Package ‘MiRAnorm’

November 3, 2016

Type Package

Title Adaptive Normalization for miRNA Data

Version 1.0.0

Description An adaptive normalization algorithm that selects housekeeping genes based on the sample level variability in the data. This is suitable for any data obtained from RT-qPCR assays. A manuscript describing the method is submitted to Genome Biology under ``MiRA-norm: An Adaptive Method for the Normalization of MicroRNA Array Data``, Yuda Zhu et al.

Depends R (>= 3.1.0)

Imports grDevices, graphics, stats, utils, ggplot2, cluster, nmpv, dendextend, parallel, MASS, plyr, reshape2, ArgumentCheck

License GPL-3

LazyData TRUE

RoxygenNote 5.0.1

NeedsCompilation no

Author Yuda Zhu [aut, cre], Amanda Zhao [aut]

Maintainer Yuda Zhu <yudazhu@gmail.com>

Repository CRAN

Date/Publication 2016-11-03 11:33:06

R topics documented:

  dist.func ................................................................. 2
  impmat ................................................................. 2
  miranorm ............................................................... 3
  scr ................................................................. 4
  simData ............................................................... 5

Index 6
Description

dist.func returns a sum total of all distances for each row of the matrix from the centroid.

Usage

dist.func(mndata, standardize = TRUE, method = "euclidean", centroid = NULL)

Arguments

- mndata: matrix dataset with numeric columns.
- standardize: is a boolean variable to denote whether columns should first be scaled to 0-1.
- method: dictates how distance is computed. Currently, only "euclidean" is implemented.
- centroid: allows user to specify the centroid values for each column. Otherwise, the mean of each column is taken by default.

Description

impmat imputes a dataset entered as rows=Genes, columns=Samples. NA values are imputed by row. Currently, impvalue function imputes these as the max value of the row, the largest observed value for a given Gene across all Samples.

Usage

impmat(data)

Arguments

data: dataset to be imputed. Should contain all numeric values with rows=Genes and columns=Samples.
miranorm

Adaptive algorithm to identify normalization genes.

Description

miranorm returns a list of suggested normalization genes based on supplied data. Various output figures are also available if requested in the parameter list.

Usage

miranorm(data = dat, group = dat$Trt, max = 15, min = 3,
method = "complex", dist.method = "Euclidean", hclust.method = "single",
ct = 25, missing = 0, clustplot = TRUE, selected = 4, ggplot = TRUE,
heatmap = TRUE, known.positives = NULL, suggested.list = NULL,
exclude = NULL)

Arguments

data

dataframe containing at a minimum: Sample, Gene, Ct, and Trt

group

treatment allocation. This should be the same length as the number of rows in
data.

max

when method is chosen as "complex", this determines the maximum selected
size at which the stability evaluation is done.

min

when method is chosen as "complex", this determines the minimum selected
size at which the stability evaluation is done.

method

choice of "simple" or "complex". Simple runs a single pass of miranorm for
suggested normalizing genes. Complex runs bootstrap samples across a range
of sizes to compute a stability metric.

dist.method

distance metric used to calculate pairwise distance between individual miRNAs
across all samples. Currently "Euclidean" and "1-Cor" are implemented.

hclust.method

the agglomerates method used to group genes. Methods are the same as de-
defined in the hclust function in the stats package and include "single", "average",
"complete", and "ward.D2". "single" is recommended as it is more robust to
small perturbations and tends to form the "chaining" phenomenon useful for
defining normalizing genes.

c

cycle threshold values at or above this level are treated as NA for the purposes
determining normalization genes. Recommend the value be set to 25.

missing

defines maximum percentage of samples missing for a given gene before that
gene is excluded from dataset during normalization.

clustplot

"true" or "false" to output or suppress stability plot. Only applicable if method
= "complex".

selected

how many adaptive normalizing genes to search for in panel. Note, actual num-
ber of genes found may be larger based on tree cut.

ggplot

"true" or "false" to output general raw data plots.
heatmap
   "True or "False" to output or suppress heatmap plot.

known.positives
   Names of miRNA that are known positive. These will be added automatically to the heatmap plot.

suggested.list
   Names of miRNA that are user suggested normalizing miRNA. These will be added automatically to the heatmap plot.

exclude
   List of miRNA to exclude from the selection process for HK genes, eg: known.positives should be included here.

Value
   A list including the following: nmean, nmed, nlcv, lcv.gene, ncls, ncls.gene.

nmean is the dataset normalized to the global mean.

nmed is the dataset normalized to the global median.

nlcv is the dataset normalized to the average of the 3 genes with the lowest CV.

lcv.gene is the names of the 3 genes with used to normalize nlcv

ncls is the dataset normalized to the adaptive normalizing genes chosen from miranorm.

ncls.gene is the names of the genes chosen as adaptive normalizing genes chosen from miranorm.

Examples

    dat = simData(n.trt=15, n.ctrl=15, n.gene=30, n.err=10, sigma.error = c(1, 0.3), mean.sample = 2, sigma.sample = 1.88, sigma.gene = 0.1, n.big.effect = 5, n.small.effect = 10, mean.big.effect = 2, mean.small.effect = 1.2)$sim

    obj = miranorm(data = dat, group = dat$Group, method="simple")

scr

Replacing values above ct to NA

Description

scr screens the dataset in wide format according to Ct threshold and missing.percent.thresh (for each gene, the percentage of samples is missing). sets all values above Ct threshold to NA and then removes all rows of genes with number of missing above missing.percent.thresh.

Usage

scr(dat, threshold, missing.percent.thresh)
simData

Arguments

dat                  the dataset with rows = Genes and columns = Samples.
threshold            Ct value at or above which all values are set to NA.
missing.percent.thresh
                         a number between 0 to 100 denoting the percentage of allowable missing. If the
                         percentage of missing is greater than this value, the row (gene) is not retained.

---

Simulating a RT-PCR miRNA dataset.

Description

simData simulates a RT-PCR miRNA dataset with user defined levels of variability and treatment
effect size.

Usage

simData(n.trt = 50, n.ctrl = 50, n.gene = 96, sigma.error, n.err = 10,
         mean.sample = 0.6, sigma.sample = 0.6, sigma.gene = 0,
         n.big.effect = 5, n.small.effect = 15, mean.big.effect = 5,
         mean.small.effect = 2)

Arguments

n.trt            Number of simulated treatment samples
n.ctrl           Number of simulated control samples
n.gene           Number of simulated genes in the panel
sigma.error      a vector of length 2 for 2 different measurement error sizes (sd).
n.err            number of genes with sigma.error[2]. The rest (n.gene - n.err) have sigma.error[1]
                  measurement error.
mean.sample      the unadjusted mean of the samples. Can generally be left as default
sigma.sample    the unadjusted sample to sample sd. Can generally be left as default.
sigma.gene       sd of gene to gene effect sizes for large and small treatment effects.
n.big.effect    Number of genes with large treatment effect
n.small.effect   Number of genes with small treatment effect
mean.big.effect
                  Average effect size for a "large" treatment effect
mean.small.effect
                  Average effect size for a "small" treatment effect
Index

dist.func, 2
impmat, 2
miranorm, 3
scr, 4
simData, 5