Package 'FieldEffectCrc'

July 11, 2024

Version 1.14.0

Date 2020-0920

Title Tumor, tumor-adjacent normal, and healthy colorectal transcriptomes as SummarizedExperiment objects

Depends utils

Suggests knitr, rmarkdown, BiocGenerics, sva, BiocManager

Imports BiocStyle, RUnit, SummarizedExperiment, ExperimentHub (>= 0.99.6), AnnotationHub, DESeq2

Description Processed RNA-seq data for 1,139 human primary colorectal tissue samples across three phenotypes, including tumor, normal adjacent-to-tumor, and healthy, available as Synapse ID syn22237139 on synapse.org. Data have been parsed into SummarizedExperiment objects available via ExperimentHub to facilitate reproducibility and extension of results from Dampier et al. (PMCID: PMC7386360, PMID: 32764205).

License Artistic-2.0

URL http://bioconductor.org/packages/release/bioc/html/FieldEffectCrc.html

biocViews ExperimentData, ReproducibleResearch, Tissue, Homo_sapiens_Data, ColonCancerData, RNASeqData, ExpressionData, ExperimentHub, SequencingData

VignetteBuilder knitr

NeedsCompilation no

BugReports https://github.com/Bioconductor/FieldEffectCrc/

Encoding UTF-8

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

git_branch RELEASE_3_19

git_last_commit 7a7f230

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-07-11

Author Christopher Dampier [aut, cre]
(<https://orcid.org/0000-0003-3099-6462>),
Bioconductor Package Maintainer [ctb]

Maintainer Christopher Dampier <chd5n@virginia.edu>

Contents

FieldEffectCrc-package 2
cohort_A 3
cohort_B 6
cohort_C 8
make_txi 10
reorder_assays 11

Index **13**

`FieldEffectCrc-package`

Tumor, tumor-adjacent normal, and healthy colorectal transcriptomes available as a `SummarizedExperiment` *object*

Description

Processed RNA-seq data for 1,139 human primary colorectal tissue samples across three phenotypes, including tumor, normal adjacent-to-tumor, and healthy, available as Synapse ID syn22237139 hosted on Synapse:

<https://www.synapse.org/#!Synapse:syn22237139/files/>

Data have been parsed into three `SummarizedExperiment` objects available via `ExperimentHub` to facilitate reproducibility and extension of results from Dampier et al. (PMCID: PMC7386360, PMID: 32764205).

Details

Includes data as well as two simple helper functions to execute common manipulations of the data:

- `make_txi`: makes a `tximport`-style list from a `SummarizedExperiment` object for downstream use
- `reorder_assays`: reorders the assays of a `SummarizedExperiment` object so that counts is the first assay

See the package vignette for examples of using these data in differential gene expression analysis.

browseVignettes("FieldEffectCrc")

Details of how the `SummarizedExperiment` objects were created from count matrices are available in the `inst/scripts/` directory of the source package in the `make-data.R` file.

Author(s)

Chris Dampier

Examples

```
## load ExperimentHub package
library(ExperimentHub)

## load hub
hub <- ExperimentHub::ExperimentHub()

## list resources associated with package
x <- ExperimentHub::listResources(hub, "FieldEffectCrc")
x

## query hub for package resources
r <- AnnotationHub::query(hub, c("FieldEffectCrc"))
r

## download selected resource by position
## output is SummarizedExperiment
## cohort C is smaller than cohort A and quicker to load
data <- r[[3]]
data

## download selected resource by hub id
## output is SummarizedExperiment
data <- r[["EH3526"]]
data

## download selected resource by filter
## output is list of SummarizedExperiments
data <- ExperimentHub::loadResources(hub, "FieldEffectCrc", "cohort C")
data

## make a txi object
se <- r[["EH3526"]]
txi <- make_txi(se)
str(txi)

## move counts to the first assay slot
se <- r[["EH3526"]]
se <- reorder_assays(se)
se
```

---

cohort_A *cohort A from Dampier et al.*

Description

Salmon-generated transcript-level abundance estimates summarized to gene level using `tximport` along with raw counts, gene lengths, and clinical annotations for 834 human primary colorectal tissue samples across three phenotypes, including tumor, normal adjacent-to-tumor, and healthy, represented as a `SummarizedExperiment`. Abundance estimates derived from paired-end RNA-seq.

Format

A `SummarizedExperiment` object containing 3 `assays` of matrices, each 37,361 rows x 834 columns. Each row is a gene and each column is a sample.

The `SummarizedExperiment` object also includes a `colData` `S4Vectors::DFrame` object with 834 rows and 27 columns. Each row is a sample and each column is a field. The fields are described below.

- dirName: name of directory into which raw data for sample was downloaded, serves as unique identifier
- projId: NCBI BioProject identifier for projects registered in the BioProject database, or common name of projects listed in other databases
- subId: subject identifier
- sampId: sample identifier
- sampType: sample type, which indicates the phenotype of the sample
- dist_cm: relative distance in centimeters from tumor from which given sample was obtained, NA for healthy samples and tumor-adjacent samples without measurements, 0 for tumor samples
- sex: reported sex of subject, NA for missing values
- race: reported ancestry of subject, NA for missing values
- tStage: stage of tumor associated with sample, NA for healthy samples and tumor or tumor-adjacent samples with missing values
- ageAtDiagDays: age in days at time of diagnosis (for subjects with tumors) or biopsy collection (for healthy subjects), NA for missing values
- daysToDeath: time in days from diagnosis to death for subjects with tumors, NA for survivors in TCGA data set and missing values in other data sets
- sampSite: anatomic subsite, where right refers to cecum and ascending, transverse refers to transverse, left refers to descending and sigmoid, rectum refers to rectum, NA for missing values
- wt_kg: subject weight in kilograms, NA for missing values
- ht_cm: subject height in centimeters, NA for missing values
- rnaMethod: method of enriching for mRNA during library preparation, either polyA for oligo(dT) selection or riboD for ribosomal depletion
- rin: RNA integrity number for sample, NA for missing values
- format: RNA sequencing read format, paired for paired-end, single for single-end

- sequencer: identifier of instrument used for sequencing, taken from FASTQ header, NA for missing values
- platform: name of Illumina instrument model used for sequencing
- study: name assigned to data set for purpose of identifying data source
- percDup: duplication level of reads on a single-end basis as measured by FastQC, presented as a percentage of total single-end reads per individual FASTQ file
- percGc: GC content as a percentage of all nucleotides sequenced as measured by FastQC
- seqLen: length in nucleotides of reads (for single-end) or fragments (for paired-end) for a given sample
- rdProc: number of reads processed by Salmon, where processed means an attempt at quasi-mapping was performed
- rdMap: number of reads quasi-mapped to the transcriptome by Salmon
- percMap: reads quasi-mapped to the transcriptome as a percentage of all reads processed
- data: abbreviated name of repository from which raw FASTQ files were downloaded, gdc means Genomic Data Commons, sradbg means Sequence Read Archive via dbGaP, srapub means Sequence Read Archive directly, bcuva means BarcUVa-Seq

Author(s)

Chris Dampier

Source

See `inst/scripts/make-data.R` for full details on generating this dataset from source files.

References

Dampier, C.H., Devall, M., Jennelle, L.T., Diez-Obrero, V., Plummer, S.J., Moreno, V., Casey, G. Oncogenic Features in Histologically Normal Mucosa: Novel Insights Into Field Effect From a Mega-Analysis of Colorectal Transcriptomes. Clinical and Translational Gastroenterology. 2020 Jul; 11(7): e00210.

Examples

```
library(ExperimentHub)
hub <- ExperimentHub::ExperimentHub()
data <- ExperimentHub::loadResources(hub, "FieldEffectCrc", "cohort A")
se <- data[[1]]
se
```

cohort_B *cohort B from Dampier et al.*

Description

Salmon-generated transcript-level abundance estimates summarized to gene level using `tximport` along with raw counts, gene lengths, and clinical annotations for 30 human primary colorectal tumors and matched normal tissue samples represented as a `SummarizedExperiment`. Abundance estimates derived from paired-end RNA-seq.

Format

A `SummarizedExperiment` object containing 3 `assays` of matrices, each 37,361 rows x 30 columns. Each row is a gene and each column is a sample.

The `SummarizedExperiment` object also includes a `colData` `S4Vectors::DFrame` object with 30 rows and 27 columns. Each row is a sample and each column is a field. The fields are described below.

- dirName: name of directory into which raw data for sample was downloaded, serves as unique identifier
- projId: NCBI BioProject identifier for projects registered in the BioProject database, or common name of projects listed in other databases
- subId: subject identifier
- sampId: sample identifier
- sampType: sample type, which indicates the phenotype of the sample
- dist_cm: relative distance in centimeters from tumor from which given sample was obtained, NA for healthy samples and tumor-adjacent samples without measurements, 0 for tumor samples
- sex: reported sex of subject, NA for missing values
- race: reported ancestry of subject, NA for missing values
- tStage: stage of tumor associated with sample, NA for healthy samples and tumor or tumor-adjacent samples with missing values
- ageAtDiagDays: age in days at time of diagnosis (for subjects with tumors) or biopsy collection (for healthy subjects), NA for missing values
- daysToDeath: time in days from diagnosis to death for subjects with tumors, NA for survivors in TCGA data set and missing values in other data sets
- sampSite: anatomic subsite, where right refers to cecum and ascending, transverse refers to transverse, left refers to descending and sigmoid, rectum refers to rectum, NA for missing values
- wt_kg: subject weight in kilograms, NA for missing values
- ht_cm: subject height in centimeters, NA for missing values
- rnaMethod: method of enriching for mRNA during library preparation, either polyA for oligo(dT) selection or riboD for ribosomal depletion

- rin: RNA integrity number for sample, NA for missing values
- format: RNA sequencing read format, paired for paired-end, single for single-end
- sequencer: identifier of instrument used for sequencing, taken from FASTQ header, NA for missing values
- platform: name of Illumina instrument model used for sequencing
- study: name assigned to data set for purpose of identifying data source
- percDup: duplication level of reads on a single-end basis as measured by FastQC, presented as a percentage of total single-end reads per individual FASTQ file
- percGc: GC content as a percentage of all nucleotides sequenced as measured by FastQC
- seqLen: length in nucleotides of reads (for single-end) or fragments (for paired-end) for a given sample
- rdProc: number of reads processed by Salmon, where processed means an attempt at quasi-mapping was performed
- rdMap: number of reads quasi-mapped to the transcriptome by Salmon
- percMap: reads quasi-mapped to the transcriptome as a percentage of all reads processed
- data: abbreviated name of repository from which raw FASTQ files were downloaded, gdc means Genomic Data Commons, sradbg means Sequence Read Archive via dbGaP, srapub means Sequence Read Archive directly, bcuva means BarcUVa-Seq

Author(s)

Chris Dampier

Source

See `inst/scripts/make-data.R` for full details on generating this dataset from source files.

References

Dampier, C.H., Devall, M., Jennelle, L.T., Diez-Obrero, V., Plummer, S.J., Moreno, V., Casey, G. Oncogenic Features in Histologically Normal Mucosa: Novel Insights Into Field Effect From a Mega-Analysis of Colorectal Transcriptomes. Clinical and Translational Gastroenterology. 2020 Jul; 11(7): e00210.

Examples

```
library(ExperimentHub)
hub <- ExperimentHub::ExperimentHub()
data <- ExperimentHub::loadResources(hub, "FieldEffectCrc", "cohort B")
se <- data[[1]]
se
```

cohort_C *cohort C from Dampier et al.*

Description

Salmon-generated transcript-level abundance estimates summarized to gene level using `tximport` along with raw counts, gene lengths, and clinical annotations for 275 human primary colorectal tissue samples across three phenotypes, including tumor, normal adjacent-to-tumor, and healthy, represented as a `SummarizedExperiment`. Abundance estimates derived from single-end RNA-seq.

Format

A `SummarizedExperiment` object containing 3 `assays` of matrices, each 37,361 rows x 275 columns. Each row is a gene and each column is a sample.

The `SummarizedExperiment` object also includes a `colData` `S4Vectors::DFrame` object with 275 rows and 27 columns. Each row is a sample and each column is a field. The fields are described below.

- dirName: name of directory into which raw data for sample was downloaded, serves as unique identifier
- projId: NCBI BioProject identifier for projects registered in the BioProject database, or common name of projects listed in other databases
- subId: subject identifier
- sampId: sample identifier
- sampType: sample type, which indicates the phenotype of the sample
- dist_cm: relative distance in centimeters from tumor from which given sample was obtained, NA for healthy samples and tumor-adjacent samples without measurements, 0 for tumor samples
- sex: reported sex of subject, NA for missing values
- race: reported ancestry of subject, NA for missing values
- tStage: stage of tumor associated with sample, NA for healthy samples and tumor or tumor-adjacent samples with missing values
- ageAtDiagDays: age in days at time of diagnosis (for subjects with tumors) or biopsy collection (for healthy subjects), NA for missing values
- daysToDeath: time in days from diagnosis to death for subjects with tumors, NA for survivors in TCGA data set and missing values in other data sets
- sampSite: anatomic subsite, where right refers to cecum and ascending, transverse refers to transverse, left refers to descending and sigmoid, rectum refers to rectum, NA for missing values
- wt_kg: subject weight in kilograms, NA for missing values
- ht_cm: subject height in centimeters, NA for missing values
- rnaMethod: method of enriching for mRNA during library preparation, either polyA for oligo(dT) selection or riboD for ribosomal depletion

- rin: RNA integrity number for sample, NA for missing values
- format: RNA sequencing read format, paired for paired-end, single for single-end
- sequencer: identifier of instrument used for sequencing, taken from FASTQ header, NA for missing values
- platform: name of Illumina instrument model used for sequencing
- study: name assigned to data set for purpose of identifying data source
- percDup: duplication level of reads on a single-end basis as measured by FastQC, presented as a percentage of total single-end reads per individual FASTQ file
- percGc: GC content as a percentage of all nucleotides sequenced as measured by FastQC
- seqLen: length in nucleotides of reads (for single-end) or fragments (for paired-end) for a given sample
- rdProc: number of reads processed by Salmon, where processed means an attempt at quasi-mapping was performed
- rdMap: number of reads quasi-mapped to the transcriptome by Salmon
- percMap: reads quasi-mapped to the transcriptome as a percentage of all reads processed
- data: abbreviated name of repository from which raw FASTQ files were downloaded, gdc means Genomic Data Commons, sradbg means Sequence Read Archive via dbGaP, srapub means Sequence Read Archive directly, bcuva means BarcUVa-Seq

Author(s)

Chris Dampier

Source

See `inst/scripts/make-data.R` for full details on generating this dataset from source files.

References

Dampier, C.H., Devall, M., Jennelle, L.T., Diez-Obrero, V., Plummer, S.J., Moreno, V., Casey, G. Oncogenic Features in Histologically Normal Mucosa: Novel Insights Into Field Effect From a Mega-Analysis of Colorectal Transcriptomes. Clinical and Translational Gastroenterology. 2020 Jul; 11(7): e00210.

Examples

```
library(ExperimentHub)
hub <- ExperimentHub::ExperimentHub()
data <- ExperimentHub::loadResources(hub, "FieldEffectCrc", "cohort C")
se <- data[[1]]
se
```

make_txi *Make a* tximport*-style list object from a* SummarizedExperiment

Description

Simple helper function to make a tximport-style txi list object for downstream use, especially with DESeq2 and the DESeqDataSetFromTximport() function.

Usage

```
make_txi(
  se
)
```

Arguments

se	A SummarizedExperiment object with three assays as would be downloaded from the FieldEffectCrc package.

Details

Meant to facilitate differential expression using DESeq2 as explained in the FieldEffectCrc package vignette.

browseVignettes("FieldEffectCrc")

Value

A list object with 4 elements, including 3 matrices and a character string. The matrices are in the following order: abundance, counts, length. The character string specifies whether tximport has scaled the counts. None of the counts have been scaled in the FieldEffectCrc package.

Author(s)

Chris Dampier

Examples

```
## make a txi object
hub <- ExperimentHub::ExperimentHub()
r <- AnnotationHub::query(hub, c("FieldEffectCrc"))
se <- r[["EH3526"]]
txi <- make_txi(se)
str(txi)
dds <- DESeq2::DESeqDataSetFromTximport(
          txi,
          SummarizedExperiment::colData(se),
          ~ sampType
        )
```

`reorder_assays` *Re-order the elements of the* `assays` *slot of a* `SummarizedExperiment` *object to make counts the first element*

Description

Simple helper function to switch the order of assays in a `SummarizedExperiment` object with 3 assays as would be downloaded from the `FieldEffectCrc` package. This function also rounds the counts assay to integer values. Useful for creating a `DESeqDataSet` directly from a downloaded `SummarizedExperiment` object with the `DESeqDataSet()` function.

Usage

```
reorder_assays(
  se,
  order = c("counts", "abundance", "length")
)
```

Arguments

`se`	A `SummarizedExperiment` object with 3 assays as would be downloaded from the `FieldEffectCrc` package. An abundance (i.e. TPM) matrix is the first element after initial download.
`order`	A character vector of length 3 with strings arranged in the order of assays desired in the output. Default setting simply switches the order of counts and abundance to make counts the first element, as is commonly expected of `SummarizedExperiment` objects.

Details

Meant to facilitate alternative uses of the `SummarizedExperiment` objects as explained in the `FieldEffectCrc` package vignette.

browseVignettes("FieldEffectCrc")

Value

A `SummarizedExperiment` object.

Author(s)

Chris Dampier

Examples

```
## move counts to the first assay slot
hub <- ExperimentHub::ExperimentHub()
r <- AnnotationHub::query(hub, c("FieldEffectCrc"))
se <- r[["EH3526"]]
se <- reorder_assays(se)
se
dds <- DESeq2::DESeqDataSet(se, design = ~ sampType)
```

Index

∗ **datasets**
    cohort_A, 3
    cohort_B, 6
    cohort_C, 8
∗ **methods**
    make_txi, 10
    reorder_assays, 11
∗ **package**
    FieldEffectCrc-package, 2

cohort A from Dampier et al.
        *(*cohort_A*)*, 3
cohort B from Dampier et al.
        *(*cohort_B*)*, 6
cohort C from Dampier et al.
        *(*cohort_C*)*, 8
cohort-A *(*cohort_A*)*, 3
cohort-B *(*cohort_B*)*, 6
cohort-C *(*cohort_C*)*, 8
cohort_A, 3
cohort_B, 6
cohort_C, 8
cohortA *(*cohort_A*)*, 3
cohortB *(*cohort_B*)*, 6
cohortC *(*cohort_C*)*, 8

FieldEffectCrc
        *(*FieldEffectCrc-package*)*, 2
FieldEffectCrc-package, 2

make_txi, 10

reorder_assays, 11