Package 'tweeDEseqCountData'

July 8, 2025

Title RNA-seq count data employed in the vignette of the tweeDEseq package

Description RNA-seq count data from Pickrell et al. (2010) employed to illustrate the use of the Poisson-Tweedie family of distributions with the tweeDEseq package.

Version 1.46.0

Depends Biobase, R (>= 4.3.0)

biocViews Genome, Homo_sapiens_Data, RNASeqData

Suggests knitr, BiocStyle, rmarkdown

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.1.2

VignetteBuilder knitr

LazyLoad yes

BugReports https://github.com/isglobal-brge/tweeDEseqCountData/issues

URL https://github.com/isglobal-brge/tweeDEseqCountData/

git_url https://git.bioconductor.org/packages/tweeDEseqCountData

git_branch RELEASE_3_21

git_last_commit bf1977b

git_last_commit_date 2025-04-15

Repository Bioconductor 3.21

Date/Publication 2025-07-08

Author Dolors Pelegri-Siso [aut, cre] (ORCID: <https://orcid.org/0000-0002-5993-3003>),
Juan R. Gonzalez [aut] (ORCID: <https://orcid.org/0000-0003-3267-2146>),
Mikel Esnaola [aut],
Robert Castelo [aut]

Maintainer Dolors Pelegri-Siso <dolors.pelegri@isglobal.org>

Contents

`tweeDEseqCountData-package`

RNA-seq count data employed in the vignette of the tweeDEseq package

---

Description

RNA-seq tables of counts employed in the vignette of the `tweeDEseq` package (Esnaola et al., submitted). These three data sets were downloaded from the ReCount repository at `http://bowtie-bio.sourceforge.net/recount` and contain tables of counts from RNA-seq experiments by Cheung et al. (2010), Montgomery et al. (2010) and Pickrell et al. (2010). The raw RNA-seq data was pre-processed according to the procedures described by Frazee et al. (2011). Please check the individual help pages of each data set for further details.

Value

void

Source

V.G. Cheung, R.R. Nayak, I.X. Wang, S. Elwyn, S.M. Cousins, M. Morley and R.S. Spielman. *Plos Biology*, 8(9), pii:e1000480, 2010.

A.C. Frazee, B. Langmead and J.T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449, 2011.

S.B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R.P. Lach, C. Ingle, J. Nisbett, R. Guigo and E.T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:773-777, 2010.

J.K. Pickrell, J.C. Marioni, A.A Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard. *Nature*, 464:768-772, 2010.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication, *submitted*.

See Also

annotEnsembl63 cheung montgomery pickrell genderGenes hkGenes

---

annotEnsembl63 *Annotation data from Ensemble Release 63*

Description

Annotation data for the human genes forming the tables of counts in pickrell.eset, montgomery.eset and cheung.eset.

Usage

```
data(annotEnsembl63)
```

Format

Symbol: gene symbol according to the HUGO Gene Nomenclature Committee (HGNC). Chr: chromosome. Start: start chromosomal position (Human genome version GRCh37). End: end chromosomal position (Human genome version GRCh37). EntrezID: Entrez gene identifier. Description: Short description of the gene. Length: Length of the longest cDNA of this gene. GCcontent: G+C content of the longest cDNA of this gene.

Details

Data for all columns except Length and GCcontent was retrieved from Ensembl release 63 using the biomaRt package. Data in columns Length and GCcontent was obtained by downloading the set of Ensembl Release 63 human cDNAs at ftp://ftp.ensembl.org/pub/release-63/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh37.63.cdna.all.fa.gz and selecting the longest cDNA for each Ensembl Gene from which length and G+C content was calculated.

Value

annotEnsembl63 dataset

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, and J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

See Also

pickrell pickrellNorm montgomery genderGenes hkGenes

Examples

```
data(annotEnsembl63)
dim(annotEnsembl63)
head(annotEnsembl63)
```

cheung *RNA-seq count data from Cheung et al. (2010)*

Description

ExpressionSet object containing RNA-seq count data from lymphoblastoid cell lines from 41 unrelated Caucasian individuals of European descent. These count data are employed in the vignette of the package tweeDEseq Esnaola et al. (submitted). The original experimental data was published by Cheung et al. (2010) and the table of counts in this ExpressionSet object corresponds to the one in the ReCount repository available at http://bowtie-bio.sourceforge.net/recount. Details on the pre-processing steps to obtain this table of counts from the raw reads of Cheung et al. (2010) are provided on that website and in the publication by Frazee et al. (2011).

Usage

```
data(cheung)
```

Format

cheung.eset: ExpressionSet object containing read counts for 52,580 Ensembl genes for each of the 41 Caucasian individuals of European descent.

Details

The table of counts is stored in the AssayData slot of an ExpressionSet object called cheung.eset whose phenotypic data contains the gender of each individual, among other bits of information.

Value

cheung dataset

Source

V.G. Cheung, R.R. Nayak, I.X. Wang, S. Elwyn, S.M. Cousins, M. Morley and R.S. Spielman. *Plos Biology*, 8(9), pii:e1000480, 2010.

A.C. Frazee, B. Langmead and J.T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449, 2011.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

See Also

annotEnsembl63 pickrell cheung genderGenes hkGenes

Examples

```
suppressMessages(library(Biobase))
data(cheung)
cheung.eset
table(cheung.eset$gender)
```

---

commonPickrell1Huang *Matching microarray and RNA-seq data from human lymphoblastoid cell lines*

Description

ExpressionSet objects containing microarray and RNA-seq count data for 36 matching samples of lymphoblastoid cell lines derived from unrelated Nigerian HapMap individuals. These microarray and count data are employed in the vignette of the package tweeDEseq and supporting analysis scripts of the article by Esnaola et al. (submitted). The original experimental data was published by Huang et al. (2007) and Pickrell et al. (2010).

The loaded objects can be divided into two classes of data. One corresponding to gene expression data from microarray and RNA-seq with matching samples, and the other where not only samples, but also genes match between the two gene expression data matrices, thus being a subset of the former.

Usage

```
data(commonPickrell1Huang)
```

Format

huangArrayRMAnoBatchCommonSamples.eset: ExpressionSet object containing filtered, normalized and batch-removed microarray expression (RMA) values for 16,323 Ensembl Gene identifiers from 36 unrelated Nigerian individuals.

huangArrayRMAnoBatchCommon.eset: ExpressionSet object containing filtered, normalized and batch-removed microarray expression (RMA) values for 15,194 Ensembl Gene identifiers from 36 unrelated Nigerian individuals.

pickrell1countsCommonSamples.eset: ExpressionSet object containing filtered RNA-seq read counts for 27,438 Ensembl Gene identifiers from 36 unrelated Nigerian individuals. This table of counts corresponds to RNA-seq data published by Pickrell et al. (2010) and processed by the pipeline described by Esnaola et al. (submitted).

pickrell1countsCommon.eset: ExpressionSet object containing filtered RNA-seq read counts for 15,194 Ensembl Gene identifiers from 36 unrelated Nigerian individuals. This table of counts corresponds to RNA-seq data published by Pickrell et al. (2010) and processed by the pipeline described by Esnaola et al. (submitted).

pickrell1countsCQNcommonSamples.eset: ExpressionSet object containing the table of read counts in pickrell1countsCommonSamples.eset normalized using the package cqn. The transformation from log CPM values and their offsets, as produced by cqn, into this table of normalized counts was done with the function normalizeCounts() of the tweeDEseq package.

pickrell1countsCQNcommon.eset: ExpressionSet object containing the table of read counts in pickrell1countsCommon.eset normalized using the package cqn. The transformation from log CPM values and their offsets, as produced by cqn, into this table of normalized counts was done with the function normalizeCounts() of the tweeDEseq package.

cqnNormCommonSamples: list object output by the cqn() function from the cqn package when normalizing the RNA-seq data in pickrell1countsCommonSamples.eset and used by the function normalizeCounts() from the tweeDEseq package to obtain the normalized count expression data matrix in the ExpressionSet object pickrell1countsCQNcommonSamples.eset. This object is necessary when using DE detection methods, such as edgeR, that employ the offsets given by cqn and the raw counts in pickrell1countsCommonSamples.eset instead of the transformed normalized counts in pickrell1countsCQNcommonSamples.eset.

cqnNormCommon: list object output by the cqn() function from the cqn package when normalizing the RNA-seq data in pickrell1countsCommon.eset and used by the function normalizeCounts() from the tweeDEseq package to obtain the normalized count expression data matrix in the ExpressionSet object pickrell1countsCQNcommon.eset. This object is necessary when using DE detection methods, such as edgeR, that employ the offsets given by cqn and the raw counts in pickrell1countsCommon.eset instead of the transformed normalized counts in pickrell1countsCQNcommon.eset.

Details

The microarray data was processed from the raw CEL files available at http://www.ncbi.nlm.nih.gov/geo under accession GSE7792. First, only Yoruba samples were considered. Second, data was processed using the Bioconductor oligo package. Quality assessment was performed by calculating NUSE and RLE diagnostics (Bolstad et al., 2005) and discarding those samples that either of the two reported diagnostics was considered below a minimum quality threshold. Third, using the RMA algorithm (Irizarry et al., 2003) implemented in the rma() function from the oligo package with argument target="core", expression values were background corrected, normalized and summarized into Affymetrix transcript clusters. Fourth, most samples formed part of family trios and only samples belonging to father or mother were kept. Fifth, using the getNetAffx() function from the oligo package, Ensembl Transcript identifiers well obtained for each Affymetrix transcript cluster identifier. Sixth, using the bioconductor package biomaRt, Ensembl Transcript identifiers were translated into Ensembl Gene identifiers, resolving multiple assignments by keeping the Ensembl Gene identifier that had a match in the Ensembl Gene identifiers forming the table of counts of the Pickrell et al. (2010) RNA-seq data, or choosing one arbitrarily, otherwise. Seven, duplicated assignments of the same Ensembl Gene identifier to multiple Affymetrix transcript cluster identifiers were resolved by keeping the transcript cluster with largest expression variability measured by its interquartile range (IQR).

At this point an expression data matrix of 16,323 Ensembl Genes by 74 samples was obtained and using the scanning date of each CEL file, samples were grouped into 4 balanced batches stored in the phenotypic variable Batch within the resulting ExpressionSet. Batch effect was removed by using the QR-decomposition method implemented in the removeBatchEffect() function from the package limma while keeping the sex-specific expression effect by setting the gender sample indicator variable within the design matrix argument. Finally, samples were further filtered to match those from the RNA-seq table of counts and this resulted into the gene expression data in huangArrayRMAnoBatchCommonSamples.eset, while an additional filtering of genes to match those from the RNA-seq table of counts resulted into the gene expression data in huangArrayRMAnoBatchCommon.

The RNA-seq data was obtained by the pipeline described in Esnaola et al. (submitted) from the raw reads deposited at http://eqtl.uchicago.edu/RNA_Seq_data/unmapped_reads. The re-

sulting table of counts is available in this data package as an ExpressionSet object under the name pickrell1.eset and consists of 38,415 Ensembl Genes by 69 samples. This table of counts was first filtered to remove genes with very low expression levels by keeping only those with a minimum average of 0.1 counts per million. Second, we further filtered this table of counts in order to match the samples obtained after processing the LCL microarray data from Huang et al. (2007). At this point we also generated a second gene expression data set by further filtering genes matching those from the LCL microarray data from Huang et al. (2007). Third, we normalized both expression data sets adjusting for gene length and G+C content using the Bioconductor package cqn (Hansen et al., 2012). The corresponding gene length and G+C content information was obtained from the data stored in the annotEnsembl63 data.frame object.

The resulting pair of sample-matching expression data matrices have 16,323 (microarray) and 27,438 (RNA-seq) genes by 36 samples, while the resulting pair of sample and gene-matching expression data matrices have 15,194 genes by 36 samples.

Value

commonPickrell1Huang dataset

Source

R.S. Huang, S. Duan, W.K. Bleibel, E.O. Kistner, W. Zhang, T.A. Clark, T.X. Chen, A.C. Schweitzer, J.E. Blume, N.J. Cox and M.E. Dolan, *Proc. Natl. Acad. Sci. USA*, 104(23):9758-9763, 2007.

J.K. Pickrell, J.C. Marioni, A.A Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard, *Nature*, 464:768-772, 2010.

References

B.M. Bolstad, F. Collin, K. Brettschneider, L. Simpson, R.A. Irizarry, and T.P. Speed. Quality assessment of Affymetrix GeneChip data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pg. 33–48, Springer, 2005.

K.D. Hansen, R.A. Irizarry and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 2012.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. A flexible count data model to fit the wide diversity of expression profiles arising from extensively replicated RNA-seq experiments, *submitted*.

See Also

pickrell1 annotEnsembl63

Examples

```
suppressMessages(library(Biobase))
data(commonPickrell1Huang)
dim(huangArrayRMAnoBatchCommonSamples.eset)
```

```
dim(pickrell1countsCQNcommonSamples.eset)
table(huangArrayRMAnoBatchCommonSamples.eset$Gender)
table(pickrell1countsCQNcommonSamples.eset$Gender)
dim(huangArrayRMAnoBatchCommon.eset)
dim(pickrell1countsCQNcommon.eset)
table(huangArrayRMAnoBatchCommon.eset$Gender)
table(pickrell1countsCQNcommon.eset$Gender)
```

---

genderGenes *Genes with documented sex-specific expression*

Description

Genes with documented sex-specific expression and occurring within the set of genes that form the tables of counts in pickrell.eset, montgomery.eset and cheung.eset.

Usage

```
data(genderGenes)
```

Format

msYgenes: Ensembl gene identifiers from genes belonging to the male-specific region of chromosome Y (Skaletsky et al., 2003).

XiEgenes: Ensembl gene identifiers from genes located in the X chromosome and which have been reported to escape X-inactivation.

Details

These two lists of genes form a gold-standard set of genes with documented sex-specific expression which have been employed in the assessment of the method for differential expression analysis implemented in the tweeDEseq package (Esnaola et al., *submitted*). Both gene lists are restricted to genes occuring within the set of genes that form the tables of counts in pickrell.eset, montgomery.eset and cheung.eset.

Value

genderGenes dataset

Source

H.S. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423:825–837, 2003.

L. Carrel and H.F. Willard. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434:400–404, 2005.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

See Also

annotEnsembl63 pickrell pickrellNorm montgomery hkGenes

Examples

```
data(genderGenes)
length(msYgenes)
length(XiEgenes)
```

---

hkGenes *Housekeeping Genes from Eisenberg and Levanon (2003)*

Description

Housekeeping genes reported by Eisenberg and Levanon (2003) and occurring within the set of genes that form the tables of counts in pickrell.eset, montgomery.eset and cheung.eset in this experimental data package.

Usage

```
data(hkGenes)
```

Format

hkGenes: Ensembl gene identifiers from the list of housekeeping genes occuring within the set of genes that form the tables of counts in pickrell.eset, montgomery.eset and cheung.eset in this experimental data package.

Details

This list of genes has been derived from mapping the original list in http://www.cgen.com/supp_info/Housekeeping_genes.html to Ensemble Gene identifiers using the org.Hs.eg.db package. This list of housekeeping genes has been employed to compare count data distributions from genes with different expression dynamics in (Esnaola et al., *submitted*) and is restricted to genes occuring within the set of genes that form the tables of counts in pickrell.eset, montgomery.eset and cheung.eset in this experimental data package.

Value

hkGenes dataset

Source

E. Eisenberg, and E.Y. Levanon. Human housekeeping genes are compact. *Trends Genet*, 19(7):362–365, 2003.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, and J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

See Also

`annotEnsembl63` `pickrell` `pickrellNorm` `montgomery` `genderGenes`

Examples

```
data(hkGenes)
length(hkGenes)
head(hkGenes)
```

`montgomery` *RNA-seq count data from Montgomery et al. (2010)*

Description

`ExpressionSet` object containing RNA-seq count data from lymphoblastoid cell lines from 60 unrelated Caucasian individuals of European descent. These count data are employed in the vignette of the package `tweeDEseq` Esnaola et al. (submitted). The original experimental data was published by Montgomery et al. (2010) and the table of counts in this `ExpressionSet` object corresponds to the one in the ReCount repository available at `http://bowtie-bio.sourceforge.net/recount`. Details on the pre-processing steps to obtain this table of counts from the raw reads of Montgomery et al. (2010) are provided on that website and in the publication by Frazee et al. (2011).

Usage

```
data(montgomery)
```

Format

`montgomery.eset`: `ExpressionSet` object containing read counts for 52,580 Ensembl genes for each of the 60 Caucasian individuals of European descent.

Details

The table of counts is stored in the `AssayData` slot of an `ExpressionSet` object called `montgomery.eset` whose phenotypic data contains the gender of each individual, among other bits of information.

Value

montgomery dataset

Source

S.B. Montgomery, M. Sammeth, M. Gutierrez-Arcelus, R.P. Lach, C. Ingle, J. Nisbett, R. Guigo and E.T. Dermitzakis. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464:773-777, 2010.

A.C. Frazee, B. Langmead and J.T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449, 2011.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

See Also

annotEnsembl63 pickrell cheung genderGenes hkGenes

Examples

```
suppressMessages(library(Biobase))
data(montgomery)
montgomery.eset
table(montgomery.eset$gender)
```

---

pickrell *RNA-seq count data from Pickrell et al. (2010)*

Description

ExpressionSet objects containing RNA-seq count data from lymphoblastoid cell lines from 69 unrelated Nigerian individuals. These count data are employed in the vignette of the package tweeDEseq Esnaola et al. (submitted). The original experimental data was published by Pickrell et al. (2010). The table of counts in pickrell.eset corresponds to the one in the ReCount repository available at http://bowtie-bio.sourceforge.net/recount. Details on the pre-processing steps to obtain this table of counts from the raw reads of Pickrell et al. (2010) are provided on that website and in the publication by Frazee et al. (2011). The other object pickrellNorm.eset contains the corresponding filtered and normalized table of counts.

The table of counts in pickrell1.eset was obtained by Esnaola et al. (2010) by the pre-processing steps described on that article and pickrell1Norm.eset contains the corresponding filtered and normalized table of counts.

Usage

```
data(pickrell)
```

Format

`pickrell.eset`: `ExpressionSet` object containing read counts for 52,580 Ensembl genes for each of the 69 Nigerian individuals. `pickrellNorm.eset`: `ExpressionSet` object containing filtered and normalized read counts for 10,231 Ensembl genes for each of the 69 Nigerian individuals. `pickrell1.eset`: `ExpressionSet` object containing read counts for 38,415 Ensembl genes for each of the 69 Nigerian individuals. `pickrell1Norm.eset`: `ExpressionSet` object containing filtered and normalized read counts for 22,060 Ensembl genes for each of the 69 Nigerian individuals.

Details

These tables of counts are stored in the `AssayData` slot of the previously enumerated `ExpressionSet` objects whose phenotypic data contains the gender of each individual, among other bits of information. The filtered and normalized table of counts was obtained from the raw counts in `pickrell.eset` and `pickrell1.eset` by first removing genes with less than 0.5 cpm (counts per million reads) in all samples but one and then applying the conditional quantile normalization procedure by Hansen et al. (2011).

Value

pickrell dataset

Source

J.K. Pickrell, J.C. Marioni, A.A Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard *Nature*, 464:768-772, 2010.

A.C. Frazee, B. Langmead and J.T. Leek. ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12:449, 2011.

References

M. Esnaola, P. Puig, D. Gonzalez, R. Castelo, J.R. Gonzalez. Gene-specific count data distributions are required in RNA-seq experiments with extensive replication. Submitted.

K.D. Hansen, R.A. Irizarry and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. Johns Hopkins University, Dept. of Biostatistics Working Papers, Paper 227. (http://www.bepress/com/jhubiostat/paper227).

See Also

`annotEnsembl63` `montgomery` `cheung` `genderGenes` `hkGenes`

Examples

```
suppressMessages(library(Biobase))
data(pickrell)
pickrell.eset
table(pickrell.eset$gender)
data(pickrellNorm)
pickrellNorm.eset
```

Index