Overview of the scRNAseq dataset collection

Davide Risso

Created: May 25, 2016; Compiled: 2018-11-01

Contents

1 Raw data availability and accession codes

This package contains a collection of three publicly available single-cell RNA-seq datasets. The data were downloaded from NCBI's SRA or from EBI's ArrayExpress (see below for Accession numbers)

The dataset `fluidigm` contains 65 cells from (Pollen et al. 2014), each sequenced at high and low coverage (SRA: SRP041736).

The dataset `th2` contains 96 T helper cells from (Mahata et al. 2014) (ArrayExpress: E-MTAB-2512).

The dataset `allen` contains 379 cells from the mouse visual cortex. This is a subset of the data published in (Tasic et al. 2016) (SRA: SRP061902).

2 Pre-processing and summary

SRA files were downloaded from the Sequence Read Archive and the SRA Toolkit was used to transform them to FASTQ. FASTQ files were downloaded from EMBL-EBI ArrayExpress.

Reads were aligned with TopHat (v. 2.0.11) (Trapnell, Pachter, and Salzberg 2009) to the appropriate reference genome (GRCh38 for human samples, GRCm38 for mouse). RefSeq mouse gene annotation (GCF_000001635.23_GRCm38.p3) was downloaded from NCBI on Dec. 28, 2014. RefSeq human gene annotation (GCF_000001405.28) was downloaded from NCBI on Jun. 22, 2015.

featureCounts (v. 1.4.6-p3) (Liao, Smyth, and Shi 2013) was used to compute gene-level read counts and Cufflinks (v. 2.2.0) (Trapnell et al. 2010) was used to compute gene-leve FPKM's.

In parallel, reads were mapped to the transcriptome using RSEM (v. xx) (Li and Dewey 2011) to compute read counts and TPM's.

FastQC (v. 0.10.1) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)) and Picard (v. 1.128) (http://broadinstitute.github.io/picard (http://broadinstitute.github.io/picard)) were used to compute sample quality control (QC) metrics. (Picard's scripts `CollectRnaSeqMetrics`, `CollectAlignmentSummaryMetrics` and `CollectInsertSizeMetrics`).

Note that all the samples available in GEO and/or ArrayExpressed were included in the data object, including the samples that were excluded in the original publication because they did not pass QC.

3 Data format and metadata

The package contains each dataset in the form of `SummarizedExperiment` objects. To illustrate the format of each dataset, we will use the `fluidigm` data.

```
library(scRNAseq)
data(fluidigm)
fluidigm
```

```
## class: SummarizedExperiment
## dim: 26255 130
## metadata(3): sample_info clusters which_qc
## assays(4): tophat_counts cufflinks_fpkm rsem_counts rsem_tpm
## rownames(26255): A1BG A1BG-AS1 ... ZZEF1 ZZZ3
## rowData names(0):
## colnames(130): SRR1275356 SRR1274090 ... SRR1275366 SRR127526
1
## colData names(28): NREADS NALIGNED ... Cluster1 Cluster2
```

We can retrieve the gene expression measures by using the `assay` contstruct.

```
head(assay(fluidigm)[,1:3])
```

```
##            SRR1275356 SRR1274090 SRR1275251
## A1BG                0          0          0
## A1BG-AS1            0          0          0
## A1CF                0          0          0
## A2M                 0          0          0
## A2M-AS1             0          0          0
## A2ML1               0          0          0
```

`assay` will return the gene-level read counts. If we want to access the FPKM values, we need the following

```
head(assay(fluidigm, 2)[,1:3])
```

```
##            SRR1275356 SRR1274090 SRR1275251
## A1BG                0  0.0000000          0
## A1BG-AS1            0  0.3256690          0
## A1CF                0  0.0687904          0
## A2M                 0  0.0000000          0
## A2M-AS1             0  0.0000000          0
## A2ML1               0  1.3115300          0
```

Alternatively, we can use the `assays` accessor to get a list with both matrices.

```
names(assays(fluidigm))
```

```
## [1] "tophat_counts"  "cufflinks_fpkm" "rsem_counts"    "rsem_
tpm"
```

```
head(assays(fluidigm)$fpkm[,1:3])
```

```
## NULL
```

Note that the all the protein-coding genes are included in the expression matrix, even if they are not expressed in these samples, hence filtering of the non-expressed genes should be performed before any statistical analysis.

```
dim(fluidigm)
```

```
## [1] 26255   130
```

```
table(rowSums(assay(fluidigm))>0)
```

```
##
## FALSE  TRUE
##  9170 17085
```

In addition to the gene expression levels, the object contains some useful QC information, as well as the available annotation of the samples. This information can be accessed through the `colData` accessor.

```
colData(fluidigm)
```

```
## DataFrame with 130 rows and 28 columns
##               NREADS  NALIGNED    RALIGN TOTAL_DUP    PRIMER
INSERT_SZ
##            <numeric> <numeric> <numeric> <numeric> <numeric>
<numeric>
## SRR1275356  10554900   7555880   71.5862   58.4931 0.0217638
208
## SRR1274090    196162    182494   93.0323   14.5122 0.0366826
247
## SRR1275251   8524470   5858130   68.7213   65.0428 0.0351827
230
## SRR1275287   7229920   5891540   81.4884   49.7609 0.0208685
222
## SRR1275364   5403640   4482910   82.9609   66.5788 0.0298284
228
## ...              ...       ...       ...       ...       ...
...
## SRR1275259   5949930   4181040   70.2705   52.5975 0.0205253
224
## SRR1275253  10319900   7458710   72.2747   54.9637 0.0205342
207
## SRR1275285   5300270   4276650   80.6873   41.6394 0.0227383
222
## SRR1275366   7701320   6373600     82.76   68.9431 0.0266275
233
## SRR1275261  13425000   9554960   71.1727   62.0001 0.0200522
241
##            INSERT_SZ_STD COMPLEXITY     NDUPR PCT_RIBOSOMAL_B
ASES
##                <numeric>  <numeric> <numeric>          <nume
ric>
## SRR1275356            63   0.868928  0.343113               2
e-06
## SRR1274090           133   0.997655   0.93573
0
## SRR1275251            89   0.789252  0.201082
0
## SRR1275287            78     0.8981  0.538191
0
## SRR1275364            76   0.890693   0.39166
0
## ...                  ...        ...       ...
...
## SRR1275259            80   0.898898  0.399189               5
e-06
## SRR1275253            62   0.863618  0.344744
0
## SRR1275285            76   0.920068  0.638765               2
e-06
## SRR1275366            83   0.860359  0.343122
0
## SRR1275261           105   0.806833  0.234551
0
##            PCT_CODING_BASES PCT_UTR_BASES PCT_INTRONIC_BASES
##                   <numeric>     <numeric>          <numeric>
## SRR1275356         0.125806      0.180954           0.613229
## SRR1274090         0.309822      0.412917           0.205185
## SRR1275251         0.398461      0.473884           0.039886
## SRR1275287          0.19642      0.227592           0.498944
## SRR1275364         0.138617      0.210406           0.543941
## ...                     ...           ...                ...
## SRR1275259         0.261384      0.383665            0.26425
## SRR1275253         0.110732      0.190036           0.606814
## SRR1275285         0.143667      0.231103            0.54007
## SRR1275366         0.215696      0.307817           0.409437
## SRR1275261         0.408881      0.391068           0.147748
##            PCT_INTERGENIC_BASES PCT_MRNA_BASES MEDIAN_CV_COVE
RAGE
##                       <numeric>      <numeric>          <nume
ric>
## SRR1275356             0.080008        0.30676            1.4
9577
## SRR1274090             0.072076       0.722739            1.0
0758
## SRR1275251              0.08777       0.872345            1.2
4299
## SRR1275287             0.077044       0.424013           0.77
5981
## SRR1275364             0.107035       0.349024            1.4
4137
## ...                         ...            ...
...
## SRR1275259             0.090696       0.645049            1.1
0104
## SRR1275253             0.092418       0.300768            1.7
0169
## SRR1275285             0.085158        0.37477           0.71
4087
## SRR1275366              0.06705       0.523513            1.2
5198
## SRR1275261             0.052302       0.799949           0.93
9066
##            MEDIAN_5PRIME_BIAS MEDIAN_3PRIME_BIAS
##                     <numeric>          <numeric>
## SRR1275356                  0           0.166122
## SRR1274090           0.181742           0.698991
## SRR1275251                  0           0.340046
## SRR1275287           0.010251           0.350915
## SRR1275364                  0           0.204074
## ...                       ...                ...
## SRR1275259                  0            0.31555
## SRR1275253                  0           0.106902
## SRR1275285           0.019578           0.419987
## SRR1275366                  0           0.281554
## SRR1275261           0.000292           0.290117
##            MEDIAN_5PRIME_TO_3PRIME_BIAS sample_id.x
Lane_ID
##                               <numeric> <character>       <ch
aracter>
```

```
## SRR1275356                      1.03625   SRX534610 D24VYACXX
130502:4
## SRR1274090                      0.29351   SRX534823
1
## SRR1275251                     0.201518   SRX534623 D24VYACXX
130502:4
## SRR1275287                     0.292838   SRX534641 D24VYACXX
130502:1
## SRR1275364                     0.619863   SRX534614 D24VYACXX
130502:7
## ...                                 ...         ...
...
## SRR1275259                     0.350391   SRX534627 D24VYACXX
130502:4
## SRR1275253                     0.944856   SRX534624 D24VYACXX
130502:3
## SRR1275285                     0.194939   SRX534640 D24VYACXX
130502:1
## SRR1275366                     0.388272   SRX534615 D24VYACXX
130502:8
## SRR1275261                     0.384402   SRX534628 D24VYACXX
130502:3
##            LibraryName avgLength     spots Biological_Conditi
on
##            <character> <integer> <integer>          <characte
r>
## SRR1275356      GW16_2       202   9818076                 GW
16
## SRR1274090       NPC_9        60     95454                  N
PC
## SRR1275251      GW16_8       202   7935952                 GW
16
## SRR1275287    GW21+3_2       202   6531944               GW2
1+3
## SRR1275364     GW16_23       202   4919561                 GW
16
## ...                ...       ...       ...
...
## SRR1275259      GW21_3       202   5528916                 GW
21
## SRR1275253      GW16_9       202   9562204                 GW
16
## SRR1275285   GW21+3_16       202   4860721               GW2
1+3
## SRR1275366     GW16_24       202   7153688                 GW
16
## SRR1275261      GW21_4       202  12142387                 GW
21
##            Coverage_Type Cluster1 Cluster2
##              <character> <factor> <factor>
## SRR1275356          High     IIIb      III
## SRR1274090           Low       1a        I
## SRR1275251          High       NA      III
## SRR1275287          High       1c        I
## SRR1275364          High     IIIb      III
## ...                  ...      ...      ...
## SRR1275259          High       NA      III
## SRR1275253          High     IIIb      III
## SRR1275285          High      Iva       IV
## SRR1275366          High       NA      III
## SRR1275261          High       II       II
```

The first columns are related to sample quality, while other fields include information on the samples, provided by the original authors in their GEO/SRA submission and/or as Supplementary files in the pubblication.

Finally, the object contains a list of `metadata` that provide additional information on the experiment.

```
names(metadata(fluidigm))
```

```
## [1] "sample_info" "clusters"    "which_qc"
```

```
metadata(fluidigm)$which_qc
```

```
##  [1] "NREADS"                       "NALIGNED"
##  [3] "RALIGN"                       "TOTAL_DUP"
##  [5] "PRIMER"                       "INSERT_SZ"
##  [7] "INSERT_SZ_STD"                "COMPLEXITY"
##  [9] "NDUPR"                        "PCT_RIBOSOMAL_BASES"
## [11] "PCT_CODING_BASES"             "PCT_UTR_BASES"
## [13] "PCT_INTRONIC_BASES"           "PCT_INTERGENIC_BASES"
## [15] "PCT_MRNA_BASES"               "MEDIAN_CV_COVERAGE"
## [17] "MEDIAN_5PRIME_BIAS"           "MEDIAN_3PRIME_BIAS"
## [19] "MEDIAN_5PRIME_TO_3PRIME_BIAS"
```

In particular, in all datasets, the metadata list includes an element called `which_qc` that contains the names of the `colData` columns that relate to sample QC.

4 ERCC spike-ins

The `th2` and `allen` datasets contain the expression of the ERCC spike-ins. Note that these are **included in the same matrices** as the endogenous genes, hence users must use caution to avoid when using the data, to avoid mistreat external spike-ins as endogenous genes. One may wish to split the datasets in two, e.g.:

```
data(th2)
ercc_idx <- grep("^ERCC-", rownames(th2))
th2_endogenous <- th2[-ercc_idx,]
th2_ercc <- th2[ercc_idx,]

head(assay(th2_ercc)[,1:4])
```

```
##            ERR488983 ERR488967 ERR488989 ERR489021
## ERCC-00002      7775     14356      3868     15478
## ERCC-00003         1        75         1      2114
## ERCC-00004      1167      2468      1960      3914
## ERCC-00009       237         4      1167      1318
## ERCC-00012         0         0         0         0
## ERCC-00013         0         0         0         0
```

References

Li, B, and CN Dewey. 2011. “RSEM: Accurate Transcript Quantification from RNA-Seq Data with or Without a Reference Genome.” *BMC Bioinformatics* 12 (1):1.

Liao, Y, GK Smyth, and W Shi. 2013. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics*, btt656.

Mahata, B, X Zhang, AA Kolodziejczyk, V Proserpio, L Haim-Vilmovsky, AE Taylor, D Hebenstreit, et al. 2014. “Single-Cell RNA Sequencing Reveals T Helper Cells Synthesizing Steroids de Novo to Contribute to Immune Homeostasis.” *Cell Reports* 7 (4):1130–42.

Pollen, AA, TJ Nowakowski, J Shuga, X Wang, AA Leyrat, JH Lui, N Li, et al. 2014. “Low-Coverage Single-Cell mRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex.” *Nature Biotechnology* 32 (10):1053–8.

Tasic, B, V Menon, TN Nguyen, TK Kim, T Jarsky, Z Yao, B Levi, et al. 2016. “Adult Mouse Cortical Cell Taxonomy Revealed by Single Cell Transcriptomics.” *Nature Neuroscience* 19:335–46.

Trapnell, C, L Pachter, and SL Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics* 25 (9):1105–11.

Trapnell, C, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ Van Baren, SL Salzberg, BJ Wold, and L Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching During Cell Differentiation.” *Nature Biotechnology* 28 (5):511–15.