

Package ‘cytoMEM’

March 30, 2023

Type Package

Title Marker Enrichment Modeling (MEM)

Version 1.3.0

Description MEM, Marker Enrichment Modeling, automatically generates and displays quantitative labels for cell populations that have been identified from single-cell data. The input for MEM is a dataset that has pre-clustered or pre-gated populations with cells in rows and features in columns. Labels convey a list of measured features and the features' levels of relative enrichment on each population. MEM can be applied to a wide variety of data types and can compare between MEM labels from flow cytometry, mass cytometry, single cell RNA-seq, and spectral flow cytometry using RMSD.

License GPL-3

Imports gplots, tools, flowCore, grDevices, stats, utils, matrixStats, methods

Collate

'MEM_function.R'create_labels_txt.R'IQR_thresh.R'build_heatmaps.R'add_cluster_ID.R'add_fileID_to_clusterID.R'ch

biocViews Proteomics, SystemsBiology, Classification, FlowCytometry, DataRepresentation, DataImport, CellBiology, SingleCell, Clustering

Depends R (>= 4.2.0)

Suggests knitr, rmarkdown

VignetteBuilder knitr

URL <https://github.com/cytolab/cytoMEM>

git_url <https://git.bioconductor.org/packages/cytoMEM>

git_branch devel

git_last_commit f03bc9f

git_last_commit_date 2022-11-01

Date/Publication 2023-03-30

Author Sierra Lima [aut] (<<https://orcid.org/0000-0001-5944-750X>>),
Kirsten Diggins [aut] (<<https://orcid.org/0000-0003-1622-0158>>),
Jonathan Irish [aut, cre] (<<https://orcid.org/0000-0001-9428-8866>>)

Maintainer Jonathan Irish <jonathan.irish@vanderbilt.edu>

R topics documented:

build_heatmaps	2
MEM	4
MEM_matrix	7
MEM_RMSD	8
MEM_values	9
PBMC	10

Index	12
--------------	-----------

build_heatmaps	<i>Build heatmaps</i>
----------------	-----------------------

Description

Takes matrix list generated from [MEM](#) as input and outputs MEM labels, heatmap of population median values, and heatmap of MEM scores.

Usage

```
build_heatmaps(
  MEM_values,
  cluster.MEM = "both",
  cluster.medians = "none",
  cluster.IQRs = "none",
  display.thresh = 1,
  output.files = FALSE,
  labels = FALSE,
  only.MEMheatmap = FALSE)
```

Arguments

MEM_values	List of matrices generated as output from MEM aka MEM_values
cluster.MEM	"none", "row", "col", or "both"; which dimension(s) of the MEM heatmap to hierarchically cluster. See Details for more information.
cluster.medians	"none", "row", "col", or "both"; which dimension(s) of the median heatmap to cluster. If "none", the median row and column order will match order of the MEM heatmap.
cluster.IQRs	"none", "row", "col", or "both"; which dimension(s) of the IQR heatmap to cluster. If "none", the IQR row and column order will match order of the MEM heatmap.
display.thresh	Numeric; 0-10. Markers with enrichment scores that are equal to or greater than display.thresh will be displayed as row names on MEM heatmap.

output.files	TRUE or FALSE; If TRUE, output of build_heatmaps will be written to folder output folder that is created as a subdirectory of the working directory. Written files include a PDF of the MEM heatmap as well as txt files containing MEM labels (row names), MEM scores, medians, and IQR values.
labels	TRUE or FALSE; whether or not to include MEM labels in the MEM heatmap
only.MEMheatmap	TRUE or FALSE; whether to only show MEM heatmap or all relevant heatmaps (MEM, median, IQR)

Details

Heatmaps are clustered using the default complete linkage hierarchical clustering in the [hclust](#) function. See [heatmap.2](#) and [hclust](#) for more information.

Value

Heatmaps of median, IQR, and MEM values on each population; optionally written to file.

Author(s)

Kirsten Diggins, Sierra Lima, Jonathan Irish

References

Diggins et al., Nature Methods, 2017

See Also

[MEM](#), [heatmap.2](#), [hclust](#)

Examples

```
# Use output from MEM function or use example data with
data(MEM_values)

build_heatmaps(
  MEM_values,
  cluster.MEM = "both",
  cluster.medians = "none",
  cluster.IQRs = "none",
  display.thresh = 1,
  output.files = TRUE,
  labels = FALSE,
  only.MEMheatmap = FALSE)
```

Description

The MEM function takes pre-clustered, single-cell data as input and calculates relative enrichment scores for each marker on each population.

Usage

```
MEM(exp_data,
    transform=FALSE,
    cofactor=1,
    choose.markers=FALSE,
    markers="all",
    choose.ref=FALSE,
    zero.ref=FALSE,
    rename.markers=FALSE,
    new.marker.names="none",
    file.is.clust=FALSE,
    add.fileID=FALSE,
    IQR.thresh=NULL,
    output.prescaled.MEM=FALSE,
    scale.matrix = "linear",
    scale.factor = 0)
```

Arguments

exp_data	list of file names or a matrix or data.frame object where the last column contains a numeric cluster ID for each cell (row). If exp_data is a list of files, either each file must be one cluster or each file must contain a cluster channel (column) that specifies a numeric cluster ID for each cell (row). See details for more information.
transform	TRUE or FALSE; whether or not to apply asinh transformation to the data. Default is FALSE.
cofactor	numeric; if transform is TRUE, what cofactor should be applied. Default is 1. Arcsinh transformed value = arcsinh(raw value/cofactor)
choose.markers	TRUE or FALSE; whether or not the user wants to choose the markers (columns) for analysis in the console. If data contains markers that will not be used in the analysis (e.g. SSC or FSC channels in flow data), should be set to TRUE. If FALSE, either all of the markers in the experiment data will be used in MEM or the user can pass a character string of the markers to be used in the analysis using the function call (markers) below.
markers	"all" or ex."1:2,7,11:12,25"; if the user wants to choose markers to be used in the MEM analysis without having the console ask for a user input, enter a character string similar to the one shown in the example. The markers chosen

should be separated with colons or commas, without spaces between. If "all", all of the markers will be used in MEM.

choose.ref	TRUE or FALSE; Default reference for each population is all other populations in the dataset. For example, in a dataset containing 7 clusters, reference for population 1 would include clusters 2-7. If set to TRUE, user will be prompted in the console to enter which cluster(s) should be used as reference instead of the default bulk non-population reference.
zero.ref	TRUE or FALSE; If set to TRUE, a zero, or synthetic negative, reference will be used for all populations. MAGref therefore is 0 and IQRref is the median IQR across all markers chosen. MEM scores will go from 0 to +10 instead of -10 to +10.
rename.markers	TRUE or FALSE; if TRUE, user will be prompted to enter new column names in the console. Default FALSE. If FALSE, either the column names will not be changed or the user can pass a character string of the new column names using the function call new.marker.names below.
new.marker.names	"none" or ex."CD4, CD19, HLA-DR, CD8, CD14, CD16"; if user wants new column names for channels without having the console ask for a user input, enter a character string like the one shown in the example. Each new column name should be separated by a comma, without spaces between names. If "none", the column names will not be changed.
file.is.clust	TRUE or FALSE; if multiple files are entered as input and each file contains cells from only one cluster, should be set to TRUE. This prompts function to merge data into one matrix for analysis and to add a file ID for each file that will stand in as the cluster ID. A text file indicating which file corresponds to which cluster number will be written to the output files folder (by default will be created as a subdirectory in your working directory).
add.fileID	TRUE or FALSE; if multiple files are entered but file.is.clust is FALSE, this indicates that there are multiple files but each contains cells from multiple clusters and that there is already a cluster channel included as the last column in each file. If add.fileID is TRUE, a file ID will be appended to the cluster ID so user can identify the file as well as cluster from which each population came.
IQR.thresh	Default NULL. Optionally can be set to a numeric value. See Details for more information.
output.prescaled.MEM	Default FALSE. If TRUE, creates folder in working directory called "output files" containing a TXT file with pre-scaled MEM values. The MEM matrix output by build_heatmaps contains post-scaled (-10 to +10 scale) MEM values.
scale.matrix	Default "linear". Choose how to scale the MEM matrix. Can choose from "linear" "log" or "arcsinh" for the MEM matrix scale, apply scale, and then transform from -10 to 10 or 0 to 10.
scale.factor	Default 0. Choose the factor for the MEM matrix scaling. For example, choosing 2 will apply a log2 scale if "log" is chosen for scale.matrix, if "arcsinh" is chosen then choosing 2 for scale.factor will use arcsinh scale with a cofactor of 2.

Details

For each population and its reference, MEM first calculates [median](#) marker levels and marker interquartile ranges (IQR), and then calculates MEM scores according to the equation

$$\text{MEM} = |\text{Median_Pop} - \text{Median_Ref}| + \text{IQR_Ref}/\text{IQR_Pop} - 1 ; \text{ if } \text{Median_Pop} - \text{Median_ref} < 0, -\text{MEM}$$

A dataset is provided as an example to be used with [MEM](#) and [build_heatmaps](#). Please see dataset [PBMC](#) for more details.

Input data can be file type .txt, .fcs, or .csv. A matrix or data.frame object where the last column contains cluster identity per cell is also accepted. In all cases, the expected data structure is cells (datapoints) in rows and measured markers (i.e. features, parameters) in columns of the input data.

IQR threshold: The MEM equation takes the ratio of population and reference IQRs and adds this value to the difference in medians. Low IQR values below 1, like those resulting from background noise level measurements, can therefore artificially inflate the overall MEM score. In order to correct this, a threshold of 0.5 is automatically applied. However, the function can calculate an IQR threshold using the input data. If IQR_thresh is set to "auto", the threshold will be calculated as the IQR associated with the 2nd quartile median value across all populations and corresponding reference populations. This should be used if the user anticipates that 0.5 will not be an adequate threshold for the particular dataset.

Value

MAGpop	Matrix; Median expression level of markers on each population
MAGref	Matrix; Median expression on each population's corresponding reference population
IQRpop	Matrix; IQR of markers on each population
IQRref	Matrix; IQR on each population's corresponding reference population

Note

The object generated from [MEM](#) is meant to be passed to [build_heatmaps](#) which will generate MEM labels and heatmaps.

Author(s)

Kirsten Diggins, Sierra Lima, and Jonathan Irish

References

Diggins et al., Nature Methods, 2017

See Also

[build_heatmaps](#)

Examples

```
## For multiple file input, set working directory to folder containing files, then
## infiles <- dir()

## For single file or object input (e.g. PBMC), input data directly into MEM function

## User inputs
data(PBMC)
MEM_values = MEM(
  PBMC,
  transform=TRUE,
  cofactor=15,
  choose.markers=FALSE,
  markers="all",
  choose.ref=FALSE,
  zero.ref = FALSE,
  rename.markers=FALSE,
  new.marker.names="none",
  IQR.thresh=NULL,
  output.prescaled.MEM=FALSE,
  scale.matrix = "linear",
  scale.factor = 0)
```

MEM_matrix

MEM matrix

Description

This matrix is the output generated from [MEM](#) analysis of the [PBMC](#) dataset. It is meant to be used as input for the [MEM_RMSD](#) function to generate RMSD scores of similarity.

Usage

```
data(MEM_matrix)
```

Format

The format is the 7 populations in rows and the MEM scores for all 25 measured markers in columns. See [PBMC](#) dataset for more details.

Examples

```
data(MEM_matrix)
```

MEM_RMSD

*MEM RMSD similarity between populations***Description**

MEM_RMSD calculates a normalized average RMSD score pairwise between populations given their MEM scores as input. This is meant to serve as a metric of similarity between populations.

The function calculates the sum of squares for all shared markers between two populations, then takes the square root of the average.

For "a" through n markers, the sum of squares is calculated as: $\text{sum of squares} = (a_2 - a_1)^2 + (b_2 - b_1)^2 \dots (n_2 - n_1)^2$

Root-mean-square deviation (RMSD) is calculated as: $\text{RMSD} = \sqrt{\text{sum of squares} / \text{number of markers}}$

The RMSD values are then converted to percentages with the maximum RMSD in the matrix set as 100 percent, so that the final RMSD score is the percent of the maximum RMSD.

$\text{Percent_max_RMSD} = 100 - \text{RMSD} / \text{max_RMSD} * 100$

The function then outputs a clustered heatmap of Percent_max_RMSD values and the matrix of numerical values used to build the heatmap.

Usage

```
MEM_RMSD(
  MEM_matrix,
  format=NULL,
  output.matrix=FALSE)
```

Arguments

MEM_matrix	The input to MEM_RMSD can be either 1) a matrix of values, where populations are in rows and their MEM scores are in columns, 2) the list of matrices output by MEM, or 3) a file path pointing to a folder containing tab-delimited text files, one file for each population, where each file lists marker names in the first column and the corresponding MEM scores in the second column.
format	Default is NULL. When format is equal to "pop files", the function expects a file path as input where the designated folder contains one file for each population's set of MEM scores.
output.matrix	If TRUE, the RMSD heatmap in PDF format and txt file with matrix of values calculated by the function will be output and located in a folder called "output files" that is generated in the working directory.

Details

If you are calculating MEM_RMSD on population files, populations do not have to include all of the same markers. The function will determine which markers each pair of populations has in common

and will use those common markers to calculate RMSD. If the populations have no markers in common, the function will terminate with an error. Note that population names must match exactly between files in order for them to be considered the same.

Value

RMSD_vals	Matrix of the calculated pairwise percent max RMSD scores
RMSD heatmap	Hierarchically clustered heatmap of RMSD_vals

Author(s)

Kirsten Diggins, Sierra Lima, Jonathan Irish

References

Diggins et al., Nature Methods, 2017

Examples

```
## For single matrix, input data directly into RMSD function

## User inputs
data(MEM_matrix)

MEM_RMSD(
  MEM_matrix,
  format=NULL,
  output.matrix=FALSE)
```

MEM_values

MEM values

Description

This list of 5 matrices is the output generated from MEM analysis of the PBMC dataset. It is meant to be used as input for the [build_heatmaps](#) function to generate population-specific MEM labels and clustered median and MEM score heatmaps.

Usage

```
data(MEM_values)
```

Format

The format is: List of 6 \$ MAGpop :List of 1 ..\$: num [1:7, 1:25] 0.0254 0.0189 0.0207 2.5075 2.4995- attr(*, "dimnames")=List of 2\$: chr [1:7] "1" "2" "3" "4"\$: chr [1:25] "CD19" "CD117" "CD11b" "CD4" ... \$ MAGref :List of 1 ..\$: num [1:7, 1:25] 0.0146 0.0209 0.0206 0.0213 0.0235- attr(*, "dimnames")=List of 2\$: chr [1:7] "1" "2" "3" "4"\$: chr [1:25] "CD19" "CD117" "CD11b" "CD4" ... \$ IQRpop :List of 1 ..\$: num [1:7, 1:25] 0.5 0.5 0.5 0.68 0.655- attr(*, "dimnames")=List of 2\$: chr [1:7] "1" "2" "3" "4"\$: chr [1:25] "CD19" "CD117" "CD11b" "CD4" ... \$ IQRref :List of 1 ..\$: num [1:7, 1:25] 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5- attr(*, "dimnames")=List of 2\$: chr [1:7] "1" "2" "3" "4"\$: chr [1:25] "CD19" "CD117" "CD11b" "CD4" ... \$ MEM_matrix:List of 1 ..\$: num [1:7, 1:25] 0.014421 -0.002766 0.000195 3.309071 3.295596- attr(*, "dimnames")=List of 2\$: chr [1:7] "1" "2" "3" "4"\$: chr [1:25] "CD19" "CD117" "CD11b" "CD4" ... \$ File Order:List of 1 ..\$: num 0

Examples

```
data(MEM_values)
```

 PBMC

Normal Human Peripheral Blood Mononuclear Cells (PBMCs)

Description

This dataset contains 49651 events from a 25-marker panel CyTOF analysis of normal human PBMCs. Expression values are the raw (pre-transformation) median intensity (MI) values.

Data has been pre-gated to include only DNA-intercalator (Iridium) positive and CD45-high events. Expert biaxial gating was used to separate these events into 7 major blood cell populations: CD4+ T cells (cluster 1), CD8+ T cells (cluster 2), IgM+ B cells (cluster 5), IgM- B cells (cluster 4), dendritic cells (DCs) (cluster 3), natural killer (NK) cells (cluster 7), and monocytes (cluster 6). Per-cell population identity is specified in the `cluster` channel (variable).

This dataset is meant to be used as an example with the [MEM](#) package.

See refs for experimental protocol and further details.

Usage

```
data("PBMC")
```

Format

A data frame with 49651 observations on the following 26 variables.

The following 25 surface markers were measured by CyTOF.

CD19 B cell receptor (BCR)

CD117 c-Kit; RTK expressed by stem and progenitor cells

CD11b ITGAM, macrophage-1 antigen; complement receptor 3

CD4 T cell receptor (TCR) co-receptor; binds antigens presented by MHC II
CD8 T cell receptor (TCR) co-receptor; binds antigens presented by MHC I
CD20 B cell surface protein
CD34 surface protein expressed by hemotopoeitic stem cells and lost over course of differentiation
CD61 surface marker expressed on platelets
CD123 Interleukin-3 receptor; expressed by progenitor cells
CD45RA CD45 isoform expressed by Naive T lymphocytes
CD45 protein tyrosine phosphatase; expressed by all mature leukocytes
CD10 membrane metallo-endopeptidase expressed by common lymphoid progenitors
CD33 Siglec-3; expressed by myeloid cells
CD11c complement receptor; highly expressed on dendritic cells and myeloid cells
CD14 pattern recognition receptor expressed by innate lymphoid cells
CD69 involved in signaling and proliferation of activated t-lymphocytes and natural killer cells
CD15 plays role in phagocytosis and chemotaxis; expressed in multiple blood cell malignancies
CD16 low affinity Fc receptor for IgG; expressed by natural killer cells, neutrophils, and myeloid cells
CD44 cell adhesion molecule and hyaluronic acid receptor
CD38 highly expressed on germinal center B cells and plasma cells
CD25 IL-2 receptor; expressed by activated T cells
CD3 T cell receptor (TCR)
IgM heavy chain isoform of BCR
HLADR MHC class II receptor
CD56 NCAM; expressed by natural killer cells
cluster 1: CD4+ T cells 2: CD8+ T cells 3: Dendritic cells (DCs) 4: IgM- B cells 5: IgM+ B cells 6: Monocytes 7: Natural killer (NK) cells

Details

The dataset should be arcsinh transformed with cofactor of 15. See [MEM](#) for more details.

Source

Leelatian et al., Methods Mol Biol, 2015.

References

Diggins et al., Methods, 2016. Diggins et al., Nature Methods, 2017

Examples

`data(PBMC)`

Index

* datasets

MEM_matrix, 7

MEM_values, 9

PBMC, 10

asinh, 4

build_heatmaps, 2, 5, 6, 9

hclust, 3

heatmap.2, 3

IQR, 6

median, 6

MEM, 2, 3, 4, 6–11

MEM_matrix, 7

MEM_RMSD, 7, 8

MEM_values, 9

PBMC, 6, 7, 9, 10