Package 'DoubleExpSeq'

January 20, 2025

Type Package

Title Differential Exon Usage Test for RNA-Seq Data via Empirical Bayes Shrinkage of the Dispersion Parameter

Version 1.1

Date 2015-09-01

Author Sean Ruddy

Maintainer Sean Ruddy <s.ruddy@yahoo.com>

Description Differential exon usage test for RNA-Seq data via an empirical Bayes shrinkage method for the dispersion parameter the utilizes inclusion-exclusion data to analyze the propensity to skip an exon across groups. The input data consists of two matrices where each row represents an exon and the columns represent the biological samples. The first matrix is the count of the number of reads expressing the exon for each sample. The second matrix is the count of the number of reads that either express the exon or explicitly skip the exon across the samples, a.k.a. the total count matrix. Dividing the two matrices yields proportions representing the propensity to express the exon versus skipping the exon for each sample.

License GPL-3

Imports numDeriv, datasets, grDevices, graphics, stats, utils

NeedsCompilation no

Repository CRAN

Date/Publication 2015-09-04 09:25:45

Contents

`DoubleExpSeq-package` *DoubleExpSeq is a package with application to RNA-Seq experiments that tests for differential exon usage.*

Description

The functions take inclusion and total counts. Inclusion counts are counts that express the exon. Exclusion counts are counts that explicitly skip the exon. The Total count is the sum of inclusion and exclusion. The package provides 2 methods for the analysis of differential exon usage in RNA-Seq data: 1) DEB-Seq and 2) WEB-Seq. Each of these assume a double-binomial distribution with the normalization constant equal to 1. A simple empirical bayes strategy is used to shrink the dispersion parameter toward a common consensus from all exons. DEB-Seq implements a 2-parameter empirical bayes strategy to estimate shrunken dispersion estimates. WEB-Seq implements a weighted likelihood approach and estimates the weight parameter via the empirical bayes strategy after reparameterizing the posterior predictive distribution in terms of only the weight parameter. Once the dispersion estimates are found, the count data is fit with a double binomial GLM (ignoring the normalization constant) and the dispersion estimates used within the test statistic calculation, but have no bearing on the estimates of the proportion. The main function returns p-values, adjusted p-values using the Benjamini-Hochberg procedure for multiple testing, and proportion estimates along with other relevant information for each exon.

Details

Package:	DoubleExpSeq
Type:	Package
Version:	1.0
Date:	2014-05-13
License:	GPL-3

Author(s)

Sean Ruddy Maintainer: Sean Ruddy <s.ruddy@yahoo.com>

Examples

```
## Toy exon data: "counts"=inclusion counts, "offsets"=total counts, "groups"=experiment design
  data(exon)

  ## Default will compare "G1" & "G2" using the WEB-Seq method
  ## and uses ALL groups to estimate dispersion
  results.G1G2.WEB <- DBGLM1( counts, offsets, groups)

  # Compare G1 & G3
  results.G1G3.WEB <- DBGLM1( counts, offsets, groups, contrast=c(1,3))

  # Compare G1 & G3. Does not use G2 for dispersion estimation.
 results.G1G3.noG2.WEB <- DBGLM1(counts, offsets, groups, contrast=c(1,3), use.all.groups=FALSE)

  # Global minimum check of the weight parameter estimate in the WEB-Seq method.
  optimPlot( counts, offsets, groups, contrast=c(1,3), use.all.groups=FALSE)

## The DEB-Seq method. Less conservative, more powerful. Very similar ranks to WEB-Seq.
 results.G1G2.DEB <- DBGLM1( counts, offsets, groups, shrink.method="DEB")

## M-A Plot
 WEB.sig <- rownames(results.G1G2.WEB$Sig)
 DB.MAPlot( counts, offsets, groups, de.tags=WEB.sig, main="WEB-Seq")
```

`counts` *Exon Inclusion Counts*

Description

A toy data set of inclusion counts consisting of 3 groups each with 5 samples.

Usage

```
data(exon)
```

Format

numeric matrix

`DB.MAPlot` *Plots Log-Fold Change versus Log-Concentration for Inclusion/Exclusion Data*

Description

M-A Plot

Usage

```
DB.MAPlot( y, m, groups, contrast=c(1,2), de.tags=NULL,
 col="lightgrey", deCol="red", deCex=0.2,
 xlab="Average Over Groups of log2 Mean Total Count",
 ylab="logFC of Odds Ratio", pch=19, cex=0.2,
 panel.last=grid(col = "red", lwd = 0.2) , ylim = c(-15, 15), ...)
```

Arguments

`y`	numeric matrix of inclusion counts.
`m`	numeric matrix of total counts: inclusion + exclusion.
`groups`	vector or factor giving the experimental group/condition for each sample/library.
`contrast`	numeric vector of length 2 specifying which levels of the "groups" factor should be compared.
`de.tags`	rownames for events identified as being differentially expressed.
`col`	color given to the points.
`deCol`	color for the events given in "de.tags".
`deCex`	cex for the events given in "de.tags".
`xlab`	x-label of plot
`ylab`	y-label of plot
`pch`	pch given to the points.
`cex`	cex given to the points.
`panel.last`	an expression to be evaluated after plotting; the default grid() draws a background grid to aid interpretation of the plot.
`ylim`	y-limits for the plot
`...`	further arguments passed to plot().

Details

The total counts are used to determine A, and the log-fold change of the odds ratio is used to determine M. In the case where a group has proportions all 1 or all 0, resulting in an infinite value for M, these points are plotted in orange and away from the main plot. Significant calls made for such events are still colored in red.

Value

A plot to the current device

Author(s)

Sean Ruddy

Examples

```
data(exon)
results.G1G2 <- DBGLM1( counts, offsets, groups)
de.tags.G1G2 <- rownames(results.G1G2$Sig)
DB.MAPlot(counts, offsets, groups, contrast=c(1,2), de.tags=de.tags.G1G2)
```

DBGLM1 — *Double Binomial Generalized Linear Model with Shrinkage of the Dispersion Parameter*

Description

Fits a double binomial GLM with the normalization constant set to 1 and uses shrinkage to obtain estimates of dispersion used for p-value calculation.

Usage

```
DBGLM1( y, m, groups, shrink.method=c(”WEB”,”DEB”),
        contrast=c(1,2), fdr.level=0.05, use.all.groups=TRUE)
```

Arguments

`y` numeric matrix of inclusion counts.

`m` numeric matrix of total counts: inclusion + exclusion.

`groups` vector or factor giving the experimental group/condition for each sample/library.

`shrink.method` for shinkage estimation of the dispersion parameter. "WEB" implements the WEB-Seq method. "DEB" implements the DEB-Seq method. Default is "WEB".

`contrast` numeric vector of length 2 specifying which levels of the "groups" factor should be compared.

`fdr.level` a numeric constant. The FDR level to determine the list of significant events. Default is 0.05.

`use.all.groups` logical. If TRUE, all data in "y" is used to estimate dispersions. If FALSE, only the 2 groups given in "contrasts" are used to estimate dispersions. Only makes a difference if "y" contains more than 2 groups. Default is TRUE.

Details

This function tests for group differences for a two group comparison via a double binomial GLM with the normalization constant set to 1, and utilizes shrinkage estimates of the dispersion parameter for p-value calcuation which is done using a likelihood ratio test. The shrinkage estimates of the dispersion are found according to the selection of "shrink.method". "DEB" implements the DEB-Seq method which uses an empirical bayes strategy to obtain shrunken estimates of the dipersion parameter. "WEB" implements the WEB-Seq method which reparameterizes the empirical bayes strategy in terms of the weight parameter in the weighted liklelihood formulation. An emprical bayes estimate of the weight parameter is found and plugged into the weighted likelihood which is then maximized to obtain shrunken estimates of the dipsersion parameter. DEB-Seq has shown to be more powerful than WEB-Seq; however, WEB-Seq is more conservative thus being more robust against departures from assumptions and therefore maintains the required FDR better in moderate to larger sample sizes.

In the case when "groups" specifies more than two groups the default procedure is to use all groups to calculate the shrunken dispersion estimates. The argument "constrast" is used to specify a particular comparison of two of the groups. If "use.all.groups" is FALSE, only the data for the groups specified in "contrasts" are used to estimate the dispersions.

Value

`Sig`	a matrix consisting of the significant events at the specified FDR level. The matrix contains the proportion estimates, unadjusted and adjusted p-values, the effective sample size, mean total count and dispersion estimates.
`All`	the same matrix as above but consisting of all events.

Author(s)

Sean Ruddy

Examples

```
## Toy exon data: "counts"=inclusion counts, "offsets"=total counts, "groups"=experiment design
  data(exon)

  ## Default will compare "G1" & "G2" using the WEB-Seq method
  ## and uses ALL groups to estimate dispersion
    results.G1G2.WEB <- DBGLM1( counts, offsets, groups)

    # Compare G1 & G3
    results.G1G3.WEB <- DBGLM1( counts, offsets, groups, contrast=c(1,3))

    # Compare G1 & G3. Does not use G2 for dispersion estimation.
   results.G1G3.noG2.WEB <- DBGLM1(counts, offsets, groups, contrast=c(1,3), use.all.groups=FALSE)

    # Global minimum check of the weight parameter estimate in the WEB-Seq method.
    optimPlot( counts, offsets, groups, contrast=c(1,3), use.all.groups=FALSE)

## The DEB-Seq method. Less conservative, more powerful. Very similar ranks to WEB-Seq.
  results.G1G2.DEB <- DBGLM1( counts, offsets, groups, shrink.method="DEB")

## M-A Plot
  WEB.sig <- rownames(results.G1G2.WEB$Sig)
  DB.MAPlot( counts, offsets, groups, de.tags=WEB.sig, main="WEB-Seq")
```

`EstimateDEBDisp` *DEB-Seq: Empirical Bayes Estimates of Dispersion for a Double Binomial Distribution*

Description

Calculation of shrunken dispersion estimates via a 2-parameter empirical bayes method.

Usage

```
EstimateDEBDisp(y,m,groups=NULL,neff=NULL,S=NULL,optim.method=c("BFGS","Nelder-Mead"))
```

Arguments

`y`	numeric matrix of inclusion counts.
`m`	numeric matrix of total counts: inclusion + exclusion.
`groups`	vector or factor giving the experimental group/condition for each sample/library. Default NULL.
`neff`	numeric vector of length equal to the number of rows of "y" where each value is the effective sample size for the event. Default is NULL in which case the effective sample size is calculated within the function.
`S`	numeric vector of length equal to the number of rows of "y" where each value is the random variable for each event whose distribution across exons is gamma. Default is NULL in which case the vector is calculated internally.
`optim.method`	character string determining which optimization routine to use for estimating the parameters of the prior distribution. Default is "BFGS".

Details

The shrunken dispersion estimates are a function of 2 parameters of the generalized beta prime distribution which are estimated via maximum likelihood resulting in empricaly bayes shinkage of the dispersion parameter.

Value

vector of length equal to the number of rows of "y" where each value is the estimate of dispersion.

Author(s)

Sean Ruddy

Examples

```
data(exon)
dispersions <- EstimateDEBDisp( counts, offsets, groups)
```

EstimateWEBDisp *WEB-Seq: Weighted Likelihood Empirical Bayes Estimates of Dispersion for a Double Binomial Distribution*

Description

Calculation of shrunken dispersion estimates via weighted likelihood where the weight parameter is estimated using an empirical bayes strategy.

Usage

```
EstimateWEBDisp(y, m, groups, neff = NULL, S = NULL)
```

Arguments

`y` numeric matrix of inclusion counts.

`m` numeric matrix of total counts: inclusion + exclusion.

`groups` vector or factor giving the experimental group/condition for each sample/library.

`neff` numeric vector of length equal to the number of rows of "y" where each value is the effective sample size for the event. Default is NULL in which case the effective sample size is calculated within the function.

`S` numeric vector of length equal to the number of rows of "y" where each value is the random variable for each event whose distribution across exons is gamma. Default is NULL in which case the vector is calculated internally.

Details

Shrunken dispersion estimates are obtained by maximizing the weighted sum of the likelihood for a given event and the sum of likelihoods for all events, the common likelihood. The weight given to the common likelihood is estimated via empirical bayes.

Value

vector of length equal to the number of rows of "y" where each value is the estimate of dispersion.

Author(s)

Sean Ruddy

Examples

```
data(exon)
dispersions <- EstimateWEBDisp( counts, offsets, groups)
```

`exon` *Toy Exon Inclusion and Total Counts Used for Examples*

Description

A toy RNA-Seq count data set consisting of 3 groups of 5 samples each. Goal is to detect differential exon usage between groups.

Usage

```
data(exon)
```

`groups` *Group Structure of the Toy Data Set*

Description

A vector signifying which samples of the toy exon data set belong to which of the 3 groups.

Usage

```
data(exon)
```

Format

character vector

`offsets` *Exon Total Counts*

Description

A toy data set of Total (inclusion+exclusion) counts consisting of 3 groups each with 5 samples.

Usage

```
data(exon)
```

Format

numeric matrix

`optimPlot` *Plot the WEB-Seq Maximum Likelihood Solution for the Weight Parameter in the Weighted Likelihood*

Description

The MLE solution is signified on a plot as the minimum of the negative log likelihood of the generalized beta prime distribution, parameterized in terms of the weight parameter.

Usage

```
optimPlot(y, m, groups, contrast=c(1,2), use.all.groups=TRUE,...)
```

Arguments

`y` numeric matrix of inclusion counts.

`m` numeric matrix of total counts: inclusion + exclusion.

`groups` vector or factor giving the experimental group/condition for each sample/library.

`contrast` numeric vector of length 2 specifying which levels of the "groups" factor should be compared. This is only relevant if "use.all.groups" is FALSE.

`use.all.groups` logical. If TRUE, all data in "y" is used to estimate dispersions. If FALSE, only the 2 groups given in "contrasts" are used to estimate dispersions. Only makes a difference if "y" contains more than 2 groups. Default is TRUE.

`...` further arguments passes to plot()

Details

The MLE estimate of the (transformed) weight parameter in the WEB-Seq method is checked to be a true global minimum of the negative log likelihood of the generalized beta prime distribution. The weight parameter is transformed from an infinite range to the (0,1) range before optimization and this is the range on which the estimate is checked.

Value

A plot to the current device

Author(s)

Sean Ruddy

Examples

```
data(exon)
# If all groups were used to estimate dispersions
  optimPlot(counts, offsets, groups)
# If only the 2 groups being compared were used to estimate dispersions
  optimPlot(counts, offsets, groups, contrast=c(1,3), use.all.groups=FALSE)
```

Index