rCube - RNA-Rates in R

Leonhard Wachutka[1,*], Carina Demel[2], Julien Gagneur[1]

[1] Department of Informatics, Technical University of Munich, Munich, Germany

[2] Max Planck Institute for biophysical Chemistry, Göttingen, Germany

[*] wachutka (at) in.tum.de

July 25, 2017

Abstract

A common problem of RNA-Seq experiments is the normalization of different samples, especially when they were also treated differently. A special case is the 4sU-labeling followed by deep-sequencing, where two libraries are generated from each RNA fraction: First, in one part of the fraction, the 4sU-labeled RNAs are pulled down and gives the labeled ("L") fraction of the experiment for library preparation. Second, the total RNA (labeled and unlabeled, "T") is used for the preparation of another RNA-Seq library. As the RNA amount for library preparation is usually stringent, the sequencing results do not reflect reality. Here, we have to account for the sequencing depth of different samples and especially adjust the ratio of Labeled to Total RNA-Seq libraries. Additionally, the labeled RNA extraction is not perfect and also a little bit of unlabeled RNA may contaminate the labeled fraction. As the gene-expression also always varies a little bit for biological samples, read counts from real genes might lead to confusing estimations. Therefore, we apply our normalization approach only to artificial spike-ins from the External RNA Control Consortium (ERCC), for which the initial amount of each spike-in is known and the same across all samples. Some of the spike-ins were 4sU-labeled by in-vitro transcription, so the cross-contamination of unlabeled spike-ins in labeled samples can be monitored. The goal of this package is to reliably estimate sequencing depths and cross-contamination rates per sample, given 4sU- and total RNA-Seq data.

rCube version: 1.1.0

If you use rCube in published research, please cite:

L. Wachutka, C. Demel, J. Gageur:

Contents

1 Background

As described in .... 4sU-seq allows to monitor changes in the RNA metabolism. If cells are exposed to 4sU, they rapidly take up this Uridine analog and incorporate it into newly-synthesized RNAs. This way, newly-synthesized RNAs are labeled and can be extracted from the total RNA in the sample. The longer the labeling time, e.i. the time from 4sU addition to harvesting the cells, the bigger is the proportion of labeled RNAs among all RNAs.

2 Getting started

This vignette provides a pipeline how to... starting from BAM files... You will learn how to estimate sample specific sequencing depths and cross-contamination rates from spike-in counts. These values can be used to normalize gene expression values obtained by RNA-Seq and thus estimate gene-specific synthesis and degradation rates. By extracting reads spanning junctions, splicing times can be estimated. For more robust estimation, multiple samples with different labeling times are taken into account. Before starting, the package must be loaded by:

```
library("rCube")
```

2.1 Input Data

The rCube package works on *rCubeExperiment* containers, that rely on the *SummerizedExperiment* class. The `rowRanges` of the *rCubeExperiment* is a *GRanges* object of features, for which RNA rates should be estimated. Experimental sample information can be either provided by a design matrix or this information can be extracted from the BAM-file names (when they fulfil the required structure). Then, an empty *rCubeExperiment*,e.g. for the artificial spike-ins can be constructed as follows:

```
data("spikeins")
data("spikeinLabeling")
data("spikeinLengths")
data("designMatrix")

spikeinCounts <- setupExperimentSpikeins(rows=spikeins,
                                         designMatrix=designMatrix,
                                         length=spikeinLengths,
                                         labelingState=spikeinLabeling)
```

The `setupExperiment` can be used analogically for genes/exons/introns/junctions. Here, only the `rows` and either `designMatrix` or `files` has to be set. See also

The individual information from the *rCubeExperiment* can be assessed by:

```
# feature information
rowRanges(spikeinCounts)
```

```
## GRanges object with 6 ranges and 9 metadata columns:
##           seqnames    ranges strand |       source       type     score
##              <Rle> <IRanges>  <Rle> |     <factor>   <factor> <numeric>
##    Spike2    chrS2 [1,  982]      + | Fruehauf2013 transcript      <NA>
##   Spike12   chrS12 [1,  947]      + | Fruehauf2013 transcript      <NA>
##    Spike4    chrS4 [1, 1011]      + | Fruehauf2013 transcript      <NA>
##    Spike5    chrS5 [1, 1012]      + | Fruehauf2013 transcript      <NA>
##    Spike8    chrS8 [1, 1076]      + | Fruehauf2013 transcript      <NA>
##    Spike9    chrS9 [1, 1034]      + | Fruehauf2013 transcript      <NA>
##               phase     gene_id transcript_id    length labelingState
```

```
##           <integer> <character>   <character> <numeric>      <factor>
##    Spike2      <NA>      Spike2        Spike2      1023          TRUE
##   Spike12      <NA>     Spike12       Spike12      1023         FALSE
##    Spike4      <NA>      Spike4        Spike4      1033          TRUE
##    Spike5      <NA>      Spike5        Spike5      1042         FALSE
##    Spike8      <NA>      Spike8        Spike8      1124          TRUE
##    Spike9      <NA>      Spike9        Spike9      1061         FALSE
##           labeledSpikein
##                <logical>
##    Spike2          FALSE
##   Spike12          FALSE
##    Spike4          FALSE
##    Spike5          FALSE
##    Spike8          FALSE
##    Spike9          FALSE
##   -------
##   seqinfo: 6 sequences from an unspecified genome; no seqlengths
```

```
# sample information
colData(spikeinCounts)
```

```
## DataFrame with 8 rows and 5 columns
##           sample condition       LT labelingTime replicate
##         <factor>  <factor> <factor>    <numeric>  <factor>
## A_L_5_1  A_L_5_1         A        L            5         1
## A_L_5_2  A_L_5_2         A        L            5         2
## B_L_5_1  B_L_5_1         B        L            5         1
## B_L_5_2  B_L_5_2         B        L            5         2
## A_T_5_1  A_T_5_1         A        T            5         1
## A_T_5_2  A_T_5_2         A        T            5         2
## B_T_5_1  B_T_5_1         B        T            5         1
## B_T_5_2  B_T_5_2         B        T            5         2
```

```
# read counts
assay(spikeinCounts)
```

```
##         A_L_5_1 A_L_5_2 B_L_5_1 B_L_5_2 A_T_5_1 A_T_5_2 B_T_5_1 B_T_5_2
## Spike2       NA      NA      NA      NA      NA      NA      NA      NA
## Spike12      NA      NA      NA      NA      NA      NA      NA      NA
## Spike4       NA      NA      NA      NA      NA      NA      NA      NA
## Spike5       NA      NA      NA      NA      NA      NA      NA      NA
## Spike8       NA      NA      NA      NA      NA      NA      NA      NA
## Spike9       NA      NA      NA      NA      NA      NA      NA      NA
```

All RNA rate estimations of this package rely on read counts. These can be either provided as count matrices, or read counts can be obtained from BAM files using the rCube pipeline.

```
#TODO
```

3 Normalizing RNA-Seq samples using spike-in counts

The sample specific parameters like sequencing depth and cross-contamination rate are estimated from spike-in counts only. Therefore, we fit a generalized linear model (GLM) of the Negative Binomial family with a log link function. The response of the GLM are the observed spike-in counts, and the terms that specify the linear predictor of the response are comprised of:

- a sample specific factor (that reflects the sample specific sequencing depth),
- a labeled sample specific factor (that reflects the control for cross contamination (only estimated for unlabeled spike-ins in labeled samples)), and
- a spike-in specific factor to allow for some spike-in specific variation e.g. due to sequence biases.

Additionally, the length of each spike-in is used as an offset, i.e. a known slope for the covariate.

```
data(geneCounts)
data(spikeinCounts)
geneCounts <- estimateSizeFactors(geneCounts, spikeinCounts, method="spikeinGLM")
colnames(colData(geneCounts))
```

```
## [1] "sample"              "condition"           "LT"
## [4] "labelingTime"        "replicate"           "sequencing.depth"
## [7] "cross.contamination"
```

```
geneCounts$sequencing.depth
```

```
## A_L_5_1 A_L_5_2 B_L_5_1 B_L_5_2 A_T_5_1 A_T_5_2 B_T_5_1 B_T_5_2
## 1.00000 0.90453 0.08914 1.10879 0.16871 0.00919 0.01453 0.14209
```

```
geneCounts$cross.contamination
```

```
## A_L_5_1 A_L_5_2 B_L_5_1 B_L_5_2 A_T_5_1 A_T_5_2 B_T_5_1 B_T_5_2
## 0.01200 0.00445 0.00721 0.00411 1.00000 1.00000 1.00000 1.00000
```

Note, the cross-contamination value for all total RNA-seq samples is 1, as 100% of the unlabeled RNAs are supposed to be in the sample. Additional fitting results are stored in the `metadata` of the resulting *rCubeExperiment* object.

```
metadata(geneCounts)
```

4 Estimating gene-specific synthesis and degradation rates

Using the sample-specific values for sequencing depth and cross-contamination as estimated in the previous section, we can now normalize all the samples. It is especially important to bring Labeled and Total samples to comparable scales. Labeled and Total samples can be sequenced at the same depth, and the same amount of RNA is used for library preparation, but the resulting read counts do not reflect the true ratio of Labeled and Total RNAs in the cells, where the amount of newly-synthesized, labeled RNA should be much less than the total RNA amount. Therefore it is necessary to upscale the Total RNA read counts compared to the Labeled RNA read counts.

4.1 Providing gene-wise dispersion estimates

Usually, read counts in different RNA-seq samples underly fluctuations due to biological or technical variances. To take these fluctuations into account, we estimate each gene's dispersion. For each gene, a single dispersion estimate for all 4sU-Seq samples and for all Total RNA-Seq samples is needed. Here, we can use the method provided in the DESeq2 package [1] The wrapper function estimateGeneDispersion applies the DESeq algorithm to all genes, while separating the count table according to the RNA-Seq protocol (labeled or total RNA). It is possible to choose between all provided DESeq dispersion estimates, namely the genewise maximum likelihood dispersion estimate ("dispGeneEst"), the smooth curve fitted through the gene-wise disperion estimates ("dispFit") and the genewise dispersion estimates shrunken towards the fitted curve ("dispMAP", default). The input for the estimateGeneDispersion function is therefore only a n*m matrix, providing count data for n genes under m conditions, and a vector of length m, indiciating for each condition if it was a labeled 4sU-Seq sample ("L") or a total RNA-Seq sample ("T"). The function returns a matrix consisting of n rows and 2 columns, the Labeled ("L") and Total ("T") dispersion estimates for all n genes.

TODO BEISPIEL

4.2 Gene-wise synthesis and degratation rate estimates

Plug-in sequencing depth and cross-contamination as estimated in Section 3 and gene-wise dispersion estimates as described in Section 4.1.

5 Estimating splicing times from junction read counts

6 Session Information

This vignette was generated using the following package versions:

```
sessionInfo()
```

```
## R version 3.4.1 (2017-06-30)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: OS X El Capitan 10.11.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
```

```
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] rCube_1.1.0               SummarizedExperiment_1.7.5
##  [3] DelayedArray_0.3.16       matrixStats_0.52.2
##  [5] Biobase_2.37.2            GenomicRanges_1.29.11
##  [7] GenomeInfoDb_1.13.4       IRanges_2.11.12
##  [9] S4Vectors_0.15.5          BiocGenerics_0.23.0
## [11] knitr_1.16
##
## loaded via a namespace (and not attached):
##  [1] locfit_1.5-9.1          Rcpp_0.12.12            lattice_0.20-35
##  [4] Rsamtools_1.29.0        Biostrings_2.45.3       rprojroot_1.2
##  [7] digest_0.6.12           plyr_1.8.4              backports_1.1.0
## [10] acepack_1.4.1           RSQLite_2.0             evaluate_0.10.1
## [13] highr_0.6               ggplot2_2.2.1           zlibbioc_1.23.0
## [16] rlang_0.1.1             lazyeval_0.2.0          data.table_1.10.4
## [19] annotate_1.55.0         blob_1.1.0              rpart_4.1-11
## [22] Matrix_1.2-10           checkmate_1.8.3         rmarkdown_1.6
## [25] splines_3.4.1           BiocParallel_1.11.4     geneplotter_1.55.0
## [28] stringr_1.2.0           foreign_0.8-69          htmlwidgets_0.9
## [31] RCurl_1.95-4.8          bit_1.1-12              munsell_0.4.3
## [34] compiler_3.4.1          base64enc_0.1-3         htmltools_0.3.6
## [37] nnet_7.3-12             tibble_1.3.3            gridExtra_2.2.1
## [40] htmlTable_1.9           GenomeInfoDbData_0.99.1 Hmisc_4.0-3
## [43] XML_3.98-1.9            MASS_7.3-47             GenomicAlignments_1.13.4
## [46] bitops_1.0-6            grid_3.4.1              xtable_1.8-2
## [49] gtable_0.2.0            DBI_0.7                 magrittr_1.5
## [52] scales_0.4.1            stringi_1.1.5           XVector_0.17.0
## [55] genefilter_1.59.0       latticeExtra_0.6-28     Formula_1.2-2
## [58] BiocStyle_2.5.8         RColorBrewer_1.1-2      tools_3.4.1
## [61] bit64_0.9-7             DESeq2_1.17.11          survival_2.41-3
## [64] yaml_2.1.14             AnnotationDbi_1.39.1    colorspace_1.3-2
## [67] cluster_2.0.6           memoise_1.1.0
```

7 References

[1] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.