

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 HeatMap Custom Descriptors Row Descriptors

Data should be input as a .csv or .xlsx or .xls file. The first two rows of the data file have information about the patient/assay and then expression data. All remaining rows have gene expression data, one row per gene. In the case of clustering 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_L1_1', 'Gene_L1_2' including Gene 1 has two probes. The columns represent the different experimental samples. A maximum of 4000 different sample groups and 4 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 ID (e.g. TCGA-01-1428, see 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. 'Pathologic 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.
b) Column 2: This should contain the gene group classification (e.g. C10 for Over-expressed/Under-expressed or pathologic favorable/unfavorable) 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 values 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-01-1428	TCGA-01-1429	TCGA-01-1430	TCGA-01-1431	TCGA-01-1432	TCGA-01-1433	TCGA-01-1434		
1	Normal	Normal	Normal	Tumor	Tumor	Tumor	Tumor	Tumor		
2	BRCA1	over	1.47	2.18	5.57	7.54	3.4	7.77	5.15	1.55
3	TP53	over	7.52	2.78	9.11	6.96	5.98	8.19	9.91	3.95
4	TP53	under	8.52	9	2.97	1.42	1.78	0.02	3.67	9.76
5	BRCA1	under	2.75	5.95	3.19	3.55	2	0.99	1.25	5.17

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

gene_id	Group	GSM99591	GSM99592	GSM99593	GSM99594	GSM99595	GSM99596	GSM99597	GSM99598	GSM99599	GSM99600
1	BRCA1	over	1.47	2.18	5.57	7.54	3.4	7.77	5.15	1.55	1.55
2	TP53	over	7.52	2.78	9.11	6.96	5.98	8.19	9.91	3.95	3.95
3	TP53	under	8.52	9	2.97	1.42	1.78	0.02	3.67	9.76	9.76
4	BRCA1	under	2.75	5.95	3.19	3.55	2	0.99	1.25	5.17	5.17

Table 2: Example dataset for one gene group (BRCA1) and four patient groups (BRCA1, BRCA2, TP53, TP53).

Both row and column descriptors can be extracted in the specific tabs. Using the options on the right panel, descriptors can be outputted as a .csv or .xlsx (default) or .pdf and a profile 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

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.

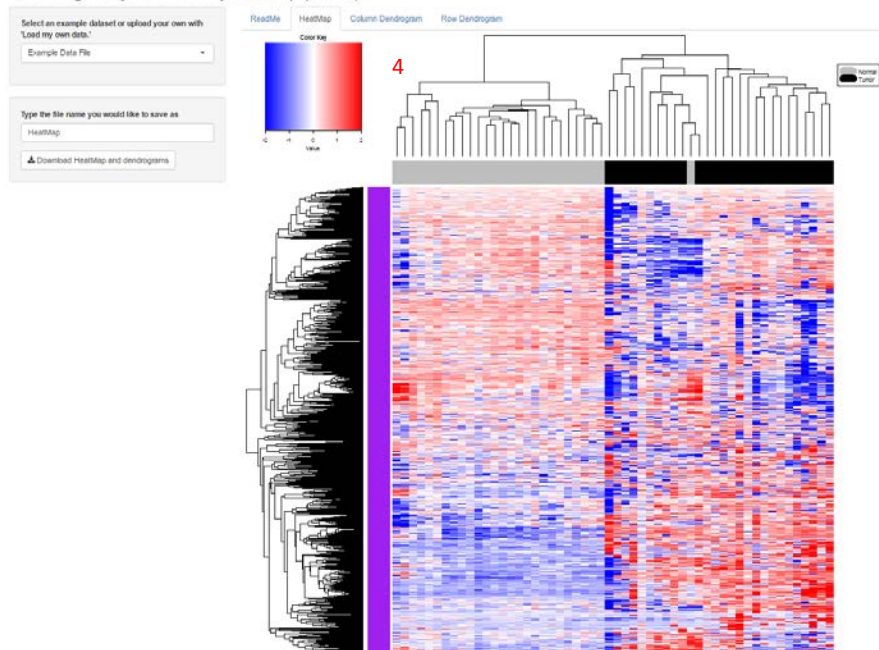
2: If using the example dataset (you will need to go straight to 3), but if you wish to download it, the 'Download Example Dataset' button will download the dataset in .csv format as seen below when using Chrome browser.

Example data set_TC....csv

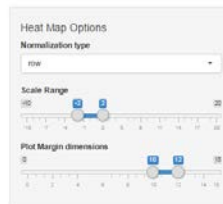
3: Click on the HeatMap tab in order to view it.

Clustering Analysis with Shiny HeatMap (CASH)

15



5



6

7

8

9

10

11

12

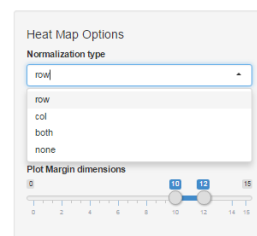
13

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.



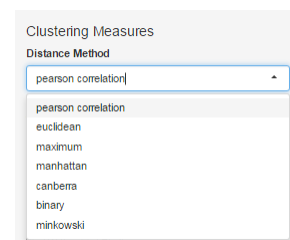
6 (optional): Drag slider to change scale range for the colors. Heatmap will be updated on movement.

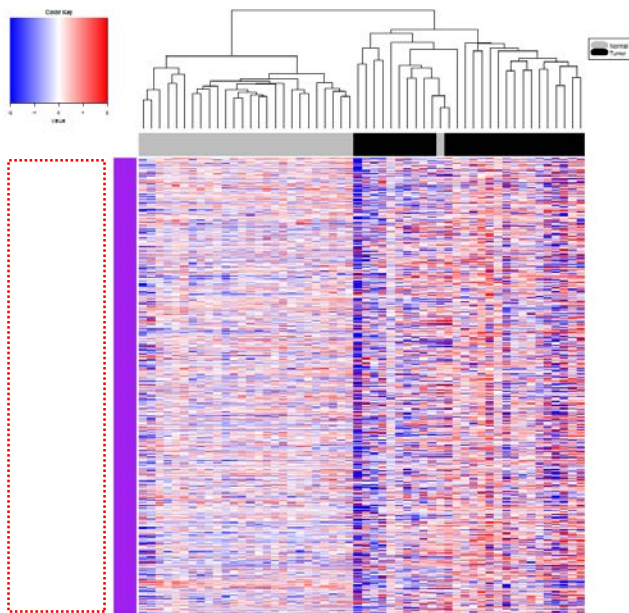


7: Select the Plot margins. If column dendrogram overlaps the legend, increase both margin points and vice versa until desired.



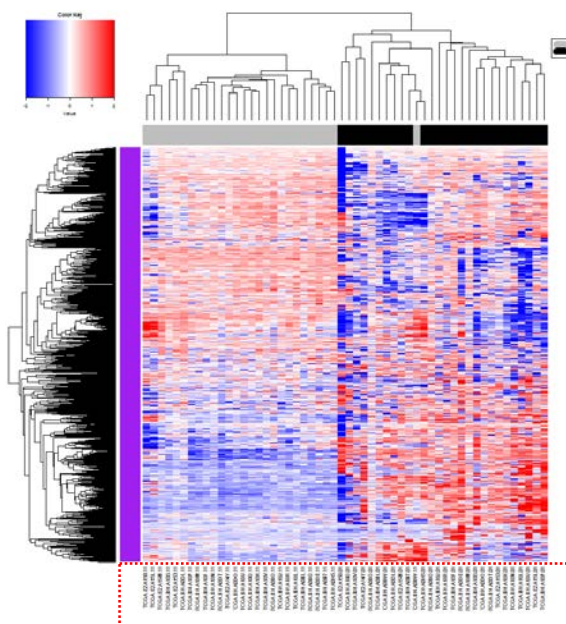
8, 9: Select Distance method and linkage method of choice using the drop down options. Each selection will display modified heatmap.





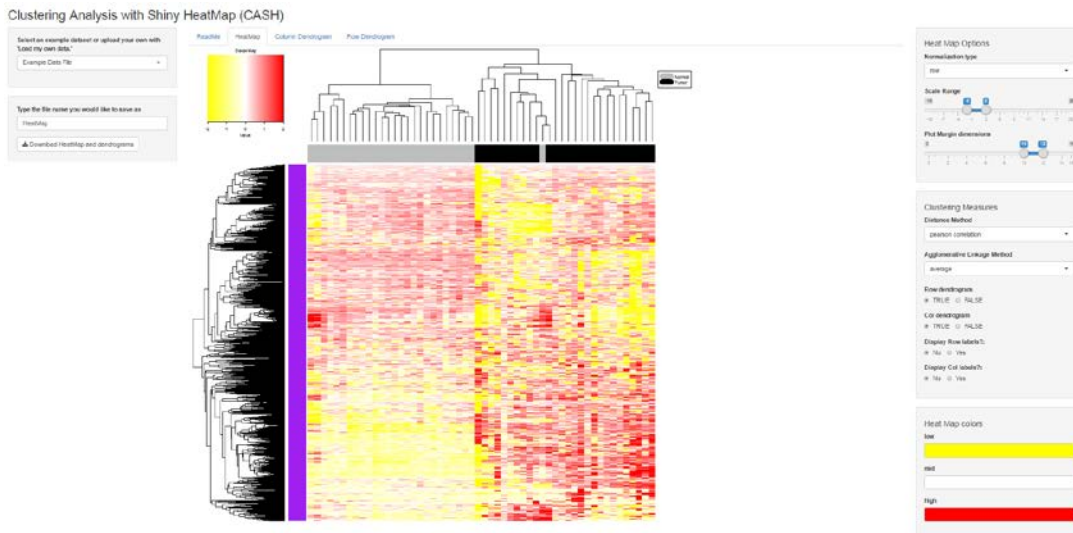
10, 11: Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.

Row dendrogram
☐ TRUE ☒ FALSE

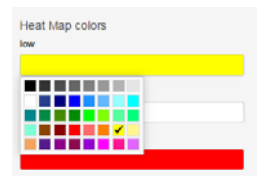


12, 13: Select Display Row labels = 'Yes' to see the corresponding CpG sites. Additional slider appears to select, font size. Same applies to Sample labels.

Display Row labels?:
☒ No ☐ Yes
 Display Col labels?:
☐ No ☒ Yes
 If yes, Col Label font size:



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.

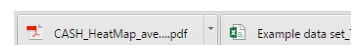


15: Input file name and click on Download button to save heatmap and the corresponding row and column dendrograms in pdf format as shown below using Chrome browser.

Type the file name you would like to save as

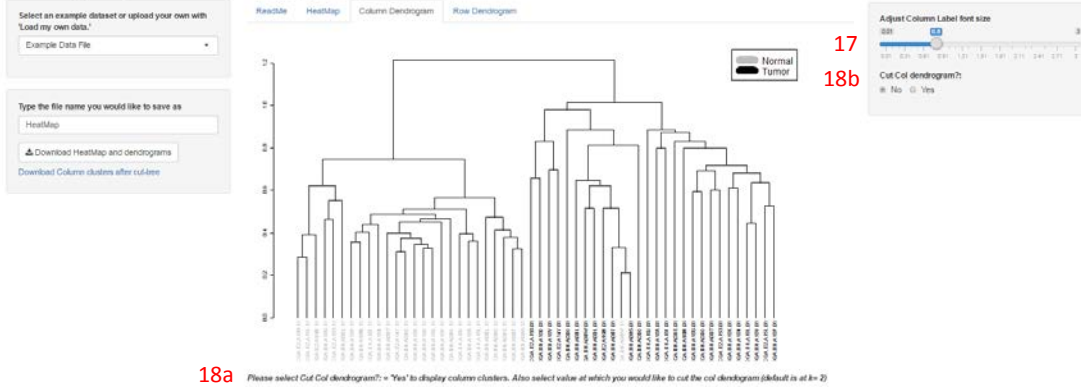
HeatMap

Download HeatMap and dendrograms



Clustering Analysis with Shiny HeatMap (CASH) 16

20



Show 5 entries

Sample	Group	Cluster
1	TCGA.E2.A15L.11	Normal
2	TCGA.E2.A15L.11	Normal
3	TCGA.E2.A15M.11	Normal
4	TCGA.BM.A18S.11	Normal
5	TCGA.E2.A15S.11	Normal

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, to display 5/10/All rows of the table.

19: Option to assess gene set significance in separation of the two clusters (Tumor vs Normal). Applicable only when ≥ 2 clusters are available for analysis.

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 75 MB is allowed). Large .csv and .txt files can be converted to .RDS file contain file size within 75 MB limit.

```
> # Rcode to convert .csv file to RDS
> samplingdata <- read.csv("file_path/file_name.csv", header = T,
stringsAsFactors = F, sep = ",")
> saveRDS(samplingdata, "file_path/file_name.rds")
```

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

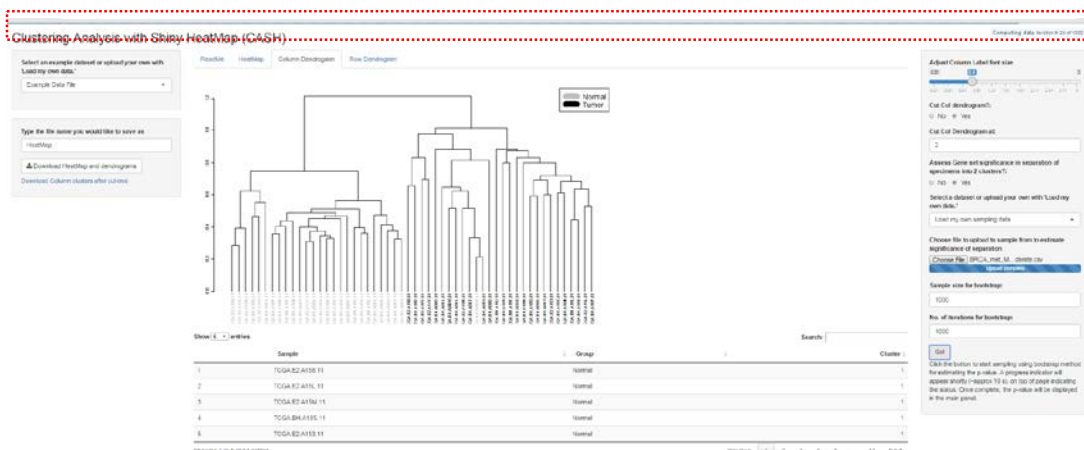
Upload complete

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 10 seconds, a progress indicator and counter (top of



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.

Type the file name you would like to save as

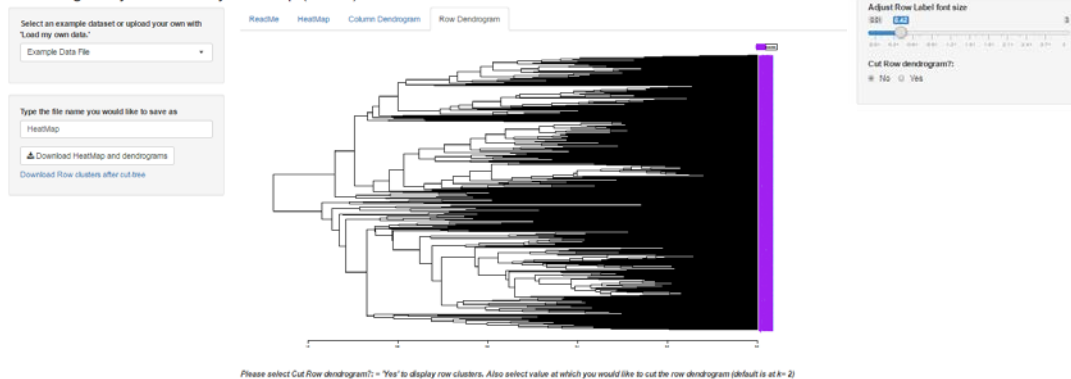
HeatMap

Download HeatMap and dendrograms

Download Column clusters after cut-tree

Clustering Analysis with Shiny HeatMap (CASH)

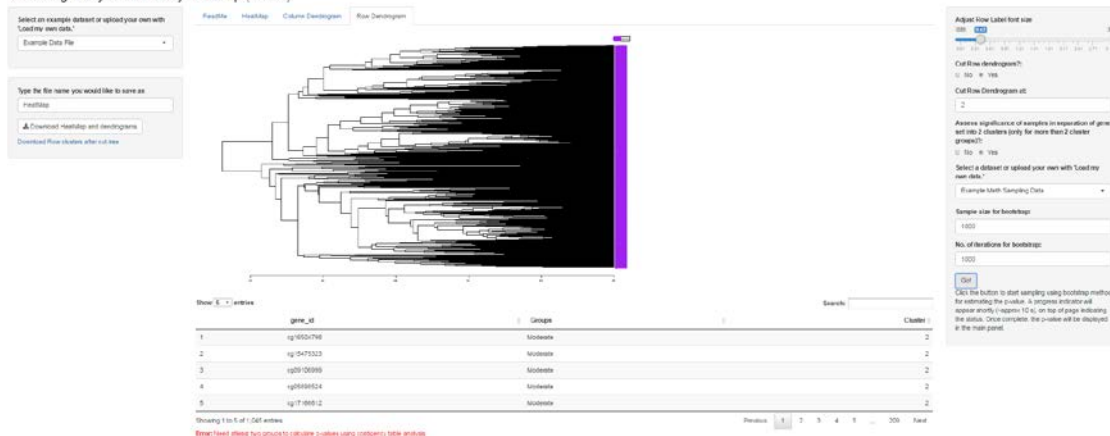
21



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.

Clustering Analysis with Shiny HeatMap (CASH)



If you attempt to calculate p-value where there is just a single group of CpG sites, an error will appear like shown. Since analysis was not performed here, 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'.