



SOFTWARE TOOL ARTICLE

REVISED Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps [version 2; peer review: 2 approved]

Michael C. Ryan ¹, Mark Stucky¹, Chris Wakefield², James M. Melott², Rehan Akbani², John N. Weinstein^{2,3*}, Bradley M. Broom ^{2*}

¹In Silico Solutions, Fairfax, VA, 22031, USA

²Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

³Department of Systems Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

* Equal contributors

v2 First published: 14 Oct 2019, 8(ISCB Comm J):1750 (<https://doi.org/10.12688/f1000research.20590.1>)

Latest published: 19 Mar 2020, 8(ISCB Comm J):1750 (<https://doi.org/10.12688/f1000research.20590.2>)

Abstract

Clustered heat maps are the most frequently used graphics for visualization and interpretation of genome-scale molecular profiling data in biology. Construction of a heat map generally requires the assistance of a biostatistician or bioinformatics analyst capable of working in R or a similar programming language to transform the study data, perform hierarchical clustering, and generate the heat map. Our web-based Interactive Heat Map Builder can be used by investigators with no bioinformatics experience to generate high-caliber, publication quality maps. Preparation of the data and construction of a heat map is rarely a simple linear process. Our tool allows a user to move back and forth iteratively through the various stages of map generation to try different options and approaches. Finally, the heat map the builder creates is available in several forms, including an interactive Next-Generation Clustered Heat Map that can be explored dynamically to investigate the results more fully.

Keywords

Bioinformatics, Genomics, Heat Map, Web Tool, Website, Hierarchical Clustering



This article is included in the International Society for Computational Biology Community Journal gateway.

Open Peer Review

Reviewer Status

	Invited Reviewers	
	1	2
version 2 (revision) 19 Mar 2020		 report
version 1 14 Oct 2019	 report	 report

1 **Natasha Caplen**, National Cancer Institute, Bethesda, USA

Soumya Sundara Rajan , National Cancer Institute, Bethesda, USA

2 **Melissa S. Cline** , University of California, Santa Cruz, Santa Cruz, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Bradley M. Broom (bmbroom@mdanderson.org)

Author roles: **Ryan MC:** Conceptualization, Investigation, Software, Visualization, Writing – Original Draft Preparation; **Stucky M:** Conceptualization, Software; **Wakefield C:** Investigation, Software, Writing – Review & Editing; **Melott JM:** Software, Validation; **Akbani R:** Investigation; **Weinstein JN:** Funding Acquisition, Investigation, Writing – Review & Editing; **Broom BM:** Funding Acquisition, Investigation, Software, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported in part by Grant Numbers U24CA143883, U24CA199461, U24CA210949 and U24CA210950 from the National Cancer Institute, as well as generous gifts from the Mary K. Chapman Foundation and the Michael & Susan Dell Foundation. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2020 Ryan MC *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ryan MC, Stucky M, Wakefield C *et al.* **Interactive Clustered Heat Map Builder: An easy web-based tool for creating sophisticated clustered heat maps [version 2; peer review: 2 approved]** F1000Research 2020, 8(ISCB Comm J):1750 (<https://doi.org/10.12688/f1000research.20590.2>)

First published: 14 Oct 2019, 8(ISCB Comm J):1750 (<https://doi.org/10.12688/f1000research.20590.1>)

REVISED Amendments from Version 1

This version of the article has been revised to address the comments and questions of reviewers and to update the manuscript to reflect a few updates in the latest release of the Interactive CHM Builder.

Any further responses from the reviewers can be found at the end of the article

Introduction

Many thousands of publications on genomics studies include clustered heat maps (CHMs) because the hierarchical clustering and intuitive visualization provide insight into the relationships among sample sub-groups and key biological processes¹⁻⁸. Construction of a CHM requires data transformation, application of clustering methods, association of covariate (classification) data, and production of the heat map visualization. Generally, those tasks require the assistance of an analyst with biostatistics or bioinformatics skills who can work in R or a similar language to manipulate the study data and generate the map. This is usually not a simple linear process because data transformation and clustering methods are often revisited to find the ideal match for the study, and modifications are often made to heat map visualizations to select the best colors, adjust covariates, insert gaps, etc. Our Interactive CHM Builder is a web-based tool for data transformation, clustering, and generation of high-quality heat maps. It can be used by investigators with no bioinformatics experience and only modest exposure to biostatistical methods. The tool guides users through the steps of creating a heat map and supports iterative refinement of the map by working backward and forward through the steps to refine data transformation, annotation, clustering, and formatting options. (Caveat: Iterative exploration of different options may introduce a multiple-comparisons issue that would have to be taken into account if the map were used for formal statistical inference, rather than discovery.)

One obvious limitation of traditional heat maps is that they contain a huge amount of information but are static in nature and do not readily support a deeper exploration of the biology behind the image. The Interactive CHM Builder produces traditional heat map images as PDF files but can also produce interactive next-generation CHMs (NG-CHMs). NG-CHMs support interactive exploration of patterns in the data through zooming, panning, searching, and advanced link-outs to dozens of external resources. An NG-CHM file can be downloaded and viewed locally with the NG-CHM viewer and, importantly, can be embedded in a study results webpage or publication.

The Interactive CHM Builder⁹, available at <https://build.ngchm.net/NGCHM-web-builder/>, is easy to try out using sample data provided at the site. Other methods of producing NG-CHMs, including an R library and a set of tools for the Galaxy platform^{10,11}, are described at <https://www.ngchm.net/>.

Methods

Implementation

The Interactive Builder⁹ is web-based application that accepts an uploaded data matrix and then walks the user through

several steps to transform the data, perform hierarchical clustering, and format the resulting CHM. The application is implemented as HTML, CSS, and JavaScript on the browser-side and Java servlets on the web server. Data manipulation and heat map generation are implemented in Java classes used by the servlets. The clustering is performed by a servlet using the Renjin engine (<https://www.renjin.org>) to perform R clustering functions in Java. Browser sessions are tracked by the server to create a working area for each user and prevent users from seeing each other's data or maps. In addition to the working version of the data matrix on which transformations are performed, an original version of the matrix is preserved. Returning to a previous matrix state is accomplished by restoring the original version and then re-applying transformations until the requested state is restored. The site retains constructed heat maps and the related uploaded data only for the duration of the HTTP session.

A Java NG-CHM heat map generator .jar file is used to construct the heat map repeatedly as options are selected in each step of the builder. The heatmapProperties.json file, which contains all options selected by the user, conveys the selected options to the generator. The current NG-CHM file set is stored in a directory under the session ID. The NG-CHM file is a zipped version of the NG-CHM directory. The downloaded .ngchm file can be saved locally and viewed interactively using a local instance of the NG-CHM viewer that can also be downloaded from the builder site. An overview is given in [Figure 1](#).

The full source code for the Interactive Builder is available in [GitHub](#).

Operation

There is no need to install software to use the Interactive Builder⁹ it is available for public use on our server at <https://build.ngchm.net/NGCHM-web-builder/>. If, however, a local private installation of Interactive Builder is preferred, there are two simple installation methods.

Organizations familiar with Docker can run the Builder as a Docker container (<https://docs.docker.com/>). To do this, clone the git repository. The base folder of this repository has a docker build file. Run the docker build command in this directory with a `-t` option to name the resulting docker image. For example: `docker build . -t nghm_builder`. Then use the docker run command to start a container using the image. The heat maps created by the software are transient and last only for the duration of a user http session so there is no need to mount an external directory to the container for persistent storage. The port for connecting to the webserver in the container does need to be specified in the docker run command. Connect the desired external port to the tomcat instance in the container. For example, `docker run --name="ngchm_builder" -d -p 8888:80 nghm_builder`. Users should then be able to connect to Interactive Builder using their browser and the URL of the docker container. For example, `http://<docker machine IP or URL>/NGCHM-web-builder`.

The other option for deploying the software is to install it on an existing web server like tomcat (<https://tomcat.apache.org/tomcat-9.0-doc>). To do this, first clone the git repository

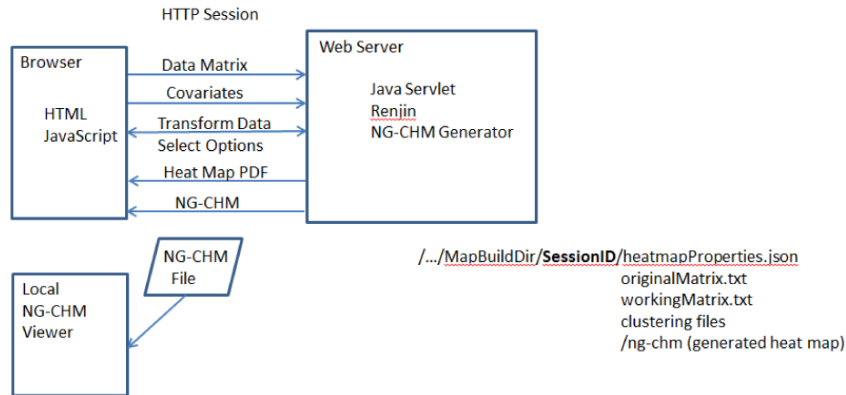


Figure 1. High-level overview of the interaction of heat map builder components. Heat maps are built on a webservice. A browser session ID is used to create a separate, temporary working area for each user. Heat map construction sessions are cleaned up when the session is ended, but PDF and NG-CHM heat map files can be downloaded.

and then use the ant script, ant_buildfile.xml in the NG-CHM_GUI_BUILDER folder to create a .war file. Then simply copy the .war file to the webapps directory of the web server. The application should then be available at `http://<server URL>/NGCHM-web-builder`.

Use case

The starting point for a CHM is a matrix of data. In this use-case example, we focus on gene expression data from The Cancer Genome Atlas (TCGA) bladder cancer project^{12,13}. The rows and columns of the matrix require identifiers, in this case sample ids and gene symbols, and the cells of the matrix must be numeric values. The builder will accept either a tab-delimited text file (*.txt), comma-separated text file (*.csv) or Excel spreadsheet (*.xlsx).

Select matrix

The Open Matrix File button on the first page of the builder (Figure 2) is used to upload the data matrix. A name and optional description to be associated with the heat map are entered. When the data have been loaded, the Select Matrix page will show the first few rows and columns of the matrix. It is important that the builder correctly identify the row labels, column labels, and matrix data; the backgrounds of labels and matrix data should be blue and green, respectively. If the input file has extra rows or columns, you may need to correct the identification of labels and matrix data by selecting the appropriate radio button and then clicking on the correct location in the matrix displayed.

Note that several screens in the builder include advanced features that are hidden by default to simplify the process for first-time users. The use-case example here does not require advanced features, but be aware that additional capabilities can be accessed using the Advanced Features checkbox.

Transform/filter the data

Creating a good heat map depends on proper data preparation. The second step in the build process is the Data Transform page (Figure 3), which provides three primary categories

of matrix transformations: functions that identify and replace missing/invalid values, filters to remove rows or columns, and transforms to perform mathematical operations on data values. There are additional choices in advanced mode for transposing the matrix and calculating correlations.

The right-hand panel of the Transform page provides summary statistics about the data matrix, including the number of rows and columns, a histogram of the data distribution, and an indication of the number of invalid cells in the matrix. The top of the page also provides suggestions about transformations that can be performed and flags any problems with the data. The use-case matrix is too large for the Interactive Builder to use in creating a heat map interactively; the clustering time, which increases approximately as the square of the larger matrix dimension for most clustering algorithms, is limiting. Currently, the website limits the heat map to no more than 5,000 total rows and columns (for example 1,000 samples and 4,000 genes) at the clustering stage. However, users can upload much larger matrices as long as filters on the transform page reduce the size to 5,000. For practical purposes, that often means extracting the most relevant data (e.g., with few enough missing values, sufficient signal, and sufficient standard deviation across samples) for clustering. We are also progressively increasing the size limit as compute power and clustering algorithms advance.

For this use case the transform tab is used to fix duplicate column headers; set a minimum threshold to reduce the influence of noise in the heat map; normalize the data with a log transform and mean center; and filter to remove rows with many missing values and to keep only rows with strong variation across samples. The transforms applied were:

- Action: Duplicates Duplicates process: Rename. Column. Suffix duplicates with underscore and instance number. Apply.
- Action: Transform Data Transform: Threshold. Set Values Below 0.00001 to NA. Apply.

THE UNIVERSITY OF TEXAS
MDAnderson
Cancer Center
 Bioinformatics and Computational Biology

NG-CHM BUILDER: Select Matrix

Open Matrix File
 Try Sample Matrix:
 Matrix: BLCA_Gene_Expression_TPM_10k.txt
 Show Advanced Features:

Heat Map Name:
 BLCA_Gene_Expression_TPM_10k

Heat Map Description:

The builder needs to know where the row labels, column labels, matrix data, and covariate data (if included) are located in the uploaded file. The labels should be blue and data should be green. If not select from the following controls and click on the grid to indicate the location of labels, covariate bars, and the location at which the matrix data begins in the imported file.

Labels Row (blue) Labels Column (blue) Data Start Location (green)

	TCGA-2F-A9KO	TCGA-2F-A9KP	TCGA-2F-A9KQ	TCGA-2F-A9KR	TCGA-2F-A9KT	TCGA-2F-A9KW
A1CF	0	0	0	0	0	0.171
AAA1	0.0669	0	0.0524	0.0584	0	0.0638
AAAS	46.2	38.8	40.4	52	33.4	56.4
AACS	18.8	14.3	23.9	15.1	17.5	30.9
AADACL2	0.038	0	0	0	0.082	0.12
AADACL4	0	0	0	0.0259	0	0
AAMP	104	114	83.1	187	71.4	80.8
AARS2	8.27	12.6	8.61	10.5	9.8	8.58
AARSD1	42.4	47.6	23.8	26.7	61.7	24.5
AASDHPPT	17.1	28.9	32.6	18.5	15.8	22.3
AASS	8.18	0.48	1.25	11	1.28	2.38
AATK	0.452	0.598	0.45	1.59	1.44	0.727
ABCA11P	2.28	5.98	2.24	9.95	2.59	5.69
ABCA12	0.0481	0.0943	0.205	10.6	4.19	8.63
ABCA13	1.37	0.13	0.0201	0.867	0.039	0.0174
ABCA1	5.13	1.56	6.4	4.86	9.13	19.1
ABCA2	2.9	7.52	5.73	5.13	27.3	2.67

Figure 2. Heat map creation starts with importing a text matrix file (e.g., *.txt, *.csv or Excel *.xlsx file) and identifying the row labels, column labels and numerical data values.

THE UNIVERSITY OF TEXAS
MDAnderson
Cancer Center
 Bioinformatics and Computational Biology

NG-CHM BUILDER: Transform Matrix
 Map Name: BLCA_Gene_Expression_TPM_10k

Show Advanced Features:

Actions:

Missing / Invalid Data:
 Select a correction:
 Replace Invalid Values With:
 N/A
 Zero
 Row Mean
 Column Mean

Matrix Change History:
 Original Version

Reset to a previous step by selecting it above and pressing

ERROR: Matrix has too many Rows (>3500) for this builder. Use Filters to remove some Rows.
 This page provides summary statistics of your matrix data including the distribution of values and row/column standard deviations. Filters and transforms can be used to manipulate the matrix to produce better heat maps. For example, a Z-norm transform could be used to normalize rows with values that differ in magnitude and a standard deviation filter could be used to remove rows with values that do not differ much across the columns.

Number of Rows: 10339
Number of Columns: 427
Missing Values: 0
Invalid Values: 0
Maximum Value: 137818.296
Minimum Value: 0.000
Minimum (Non-Zero): 7.42e-288

Distribution:

	TCGA-2F-A9KO	TCGA-2F-A9KP	TCGA-2F-A9KQ	TCGA-2F-A9KR	TCGA-2F-A9KT	TCGA-2F-A9KW
A1CF	0	0	0	0	0	0.171
AAA1	0.0669	0	0.0524	0.0584	0	0.0638
AAAS	46.2	38.8	40.4	52	33.4	56.4
AACS	18.8	14.3	23.9	15.1	17.5	30.9
AADACL2	0.038	0	0	0	0.082	0.12
AADACL4	0	0	0	0.0259	0	0
AAMP	104	114	83.1	187	71.4	80.8
AARS2	8.27	12.6	8.61	10.5	9.8	8.58
AARSD1	42.4	47.6	23.8	26.7	61.7	24.5
AASDHPPT	17.1	28.9	32.6	18.5	15.8	22.3
AASS	8.18	0.48	1.25	11	1.28	2.38
AATK	0.452	0.598	0.45	1.59	1.44	0.727
ABCA11P	2.28	5.98	2.24	9.95	2.59	5.69

Figure 3. The data transform page makes it easy to perform operations on the matrix like log transformation or filtering to reduce and normalize data.

- Action: Transform Data Transform: Logarithmic. Log Base 10. Apply.
- Action: Transform Data Transform: Mean Center Row. Apply.
- Action: Filter Data Filter: Missing Data Row. Remove if > 50% Missing Values. Apply.
- Action: Filter Data Filter: Standard Deviation Row. Keep 500 rows with highest Standard Deviation. Apply.

After applying the transformations, the matrix contains no errors and should be suitable for heat map generation (Figure 4). Note that the left-hand panel shows the history of transformations performed on the matrix, and one can ‘undo’ back to any previous state of the matrix (including the original version) by clicking the desired previous state and hitting reset. More generally, the entire process of creating a heat map is iterative; the Next and Previous buttons can be used to return to previous steps to try different options. If, after generating the heat map, it appears that there should be more or fewer rows or different transforms, one can return to the pertinent screen and use the history and Reset option to adjust the data matrix. Finally, as an added feature, the Transform screen enables the user to download the filtered, transformed matrix for use in other analyses.

Clustering

The next step is clustering (Figure 5). The row order and column order drop-down menus can be used to select the clustering algorithm and distance measure to be applied to the

rows and/or columns. Ward’s algorithm with Euclidean distance metric is one common choice, but the menus include many other possibilities, appropriate for different purposes and data characteristics. For the sample case, the Ward/Euclidean options provide strong separation in the dendrogram and interesting groups of samples. The menus also allow the rows and columns to be left in original order or randomized. Additional options will be provided in the future.

Please be aware that clustering of larger matrices may take a few minutes to complete. (The time it takes to cluster data increases approximately as the square of the number of rows or number of columns, whichever is larger.)

Covariate bars

The next page allows covariate (classification) bars to be added to the heat map (Figure 6). Covariate bars add descriptive information about the rows or columns of the heat map. A covariate bar file has the same labels as the rows or columns in the matrix and an annotation value. In this use-case we will use TCGA clinical data to add age, smoking status, gender, and tumor stage to the heat map. The covariate file contains sample ids and clinical values – one value per line. When a covariate file is added, one must identify it as a row or column covariate and specify whether it contains discrete (categorical) data or continuous values. In this case smoker status, gender, and stage are discrete column covariates, and age is a continuous column covariate.

After covariate bars have been added, the colors associated with the covariate values can be changed. If the color scheme

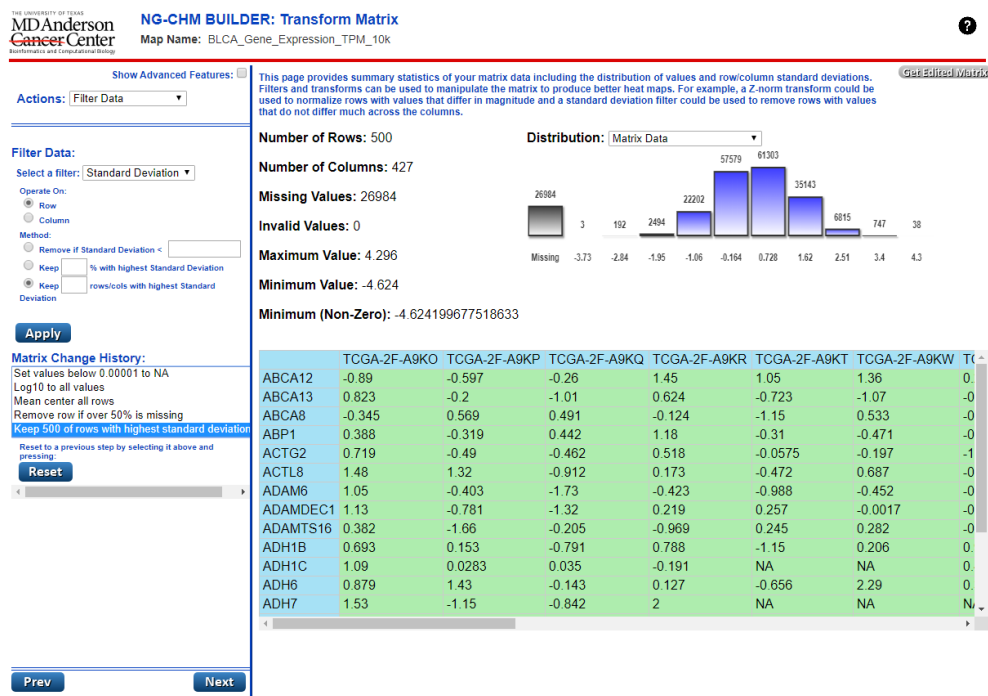


Figure 4. The transformed dataset has a better distribution and size for heat map generation than did the original. The history of transformations in the left-hand panel can be used to undo changes and revert to previous matrix states.

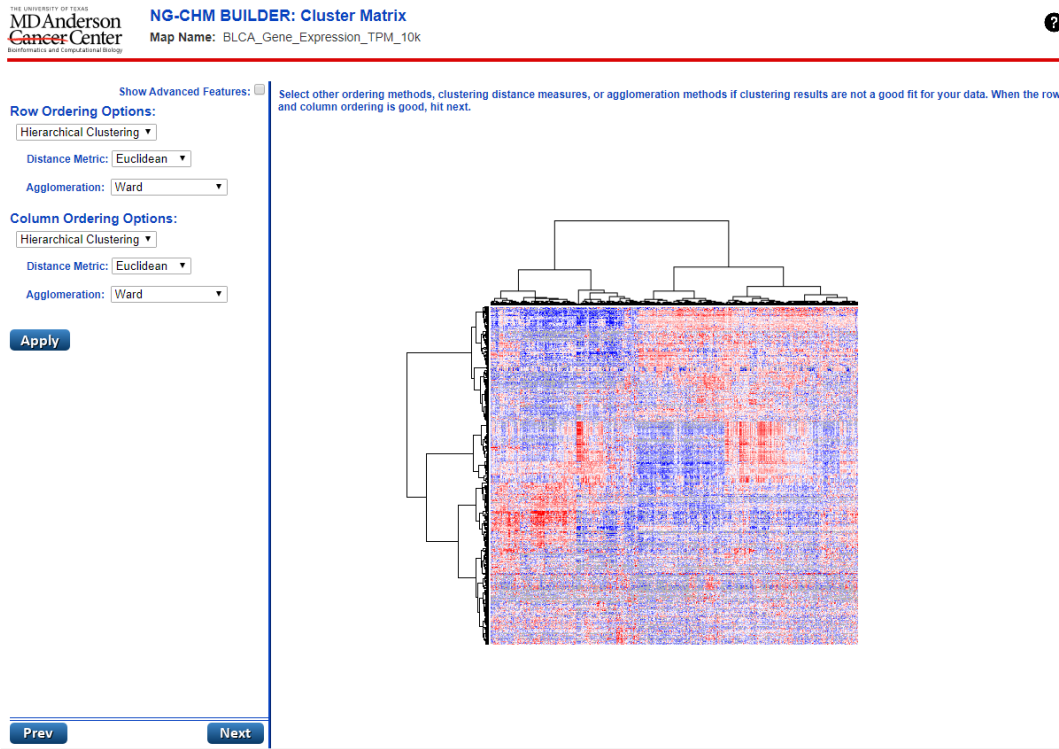


Figure 5. The clustering step supports many different clustering methods and distance measures. The Apply button performs clustering and displays the resulting dendrograms.

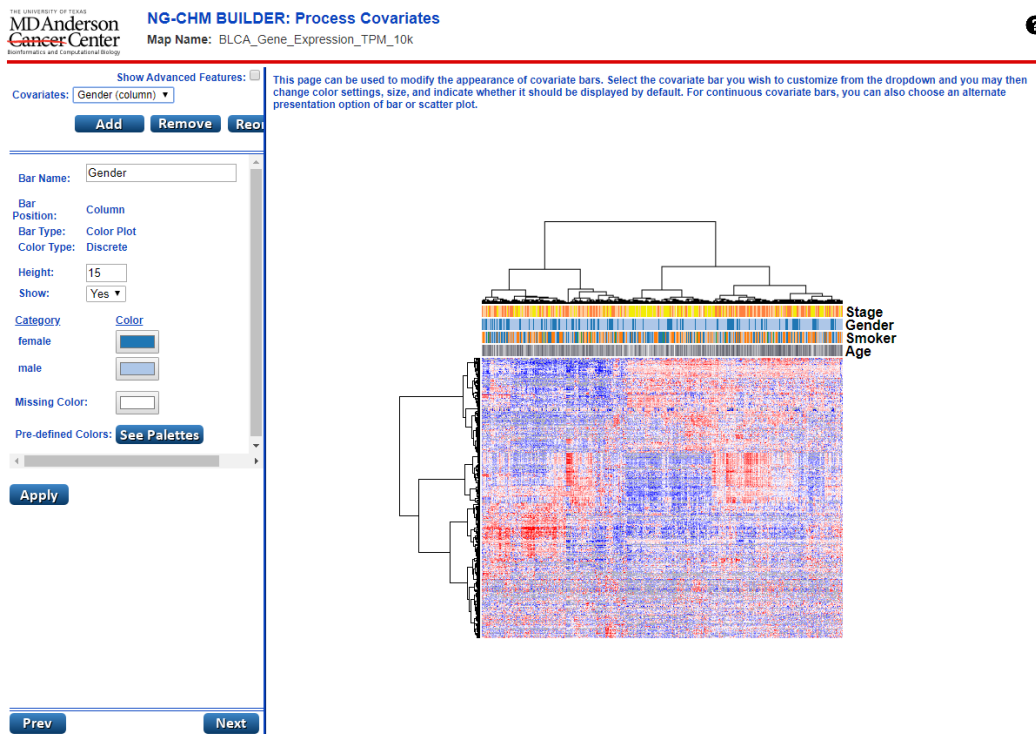


Figure 6. The covariate screen allows for the addition of supplemental data that describes the rows or columns of the data. This screen is also used to change the color of values and ordering of the covariate bars.

might be useful for other maps, the palette can be saved to the server using the See Palettes button. Covariates can be reordered on the same screen.

An advanced feature, accessed on the cluster page, is the ability to generate a covariate bar based on the clustering dendrogram. If, for example there are four distinct clusters in the data and one wants to emphasize them in discussion of the heat map, a covariate that identifies the four top clusters based on the four top branches of the dendrogram can be generated.

Another notable advanced feature is the ability to include classification data in the original matrix uploaded in the first step, rather than providing individual covariate files on the covariate page. Choosing advanced features on the first page enables the user to identify covariates as well as labels and data in the uploaded matrix.

Format heat map

The format screen (Figure 7) supports the final step in generation of a heat map, adjustments of its appearance:

- Adjustment of colors and break points in the body of the heat map.
- Formatting of labels
- Formatting of the dendrograms
- Specification of the data type of the labels for link-outs.

For this use case, several changes were made: (i) a slight adjustment to the break points to emphasize high and low values in the matrix, (ii) identification of row labels as gene symbols, and (iii) identification of column labels as TCGA sample identifiers. Associating the labels with known data types activates available type-specific link-outs to external data resources.

Interesting advanced features on the same page include the addition of ‘top items’ that will be displayed in the global (i.e., full) heat map view. For example, to show the positions of a few key genes, they can be entered on the page and will show on the global heat map display. Another powerful advanced feature is the ability to add gaps to emphasize sub-groups in the heat map.

Heat map – view and download

The heat map is now complete, but the Prev button can still be used to go back to previous build steps to try different options. On this final page of the Interactive Builder (Figure 8), the map can be explored dynamically and downloaded. The Get Heat Map PDF button downloads a PDF of the summary and/or detail views as they appear on the screen – including a version of the detailed view zoomed as desired. The legends and other metadata are shown on a separate page of the pdf. The final screen can also be used to explore the dynamic heat map by zooming, panning, searching, dendrogram selection, and link outs. Clicking the Expand Map button devotes the whole browser window to the map.

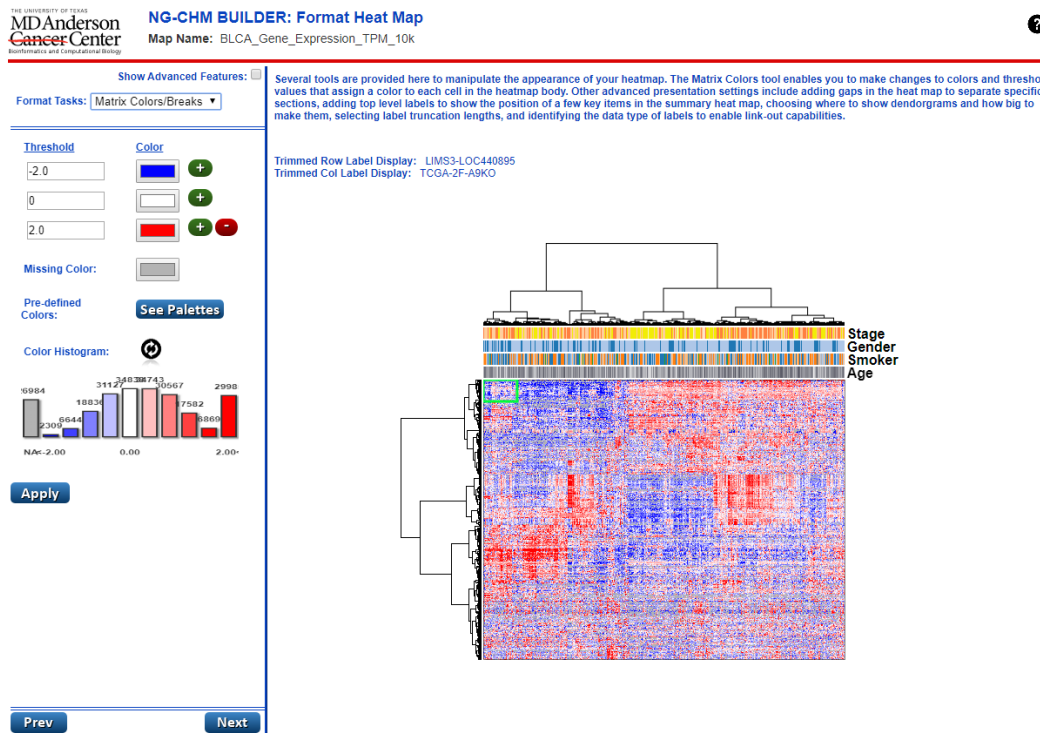


Figure 7. The format step is used to make changes to the appearance of the heat map, for example, changing the color scheme or altering the breakpoints associated with the colors. Many appearance change options are available.

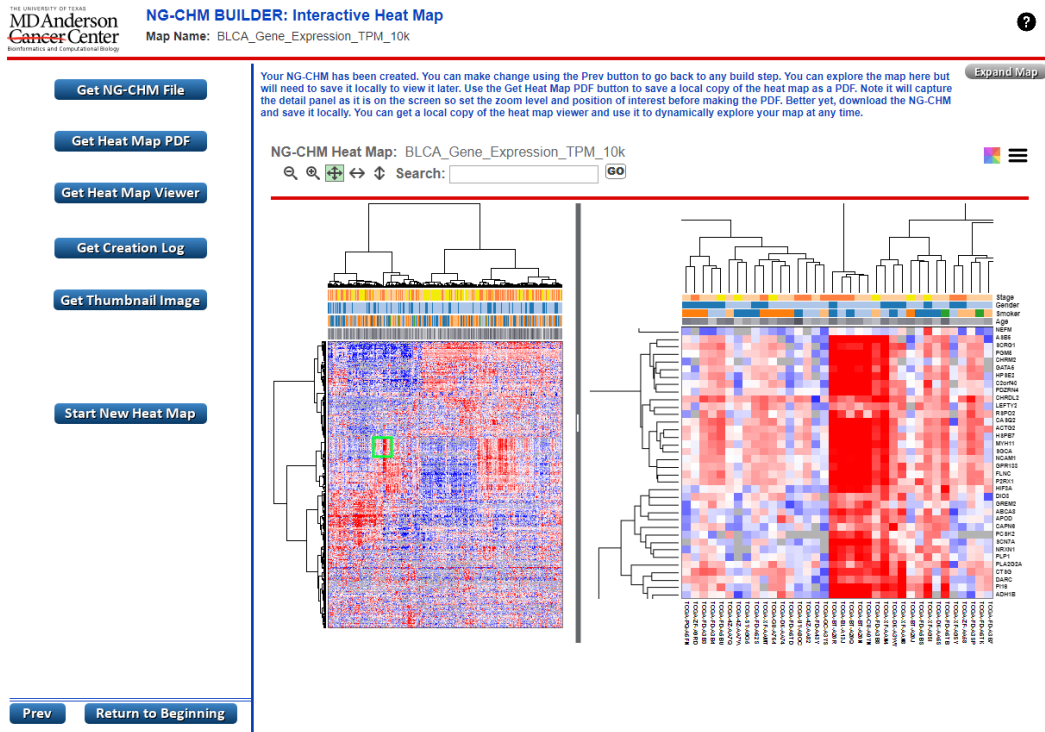


Figure 8. The heat map review and download screen shows the completed heat map, allows for dynamic exploration of the map, and provides download options for a PDF, an NG-CHM, and/or the construction history.

Heat maps constructed on the Interactive Builder website are not saved. However, NG-CHMs can be downloaded to save and explore dynamically on your own computer. Select the Get NG-CHM file to obtain a map and then select the Get Heat Map Viewer to get a stand-alone NG-CHM viewer to run on your computer. See our NG-CHM site for more details on the capabilities of dynamic heat maps, additional builders to generate NG-CHMs (Galaxy and R)², and instructions on how to embed dynamic heat maps in your websites - <https://www.ngchm.net/>. Also see our [YouTube channel](#) for tutorials on NG-CHM features.

NG-CHM

The interactive NG-CHM produced by the Builder for the use case can be viewed [here](#). Try the pan, zoom, search, and link-out features.

Reproducibility

Reproducibility of results is becoming increasingly important for publication in high-impact journals¹⁴. Therefore, it is important to be able to report the exact steps performed to transform data and create a heat map. That is particularly challenging with an iterative tool that facilitates exploration of alternative options. The Get Creation Log button on the file page of the Interactive Builder is meant to address that need. The history provided by the log shows each option, including the data transformations that were performed to produce the current map. With the original data file and the history, it is possible to recreate a heat map exactly.

Conclusions

The Interactive CHM Builder⁹ is an easy to use yet powerful tool for creating custom clustered heat maps for any type of study that generates a matrix of data. It has an intuitive step by step process to prepare the data and build high-quality CHMs. A sample dataset is built-in so it takes just seconds to try out the process and become familiar with the basic steps for heat map generation. It is also easy to back up to previous steps or data states to try alternative approaches and refine formatting. Finally, heat maps can be downloaded as either PDF files or NG-CHM files that support in-depth exploration of the maps.

Although there are many methods available to correct/normalize/filter data, perform hierarchical clustering, and present the resulting heat maps, most of them require programming and biostatistical skills. For non-programmers the options are more limited. The best-known software packages for that purpose are Cluster 3.0¹⁵ for data manipulation and clustering combined with TreeView¹⁶ for display of heat maps. Newer tools in the category include Morpheus (<https://software.broadinstitute.org/morpheus/>) and Heatmapper¹⁷. Some advantages of the Interactive CHM Builder are:

- Unlike Cluster 3.0/TreeView, no software installation and configuration are required. Interactive CHM Builder is available as a web service.
- Unlike other heat map tools, Interactive CHM Builder provides a step by step process starting with an unprocessed matrix that includes: correction of invalid/missing

values, data normalization and transformation, data filtering, clustering, addition of covariates, and advanced customization of heat map display including link outs. At each step of the process we provide histograms and incremental heat map visualizations to assist with understanding the data and the effect of option selection.

- It is a fluid tool that supports the iterative nature of heat map creation, enabling users to move easily back and forth to revisit and modify any step of the process.
- Unlike other tools, it provides a complete history of each option selected to transform the data and generate the heat map. That capability enables the user to reproduce the heat map even months or years later.
- Finally, the resulting NG-CHMs provide enhanced ability to support dynamic exploration of patterns in the data. They can be shared with collaborators and larger research communities on a website with an NG-CHM plugin or as a stand-alone heat map and viewer.

Data availability

Open Science Framework: NG-CHM Interactive Builder Use-Case Data. <https://doi.org/10.17605/OSF.IO/H7ZS2>¹³.

This project contains the sample TCGA bladder cancer matrix used in the use-case.

Data are available under the terms of the [Creative Commons Zero “No rights reserved” data waiver](#) (CC0 1.0 Public domain dedication).

Software availability

The Interactive CHM Builder is freely available for use as a web resource at: <https://build.ngchm.net/NGCHM-web-builder/>.

Source code available from: https://github.com/MD-Anderson-Bioinformatics/NG-CHM_GUI_BUILDER.

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.3460673>⁹.

License: GNU General Public License version 2.

References

- Weinstein JN, Myers T, Buolamwini J, *et al.*: **Predictive statistics and artificial intelligence in the U.S. National Cancer Institute’s Drug Discovery Program for Cancer and AIDS.** *Stem Cells.* 1994; **12**(1): 13–22.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Weinstein JN, Myers TG, O’Connor PM, *et al.*: **An information-intensive approach to the molecular pharmacology of cancer.** *Science.* 1997; **275**(5298): 343–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Myers TG, Anderson NL, Waltham M, *et al.*: **A protein expression database for the molecular pharmacology of cancer.** *Electrophoresis.* 1997; **18**(3–4): 647–53.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Eisen MB, Spellman PT, Brown PO, *et al.*: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A.* 1998; **95**(25): 14863–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Scherf U, Ross DT, Waltham M, *et al.*: **A gene expression database for the molecular pharmacology of cancer.** *Nat Genet.* 2000; **24**(3): 236–44.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ross DT, Scherf U, Eisen MB, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet.* 2000; **24**(3): 227–35.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Zeeberg BR, Qin H, Narasimhan S, *et al.*: **High-Throughput GoMiner, an ‘industrial-strength’ integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID).** *BMC Bioinformatics.* 2005; **6**: 168.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Weinstein JN: **Biochemistry. A postgenomic visual icon.** *Science.* 2008; **319**(5871): 1772–3.
[PubMed Abstract](#) | [Publisher Full Text](#)
- mstucky, flikseda, Ryan M, *et al.*: **MD-Anderson-Bioinformatics/NG-CHM_GUI_BUILDER 2.15.1.** (Version 2.15.1). *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.3460673>
- Broom BM, Ryan MC, Brown RE, *et al.*: **A Galaxy implementation of next-generation clustered heatmaps for interactive exploration of molecular profiling data.** *Cancer Res.* 2018; **77**(21): e23–e26.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Goecks J, Nekrutenko A, Taylor J: **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** *Genome Biol.* 2010; **11**(8): R86.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Robertson AG, Kim J, Al-Ahmadie H, *et al.*: **Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer.** *Cell.* 2017; **171**(3): 540–556.e25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ryan M: **NG-CHM Interactive Builder Use-Case Data.** 2019.
<http://www.doi.org/10.17605/OSF.IO/H7ZS2>
- McNutt M: **Reproducibility.** *Science.* 2014; **343**(6168): 229.
[PubMed Abstract](#) | [Publisher Full Text](#)
- de Hoon MJ, Imoto S, Nolan J, *et al.*: **Open source clustering software.** *Bioinformatics.* 2004; **20**(9): 1453–1454.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Saldanha AJ: **Java Treeview—extensible visualization of microarray data.** *Bioinformatics.* 2004; **20**(17): 3246–3248.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Babicki S, Arndt D, Marcu A, *et al.*: **Heatmapper: web-enabled heat mapping for all.** *Nucleic Acids Res.* 2016; **44**(W1): W147–53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status:  


Version 2

Reviewer Report 30 March 2020

<https://doi.org/10.5256/f1000research.25215.r61493>

© 2020 Cline M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Melissa S. Cline 

Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

The authors have faithfully addressed the reviewers' feedback. This is a well-written write up of an excellent piece of research software, and represents a fine resource for the research community.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: 20 years of experience in genomics, including RNA expression analysis and cancer data visualization.

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

Version 1

Reviewer Report 21 February 2020

<https://doi.org/10.5256/f1000research.22637.r59179>

© 2020 Cline M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Melissa S. Cline 

Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

Ryan *et al.* present a manuscript for the interactive clustered heat map builder tool that has been widely used in cancer research consortia. The tool is of excellent quality overall, is very useful, and is intuitive in its design and execution. The manuscript is well-written, with some caveats below.

Major feedback:

The authors are not doing justice to the tool, which offers much more than user-friendly heatmap generation. To put the functionality in perspective, they should contrast it with the Cluster/TreeView suite, which also offers a user-friendly interface to filtering and data transformation.

Minor feedback

The Operation subsection assumes knowledge of Docker and Tomcat. The authors should cite appropriate background reference material for readers who aren't familiar with these technologies.

For the use case, the authors summarized how they transformed the data, but did not indicate how those transformations were done with their tool. This needs to be clarified, because it's not obvious.

The sample data in the OSF Storage site is stored as a single tarball. This is awkward, as the entire tarball has to be downloaded and expanded in order to access any single file.

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.

Reviewer Expertise: 20 years of experience in genomics, including RNA expression analysis and cancer data visualization.

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

Author Response 11 Mar 2020

Michael Ryan, In Silico Solutions, Fairfax, USA

Thank you for your feedback and suggestions. Below we have described how each was addressed.

Major feedback:

The authors are not doing justice to the tool, which offers much more than user-friendly heatmap generation. To put the functionality in perspective, they should contrast it with the Cluster/TreeView suite, which also offers a user-friendly interface to filtering and data transformation.

Good suggestion. We have modified the 5th sentence in the first paragraph of the introduction to broaden the description of the scope of the tool and have included a new paragraph in the conclusion section to contrast our tool with Cluster 3.0/Treeview.

Minor feedback

The Operation subsection assumes knowledge of Docker and Tomcat. The authors should cite appropriate background reference material for readers who aren't familiar with these technologies.

We agree. Additional detail including links to the appropriate reference material for Docker and Tomcat, has been added to the Operation section.

For the use case, the authors summarized how they transformed the data, but did not indicate how those transformations were done with their tool. This needs to be clarified, because it's not obvious.

Thank you for pointing that out. The transforms section of the use case has been modified to provide the exact path through the screen options for each transform performed. That should make it easier to follow the steps exactly.

The sample data in the OSF Storage site is stored as a single tarball. This is awkward, as the entire tarball has to be downloaded and expanded in order to access any single file.

Agreed. Accordingly, we have modified the OSF Storage site to contain a folder with all of the files needed for the use case as individual, uncompressed files.

Competing Interests: No competing interests were disclosed.

Reviewer Report 29 October 2019

<https://doi.org/10.5256/f1000research.22637.r55133>

© 2019 Caplen N et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

**Natasha Caplen**

Genetics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

Soumya Sundara Rajan 

Genetics Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA

Ryan and co-workers have developed the software tool Interactive Clustered Heat Map (CHM) builder to enable investigators with minimal expertise in bioinformatics and biostatistics to generate publication-quality heatmaps. The use of heatmaps to visualize related datasets is a common feature in many reports of the results of studies that include genome or transcriptome-scale experiments. However, the statistical underpinnings of a heatmap require the application of appropriate transformation and clustering procedures. The interactive CHM tool makes use of user-uploaded data that is then processed to generate heatmaps defined by a set of standardized options; for example, the user can select different distance metrics (e.g., the calculation of Euclidean distance versus Manhattan distance) or clustering options (random versus hierarchical). The user also has the option to input possible co-variant data sets for the further stratification of the primary results. Furthermore, the user can customize the visual properties of the heatmap by selecting the output of the computational pipeline from a palette of colors. The article itself is well-written, though, as stated below, we recommend some edits to the current text. A particularly positive feature of this CHM tool is the inclusion of a dynamic capability that allows the user to explore their data in greater depth. Many of the features of the graphical user interface (GUI) are easy to use, and the user does not have to refer to the accompanying article describing the builder software continually. However, to enhance the impact of this resource, we recommend modification of the current versions of their article and software tool to address the following points.

Article

In the Introduction, the authors discuss the user's ability to use their tool to generate heatmaps iteratively, refining data transformation, annotation, clustering, and formatting. The authors also point out that this may introduce the risk of generating a multiple-comparison issue. To help the user avoid such issues, can the authors briefly mention other resources (e.g., review articles) that the user can refer to when considering which of the transformation, clustering, and distance metrics will be most applicable to their dataset?

The authors should include a discussion of how the interactive HCM builder compares to other free heatmap generators available, for example, heatmapper.ca; Babickiet al., *Heatmapper: web-enabled heat mapping for all* [Nucleic Acids Res. 2016 May 17 \(epub ahead of print\). DOI:10.1093/nar/gkw419](https://doi.org/10.1093/nar/gkw419)); and the Morpheus software from the Broad Institute (<https://software.broadinstitute.org/morpheus/>).

Some datasets require non-hierarchical clustering to obtain the most appropriate and meaningful interpretation of the results. Please explain why this software provides only either hierarchical, random, or no clustering options?

Website

Some test runs found that when the user runs through the work-flow and generates a heatmap using a dataset, the generation of a new heatmap either using the same dataset or a different dataset requires the user to close the website and re-open the homepage. The re-set function may need modification.

Some test runs found that when choosing the formatting and then palettes after adding co-variants, the apply button on the left-hand window has lines running through it.

It is easy to maneuver and resize the highlighter box over any region of the heatmap generated using the sample data. However, we noted not all heatmaps performed as well using user-uploaded data.

Please state clearly on the website's front-page that the website limits the heat map to "no more than 4,000 total rows and columns and no more than 3,500 elements on either axis." In the absence of this statement on the front-page, users may attempt to upload more complex datasets.

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?

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.

Reviewer Expertise: Functional genomics. We have relevant expertise in the generation and interpretation of complex 'Omic scale datasets, but not in the statistical analysis that underlays the tool described. Our viewpoint represents that of the potential user of the tool described in this study.

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

Author Response 11 Mar 2020

Michael Ryan, In Silico Solutions, Fairfax, USA

Thank you for your detailed comments and suggestions on the article and the tool. Each suggestion/comment is addressed below:

Article

In the Introduction, the authors discuss the user's ability to use their tool to generate heatmaps reiteratively, refining data transformation, annotation, clustering, and formatting. The authors also point out that this may introduce the risk of generating a multiple-comparison issue. To help the user avoid such issues, can the authors briefly mention other resources (e.g., review articles) that the user can refer to when considering

which of the transformation, clustering, and distance metrics will be most applicable to their dataset?

We are not aware of any review article that covers the topic comprehensively. But, in the text, we do cite an article of ours (Weinstein JN: A postgenomic visual icon. *Science*. 2008;319: 1772) that provides additional background on some of the relevant subtleties of heat map generation. As you correctly point out, the optimum approach depends on the specifics of an individual dataset and objectives of the study. In light of your comment, we will consider writing a review article that addresses those issues at more length. Also, we are contemplating a future enhancement to the Interactive CHM Builder that would provide templates or workflows based on study type as a starting point to assist with navigating data transformations and heat map generation.

The authors should include a discussion of how the interactive HCM builder compares to other free heatmap generators available, for example, heatmapper.ca; Babickiet al., *Heatmapper: web-enabled heat mapping for all Nucleic Acids Res.* 2016 May 17 (epub ahead of print). DOI:10.1093/nar/gkw419); and the Morpheus software from the Broad Institute (<https://software.broadinstitute.org/morpheus/>).

Thank you for the suggestion. We have added a paragraph to the conclusions section to enumerate what we feel are the advantages of our Interactive CHM builder compared with the other similar tools.

Some datasets require non-hierarchical clustering to obtain the most appropriate and meaningful interpretation of the results. Please explain why this software provides only either hierarchical, random, or no clustering options?

We agree. Thank you. The methods we have currently implemented are the ones that are most heavily used in publications of omics research. We will add non-hierarchical clustering methods to our requested features list for a future release.

Website

Some test runs found that when the user runs through the work-flow and generates a heatmap using a dataset, the generation of a new heatmap either using the same dataset or a different dataset requires the user to close the website and re-open the homepage. The re-set function may need modification.

Thank you for reporting this issue. We have modified the restart flow and believe the problem has been corrected.

Some test runs found that when choosing the formatting and then palettes after adding co-variants, the apply button on the left-hand window has lines running through it.

We have been unable to reproduce that issue in the latest release of the software so we believe it has been corrected. If you encounter it again, we would appreciate it if you submit a git issue, noting the browser and operating system for which it occurs.

It is easy to maneuver and resize the highlighter box over any region of the heatmap generated using the sample data. However, we noted not all heatmaps performed as well

using user-uploaded data.

Thank you for the report. Since submission of the paper, we have made several improvements to the selection/sizing features and have tested many odd sized asymmetrical matrices. We will continue to implement improvements in selection mechanics if additional issues arise.

Please state clearly on the website's front-page that the website limits the heat map to "no more than 4,000 total rows and columns and no more than 3,500 elements on either axis." In the absence of this statement on the front-page, users may attempt to upload more complex datasets.

For many studies, an important step in preparing data for clustering and heat map generation is filtering out rows and/or columns that have a high proportion of missing values or that show little variance across samples. We want to allow users to upload matrices that are above the clustering limit because the filtering step will often reduce the size of the matrices such that they can be clustered. The manuscript was not clear on this point. Thank you for pointing this out. We have modified the Use Case section "Transform/filter the data", paragraph 2, to explicitly discuss clustering limits and the use of filtering to reduce larger datasets.

The interactive nature of the tool does limit the maximum matrix we can cluster. As you know, the compute time for most clustering algorithms essentially increases as the square of the largest dimension. The tool's limit for the clustering step has been increased from 4,000 to 5,000 total rows/columns. The 3,500 axis limit has been removed. We've also added new system messages that more clearly explain those issues, and we plan to continue pursuing increases in the limits as computational power increases and clustering algorithms advance.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

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

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

F1000Research