# Package 'GSVAdata'

October 1, 2024

**Title** Data employed in the vignette of the GSVA package

**Version** 1.41.1

**Author** Robert Castelo <robert.castelo@upf.edu>

**Description** This package stores the data employed in the vignette of the GSVA package. These data belong to the following publications: Armstrong et al. Nat Genet 30:41-47, 2002; Cahoy et al. J Neurosci 28:264-278, 2008; Carrel and Willard, Nature, 434:400-404, 2005; Huang et al. PNAS, 104:9758-9763, 2007; Pickrell et al. Nature, 464:768-722, 2010; Skaletsky et al. Nature, 423:825-837; Verhaak et al. Cancer Cell 17:98-110, 2010; Costa et al. FEBS J, 288:2311-2331, 2021.

**Maintainer** Robert Castelo <robert.castelo@upf.edu>

**Depends** R (>= 3.5), Biobase, GSEABase, hgu95a.db, SummarizedExperiment

**License** Artistic-2.0

**biocViews** ExperimentData, RNASeqData, Homo_sapiens_Data, CancerData, LeukemiaCancerData

**git_url** https://git.bioconductor.org/packages/GSVAdata

**git_branch** devel

**git_last_commit** 6b190dd

**git_last_commit_date** 2024-09-27

**Repository** Bioconductor 3.20

**Date/Publication** 2024-10-01

## Contents

GSVAdata-package *Data employed in the vignette of the GSVA package.*

## Description

This package contains data employed in the vignette of the GSVA package.

## Data sets

- `leukemia` Leukemia data by Armstrong et al. (2002) from the Broad Institute.
- `c2BroadSets` C2 canonical pathways from the MSigDB 3.0 database of gene sets at the Broad Institute.
- `gbm_VerhaakEtAl` TCGA Glioblastoma Multiforme microarray expression data from Verhaak et al. (2010).
- `brainTxDbSets` Gene sets signatures specific to four different brain cell types derived from murine models (Cahoy et al., 2008).
- `commonPickrellHuang` Matching microarray and RNA-seq data from human lymphoblastoid cell lines (Huang et al., 2007; Pickrell et al., 2010).
- `annotEntrez220212` Annotation data on gene length and G+C content from NCBI at `http://www.ncbi.nlm.nih.gov`.
- `genderGenesEntrez` Entrez genes with documented sex-specific expression (Skaletsky et al., 2003; Carrel and Willard, 2005).

## Author(s)

S. Haenzelmann, J. Guinney and R. Castelo

## References

S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub and S.J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 30:41-47, 2002.

J.D. Cahoy, B. Emery, A. Kaushal, L.C. Foo, J.L. Zamanian, et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci*, 28:264-278, 2008.

L. Carrel and H.F. Willard. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434:400–404, 2005.

R.S. Huang, S. Duan, W.K. Bleibel, E.O. Kistner, W. Zhang, T.A. Clark, T.X. Chen, A.C. Schweitzer, J.E. Blume, N.J. Cox and M.E. Dolan, *Proc. Natl. Acad. Sci. USA*, 104(23):9758-9763, 2007.

J.K. Pickrell, J.C. Marioni, A.A Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard, *Nature*, 464:768-772, 2010.

H.S. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423:825–837, 2003.

R.G.W. Verhaak, K.A. Hoadley, E. Purdom, V. Wang, Y. Qi, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17:98-110, 2010.

---

`annotEntrez220212` *Annotation data on gene length and G+C content from NCBI*

---

## Description

Annotation data for human genes defined by Entrez identifiers and downloaded on 22/02/2012.

## Usage

```
data(annotEntrez220212)
```

## Format

`Length`: Length of the longest cDNA of this gene.

`GCcontent`: G+C content of the longest cDNA of this gene.

## Details

All human mRNAs were downloaded from NCBI on 22/02/12 by going first to the taxonomy browser, select "Homo Sapiens", then select mRNAs, then follow the link on "manage filters" and made sure that only mRNAs is checked. Then, we selected FASTA format and sent it to a file. We also downloaded the NCBI Entrez Gene ID to NCBI Accession mapping from `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2accession.gz` which we used to group mRNA transcripts by Entrez Gene identifier. Finally, for each Entrez Gene identifier we picked the longest mRNA and stored its length and G+C content in a data frame called `annotEntrez220212` whose row names indicate the corresponding Entrez Gene identifier.

## Source

NCBI: `http://www.ncbi.nlm.nih.gov`

## Examples

```
data(annotEntrez220212)
dim(annotEntrez220212)
head(annotEntrez220212)
```

## brainTxDbSets *Gene sets signatures of brain cell types*

### Description

Gene sets signatures specific to four different brain cell types (astrocytes, oligodendrocytes, neurons and cultured astroglial cells) derived from murine models (Cahoy et al. 2008).

### Usage

```
data(brainTxDbSets)
```

### Details

The data is contained in an `list` object called `brainTxDbSets` obtained from the Brain Transcriptome Database (Cahoy et al., 2008).

### Source

Cahoy, J.D., Emery, B., Kaushal, A., Foo, L.C., Zamanian, J.L. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci*, 28:264-278, 2008.

### Examples

```
data(brainTxDbSets)
head(lapply(brainTxDbSets, head))
```

## c2BroadSets *C2 collection of canonical pathways from MSigDB 3.0*

### Description

C2 Broad Sets.

### Usage

```
data(c2BroadSets)
```

### Details

The data is contained in an `GeneSetCollection` object called `c2BroadSets` obtained by parsing the file `c2.all.v3.0.entrez.gmt`, downloaded from http://www.broadinstitute.org/gsea, using the `getGmt()` function from the `GSEABase` package.

### Source

Subramanian, Tamayo, et al. *PNAS*, 102:15545-15550, 2005.

Mootha, Lindgren, et al. *Nat Genet*, 34:267-273, 2003.

## Examples

```
data(c2BroadSets)
c2BroadSets
```

---

commonPickrellHuang *Matching microarray and RNA-seq data from human lymphoblastoid cell lines*

---

## Description

ExpressionSet objects containing microarray and RNA-seq count data for 11,508 matching Entrez genes from 36 samples of lymphoblastoid cell lines derived from unrelated Nigerian individuals. These microarray and count data are employed in the vignette of the package GSVA Hanzelmann et al. (submitted). The original experimental data was published by Huang et al. (2007) and Pickrell et al. (2010).

## Usage

```
data(commonPickrellHuang)
```

## Format

huangArrayRMAnoBatchCommon_eset: ExpressionSet object containing filtered, normalized and batch-removed microarray expression values for 11,508 Entrez genes from 36 unrelated Nigerian individuals.

pickrellCountsArgonneCQNcommon_eset: ExpressionSet object containing filtered and normalized RNA-seq read counts for 11,508 Entrez genes from 36 unrelated Nigerian individuals. This table of counts corresponds to RNA-seq data produced at the Argonne sequencing center (see Pickrell et al., 2010).

pickrellCountsYaleCQNcommon_eset: ExpressionSet object containing filtered and normalized RNA-seq read counts for 11,508 Entrez genes from 36 unrelated Nigerian individuals. This table of counts corresponds to RNA-seq data produced at the Yale sequencing center (see Pickrell et al., 2010).

## Details

The microarray data was processed from the raw CEL files available at http://www.ncbi.nlm.nih.gov/geo under accession GSE7792. First, only Yoruba samples were considered. Second, data was processed using the Bioconductor oligo package. Quality assessment was performed by calculating NUSE and RLE diagnostics (Bolstad et al., 2005) and discarding those samples that either of the two reported diagnostics was considered below a minimum quality threshold. Third, most samples formed part of family trios and only samples belonging to father or mother were kept. Fourth, using the RMA algorithm (Irizarry et al., 2003) implemented in the rma() function from the oligo package with argument target="core", expression values were background corrected, normalized and summarized into Affymetrix transcript clusters. Fifth, using the getNetAffx function from the oligo package, Affymetrix transcript cluster identifiers were translated into Entrez Gene identifiers resolving duplicated assignments by keeping the transcript cluster with largest expression variability measured by its interquartile range (IQR).

At this point an expression data matrix of 17,324 Entrez genes by 59 samples was obtained and using the scanning date of each CEL file samples were grouped into 5 batches stored in the phenotypic variable Batch within the resulting ExpressionSet. Batch effect was removed by using the

QR-decomposition method implemented in the `removeBatchEffect()` function from the package `limma` while keeping the sex-specific expression effect by setting the gender sample indicator variable within the design matrix argument. Finally, samples and genes were further filtered to match those from the RNA-seq tables of counts.

The RNA-seq data was obtained by directly downloading the tables of counts processed by Pickrell et al. (2010) from http://eqtl.uchicago.edu/RNA_Seq_data/results, which initially consisted of 41,466 Ensembl genes by 80 and 81 samples corresponding to the RNA-seq data obtained at the Argonne and Yale sequencing centers, respectively. Some of the samples (11 from Argonne and 12 from Yale) were prepared and sequenced twice within each sequencing center. In these cases we kept the sample of deeper coverage obtaining a final number of 69 samples on each table. We further filtered genes with low expression by discarding those with a mean of less than 0.5 counts per million calculated in log2 scale resulting in tables of counts with 17,607 genes (Argonne) and 17,843 genes (Yale) by 69 samples and we kept those genes common to both tables (17,324). Next, we normalized these two tables of counts adjusting for gene length and G+C content using the Bioconductor package `cqn` (Hansen et al., 2012). The corresponding gene length and G+C content information was extracted from data deposited at the same site from were the tables of counts were downloaded. We further filtered these two normalized tables of counts in order to match the genes and samples obtained after processing the LCL microarray data from Huang et al. (2007). This step required first to translate Ensembl gene identifiers into Entrez gene identifiers and second to match gene and sample identifiers between microarray and RNA-seq data. After these two steps we obtained the two final tables of counts of 11,508 Entrez genes by 36 samples included in this package.

## Source

R.S. Huang, S. Duan, W.K. Bleibel, E.O. Kistner, W. Zhang, T.A. Clark, T.X. Chen, A.C. Schweitzer, J.E. Blume, N.J. Cox and M.E. Dolan, *Proc. Natl. Acad. Sci. USA*, 104(23):9758-9763, 2007.

J.K. Pickrell, J.C. Marioni, A.A Pai, J.F. Degner, B.E. Engelhardt, E. Nkadori, J.B Veyrieras, M. Stephens, Y. Gilad, and J.K. Pritchard, *Nature*, 464:768-772, 2010.

## References

B.M. Bolstad, F. Collin, K. Brettschneider, L. Simpson, R.A. Irizarry, and T.P. Speed. Quality assessment of Affymetrix GeneChip data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pg. 33–48, Springer, 2005.

K.D. Hansen, R.A. Irizarry and Z. Wu. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, 2012.

R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

S. Haenzelmann, J. Guinney and R. Castelo. GSVA: Gene Set Variation Analysis for microarray and RNA-Seq data, *submitted*.

## See Also

`genderGenesEntrez`

## Examples

```
suppressMessages(library(Biobase))
data(commonPickrellHuang)
dim(huangArrayRMAnoBatchCommon_eset)
```

```
dim(pickrellCountsArgonneCQNcommon_eset)
dim(pickrellCountsYaleCQNcommon_eset)
table(huangArrayRMAnoBatchCommon_eset$Gender)
table(pickrellCountsArgonneCQNcommon_eset$Gender)
table(pickrellCountsYaleCQNcommon_eset$Gender)
stopifnot(identical(featureNames(huangArrayRMAnoBatchCommon_eset),
                    featureNames(pickrellCountsArgonneCQNcommon_eset)))
stopifnot(identical(sampleNames(huangArrayRMAnoBatchCommon_eset),
                    sampleNames(pickrellCountsArgonneCQNcommon_eset)))
```

---

gbm_eset *Glioblastoma Multiforme (GBM) Data by Verhaak et al. (2010)*

---

## Description

Microarray data from Glioblastoma multiforme (GBM) downloaded from the TCGA website (http://cancergenome.nih.gov). The data is provided as an ExpressionSet object containing RMA-processed expression values.

## Usage

```
data(gbm_VerhaakEtAl)
```

## Details

The data is contained in an ExpressionSet object called gbm_eset and was obtained using RMA (Irizarry et al. 2003).

## Source

Verhaak, R.G.W., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17:98-110, 2010.

## References

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

## Examples

```
data(gbm_VerhaakEtAl)
gbm_eset
head(pData(gbm_eset))
table(gbm_eset$subtype)
```

genderGenesEntrez *Entrez genes with documented sex-specific expression*

## Description

Entrez genes with documented sex-specific expression.

## Usage

```
data(genderGenes)
```

## Format

`msYgenesEntrez`: Entrez gene identifiers from genes belonging to the male-specific region of chromosome Y (Skaletsky et al., 2003).

`XiEgenesEntrez`: Entrez gene identifiers from genes located in the X chromosome and which have been reported to escape X-inactivation (Carrel and Willard, 2005).

## Details

These are two vectors of Entrez gene identifiers corresponding to genes with sex-specific expression documented by Skaletsky et al. (2003) and Carrel and Willard (2005).

## Source

H.S. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, T. Graves, S.F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423:825–837, 2003.

L. Carrel and H.F. Willard. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434:400–404, 2005.

## Examples

```
data(genderGenesEntrez)
length(msYgenesEntrez)
length(XiEgenesEntrez)
```

`geneExpCostaEtAl2021` *RNA-seq Data by Costa et al. (2021)*

## Description

Normalized log-CPM units of expression derived from the bulk RNA sequencing (RNA-seq) data published by Costa et al. (2021). The RNA-seq data was obtained by extracting and sequencing total RNA from archived neonatal dried blood spots (DBS) specimens from 21 extremely low gestational age newborns (ELGANs, < 28 weeks of gestation). The DBS specimens were obtained from 8 females and 13 males, among which 10 where exposed to a fetal inflammatory response (FIR) before birth, and 11 were not exposed.

## Usage

```
data(geneExpCostaEtAl2021)
```

## Details

The data is contained in an `SummarizedExperiment` object called `geneExpCostaEtAl2021` obtained as follows:

- Raw 2x75nt paired-end reads in FASTQ files were aligned to the GRCh38 version of the reference human genome, without alternate locus scaffolds (GCA_000001405.15) and including human decoy sequences from hs38d1 (GCA_000786075.2), using STAR version 2.6.0c (Dobin et al., 2013) with default parameters, except for `--peOverlapNbasesMin 10` and `--sjdbOverhang 74`.
- Aligned reads in BAM files were reduced to a table of counts of 25,221 genes by 84 samples using gene annotations from Gencode v24 and the R/Bioconductor package GenomeAlignments version 1.16.0, and its function `summarizeOverlaps`. We used specific arguments in the call to this function to restrict the count of genic reads to only those that fell entirely within the exonic regions and aligned to a unique site on the genome, to reflect library preparation protocols, and to avoid counting reads without a matching pair or overlapping multiple features.
- Lowly-expressed genes were filtered out by discarding those that did not show a minimum reliable level of expression of 10 counts per million reads of the smallest library size, in at least 6 samples. This filtering step led to a final table of counts of 11,279 genes by 21 samples.
- Normalized log-CPM units of expression were obtained by using the edgeR package (Robinson et al., 2010) as follows. First, we calculated library factors using the TMM algorithm implemented in the function `calcNormFactors()`, and second we used the function `cpm()` with default parameters, except for `log=TRUE`.

## Source

D. Costa, N. Bonet, A. Solé, J.M. González de Aledo-Castillo, E. Sabidó, F. Casals, C. Rovira, A. Nadal, J.L. Marín, T. Cobo and R. Castelo. Genome-wide postnatal changes in immunity following fetal inflammatory response. *FEBS Journal*, 288:2311-2331, 2021. https://doi.org/10.1111/febs.15578.

## References

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, and Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15-–21, 2013.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.

## Examples

```
data(geneExpCostaEtAl2021)
geneExpCostaEtAl2021
```

---

`leukemia_eset` *Leukemia Data by Armstrong et al. (2002) from the Broad Institute*

---

## Description

Microarray data hybridized on the Affymetrix Human Genome U95 Set chip, for 37 different individuals with human acute leukemias, where 20 of them had conventional childhood acute lymphoblastic leukemia (ALL) and the other 17 were affected with the MLL (mixed-lineage leukemia gene) translocation. The data is provided as an `ExpressionSet` object containing RMA-processed expression values.

## Usage

```
data(leukemia)
```

## Details

The data is contained in an `ExpressionSet` object called `leukemia_eset` obtained as follows:

- Raw CEL files corresponding to the data of the entire study (72 individuals) were downloaded from http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63
- 41 ALL and MLL samples with the same scanning date were kept and the rest were discarded.
- Based on quality assessments by NUSE and RLE diagnostics (Bolstad et al., 2005), 4 additional samples were discarded such that 20 ALL and 17 MLL samples were finally kept.
- Probe-level data from these 37 samples were background corrected, normalized and summarized using RMA (Irizarry et al., 2003) providing this final `ExpressionSet` object.

## Source

Scott A. Armstrong, Jane E. Staunton, Lewis B. Silverman, Rob Pieters, Monique L. den Boer, Mark D. Minden, Stephen E. Sallan, Eric S. Lander, Todd R. Golub and Stanley J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 30:41-47, 2002.

## References

Bolstad, B.M., Collin, F., Brettschneider, K., Simpson, L., Irizarry, R., and Speed, T.P. Quality assessment of Affymetrix GeneChip data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pg. 33–48, Springer, 2005.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 2003.

## Examples

```
data(leukemia)
leukemia_eset
head(pData(leukemia_eset))
```

# Index