

# Package ‘HumanTranscriptomeCompendium’

October 16, 2019

**Title** Tools to work with a Compendium of 181000 human transcriptome sequencing studies

**Description** Provide tools for working with a compendium of human transcriptome sequences (originally htxcomp).

**Version** 1.0.1

**Author** Sean Davis, Vince Carey

**Suggests** knitr, BiocStyle, beeswarm, tximportData, DT, tximport, dplyr, magrittr, BiocFileCache, testthat, BiocManager

**Imports** shiny, ssrch, S4Vectors, SummarizedExperiment, utils

**Depends** R (>= 3.6)

**Maintainer** VJ Carey <stvjc@channing.harvard.edu>

**License** Artistic-2.0

**LazyLoad** yes

**LazyData** yes

**biocViews** Transcriptomics, Infrastructure

**VignetteBuilder** knitr

**RoxygenNote** 6.1.1

**Encoding** UTF-8

**git\_url** <https://git.bioconductor.org/packages/HumanTranscriptomeCompendium>

**git\_branch** RELEASE\_3\_9

**git\_last\_commit** 54511b0

**git\_last\_commit\_date** 2019-09-25

**Date/Publication** 2019-10-15

## R topics documented:

addRD . . . . .	2
get_ds4841 . . . . .	2
htx_app . . . . .	3
htx_load . . . . .	4
htx_query_by_study_accession . . . . .	4
htx_query_by_text . . . . .	5
HumanTranscriptomeCompendium.colnames . . . . .	6

load_bigrnaFiles . . . . .	6
load_experTable . . . . .	7
load_studTable . . . . .	7
path_doc4842 . . . . .	8
procExpToGene . . . . .	8
tx2gene_gencode27 . . . . .	9
uniqueAcc_120518 . . . . .	10

## Index 11

---

addRD	<i>add gene-level rowData derived from transcript level rowRanges</i>
-------	---

---

### Description

add gene-level rowData derived from transcript level rowRanges

### Usage

```
addRD(x)
```

### Arguments

x	result of htx_load()
---	----------------------

### Value

RangedSummarizedExperiment with enhanced rowData

### Examples

```
# this function operates on a SummarizedExperiment that has
# transcript-level rowRanges but gene-level quantifications
addRD
```

---

get_ds4841	<i>return instance of ssrch::DocSet with metadata on 4841 human transcriptome studies in NCBI SRA</i>
------------	---

---

### Description

return instance of ssrch::DocSet with metadata on 4841 human transcriptome studies in NCBI SRA

### Usage

```
get_ds4841(cache = BiocFileCache::BiocFileCache(),
           csv_zip_path = path_doc4842())
```

### Arguments

cache	instance of 'BiocFileCache', defaults to 'BiocFileCache::BiocFileCache()'
csv_zip_path	a path leading to the zip file of CSV for metadata in the DocSet instance

**Value**

instance of DocSet as defined in `ssrch` package

**Note**

will bind the correct value of `'zipf'` in `'environment(ds4841@doc_retriever)'`, which depends on details of installation

**Examples**

```
get_ds4841()
```

---

<code>htx_app</code>	<i>explore SRA metadata</i>
----------------------	-----------------------------

---

**Description**

explore SRA metadata

**Usage**

```
htx_app()
```

**Value**

a SummarizedExperiment can be requested through an event

**Note**

This function deals with extraction of compendium elements. The overall scope is determined by `HumanTranscriptomeCompendium::studTable` which is the list of all studies with taxon 9606, strategy RNA-seq, source transcriptomic. Some studies will not have experiments in the compendium, and if such are selected, a warning will be generated in the session.

**Examples**

```
if (interactive()) htx_app()
```

---

htx_load	<i>load a SummarizedExperiment shell for the Human Transcriptome Compendium</i>
----------	---

---

**Description**

load a SummarizedExperiment shell for the Human Transcriptome Compendium

**Usage**

```
htx_load(remotePath = "https://s3.amazonaws.com/bcfound-bigrna/htxcompSE.rds",
  cache = BiocFileCache::BiocFileCache(), genesOnly = TRUE)
```

**Arguments**

remotePath	path to an RDS representation of the DelayedArray-based SummarizedExperiment
cache	a BiocFileCache instance, defaulting to value of BiocFileCache()
genesOnly	logical(1) if TRUE return reference to SummarizedExperiment with gene-level quantifications; in this case the remotePath value is set to 'https://s3.amazonaws.com/bcfound-bigrna/rangedHtxGeneSE.rds'.

**Value**

a RangedSummarizedExperiment instance

**Examples**

```
htx_load
```

---

htx_query_by_study_accession	<i>retrieve 'restfulSE' SummarizedExperiment instance for selected studies in htx compendium</i>
------------------------------	--

---

**Description**

retrieve 'restfulSE' SummarizedExperiment instance for selected studies in htx compendium

**Usage**

```
htx_query_by_study_accession(studies, htxSE, ...)
```

**Arguments**

studies	character vector of study accessions
htxSE	SummarizedExperiment instance, typically the result of htx_load(), which we don't want to repeat needlessly
...	passed to 'htx_load', ignored if 'se' is nonmissing

**Value**

SummarizedExperiment instance

**Note**

This function was designed to perform a single query on a 'fresh' compendium image from 'htx\_load()'. However, one could consider iterating the process to build up metadata on multiple series of studies. This is not likely to succeed without careful manipulation of the colData of the input SummarizedExperiment. A message will be written if the input SummarizedExperiment appears to be other than a 'fresh' 'htx\_load' result.

**Examples**

```
htx_query_by_study_accession("ERP011411")
```

---

htx_query_by_text	<i>subset compendium through keyword lookup</i>
-------------------	---

---

**Description**

subset compendium through keyword lookup

**Usage**

```
htx_query_by_text(query, ..., tryGrep = TRUE, ignore.case = TRUE)
```

**Arguments**

query	character(1) to be found in ls(ssrch::kw2docs(get_ds4841()))
...	passed to 'htx_query_by_study_accession'
tryGrep	logical(1) if TRUE, 'query' does not match any keyword directly, it will be treated as a regular expression and the vector of keywords will be grepped for pattern 'query'; defaults to TRUE
ignore.case	logical(1) used when tryGrep is TRUE, defaults to TRUE

**Value**

SummarizedExperiment instance

**Note**

The DocSet instance returned by 'get\_ds4841()' is used. Lookups are case-sensitive. Look carefully at note for 'htx\_query\_by\_study\_accession' to understand logic of incrementing metadata on a given input SummarizedExperiment.

**Examples**

```
htx_query_by_text("HNRNPC")
```

---

HumanTranscriptomeCompendium.colnames  
*character vector of available samples in HDF cloud assay representation*

---

**Description**

character vector of available samples in HDF cloud assay representation

**Usage**

```
HumanTranscriptomeCompendium.colnames
```

**Format**

character vector

**Source**

compendium processing

**Examples**

```
head(HumanTranscriptomeCompendium::HumanTranscriptomeCompendium.colnames)
```

---

load\_bigrnaFiles      *obtain listing of contents of BigRNA compendium (salmon runs)*

---

**Description**

obtain listing of contents of BigRNA compendium (salmon runs)

**Usage**

```
load_bigrnaFiles(cache = BiocFileCache::BiocFileCache())
```

**Arguments**

cache                  instance of 'BiocFileCache', defaults to 'BiocFileCache::BiocFileCache()'

**Value**

a named vector

**Examples**

```
if (interactive()) head(load_bigrnaFiles())
```

---

load_experTable	<i>obtain listing of experiments and submission date/time in compendium</i>
-----------------	---

---

**Description**

obtain listing of experiments and submission date/time in compendium

**Usage**

```
load_experTable(cache = BiocFileCache::BiocFileCache())
```

**Arguments**

cache            instance of 'BiocFileCache', defaults to 'BiocFileCache::BiocFileCache()'

**Value**

a data.frame

**Examples**

```
if (interactive()) head(load_experTable())
```

---

load_studTable	<i>obtain listing of all studies in compendium</i>
----------------	--

---

**Description**

obtain listing of all studies in compendium

**Usage**

```
load_studTable(cache = BiocFileCache::BiocFileCache())
```

**Arguments**

cache            instance of 'BiocFileCache', defaults to 'BiocFileCache::BiocFileCache()'

**Value**

a data.frame

**Examples**

```
if (interactive()) head(load_studTable())
```

---

path\_doc4842                    *return path to metadata csvs in zip file*

---

### Description

return path to metadata csvs in zip file

### Usage

```
path_doc4842(cache = BiocFileCache::BiocFileCache())
```

### Arguments

cache                    instance of 'BiocFileCache', defaults to 'BiocFileCache::BiocFileCache()'

### Value

path to zipfile

### Note

CSVs were retrieved using methods provided at <https://api-omicidx.cancerdatasci.org/sra/1.0/ui/> and zipped together. Function will lodge zipfile in 'cache' if not present.

### Examples

```
path_doc4842()
```

---

procExpToGene                    *acquire a single sample from bigRNA compendium specified by accession and develop gene-level quantifications using tximport*

---

### Description

acquire a single sample from bigRNA compendium specified by accession and develop gene-level quantifications using tximport

### Usage

```
procExpToGene(acc, tx2gene = tx2gene_gencode27(),
  urlprefix = "http://bigrna.cancerdatasci.org/",
  manifestdata = HumanTranscriptomeCompendium::load_bigrnaFiles(),
  regexp = "quant.sf.bz2|json")
```



**Arguments**

acc	character(1) sample-level accession as defined in SRA
tx2gene	a data.frame instance mapping transcript identifiers used in the compendium to gene identifiers. See note.
urlprefix	character(1) where the salmon run outputs are lodged, with acc a subfolder defined through the manifestData parameter.
manifestdata	a character vector defining folders (under results/human/27/ with salmon outputs).
regexp	a character(1) regular expression for filtering filename elements in manifestdata to define which salmon output components in the bigrna compendium are retrieved.

**Value**

the result of a tximport run

**Note**

The tx2gene\_gencode function supplied with this package uses the tximportData package contents to create the data.frame for use as tx2gene. The system2 function is used to generate folders to be used by tximport.

**Examples**

```
# this example involves nontrivial internet communications
args(procExpToGene)
## Not run:
td = tempdir()
od = getwd()
setwd(td)
nn = procExpToGene("ERX1097381")
str(nn)
setwd(od)

## End(Not run)
```

---

tx2gene_gencode27	<i>generate a data.frame mapping gencode 27 ensembl transcript identifiers to ensembl gene identifiers</i>
-------------------	--

---

**Description**

generate a data.frame mapping gencode 27 ensembl transcript identifiers to ensembl gene identifiers

**Usage**

```
tx2gene_gencode27()
```

**Value**

a data.frame with 200401 rows mapping transcript identifiers in column 1 to 58288 gene symbols in column 2.

**Note**

Uses CSV in tximportData to acquire the information.

**Examples**

```
head(tx2gene_gencode27())
```

---

uniqueAcc_120518	<i>experiment accessions available in compendium as of may 12 2018</i>
------------------	--

---

**Description**

experiment accessions available in compendium as of may 12 2018

**Usage**

```
uniqueAcc_120518
```

**Format**

```
data.frame
```

**Source**

SRAdBV2 may 12 2018

**Examples**

```
head(HumanTranscriptomeCompendium::uniqueAcc_120518)
```

# Index

## \*Topic **datasets**

HumanTranscriptomeCompendium.colnames,

[6](#)

uniqueAcc\_120518, [10](#)

addRD, [2](#)

get\_ds4841, [2](#)

htx\_app, [3](#)

htx\_load, [4](#)

htx\_query\_by\_study\_accession, [4](#)

htx\_query\_by\_text, [5](#)

HumanTranscriptomeCompendium.colnames,

[6](#)

load\_bigrnaFiles, [6](#)

load\_experTable, [7](#)

load\_studTable, [7](#)

path\_doc4842, [8](#)

procExpToGene, [8](#)

tx2gene\_gencode27, [9](#)

uniqueAcc\_120518, [10](#)