RNAprobR: An R package for analysis of the massive parallel sequencing based methods of RNA structure probing

Lukasz Jan Kielpinski, Nikos Sidiropoulos, Jeppe Vinther

Modified: 27 October, 2014. Compiled: April 24, 2017

Contents

1 Introduction

RNA structure probing is more and more often conducted with the use of high-throughput sequencing. Analysis of data coming from those experiments can be challenging and time consuming. Here we describe an R package - *RNAprobR* - which intends to standardize and simplify processing of experiments for which information on RNA susceptibility to different probing conditions is encoded in a location of sequenced reads termini.

2 Galaxy workflow

RNAprobR takes as input "Unique Barcodes" and "k2n" (or "Read Counts") files generated by "RNA probing" Galaxy workflow [1]. The package comes with sample data from HRF-Seq experiment [2]. This data was generated from raw sequencing reads (available under: http://people.binf.ku.dk/jvinther/data/HRF-Seq/) with the Galaxy workflow specifying: barcode sequence to NNNNNNN, 3' trimming length to 0 and using trimming untemplated nucleotides.

3 RNAprobR workflow

The package was designed with a specific processing workflow in mind (Fig. 1). It reads-in the Unique Barcodes files and calculates EUCs (`readsamples()`), compiles positional information based on sequenced fragments (`comp()`), performs data normalization (`dtcr()`, `slograt()`, `swinsor()`) and exports data in various formats (`GR2norm_df()`, `plotRNA()`, `norm2bedgraph()`)

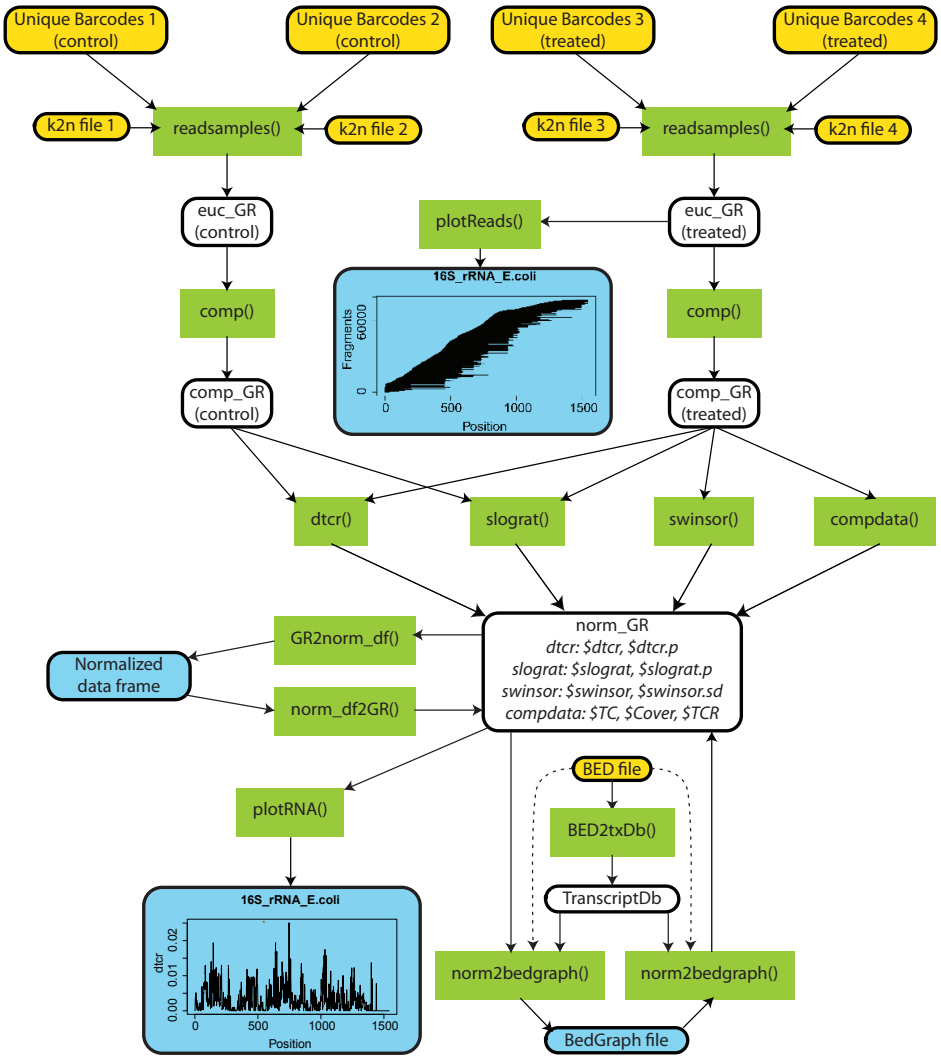


Figure 1: RNAprobR workflow

4 EUC Calculation

Proposed data analysis takes advantage of removing PCR duplicates by looking at the sequences of random barcodes attached to the beginning of each read. Counting each unique barcode only once is correct for fragments present in low counts range but leads to erroneous quantification in the high range (higher probability of physically distinct barcodes bearing the same sequence). We define Estimated Unique Counts (EUCs) as the number of molecules expected to give rise

to the observed number of unique barcodes. In function `readsamples()` we have implemented two methods to calculate EUCs: `euc="Fu"` [3] or `euc="HRF-Seq"` (Kielpinski and Vinther, 2014). The latter requires supplying precomputed k2n files (generated by "RNA probing" Galaxy workflow if "Produce k2n file" checked or `k2n_func()` function from the package).

Note: if no random barcode was used in experiment, "Unique Barcodes" and "Read Counts" files are identical and each can be used for subsequent analysis. In this case for the `readsamples()` function use option: `euc="counts"`.

5 Reading in datasets

The first step in data processing is reading in the Unique Barcodes files, combining samples of the same treatment (if we had e.g. repeated control) and calculating EUCs. All of those steps are performed by `readsamples()` function. Start with specifying paths to different Unique Barcodes files and their respective k2n files (order matters!), followed by reading-in the data and calculating EUCs according to HRF-Seq method:

```
> library(RNAprobR)
```

```
> treated <- c(system.file("extdata", "unique_barcodes14.gz", package="RNAprobR"),
+              system.file("extdata", "unique_barcodes22.gz", package="RNAprobR"))
> control <- c(system.file("extdata", "unique_barcodes16.gz", package="RNAprobR"),
+              system.file("extdata", "unique_barcodes24.gz", package="RNAprobR"))
> k2n_treated <- c(system.file("extdata", "k2n_14", package="RNAprobR"),
+                  system.file("extdata", "k2n_22", package="RNAprobR"))
> k2n_control <- c(system.file("extdata", "k2n_16", package="RNAprobR"),
+                  system.file("extdata", "k2n_24", package="RNAprobR"))
> control_euc <- readsamples(control, euc="HRF-Seq", k2n_files=k2n_control)
> treated_euc <- readsamples(treated, euc="HRF-Seq", k2n_files=k2n_treated)
```

`control_euc` and `treated_euc` are *GenomicRanges* (GRanges) objects holding information on sequenced fragments span and EUC.

6 Compiling positional data

`comp()` function takes as input an object imported by `readsamples()` function and uses it to compile needed data for each nucleotide in analyzed RNA molecules. Specifically it computes:

- termination count (TC),
- coverage (Cover),
- termination-coverage ratio (TCR),
- priming count (PC).

If one specifies path to FASTA file which was used for mapping (option: `fasta_file`) it also adds nucleotide identity (nt). By specifying "cutoff" value, function discards fragments which are shorter than the provided value. Usage:

```
> treated_comp <- comp(treated_euc, cutoff=101, fasta_file =
+                          system.file("extdata", "hrfseq.fa", package="RNAprobR"))
> control_comp <- comp(control_euc, cutoff=101, fasta_file =
+                          system.file("extdata", "hrfseq.fa", package="RNAprobR"))
```

`treated_comp` and `control_comp` are *GRanges* objects with each range being of length 1.

7 Normalization

Positionally compiled GRanges objects can be used as input for normalization. In this package we have implemented three normalization functions:

- `dtcr()`, which performs ΔTCR normalization as described in (Kielpinski and Vinther, 2014),
- `slograt()`, which calculates smooth log-ratio as described in (Wan et al., 2014),
- `swinsor()`, which calculates smooth Winsor normalization. First it calculates Winsorized values as described in (Rouskin et al., 2013) but in 1-nt sliding windows, and then for each nucleotide it returns mean and standard deviation of all predictions that overlapped given nucleotide.

Single *GRanges* object can hold data normalized by all of the abovementioned methods - usually the first call of normalization functions will create the GRanges object and subsequent calls will add the data to already existing object (via `add_to` option). Let's normalize HRF-Seq data with all three methods:

```
> hrfseq_norm <- dtcr(control_GR=control_comp, treated_GR=treated_comp,
+                     window_size=3, nt_offset=1)
```

```
> hrfseq_norm <- slograt(control_GR=control_comp, treated_GR=treated_comp,
+                        add_to=hrfseq_norm)
```

```
> hrfseq_norm <- swinsor(Comp_GR=treated_comp, add_to=hrfseq_norm)
```

One can add data from compiled *GRanges* object (TC, Cover, TCR) to normalized *GRanges* object with `compdata()` function:

```
> hrfseq_norm <- compdata(Comp_GR=treated_comp, add_to=hrfseq_norm)
```

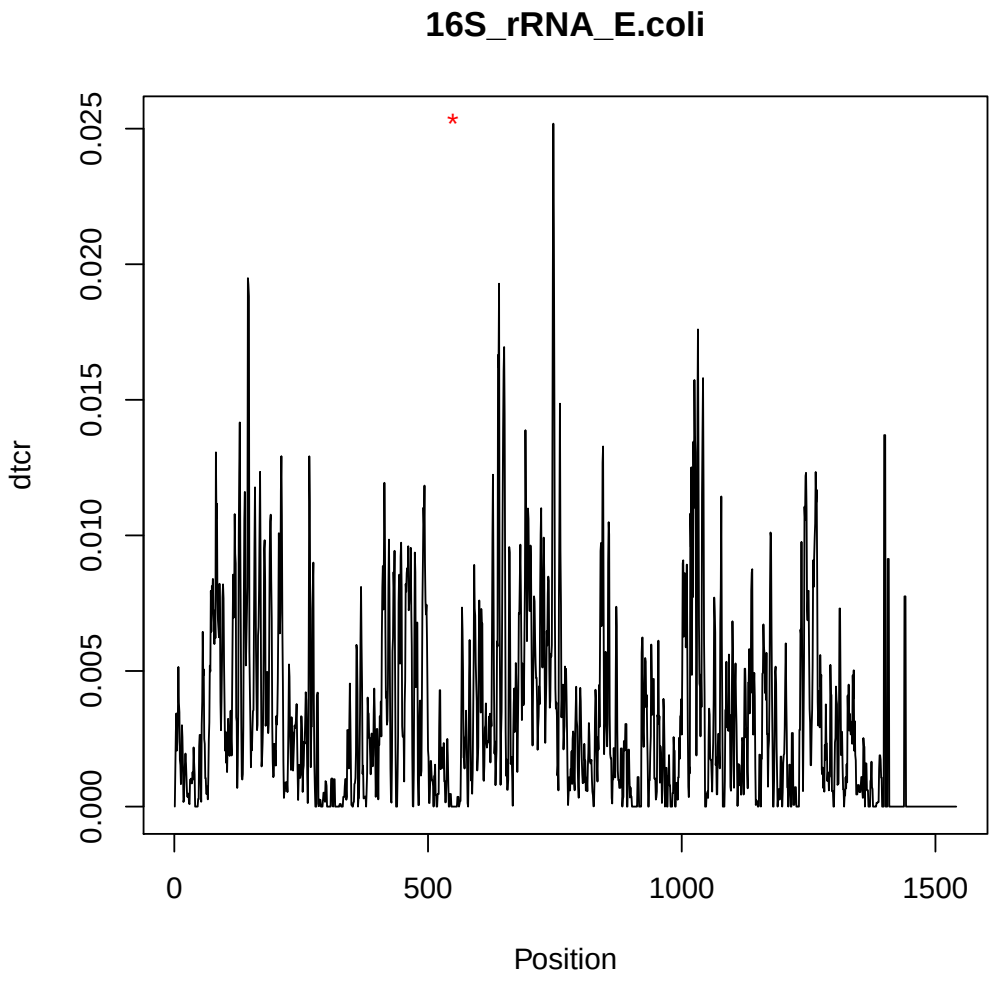
8 Export

The package allows for data export in multiple formats. First, let say we are interested in obtaining the slograt normalized table for RNase_P. Simply type:

```
> norm_df <- GR2norm_df(hrfseq_norm, RNAid = "RNase_P", norm_methods = "slograt")
```

Or, we could make a plot of ΔTCR values over 16S rRNA:

```
> plotRNA(norm_GR=hrfseq_norm, RNAid="16S_rRNA_E.coli", norm_method="dtcr")
```



At last, if we want to visualize the data in UCSC Genome Browser we can generate the BedGraph file:

```
> RNase_P_BED <- system.file("extdata", "RNaseP.bed", package="RNAprobR")
> norm2bedgraph(hrfseq_norm, bed_file = RNase_P_BED, norm_method = "dtcr",
+               genome_build = "baciSubt2", bedgraph_out_file = "RNaseP_dtcr",
+               track_name = "dtcr", track_description = "deltaTCR Normalization")
```

9 Session Info

```
R version 3.4.0 (2017-04-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.2 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=C               LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats4    parallel  stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] RNAprobR_1.8.0         plyr_1.8.4             GenomicFeatures_1.28.0
 [4] AnnotationDbi_1.38.0   Biobase_2.36.0         GenomicRanges_1.28.0
 [7] GenomeInfoDb_1.12.0    IRanges_2.10.0         S4Vectors_0.14.0
[10] BiocGenerics_0.22.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.10               compiler_3.4.0             XVector_0.16.0
 [4] bitops_1.0-6               tools_3.4.0                zlibbioc_1.22.0
 [7] biomaRt_2.32.0             digest_0.6.12              lattice_0.20-35
[10] evaluate_0.10              RSQLite_1.1-2              memoise_1.1.0
[13] Matrix_1.2-9               DelayedArray_0.2.0         DBI_0.6-1
[16] yaml_2.1.14                GenomeInfoDbData_0.99.0    rtracklayer_1.36.0
[19] stringr_1.2.0              knitr_1.15.1               Biostrings_2.44.0
[22] grid_3.4.0                 rprojroot_1.2              XML_3.98-1.6
[25] BiocParallel_1.10.0        rmarkdown_1.4              magrittr_1.5
[28] backports_1.0.5            Rsamtools_1.28.0           htmltools_0.3.5
[31] matrixStats_0.52.2         GenomicAlignments_1.12.0   SummarizedExperiment_1.6.0
[34] BiocStyle_2.4.0            stringi_1.1.5              RCurl_1.95-4.8
```

References

[1] Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86, January 2010. URL: http://genomebiology.com/2010/11/8/R86, doi:10.1186/gb-2010-11-8-r86.

[2] Lukasz Jan Kielpinski and Jeppe Vinther. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic acids research*, 42(8):e70, April 2014. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4005689&tool=pmcentrez&rendertype=abstract, doi:10.1093/nar/gku167.

[3] Glenn K Fu, Jing Hu, Pei-Hua Wang, and Stephen P A Fodor. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9026–31, May 2011. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3107322&tool=pmcentrez&rendertype=abstract, doi:10.1073/pnas.1017621108.