




SOFTWARE TOOL ARTICLE

Neopeptide Analyser: A software tool for neopeptide discovery in proteomics data [version 1; referees: 2 approved]

Mandy Peffers ¹, Andrew R. Jones², Antony McCabe², James Anderson¹

¹Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, L7 9TX, UK

²Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK

v1 First published: 07 Apr 2017, 2:24 (doi: [10.12688/wellcomeopenres.11275.1](https://doi.org/10.12688/wellcomeopenres.11275.1))




Latest published: 07 Apr 2017, 2:24 (doi: [10.12688/wellcomeopenres.11275.1](https://doi.org/10.12688/wellcomeopenres.11275.1))

Abstract

Experiments involving mass spectrometry (MS)-based proteomics are widely used for analyses of connective tissues. Common examples include the use of relative quantification to identify differentially expressed peptides and proteins in cartilage and tendon. We are working on characterising so-called 'neopeptides', i.e. peptides formed due to native cleavage of proteins, for example under pathological conditions. Unlike peptides typically quantified in MS workflows due to the *in vitro* use of an enzyme such as trypsin, a neopeptide has at least one terminus that was not due to the use of trypsin in the workflow. The identification of neopeptides within these datasets is important in understanding disease pathology, and the development of antibodies that could be utilised as diagnostic biomarkers for diseases, such as osteoarthritis, and targets for novel treatments. Our previously described neopeptide data analysis workflow was laborious and was not amenable to robust statistical analysis, which reduced confidence in the neopeptides identified. To overcome this, we developed 'Neopeptide Analyser', a user friendly neopeptide analysis tool used in conjunction with label-free MS quantification tool Progenesis QIP for proteomics. Neopeptide Analyser filters data sourced from Progenesis QIP output to identify neopeptide sequences, as well as give the residues that are adjacent to the peptide in its corresponding protein sequence. It also produces normalised values for the neopeptide quantification values and uses these to perform statistical tests, which are also included in the output. Neopeptide Analyser is available as a Java application for Mac, Windows and Linux. The analysis features and ease of use encourages data exploration, which could aid the discovery of novel pathways in extracellular matrix degradation, the identification of potential biomarkers and as a tool to investigate matrix turnover. Neopeptide Analyser is available from <https://github.com/PGB-LIV/neo-pep-tool/releases/>.

Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
version 1		
published 07 Apr 2017	report	report
<p>1 Timothy E. Hardingham, University of Manchester UK, Jamie Soul, University of Manchester UK</p> <p>2 Stephanie G. Dakin , University of Oxford UK</p>		

Discuss this article

[Comments](#) (0)

Corresponding author: Mandy Peffers (peffs@liv.ac.uk)

How to cite this article: Peffers M, Jones AR, McCabe A and Anderson J. **Neopeptide Analyser: A software tool for neopeptide discovery in proteomics data [version 1; referees: 2 approved]** Wellcome Open Research 2017, 2:24 (doi: [10.12688/wellcomeopenres.11275.1](https://doi.org/10.12688/wellcomeopenres.11275.1))

Copyright: © 2017 Peffers M *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: This work was supported by the Wellcome Trust [107471]; and a University of Liverpool Technical Directorate Voucher. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Competing interests: No competing interests were disclosed.

First published: 07 Apr 2017, 2:24 (doi: [10.12688/wellcomeopenres.11275.1](https://doi.org/10.12688/wellcomeopenres.11275.1))

Introduction

Mass spectrometry (MS)-based proteomics can generate large amounts of data for downstream analyses, such as protein discovery, relative quantification and novel peptide fragment (neopeptide) discovery. The identification of neopeptides provides a platform for the development of antibodies that could assist in the discovery of molecular markers for diseases, such as osteoarthritis¹, as well as the identification of basic processes underlying disease, such as matrix turnover². Generating neopeptide antibodies enables the detection and monitoring of cartilage degeneration and therapeutic responses to treatment, in addition to providing treatment targets.

We have undertaken a number of studies to identify neopeptides following MS of ageing or diseased cartilage³, tendon^{2,4}, and synovial fluid⁵, as well as following specific exogenous protease-driven digestion of cartilage extracts and in an *in vitro* model of early osteoarthritis¹. From these studies we have identified both novel and previously characterised neopeptides.

There are no available tools at present to interrogate the identified neopeptides. Therefore, in order to identify relevant neopeptides, we previously developed a novel LC-MS/MS data processing workflow. Under the previous workflow, we undertook “semi-trypsin” searches (i.e. only one terminus of the peptide was required to be the result of tryptic cleavage) with the relevant Uniprot databases using Mascot (Matrix Science, London, UK) or PEAKS Studio (Bioinformatics Solutions, Inc., Waterloo, Canada). The resulting identified peptides from individual samples were input into spreadsheets for further filtering. This data analysis was laborious and the inclusion of neopeptides to take forward user dependant. These factors inhibited users from generating neopeptide data with statistical confidence for further exploration.

To address this, we developed ‘Neopeptide Analyser’, a user-friendly interface for neopeptide discovery [in association with Progenesis QIP software for relative quantification (Waters, Manchester, UK)] that rapidly identifies neopeptides and provides a p-value to indicate differential expression. A key feature of Neopeptide Analyser is the ability to apply a statistical value to neopeptide discovery whilst also enabling the user to apply less stringent cut offs if required.

Methods

The tool parses data files that are exported from Progenesis QIP (Version 2) (<http://www.nonlinear.com/progenesis/qi-for-proteomics/>) in csv format. This is known as the peptide measurements csv file. The tool can also take a protein database (fasta) file as input, in order to search for the peptide locations. Two output files are produced by the tool. The first file is in the same format as the Progenesis QIP file, with the addition of three data columns (file input name suffixed ‘with_filter’ as default). For each peptide in the Progenesis output, these columns contain the residues preceding and following the peptide within its parent protein, and whether the peptide is fully tryptic (two termini resulting from trypsin digestion) or semi-tryptic (one terminus resulting from trypsin digestion). For some peptides, the sequence can be found

in multiple proteins. For these cases, if the peptide could either be fully tryptic, or semi tryptic, it is assumed that the fully tryptic peptide is the most likely source.

The tool creates a second output file, also in csv format, which describes just the neopeptides that were found in the input file, and normalises the quantification values for each peptide (suffixed with ‘processed’). The normalisation method aims to remove the effect of changes in the overall parent protein abundance from the quantification value for the neopeptide, such that changes in abundance for the normalised neopeptide can be assumed to be the result of different extents of *in vivo* cleavage of the parent protein. Normalisation is thus achieved by dividing the candidate neopeptide (semi-tryptic) abundance by the sum of the abundance of all the tryptic peptides for the parent protein.

Where an experiment is setup with two conditions, these are read from the Progenesis QIP input file and a Student’s t-test is performed. This uses a normalised quantification value for each neopeptide, across the two conditions, and a p-value is produced so that the user can determine if the change in normalised abundance across the conditions may be significant. For discovery proteomics, this t-test may not be particularly meaningful by itself, as some peptides amongst many are likely to score a low value purely by random chance, so the output file also gives the Bonferroni (BF) corrected result, based on a user-supplied false discovery rate (FDR), as well as Benjamini-Hochberg (BH) corrected p-values. BH is generally the preferred method of global correction in quantitative proteomics, as BF is often too conservative to gain significance amongst large numbers of peptides/proteins.

Implementation

The tool was developed as a Java application, with a ‘Swing’ graphical user interface (for compatibility with almost all desktop computers, requiring only a Java SE Version 7 Runtime environment installed, which has been available since 2011). An executable Java archive file can be downloaded, and opening this file will show the user interface.

The user interface allows the user to select a Progenesis QIP export file, as well as a fasta file to use as a protein database. These choices, and all other settings, are saved and restored each time the tool is started.

There are settings to allow the user to specify the format of the input file. The default settings are correct for files that are currently produced by Progenesis QIP, but these may change in future, or the data may have been manipulated in some other program (such as Microsoft Excel), before being used by the tool. The auto-detect features will usually be able to correctly identify the format of the file, by searching for columns that contain numerical data or only strings of amino-acids. The two output data files that are produced are given default names automatically based upon the input file name, but these can be changed via the user interface. The user can then click to process the input file, which will perform the computations needed to produce the two output files.

The default method for the tool is to search for the peptide in any matching proteins within the fasta file; it can then look at the previous and next residues in a matching protein, bearing in mind that the peptide could align with the C-terminus of the protein (and hence be fully tryptic even if it does not end with Arginine or Lysine). Similarly, if the peptide is at the N-terminus (peptide start position within the protein =1), the previous residue does not need to be tested, and where the previous residue is a Methionine (peptide start position =2), this is also not evidence of a non-tryptic cleavage, but N-terminal methionine cleavage, which is very common *in vivo*.

For large data files, a faster lookup method may also be employed, and can be selected from the options section. The database fasta file is used to build an internal dictionary of all possible tryptic peptides, including the required number of missed cleavages, which is determined by examining the input data file. The tool can then quickly see if any of the input peptides are in this fully-tryptic dictionary, or not, and even large files with many tens of thousands of peptides can be processed in less than a minute on a standard desktop computer. However, this method does not allow finding the preceding and following residues, as it is not feasible to create a dictionary of every possible semi-tryptic peptide.

In order to produce more meaningful statistical results, normalised neopeptide abundance values are created, and are included in the output file. The tool can detect two conditions that are present in the Progenesis QIP export file, and groups the sample data according to the relevant condition. It uses the calculated

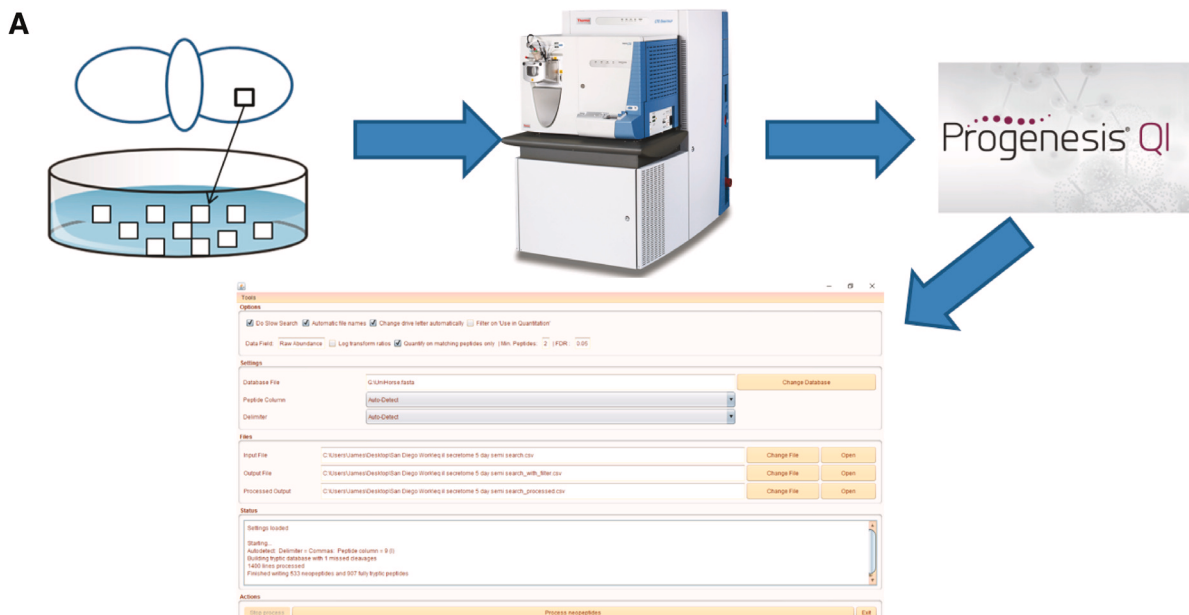
normalised quantification value for each neopeptide, in each sample, to produce a p-value, indicating the statistical significance of the variation across the two conditions (using Student's t-test). The tool uses the FDR supplied by the user to output the result of BF correction, and follows the standard BH procedure for multiple testing to also give a corrected p-value in the output.

Operation

The Neopeptide Analyser is available both as a pre-compiled Java executable file (NeopeptideTool.jar) and as java source code. No external libraries are used and the tool can be compiled with any compiler supporting Java SE version 7 or above. The pre-compiled Java executable file is compatible with any computer that has a Java runtime environment installed of version 7 or above. If it is not already installed, the runtime environment needed can be freely downloaded from <https://java.com/en/download/>.

Use case

Figure 1 illustrates a typical use for Neopeptide Analyser. Data used for input were label-free quantification results following analysis of the secretome of equine metacarpophalangeal cartilage explants treated with interleukin 1 β for 5 days¹. Progenesis QIP was used to undertake label-free quantification of the proteins within the secretome following liquid chromatography tandem MS. Following feature picking, we exported the top three spectra for each feature. These were exported from Progenesis QIP and utilized for peptide identification with a locally implemented Mascot server in Unihorse database (<http://www.uniprot.org/uniprot/?query=equus%20caballus>). Search parameters used were:



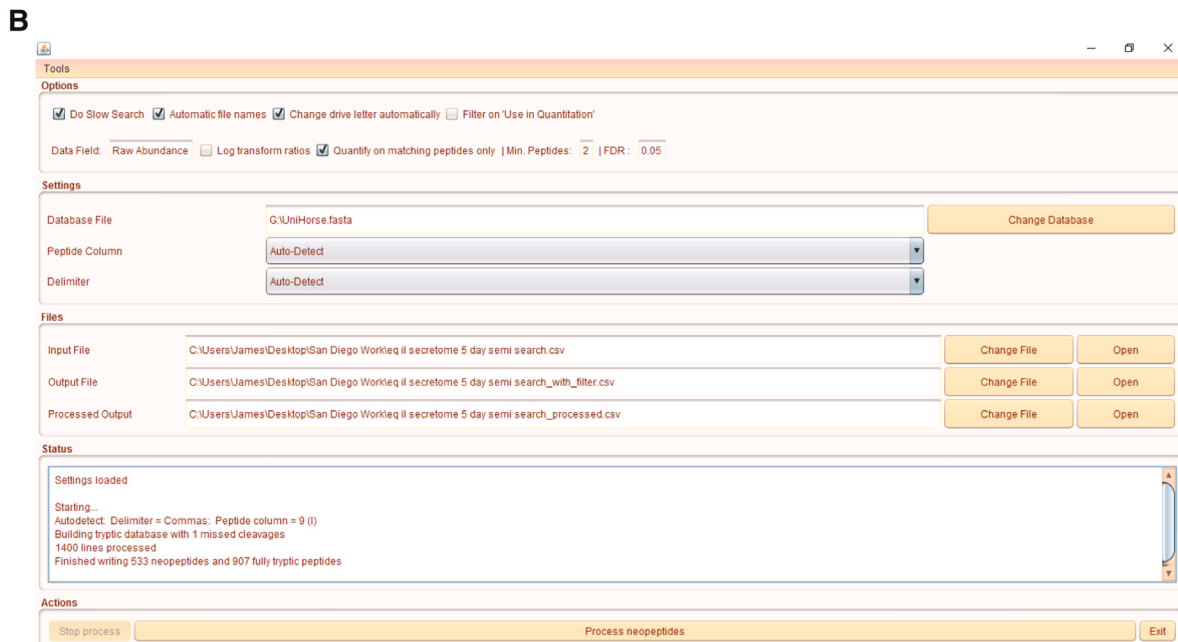


Figure 1. An example of Neopeptide Analyser workflow to analyse an equine cartilage secretome following treatment with IL-1 β . (A) Diagram of workflow incorporating liquid chromatography tandem mass spectrometry analysis with label-free quantification using Progenesis QIP and Neopeptide Analyser. In this example, cartilage explants are used from the metacarpophalangeal joint of the horse and grown *in vitro* with and without IL-1 β . (B) Neopeptide Analyser interface showing file input and outputs. The data file used to generate this figure and Neopeptide Analyser output data files are available in Supplementary File 1 (<https://doi.org/10.6084/m9.figshare.4769746.v1>) and Supplementary File 2 (<https://doi.org/10.6084/m9.figshare.4772131.v1>). The 'Options' tab includes selection for slow search, 'automatic file names' (taken from input file name), drive letter selection, 'use in quantitation' filter, data field, 'log transform ratios', 'quantify on matching peptides only'. Additionally, the minimum number of peptides and false discovery rate can be set manually. The 'Settings' tab enables the database file to be selected as well as peptide column and delimiter (default as auto-detect). The 'Files' tab contains options for the 'Input File', 'Output File' and 'Processed Output File'. The 'Status' tab updates the user on the stage of analysis. Finally in the 'Actions' tab, the 'Process neopeptides' button is selected to start the analysis.

10 ppm peptide mass tolerance and 0.6 Da fragment mass tolerance; one missed cleavage allowed; fixed modification; carbamidomethylation; variable modifications; methionine oxidation and enzyme semitrypsin.

In [Figure 1A](#) the workflow of a typical experiment is demonstrated. [Figure 1B](#) details the input and output options on Neopeptide Analyser. The peptide measurement csv generated from Progenesis QIP was used as input (Supplementary File 1; <https://doi.org/10.6084/m9.figshare.4769746.v1>). The Unihorse fasta file is applied to search for matching proteins. The two output files are evident as 'Output file' and 'Processed output file' (Supplementary File 2; <https://doi.org/10.6084/m9.figshare.4772131.v1>). The processed output file details the protein, neopeptide sequence, preceding and following residues, p-value and FDR-adjusted p-value.

Conclusions

Neopeptide Analyser enables rapid neopeptide detection from many thousands of peptides to be analysed within a minute using

a standard computer. This will facilitate wider exploration of high-throughput proteomics data, leading to the identification of known neopeptides and the discovery of novel neopeptides. These may be used as indicators of matrix turnover, and as diagnostic or prognostic biomarkers. Whilst the tool enables the statistical significance of the variation across the conditions to be applied, the output enables the data to be interrogated with less stringent cut-offs that may be more applicable in some experiments.

Data and software availability

The data input file for the Neopeptide Analyser is available in Supplementary File 1 in Figshare (<https://doi.org/10.6084/m9.figshare.4769746.v1>). The output and processed output files are available in Supplementary File 2 in Figshare (<https://doi.org/10.6084/m9.figshare.4772131.v1>).

Version 1.0 of Neopeptide Analyser is available to download from <https://github.com/PGB-LIV/neo-pep-tool/releases/> both as a

pre-compiled Java executable file (NeopeptideTool.jar) and as java source code.

Archived source code as at time of publication: <https://doi.org/10.5281/zenodo.438664>⁸

License: MIT

Author contributions

MJP and JA conceived the software. MJP designed the studies. AJ and AM designed and wrote the software. All authors wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust [107471]; and a University of Liverpool Technical Directorate Voucher.

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

Acknowledgements

We thank the Centre for Proteomic Research, University of Liverpool for the mass spectrometry facilities. We are grateful to members of the Clegg lab, University of Liverpool for testing the tool.

References

1. Peffers MJ, Thornton DJ, Clegg PD: **Characterization of neopeptides in equine articular cartilage degradation.** *J Orthop Res.* 2016; **34**(1): 106–20. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Thorpe CT, Peffers MJ, Simpson D, *et al.*: **Anatomical heterogeneity of tendon: Fascicular and interfascicular tendon compartments have distinct proteomic composition.** *Sci Rep.* 2016; **6**: 20455. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Peffers MJ, Cillero-Pastor B, Eijkel GB, *et al.*: **Matrix assisted laser desorption ionization mass spectrometry imaging identifies markers of ageing and osteoarthritic cartilage.** *Arthritis Res Ther.* 2014; **16**(3): R110. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Peffers MJ, Thorpe CT, Collins JA, *et al.*: **Proteomic analysis reveals age-related changes in tendon matrix composition, with age- and injury-specific matrix fragmentation.** *J Biol Chem.* 2014; **289**(37): 25867–78. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Peffers MJ, McDermott B, Clegg PD, *et al.*: **Comprehensive protein profiling of synovial fluid in osteoarthritis following protein equalization.** *Osteoarthritis Cartilage.* 2015; **23**(7): 1204–13. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Peffers M, Jones A, Anderson J: **Equine cartilage secretome 5 days IL-1 treatment *in vitro*; Progenesis semi tryptic peptides measurements file for input into the Neopeptide Analyser.** *figshare.* 2017. [Data Source](#)
7. Peffers M, Jones A, Anderson J: **Equine cartilage secretome 5 days IL-1 treatment *in vitro*; 'Neopeptide Analyser' output files.** *figshare.* 2017. [Data Source](#)
8. tonyattiv: **PGB-LIV/neo-pep-tool: First release, with MIT licence [Data set].** *Zenodo.* 2017. [Data Source](#)

Open Peer Review

Current Referee Status:  

Version 1

Referee Report 10 May 2017

doi:[10.21956/wellcomeopenres.12164.r22115](https://doi.org/10.21956/wellcomeopenres.12164.r22115)



Stephanie G. Dakin 

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS), Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK

In this manuscript, the authors develop a neopeptide analysis tool used that may be used in conjunction with label-free MS quantification tool Progenesis QIP for proteomics. Given the extensive use of MS based proteomics to analyse connective tissues in health and disease, and the difficulties associated with analyses of generated neopeptide data, the rationale for this study is clearly stated and the development of such a tool is timely. The findings from this study are likely to translate and be broadly applicable to MS analyses of other body tissues from human and animal species. The authors may wish to mention the potential broader utility of this tool in the manuscript.

In addition to the description of the software tool and details of the code, methods and analyses outlined in the manuscript, the authors may wish to provide material to support users that may encounter any problems during use of the software, for example a series of frequently asked questions or trouble shooting guide in the form of supplementary material.

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 02 May 2017

doi:[10.21956/wellcomeopenres.12164.r21632](https://doi.org/10.21956/wellcomeopenres.12164.r21632)



Timothy E. Hardingham, Jamie Soul

Wellcome Trust Centre for Cell-Matrix Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

This paper presents the tool Neopeptide Analyser to detect neopeptides from proteomics data analysed with Progenesis QIP. The paper is well written and gives a clear explanation of the tools structure and application. There is no stand-alone documentation provided, but the tool was tested and found simple to use and ran quickly using the provided test data.

Although dependant on the use of the Progenesis QIP software for the initial proteomic analysis, the purpose of the tool appeared well thought out to fulfil a specific need to detect unique peptides generated by different proteinases. It therefore appears to have strong potential for wide use to detect novel proteinase generated fragments and to perform differential quantification analysis of these neopeptides. Recent evidence suggests that different technology/software combinations generate some differences in results (e.g. Al Shweiki et al 2017), but results using this package are likely to be consistent and reproducible when using the same analysis format.

The methods state that the output file also gives the Bonferroni (BF) corrected result, based on a user-supplied false discovery rate (FDR), as well as Benjamini-Hochberg (BH) corrected p-values, but only one adjusted p-value column is given in the output file?

A sentence or two giving an example of the utility of the results would be useful and an example might be the differential analysis of fibronectin fragments.

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.

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.
