Package ‘QuantNorm’

February 1, 2019

Title Mitigating the Adverse Impact of Batch Effects in Sample Pattern Detection

Description Modifies the distance matrix obtained from data with batch effects, so as to improve the performance of sample pattern detection, such as clustering, dimension reduction, and construction of networks between subjects. The method has been published in Bioinformatics (Fei et al, 2018, <doi:10.1093/bioinformatics/bty117>). Also available on 'GitHub' <https://github.com/tengfei-emory/QuantNorm>.

Version 1.0.5

Depends R (>= 3.3.0)
Imports stats, utils
Suggests GGally, ggplot2, network, pheatmap, rgl, sna

Date 2019-02-01

Author Teng Fei, Tianwei Yu

Maintainer Teng Fei <tfei@emory.edu>

BugReports https://github.com/tengfei-emory/QuantNorm/issues

License GPL (>= 2)

RoxygenNote 6.0.1
NeedsCompilation no

Repository CRAN

Date/Publication 2019-02-01 20:23:25 UTC

R topics documented:

  brain ................................................................. 2
  connection.matrix .................................................. 2
  ENCODE ............................................................. 3
  QuantNorm ......................................................... 4

Index 6
### connection.matrix

**brain**  
*Brain RNA-Seq data for both human and mouse.*

**Description**  
Brain RNA-Seq data for both human and mouse.

**Usage**  
```r
data(brain)
```

**Format**  
Large matrix with 15041 rows and 62 columns

**Source**  

---

### connection.matrix

*Construct connection matrix for network analysis*

**Description**  
For data with known labels, this function constructs a connection matrix between unique labels, such as unique cell types. The returned matrix can be used for subject-wise network construction.

**Usage**  
```r
connection.matrix(mat, label, threshold = 0.15, closest = TRUE)
```

**Arguments**

- **mat**: n*n dissimilarity (1-correlation) matrix (e.g. obtained by QuantNorm).
- **label**: n-dimension vector for the labels of the n subjects. Replicates share the same label.
- **threshold**: A number between 0 to 1. Two groups will be regarded as connected if average 1-correlation < threshold.
- **closest**: True or False. Whether connect the closest group or not if the closest group cannot satisfy the threshold condition.

**Value**  
Returns the connection matrix between unique labels.
Author(s)
Teng Fei. Email: tfei@emory.edu

References

Examples

library(network); library(ggplot2); library(sna); library(GGally) #drawing network graph
data("ENCODE")

#Assigning the batches based on species
batches <- c(rep(1,13),rep(2,13))

#QuantNorm correction
corrected.distance.matrix <- QuantNorm(ENCODE,batches,method='row/column', cor_method='pearson', logdat=FALSE,standardize = TRUE, tol=1e-4)

#Constructing connection matrix
mat <- connection.matrix(mat=corrected.distance.matrix,label=colnames(corrected.distance.matrix))

#Creating network object and plot
ENCODE.net=network(mat, directed=FALSE)
ENCODE.net %V% "Species" <- c(rep('Human',13),rep('Mouse',13))
p0 <- ggnet2(ENCODE.net,label=TRUE,color = 'Species', palette = "Set2",
             size = 3, vjust = -0.6,mode = "kamadakawai",label.size = 3,
             color.legend = 'Species')+theme(legend.position = 'bottom')
plot(p0)

Normalization ENCODE raw counts data for both human and mouse.

Description
Normalized ENCODE raw counts data for both human and mouse.

Usage
data(ENCODE)

Format
Large matrix with 10309 rows and 26 columns
QuantNorm

Source

QuantNorm

Adjust the distance matrix by quantile normalization for data with batch effect

Description
This function applies quantile normalization on the distance matrix (dissimilarity matrix) and return the corrected distance matrix.

Usage
QuantNorm(dat, batch, method = "row/column", cor_method = "spearman", 
tol = 0.01, max = 50, logdat = TRUE, standardize = FALSE)

Arguments

dat
The original p*n batch effect data with n subjects and p RNA-seq measurements or the n by n distance matrix.

batch
The vector of length n indicating which batch the subjects belong to.

method
Method for the quantile normalization. There are two options: "row/column" and "vectorize".

cor_method
Method to calculate the correlation matrix, can be 'spearman'(default), 'pearson' or 'kendall'.

tol
The tolerance for the iterative method "row/column", which is the Euclidean distance of the vectorized two dissimilarity matrices before and after each iteration.

max
Maximum number of the iteration if the tolerance is not reached.

logdat
Whether conducting log transformation to data or not.

standardize
Whether conducting standardization [(dat - mean)/sqrt(var)] to data or not.

Value
Returns the corrected 1-correlation matrix between subjects.

Author(s)
Teng Fei. Email: tfei@emory.edu

References
Examples

library(pheatmap) #drawing heatmap

data("ENCODE") #load the ENCODE data

#Before correction, the subjects are clustered by species
pheatmap(cor(ENCODE))

#Assigning the batches based on species
batches <- c(rep(1,13),rep(2,13))

#QuantNorm correction
corrected.distance.matrix <- QuantNorm(ENCODE,batches,method='row/column', cor_method='pearson',
                                      logdat=FALSE, standardize = TRUE, tol=1e-4)

pheatmap(1-corrected.distance.matrix)
Index

*Topic **datasets**
  brain, 2
  ENCODE, 3

brain, 2

collection_matrix, 2

ENCODE, 3

QuantNorm, 4