Package 'tximport'

August 21, 2017

Version 1.4.0

Title Import and summarize transcript-level estimates for transcript- and gene-level analysis

Description Imports transcript-level abundance, estimated counts and transcript lengths, and summarizes into matrices for use with downstream gene-level analysis packages. Average transcript length, weighted by sample-specific transcript abundance estimates, is provided as a matrix which can be used as an offset for different expression of gene-level counts.

Author Michael Love [cre,aut], Charlotte Soneson [aut], Mark Robinson [aut], Rob Patro [ctb], Andrew Parker Morgan [ctb], Ryan C. Thompson [ctb], Matt Shirley [ctb]

Maintainer Michael Love <michaelisaiahlove@gmail.com>

License GPL (>=2)

VignetteBuilder knitr

Imports utils

Suggests knitr, rmarkdown, testthat, tximportData, TxDb.Hsapiens.UCSC.hg19.knownGene, readr (>= 0.2.2), limma, edgeR, DESeq2 (>= 1.11.6), rhdf5, rjson

biocViews RNASeq, Transcription, GeneExpression, DataImport

RoxygenNote 6.0.1

NeedsCompilation no

R topics documented:

`summarizeToGene` *Import transcript-level abundances and counts for transcript- and gene-level analysis packages*

Description

txiimport imports transcript-level estimates from various external software and optionally summarizes abundances, counts, and transcript lengths to the gene-level (default) or outputs transcript-level matrices (see txOut argument).

Usage

```
summarizeToGene(txi, tx2gene, ignoreTxVersion = FALSE,
  countsFromAbundance = c("no", "scaledTPM", "lengthScaledTPM"))

tximport(files, type = c("none", "salmon", "sailfish", "kallisto", "rsem"),
  txIn = TRUE, txOut = FALSE, countsFromAbundance = c("no", "scaledTPM",
  "lengthScaledTPM"), tx2gene = NULL, varReduce = FALSE,
  dropInfReps = FALSE, ignoreTxVersion = FALSE, geneIdCol, txIdCol,
  abundanceCol, countsCol, lengthCol, importer = NULL)
```

Arguments

txi	list of matrices of trancript-level abundances, counts, and lengths produced by tximport, only used by summarizeToGene
tx2gene	a two-column data.frame linking transcript id (column 1) to gene id (column 2). the column names are not relevant, but this column order must be used. this argument is required for gene-level summarization for methods that provides transcript-level estimates only (kallisto, Salmon, Sailfish)
ignoreTxVersion	logical, whether to split the tx id on the '.' character to remove version information, for easier matching with the tx id in gene2tx (default FALSE)
countsFromAbundance	character, either "no" (default), "scaledTPM", or "lengthScaledTPM", for whether to generate estimated counts using abundance estimates scaled up to library size (scaledTPM) or additionally scaled using the average transcript length over samples and the library size (lengthScaledTPM). if using scaledTPM or lengthScaledTPM, then the counts are no longer correlated with average transcript length, and so the length offset matrix should not be used.
files	a character vector of filenames for the transcript-level abundances
type	character, the type of software used to generate the abundances. Options are "salmon", "sailfish", "kallisto", "rsem". This argument is used to autofill the arguments below (geneIdCol, etc.) "none" means that the user will specify these columns.
txIn	logical, whether the incoming files are transcript level (default TRUE)
txOut	logical, whether the function should just output transcript-level (default FALSE)
varReduce	whether to reduce per-sample inferential replicates information into a matrix of sample variances variance (default FALSE)
dropInfReps	whether to skip reading in inferential replicates (default FALSE)
geneIdCol	name of column with gene id. if missing, the gene2tx argument can be used
txIdCol	name of column with tx id
abundanceCol	name of column with abundances (e.g. TPM or FPKM)
countsCol	name of column with estimated counts
lengthCol	name of column with feature length information
importer	a function used to read in the files

Details

`tximport` will also load in information about inferential replicates – a list of matrices of the Gibbs samples from the posterior, or bootstrap replicates, per sample – if these data are available in the expected locations relative to the `files`, and if `txOut=TRUE`. The inferential replicates, stored in `infReps` in the output list, are on estimated counts, and therefore follow `counts` in the output list. By setting `varReduce=TRUE`, the inferential replicate matrices will be replaced by a single matrix with the sample variance per transcript and per sample. Inferential replicate information is not summarized to the gene level.

While `tximport` summarizes to the gene-level by default, the user can also perform the import and summarization steps manually, by specifing `txOut=TRUE` and then using the function `summarizeToGene`. Note however that this is equivalent to `tximport` with `txOut=FALSE` (the default).

Solutions to the error "tximport failed at summarizing to the gene-level":

1. provide a `tx2gene` data.frame linking transcripts to genes (more below)
2. avoid gene-level summarization by specifying `txOut=TRUE`
3. set `geneIdCol` to an appropriate column in the files

See `vignette('tximport')` for example code for generating a `tx2gene` data.frame from a `TxDb` object. Note that the `keys` and `select` functions used to create the `tx2gene` object are documented in the man page for AnnotationDb-class objects in the AnnotationDbi package (TxDb inherits from AnnotationDb). For further details on generating TxDb objects from various inputs see `vignette('GenomicFeatures')` from the GenomicFeatures package.

Value

a simple list containing matrices: abundance, counts, length. Another list element 'countsFromAbundance' carries through the character argument used in the tximport call. If detected, and `txOut=TRUE`, inferential replicates for each sample will be imported and stored as a list of matrices, itself an element `infReps` in the returned list. If `varReduce=TRUE` the inferential replicates will be summarized according to the sample variance, and stored as a matrix `variance`. The length matrix contains the average transcript length for each gene which can be used as an offset for gene-level analysis.

Functions

- `summarizeToGene`: Summarize transcript-level to gene-level
- `tximport`: Import estimates of abundances and counts

References

Charlotte Soneson, Michael I. Love, Mark D. Robinson (2015): Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. http://dx.doi.org/10.12688/f1000research.7563.1

Examples

```
# load data for demonstrating tximport
# note that the vignette shows more examples
# including how to read in files quickly using the readr package

library(tximportData)
dir <- system.file("extdata", package="tximportData")
```

```
samples <- read.table(file.path(dir,"samples.txt"), header=TRUE)
files <- file.path(dir,"salmon", samples$run, "quant.sf")
names(files) <- paste0("sample",1:6)

# tx2gene links transcript IDs to gene IDs for summarization
tx2gene <- read.csv(file.path(dir, "tx2gene.csv"))

txi <- tximport(files, type="salmon", tx2gene=tx2gene)
```

Index