

Analyzing data using CASH

Clustering Analysis with Shiny HeatMap (CASH)

1 Select an example dataset or upload your own with 'Load my own data.'

2 Example Data File

3 Download Example Dataset

After data selection, to view generated Heatmap, click on Heatmap tab.

Heatmap Column Dendrogram Row Dendrogram

Data should be input as a .txt or .csv file. The first two rows of the data file have information about the patients/specimens and their response/outcome. 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', 'Gene_1_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 'Group' (column 2) followed by the patient IDs e.g. TCGA.01.142B, one per column, starting at column 3. Column 1 gene identifier has to be labelled 'gene_id' and column 2 header should be labelled 'Group' for using this tool. Other files may cause the program to display errors.
2. The second line of the file contains the patient response classification e.g. 'Favorable' for favorable 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.
b) Column_2: This should contain the gene group classification e.g. CNV for Over-expressed/Under-expressed or Hypermethylated/Hypomethylated in alphabetical order. If only one gene group, use any alphabet e.g. A 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.

NOTE: Clustering is based on scaled data, if the user chooses this option, prior to input into heatmap R function.

Example format for Data

gene_id	Group	TCGA.01.38GF	TCGA.08.045D	TCGA.02.02ZF	TCGA.01.127D	TCGA.02A0KD	TCGA.12.T37D	TCGA.1L.Y25S	TCGA.01.K0TD
1		Normal	Normal	Normal	Tumor	Tumor	Tumor	Tumor	
2	BRCA1	over	1.47	2.18	5.87	7.64	3.4	7.77	5.15
3	YRNA6	over	7.93	2.78	9.91	6.96	5.96	8.19	8.91
4	SPIN	under	8.02	9	2.17	1.12	3.76	0.02	3.67
5	BRAF	under	2.75	5.99	3.19	3.09	2	0.99	1.28

Table 1: Example dataset for two gene groups (over and under-expressed) and two patient groups (Normal, Tumor).

gene_id	Group	GSM9991	GSM9992	GSM9993	GSM9994	GSM9995	GSM9996	GSM9997	GSM9998	GSM9999	GSM10000	GSM10001	GSM10002	GSM10003	GSM10004	GSM10005	GSM10006	GSM10007	GSM10008	GSM10009	GSM10010
1		MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM
2	YRNA6	over	1.47	2.18	5.87	9.12	7.34	1.56	3	7.77	3.4	1.56									
3	YRNA6	over	7.93	2.78	9.91	6.96	5.96	8.19	8.91	5.96											
4	SPIN	under	8.02	9	2.17	1.12	3.76	0.02	3.67	9.76											
5	BRAF	under	2.75	5.99	3.19	11.86	6.54	8.17	2	0.99	2	1.17									
6	YRNA6	over	7.93	2.78	9.91	6.96	5.96	8.19	8.91	5.96											

Table 2: Example dataset for one gene group (BRCA1) and four patient groups (MM, MM, MM, MM).

Both row and column dendrograms can be extracted in this specific tabs. Using the options on the right panel, dendrograms can be cut into desired no. of clusters (default at 2) and a p-value of significance between the clusters can be determined using bootstrap method.

1: Select dataset of interest. Using the dropdown, you can choose the example or upload your own. If uploading your own, follow instructions on right of how to format data before inputting.

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

Example Data File

Load my own data

After data selection, to view generated heatmap, click on Heatmap tab.

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

Load my own data

Choose file to upload (maximum size 10 MB)

Choose File No file chosen

If loading your own data, click on Choose File and browse to the location of input file and click open. Upload progress bar will initiate.

Choose file to upload (maximum size 10 MB)

Choose File BRCA.Example.data.csv

Upload complete

Wait for processing.

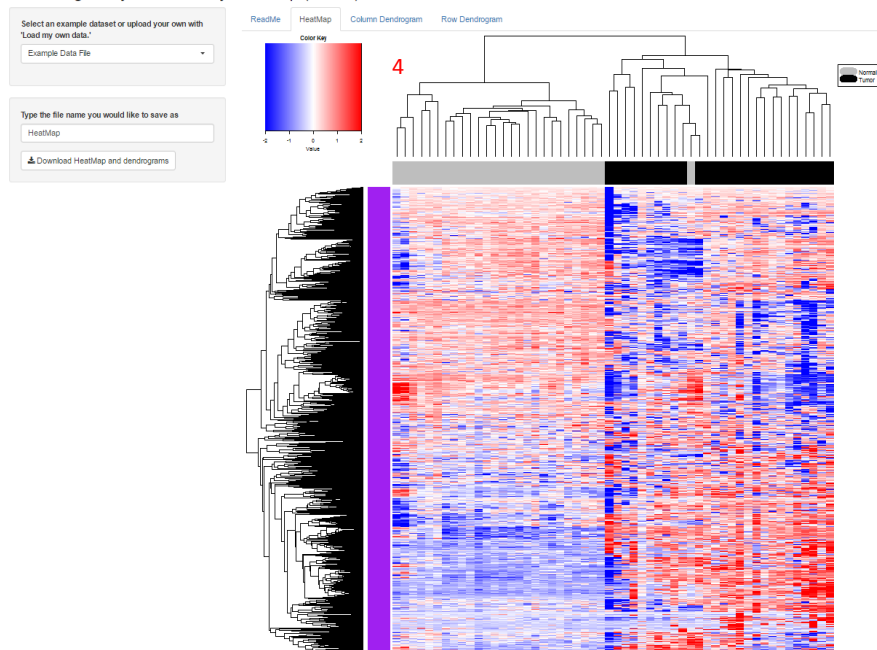
2: If using the example dataset (you will need to go straight to 3), but if you wish to download it, the 'Download Example Dataset' button will download the dataset in .csv format as seen below when using Chrome browser.

Example data set_TC....csv

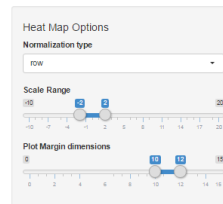
3: Click on the HeatMap tab in order to view it.

Clustering Analysis with Shiny HeatMap (CASH)

15



5



6

7

8

9

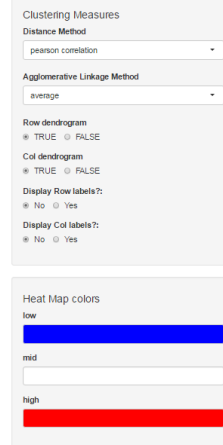
10

11

12

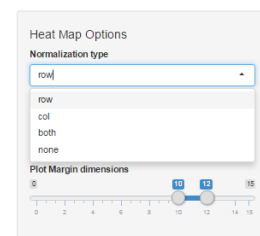
13

14



4: HeatMap created using 'row' normalization, 'Pearson correlation' distance and 'average' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

5: Select a different normalization/scaling you'd like for the data using drop down options. Each time a different type is chosen, the heatmap will be updated.



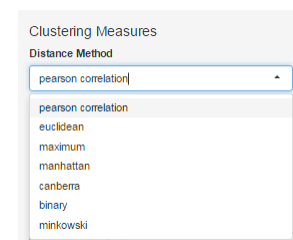
6 (optional): Drag slider to change scale range for the colors. Heatmap will be updated on movement.

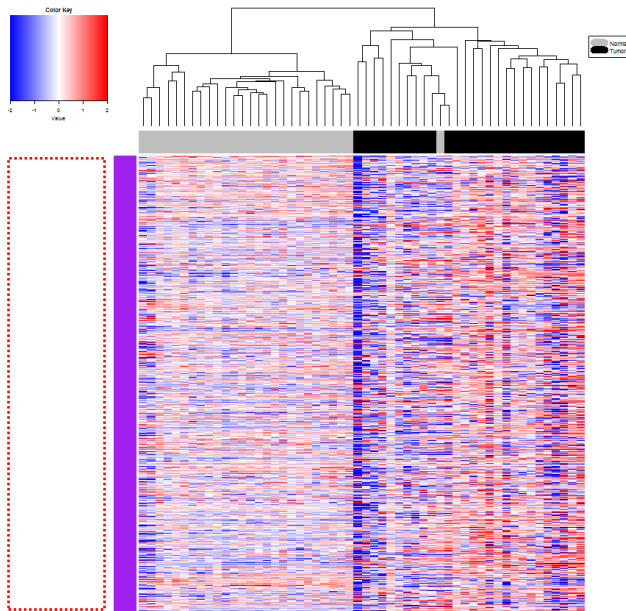


7: Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice versa until desired.



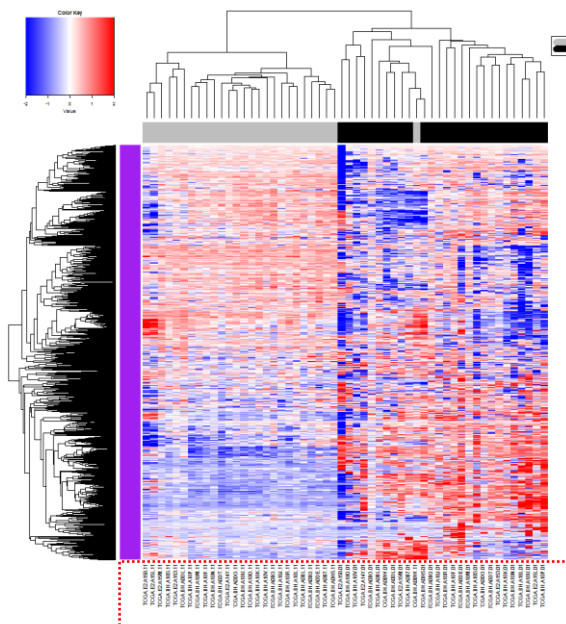
8, 9: Select Distance method and linkage method of choice using the drop down options. Each selection will display modified heatmap.





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.

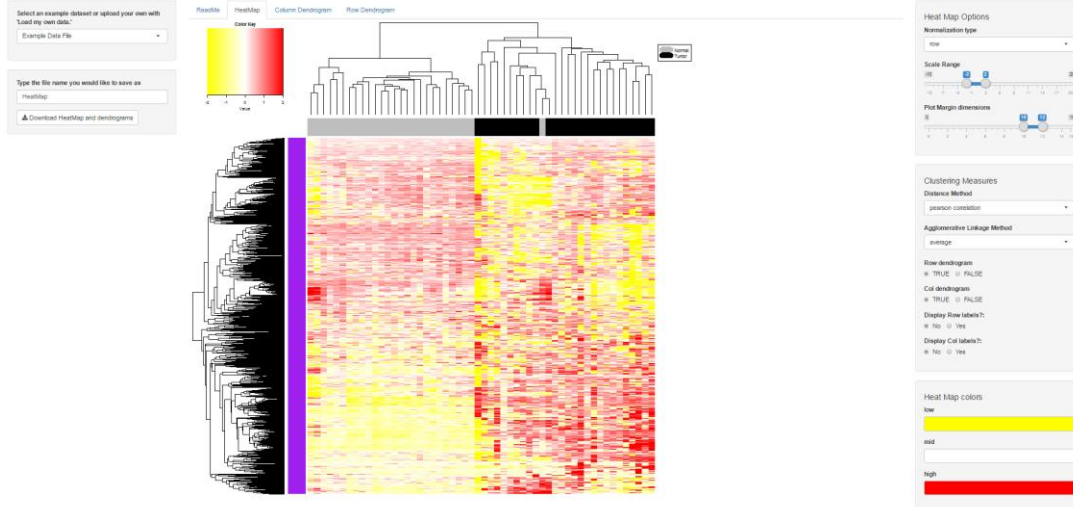
Row dendrogram
☐ TRUE ☒ FALSE



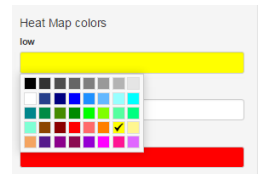
12, 13: Select Display Row labels = 'Yes' to see the corresponding CpG sites. Additional slider appears to select, font size. Same applies to Sample labels.

Display Row labels?:
☒ No ☐ Yes
 Display Col labels?:
☐ No ☒ Yes
 If yes, Col Label font size:
 0.01 0.31 1.21 1.31 1.81 2.11 2.41 2.71 3

Clustering Analysis with Shiny HeatMap (CASH)



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.

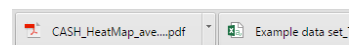


15: Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.

Type the file name you would like to save as

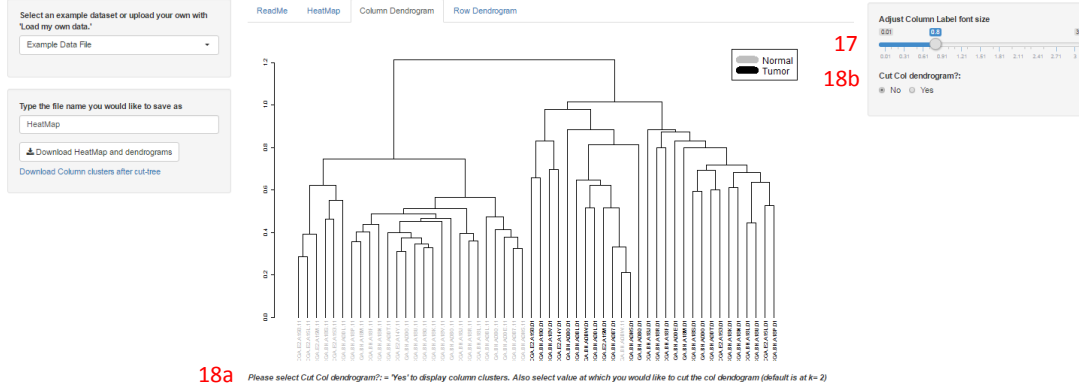
HeatMap

Download HeatMap and dendrograms



Clustering Analysis with Shiny HeatMap (CASH) 16

20



16: View in column dendrogram tab

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

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)

Show ☒ 5 ☐ 10 ☐ All

Sample	Group	Cluster
1 TCGA.E2.A158.11	Normal	1
2 TCGA.E2.A15L.11	Normal	1
3 TCGA.E2.A15M.11	Normal	1
4 TCGA.BH.A185.11	Normal	1
5 TCGA.E2.A153.11	Normal	1

Showing 1 to 5 of 54 entries

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

Use the drop down on upper left ☒ 5 ☐ 10 ☐ All, 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)

19a

19b

19c

19d

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

☒ No ☐ Yes

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

Meth 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 s), 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 Methylation Sampling data is available or user can input their own (up to 10 MB is allowed).

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

Load my own sampling data

Choose file to upload to sample from to estimate significance of separation

Choose File BRCA_met_M...derate.csv

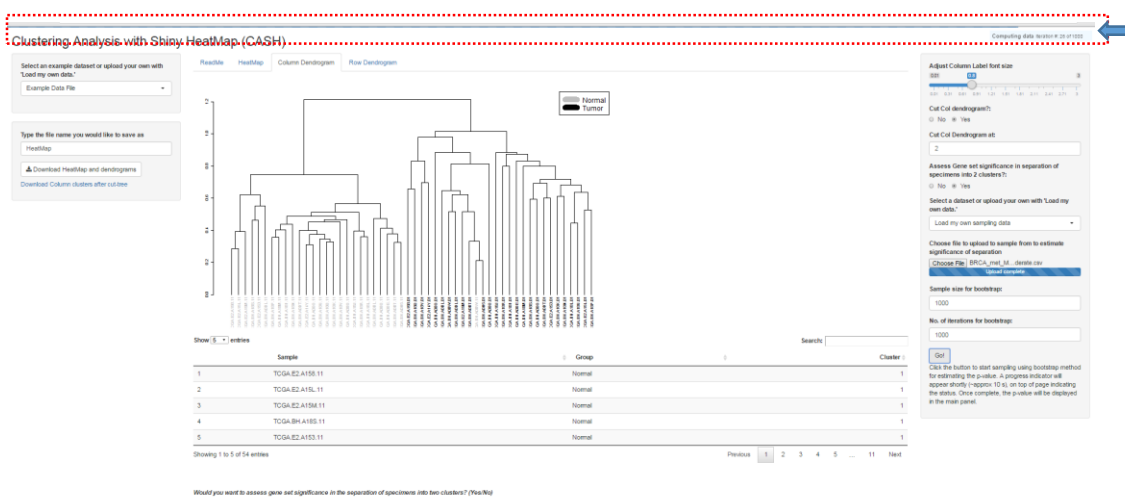
Upload complete

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

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 10s, a progress indicator and counter (top of page) will appear to track the time remaining for the analysis to get completed.



The p-value to test the gene set significance in the separation of specimens into 2 clusters is = 0

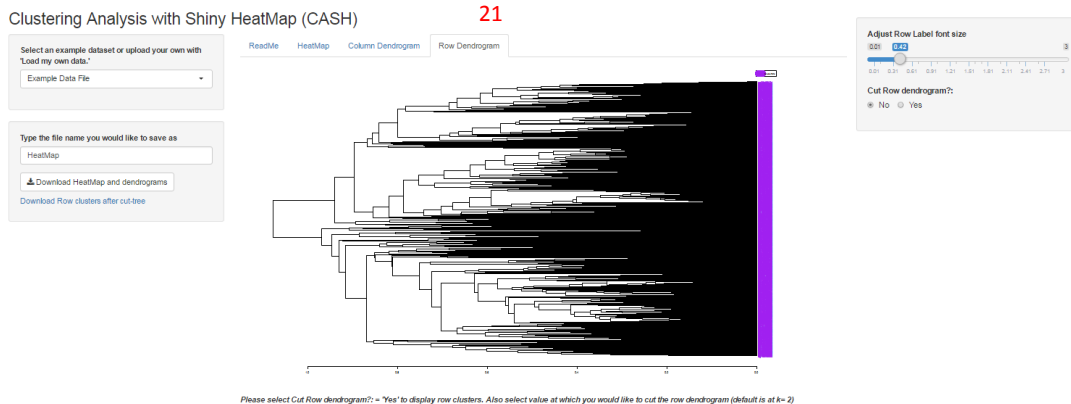
The gene set cluster is statistically significant, i.e., a random sample of CpG probes/gene sets of the same number is Not able to separate the specimens when compared to the CpG probes/gene sets of interest of the same class

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 name 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.

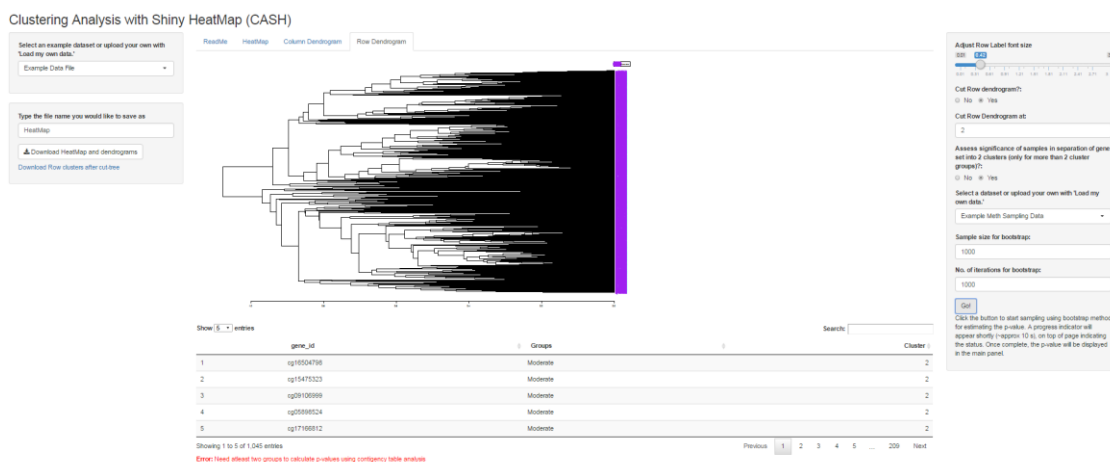
To download the table for the classification of samples by clusters, click on link and the table will be saved as an .csv file.

The screenshot shows the download options section. There is a text input field labeled 'Type the file name you would like to save as' with the value 'HeatMap'. Below it are two buttons: 'Download HeatMap and dendrograms' and 'Download Column clusters after cut-tree'. A blue arrow points to the second button.



21: When in Row Dendrogram tab, follow the same steps as when in the column dendrogram tab.

If p-value is calculated here (not applicable here because of a single group), the results will automatically be incorporated in the downloaded pdf file.



If you attempt to calculate p-value where there is just a single group of CpG sites, an error will appear like shown but since analysis was not performed, the result table will not be included for download.

Error: Need atleast two groups to calculate p-values using contingency table analysis

Output using the example dataset is available in a separate file titled 'CASH_HeatMap_2016-07-27 13-28-11.pdf'.