Check for updates

SOFTWARE TOOL ARTICLE

Extending TCGA queries to automatically identify analogous genomic data from dbGaP [version 1; referees: 2 approved, 1 approved with reservations]

Erin K. Wagner[1], Satyajeet Raje[2], Liz Amos[2], Jessica Kurata[3], Abhijit S. Badve[4], Yingquan Li[5], Ben Busby[6]

[1]BioStat Solutions, Frederick, USA
[2]National Library of Medicine, National Institutes of Health, Bethesda, USA
[3]Department of Molecular and Cellular Biology, City of Hope, Duarte, USA
[4]GeneDx, Gaithersburg, USA
[5]Corporate Executive Board (CEB), Arlington, USA
[6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, USA

v1 **First published:** 24 Mar 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)
Latest published: 24 Mar 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)

Abstract

Data sharing is critical to advance genomic research by reducing the demand to collect new data by reusing and combining existing data and by promoting reproducible research. The Cancer Genome Atlas (TCGA) is a popular resource for individual-level genotype-phenotype cancer related data. The Database of Genotypes and Phenotypes (dbGaP) contains many datasets similar to those in TCGA. We have created a software pipeline that will allow researchers to discover relevant genomic data from dbGaP, based on matching TCGA metadata. The resulting research provides an easy to use tool to connect these two data sources.

This article is included in the Hackathons collection.

Open Peer Review

Referee Status: ? ✓ ✓

	Invited Referees		
	1	**2**	**3**
version 1 published 24 Mar 2017	? report	✓ report	✓ report

1 **Yussanne Ma**, BC Cancer Agency, Canada
Victoria Trinh, BC Cancer Agency, Canada

2 **Konstantinos Krampis**, Hunter College of The City University of New York, USA

3 **Tsung-Jung Wu**, Baylor College of Medicine, USA

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: Wagner EK, Raje S, Amos L *et al.* **Extending TCGA queries to automatically identify analogous genomic data from dbGaP [version 1; referees: 2 approved, 1 approved with reservations]** *F1000Research* 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)

Grant information: Funding was provided by the NIH intramural research program at the NLM. JK was supported in part by the Frances Berger Foundation Fellowship and her lab was supported in part by Ramesh Kesanupalli family and Beckman Research Institute Funds.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 24 Mar 2017, **6**:319 (doi: 10.12688/f1000research.9837.1)

Introduction

Many large funding organizations, including the National Institutes of Health (NIH), encourage researchers to make their data available in public databases. Policies like the NIH's Genomic Data Sharing policy (https://gds.nih.gov/03policy2.html) and other incentives around data sharing have promoted the development of several public data repositories. However, in spite of the availability of data, it can still be challenging to harness the power of these public databases, and researchers are faced with a variety of barriers in accessing shared data (van Schaik *et al.*, 2014).

A major obstacle to data discovery is the disconnectedness of various data sharing resources. Automated tools that can connect these databases and reduce the time that researchers spend on data discovery are critically needed (Dudley & Butte, 2008; Ruau *et al.*, 2011). Such tools will promote reproducibility, increase the efficiency of research, and aid in solving the problem of small sample sizes. These issues are especially relevant to genomic data, which is typically expensive to gather.

Here, we focus on connecting two popular genomic data repositories, the Database of Phenotypes and Genotypes (dbGaP) (Tryka *et al.*, 2014) and The Cancer Genome Atlas (TCGA), data hosted by the Genomic Data Commons (GDC; https://gdc.cancer.gov/). These two popular data sharing resources both house genomic datasets related to cancer, but despite containing similar data, these repositories have no direct connection to allow researchers to link them together. In the case of these two repositories the only way to find projects with analogous metadata is to manually search each repository. The key contribution of this work is a tool that acts as an interface between the GDC and dbGaP, which allows researchers to discover dbGaP datasets with similar metadata to a TCGA dataset of interest.

Methods

1. Resources

GDC. The GDC (https://gdc.cancer.gov/) is a highly curated resource for datasets from cancer related genomic studies from the National Cancer Institute (NCI). Its primary function is to provide a centralized repository for accessibility to data from large-scale NCI programs, such as ***TCGA*** and its pediatric equivalent, Therapeutically Applicable Research to Generate Effective Treatments. As of September 2016, GDC held over 260K sequence files with different genomic data-types (whole genome, RNA, etc.) of over 14K patients.

dbGaP. The National Center for Biotechnology Information (NCBI) dbGaP (https://www.ncbi.nlm.nih.gov/gap) is the largest collection of genomic data. It is not limited to cancer data or human data. While the metadata fields are fixed, unlike the GDC, the entries in these fields are not curated. This is a challenge for harmonizing the metadata across the two datasets. The NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) is a collection of sequence data associated with the studies in dbGaP.

2. Development

As the tool was developed as part of a hackathon, we used a development methodology similar to the Rapid Application Development model suitable for prototype development (Kerr & Hunter, 1994). This subsection is organized as steps within this methodology.

Defining the scope. We first identified the end users of our tool to be molecular and computational biologists and bioinformaticians with limited programming experience. Thus, the tools had to be easy to setup and execute. Next, we identified the use-cases as follows:

- The tool should take TCGA study identifiers or study-level metadata values from the GDC and identify dbGaP studies with analogous data.
- The tool should subsequently provide the capability of fetching the sequence level genomic data directly for these studies from the NCBI SRA data repository.

This gave us the necessary modules that needed to be developed.

Mapping the metadata. We first extracted the required metadata by parsing the raw XML data and also scrapping the website data from both TCGA (GDC) and dbGaP. This metadata is stored as mapping tables in CSV format. Based on the extracted metadata, we developed two mapping dictionaries to translate between 1) disease terms and 2) genomic data-types, as defined separately within dbGaP and the GDC.

Accomplishing this mapping was challenging, as the allowable values for these fields is strictly controlled in the GDC, but completely user-defined in dbGaP. We designed a rule-based mapper to generate an initial map between search values from each repository, then manually curated these mappings to refine and rank mapped terms. These mappings are stored and used during the execution of our tool.

Developing the required modules. Both the TCGA data (through GDC; https://gdc.cancer.gov/developers/gdc-application-programming-interface-api) and dbGaP (through NCBI Eutils; https://eutils.ncbi.nlm.nih.gov/entrez/eutils/) provide APIs to access their respective data that allow metadata transfer in the XML or JSON formats. An API or Application Programming Interface provide an interface to data and services that other programs can directly use.

The SRA toolkit is a software tool that allows researchers to obtain the sequence data (with appropriate access rights) from the SRA database. The search can be narrowed by various parameters, including the genomic region and type of sequence (e.g. mRNA and whole genome shotgun).

We used Python (version 2.7; https://www.python.org/) for the development of our tool. We wanted to keep the tool as platform agnostic as possible. As the SRA toolkit is Unix-based, only the final part of the implementation pipeline, as discussed subsequently, is a shell script (not directly compatible in Windows environment).

Results

We developed an easy-to-use tool that can be used to find additional data from dbGaP (and SRA) by expanding TCGA queries automatically. The first part of the pipeline allows researchers to query either repository by TCGA Project ID, File ID, Case ID, disease type, or experimental strategy via a metadata mapping dictionary. It returns not only a list of TCGA IDs, but also a list of related dbGaP study IDs. For dbGaP studies with NCBI SRA data, the second part of the pipeline will return the .sam files that contains reads aligned to a genomic region of interest to be used with the SRA Toolkit. Our tool is divided into three modules as illustrated in Figure 1. Below, each module is discussed in detail.

1. Fetching dbGaP studies using TCGA data

This component of the pipeline queries the GDC in multiple ways, including a direct ID search for projects, cases, samples, or files, or a custom search by the cancer type or experimental methods. Currently, the scope of custom search is limited to the available terms in the GDC data portal (Table 1). The module fetches the metadata using the GDC API and extracts the metadata terms related to the specified ID (i.e. the cancer type and experiment method). It then translates these terms to corresponding dbGaP search terms and returns the relevant dbGaP study IDs using the NCBI Eutils API. While executing the pipeline, the XML/JSON outputs of the APIs are processed in-memory behind the scenes. Thus, the end-users are not exposed to the API directly.

For custom searches, this module returns results from both the GDC and dbGaP simultaneously. Thus, this module also provides consolidated search capability over the TCGA and dbGaP data. The output from this module includes two files:

- a list of the TCGA cases for the given project or search criteria, and
- a list of dbGaP studies (with links) that are analogous to the input query.

2. Fetch SRRs for given dbGaP studies

The second component of the pipeline takes the list of dbGaP studies IDs and returns the list of sequence read run (SRR) files from the NCBI SRA from the dbGaP studies, when available. The users can specify the genomic region of interest as an additional parameter.

3. Fetch the sequence files from SRRs

The final part of the pipeline takes a list of SRRs and uses the SRA-toolkit to return sequencing level genomic data for a genomic region of interest directly from NCBI SRA data repository. This

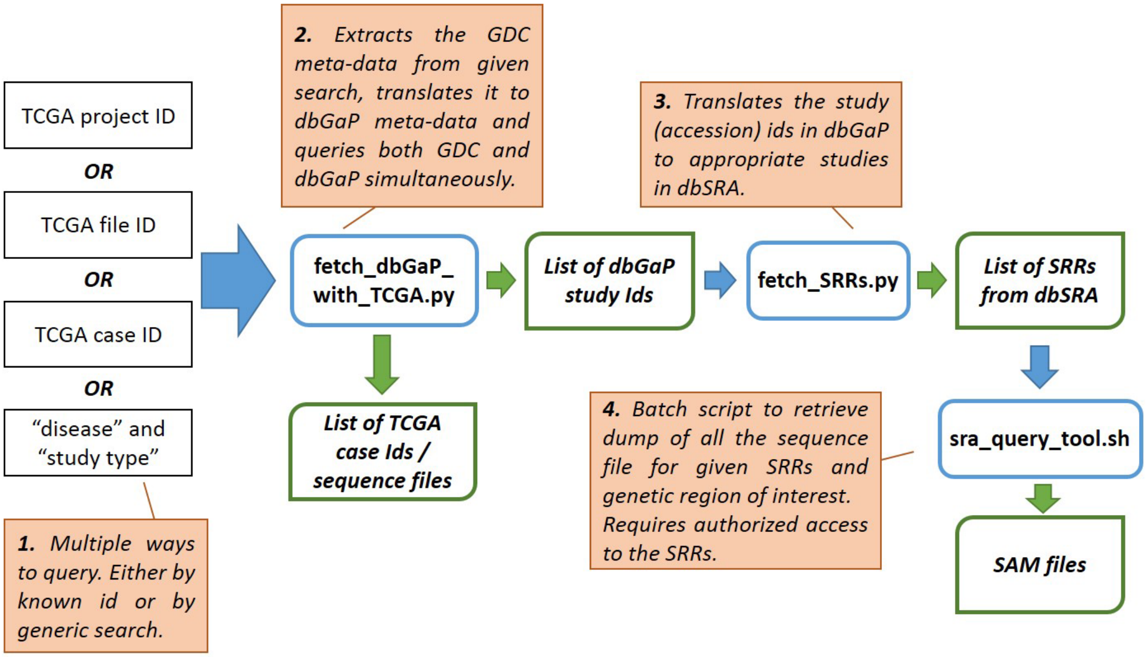
TCGA project ID
OR
TCGA file ID
OR
TCGA case ID
OR
"disease" and "study type"
1. Multiple ways to query. Either by known id or by generic search.
2. Extracts the GDC meta-data from given search, translates it to dbGaP meta-data and queries both GDC and dbGaP simultaneously.
fetch_dbGaP_ with_TCGA.py
List of TCGA case Ids / sequence files
List of dbGaP study Ids
3. Translates the study (accession) ids in dbGaP to appropriate studies in dbSRA.
fetch_SRRs.py
List of SRRs from dbSRA
sra_query_tool.sh
4. Batch script to retrieve dump of all the sequence file for given SRRs and genetic region of interest. Requires authorized access to the SRRs.
SAM files

Figure 1. **Module organization and a typical end-to-end workflow.**

Table 1. List of allowable values for Study Type, Primary Site and Disease in the Genomic Data Commons (The Cancer Genome Atlas) data. The mapping between the Disease and Primary Site can be found in our GitHub repository.

Study Type	Primary Site	Disease
Genotyping Array	Adrenal Gland	Pheochromocytoma and Paraganglioma
miRNA-Seq	Bile Duct	Adrenocortical Carcinoma
RNA-Seq	Bladder	Cholangiocarcinoma
WXS (Whole Exome Sequencing)	Blood	Bladder Urothelial Carcinoma
	Bone	Acute Myeloid Leukemia
	Brain	Osteosarcoma
	Breast	Glioblastoma Multiforme
	Cervix	Brain Lower Grade Glioma
	Colorectal	Breast Invasive Carcinoma
	Esophagus	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma
	Eye	Colon Adenocarcinoma
	Head and Neck	Rectum Adenocarcinoma
	Kidney	Esophageal Carcinoma
	Liver	Uveal Melanoma
	Lung	Head and Neck Squamous Cell Carcinoma
	Lymph Nodes	High-Risk Wilms Tumor
	Nervous System	Kidney Renal Clear Cell Carcinoma
	Ovary	Kidney Renal Papillary Cell Carcinoma
	Pancreas	Kidney Chromophobe
	Pleura	Rhabdoid Tumor
	Prostate	Clear Cell Sarcoma of the Kidney
	Skin	Liver Hepatocellular Carcinoma
	Soft Tissue	Lung Adenocarcinoma
	Stomach	Lung Squamous Cell Carcinoma
	Testis	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
	Thymus	Neuroblastoma
	Thyroid	Ovarian Serous Cystadenocarcinoma
	Uterus	Pancreatic Adenocarcinoma
		Mesothelioma
		Prostate Adenocarcinoma
		Skin Cutaneous Melanoma
		Sarcoma
		Stomach Adenocarcinoma
		Testicular Germ Cell Tumors
		Thymoma
		Thyroid Carcinoma
		Uterine Corpus Endometrial Carcinoma
		Uterine Carcinosarcoma

module assumes the required authorization has been granted prior to accessing the sequencing data.

Conclusion

To our knowledge, this is the first easy-to-use tool for harmonizing TCGA and dbGaP study metadata for the purpose of data discovery and consolidated querying. We would like to continue to work with the cancer biology community to develop this interface tool. Future improvements include extending our search capabilities to include other metadata, the option to query multiple genomic regions simultaneously, and a user-friendly GUI. Feature requests or contributions of code can be made on our GitHub site, which will be monitored for such activity.

Software availability

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

Archive source code as at the time of publication: doi, 10.5281/zenodo.160551 (Kurata, 2016) (https://zenodo.org/record/160551#.WE7Lz9WLTcs)

License: CC0 1.0 Universal

Author contributions

BB, LA and SR conceived the idea. All authors participated in the background research, design and implementation of the software tool. JK, EW and SR developed the variable mappings across dbGaP and TCGA. JK, AB, YL, EW and SR were primarily involved in implementation of the software components. EW and LA prepared the first draft of the manuscript. SR contributed to subsequent drafts.

Competing interests

No competing interests were disclosed.

Grant information

Funding was provided by the NIH intramural research program at the NLM. JK was supported in part by the Frances Berger Foundation Fellowship and her lab was supported in part by Ramesh Kesanupalli family and Beckman Research Institute Funds.

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

Acknowledgements

We would like to acknowledge Lisa Federer, NIH Library Writing Center, for manuscript editing assistance. The work described here was done as part of the NCBI Hackathon (August 2016). This research was supported in part by an appointment to the National Library of Medicine (NLM) Research Participation Program administered by ORISE through an interagency agreement between the U.S Dept. of Energy and the NLM and supported in part by the Intramural Research Program of the U.S. National Institutes of Health, NLM.

References

Dudley J, Butte AJ: **Enabling integrative genomic analysis of high-impact human diseases through text mining.** *Pac Symp Biocomput.* 2008; 580–591.
PubMed Abstract | Free Full Text

Kerr J, Hunter R: **Inside RAD: How to Build Fully Functional Computer Systems in 90 Days or Less**. McGraw-Hill Inc, New York, NY USA, 1994.

Kurata J, Badve A, Raje S, *et al.*: **NCBI-Hackathons/TCGA_dbGaP: TCGA_dbGaP_v1.0 2016 [Data set].** *Zenodo.* 2016.
Data Source

Ruau D, Mbagwu M, Dudley JT, *et al.*: **Comparison of automated and human assignment of MeSH terms on publicly-available molecular datasets.** *J Biomed Inform.* 2011; **44**(Suppl 1): S39–43.
PubMed Abstract | Publisher Full Text | Free Full Text

Tryka KA, Hao L, Sturcke A, *et al.*: **NCBI's Database of Genotypes and Phenotypes: dbGaP.** *Nucleic Acids Res.* 2014; **42**(Database issue): D975–9.
PubMed Abstract | Publisher Full Text | Free Full Text

Van Schaik TA, Kovalevskaya NV, Protopapas E, *et al.*: **The need to redefine genomic data sharing: A focus on data accessibility.** *Appl Transl Genomics.* 2014; **3**(4): 100–104.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ✓ ✓

Version 1

Referee Report 28 July 2017

doi:10.5256/f1000research.10605.r23937

✓ **Tsung-Jung Wu**
Baylor College of Medicine, Houston, TX, USA

The software can fulfill the requirement of authors designated task. Since the original design of this software is not for general usage, it might be difficult for a general user to access. However, with this tools' help, a cancer genomics researcher can download database from both GDC and dbGAP sample info easily. The most important part of this software is manually curated terms. This approach can assure more accurate search outcome and save user's valuable time. By introducing Disease Ontology to the mapping between the Disease and Primary Site step, it might be able to help more accurate curation and cancer type determination. Articles provide here are about Disease Ontology and Disease Ontology Cancer Slim. The tool is designed for cancer genomics community and with these terminologies and ID. This tool will be able to further expand its usage and application.

Some more documentations of this software and outcome interpretations will be helpful.

References

1. Kibbe WA, Arze C, Felix V, Mitraka E, Bolton E, Fu G, Mungall CJ, Binder JX, Malone J, Vasant D, Parkinson H, Schriml LM: Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data.*Nucleic Acids Res*. 2015; **43** (Database issue): D1071-8 PubMed Abstract | Publisher Full Text
2. Wu TJ, Schriml LM, Chen QR, Colbert M, Crichton DJ, Finney R, Hu Y, Kibbe WA, Kincaid H, Meerzaman D, Mitraka E, Pan Y, Smith KM, Srivastava S, Ward S, Yan C, Mazumder R: Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis.*Database (Oxford)*. 2015; **2015**: bav032 PubMed Abstract | Publisher Full Text

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?
Partly

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

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 10 July 2017

doi:10.5256/f1000research.10605.r23935

✓ **Konstantinos Krampis**
Department of Biological Sciences, Hunter College of The City University of New York, New York, NY, USA

The software is adequately explained and is a useful tool for a specialized task, fitting the format and section of the F1000. It is great work given that this was all completed during the hackathon. However I would suggest some polishing of the readme on the Github. This will still not make it any easier to use for not expert users, but for this purpose it would be ideal if the authors can add a short software readme as supplementary to the manuscript. The citation I have provided with this review points to a paper for the BioDocklets software that includes a manual that the authors could use as example. Other than that this is a great article that provides a very useful tool integrating key aspects of two important databases for the community.

References

1. Kim B, Ali T, Lijeron C, Afgan E, Krampis K: Bio-Docklets: virtualization containers for single-step execution of NGS pipelines. *GigaScience*. 2017. Publisher Full Text

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 28 April 2017

doi:10.5256/f1000research.10605.r21260

? **Yussanne Ma** , **Victoria Trinh**
Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada

The authors present a clear description of a tool that is simple in concept but will be of use to the cancer genomics community. They identified a clear need for researchers to be able to easily identify and download datasets from GDC and dbGAP sample attributes and have designed a solution that is conceptually sound. Most helpfully, they have taken care of not only the manual search but the mapping of the non-homogeneous metadata fields to save users time.

My detailed comments are listed below:

1. (Major) After installing the code and running the first command, we received this error:

```
$ bin/python bin/fetch_dbGaP_with_TCGA.py id -i TCGA-BRCA -s project -l low Traceback (most recent
call last):
  File "bin/fetch_dbGaP_with_TCGA.py", line 332, in
    sys.exit(main())
  File "bin/fetch_dbGaP_with_TCGA.py", line 322, in main

+outDict[outStringKeys[returnType][2]]+","+"\\".join(outDict[outStringKe
+ys[
returnType][3]])+","+outDict[outStringKeys[returnType][4]]+"\n"
TypeError: coercing to Unicode: need string or buffer, list found
```

As the authors are targeting users with minimal coding experience, it is important that error messages are far more informative. From the above it's completely unclear without looking into the code itself whether this is due to the input being in an incorrect format or if there is an actual problem with the code itself. Much more user testing and error handling is needed.

2. (Major) In general, more user documentation is needed. The instructions are brief and again, do not account for the case of everything not working on the first try. Input examples and example commands should also be provided for fetch_dbGaP_with_TCGA.py. Notes on how to interpret the output would also be helpful, perhaps by annotating the output file examples provided, which are a good inclusion.

3. (Minor) Currently the only way to see the re-mapping of metadata is to look in the code itself on github. Could a txt file of the field mappings be provided? This would be helpful for users to understand the assumptions being made with the metadata vocabulary.

4. (Minor) In the results section, could the authors summarize the results of the testing they did to ensure that the results being returned are correct? At minimum, it is important that users are assured of the completeness of the search. This could, for example, be demonstrated by searching on a TCGA disease type in GDC and comparing the number of results with the TCGA cohort size. Specificity and accuracy of the search results should also be demonstrated, perhaps by showing and summarizing the results of some example searches in both sites.

5. (Minor) GDC currently hosts not only TCGA, but also TARGET data (which is mentioned in the manuscript), and will soon be hosting other datasets. Is this tool limited to the TCGA datasets in GDC? As far as I know from our submissions to the GDC, the other datasets will have the same controlled metadata fields so the functionality should extend naturally to all of the data hosted at the GDC and it would greatly increase the utility if this were the case.

I think this tool will be beneficial to the cancer genomics community and will facilitate and encourage users to mine the rich NGS datasets that have been made available in the past decade. It would be good if the authors are able to provide a revision of the code that works and make some improvements to the user documentation, so that these benefits can be realized.

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?
No

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

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

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.