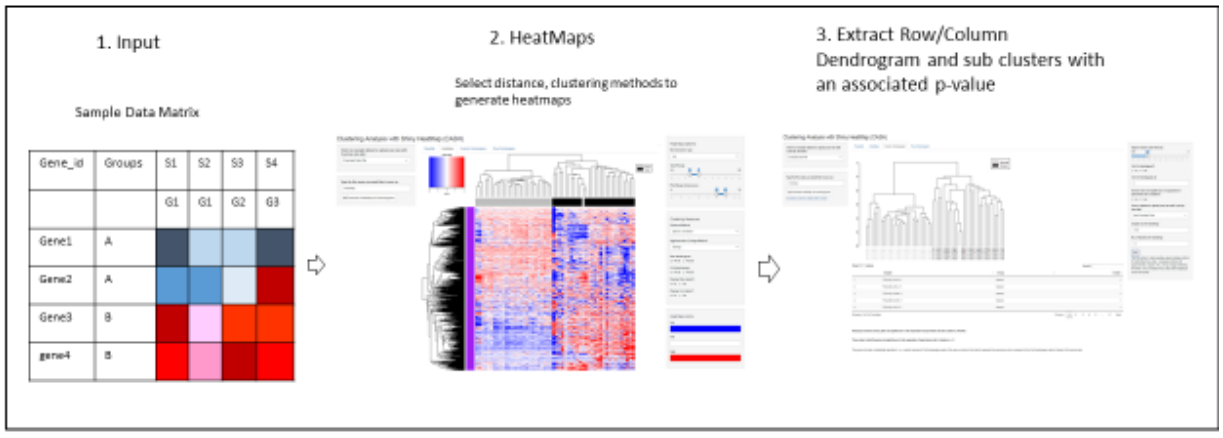
Clustering Analysis with Shiny HeatMap (CASH)

Manali Rupji, Bhakti Dwivedi and Jeanne Kowalski

Winship Cancer Institute of Emory University, Atlanta, GA 30322

Introduction

CASH is a tool designed to perform clustering analysis with just minimal knowledge and coding experience. This point and click tool allows the user to enter genomic data of their choice (expression or methylation) to obtain a beautiful HeatMap displaying the corresponding clusters. The user can choose from the various different distance and clustering methods available. Clustering is based on scaling the data before input into the heatmap function in this tool. A limitation to this tool is, up to ten different sample groups and in addition up to **six** different gene groups can be used with this tool. A minimum of one sample group and one row group needs to be supplied. The column and row dendograms can also be extracted separately. Furthermore, these can be cut to extract the corresponding subgroup members. A novel feature of this tool is the ability to test the significance between the clusters based on a bootstrap approach with an associated p-value. Figure 1 shows a schematic representation of the HeatMap tool using a TCGA BRCA example dataset.



TCGA BRCA Methylation Example dataset is available for download using the download example dataset button on the left hand side sidebar.

Data Requirements

The only requirement for this tool is the data input should be in a particular format. Data should be input as a .txt or .xlsx or .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.

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 has to 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*.

* The second line of the file contains the patient response classification e.g. Unf/Fav for Unfavorable outcome group vs the favorable outcome group or Tumor/Normal, etc., starting at column 3. The first two columns for this row should be missing.

* The remaining lines contain gene expression measurements one line per gene, described in the format below.

o **Column 1**: This should contain the gene name, for the user's reference.

o **Column 2**: This should contain the gene group classification e.g. O/U for Overexpressed and Under-expressed or Hyper/Hypo for hyper-regulated vs hypo-regulated cpG islands. If only one gene group, use any alphabet e.g. A for each row instead.

o **Remaining Columns**: These should contain the expression measurements as numbers. Missing expression measurements should be noted blank. Data inputted should be non-negative. Columns and rows with zero variance should be removed from the data.

Example format for Data:

gene_id	**Groups**	TCGA.01.12TD	TCGA.02.A0KO	TCGA.12.T37D	TCGA.16.Y2S5	TCGA.01.KITD	TCGA.01.98GF	TCGA.08.U5TD	TCGA.02.D23F
		Tumor	Tumor	Tumor	Tumor	Tumor	Normal	Normal	Normal
Gene1	Over	7.64	3.40	7.77	5.15	1.56	1.47	2.18	5.87
Gene2	Over	6.96	5.98	8.19	8.91	0.98	7.93	2.76	9.11
Gene3	Under	1.12	3.76	0.02	3.67	9.76	8.02	8.00	2.17
Gene4	Under	3.09	2.00	0.99	1.28	8.17	2.75	5.99	3.19

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

gene_id	**Groups**	GSM1121	GSM1250	GSM3112	GSM4987	GSM1277	GSM9981	GSM1870	GSM4618	GSM7689
		MUGS	MUGS	NPC	SM	SM	MM	MM	MM	MM
Gene1	A	1.56	3.0	7.77	3.40	1.56	1.47	2.18	5.87	9.12
Gene2	A	0.98	5.98	8.19	5.98	1.98	7.93	2.76	9.11	8.46
Gene3	A	9.76	3.76	0.02	3.76	7.94	8.02	8.00	2.17	10.12
Gene4	A	8.17	2.00	0.99	2.00	1.17	2.75	5.99	3.19	11.86

Table 2: Example dataset for 1 gene group (marked A) and 4 patient groups (MUGS, NPC, SM and MM)

Using HeatMap tool

This tool will generate a heatmap in the HeatMap tab as soon as the user inputs the data using the browse button (or chooses the example data) using the default settings, i.e. Pearson correlation distance and average agglomerative linkage method for clustering. Using the right hand side panel, the user can change the distance and the agglomerative linkage methods to display a new heatmap reflecting the changes. Row and column dendograms are displayed on the left side and the top of the plot respectively. The same Dendograms with their corresponding sample clusters can be found on separate tabs for Column Dendograms or Row Dendograms.

- **Normalization type**

CASH tool allows users to change various parameters when generating the HeatMap. Data can be scaled using any of the normalization type options provided on the left hand sidebar i.e. “row” (default), “column”, “both” rows and columns or “none”. No normalization scales the heat map using the original log2 transformed data.

- **Scale Range**

The color scale is set to (-2, 2) by default but can be increased to up to (-10, 10) using the slider. If no normalization is used, the scale can be set to the positive values to represent the original log2 transformed data.

- **Clustering Measures**

The user can choose between eight different agglomerative linkage methods (complete, ward.D, ward.D2, single, average, mcquitty, median, centroid) and seven distance methods (euclidean, maximum, manhattan, canberra, binary, minkowski, pearson correlation) for generating the HeatMap. The underline options are default settings.

By default, both the Row and Column dendrograms are reordered based on means. If no row reordering is desired, the user can select Row dendograms = FALSE and no row dendrogram will be displayed. The same applies for column dendograms.

When in the HeatMap tab, along with the dendrograms if the user wants to identify the genes and samples, using the row and column labels, the user can select ***Display Row labels? = ‘Yes’ and***/or ***Display Col labels? = “Yes”*** option to see them. The font size for the row and column labels can be adjusted depending on the size of the heatmap using the If yes, Row Label font size and/or If yes, Col Label font size bar. By default, the size for each of the labels in set to 0.5.

- **HeatMap colors**

Depending on the type of data input, expression or methylation, the user can choose the desired coloring scheme to represent the data. The user will have to choose three different colors. For expression data, generally Green-Black-Red is preferred whereas for methylation data, Blue-White-Red is considered the norm. By default, the Blue-White-Red colors have been chosen to represent the example methylation dataset.

- **Subgroup based on clusters**

In some instances, the user may be interested to cut the Column into 'n' no. of clusters and observe which samples and their corresponding clusters. When on the column tab, a message will be displayed under the tree '***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)***'. The user can select that option and a group clustering information will be displayed on the same tab screen. The same applies to subgroups gene sets/ cpG islands when on the row dendrogram tab.

- **Access gene set significance in separation of Samples into two or more clusters**

The user may also be interested in estimating the statistical significance of a gene set in being able to separate two groups or greater (eg. Tumor vs Normal or Favorable v/s Unfavorable) based on a cluster analysis in comparison to a random sample set. When on the column dendrograms tab, a message will be displayed '**Would you want to assess gene set significance in the separation of specimens into two clusters? (Yes/No)**'. The user can select Yes option and the user will be asked to provide specific information for the estimation of the p-values using the bootstrap method such as the dataset to resample from (available example methylation dataset or load your own data in the same format as the original data), the sample size for the resampling and the no. of iterations the user wishes to perform. It is recommended that the user chooses at least 1000 iterations and a sample size of at least 1000 genes/cpG islands. Only once the user hits the 'go' button, will the p-value calculations begin. The time required for the estimation of the p-value can also be tracked using this tool. After approximately a 10 second lag period, the user will see a small box appear on the top right hand corner of the page that indicates the iteration number the analysis is on. Also a progress bar will appear on the top of the page just under the address bar. After the analysis is complete, the p-value and its interpretation will be posted on the tab screen along with the distribution of the permuted p-values in comparison to the p-value obtained from the original data. The p-value calculation will not be carried out in instances where the original data p-value is not significant and an error will appear on the screen that "Testing significance does not make sense". Similar p-value analysis can be conducted when in the Row dendrogram tab for the estimation of significance in separation of gene sets into two or more clusters.

Terms of Use

Authors- Manali Rupji, dual M.S., Bhakti Dwivedi Ph.D. & Jeanne Kowalski Ph.D.
Maintainer- Manali Rupji 'manali(dot)rupji(at)emory(dot)edu'