Package 'AnnotatedBCGEData'

November 10, 2026

Title 100+ Curated Breast Cancer Gene Expression Data sets

Version 0.99.13

Date 2025-10-16

Description AnnotatedBCGEData is an R Bioconductor package that provides access to more than 100 gene-expression datasets generated using microarray and RNA-sequencing technologies. We collected these datsets from public repositories, including Gene Expression Omnibus, ArrayExpress, and The Cancer Genome Atlas. Where feasible, we obtained and reprocessed the raw data using computational pipelines and annotations that were standardized and more updated than those used to process the data originally. Additionally, we annotated the sample metadata using oncology terms from the National Cancer Institute Thesaurus. This was necessary because different data submitters used different terminology to describe breast-cancer patients and samples. Mapping the metadata to ontology terms makes it possible for users to search the data using these standardized terms.

License CC0

URL https://github.com/srp33/AnnotatedBCGEData

BugReports https://github.com/srp33/AnnotatedBCGEData//issues

biocViews ExperimentData, CancerData, ExpressionData, Homo_sapiens_Data, GEO, ArrayExpress, Tissue, MicroarrayData, SequencingData, BreastCancerData, Genome, TissueMicroarrayData

Encoding UTF-8

Roxygen list(markdown=TRUE)

Suggests BiocStyle, knitr, rmarkdown, sessioninfo, ggplot2, testthat (>= 3.0.0)

VignetteBuilder knitr

Depends R (>= 4.6)

Imports SummarizedExperiment, stringr, readr, dplyr, tibble, tidyr, zen4R, BiocFileCache, cli, rlang, readxl

LazyData false

Config/testthat/edition 3

Config/roxygen2/version 8.1.0

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

git_branch devel

git_last_commit 3f5c4df

git_last_commit_date 2026-10-22

Repository Bioconductor 3.24

Date/Publication 2026-11-10

Author Heidi Steadman [aut],
Shu-Fei Lo [aut],
Stephen Piccolo [cre, aut] (ORCID:
<https://orcid.org/0000-0003-2001-5640>),
BYU College Undergraduate Research Award [fnd]

Maintainer Stephen Piccolo <stephen_piccolo@byu.edu>

Contents

AnnotatedBCGEData-package 2
getDataset 3
getDatasetMetadata 4
searchForDatasetsByField 5
searchForDatasetsByValue 5
searchOntologyTerms 6

Index **7**

`AnnotatedBCGEData-package`

AnnotatedBCGEData: 100+ Curated Breast Cancer Gene Expression Data sets

---

Description

AnnotatedBCGEData is an R Bioconductor package that provides access to more than 100 gene-expression datasets generated using microarray and RNA- sequencing technologies. We collected these datsets from public repositories, including Gene Expression Omnibus, ArrayExpress, and The Cancer Genome Atlas. Where feasible, we obtained and reprocessed the raw data using computational pipelines and annotations that were standardized and more updated than those used to process the data originally. Additionally, we annotated the sample metadata using oncology terms from the National Cancer Institute Thesaurus. This was necessary because different data submitters used different terminology to describe breast-cancer patients and samples. Mapping the metadata to ontology terms makes it possible for users to search the data using these standardized terms.

Overview

AnnotatedBCGEData is an R Bioconductor package that provides access to more than 100 gene-expression datasets generated using microarray and RNA- sequencing technologies. We collected these datsets from public repositories, including GEO, ArrayExpress, and TCGA. Where feasible, we obtained the and reprocessed the raw data using computational pipelines and annotations that were standardized and more updated than those used to process the data originally. Additionally, we annotated the sample metadata using oncology terms from the National Cancer Institute Thesaurus. This was necessary because different data submitters used different terminology to describe breast-cancer patients and samples. Mapping the metadata to ontology terms makes it possible for users to search the data using these standardized terms.

- `getDataset()` allows users to get a SummarizedExperiment object of the chosen dataset.
- `searchOntologyTerms()` allows users to browse our custom ontology.
- `searchForDatasetsByField()` is one function that finds datasets related to a given term.

Read the vignette for more information and detailed examples of the package functions.

Examples

```
# Searching the ontology by a specific term name
searchOntologyTerms("Progesterone Receptor Status")

# Find associated datasets in the ontology mappings by a field code
searchForDatasetsByField("C16149")

# Retrieving a dataset
getDataset("GSE41197")
```

Author(s)

Maintainer: Stephen Piccolo <stephen_piccolo@byu.edu> (ORCID)

Authors:

- Stephen Piccolo <stephen_piccolo@byu.edu> (ORCID)
- Heidi Steadman <hstead05@byu.edu>
- Shu-Fei Lo <shufeilo@byu.edu>

Other contributors:

- BYU College Undergraduate Research Award [funder]

See Also

Useful links:

- https://github.com/srp33/AnnotatedBCGEData
- Report bugs at https://github.com/srp33/AnnotatedBCGEData//issues

---

`getDataset` *Function that retrieves a breast cancer gene expression dataset. Accepts the identifier of a desired dataset, an optional directory path for the data to be cached in, and an optional version number. The function looks for the data in the cache. If the data is not found, it is downloaded and cached. The data are packaged as SummarizedExperiment objects. If the user specifies a cache directory, downloaded files are stored in this location so they don't need to be re-downloaded. The default is to use the R session's temporary directory, which is removed after the session ends.*

Description

Function that retrieves a breast cancer gene expression dataset. Accepts the identifier of a desired dataset, an optional directory path for the data to be cached in, and an optional version number. The function looks for the data in the cache. If the data is not found, it is downloaded and cached. The data are packaged as SummarizedExperiment objects. If the user specifies a cache directory, downloaded files are stored in this location so they don't need to be re-downloaded. The default is to use the R session's temporary directory, which is removed after the session ends.

Usage

```
getDataset(datasetID, cacheDirPath = tempdir(), v = NULL)
```

Arguments

datasetID	The identifier of the data set.
cacheDirPath	The optional directory path to where the cache should be saved.
v	Version of data to download. Defaults to most recent.

Value

A SummarizedExperiment object for the dataset.

Examples

```
getDataset("GSE41197")
```

---

getDatasetMetadata *A function that accepts a vector of dataset IDs and returns a tibble with metadata about the overall study.*

Description

A function that accepts a vector of dataset IDs and returns a tibble with metadata about the overall study.

Usage

```
getDatasetMetadata(datasets)
```

Arguments

datasets	A vector of strings consisting of dataset identifier(s).

Value

Tibble with metadata.

Examples

```
getDatasetMetadata(c("GSE41197", "GSE10797"))
```

`searchForDatasetsByField` *Function that searches breast cancer datasets based on ontology terms mapped to data fields.*

Description

Accepts an ontology term code and searches for datasets based on fields (metadata columns) that have been mapped to that ontology term. Returns a tibble with the dataset identifier, field name, and code. Users can then pass the dataset identifiers to the getDataset function.

Usage

```
searchForDatasetsByField(code)
```

Arguments

`code`	is an ontology term code retrieved using the searchOntologyTerms function.

Value

a tibble providing information about any identified datasets.

Examples

```
searchForDatasetsByField("C16149")
```

`searchForDatasetsByValue` *Function that searches breast cancer datasets based on ontology terms mapped to data value.*

Description

Accepts an ontology term code and searches for datasets based on metadata values that have been mapped to that ontology term. Returns a tibble with the dataset identifier, field name, and code. Users can then pass the dataset identifiers to the getDataset function.

Usage

```
searchForDatasetsByValue(code)
```

Arguments

`code`	is an ontology term code retrieved using the searchOntologyTerms function.

Value

a tibble providing information about any identified datasets.

Examples

```
searchForDatasetsByValue("C15496")
```

`searchOntologyTerms` *Function that searches for ontology terms.*

Description

Accepts a string to search by and which ontology field to search ontology terms associated with that string. This could be the name of the term itself, the term's definition, the unique identifier (code) associated with that term, or its URL. The function returns a tibble with information about the matching ontology term(s).

Usage

```
searchOntologyTerms(term, term_type = "Name")
```

Arguments

`term`	String to search by.
`term_type`	Ontology field to search: Name, URI, Code, or Definition. When Name (default) is provided, it also searches based on synonyms.

Value

A tibble with a row for each matching ontology term.

Examples

```
searchOntologyTerms("Progesterone Receptor Status")
```

Index

∗ **internal**
AnnotatedBCGEData-package, 2

AnnotatedBCGEData
(AnnotatedBCGEData-package), 2
AnnotatedBCGEData-package, 2

getDataset, 3
getDataset(), *3*
getDatasetMetadata, 4

searchForDatasetsByField, 5
searchForDatasetsByField(), *3*
searchForDatasetsByValue, 5
searchOntologyTerms, 6
searchOntologyTerms(), *3*