

Introduction to the *sampleClassifierData* Package

Khadija El Amrani

April 17, 2025

Contents

1	Introduction	1
2	Data overview.	1
3	Data pre-processing	2

1 Introduction

sampleClassifierData contains a collection of publicly available microarray and RNA-seq datasets that have been pre-processed for use with the *sampleClassifier* package. These pre-processed datasets can be used as reference matrices for gene expression profile classification using *sampleClassifier*. This introduction contains a brief overview of the datasets included in the package. For more examples on how to use *sampleClassifier* and *sampleClassifierData*, please refer to the *sampleClassifier* Vignette.

2 Data overview

First, we load the package *sampleClassifierData*:

```
> library(sampleClassifierData)
```

The *sampleClassifierData* package contains two microarray datasets and two RNA-seq datasets that have been pre-processed for use with *sampleClassifier*.

The datasets are stored as `SummarizedExperiment` objects. The numeric matrices to use with the *sampleClassifier* can be extracted using the `assay()` function from *SummarizedExperiment* package.

The object `se_rnaseq_refmat` contains pre-processed RNA-seq data from the study E-MTAB-1733 [1]. The data are available from the ArrayExpress [2] (<http://www.ebi.ac.uk/arrayexpress/>) database. The provided dataset contains gene expression profiles from 24 tissue types. Each tissue is represented by 3 replicates, except ovary which is represented by 2 replicates.

To download and load this dataset, run the following code:

```
> data("se_rnaseq_refmat")
> rnaseq_refmat <- assay(se_rnaseq_refmat)
> dim(rnaseq_refmat)
```

```
[1] 43819    71
```

The object *se_micro_refmat* contains normalized microarray data from the study GSE3526 [3]. The dataset is available from GEO [4] (<https://www.ncbi.nlm.nih.gov/geo/>). The provided dataset contains gene expression profiles from 26 tissues. Each tissue is represented by 3 replicates. To download and load this dataset, run the following code:

```
> data("se_micro_refmat")
> micro_refmat <- assay(se_micro_refmat)
> dim(micro_refmat)

[1] 54675    78
```

The object *se_rnaseq_testmat* contains pre-processed RNA-seq data derived from the study E-MTAB-513 [5]. The data are available from the ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) database. The provided dataset contains gene expression profiles from 12 tissues. To download and load this dataset, run the following code:

```
> data("se_rnaseq_testmat")
> rnaseq_testmat <- assay(se_rnaseq_testmat)
> dim(rnaseq_testmat)

[1] 43819    12
```

The object *se_micro_testmat* contains normalized microarray data derived from the study GSE2361 [6]. The dataset is available from GEO. The provided dataset contains gene expression profiles from 16 tissues. To download and load this dataset, run the following code:

```
> data("se_micro_testmat")
> micro_testmat <- assay(se_micro_testmat)
> dim(micro_refmat)

[1] 54675    78
```

3 Data pre-processing

The reads from the studies E-MTAB-1733 and E-MTAB-513 were mapped to the GRCh37 version of the human genome with Tophat v2.1.0 [7]. FPKM (fragments per kilobase of exon model per million mapped reads) values were calculated using cuffnorm v2.2.1 [8]. The used data from E-MTAB-1733 were extracted after processing of all samples and averaging across technical replicates.

The microarray data from the studies GSE3526 and GSE2361 were normalized using YuGene [9].

References

- [1] Linn Fagerberg, Björn M Hallström, Per Oksvold, Caroline Kampf, Dijana Djureinovic, Jacob Odeberg, Masato Habuka, Simin Tahmasebpour, Angelika Danielsson, Karolina Edlund, Anna Asplund, Evelina Sjöstedt, Emma Lundberg, Cristina Al-Khalili Szgyarto, Marie Skogs, Jenny Ottosson Takanen, Holger Berling, Hanna Tegel, Jan Mulder, Peter Nilsson, Jochen M Schwenk, Cecilia Lindskog, Frida Danielsson, Adil Mardinoglu, Asa

- Sivertsson, Kalle von Feilitzen, Mattias Forsberg, Martin Zwahlen, IngMarie Olsson, Sanjay Navani, Mikael Huss, Jens Nielsen, Fredrik Ponten, and Mathias Uhlén. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP*, 13(2):397–406, feb 2014. URL: <http://www.ncbi.nlm.nih.gov/pubmed/24309898><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3916642>, doi:10.1074/mcp.M113.035600.
- [2] Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra, and Susanna-Assunta Sansone. ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, 31(1):68–71, jan 2003. URL: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=165538&tool=pmcentrez&rendertype=abstract>.
- [3] Richard B Roth, Peter Hevezi, Jerry Lee, Dorian Willhite, Sandra M Lechner, Alan C Foster, and Albert Zlotnik. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7(2):67–80, may 2006. URL: <http://www.ncbi.nlm.nih.gov/pubmed/16572319>, doi:10.1007/s10048-006-0032-6.
- [4] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic acids research*, 35(Database issue):D760–5, jan 2007. URL: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1669752&tool=pmcentrez&rendertype=abstract>, doi:10.1093/nar/gkl887.
- [5] The illumina body map 2.0 data. <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/>.
- [6] Xijin Ge, Shogo Yamamoto, Shuichi Tsutsumi, Yutaka Midorikawa, Sigeo Ihara, San Ming Wang, and Hiroyuki Aburatani. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 86(2):127–41, aug 2005. URL: <http://www.sciencedirect.com/science/article/pii/S0888754305001114>, doi:10.1016/j.ygeno.2005.04.008.
- [7] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, may 2009. URL: <http://www.ncbi.nlm.nih.gov/pubmed/19289445><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2672628>, arXiv:9605103, doi:10.1093/bioinformatics/btp120.
- [8] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, may 2010. URL: <http://www.nature.com/doifinder/10.1038/nbt.1621><http://www.nature.com/nbt/journal/v28/n5/full/nbt.1621.html>{%}5Cn<http://www.nature.com/nbt/journal/v28/n5/pdf/nbt.1621.pdf>, arXiv:171, doi:10.1038/nbt.1621.
- [9] Kim-Anh Lê Cao, Florian Rohart, Leo Mchugh, Othmar Korn, and Christine A Wells. YuGene: A simple approach to scale gene expression data derived from different platforms for integrated analyses. *Genomics*, 103:239–251, 2014. doi:10.1016/j.ygeno.2014.03.001.