatics, Shreveport, LA, 71105, USA

v1 **First published:** 28 Apr 2017, **6**:596 (doi: 10.12688/f1000research.9762.1)
Latest published: 28 Apr 2017, **6**:596 (doi: 10.12688/f1000research.9762.1)

Abstract

RNA-Seq Viewer is a web application that enables users to visualize genome-wide expression data from NCBI's Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) databases. The application prototype was created by a small team during a three-day hackathon facilitated by NCBI at Brandeis University. The backend data pipeline was developed and deployed on a shared AWS EC2 instance. Source code is available at https://github.com/NCBI-Hackathons/rnaseqview.

This article is included in the Hackathons collection.

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	 2	 3
version 1 published 28 Apr 2017	✓ report	✓ report	✓ report

1 **Chase A. Miller**, University of Utah, USA

2 **Christopher J. Fields**, University of Illinois at Urbana–Champaign, USA

3 **Philip Ewels**, Stockholm University, Sweden

Discuss this article

Comments (0)

Corresponding author: Ben Busby (ben.busby@nih.gov)

Competing interests: No competing interests were disclosed.

How to cite this article: Weitz EM, Pantano L, Zhu J *et al.* **Viewing RNA-seq data on the entire human genome [version 1; referees: 3 approved]** *F1000Research* 2017, **6**:596 (doi: 10.12688/f1000research.9762.1)

Grant information: Work on this project by Eric Weitz and Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH)/National Library of Medicine (NLM)/National Center for Biotechnology Information (NCBI).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 28 Apr 2017, **6**:596 (doi: 10.12688/f1000research.9762.1)

Introduction

Interactive visualizations can yield insights from the deluge of gene expression data brought about by RNA-seq technology. Several genome browsers enable users to see such data conveniently plotted within a single chromosome in a web application (Broad Institute, 2014; Kent *et al.*, 2002; National Center for Biotechnology Information: Genome Data Viewer (2016)). While such single-chromosome views excel at displaying local features, depicting RNA-seq data across all chromosomes in a genome, i.e. in an ideogram, has the potential to intuitively highlight global patterns of gene expression (such as in Figure 2a in Parker *et al.*, 2016).

In this paper we describe RNA-Seq Viewer, a web application that enables users to visualize genome-wide expression data from the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) (Kodama *et al.*, 2012) and Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013) databases. The application consists of a backend data pipeline written in Python and a web frontend powered by Ideogram.js, a JavaScript library for chromosome visualization (Weitz, 2015).

The data pipeline, developed by a small team of software engineers in a three-day NCBI hackathon at Brandeis University, extracts aligned RNA-seq data from SRA or GEO and transforms it into a format used by Ideogram. Ideogram then displays the distribution of genes in chromosome context across the entire human genome and enables users to filter those genes by gene type or expression levels in the given SRA/GEO sample.

Methods

The primary task of the hackathon was to develop a prototype data pipeline to extract aligned RNA-seq data from SRA, determine genomic coordinates for the sampled genes, and transform the combined result into the JSON format used by Ideogram.js annotation sets. The formatted annotation data was then plugged into a lightly modified example from the Ideogram repository to provide an interactive, faceted search application for exploring genome-wide patterns of gene expression.

Ideogram.js uses JavaScript and SVG to draw chromosomes and associated annotation data in HTML documents. It leverages D3.js, a popular JavaScript visualization library, for data binding, DOM manipulation, and animation (Bostock *et al.*, 2011). Faceted search in Ideogram is enabled by Crossfilter, a JavaScript library for exploring large multivariate datasets (Square Inc., 2012). By relying only on JavaScript libraries, HTML and CSS, Ideogram can function entirely in a web browser, with no server-side code required, which simplifies embedding ideograms in a web application.

Annotation data for Ideogram leverages space-efficient data structures and the compact nature of JSON to minimize load time in web pages. For example, the gzip-compressed set of 31,148 human gene feature annotations, including data on expression level and gene type, output by our pipeline for SRA run SRR562646 (National Center for Biotechnology: Sequence Read Archive Run Browser) is 399 KB in size and takes less than 285 ms to download on an average US Internet connection (14 Mb/s download bandwidth, 50 ms latency) (Belson *et al.*, 2016) as measured using Chrome Developer Tools (Basques & Kearney, 2016). Under the same network-throttled conditions using Chrome version 51 on a Mac OS X laptop with a 2.9 GHz Intel Core i5 CPU, the Chrome DevTools Timeline tab reports that an uncached, interactive genome-wide histogram of expression for 31,148 gene features takes Ideogram between 830 ms and 1044 ms to completely load and render after the start of navigation to the web page.

Broadly, the pipeline developed to produce Ideogram annotation data works as follows:

1. Get data for an SRR accession from NCBI SRA (National Center for Biotechnology Information: Sequence Read Archive).
2. Count reads for each gene and normalize expression values to TPM units (Wagner *et al.*, 2012)
3. Get coordinates and type for each gene from a GFF file in the NCBI *Homo sapiens* Annotation Release
4. Format coordinates and TPM values for each gene into JSON used by Ideogram.js

The data pipeline exists in two parts: one for data in SRA and one for data in GEO.

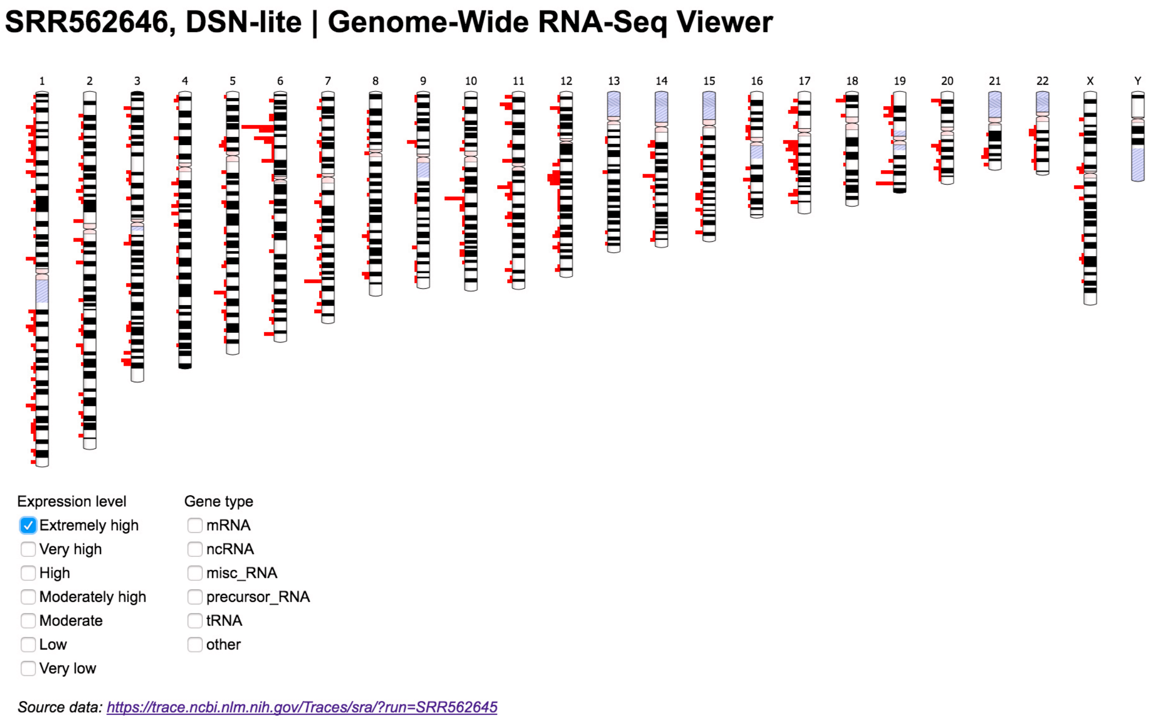
The tool reads a list of SRR accession numbers (National Center for Biotechnology Information, 2011; National Center for Biotechnology Information: SRA Handbook (2011)) and identifies the ones that have alignment. It then retrieves the genome reference used for the creation of the BAM/SAM file to download the gene annotation for quantification. Only genome assemblies GRCh37 (GCA_000001405.1) and GRCh38 (GCA_000001405.15) are supported, and the annotations used for each of them are NCBI *Homo sapiens* Annotation Release 105 and 107, respectively (National Center for Biotechnology Information, 2013; National Center for Biotechnology Information, 2015).

Alternatively, the tool can read a BAM/SAM file in case of local files. In one single command, the tool quantifies gene expression using HTSeq-count version 0.6.1p1 (Anders *et al.*, 2015) after sam-dump version 1.3 (National Center for Biotechnology Information, 2011). To avoid possible errors due to non-standard SAM files, our filtering steps in the middle sort the BAM file and keep only properly paired reads. The output from HTSeq-count is a tabular file, where the first column is the gene symbol and the second is the read counts. Finally, we normalize the expression by the length of the mature transcript using the longest transcript as the size of the gene.

After obtaining TPM values for each gene's expression level (Step 2) as described above, the next step in the pipeline parses genomic coordinates (chromosome name, start and stop) and gene type (e.g. mRNA, ncRNA) from a GFF file in the NCBI *Homo sapiens* Annotation Release. These data are combined with each gene's TPM value, formatted into a compressed JSON structure, and written to a file containing symbols, genomic coordinates, expression levels and gene types for every human gene. This file, e.g. SRR562646.json, represents the final output of the RNA-Seq Viewer data pipeline, and contains all the data used by the fast client-side faceted search in Ideogram.js.

Results

The resulting RNA-Seq Viewer web application prototype was demonstrated at the conclusion of the three-day hackathon at Brandeis University. The application provides an interactive data visualization in which users can filter genes by expression level and gene type across the entire human genome (Figure 1) or within a single chromosome (Figure 2).



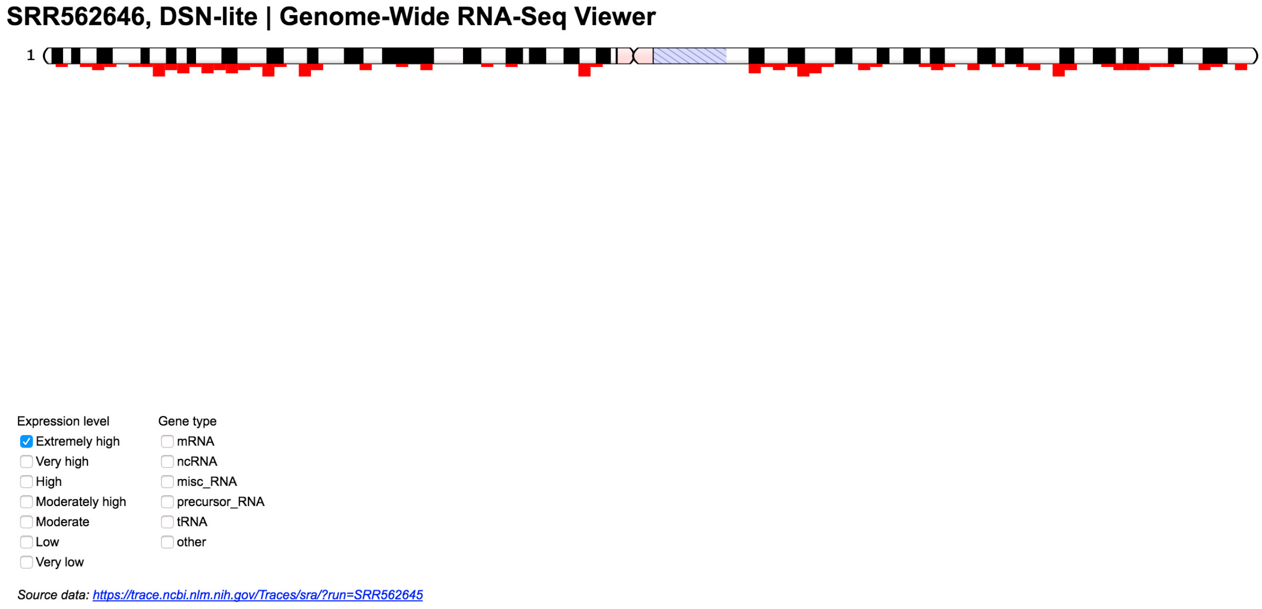
SRR562646, DSN-lite | Genome-Wide RNA-Seq Viewer

Expression level
☑ Extremely high
□ Very high
□ High
□ Moderately high
□ Moderate
□ Low
□ Very low

Gene type
□ mRNA
□ ncRNA
□ misc_RNA
□ precursor_RNA
□ tRNA
□ other

Source data: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR562645

Figure 1. RNA-seq data for all chromosomes can be filtered on demand and viewed en masse.



SRR562646, DSN-lite | Genome-Wide RNA-Seq Viewer

Expression level
☑ Extremely high
□ Very high
□ High
□ Moderately high
□ Moderate
□ Low
□ Very low

Gene type
□ mRNA
□ ncRNA
□ misc_RNA
□ precursor_RNA
□ tRNA
□ other

Source data: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR562645

Figure 2. RNA-seq data for a single chromosome can also be filtered and viewed.

Discussion

The RNA-Seq Viewer prototype demonstrates a pipeline for transforming aligned RNA-seq data from SRA into a format used for genome-wide visualization.

Next steps for this data pipeline include supporting RNA-seq alignment and normalization when using multiple samples, such as from different tissues. Filters for those different tissues could also be added as filters in the display. The resulting genome-wide visualizations could then be embedded in genome browsers, e.g. NCBI Genome Data Viewer (National Center for Biotechnology Information: Genome Data Viewer), or any genomics-oriented application that supports HTML, CSS, and JavaScript.

The prototype implemented in the hackathon only supports RNA-seq datasets from SRA that are already aligned to a reference genome, e.g. GRCh37 or GRCh38. Salmon (Patro *et al.*, 2015) and Kallisto (Bray *et al.*, 2016) are two popular alignment programs that could be used for this task. Both alignment programs can generate gene expression, with low memory and CPU requirements.

Software availability

Latest source code: https://github.com/NCBI-Hackathons/rnaseqview

Archived source code as at the time of publication: https://dx.doi.org/10.5281/zenodo.377055 (Weitz *et al.*, 2017)

License: CC0 1.0 Universal

Author contributions

All of the authors participated in designing the study, carrying out the research, and preparing the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

Competing interests

No competing interests were disclosed.

Grant information

Work on this project by Eric Weitz and Ben Busby was supported by the Intramural Research Program of the National Institutes of Health (NIH)/National Library of Medicine (NLM)/National Center for Biotechnology Information (NCBI).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors thank Francesco Pontiggia and Brandeis University for provisioning computing resources and facilities for development. The authors thank Lisa Federer, NIH Library Writing Center, for manuscript editing assistance. The authors thank Valerie Schneider, NCBI, for insightful comments and suggestions.

References

Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics.* 2015; **31**(2): 166–9.
PubMed Abstract | Publisher Full Text | Free Full Text

Barrett T, Wilhite SE, Ledoux P, *et al.*: **NCBI GEO: archive for functional genomics data sets--update.** *Nucleic Acids Res.* 2013; **41**(Database issue): D991–5.
PubMed Abstract | Publisher Full Text | Free Full Text

Basques K, Kearney M: **Chrome Developer Tools: Network Panel Overview [Online].** 2016; [Accessed: 26 September 2016].
Reference Source

Basques K, Kearney M: **Chrome Developer Tools: Network Panel Overview. [Online].** 2016; [Accessed: 26 September 2016].
Reference Source

Belson D, Thompson J, Sun J, *et al.*: **Q4 2015 State of the Internet Report.** *Akamai Technologies.* 2016; **8**(4).
Reference Source

Bostock M, Ogievetsky V, Heer J: **D^3: Data-Driven Documents.** *IEEE Trans Vis Comput Graph.* 2011; **17**(12): 2301–2309.
PubMed Abstract | Publisher Full Text

Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.
PubMed Abstract | Publisher Full Text

Broad Institute: **Lightweight html5 version of the Integrative Genomics Viewer [Online].** 2014; [Accessed: 26 September 2016].
Reference Source

Kent WJ, Sugnet CW, Furey TS, *et al.*: **The human genome browser at UCSC.** *Genome Res.* 2002; **12**(6): 996–1006.
PubMed Abstract | Publisher Full Text | Free Full Text

Kodama Y, Shumway M, Leinonen R, *et al.*: **The Sequence Read Archive: explosive growth of sequencing data.** *Nucleic Acids Res.* 2012; **40**(Database issue): D54–6.
PubMed Abstract | Publisher Full Text | Free Full Text

National Center for Biotechnology Information: **Genome Data Viewer [Online].** 2016; [Accessed: 26 September].
Reference Source

National Center for Biotechnology Information: **SRA Handbook [Online].** 2011; [Accessed: 26 September 2016].
Reference Source

National Center for Biotechnology Information Sequence Read Archive Run Browser [Online]: **GSM999527: DSN-lite; Homo sapiens; RNA-Seq (SRR562645).** [Accessed: 14 February 2017].
Reference Source

National Center for Biotechnology Information: **Using the SRA Toolkit to convert .sra files into other formats.** In *SRA Knowledge Base.* 2011; [Accessed: 26 September 2016].
Reference Source

National Center for Biotechnology Information: **Homo sapiens Annotation Release 105 [Online].** 2013; [Accessed: 26 September 2016].
Reference Source

National Center for Biotechnology Information: **Homo sapiens Annotation Release 107 [Online].** 2015; [Accessed: 26 September 2016].
Reference Source

Parker CC, Gopalakrishnan S, Carbonetto P, *et al.*: **Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice.** *Nat Genet.* 2016; **48**(8): 919–926.
PubMed Abstract | Publisher Full Text | Free Full Text

Patro R, Duggal G, Kingsford C: **Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference.** *BioRxiv.* 2015.
Publisher Full Text

Square Inc: **Crossfilter [Online].** 2012; [Accessed: 26 September 2016].
Reference Source

Wagner GP, Kin K, Lynch VJ: **Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.** *Theory Biosci.* 2012; **131**(4): 281–285.
PubMed Abstract | Publisher Full Text

Weitz EM: **Ideogram [Online].** 2015; [Accessed: 26 September 2016].
Reference Source

Weitz EM, Pantano L, Zhu J, *et al.*: **NCBI-Hackathons/rnaseqview 1.1.** *Zenodo.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ✓

Version 1

Referee Report 20 September 2017

doi:10.5256/f1000research.10524.r25863

✓ **Philip Ewels**

Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, 106 91, Sweden

This manuscript does a great job of describing RNA-Seq Viewer, a tool to visualise genome-wide co-ordinate based clustering of gene expression in a sample. It's a neat project and great to see such a well polished product coming out of a hackathon. Such a responsive tool is impressive, and the user interface is simple to use.

Minor concerns:

- The abstract says that RNA-Seq Viewer is a *"web application"*. However, to use it users are required to run a series of command line tools to prepare data and then edit a HTML file before getting to the web page. So whilst the tool certainly uses web technologies, I would not say that it's a fully fledged (eg. online only) web application yet. A minor change in wording would be sufficient to clear this up.

- The abstract says that *"The backend data pipeline was developed and deployed on a shared AWS EC2 instance."* - however, this seems to be the only mention of AWS in the manuscript or repository. If the authors mean that they deployed it for a one-off run, I think it's a little misleading (my assumption was that it is running as a service on AWS for anyone to use).
  - Additional documentation as to how users can use AWS to run the tool would also be useful.

- There are example reports in the GitHub repository, but they're not mentioned anywhere that I can see. It would be nice if the readme clearly pointed towards these in the introduction with links using http://rawgit.com so that they can be loaded directly.

Other than this, I think that the manuscript fairly describes the project. I'd love to see the additions that the authors propose at the end (support for multiple samples and use with a wider range of input data) and hope that the manuscript may get a future revision with such additional features!

I see that another reviewer mentions the generic name - I agree that 'RNAideogram' or something else may be a little more specific and useful!

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

Is the description of the software tool technically sound?
Yes

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

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

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

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 18 September 2017

doi:10.5256/f1000research.10524.r25864

✓ **Christopher J. Fields**
High-Performance Biological Computing Group, Roy J. Carver Biotechnology Centre, University of Illinois at Urbana–Champaign, Urbana, IL, USA

The authors have succeeded in integrating multiple disparate software resources into a useful and well-motivated tool, RNA-Seq Viewer, that was largely put together within a three-day hackathon, which is even more impressive and speaks to the strengths of hackathons in general, particularly when there is a clear motivation and goal in mind that play to the strengths of everyone involved. I'm particularly happy to see that NCBI is more actively engaging the open-source community though organization of workshops and events such as this.

One key item: I couldn't find an online example, if there is one available this would be very useful as a live demo and would be particularly useful in garnering feedback, including possible directions for future development.

I should note: there are already a few 'RNA-Seq Viewer' tools out there, not sure if you would need to change the name (if so please no horrible backronyms):

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320319/
http://bioinfo.au.tsinghua.edu.cn/software/RNAseqViewer/

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

Is the description of the software tool technically sound?
Yes

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

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

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

Competing Interests: No competing interests were disclosed.

Referee Expertise: Computational biology, genomics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 08 May 2017

doi:10.5256/f1000research.10524.r22368

✓

Chase A. Miller
Department of Human Genetics, University of Utah, Salt Lake City, UT, USA

The RNA-Seq Viewer tool is well motivated and clearly described in this article. RNA-Seq Viewer does a good job of combining data from multiple public sources into a single coherent visualization and interface. There is a great need for more tools like this that make use of the huge amounts of public genomic data available.

I found an online example of RNA-Seq Viewer here. It would be very useful if this link was included in the Abstract so that potential users can quickly try out the tool.

Although beyond the scope of a 3-day hackathon, in the future it would be valuable to turn this tool into a fully hosted web app so that no download or command line knowledge would be required.

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

Is the description of the software tool technically sound?
Yes

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

Yes

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

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

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.