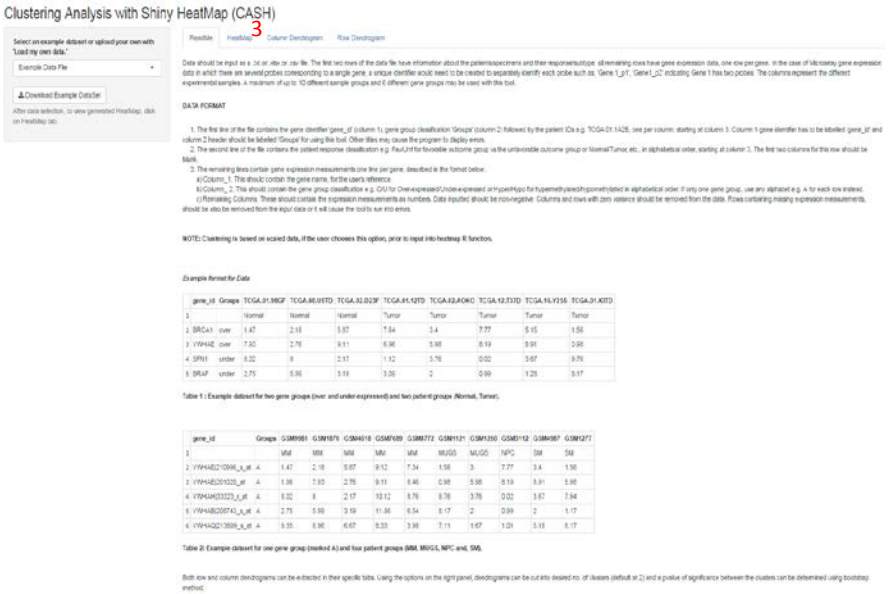
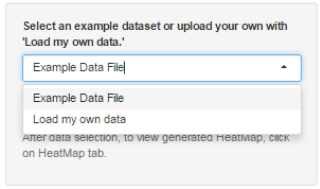
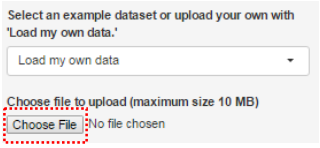
Analyzing data using CASH

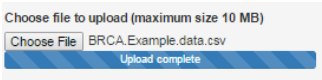


1: Select dataset of interest. Using the dropdown, you can choose the example or upload your own. If uploading your own, follow instructions on right of how to format data before inputting.



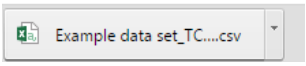


If loading your own data, click on Choose File and browse to the location of input file and click open. Upload progress bar will initiate.

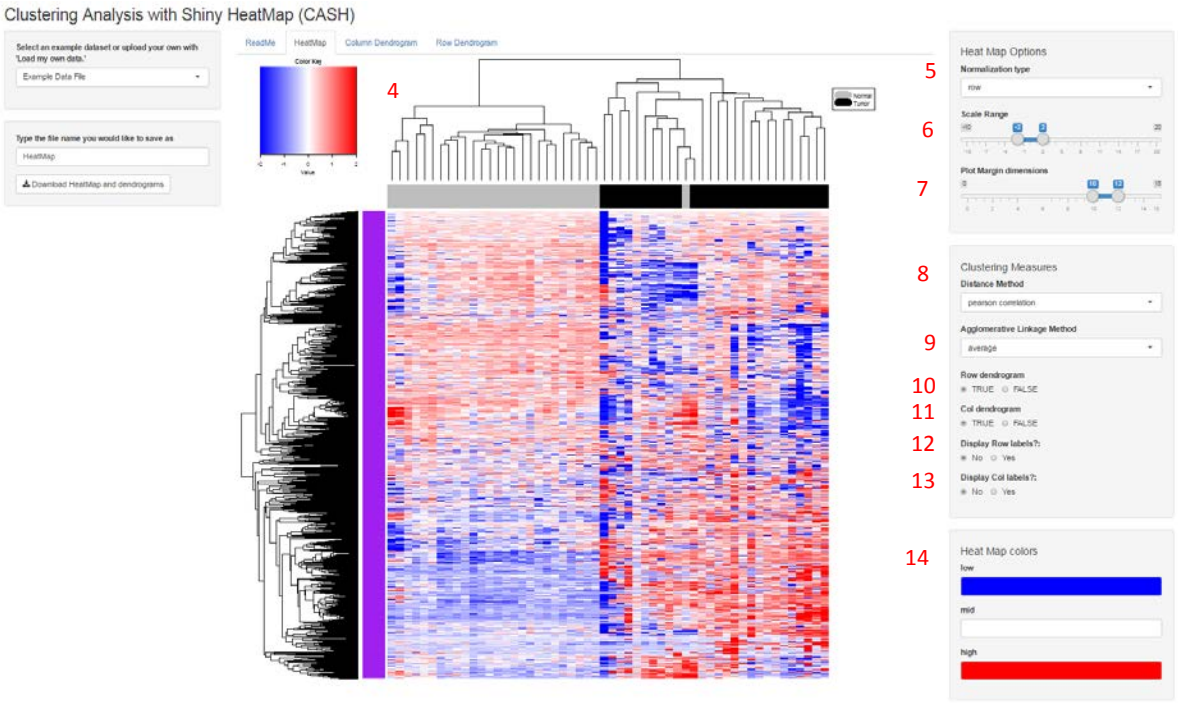


Wait for processing.

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.

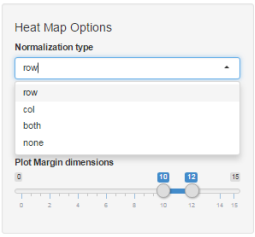


3: Click on the HeatMap tab in order to view it.



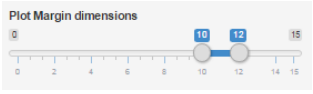
4: HeatMap created using 'row' normalization, 'Pearson correlation' distance and 'average' agglomerative linkage method (i.e. default settings). Depending on dataset may take several minutes to load.

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.

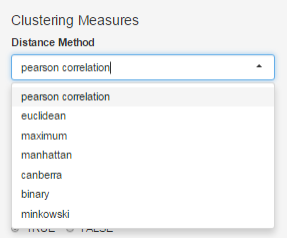


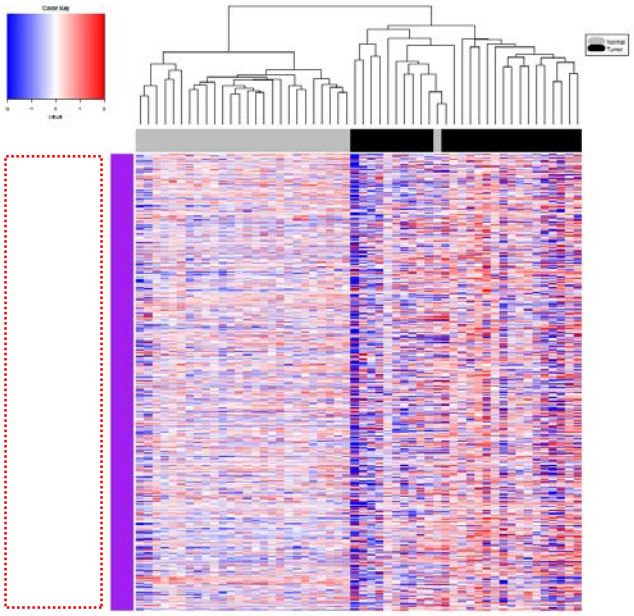
6 (optional): Drag slider to change scale range for the colors. Heatmap will be updated on movement.

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

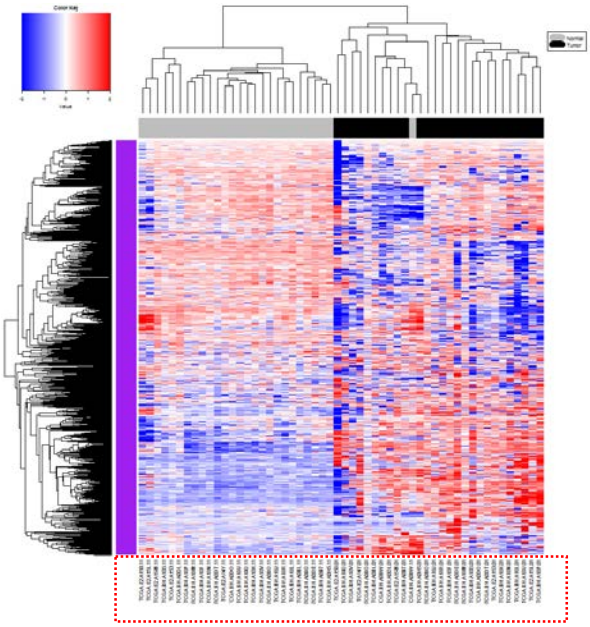


8, 9: Select Distance method and linkage method of choice using the drop down options. Each selection will display modified heatmap.

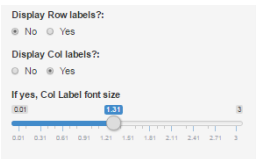


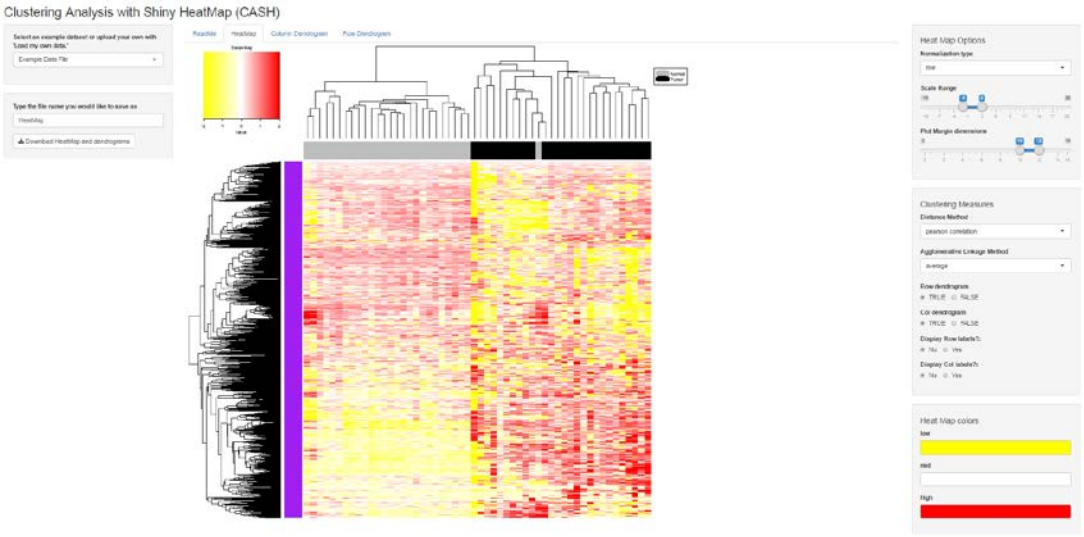


10, 11: Select either to display Row dendrogram or not. If FALSE is chosen, row dendrogram will disappear and data will not be ordered based on means. Same applies to Column dendrogram.

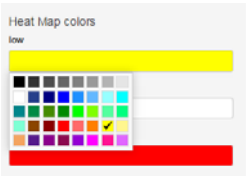


12, 13: Select Display Row labels = 'Yes' to see the corresponding CpG sites. Additional slider appears to select, font size. Same applies to Sample labels.

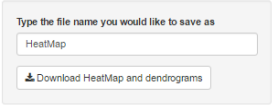


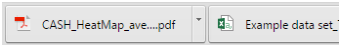


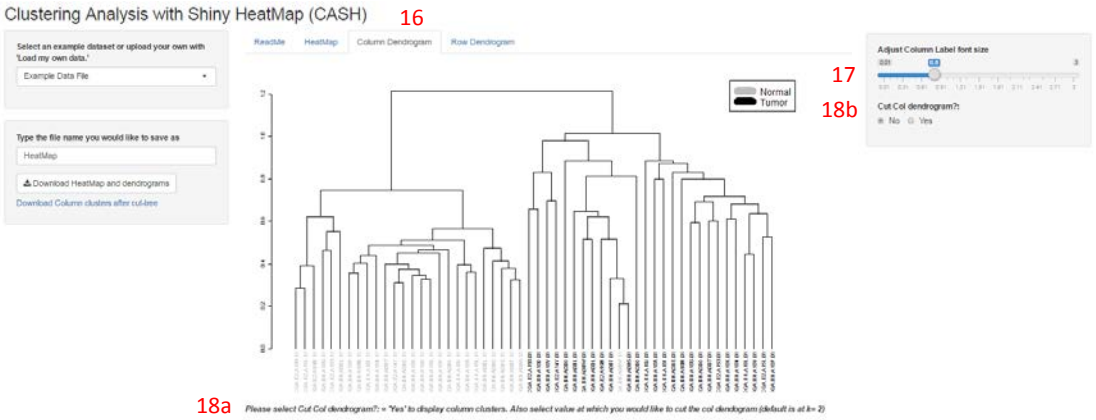
14: Select color scheme. Red-Black-Green is typically used for Expression data and Blue-White-Red is used to represent methylation data. Heatmap will update as soon as color is chosen. After choosing desired color(s), click anywhere on screen to come out of color selection panel.



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.



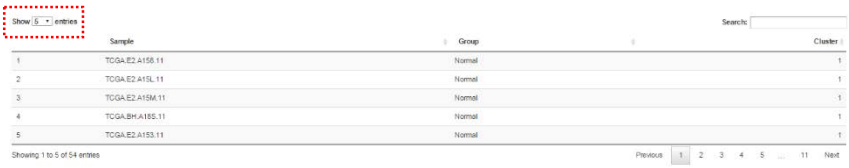




16: View in column dendrogram tab

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

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)

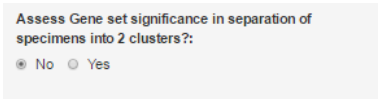


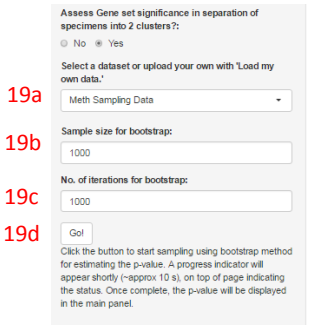
When selected, a table will appear that classifies Samples, their Groups, and their corresponding clusters.

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

19: 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)*

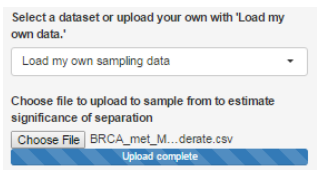




When 'Yes' is selected, parameters for Monte Carlo p-value estimation will be made available.

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")
```

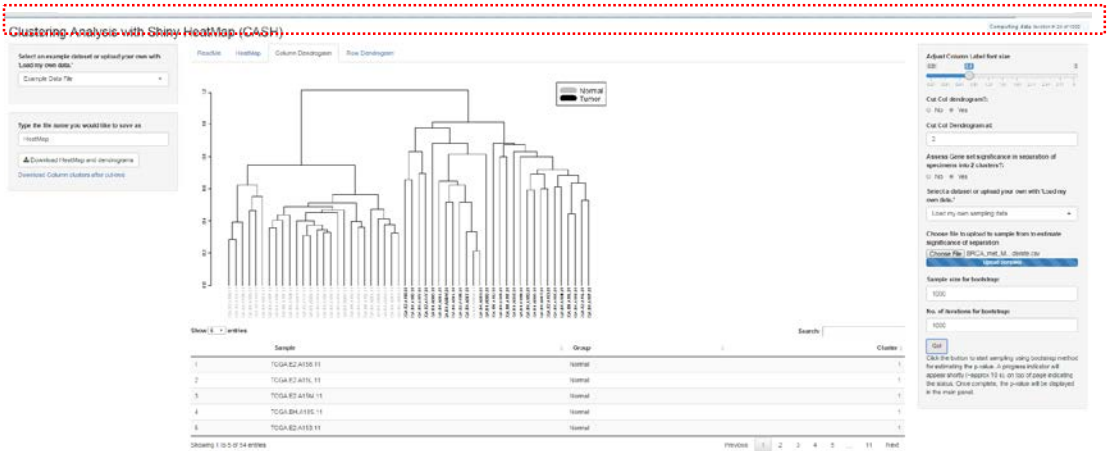


19b: Choose Sample size of the data for bootstrap. Use a size that does not exceed the original sampling data itself.

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 page) will appear to track

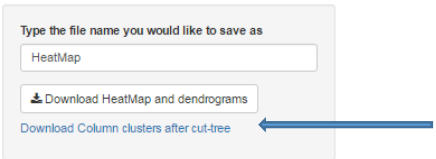


The p-value to test the gene set significance in the separation of specimens into 2 clusters is = 0

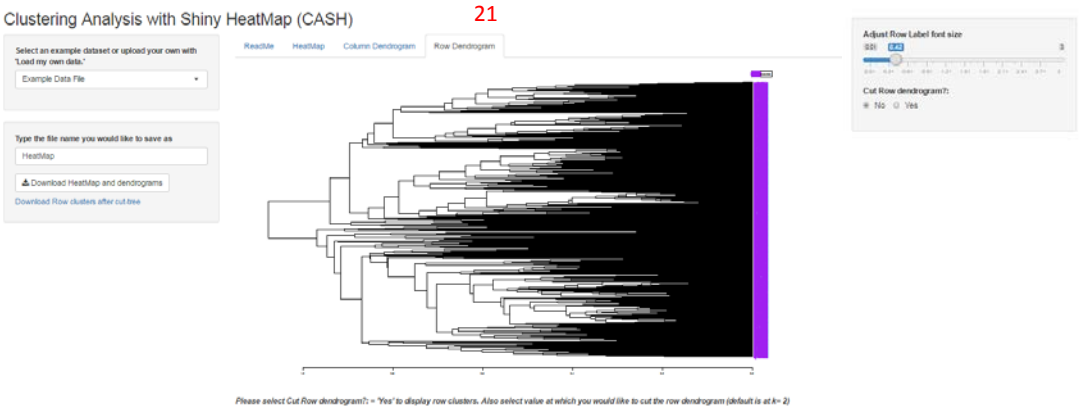
The gene set cluster is statistically significant, i.e., a random sample of CpG probes/gene sets of the same number is Not able to separate the specimens when compared to the CpG probes/gene sets of interest of the same class

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.

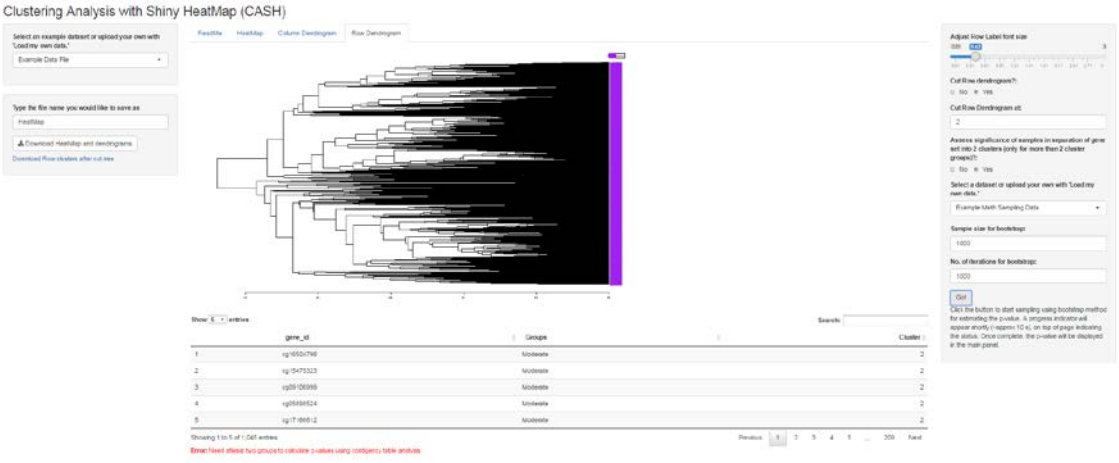


To download the table for the classification of samples by clusters, click on link and the table will be saved as an .csv file.



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. Since analysis was not performed here, the result table will not be included for download.

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