

Analyzing Genomic Data with NOJAH

Introduction

Not Just Another Heatmap (NOJAH) is a web-based comprehensive tool for heatmap analysis. It is intended for the researchers who are interested in (A) defining a core gene subset with core sample-set with the Genome-Wide Heatmap Analysis (GWH) (B) combining the results of the clustering analysis for up to three data types (such as Expression, Methylation and Copy Number) with the Combined results Clustering (CrC) Analysis, and (C) defining the significance of a gene-set in separation of the clusters with the Significance of Clustering (SoC) analysis. Each analysis is independent and is hosted in an individual tab. The *Supplementary Methods* of the NOJAH manuscript (see citation tab for manuscript) includes details of individual methods.

The tool is hosted on a 64 bit CentOS 6 server with a 8 core processor and 64 GB RAM (<http://bbisr.shinyapps.winship.emory.edu/NOJAH/>) running the Shiny Server program designed to host R Shiny applications. This tool has been extensively tested on Windows 7 and MacBook Pro 10 operating system with chrome and firefox browser (versions 70.0.3538.77 (official Build) (64 bit) and quantum 62.0.3 (64-bit), respectively). The example data provided on NOJAH includes Genome-Wide RNA-Seq TCGA-BRCA dataset consisting of 60, 000 genes and 25 samples in the GWH tab. The core sample with core gene-set output from the GWH tab is used as input in the CrC and SoC tabs. Given the genome-wide dataset, it took less than two minutes to identify a core gene-set with core sample set. Similarly, each expression, methylation and copy number dataset of 1000 genes and 25 samples, within the CrC analysis took less than a couple of seconds to compute cluster results and less than five seconds to display the combined results clustering results. The time taken to generate the significance of cluster analysis depends on the number of genes and the number of bootstrap iterations requested; it took less than two minutes for the expression data with 1000 bootstrap iterations.

Running NOJAH on a local computer

1) Download and install R and/or RStudio (version 3.3.2. or later) from <https://cran.rproject.org>

2) Open Rstudio and install the below required packages:

```
> install.packages(c("shiny", "shinythemes", "shinycssloaders", "RColorBrewer", "gplots", "gdata",  
"plyr", "DT", "dendextend", "colourpicker", "cluster", "reshape", "gridExtra", "Diagrammer",  
"changepoint", "rhandsontable", "matrixStats", "ggplot2"))  
> if(!requireNamespace("BiocManager", quietly = TRUE))  
> install.packages("BiocManager")  
> BiocManager::install("ConsensusClusterPlus", version = "3.8")  
> BiocManager::install("impute")
```

3) Users can run NOJAH locally using the source code available from the GitHub:

<https://github.com/bbisr-shinyapps/NOJAH>, by typing the below commands in R console:

```
> library(shiny)  
> runApp("NOJAH")
```

4) Users can also download and run the app from GitHub directly using:

```
> shiny::runGitHub('NOJAH', 'bbisr-shinyapps')
```

Data Input Requirements

Data should be input as a TXT or a CSV file. The first two rows of the data file have information about the patients/specimens and their response/subtype; all remaining rows have gene expression data, one row per gene. In the case of Microarray gene expression data in which there are several probes corresponding to a single gene, a unique identifier would need to be created to separately identify each probe such as, 'Gene 1_p1', 'Gene1_p2' indicating Gene 1 has two probes. The columns represent the different experimental samples. A maximum of up to 10 different sample groups and 6 different gene groups may be used with this tool.

Data Format

1. The first line of the file contains the gene identifier 'gene_id' (column 1), gene group classification 'Groups' (column 2) followed by the patient IDs e.g. TCGA.01.1A2B, one per column, starting at column 3. Column 1 gene identifier must be labelled 'gene_id' and column 2 header should be labelled 'Groups' for using this tool. Other titles may cause the program to display errors.
2. The second line of the file contains the patient response classification e.g. Fav/Unf for favorable outcome group vs the unfavorable outcome group or Normal/Tumor, etc., in alphabetical order, starting at column 3. The first two columns for this row should be blank.
3. The remaining lines contain gene expression measurements one line per gene, described in the format below.
 - a) Column_1. This should contain the gene name, for the user's reference. Each row should be unique. For microarray data, for example, gene and probes can be combined into a single identifier using any delimiter except the '|'. Delimiters such as >,;,:#%&(!)_+ are acceptable.
 - b) Column_2. This should contain the gene group classification e.g. O/U for Over-expressed/Under-expressed or Hyper/Hypo for hypermethylated/hypomethylated in alphabetical order. If only one gene group, use any alphabet e.g. A or even na for each row instead.
 - c) Remaining Columns. These should contain the expression measurements as numbers. Data inputted should be non-negative. Columns and rows with zero variance should be removed from the data. Rows containing missing expression measurements, should be also be removed from the input data or it will cause the tool to run into errors.

Example format for Data

gene_id	Groups	GSM998	GSM187	GSM461	GSM768	GSM877	GSM112	GSM125	GSM311	GSM498	GSM127
		MM	MM	MM	MM	MM	MUGS	MUGS	NPC	SM	SM
Subtype		Classical	Classical	Neural	Pro-neural	Pro-neural	Classical	Mesenchymal	Classical	Neural	Neural
YWHAE>210996_s_at	na	1.47	2.18	5.87	9.12	7.34	1.56	3	7.77	3.4	1.56
YWHAE>201020_at	na	1.98	7.93	2.76	9.11	8.46	0.98	5.98	8.19	8.91	5.98
YWHAH>33323_r_at	na	8.02	8	2.17	10.12	8.76	9.76	3.76	0.02	3.67	7.94

Table 1: Example dataset for one gene group (marked na) and four patient groups (MM, MUGS, NPC and, SM). Additional subtype information is appended above the numeric data. Numeric data starts at row four and column three. Any missing entries in subtype should be coded as none. NA or blank may display errors.

Note:

- Duplicate gene names are not allowed and will cause program to throw errors. In such cases, using full id is recommended. For example, YWHAE>210996_s_at and YWHAE>201020_at instead of duplicating gene id, YWHAE.
- Rows with zero variance may cause the program to run into errors and should be removed before clustering.

Getting Started:

TAB A) GENOME-WIDE HEATMAP (GWH) ANALYSIS

Genome-Wide Heatmap analysis is available for users who wish to perform heatmap clustering on GW of any data type, for example, expression, methylation, variant or copy number. The GWH analysis workflow allows user to (1) define a most variable gene set (a.k.a., 'core genes'); 2) perform cluster analysis using core genes and construct heatmap of results; 3) estimate the number of clusters; 4) define a core sample set and update the heatmap using both core genes and core samples. The analysis workflow is created to be performed sequentially. Each step of the workflow can also be used individually with the own user defined data.



A Genome-Wide Heatmap can be very dense. Given the limitation with the computational power required to construct a genome wide heatmap, NOJAH showcases a [Genome-Wide Dendrogram](#).

Genome-Wide Heatmap Analysis workflow is divided into four main subparts:

1. [Define Core Features with Most Variable Approach](#)
2. [Heatmap of Core Features](#)
3. [Define Cluster Number](#)
4. [Define Core Samples](#)

Heatmap is *updated* based on the Core Features with Core Samples.

When using the analysis workflow, each step of the workflow is intended to be used sequentially i.e. the output of step 1 is fed into step 2 as input and so on. However each of these components can also be used independently. For example, if only consensus clustering needs to be performed then the 'Cluster Number' tab can be used.

Step 1:

Select the example dataset or upload your own. Genome-Wide TCGA-BRCA Expression datasets is available.

To view example data and format, use download button to view CSV file.

Step 2:

Choose to display Genome-Wide dendrogram. Depending on size of data, it may take a few seconds to minutes.

Step 3a:

Select method of sub-setting. You can use the boxplot on the main panel to help choose the method. In the TCGA BRCA ds, IQR shows relatively larger spread in comparison to VAR and MAD.

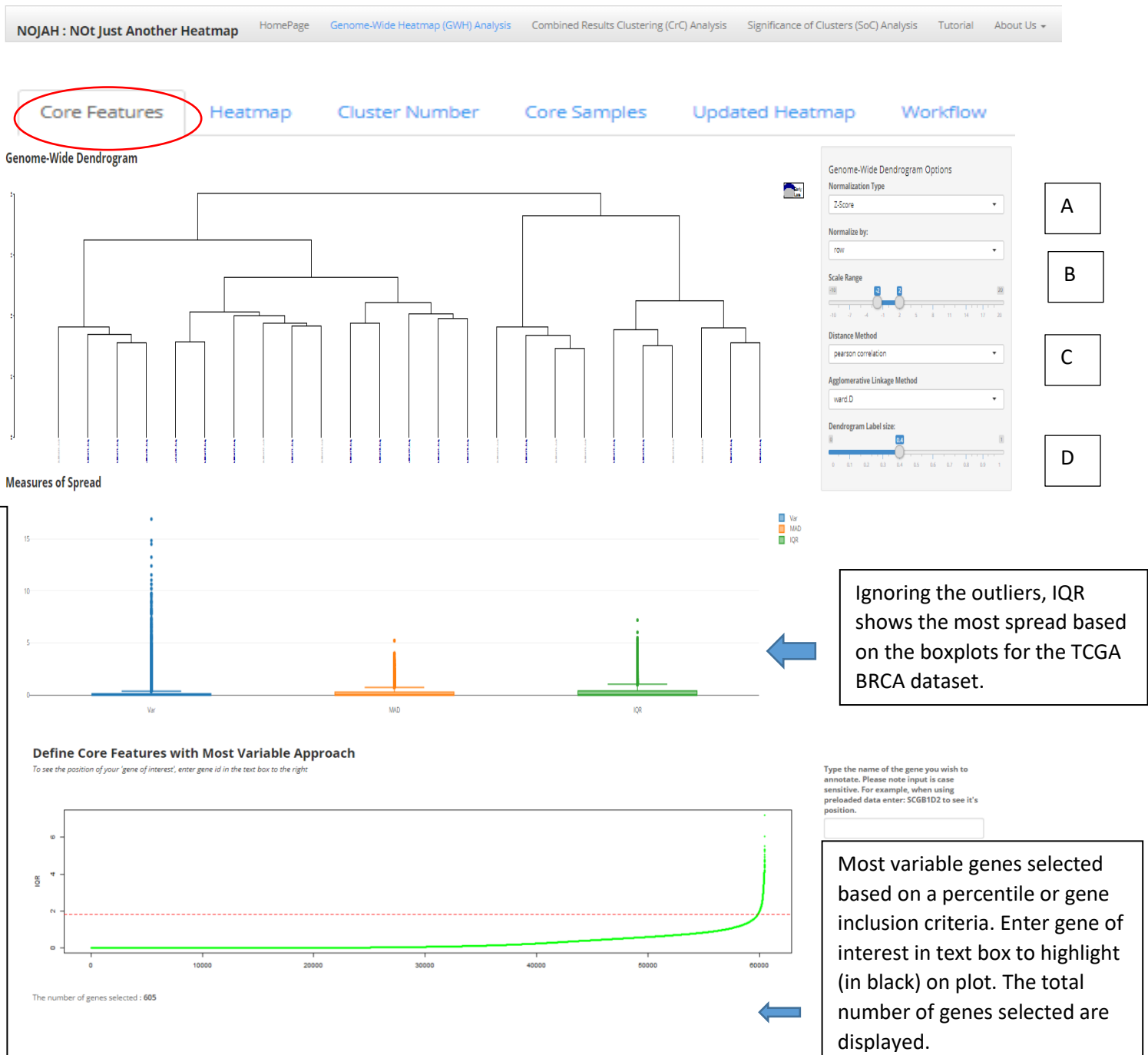
Step 3b:

Choose a percentile cut-off to select the top most variable number of features (genes in this case). You can also choose cut-off based on the inclusion of a gene. Here 99th percentile i.e. top 1% shows increased variability.

Click Run button to display results in main panel each time any parameters are updated. The total number of selected genes are displayed in the main panel.

Core Features

A Genome wide dendrogram can be quite dense and not necessarily informative. When requested, NOJAH displays an Interactive Genome-Wide dendrogram based on three different normalization and scaling methods each, eight-different distance, and seven different agglomerative linkage methods (Description on page #4). Only a handful of genes are variable across the samples. These genes can be filtered out based on measures of spread calculated for each gene across all samples. The three measures of spread available are variance (VAR), median absolute deviation (MAD) and inter-quartile range (IQR). Based on the spread of the individual measures (the larger, the better), user selects a single or a combination of these measures (for example, VAR and IQR). Additionally, a scatter plot of the ordered statistic (for example, ordered IQR for a single measure vs sum of ranks of VAR and IQR for a combined measure) aids user determine the percentile cut-off to choose the number of genes to define this core gene set. A single gene can be marked on the scatter plot and used as inclusion or exclusion criteria to select the core gene set.



Heatmap

Once the core gene set is chosen, data is automatically transferred to the heatmap tab for clustering analysis. This step helps visualize if the core gene-set is able to cluster samples based on their phenotype or sample groups. Data is scaled before input into the heatmap. NOJAH displays an Interactive Genome-Wide dendrogram based on three different normalization and scaling methods each. User can also choose between the eight-different distance and seven different agglomerative linkage methods. The choice of the normalization, scaling, and clustering measures is dependent on the data. From, the heatmap of the core gene set, a subset set of genes characterizing a cluster can be obtained (for example, the left quadrant includes genes that are upregulated in the left cluster dominated by the Early group and so on). This publication quality gene-set heatmaps characterizing the sample groups can be directly reported. Separate tabs display column and row dendrograms. The gene-sets and sample sets with their cluster annotations are also available for download.

Heatmap of Core Features

Subset HeatMap Input

Download subset HeatMap

Choose pre-loaded data or Input own subset data and Download Subset Heatmap as pdf file with minimal information required for reproducibility. User can choose file name.

HeatMap

Column Dendrogram

Row Dendrogram

Heatmap of Core Features

Heat Map Options

Clustering Measures

Heat Map colors

Interactive Heatmap of the top most variable genes selected using the above criteria. Separate tabs display column and row dendrograms. (See tutorial for part C for detailed run-through of heatmap options, Page #14)

A. By default, data is z-scored before input into the heatmap.2 function. Alternatively, modified z-scored data or unscaled can be used.

B. Data can be normalized by row, column or both, default: both

C. Scale is set from -2 to 2 but can be changed by the user

D. Choose clustering and distance measure, pearson and ward.D respectively are defaults

E. Supervised row-wise or column wise clustering can be selected using FALSE option. Row and column labels can be displayed using TRUE option

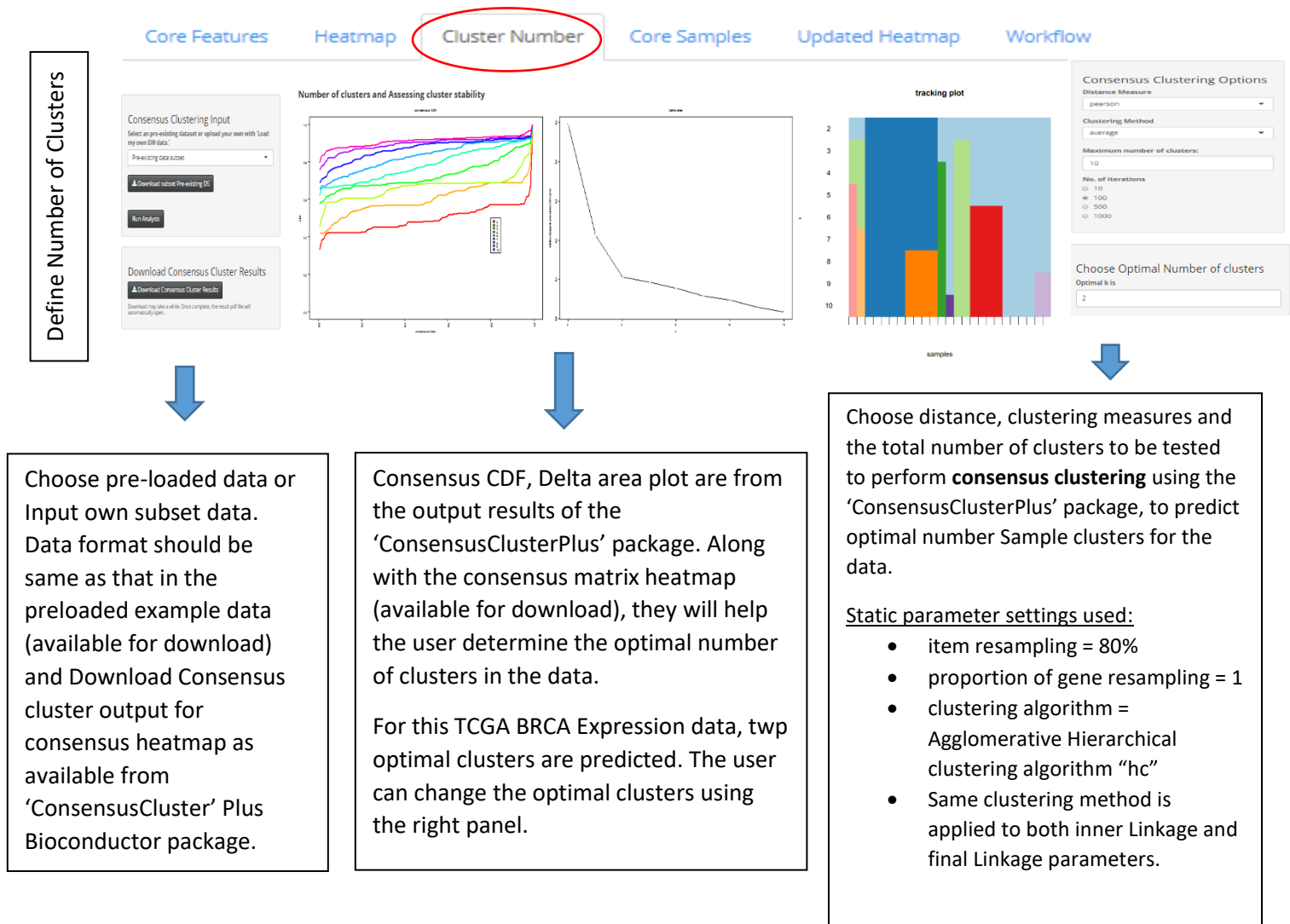
F. Change color of Heatmap by clicking on the high, mid and low colors

Details on heatmap options are available on page #12.

5

Cluster Number

The cluster number tab, allows the user to validate the actual numbers of sample clusters in the core gene-set based on the consensus clustering approach from the ConsensusClusterPlus bioconductor package [1]. The core gene-set output data from the core features and the heatmap tab is automatically input into the cluster number tab. User can perform consensus clustering analysis using the same eight distance and seven clustering measures. The choice of the clustering measures is data dependent. Additionally, the number of iterations and the maximum number of clusters to be tested also need to be provided. The output from the consensus clustering package are the consensus clustering heatmaps based on each number of cluster tested, CDF and delta plots, which aid the user to determine the number of the sample clusters in their core dataset (see [1] for interpretation details). The consensus clustering heatmap can be visualized after downloading the PDF results.



Core Samples

The core samples tab aids users to visualize how the distribution of the samples within each cluster based on the number of clusters. The output data from the cluster number tab namely, the core gene set and the number of optimal sample clusters is used as input in this tab. A silhouette plot is provided as output to help users visualize the closeness of each sample within each cluster and corresponding to the other clusters based on a silhouette width. The larger the silhouette width, the better the stability. Samples with negative and/or lower silhouette width signify poor cluster stability and are

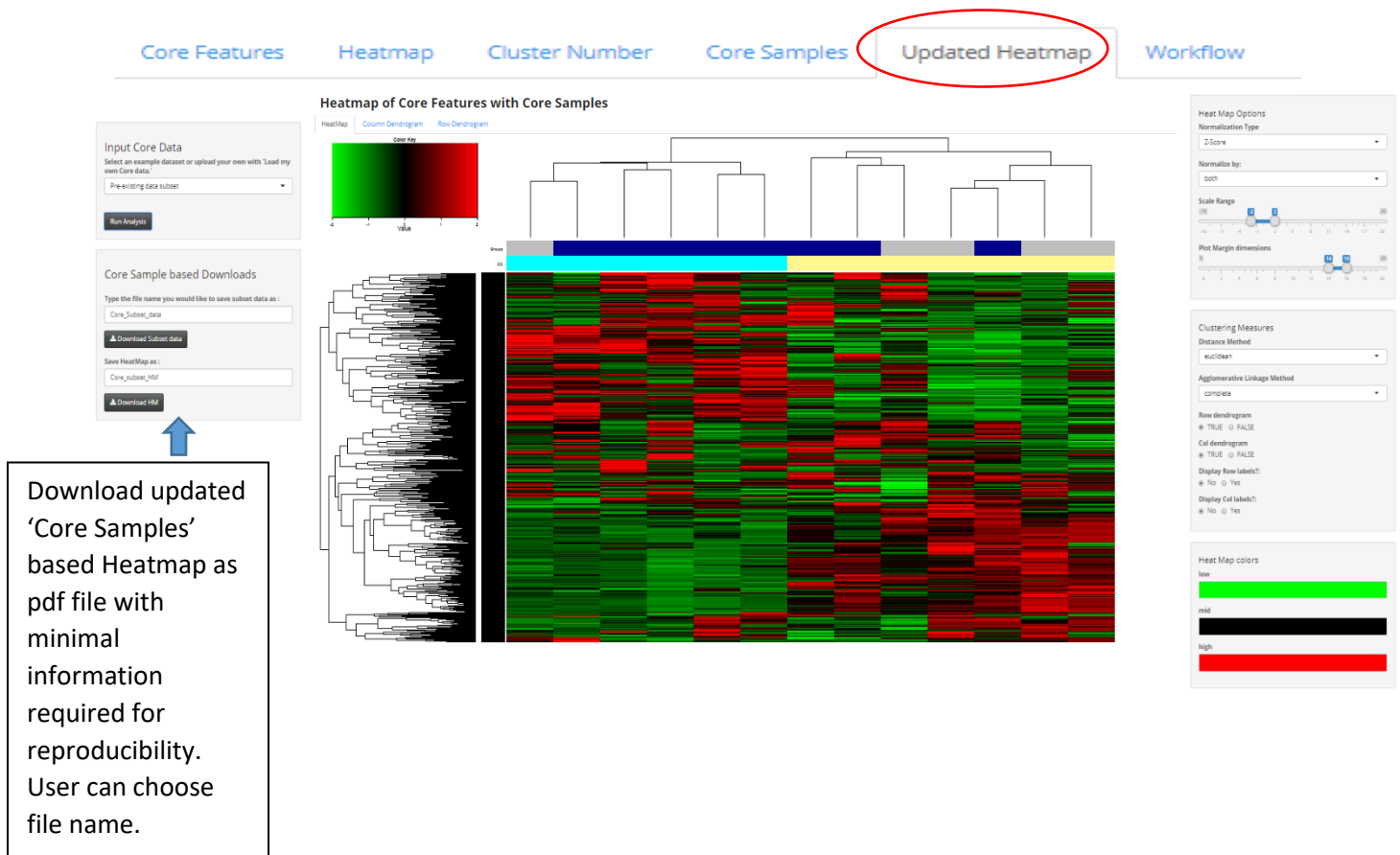
removed to create the core sample set. These samples could be identified as having the potential to misclassify into a different sample cluster given that lack of tightness with their own sample type.

Three different options are provided to help users to define the core sample set, namely, the silhouette width cut off, percentile cut off and a change point method for each cluster. The choice of the method is data dependent and specifically depends on the number of samples in each cluster. For example, the change point model is better suited for greater than five samples per cluster to provide optimal results. The percentile and manual cut off can be used when a relative or a specific number of samples from a cluster need to be removed but neither of the methods need a larger number of samples in each cluster.



Updated Heatmap

The updated heatmap helps visualize the clustering of the core gene set after the core sample set is defined. The idea is that core sample with the core gene set data, would show better clustering than in step 2. For example, fewer samples would be miss-classified into a different cluster resulting in tighter gene and sample clusters which can also be visualized by quadrants or areas in the actual heatmap. The updated heatmap is based on the consensus clusters (CC) from the core samples tab. The original sample clustering is available as the column color bar over the CC. The same choices for normalization, scaling, and clustering and the row and column dendrograms are available and can be used depending on the actual data.

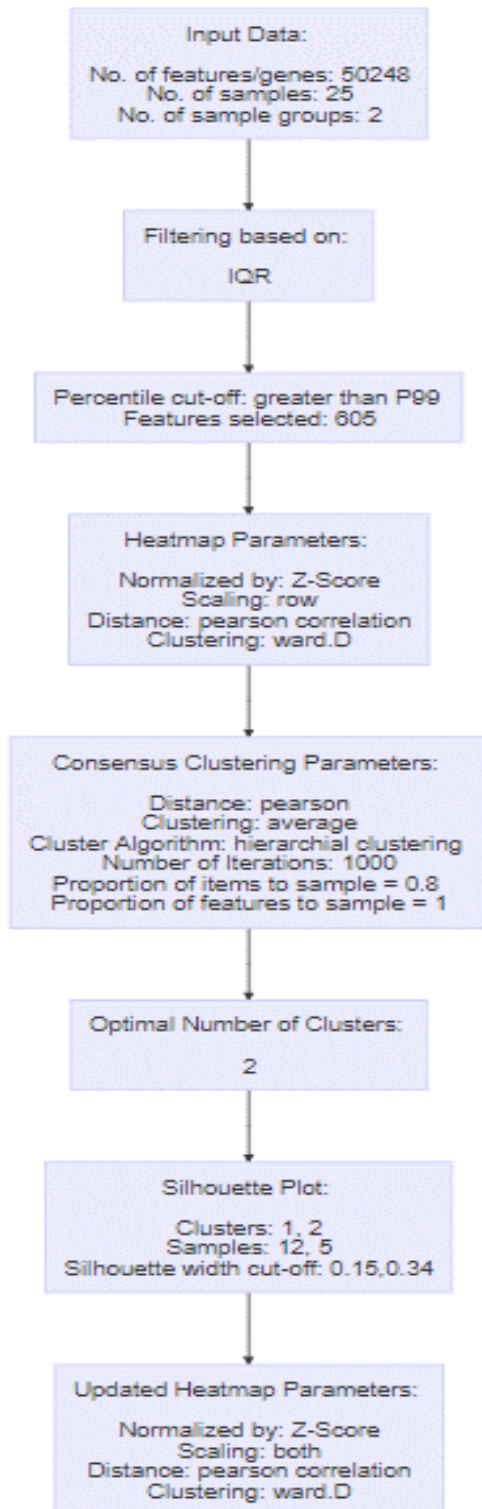


The GWH analysis tab of NOJAH should be used by the users as an iterative tool. In cases where a clear separation of the core samples and/or core-gene clusters is not seen, a different set of parameters can be used upstream (for example in using different set of filtering options such as VAR and IQR instead of IQR, or trying different clustering options when defining cluster number and in turn core samples) and the process can be repeated to yield a crisper heatmap of the core sample with core gene-set.

Workflow

One of the novelties of NOJAH is that it produces a workflow diagram with the exact parameters that the user used in each step of the workflow. This workflow image stores the exact settings that were used which aids replicability. It is available even when a single step or all steps of the GWH workflow are used. It can be downloaded using a snipping tool for Windows users and the equivalent grab tool for MacOS users.

Workflow for Genome-Wide Heatmap (GWH) Analysis



A complete workflow for the Genome-wide analysis is available in the workflow tab based on parameters selected on each tab. The workflow would be displayed for each tab in which the analysis was performed.

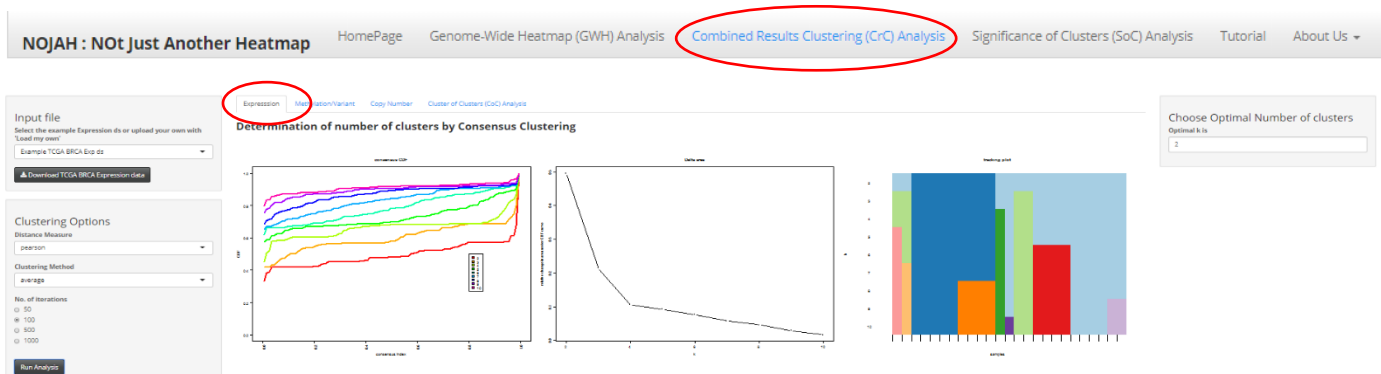
TAB B) Combined Results Cluster (CrC) ANALYSIS

Combined results Cluster Analysis is available for users who wish to perform cluster of cluster analysis which combines the results of clustering for multiple data types with the same samples in each datatype. If the same samples are not used, the CrC analysis can't be performed. In NOJAH up to three data types can be used, for example, expression, methylation/variant and copy number. Typically, the most variable or core gene set for each data type is used for clustering which can be obtained using the GWH analysis tab of NOJAH. Although, clustering for each data type can be performed using on any of the commonly available clustering tools such as consensus clustering, agglomerative heatmap clustering, k-means clustering, etc., NOJAH uses the ConsensusClusteringPlus Bioconductor package to define sample clusters based on each data type. The choice of consensus clustering is because consensus clustering is an iterative process which allows cluster visualization through consensus heatmaps, CDF and delta plots.

Data Type 1: Expression

Consensus clustering for Expression data is based on the same core gene set that was output in step 2 of the GWH analysis. The same parameters for consensus clustering are available and can be tweaked based on the data. In addition to performing consensus clustering, a silhouette plot for the sample clusters is available to assess homogeneity among the samples.

**Note: The same steps apply to Expression, Variant and Copy Number tabs.*



Input file

Select the example Expression ds or upload your own with 'Load my own'

Example TCGA BRCA Exp ds

Download TCGA BRCA Expression data

Clustering Options

Distance Measure

pearson

Clustering Method

average

No. of iterations

☐ 50
☒ 100
☐ 500
☐ 1000

Run Analysis

Download Results

Consensus Clustering

Download Expression clusters

Download may take a while. Once complete, the result pdf file will automatically open.

Step 1:

Select a pre-filtered RNASeq **Expression** TCGA BRCA file or input your own expression file.

Step 2:

Choose distance and clustering measures to perform **consensus clustering** using the 'ConsensusClusterPlus' Bioconductor package, to predict optimal number of Sample clusters for the data.

Step 3:

Choose the no. of iterations. Default is set to 100 for faster computing. In practice, set this to 1000 iterations.

Static parameter settings used:

- item resampling = 80%
- proportion of gene resampling = 1
- maximum evaluated k = 9
- clustering algorithm = Agglomerative Hierarchical clustering algorithm "hc"
- Same clustering method is applied to both inner Linkage and final Linkage parameters.

Data Type 2: Methylation/Variant

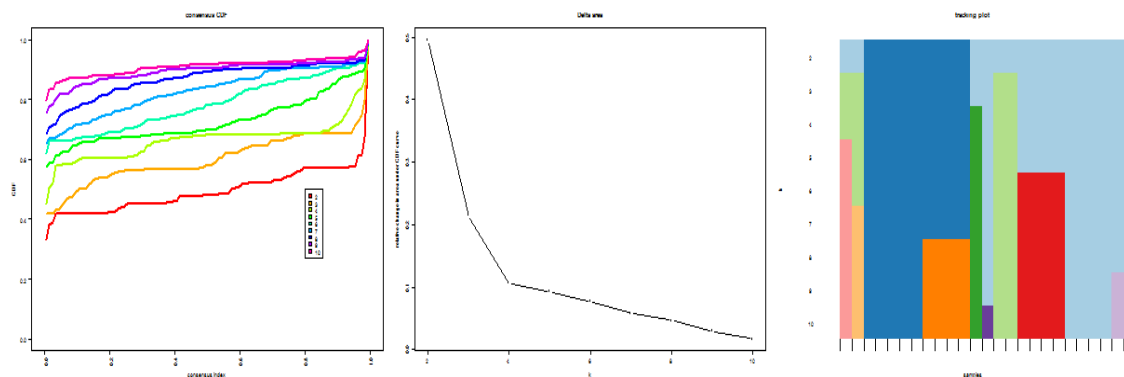
Expression

Methylation/Variant

Copy Number

Cluster of Clusters (CoC) Analysis

Determination of number of clusters by Consensus Clustering

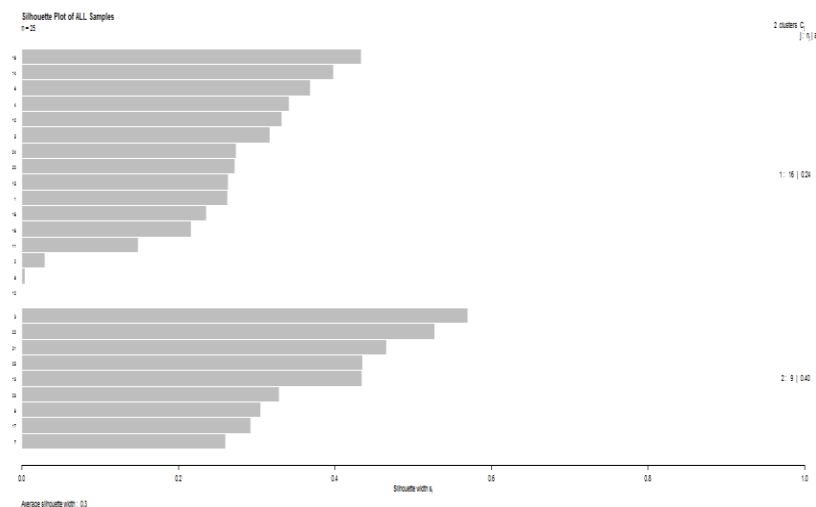


Choose Optimal Number of clusters

Optimal k is

2

Silhouette Plot



These plots will be displayed using the parameter setting above like the GWH tab.

Consensus CDF, Delta area plot are from the output results of the 'ConsensusClusterPlus' package. Along with the consensus matrix heatmap (available for download), they will help the user determine the optimal number of clusters in the data.

For this Expression data, two optimal clusters are predicted. The user can change the optimal clusters using the right panel.

Silhouette Plot can further help confirm the identification of the number of clusters visually. The larger the average silhouette width, the more reliable the cluster structures are.

Expression

Methylation/Variant

Copy Number

Cluster of Clusters (CoC) Analysis

Data type

Choose data used

☒ Methylation

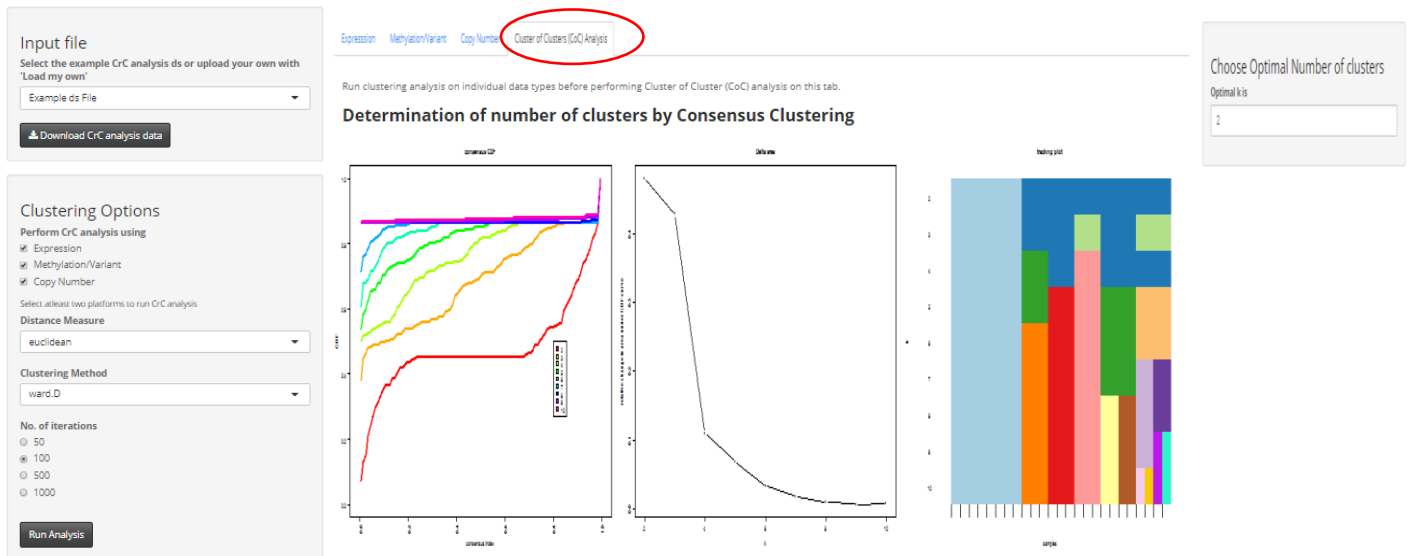
☐ Variant

In the Methylation or Variant tab, the user can further select the data type used. This choice will be displayed in the CrC heatmap row label.

CrC Analysis

In the CrC analysis tab of NOJAH, the consensus clustering based results performed on each data type are then combined to perform a cluster of cluster analysis (details are available in the Supplementary section of NOJAH). Alternately, the user can input sample clustering based on different methods such as k-means, heatmap, etc. in the CrC tab to perform cluster results cluster analysis. The output includes the consensus clustering plots based on the 1-0 matrix of clusters as rows and samples as columns. The heatmap of the same data with the consensus clusters is available. The blue areas of the heatmap show which data type cluster (example E1 or V1) is present (instead of upregulated) in the consensus cluster. Depending on the blue regions conclusions can be made about the contributions of specific data type clusters in a consensus cluster. Additional clinical features such as sample based clinical or mutation information when appended on the top of the heatmap can help users understand if one cluster includes majority of the patients of one clinical category.


The novelty of the NOJAH tool is that in addition to the heatmap with the consensus clusters based on the clustered 1-0 matrix, if the data points for each data are available, a boxplot and/or scatter plots of variance vs mean, by data type, are available. The scatter plots aid users to understand the separation of clusters in each data type (the farther the better). The boxplots show the relative difference between the clusters within that data type. For example, if E1 has samples with median gene expression greater than those in E2 and so on. This helps users better understand the blue areas of the heatmap are due to increased or decreased expression and/or copy number and/or variant. A contingency table is also available which helps users understand the sample distribution between the two data types when stratified on the third. A fisher's exact test can help determine if there is there is any significant association between the data types.



Input file

Select the example CoC analysis ds or upload your own with 'Load my own'

Example ds File

 Download CoC analysis data

Clustering Options

Perform CoC analysis using

- ☒ Expression
- ☒ Variant
- ☒ Copy Number

Select atleast two platforms to run CoC analysis

Distance Measure

euclidean

Clustering Method

average

No. of iterations

- ☒ 50
- ☐ 100
- ☐ 500
- ☐ 1000

Run Analysis

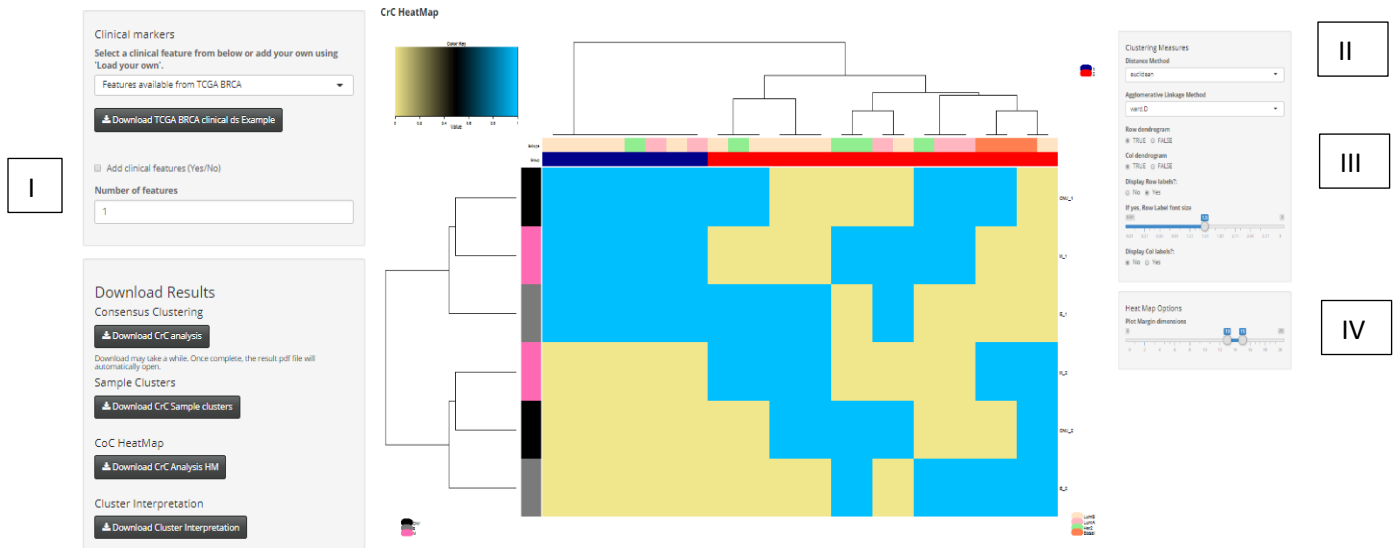
Step 1:

Select the computed number of optimal RNASeq Expression, Methylation and CNV data from the previous 3 tabs or upload your own clusters using the load my own option. The user-uploaded cluster data should be in the same format as the example data. Example data is available for download. In addition, the same patients should be input in the same order for each platform.

Step 2:

Select the platforms to base CrC analysis.
User should select at least two platforms.

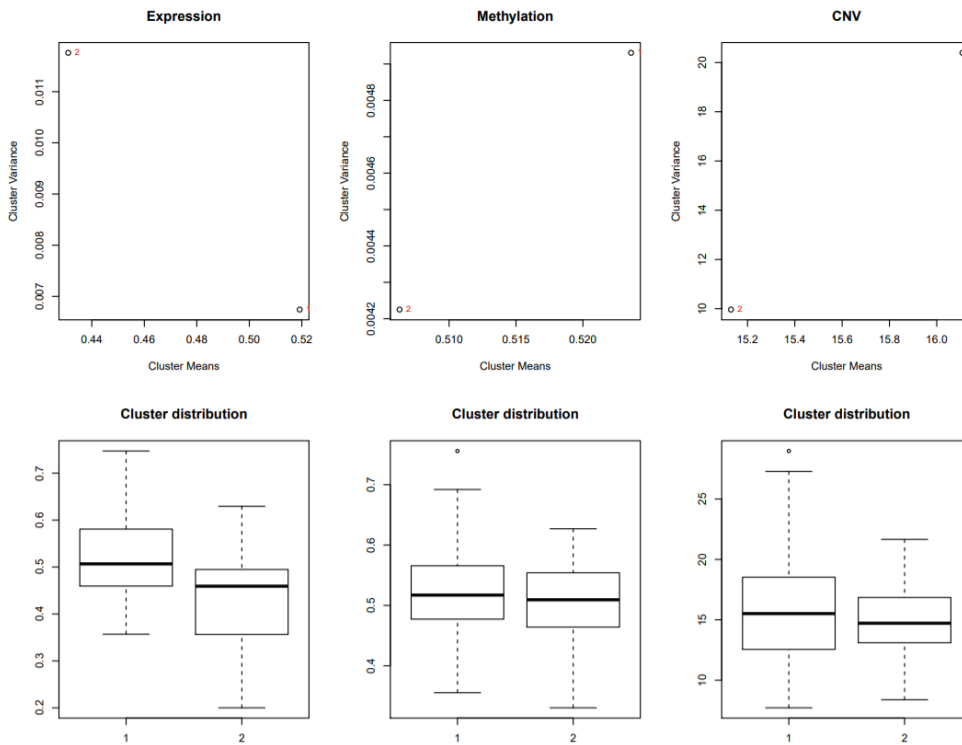
Select other parameter setting in the similar fashion to the previous tabs to run 'ConsensusClusterPlus' package.



Interactive HeatMap for Cluster of Cluster Analysis.

1-0 Transformed matrix data based on the individual platform clusters is used as input into the modified heatmap.2 function.

- I. Add Single or multiple clinical feature(s) as bars just below the dendrogram. As an example, sample risk status is displayed above the predicted consensus cluster bar (when the box I is checked) and can be downloaded using the download button.
- II. Choose clustering and distance measure
- III. Supervised row-wise or column wise clustering can be selected using FALSE option. Display Row and column labels using TRUE option. Adjust size of the labels using the slider.
- IV. Adjust Plot margins using the slider.



Interpretation of CoC Analysis Cluster HM based on the individual platform clusters

This option is available only when the actual expression, methylation or variant data is used. These plots will not be available if load my own option is used in the Cluster of cluster analysis tab.

Variance vs the mean plot of the lower triangular distance matrix serves as a relative measure of each cluster relative of the others within the same platform.

Boxplot of the individual clusters also helps determine which cluster has a relatively higher or lower median Expression (or median methylation or median CNV segment mean). The spread among the clusters is also informative.

Contingency Table(s)

Stratified by CNV

	CNV1M1	CNV1M2	CNV2M1	CNV2M2
E1	8	3	2	3
E2	3	2	2	2

Contingency Table(s) options:

Stratify by:

- ☐ Expression
- ☐ Methylation/Variant
- ☒ Copy Number

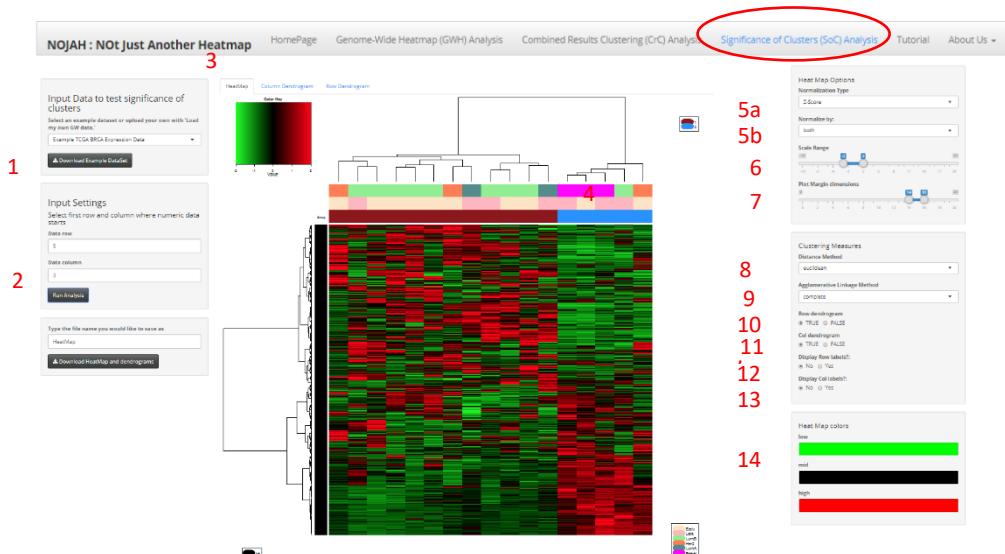
Contingency table displays the distribution of samples among the clusters. When all three platforms are used, the contingency table is stratified by the third platform i.e. CNV (by default, but can be changed using the right panel).

TAB C) SIGNIFICANCE OF CLUSTER ANALYSIS

NOJAH's significance of cluster analysis tab allows the user to perform heatmap clustering analysis on subset data. Like the CrC analysis tab, it is not suited to input genome-wide data. The exact column and row where the numerical data starts needs to be provided, for the tool to run correctly. For example, for data as in Table 1 provided above, numerical data points start at row four and column 3. If additional clinical rows are included, the row dropdown need to be increased to include these. Similarly, for columns. Depending on the data, the user selects normalization, scaling and clustering inputs. The same choices as in the GWH Analysis tab are available.

Although this functionality is already available in the GWH Analysis tab, the additional and the main feature of the SoC tab is that, it can help users test the significance of the gene set in obtaining the sample clusters. As this significance testing is based on a bootstrap approach which samples a gene set of the same size as the actual data from the genome-wide data typically 1000 times, it is computationally heavy. Thus, a separate tab for such analysis is available. The output from the SoC analysis is a significance test p-value of whether a random gene set of the same size could differentiate the samples into similar clusters ($p > 0.05$). A similar analysis can be performed for the gene clusters.

A step by step tutorial for implementing heatmap clustering analysis followed by SoC analysis is outlined below:



1: Select dataset of interest. Using the dropdown, you can choose the example most variable TCGA BRCA (from GWH tab)/CoMMpass Expression dataset or upload your own. If uploading your own, format data in same format as in the example file. Also input numeric data start: row and column. In example data, numeric data starts on row 5 and column 3. Depending on each dataset, this needs to be adjusted.

2: Download example data using download button to view contents/formatting of example file.

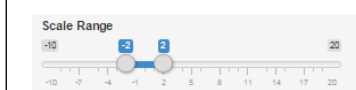
3: If example file is chosen, Heatmap automatically displayed in the HeatMap tab.

4: HeatMap created using Z-score 'both row and column' normalization, 'Euclidean' distance and 'complete' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

5a, b: Select a different normalization method you'd like for the data using drop down options. After choosing a different type, hit 'Run Analysis' button to update the heatmap.

This screenshot shows the 'Normalization Type' dropdown menu. The selected option is 'Z-Score'. Below it, the 'Normalize by:' dropdown menu is open, showing three options: 'row', 'col', and 'both'. The 'both' option is currently selected.

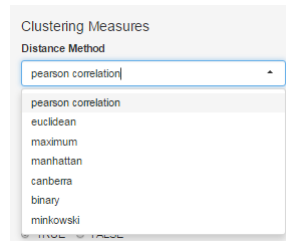
6 (optional): Drag slider to change scale range for the colors. Hit 'Run analysis' button to update Heatmap.



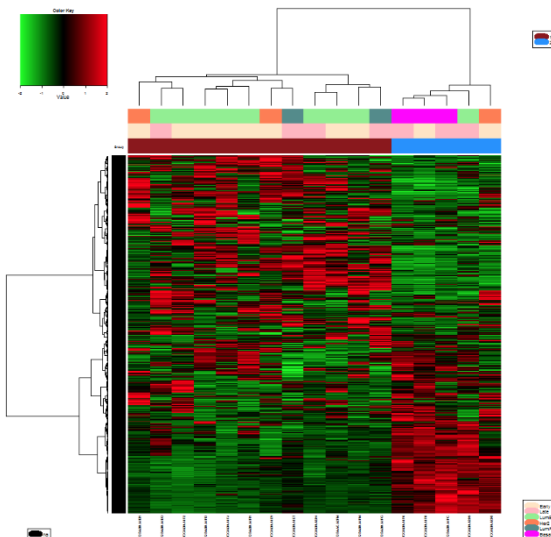
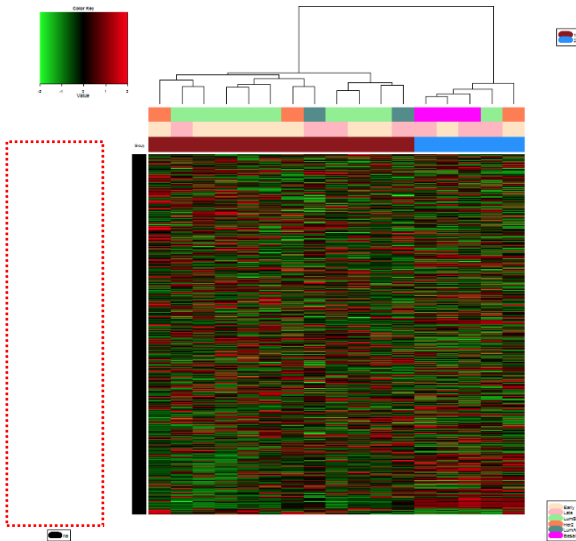
7: Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice versa until desired.



8, 9: Select Distance method and linkage method of choice using the drop-down options. Each selection will display modified heatmap.

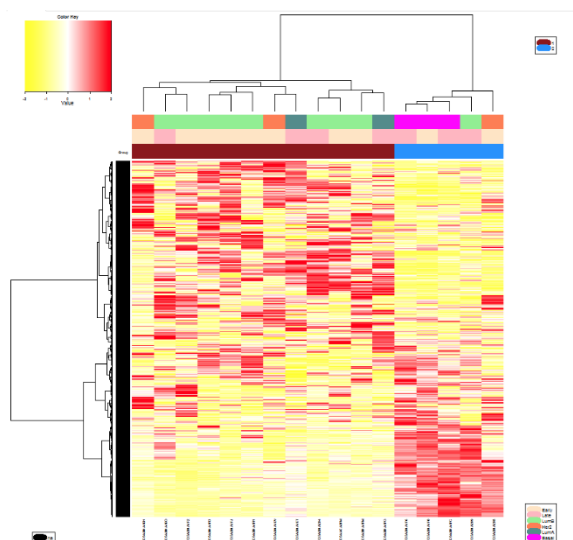


10, 11: Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.



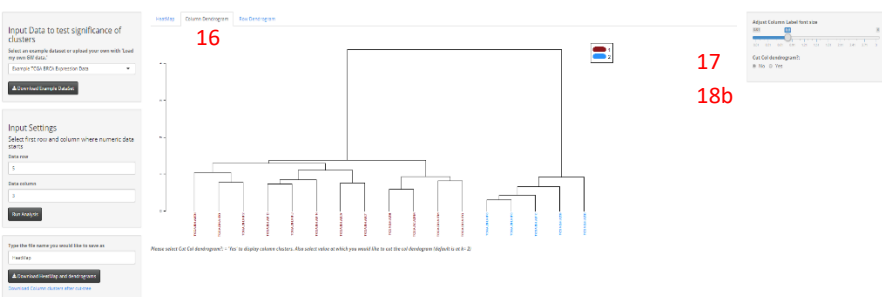
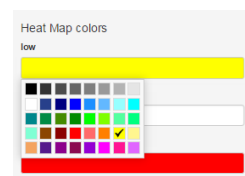
12, 13: Select Display Row labels = 'Yes' to see the corresponding genes. Additional slider appears to select, font size. Same applies to Sample labels.





14: Select color scheme. Red-Black-Green is typically used for Expression data and Blue-White-Red is used to represent methylation data. Heatmap will update as soon as color is chosen. After choosing desired color(s), click anywhere on screen to come out of color selection panel.

15: Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.



16: View in column dendrogram tab

17: Slider to adjust font size of the column dendrogram labels

18a Please select Cut Col dendrogram?: "Yes" to display column clusters. Also select value at which you would like to cut the col dendrogram (default is at k=2)

Show entries

Sample	Group	Cluster
1	TCGA-BH-A3M	1
2	TCGA-A4-A3M	1
3	TCGA-BH-A1M	1
4	TCGA-BH-A1G	1
5	TCGA-BH-A1G	1

Showing 1 to 5 of 13 entries

18 a, b: a. Option to cut the tree. b. If yes is chosen, user is asked at which position they want to cut the tree (default at 2)

When selected, a table will appear that classifies Samples, their Groups, and their corresponding clusters.

Use the drop down on upper left to display 5/10/All rows of the table.

19 Would you want to assess gene set significance in the separation of specimens into two clusters? (Yes/No)

19: Option to assess gene set significance in separation of the two clusters (Group1 vs Group2). Applicable only when >=2 clusters are available for analysis.

When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.

19a

19b

19c

19d

Select a dataset or upload your own with 'Load my own data.'

Example TCGA BRCA Exp Sampling Data

Sample size for bootstrap:

1000

No. of iterations for bootstrap:

1000

Go!

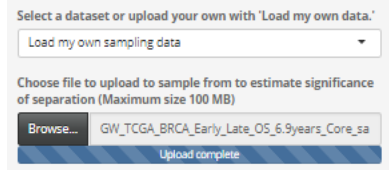
Click the button to start sampling using bootstrap method for estimating the p-value. A progress indicator will appear shortly (~approx 10 seconds), on top of page indicating the status. Once complete, the p-value will be displayed in the main panel.

Assess Gene set significance in separation of specimens into 2 clusters?:

☒ No ☐ Yes

When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.

19a: Select Sampling dataset for bootstrap. An example GW TCGA BRCA Sampling data is available or user can input their own (up to 75 MB is allowed). Large CSV and TXT files can be converted to RDS file contain file size within 75 MB limit.

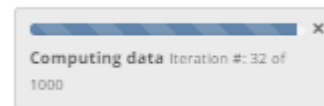


19b: Choose Sample size of the data for bootstrap. Use a size that does not exceed the original sampling data itself. For example, 1000.

19c: Select number of iterations you wish to perform. A good practice is to perform at least 1000 iterations for accuracy of analysis.

19d: Once all options are selected, press 'Go' button to start analysis.

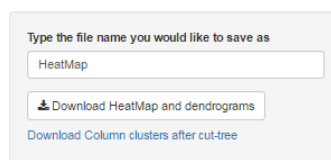
After approximately 10 seconds, a progress indicator will appear to track the time remaining for the analysis to be completed.



p-value results from the boot strap approach for calculation significance of clusters using Fisher's exact test will be displayed under the table along with the interpretation.

20: To download the p-value results as well, input the file names and click on Download button. The heatmap and the corresponding row and column dendrograms followed by the p-value results will be downloaded in pdf format.

To download the table for the classification of samples by clusters, click on link and the table will be saved as a CSV file.



Similar analysis can be performed on Row Dendrogram, provided you have at-least two row groups.

References:

1. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking*. Bioinformatics, 2010. **26**(12): p. 1572-3.