Package 'seqc'

October 7, 2015

Type Package

Title RNA-seq data generated from SEQC (MAQC-III) study

Version 1.2.0

Date 2014-11-19

Author Yang Liao and Wei Shi with contributions from Steve Lianoglou

Maintainer Yang Liao <liao@wehi.edu.au> and Wei Shi <shi@wehi.edu.au>

Description The SEQC/MAQC-III Consortium has produced benchmark RNA-seq data for the assessment of RNA sequencing technologies and data analysis methods (Nat Biotechnol, 2014). Billions of sequence reads have been generated from ten different sequencing sites. This package contains the summarized read count data for ~2000 sequencing libraries. It also includes all the exon-exon junctions discovered from the study. TaqMan RT-PCR data for ~1000 genes and ERCC spike-in sequence data are included in this package as well.

URL http://bioconductor.org/packages/release/data/experiment/html/seqc.html

License GPL-3

LazyData yes

biocViews ExperimentData, RNASeqData, qPCRData, SequencingData

Depends R (>= 2.10)

NeedsCompilation no

R topics documented:

SEQC-package 2
SEQC-features 3
SEQC-junction-tables 4
seqc.eSet 5
seqc.samples 6
TaqMan-RTPCR 7

Index **8**

SEQC-package *RNA-seq data employed in the Sequence Quality Control (SEQC) Project*

Description

This package contains 69 data frames in total, in which there are 22 read counting tables, 46 exon-exon junction tables and a gene intensity table. All these tables were generated from the RNA-Seq libraries of human brain RNA and universal human reference RNA samples employed in the SEQC/MAQC study (Nat Biotech, 2014), except that the gene intensity table was derived by using the TaqMan RT-PCR technology on the same samples. Three platforms (Illumina, Roche 454 and SOLiD) were examined in the project at twelve different sites (the Illumina platform at AGR, BGI, CNL, COH, MAY and NVS, the Roche 454 platform at MGP, NYU and SQW, the SOLiD platform at LIV, NWU, PSU and SQW), and the libraries were mapped by using Subread-1.3.0. The reads were then assigned to the annotated genes by using featureCounts.

Please read the vignette document before using this package by using R command "vignette('seqc')".

Details

Package:	SEQC
Type:	Package
Version:	1.0
Date:	2014-08-28
License:	The GNU General Public License

The 22 read counting tables provide the numbers of single-end reads or paired-end fragments assigned to human genes for each library. Two annotation sets were used in read/fragment assignment: the RefSeq annotations and the AceView annotations, both from NCBI. In each read counting table, the first four columns record the entrez IDs, the symbols and the summed exon lengths of the genes, plus a column of Boolean values indicating if the gene is an ERCC spike-in sequence, and the following columns each corresponds to a RNA-seq library. These tables are named as (Platform)_(Annotation)_gene_(Site).

The 46 exon-exon junction tables are named as (Platform)_junction_(Site)_(Sample). The four columns in each table are the chromosome names, the two ends of the junction and the number of supporting reads to the junctions. It is noticed that a read may supporting zero, one or multiple exon-exon junctions depending on the number of exon-exon junctions in it.

The TaqMan RT-PCR results on the SEQC samples are also provided in the package. The data object is named as "taqman".

Author(s)

Yang Liao and Wei Shi

References

Liao Y, Smyth GK and Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108, 2013

Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30, 2014

Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF, et al. (SEQC/MAQC-III Consortium). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nature Biotechnology, Published online 24 August 2014

See Also

Rsubread, voom

Examples

```
ls(2)
ILM_aceview_gene_AGR[1:10,1:10]
taqman[1:10, 1:10]
```

`SEQC-features` *Read count tables from the SEQC project*

Description

This package contains 22 read counting tables (data frames). All these tables were generated from the RNA-Seq libraries from the SEQC project. Three platforms (Illumina, Roche 454 and SOLiD) were examined in the project at twelve different sites, and the libraries were mapped by using Subread. The reads (in the single-end libraries) or fragments (in the paired-end libraries) were then assigned to the annotated genes by using the featureCounts program on the single-end mode or the paired-end mode accordingly.

All the platforms and sequencing sites are listed in Table 1 in the vignette document of this package. This table also describes if each library is either single-end or paired-end. Please read the vignette document before using this package by using R command "vignette('seqc')".

Details

The 22 read counting tables are data frames providing the numbers of reads or fragments overlapping with the exon regions of each of the known human genes in the annotations. Two annotation sets were used in this package: the RefSeq annotations and the AceView annotations, both from the NCBI. The 22 data frames are named as to its origin; the naming scheme is (Platform)_(Annotation)_gene_(Site). For example, "LIF_refseq_gene_NMU" contains the read counts from all samples sequenced at NMU by using the SOLiD systems; the mapping results were then assigned to the RefSeq annotations.

Each of the data frames has many columns. The first three columns give the entrez ids, symbols and lengths of the genes, followed by a column of Boolean values indicating if each gene is actually an

ERCC RNA Spike-In. After the fourth column are the columns for the read/fragment counts in the libraries; the column names are concatenated from three or four parts: (Sample)_(Replicate)_(Lane)_(FlowCell) for the tables generated by the Illumina platform or the SOLiD systems, or (Sample)_(Replicate)_(Region) for the tables generated by the Roche 454 platform. For example, column "ILM_aceview_gene_AGR$B_3_L02_FlowCellA" is a column of fragment counts on the third replicate of sample B, sequenced at AGR by using the Illumina HiSeq 2000 device (the second lane and flow cell A); the fragments were then assigned to the AceView annotations.

Author(s)

Yang Liao and Wei Shi

References

Liao Y, Smyth GK and Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108, 2013

Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7):923-30, 2014

Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF, et al. (SEQC/MAQC-III Consortium). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nature Biotechnology, Published online 24 August 2014

See Also

Rsubread, voom

Examples

```
ls(2)
colnames(ILM_aceview_gene_AGR)
ILM_aceview_gene_AGR[1:10,1:10]
```

SEQC-junction-tables *Junction tables from the SEQC project*

Description

This package contains 46 exon-exon junction tables (data frames). All these tables were generated from the RNA-Seq libraries for the SEQC project. Three platforms were examined in the project at twelve different sites, and the libraries were mapped by using Subread-1.3.0. The exon-exon junctions were further detected from the mapping results by the Subjunc program. After the exon-exon junctions were detected from each library, we combined the junction lists from the same "sample and sequencing site" combinations by adding up the supporting read numbers to the same junctions (i.e., junctions connecting the same exon-pairs); every junction is reported only once in each table.

Please read the vignette document before using this package by using R command "vignette('seqc')".

Details

The 46 exon-exon junction tables are data frames named as (Platform)_junction_(Site)_(Sample). For example, data frame "ILM_junction_BGI_C" contains all junctions detected from the RNA-seq libraries of sample C sequenced at BGI by using the Illumina HiSeq 2000 device.

Each of the 46 data frames has four columns: the chromosome names, the two sides of the junctions and the number of supporting reads to the junctions. More precisely, the two sides of a junction are defined as the chromosomal locations (coordinates starting from one) of the last base in the exon before this junction, and the first base in the exon after this junction. It is also noted that a read may supporting zero, one or multiple exon-exon junctions depending on the number of exon-exon junctions in it.

Author(s)

Yang Liao and Wei Shi

References

Liao Y, Smyth GK and Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108, 2013

Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF, et al. (SEQC/MAQC-III Consortium). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nature Biotechnology, Published online 24 August 2014

See Also

Rsubread, voom

Examples

```
ls(2)
LIF_junction_NWU_B[1:10,]
```

---

seqc.eSet *Create an ExpressionSet of counts for a given subset of the samples provided by the seqc data package*

Description

Create an ExpressionSet object.

Usage

```
seqc.eSet(feature=c('gene', 'junction', 'taqman'),annotation=c('refseq', 'aceview'),
platform=NULL,center=NULL)
```

Arguments

feature	the type of features you want to build an ExpressionSet.
annotation	If feature == 'gene', then this determines which set of gene features you want.
platform	Subset the data based on the platform it was sequenced on. Possible values include “ILM", “ROC" and “LIF".
center	Subset the data based on the center that sequenced the libraries. Possible values include “AGR”, “BGI”, “CNL”, “COH”, “MAY”, and “NVS” for Illumina platform, “LIV”, “NWU”, “PSU” and “SQW” for LifeTech SOLiD platform and “MGP”, “NYU”, “SQW” for Roche 454 platform.

Details

Currently this only works for feature == 'gene'.

Value

An ExpressionSet with the counts from the samples that satisfy the critera set by the function parameters.

Author(s)

Steve Lianoglou

---

seqc.samples *Return a character vector of the available SEQC samples*

Description

Get sample names.

Usage

```
seqc.samples(feature=NULL,annotation=NULL,platform=NULL,center=NULL)
```

Arguments

feature	Should be either ’gene’ or ’junction’.
annotation	Should be either ’refseq’ or ’aceview’.
platform	Subset the data based on the platform it was sequenced on. Possible values include “ILM", “ROC" and “LIF".
center	Subset the data based on the center that sequenced the libraries. Possible values include “AGR”, “BGI”, “CNL”, “COH”, “MAY”, and “NVS” for Illumina platform, “LIV”, “NWU”, “PSU” and “SQW” for LifeTech SOLiD platform and “MGP”, “NYU”, “SQW” for Roche 454 platform.

Value

A character vector of the sample names/objects in the package:seqc namespace.

Author(s)

Steve Lianoglou

TaqMan-RTPCR *TaqMan RT-PCR results from the SEQC project*

Description

RNA samples A, B, C and D were analyzed by using the TaqMan RT-PCR technology in the SEQC project. A data frame, named "taqman", is provided in this package, containing the expression intensity of 1044 selected genes across the 16 replicates of the four samples, derived in the TaqMan RT-PCR analysis. The columns in the data frame include the entrez ids (taqman$EntrezID), the gene symbols (taqman$Symbol) and the columns for the intensity values, named as (Sample).(Sample)(Replicate)_value. For example, column "taqman$C.C1_value" contains the expression intensity of the 1044 genes in the first replicate of sample C.

Please read the vignette document before using this package by using R command "vignette('seqc')".

Author(s)

Yang Liao and Wei Shi

References

Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, Thompson JF, et al. (SEQC/MAQC-III Consortium). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequence Quality Control consortium. Nature Biotechnology, Published online 24 August 2014

See Also

Rsubread, voom

Examples

```
ls(2)
rownames(taqman)
taqman[1:10, 1:10]
```

Index

∗Topic **SEQC**
SEQC-features, 3
SEQC-junction-tables, 4
SEQC-package, 2
TaqMan-RTPCR, 7

ILM_aceview_gene_AGR *(*SEQC-features*)*, 3
ILM_aceview_gene_BGI *(*SEQC-features*)*, 3
ILM_aceview_gene_CNL *(*SEQC-features*)*, 3
ILM_aceview_gene_COH *(*SEQC-features*)*, 3
ILM_aceview_gene_MAY *(*SEQC-features*)*, 3
ILM_aceview_gene_NVS *(*SEQC-features*)*, 3
ILM_junction_AGR_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_AGR_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_AGR_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_AGR_D
*(*SEQC-junction-tables*)*, 4
ILM_junction_BGI_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_BGI_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_BGI_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_BGI_D
*(*SEQC-junction-tables*)*, 4
ILM_junction_CNL_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_CNL_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_CNL_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_CNL_D
*(*SEQC-junction-tables*)*, 4
ILM_junction_COH_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_COH_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_COH_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_COH_D
*(*SEQC-junction-tables*)*, 4
ILM_junction_MAY_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_MAY_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_MAY_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_MAY_D
*(*SEQC-junction-tables*)*, 4
ILM_junction_NVS_A
*(*SEQC-junction-tables*)*, 4
ILM_junction_NVS_B
*(*SEQC-junction-tables*)*, 4
ILM_junction_NVS_C
*(*SEQC-junction-tables*)*, 4
ILM_junction_NVS_D
*(*SEQC-junction-tables*)*, 4
ILM_refseq_gene_AGR *(*SEQC-features*)*, 3
ILM_refseq_gene_BGI *(*SEQC-features*)*, 3
ILM_refseq_gene_CNL *(*SEQC-features*)*, 3
ILM_refseq_gene_COH *(*SEQC-features*)*, 3
ILM_refseq_gene_MAY *(*SEQC-features*)*, 3
ILM_refseq_gene_NVS *(*SEQC-features*)*, 3

LIF_aceview_gene_LIV *(*SEQC-features*)*, 3
LIF_aceview_gene_NWU *(*SEQC-features*)*, 3
LIF_aceview_gene_PSU *(*SEQC-features*)*, 3
LIF_aceview_gene_SQW *(*SEQC-features*)*, 3
LIF_junction_LIV_A
*(*SEQC-junction-tables*)*, 4
LIF_junction_LIV_B
*(*SEQC-junction-tables*)*, 4
LIF_junction_LIV_C
*(*SEQC-junction-tables*)*, 4
LIF_junction_LIV_D
*(*SEQC-junction-tables*)*, 4

LIF_junction_NWU_A
(SEQC-junction-tables), 4
LIF_junction_NWU_B
(SEQC-junction-tables), 4
LIF_junction_NWU_C
(SEQC-junction-tables), 4
LIF_junction_NWU_D
(SEQC-junction-tables), 4
LIF_junction_PSU_A
(SEQC-junction-tables), 4
LIF_junction_PSU_B
(SEQC-junction-tables), 4
LIF_junction_PSU_C
(SEQC-junction-tables), 4
LIF_junction_PSU_D
(SEQC-junction-tables), 4
LIF_junction_SQW_A
(SEQC-junction-tables), 4
LIF_junction_SQW_B
(SEQC-junction-tables), 4
LIF_junction_SQW_C
(SEQC-junction-tables), 4
LIF_junction_SQW_D
(SEQC-junction-tables), 4
LIF_refseq_gene_LIV (SEQC-features), 3
LIF_refseq_gene_NWU (SEQC-features), 3
LIF_refseq_gene_PSU (SEQC-features), 3
LIF_refseq_gene_SQW (SEQC-features), 3

ROC_aceview_gene_MGP (SEQC-features), 3
ROC_aceview_gene_NYU (SEQC-features), 3
ROC_aceview_gene_SQW (SEQC-features), 3
ROC_junction_MGP_A
(SEQC-junction-tables), 4
ROC_junction_MGP_B
(SEQC-junction-tables), 4
ROC_junction_NYU_A
(SEQC-junction-tables), 4
ROC_junction_NYU_B
(SEQC-junction-tables), 4
ROC_junction_SQW_A
(SEQC-junction-tables), 4
ROC_junction_SQW_B
(SEQC-junction-tables), 4
ROC_refseq_gene_MGP (SEQC-features), 3
ROC_refseq_gene_NYU (SEQC-features), 3
ROC_refseq_gene_SQW (SEQC-features), 3

SEQC (SEQC-package), 2
seqc (SEQC-package), 2
SEQC-features, 3
SEQC-junction-tables, 4
SEQC-package, 2
seqc-package (SEQC-package), 2
seqc.eSet, 5
seqc.samples, 6

taqman (TaqMan-RTPCR), 7
TaqMan-RTPCR, 7