endoSeqR Practice

Kelly Daescu

January 14, 2018

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 (https://github.com/kjdaescu/endoSeqR). The bowtie builds can be obtained from google drive at the following link: https://drive.google.com/drive/folders/16xoPxbzVu9Zo9_gqF16umHO_swB91kxV?usp=sharing
These bowtie builds are too large to be hosted on github, so they need to be transferred into the “endo_Indexes” sub-directory in the endoSeqR package prior to analysis.

Available species are *C. elegans*, *S. scrofa*, *D. melanogaster*, *D. rerio*, *H. sapiens*, *M. musculus*, *G. gallus*, *A. thaliana*, *X. tropicalis*. To save space, only transferring the species of interest is recommended.

Dependencies:

Linux: samtools, bedtools, bowtie

R: openxlsx, Biostrings, plyr

First time only installation procedure:

1. Download the package from github and place into the desired directory
2. implement the following commands in the parent directory.

```
library(devtools)
install("endoSeqR", lib=".")
```

3. Transfer indexes from Google Drive to endo_Indexes sub-folder after package installation
4. In the endoSeqR package, there are two linux executables - endo_MAP and clean_up. It may be necessary to use command line within the package directory and type in the following:

chmod u+x endo_MAP

chmod u+x clean_up

The library will need to be loaded at the start of every session. To load all the dependencies, use the Package_Load command.

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

To prepare the fasta file, the generate_fasta command requirest two parameters:

1. The file of interest: supported formats (fasta, fa, fasta.gz, fa.gz, fastq, fq, fastq.gz, fq.gz, txt, txt.gz). If a text file is the input file, it must be a list of sequences with no headers for correct results.
2. The desired name of the resulting fasta file

```
generate_fasta("cel.Asikainin_2008.endosiRNA.fasta","Mixed_Stage")
```

Specify the path the path to the endo_Indexes folder within the endoSeqR software, followed by species/species. See the index below for an example of the path.

```
index<-paste0("../endo_Indexes/elegans/elegans")
```

Then, the Map_endosiRNA command will do a three-tiered alignment strategy discussed in the following paper:

Kelly Daescu. endoSeqR: A simple R package to facilitiate the identification of endo-siRNAs from small RNA libraries. Under Review.

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

The following output will be saved in the working directory:

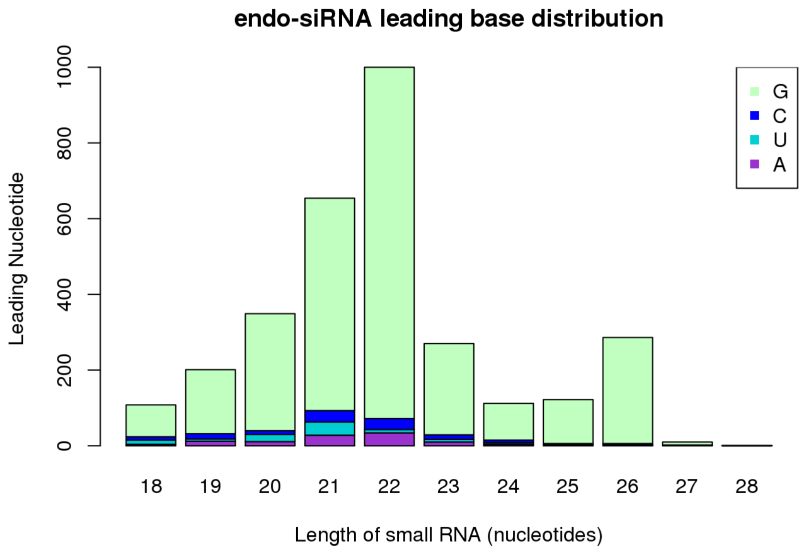
1. Bed file (xlsx format)
2. Detailed bed file that includes sequence information (xlsx format)
3. RData file that can be uploaded into any R session for further analysis

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

The PlotNTLetter output is the following:

1. Data frame with length/leading base information
2. Length/leading base histogram

```
load("Mixed_Stage_endosiRNA.RData")
df2<-PlotNTLetter(df)
```


endo-siRNA leading base distribution
Leading Nucleotide
G
C
U
A
Length of small RNA (nucleotides)