This report was automatically generated with the R package **knitr** (version 1.20).

```
---
title: "endoSeqR Practice"
author: "Kelly Daescu"
date: "October 23, 2018"
output: html_document
---
```

```
## Error: <text>:8:0: unexpected end of input
## 6: ---
## 7:
##   ^
```

This tutorial will implement the endoSeqR pipeline on a small RNA dataset. The dataset analyzed is from the following paper: Asikainen, S., et al., Functional characterization of endogenous siRNA target genes in Caenorhabditis elegans. BMC Genomics, 2008. 9: p. 270.

The R package can be cloned from github @ https://github.com/kjdaescu/endoSeqR. The bowtie builds and dataset and information can be obtained from google drive at the following links: https://drive.google.com/drive/folders/16xoPxbzVu9Zo9_gqF16umHO_swB91kxV?usp=sharing. Available species are C elegans, S scrofa, D melanogaster, D rerio, H sapiens, M musculus, G gallus, A thaliana, X tropicalis.

Must have the following dependencies: Linux: samtools, bedtools, bowtie R: openxlsx, Biostrings, plyr

First time only - Installation

library(devtools) install(“endoSeqR”, lib=“.”)

To use, first load the package and its dependencies.

```
library(endoSeqR)
Package_Load()
```

```
## Loading required package: BiocGenerics
```

```
## Loading required package: parallel
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, lengths, Map, mapply, match, mget, order,
##     paste, pmax, pmax.int, pmin, pmin.int, Position, rank, rbind,
##     Reduce, rowMeans, rownames, rowSums, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which, which.max,
##     which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:plyr':
##
##     rename
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
## Loading required package: IRanges
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:plyr':
##
##     desc
```

```
## Loading required package: XVector
```

```
##
## Attaching package: 'XVector'
```

```
## The following object is masked from 'package:plyr':
##
##     compact
```

```
##
## Attaching package: 'Biostrings'
```

```
## The following object is masked from 'package:base':
##
##     strsplit
```

For bowtie to work, the parent path of the index files must be specified. The index suffixes are fixed (genomic.index.ebwt, transcriptomic.index.ebwt, reference.index.ebwt), so all that is needed is the path and species. The fasta/fastq/text file of read sequences must also be specified.

```
index<-paste0("/cb/130/kxr131930/endo.psirbase/endoSeqR_Indexes/elegans/ele
generate_fasta("cel.Asikainin_2008.endosiRNA.fasta","Mixed_Stage")
```

```
## [1] "Read length distribution:"
##     x freq
## 1  18  108
## 2  19  201
## 3  20  349
## 4  21  654
## 5  22 1000
## 6  23  270
## 7  24  112
## 8  25  122
## 9  26  286
## 10 27   10
## 11 28    1
## [1] "Starting reads:"
## [1] ">Mixed_Stage_1"        "AAAAAGAGTGGTAGAGTGTCG" ">Mixed_Stage_2"
## [4] "AAAAAGGAAAGATTTGGATTG" ">Mixed_Stage_3"        "AAAAATTCCTGCTTTGTGC
```

Once the package is loaded, the index is specified, and the reads are fasta formatted, the next step is to find the endo-siRNA genomic coordinates and sequences using the Map_endosiRNA function.

```
Map_endosiRNA(index, "Mixed_Stage")
```

```
## arguments 'show.output.on.console', 'minimized' and 'invisible' are for
```

To easily access the genomic coordinates and endo-siRNA sequences, xlsx file and ".RData" files will be generated using the Import_endosiRNA command. The endosiRNA reverse complementary sequence will also be included.

```
df<-Import_endosiRNA("Mixed_Stage")
```

```
## [1] "Length siRNA"
##     x freq
## 1  18  108
## 2  19  201
## 3  20  349
## 4  21  654
## 5  22 1000
## 6  23  270
## 7  24  112
## 8  25  122
## 9  26  286
## 10 27   10
## 11 28    1
```

```
colnames(df)
```

```
## [1] "Read_ID"     "Sequence"    "Length"      "Chr"         "Start"
## [6] "Stop"        "Strand"      "target.mRNA"
```

```
head(df)
```

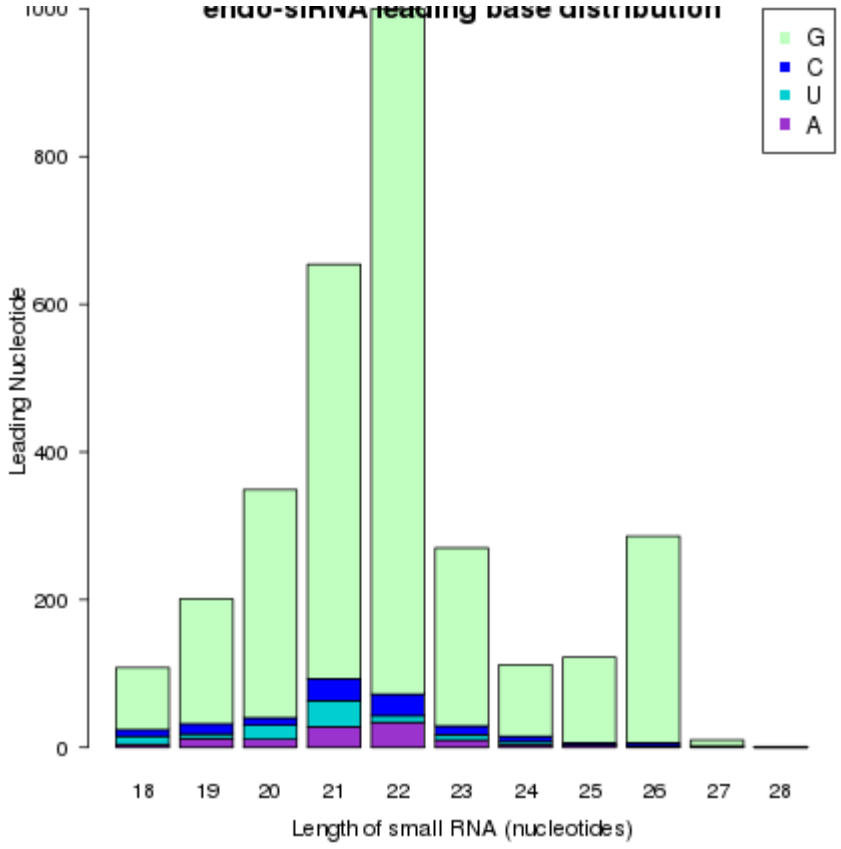
```
##            Read_ID                Sequence Length Chr    Start     Stop
## 1    Mixed_Stage_1  AAAAAGAGTGGTAGAGTGTCG     21  II 11619330 11619351
## 2   Mixed_Stage_10  AAATTTCTTGAAACATCTCCC     21  IV 17217240 17217261
## 3  Mixed_Stage_100 ATTCACCAGACCATTGTTTCCA     22   V  5183293  5183315
## 4 Mixed_Stage_1000  GATATTTTCGGCGAACCTCGA     21  II  1843840  1843861
## 5 Mixed_Stage_1001  GATCAATCGATAGCTCGGAAC     21 III  1880643  1880664
## 6 Mixed_Stage_1002  GATCAATTCCTTGATACTTGG     21  IV 16240081 16240102
##   Strand            target.mRNA
## 1      +  CGACACTCTACCACTCTTTTT
## 2      -  GGGAGATGTTTCAAGAAATTT
## 3      - TGGAAACAATGGTCTGGTGAAT
## 4      -  TCGAGGTTCGCCGAAAATATC
## 5      -  GTTCCGAGCTATCGATTGATC
## 6      -  CCAAGTATCAAGGAATTGATC
```

Finally, in order to visualize the endo-siRNA population, the PlotNTLetter command can be run, as demonstrated below.

```
PlotNTLetter(df)
```

```
## [1] "endo-siRNA length:  18"
## [1] "leading base distribution"
##   N A  U C  G
## 1 0 4 11 9 84
## [1] "endo-siRNA length:  19"
## [1] "leading base distribution"
##   N  A U  C   G
## 1 0 12 6 14 169
## [1] "endo-siRNA length:  20"
## [1] "leading base distribution"
##   N  A  U  C   G
## 1 0 11 19 10 309
## [1] "endo-siRNA length:  21"
## [1] "leading base distribution"
##   N  A  U  C   G
## 1 0 28 35 30 561
## [1] "endo-siRNA length:  22"
## [1] "leading base distribution"
##   N  A U  C   G
## 1 0 34 9 29 928
## [1] "endo-siRNA length:  23"
## [1] "leading base distribution"
##   N  A U  C   G
## 1 0 10 7 12 241
## [1] "endo-siRNA length:  24"
## [1] "leading base distribution"
##   N A U C  G
## 1 0 4 4 7 97
## [1] "endo-siRNA length:  25"
## [1] "leading base distribution"
##   N A U C   G
## 1 0 3 0 3 116
## [1] "endo-siRNA length:  26"
## [1] "leading base distribution"
##   N A U C   G
## 1 0 2 0 4 280
## [1] "endo-siRNA length:  27"
## [1] "leading base distribution"
##   N A U C G
## 1 0 1 1 0 8
## [1] "endo-siRNA length:  28"
## [1] "leading base distribution"
##   N A U C G
## 1 0 0 0 0 1
##    x.N x.A x.U x.C x.G freq
## 1    0   0   0   0   1    1
```

```
## 2     0    1    1    0    8     1
## 3     0    2    0    4  280     1
## 4     0    3    0    3  116     1
## 5     0    4    4    7   97     1
## 6     0    4   11    9   84     1
## 7     0   10    7   12  241     1
## 8     0   11   19   10  309     1
## 9     0   12    6   14  169     1
## 10    0   28   35   30  561     1
## 11    0   34    9   29  928     1
##    18  19  20  21  22  23 24  25  26 27 28
## N   0   0   0   0   0   0  0   0   0  0  0
## A   4  12  11  28  34  10  4   3   2  1  0
## U  11   6  19  35   9   7  4   0   0  1  0
## C   9  14  10  30  29  12  7   3   4  0  0
## G  84 169 309 561 928 241 97 116 280  8  1
```




```
##    18  19  20  21  22  23 24  25  26 27 28
## N   0   0   0   0   0   0  0   0   0  0  0
## A   4  12  11  28  34  10  4   3   2  1  0
## U  11   6  19  35   9   7  4   0   0  1  0
## C   9  14  10  30  29  12  7   3   4  0  0
## G  84 169 309 561 928 241 97 116 280  8  1
```

The R session information (including the OS info, R version and all packages used):

```r
sessionInfo()
```

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Red Hat Enterprise Linux
##
## Matrix products: default
## BLAS/LAPACK: /usr/lib64/R/lib/libRblas.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
## [1] Biostrings_2.48.0   XVector_0.20.0      IRanges_2.14.10
## [4] S4Vectors_0.18.3    BiocGenerics_0.26.0 plyr_1.8.4
## [7] openxlsx_4.1.0      endoSeqR_0.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.18    magrittr_1.5    evaluate_0.11   highr_0.7
##  [5] zip_1.0.0       zlibbioc_1.26.0 stringi_1.2.4   tools_3.5.0
##  [9] stringr_1.3.1   compiler_3.5.0  knitr_1.20
```

```
Sys.time()
```

```
## [1] "2018-10-23 18:01:55 CDT"
```