Package 'cqn'

October 9, 2013

Version 1.6.0

Title Conditional quantile normalization

Description A normalization tool for RNA-
Seq data, implementing the conditional quantile normalization method.

Author Jean (Zhijin) Wu, Kasper Daniel Hansen

Maintainer Kasper Daniel Hansen <khansen@jhsph.edu>

Depends R (>= 2.10.0), mclust, nor1mix, stats, preprocessCore,splines, quantreg

Imports splines

Suggests scales, edgeR

License Artistic-2.0

LazyLoad yes

biocViews RNAseq, Preprocessing, DifferentialExpression

R topics documented:

cqn *CQN (conditional quantile normalization) for RNA-Seq data*

Description

This function implements CQN (conditional quantile normalization) for RNA-Seq data.

Usage

```
cqn(counts, x, lengths, sizeFactors = NULL, subindex = NULL, tau = 0.5, sqn = TRUE,
    lengthMethod = c("smooth", "fixed"), verbose = FALSE)
## S3 method for class 'cqn'
print(x, ...)
```

Arguments

counts	An object that can be coerced to a matrix of region by sample counts. Ought to have integer values.
x	This is a covariate whose systematic influence on the counts will be removed. Typically the GC content. Has to have the same length as the number of rows of counts.
lengths	The lengths (in bp) of the regions in counts. Has to have the same length as the number of rows of counts.
sizeFactors	An optional vector of sizeFactors, ie. the sequencing effort of the various samples. If NULL this is calculated as the column sums of counts.
subindex	An optional vector of indices into the rows of counts. If not given, this becomes the indices of genes with row means of counts greater then 50.
tau	This argument is passed to rq, it indicates what quantile is being fit. The default should only be changed by expert users..
sqn	This argument indicates whether the residuals from the systematic fit are (subset) quantile normalized. The default should only be changed by expert users.
lengthMethod	Should length enter the model as a smooth function or not.
verbose	Is the function verbose?
...	Not used.

Details

These functions implement the CQN (conditional quantile normalization) for RNA-Seq data. The functions remove a single systematic effect, contained in the argument x, which will typicall be GC content. The effect of lengths will either be modelled as a smooth function (which we recommend), if you are using lengthMethod = "smooth" or as an offset (equivalent to modelling using RPKMs), if you are using lengthMethod = "fixed". Length can be complete removed from the model by having lengthMethod = "fixed" and setting all lengths to 1000.

Final corrected values are equal to value$y + value$offset.

Value

A `list` with the following components

`counts`	The value of argument `counts`.
`x`	The value of argument `x`.
`lengths`	The value of argument `lengths`.
`sizeFactors`	The value of argument `sizeFactors`. In case the argument was `NULL`, this is the value used internally.
`subindex`	The value of argument `subindex`. In case the argument was `NULL`, this is the value used internally.
`y`	The dependent value used in the systematic effect fit. Equal to log2 tranformed reads per millions.
`offset`	The estimated offset.
`offset0`	A single number used internally for identifiability.
`glm.offset`	An offset useful for supplying to a GLM type model function. It is on the natural log scale and includes correcting for sizeFactors.
`func1`	The estimated effect of function 1 (argument `x`). This is a matrix of function values on a grid. Columns are samples and rows are grid points.
`grid1`	The grid points on which function 1 (argument `x`) was evaluated.
`knots1`	The knots used for function 1 (argument `x`).
`func2`	The estimated effect of function 2 (lengths). This is a matrix of function values on a grid. Columns are samples and rows are grid points.
`grid2`	The grid points on which function 2 (lengths) was evaluated.
`knots2`	The knots used for function 2 (lengths).
`call`	The call.

Note

Internally, the function uses a custom implementation of subset quantile normalization, contained in the (not exported) `SQN2` function.

Author(s)

Kasper Daniel Hansen, Zhijin Wu

References

KD Hansen, RA Irizarry, and Z Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 2012 vol. 13(2) pp. 204-216.

See Also

The package vignette.

Examples

```
data(montgomery.subset)
data(sizeFactors.subset)
data(uCovar)
cqn.subset <- cqn(montgomery.subset, lengths = uCovar$length,
                  x = uCovar$gccontent, sizeFactors = sizeFactors.subset,
                  verbose = TRUE)
```

`cqnplot` *Plot the systematic effect estimated as part of a CQN normalization.~*

Description

This function plots the estimated systematic effect which are removed suring CQN normalization.

Usage

```
cqnplot(x, n = 1, col = "grey60", ylab = "QR fit", xlab = "", type = "l", lty = 1, ...)
```

Arguments

`x`	The result of a call to `cqn`; an object of class `cqn`.
`n`	Which systematic effect is plotted.
`col`	A vector of colors, as in `plot`.
`ylab`	y-label as in `plot`.
`xlab`	x-label as in `plot`.
`type`	type, as in `plot`.
`lty`	line type, as in `plot`.
`...`	These arguments are passed to `matplot`

Value

This function is invoked for its side effect.

Author(s)

Kasper Daniel Hansen

Examples

```
data(montgomery.subset)
data(sizeFactors.subset)
data(uCovar)
cqn.subset <- cqn(montgomery.subset, lengths = uCovar$length,
                  x = uCovar$gccontent, sizeFactors = sizeFactors.subset,
                  verbose = TRUE)
cqnplot(cqn.subset, n = 1)
```

`montgomery.subset` *Mongtomery RNA-seq data.*

Description

A gene by sample count matrix for 10 samples from from Montgomery et al. Also included is information about these genes (length and gc content) as well as sequencing depth for each of the samples.

Usage

```
data(montgomery.subset)
data(sizeFactors.subset)
data(uCovar)
```

Format

`montgomery.subset` is a data frame with 23552 observations on 10 different samples, the column names are the sample ids. `sizeFactors.subset` a a named vector of length 10 containing the number of mapped reads for each of the 10 samples. `uCovar` is a data frame with 23552 observations on 2 different covariates: gc content and genic length in bp.

Details

Gene models are union models based on Ensembl 61. These gene models were constructed using Genominator. Genes that have zero counts in all 10 samples were excluded.

References

SB Montgomery, M Sammeth, M Gutierrez-Arcelus, RP Lach, C Ingle, J Nisbett, R Guigo, ET Dermitzakis, (2010) "Transcriptome genetics using second generation sequencing in a Caucasian population". Nature 464(7289), 773-777.

Index