The patel2014gliohuman data user's guide

Will Townes (will.townes@gmail.com)   Stephanie C. Hicks (shicks@jimmy.harvard.edu)

Modified: January 23, 2017. Compiled: January 23, 2017

1 Overview

The patel2014gliohuman package contains gene expression data on 875 RNA-Seq samples from a study investigating tumor heterogeneity in five primary glioblastoma tumors [Patel et al. (2014)]. This study includes both single-cell RNA-Seq samples and bulk RNA-Seq samples from tumors and cells lines. Metadata was obtained from Gene Expression Omnibus (GSE57872) and raw FASTQ files were downloaded from Sequence Read Archive. The metadata includes the sequence identifier information provided in the header of the FASTQ files which can be used as a surrogate for batch. All samples were processed using Kallisto for gene expression quantification. The data provided are the estimated counts for ENSEMBL genes for all samples, not just the ones that were included in the authors' analysis. These samples have not been normalized or pre-processed. The data are provided as an SummarizedExperiment.

The data can be accessed as follows:

```
library(SummarizedExperiment)
library(patel2014gliohuman)
```

```
data(patel2014gliohuman)

# Get the expression data
counts = assay(patel_glio_2014)

## Loading required package:  Matrix

##
## Attaching package:  'Matrix'

## The following object is masked from 'package:S4Vectors':
##
##     expand

counts[1:5, 1:5]

## 5 x 5 sparse Matrix of class "dgCMatrix"
##                 MGH26_GSM1395399 MGH26_GSM1395400 MGH26_GSM1395401 MGH26_GSM1395402
## ENSG00000000003          1.10909          1.36451         233.6146          1.00000
## ENSG00000000005           .                .                 .                .
## ENSG00000000419         50.03170          1.00000          52.0000           .
## ENSG00000000457           .               1.06708            .                .
## ENSG00000000460          1.00000        277.14584            .               2.55005
##                 MGH26_GSM1395403
## ENSG00000000003              654
## ENSG00000000005                .
## ENSG00000000419                .
## ENSG00000000457                .
## ENSG00000000460                .

dim(counts)

## [1] 36579   875

# Get the pheno data
pdata = colData(patel_glio_2014)
head(pdata)

## DataFrame with 6 rows and 27 columns
##                              cell total_reads aligned_reads   sample          cell_id
##                          <factor>   <integer>     <integer> <factor>         <factor>
## MGH26_GSM1395399 MGH26_GSM1395399     3186452        409686    MGH26 MGH26_GSM1395399
## MGH26_GSM1395400 MGH26_GSM1395400     2141457        282986    MGH26 MGH26_GSM1395400
## MGH26_GSM1395401 MGH26_GSM1395401     2872050        279391    MGH26 MGH26_GSM1395401
## MGH26_GSM1395402 MGH26_GSM1395402     2340059         88594    MGH26 MGH26_GSM1395402
## MGH26_GSM1395403 MGH26_GSM1395403     3678419       1027176    MGH26 MGH26_GSM1395403
## MGH26_GSM1395404 MGH26_GSM1395404     3022848       1234108    MGH26 MGH26_GSM1395404
##                  pct_aligned detection        Run geo_accession  sampleName  sampleType
##                    <numeric> <integer>   <factor>      <factor> <character> <character>
## MGH26_GSM1395399  0.12857121      3662 SRR1294492    GSM1395399   MGH26_A01          SC
## MGH26_GSM1395400  0.13214648      5471 SRR1294493    GSM1395400   MGH26_A02          SC
## MGH26_GSM1395401  0.09727930      5134 SRR1294494    GSM1395401   MGH26_A03          SC
## MGH26_GSM1395402  0.03785973      2393 SRR1294495    GSM1395402   MGH26_A04          SC
## MGH26_GSM1395403  0.27924388      4981 SRR1294496    GSM1395403   MGH26_A05          SC
## MGH26_GSM1395404  0.40826002      4341 SRR1294497    GSM1395404   MGH26_A06          SC
##                    tumorName     cellType  subType includeSample avgLength Experiment
##                  <character>     <factor> <factor>     <logical> <integer>   <factor>
## MGH26_GSM1395399       MGH26 Glioblastoma       NA         FALSE        50  SRX549106
## MGH26_GSM1395400       MGH26 Glioblastoma      Pro          TRUE        50  SRX549107
## MGH26_GSM1395401       MGH26 Glioblastoma  Pro+Cla          TRUE        50  SRX549108
## MGH26_GSM1395402       MGH26 Glioblastoma       NA         FALSE        50  SRX549109
## MGH26_GSM1395403       MGH26 Glioblastoma  Pro+Cla          TRUE        50  SRX549110
## MGH26_GSM1395404       MGH26 Glioblastoma       NA         FALSE        50  SRX549111
##                     Sample    BioSample
##                   <factor>     <factor>
## MGH26_GSM1395399 SRS617086 SAMN02796848
## MGH26_GSM1395400 SRS617087 SAMN02796844
## MGH26_GSM1395401 SRS617088 SAMN02796850
## MGH26_GSM1395402 SRS617089 SAMN02796851
## MGH26_GSM1395403 SRS617090 SAMN02796849
## MGH26_GSM1395404 SRS617112 SAMN02796854
##                                                              download_path  instrument
##                                                                   <factor> <character>
## MGH26_GSM1395399 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294492  GLPB22-B5C
## MGH26_GSM1395400 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294493  GLPB22-B5C
## MGH26_GSM1395401 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294494  GLPB22-B5C
## MGH26_GSM1395402 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294495  GLPB22-B5C
## MGH26_GSM1395403 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294496  GLPB22-B5C
## MGH26_GSM1395404 http://sra-download.ncbi.nlm.nih.gov/srapub/SRR1294497  GLPB22-B5C
##                        runID        fcID      fcLane        tile       xtile       ytile
##                  <character> <character> <character> <character> <character> <character>
## MGH26_GSM1395399         556   H0PFYADXX           1        1101        1445        2150
## MGH26_GSM1395400         556   H0PFYADXX           1        1101        1007        2100
## MGH26_GSM1395401         556   H0PFYADXX           1        1101        2230        2191
## MGH26_GSM1395402         556   H0PFYADXX           1        1101        1169        2225
## MGH26_GSM1395403         556   H0PFYADXX           1        1101        1632        2094
## MGH26_GSM1395404         556   H0PFYADXX           1        1101        1245        2161
```

2 References

1. Patel et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344** (6190): 1396 - 1401. PMID: 24925914 PMCID: PMC4123637.