Package ‘contamDE’

November 4, 2016

Type Package

Title DE analysis with heterogeneous tumor samples

Version 1.0

Date 2014-11-16

Author Qi Shen, Jiyuan Hu, Hong Zhang

Maintainer Hong Zhang <zhangh@fudan.edu.cn>

Depends R (>= 2.10), edgeR, nloptr

Description contamDE is an R package for identifying differentially expressed genes between heterogeneous tumor samples and matched or unmatched normal samples under 2 or more conditions.

License GPL (>= 2)

LazyLoad yes

RoxygenNote 5.0.1

URL http://github.com/zhanghfd/contamDE

BugReports http://github.com/zhanghfd/contamDE/issues

R topics documented:

`contamDE-package` *Differentially expression analysis for contaminated tumor samples and normal samples.*

Description

This package conducts statistical differential expression (DE) analysis between tumor samples and normal samples using RNA-seq count data from contaminated tumor samples and normal samples. The tumor samples could be either matched or unmatched with normal samples.

Details

Package:	contamDE
Type:	contamDE
Version:	1.0
Date:	2015-01-13
License:	GPL (>= 2)

Author(s)

Qi Shen, Jiyuan Shen, Hong Zhang

Maintainer: Hong Zhang <zhanghd@fudan.edu.cn>

References

SHEN Qi, HU Jiyuan, JIANG Ning, HU Xiaohua, LUO Zewei, and ZHANG Hong (2016) contamDE: Differential expression analysis of RNA-seq data for contaminated tumor samples. Bioinformatics 32(5): 705-712.

`contamDE` *DE analysis using contaminated tumor samples*

Description

Conduct differentially expression analysis between contaminated tumor samples and normal samples. This function provides estimates of relative proportions of pure tumor cells in the contaminated samples, and pesudo likelihood ratio statistics, p-values, and log2-fold change for detecting differentially expressed genes between tumor samples and normal samples. Here the reported proportions are scaled to have average value 1, so that some of them are greater than 1 and some of them are smaller than 1. As a disadvantage, the absulute proportions cannot be estimated without using extra information, refer to Shen et al. (2016) for details. A computationally very fast algorithm implemented in the R package ‘nloptr’ is used to obtain the fold changes, which may report some warning messages that do not affect the final results so they can be safely ignored.

Usage

```
contamDE(data,R,n=NA,match=TRUE)
```

Arguments

`data` This is a G x N read count matrix, where G is the number of genes and N is the total sample size. The (g,i)th entry is the read count of the gth gene of the ith sample.

`R` The number of conditions.

`n` A list of sample sizes for all conditions(valid if match=FALSE).

`match` TRUE if the tumor samples are matched with normal samples and FALSE otherwise. If matched=TRUE, then in ‘data’, the ith, (i+N/R)th,...,(i+(R-1)N/R)th samples are matched with each other (i=1,...,N/R).

Value

W	Scaled proportions (average value = 1) of pure tumor cells in contaminated tumor samples.
LR	A J x (R+1) matrix. Column 1: pseudo likelihood ratio statistics for DE analysis; column 2: p-value for differential expression analysis; column 3~(1+R): log2 fold changes (cancer vs. normal);

References

SHEN Qi, HU Jiyuan, JIANG Ning, HU Xiaohua, LUO Zewei, and ZHANG Hong (2016) contamDE: Differential expression analysis of RNA-seq data for contaminated tumor samples. Bioinformatics 32(5): 705-712.

Examples

```
data("prostate");

## Not run
# d = contamDE(prostate[,-1],R=2,match=TRUE);

data("drosophila");

## Not run
# d = contamDE(drosophila[,-1],R=2,n=list(1:4,5:7),match=FALSE);
```

drosophila *Drosophila melanogaster dataset*

Description

The dataset consists of the RNA-seq read counts for 7 samples of Drosophila melanogaster S2 cells, of which 4 samples were untreated while 3 samples were treated with siRNA targeting the splicing factor pasilla (CG1844).

Usage

```
data("drosophila")
```

Format

A data frame with 7196 observations (genes) on the following 8 variables.

nameOfGene	Gene name
CT.PA.1	Read count from Untreated-3.
CT.PA.2	Read count from Untreated-4.
CT.SI.5	Read count from Untreated-1.
CT.SI.7	Read count from Untreated-6.
KD.PA.3	Read count from "CG8144_RNAi-3".
KD.SI.6	Read count from "CG8144_RNAi-1".
KD.PA.4	Read count from "CG8144_RNAi-4".

Source

Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E., and Graveley, B.R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Research*, 21(2), 193–202.

`lung` *Lung cancer data*

Description

The data are from a study of the lung cancer. Six patients provided tissue samples and normal samples besides the lung tissues. The read counts were summarized by RefSeq transcript, and only those transcripts with at least 50 aligned reads for at least one tissue in each condition were provided in the table. RefSeq identifiers were mapped to the latest official gene symbols by following the user guide of the Bioconductor package 'edgeR' using the Bioconductor annotation package 'org.Hs.eg.db' (version 2.7.1). Those RefSeq identifiers not in the database were discarded, and each gene was represented by the RefSeq transcript with the greatest number of exons and the other transcripts were removed. Altogether 11,597 transcripts (genes) were kept.

Usage

```
data(lung)
```

Format

A data frame with 11,597 observations on the following 13 variables.

`nameOfGene` Gene name

`N4` Read count for normal sample of patient 4.

`T4` Read count for normal sample of patient 4.

`N12` Read count for normal sample of patient 12.

`T12` Read count for tumor sample of patient 12.

`N13` Read count for normal sample of patient 13.

`T13` Read count for tumor sample of patient 13.

`N14` Read count for normal sample of patient 14.

`T14` Read count for tumor sample of patient 14.

`N15` Read count for normal sample of patient 15.

`T15` Read count for tumor sample of patient 15.

`N16` Read count for normal sample of patient 16.

`T16` Read count for tumor sample of patient 16.

`prostate` *Prostate cancer data*

Description

Prostate cancer samples and adjacent normal cell samples were provided by 14 patients from Shanghai Changhai Hospital. This dataset contains the read count of 12,699 genes for each sample.

Usage

```
data("prostate")
```

Format

A data frame with 12,699 observations on the following 29 variables.

`nameOfGene` Gene name

`N1` Read count for normal sample of patient 1.

`N2` Read count for normal sample of patient 2.

`N3` Read count for normal sample of patient 3.

`N4` Read count for normal sample of patient 4.

`N5` Read count for normal sample of patient 5.

`N6` Read count for normal sample of patient 6.

`N7` Read count for normal sample of patient 7.

`N8` Read count for normal sample of patient 8.

`N9` Read count for normal sample of patient 9.

`N10` Read count for normal sample of patient 10.

`N11` Read count for normal sample of patient 11.

`N12` Read count for normal sample of patient 12.

`N13` Read count for normal sample of patient 13.

`N14` Read count for normal sample of patient 14.

`T1` Read count for heterogeneous tumor sample of patient 1.

`T2` Read count for heterogeneous tumor sample of patient 2.

`T3` Read count for heterogeneous tumor sample of patient 3.

`T4` Read count for heterogeneous tumor sample of patient 4.

`T5` Read count for heterogeneous tumor sample of patient 5.

`T6` Read count for heterogeneous tumor sample of patient 6.

`T7` Read count for heterogeneous tumor sample of patient 7.

`T8` Read count for heterogeneous tumor sample of patient 8.

`T9` Read count for heterogeneous tumor sample of patient 9.

`T10` Read count for heterogeneous tumor sample of patient 10.

`T11` Read count for heterogeneous tumor sample of patient 11.

`T12` Read count for heterogeneous tumor sample of patient 12.

`T13` Read count for heterogeneous tumor sample of patient 13.

`T14` Read count for heterogeneous tumor sample of patient 14.

Source

Ren S, Peng Z, Mao J, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W, Chen C, Wang F, Guo X, Lu J, Yang J, Wei M, Tian Z, Guan Y, Tang L, Xu C, Wang L, Gao X, Tian W, Wang J, Yang H, Wang J and others. (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Research* 22(5), 806–821.

Index