Visualisation and functional analysis of single-nucleotide modifications in mRNAs using the RNAModR package

M. Evers[1]

[1] The John Curtin School of Medical Research,
The Australian National University, Canberra, Australia

April 18, 2016

1 Introduction

Following is a typical workflow for visualising and analysing single-nucleotide modifications:

1. Construction of a custom transcriptome.
2. Mapping of genome alignment-based single-nucleotide modifications to transcript coordinates.
3. Visualisation of basic metrics, e.g. distribution of sites across different transcript sections.
4. Enrichment analyses of single-nucleotide modifications.

The RNAModR library is loaded following

```
library("RNAModR")
```

2 Reference transcriptome and transcript annotation

RNAModR analyses the distribution of RNA modifications across mRNA transcripts and therefore requires suitable mRNA transcriptome and transcript annotation data. When RNAModR is run for the first time, a custom transcriptome is constructed based on UCSC RefSeq gene annotations and associated R data are stored locally in a single binary R file. This process can take up to 10 minutes (depending on the computer hardware), and needs to be run only once as part of a new data analysis. Subsequent re-analyses check for existing suitable transriptome data, and load the required data as part of the different anlysis routines.

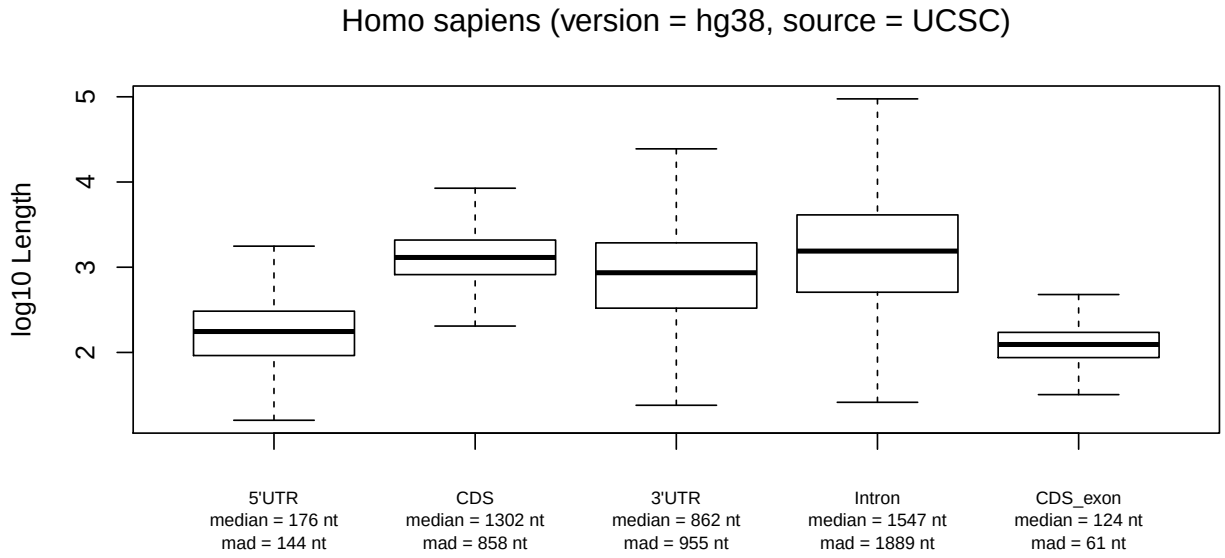
To start our example of an analysis of the m^{6}A sample data [**?**], we construct the necessary transcriptome and annotation data based on the reference genome assembly version GRCh38/hg38 using

```
BuildTx("hg38");

## Found existing transcriptome data.
## To rebuild run with force = TRUE.
```

We can plot the length distribution of the different transcript sections using

```
LoadRefTx("hg38");
PlotTxSecLength(txBySec, filter = c("5'UTR", "CDS", "3'UTR", "CDS_exon", "Intron"));
```



3 Mapping RNA modifications to transcript coordinates

Next we load a BED file containing genome locations of m^6A modifications from [**?**]:

```
bedFile <- system.file("extdata",
                       "miCLIP_m6A_Linder2015_hg38.bed",
                       package = "RNAModR");
sites <- ReadBED(bedFile);
class(sites);

## [1] "GRanges"
## attr(,"package")
## [1] "GenomicRanges"

length(sites);

## [1] 15167
```

We map genome coordinates of the 15167 sites to the transcriptome:

```
posSites <- SmartMap(sites, id = "m6A", refGenome = "hg38");
```

The return object is an RNAModR-specific object of type txLoc, that contains a list of sites with positions within any of the transcript sections (promoter, 5'UTR, CDS, 3'UTR, intron).

We can access a txLoc to get a brief summary of the mapped sites:

```
info(posSites);

## Object of class "txLoc".
##
## ID               = m6A
## Reference genome = hg38
## Version          = 2016-04-18
## Total # of sites = 14845
```

```
## Package           = RNAModR
## 5 transcript sections: Promoter, 5'UTR, CDS, 3'UTR, Intron
##    Promoter: Number of loci = 198
##       5'UTR: Number of loci = 946
##         CDS: Number of loci = 7449
##       3'UTR: Number of loci = 5784
##      Intron: Number of loci = 468
```

We can also show the distribution of sites across different transcript sections in a piechart:

```
PlotSectionDistribution(posSites, filter = c("5'UTR", "CDS", "3'UTR"));
```