Read external data into movAPA

Xiaohui Wu, Wenbin Ye, Tao Liu, Hongjuan Fu

Last modified 2020-12-25

Contents

1 Overview

This documentation describes how to read an external file of poly(A) sites and analyze it with movAPA. We used the model species – Arabidopsis for demonstration. First we can download a poly(A) site list from PlantAPAdb. Here we just downloaded poly(A) site clusters (PACs) for demostration. A PAC is already the group of nearby cleavage sites.

Demo file 1: PACs with genome annotation (3 replicates). Download the data (arabidopsis_thaliana.SRP093950_amp.high_c
here.

Demo file 2: PACs in bed format with only coordinates. Download the data here.

These data files and the Arabidopsis TAIR10 gff3 file can also be downloaded here.

2 Read the file of PACs with genome annotation

movAPA implemented the *PACdataset* object for storing the expression levels and annotation of PACs from various conditions/samples. Almost all analyses of poly(A) site data in movAPA are based on the *PACdataset*. The "counts" matrix is the first element in the array list of *PACdataset*, which stores non-negative values representing expression levels of PACs. The "colData" matrix records the sample information and the "anno" matrix stores the genome annotation or additional information of the poly(A) site data.

2.1 Data read

```
library(movAPA)
pac=read.csv('arabidopsis_thaliana.SRP093950_amp.high_confidence.PAC.annotation.tpm.csv', stringsAsFact

## Rename annotation columns.
## In a PACdataset, the annotation column names must be named as (gene/gene_type/ftr/ftr_start/ftr_end/
## Other non-sample columns will be also retained in the @anno slot of the PACdataset.
pac=dplyr::rename(pac, UPA_start = 'start', UPA_end='end', gene_type='biotype')
colnames(pac)

## Describe the sample columns and corresponding group(s) in a data.frame
colData=as.data.frame(matrix(c('Amp','Amp','Amp'), ncol=1, dimnames=list(paste0('Amp311_R',1:3), 'group

## Read the PAC file into a PACdataset
PACds=readPACds(pacFile=pac, colDataFile=colData, noIntergenic=FALSE, PAname='PA')

PACds
```

2.2 Statistics

After read the data into a PACdataset, users can use many functions in movAPA for removing internal priming artifacts, polyA signal analysis, etc. Please follow the vignette of “movAPA_on_rice_tissues” for more details.

```
# For example, users can remove internal priming artifacts
library("BSgenome.Athaliana.TAIR.TAIR9")
bsgenome <- BSgenome.Athaliana.TAIR.TAIR9

# Please make sure the chr name of your PAC data is the same as the BSgenome.
seqnames(bsgenome)

PACdsIP=removePACdsIP(PACds, bsgenome, returnBoth=TRUE,
                      up=-10, dn=10, conA=6, sepA=7)
length(PACdsIP$real)
length(PACdsIP$ip)

# Base compostions and k-grams
faFiles=faFromPACds(PACds, bsgenome, what='updn', fapre='updn',
                    up=-300, dn=100, byGrp='ftr')
```

```
faFiles=c("updn.3UTR.fa", "updn.CDS.fa", "updn.intergenic.fa", "updn.intron.fa")
## Plot single nucleotide profiles using the extracted sequences and merge all plots into one.
plotATCGforFAfile (faFiles, ofreq=FALSE, opdf=FALSE,
                   refPos=301, mergePlots = TRUE)
```

3 Read the file of PACs with only coordinates

In this section, we show how to read a list of polyA sites with only coordinates. Here we use the file in bed format for demonstration.

3.1 Data read

```
## Read a BED file
pac=read.table('arabidopsis_thaliana.SRP093950_amp.high_confidence.PAC.bed',
               header=F, stringsAsFactors =F)
head(pac)

# We only keep the chr/strand/coord, here we used the start position as the coord.
colnames(pac)=c('chr','coord','x','dot','strand')
pac=pac[,c('chr','strand','coord')]

# We don't have any expression level of the sample,
# so we only read the PAC list and set the expression as 1.
## Read the PAC file into a PACdataset
PACds=readPACds(pacFile=pac, colDataFile=NULL, noIntergenic=FALSE, PAname='PA')
PACds
```

3.2 Annotation

After read the data into a PACdataset, users can use movAPA for annotation first.

```
# Please download the genome annotation file of Arabidopsis TAIR 10
# in gff3 format from the tair website.
athGFF="Arabidopsis_thaliana.TAIR10.42.gff3"

# First we parse the gff3 file.
gff=parseGff(athGFF)

# Please make sure the chromosome name of your PAC data
# is the same as the gff file (and the BSgenome)
head(gff$anno.need)

# You can also save the parsed gff file as an rda object for further use.
# save(gff, file='TAIR10.gff.rda')
# Annotate the PAC data
PACds=annotatePAC(PACds, gff)
PACds
```

3.3 Statistics

After read the data into a PACdataset, users can use many functions in movAPA for removing internal priming artifacts, polyA signal analysis, etc. Please follow the vignette of “movAPA_on_rice_tissues” or the above example for more details.