Package 'mixtNB'

May 2, 2015

Type Package

Title DE Analysis of RNA-Seq Data by Mixtures of NB

Version 1.0

Date 2015-05-01

Author Elisabetta Bonafede, Cinzia Viroli

Maintainer Cinzia Viroli <cinzia.viroli@unibo.it>

Description Differential expression analysis of RNA-Seq data when replicates under two conditions are available is performed. First, mixtures of Negative Binomial distributions are fitted on the data in order to estimate the dispersions, then the Wald test is computed.

License GPL-3

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-02 14:11:25

R topics documented:

`mixtNB-package` *DE Analysis of RNA-Seq Data by Mixtures of Negative Binomials*

Description

A method for performing differential expression analysis of RNA-Seq data when replicates under two conditions are available is implemented. First, mixtures of Negative Binomial distributions are fitted on the data in order to estimate the dispersions, then the Wald test is performed.

Details

Package:	mixtNB
Type:	Package
Version:	1.0
Date:	2015-05-01
License:	GPL-3

Author(s)

Authros: Elisabetta Bonafede, Cinzia Viroli

Maintainer: Cinzia Viroli <cinzia.viroli@unibo.it>

References

E. Bonafede, F. Picard, S. Robin and C. Viroli (2015), Modelling overdispersion heterogeneity in differential expression analysis using mixtures, under revision.

`filter.em` *Internal function that perform pre-filtering*

Description

Internal function that perform pre-filtering

`fit.mixtNB` *Fitting mixtures of Negative Binomials in RNA-Seq data*

Description

A mixture of K Negative Binomial distributions is fitted on the data with the aim to cluster the genes according to their dispersions. The number of groups must be known in advance.

Usage

```
fit.mixtNB(y, cr, K, it = 200, eps = 1e-05, init = NULL, seme = 1,
           filter = TRUE, quiet = FALSE)
```

Arguments

y	matrix that contains the (normalized) dataset where the rows contain the set of replicates in the counts in two conditions. Given p the number of genes and n the total number of replicates, the matrix must have dimensions p x n.
cr	a vector with n elements that contains the numerical labels of the conditions
K	the number of mixture components
it	maximum number of iterations for the EM algorithm
eps	a tolerance level for checking the convergence of the EM algorithm
init	a list that may contain the initial values for the EM algorithm. It may contain: a K-vector 'a' that are the sizes or dispersions of the Negative Binomials, 'w' is the vector with the K initial values for the weights of the components; 'lambda' is the matrix of dimension p x 2 with the initial values of the lambda parameters. If init is NULL, a random initialization will be used.
seme	A numerical value to be used in the set.seed function
filter	Logical to indicate the genes with very small counts should be removed
quiet	Logical to indicate if information about the fitting should be provided

Details

A mixture of K components is fitted with the aim of clustering the genes according to their dispersions. Genes with too small number of reads across experiments are filtered out. The default is to filter out genes with no more than 5 reads totally across all experiments, AND with no more than 0.5 reads averagely across all experiments. The EM algorithm stops when the maximum number of iterations are reached or the relative increment of log-likelihood is smaller than eps.

Value

A list containing

y	The filtered data
K	The number of components
cr	Labels denoting the condition for the replicates
cl	Posterior classification of the genes to the components
likelihood	Log-likelihood at each iteration
AIC	Akaike information criterion
BIC	Bayesian information criterion
a	Estimated dispersions
lambda	Estimated means
f.z.y	Estimated posterior probabilities
time.sec	Computational time (in seconds)
variances	Estimated variances
gname	Positions of the filtered genes

Author(s)

Elisabetta Bonafede, Cinzia Viroli

References

E. Bonafede, F. Picard, S. Robin and C. Viroli (2015), Modelling overdispersion heterogeneity in differential expression analysis using mixtures, under revision.

Examples

```
# create a toy data set with 1000 genes, and 5 samples in each of the two conditions.
# The first 100 genes are DE expressed. The other 900 genes are null.

lambda.de<-matrix(runif(100,0,250),100)
lambda.de=cbind(lambda.de,lambda.de/exp(rnorm(100,0.5,0.125)))
lambda<-rbind(lambda.de,matrix(runif(900,0,250),900,2))
a<-runif(1000,0.5,600)
cr<-rep(1:2,each=5)
y<-matrix(0,1000,10)
for (i in 1:1000) for (l in 1:10) y[i,l]<-rnbinom(1,mu=lambda[i,cr[l]],size=a[i])
fit=fit.mixtNB(y,cr,K=3)
```

---

fun.a *Internal function for the Newton-Raphson step of the EM algorithm*

Description

Internal function for the Newton-Raphson step of the EM algorithm

---

wald.test *Wald test for performing DE analysis*

Description

This function implements the Wald test for performing DE according to three statistics: difference, ratio and logratio

Usage

```
wald.test(out, statistic = "diff", quiet = FALSE, alpha = 0.01)
```

Arguments

out	The fit of mixtNB
statistic	The statistic to be used: "diff" (difference, the default), "ratio" and "logratio"
quiet	Logical to indicate if the DE genes should be printed
alpha	the significance level to detect DE genes

Details

This function implements the Wald test for performing DE according to three statistics: difference, ratio and logratio. It returns the statistics, the p-values and the adjusted p-values according to the Benjamini and Hochberg (1995)

Value

A list containing

`stat`	The value of the Wald test
`pvalue`	nominal p-values for each gene
`pvalueadj`	adjusted p-values according to the Benjamini and Hochberg (1995)
`var`	estimated variances of the genes
`gname`	Positions of the filtered genes

Author(s)

Elisabetta Bonafede, Cinzia Viroli

References

E. Bonafede, F. Picard, S. Robin and C. Viroli (2015), Modelling overdispersion heterogeneity in differential expression analysis using mixtures, under revision.

Examples

```
lambda.de<-matrix(runif(100,0,250),100)
lambda.de=cbind(lambda.de,lambda.de/exp(rnorm(100,0.5,0.125)))
lambda<-rbind(lambda.de,matrix(runif(900,0,250),900,2))
a<-runif(1000,0.5,600)
cr<-rep(1:2,each=5)
y<-matrix(0,1000,10)
for (i in 1:1000) for (l in 1:10) y[i,l]<-rnbinom(1,mu=lambda[i,cr[l]],size=a[i])
fit=fit.mixtNB(y,cr,K=3)
DE.genes=wald.test(fit)
```

Index