Package ‘MANCIE’

October 12, 2022

Type Package

Title Matrix Analysis and Normalization by Concordant Information Enhancement

Version 1.4

Date 2016-03-01

Author Tao Wang, Chongzhi Zang

Maintainer Tao Wang <tao.wang@utsouthwestern.edu>

Depends R (>= 2.15.0)

Description
High-dimensional data integration is a critical but difficult problem in genomics research because of potential biases from high-throughput experiments. We present MANCIE, a computational method for integrating two genomic data sets with homogenous dimensions from different sources based on a PCA procedure as an approximation to a Bayesian approach.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2016-03-02 01:14:19

R topics documented:

mancie *Matrix Analysis and Normalization by Concordant Information Enhancement*

Description

This function removes noise in the main matrix by utilizing information available from the supplementary matrix or summarized supplementary matrix.

Usage

```
mancie(mat_main,mat_supp,cutoff1=0.5,cutoff2=0)
```

Arguments

mat_main	The main matrix or data frame. Rows are features (genes/peaks/etc) and cols are samples (conditions/replicates)
mat_supp	The supplementary matrix or data frame. mat_supp must have the same dimensions as mat_main
cutoff1	The higher cutoff. See below for explanation.
cutoff2	The lower cutoff. See below for explanation.

Details

If the supplementary dataset have the same genomic features on rows and samples on columns as mat_main, it can be directly fed to mancie. An example is RNA-Seq data of the same cell lines from two labs. If the supplementary dataset has different rows from mat_main. It need to be first summarized using summarize_mat to be compatible with mat_main. An example is RNA-Seq data and DNase-seq data of the same tissue types.

The underlying rationale for using MANCIE is that the variation of genomic features in mat_supp are concordant with and can be used to remove noise in the variation of genomic features in mat_main.

(a) If the correlation between row i of mat_main and row i of mat_supp is larger than cutoff1, the new row vector will be the first PC of the matrix formed by these two row vectors. (b) If the correlation is between cutoff1 and cutoff2, the new row vector will be the weighted average of these two rows. The weight for row i of mat_main is 1 and the weight for row i of mat_supp is the correlation between these two row vectors. (c)If the correlation is smaller than cutoff2, the new row vector is the original row i of mat_main

There should be a reasonable portion of rows that fall into the first and second category. If not, the user should check if the data they would like to try MANCIE on really fits the aforementioned rationale. The user may also vary the default values of cutoff1 and cutoff2 if they see fit. The mancie function will report percentage of rows falling into each category.

Value

A modified matrix with the same dimensions as the main matrix

See Also

summarize_mat

Examples

```
data(mancie_example,package="MANCIE")
sum_DNase=summarize_mat(exp,ann_exp,DNase,ann_DNase)
lev_exp=mancie(exp,sum_DNase)
```

mancie_example *The demo dataset for the* MANCIE *package*

Description

This demo dataset is a small portion of the Encode dataset used in our publication.

Usage

```
data(mancie_example)
```

Format

4 data frames

summarize_mat *Summarize information in the supplementary matrix*

Description

Summarize information in the supplementary matrix according to physical location into a new matrix with the same dimensions as the main matrix

Usage

```
summarize_mat(mat_main,ann_main,mat_supp,ann_supp,n_limit=50,extend=100000,method="pca")
```

Arguments

`mat_main` The main matrix or data frame. Rows are features (genes/peaks/etc) and cols are samples (conditions/replicates)

`ann_main` `ann_main` is a data frame that contains the genomic locations of features in `mat_main`. It must have the same number of rows as `mat_main`, and must have columns named as "chr", "start" and "end".

`mat_supp` The supplementary matrix or data frame. Rows are features (genes/peaks/etc) and cols are samples (conditions/replicates) ann_supp — ann_supp os a data frame that contains the genomic locations of features in mat_supp. It must have the same number of rows as mat_supp, and must have columns named as "chr", "start" and "end".

n_limit — The most number of closet features in the supplemenatry matrix that can be used for summarization for each feature in the main matrix

extend — The genomic features in the supplemenatry matrix that are no farther away than extend bp from the feature in question in the main matrix will be used for summarization

method — Which method to summarize the information in the supplementary matrix when there are >1 neighboring row vectors associated with the row vector in the main matrix. "pca" (default) or "max". In the "max" method, the row vector of these neighboring vectors with the highest correlation with the row vector in the main matrix is used. In the "pca" method, PCA is caculated for these row vectors and the first principal component is used.

Details

The main matrix and supplementary matrix must have the same columns corresponding to conditions or replicates. They have different features on rows that can be linked by physical location on genomes. The basic assumption is that one feature's variation in the main matrix is correlated with nearby feature(s)' principal variation in the supplementary matrix.

Value

A modified matrix with the same dimensions as the main matrix

See Also

mancie

Examples

```
data(mancie_example,package="MANCIE")
sum_DNase=summarize_mat(exp,ann_exp,DNase,ann_DNase)
lev_exp=mancie(exp,sum_DNase)
```

Index