Analyses of APA dynamics in mouse sperm cells with the movAPA package

Xiaohui Wu, Wenbin Ye, Tao Liu, Hongjuan Fu

Last modified 2023-10-10

Contents

1 Overview

Here we investigated the application of movAPA on poly(A) sites (or called poly(A) site clusters, PACs) from mouse sperm cells. Poly(A) sites from three stages of differentiation process were obtained from the previous study (Shulman and Elkon, 2019), including early stage (spermatocytes, SC), intermediate stage (round spermatids, RS), and late stage (elongating spermatids, ES). We used a small dataset containing 3'UTR poly(A) sites from the chromosome 12 for demonstration.

2 Preparations

2.1 PAC data of mouse sperm cells

movAPA is highly scalable and flexible in that the dataset of APA sites in single cells can be readily represented by the generic object of *PACdataset* where cells of the same cell type are regarded as replicates of a biological sample.

The moveAPA package includes an example single cell PAC dataset stored as a *PACdataset* object, containing 771 PACs from 396 genes located in chromosome 12. There are total 2042 cells from three cell types. This dataset contains the gene *Psen1* (ENSMUSG00000019969) presented in Shulman et al, 2019.

```
library(movAPA, warn.conflicts = FALSE, quietly=TRUE)
data(scPACds)
summary(scPACds)
#> PAC# 771
#> sample# 2042
#> summary of expression level of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>       1      50     228    2141    1151  126436
#> summary of expressed sample# of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     1.0    37.0   156.0   408.0   580.5  2041.0
#> gene# 396
#>            nPAC
#> 3UTR        644
#> CDS           3
#> intergenic   94
#> intron       30
head(scPACds@counts[1:2,1:5])
#> 2 x 5 sparse Matrix of class "dgCMatrix"
#>        AAACCTGAGAGGGCTT AAACCTGAGCTTATCG AAACCTGCATACGCCG AAACCTGGTTGAGTTC
```

```
#> PA3443                .                .                .                .
#> PA3446                .                .                .                .
#>        AAACCTGTCAACGAAA
#> PA3443                .
#> PA3446                .
head(scPACds@anno, n=2)
#>          chr strand     coord   peakID  ftr gene_type ftr_start   ftr_end
#> PA3443 chr12      - 100125475 peak3443 3UTR      <NA> 100125452 100125605
#> PA3446 chr12      - 100549890 peak3446 3UTR      <NA> 100549778 100551443
#>                       gene gene_start  gene_end gene_stop_codon upstream_id
#> PA3443 ENSMUSG00000021179  100125452 100159653       100125606        <NA>
#> PA3446 ENSMUSG00000021180  100549778 100725028       100551444        <NA>
#>        upstream_start upstream_end downstream_id downstream_start
#> PA3443             NA           NA          <NA>               NA
#> PA3446             NA           NA          <NA>               NA
#>        downstream_end three_UTR_length three_extend
#> PA3443             NA              131           NA
#> PA3446             NA             1554           NA
head(scPACds@colData, n=2)
#>                              group celltype        tsn1        tsn2
#> AAACCTGAGAGGGCTT AAACCTGAGAGGGCTT       SC 22.54797966 4.077467845
#> AAACCTGAGCTTATCG AAACCTGAGCTTATCG       RS 1.138437608 -32.9317999
levels(scPACds@colData$celltype)
#> [1] "ES" "RS" "SC"
```

2.2 Reference genome

The reference genome is not necessary for this case study, while it is required for removing internal priming or poly(A) signal analyses. movAPA uses reference genome sequences that are represented as a *BSgenome* object or stored in a fasta file. To use *BSgenome* object, please refer to the *BSgenome* package for obtaining a *BSgenome* object for your species.

2.3 Genome annotation

Genome annotation stored in a GFF/GTF file or a TXDB R object can be used for annotating PACs. The function *parseGenomeAnnotation* is used to parse the given annotation and the processed annotation can be saved into an rdata object for further use. The genome annotation file is not necessary for this case study as the information has been stored in scPACds.

Process the genome annotation of mm10 represented as TxDb object.

```
library(TxDb.Mmusculus.UCSC.mm10.ensGene)
txdbmm10 <- TxDb.Mmusculus.UCSC.mm10.ensGene

scPACds=createPACdataset(scPACds@counts, scPACds@anno, scPACds@colData, forceSparse = TRUE)
scPACds=annotatePAC(scPACds, txdbmm10)
```

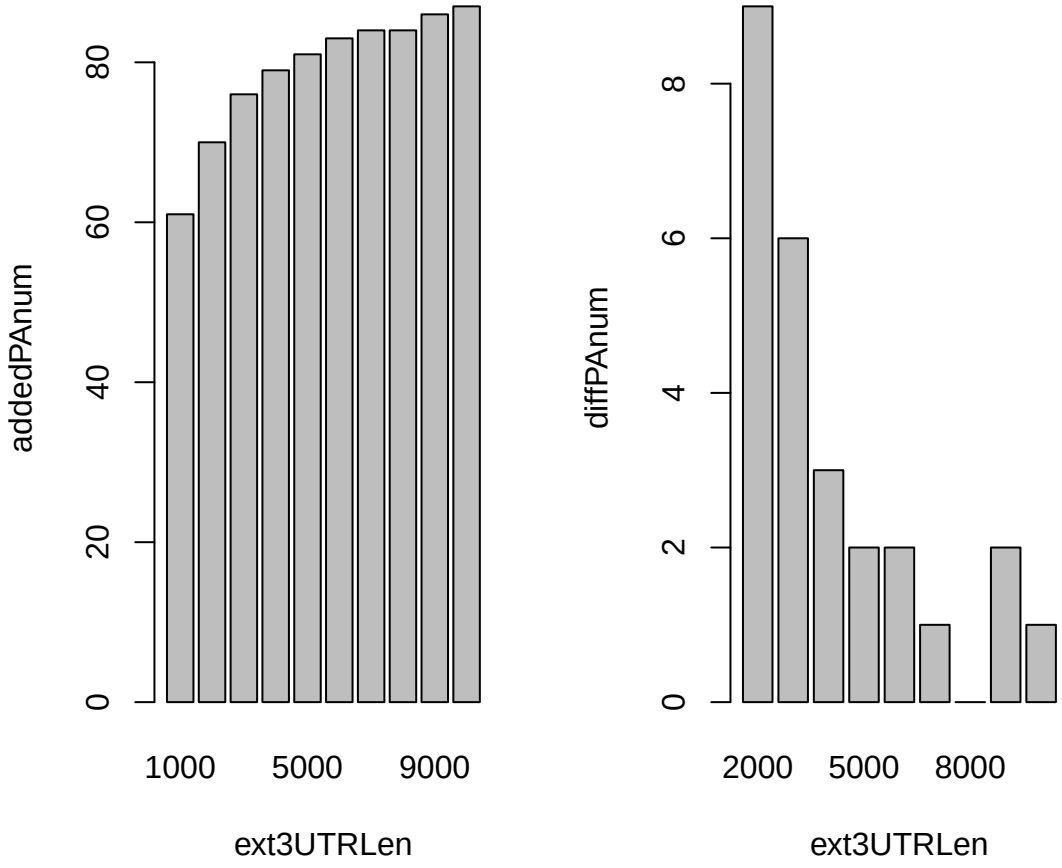
3 Preprocessing of PAC data

3.1 Extending annotated 3'UTRs

Genes with or without annotated 3'UTR could be assigned an extended 3'UTR of a given length using the function `ext3UTRPACds`, which can improve the "recovery" of poly(A) sites falling within authentic 3'UTRs.

Before extending, we can calculate the number of PACs falling into extended 3'UTRs of different lengths.

```
testExt3UTR(scPACds, seq(1000, 10000, by=1000))
```




```
#>     ext3UTRLen addedPAnum
#> 1         1000         61
#> 2         2000         70
#> 3         3000         76
```

```
#> 4         4000          79
#> 5         5000          81
#> 6         6000          83
#> 7         7000          84
#> 8         8000          84
#> 9         9000          86
#> 10       10000          87
```

Here we extended 3'UTR length for 2000 bp. After extension, 70 PACs in intergenic region are now in extended 3'UTRs.

```
table(scPACds@anno$ftr)
#>
#>       3UTR        CDS intergenic     intron
#>        644          3         94         30
scPACds=ext3UTRPACds(scPACds, ext3UTRlen=2000)
#> 70 PACs in extended 3UTR (ftr=intergenic >> ftr=3UTR)
#> Get 3UTR length (anno@toStop) for 3UTR/extended 3UTR PACs
table(scPACds@anno$ftr)
#>
#>       3UTR        CDS intergenic     intron
#>        714          3         24         30
```

3.2 Normalization

The function *normalizePACds* can be called for normalization, which implements three strategies including TPM (Tags Per Million), the normalization method of DESeq (Anders and Huber, 2010), and the TMM method used in EdgeR (Robinson, et al., 2010).

Note: normalization should be performed in caution, because different methods would have significant and different impact on the data and downstream analysis!

Here is an example to normalize the data using the TPM method.

```
scPACdsNorm2=normalizePACds(scPACds, method='TPM')
head(Matrix::colSums(scPACdsNorm2@counts))
#> AAACCTGAGAGGGCTT AAACCTGAGCTTATCG AAACCTGCATACGCCG AAACCTGGTTGAGTTC
#>              985              591             2507              486
#> AAACCTGTCAACGAAA AAACCTGTCGCGGATC
#>              592               81
```

3.3 Filter PACs or cells

We can use `subsetPACds` to filter PACs by different options. Here we filter PACs with total counts>=20 and remove intergenic PACs.

```
scPACdsFlt=subsetPACds(scPACds, totPACtag=20, choosePA=NULL,
                       noIntergenic=TRUE, verbose=TRUE)
#>                     count
#> before subsetPACds   771
#> noItg                747
#> totPACtag>=20        682
#> minExprConds>=1      682
summary(scPACdsFlt)
#> PAC# 682
#> sample# 2042
#> summary of expression level of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>      20      76     337    2411    1361  126436
#> summary of expressed sample# of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    14.0    58.0   224.5   456.1   653.0  2041.0
#> gene# 411
#>        nPAC
#> 3UTR    655
#> CDS       2
#> intron   25
```

Filter only PACs in 3'UTR and obtain PACs in 3'UTRs with >=2 PACs.

```
scPACdsFlt=get3UTRAPAds(scPACdsFlt, sortPA=TRUE, choose2PA=NULL)
summary(scPACdsFlt)
#> PAC# 411
#> sample# 2042
#> summary of expression level of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    20.0    74.5   239.0  1174.0   878.0 42558.0
#> summary of expressed sample# of each PA
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>      14      55     166     357     481    2020
#> gene# 165
#>      nPAC
#> 3UTR  411
```

4 Statistics of PACds

4.1 PAC distributions among cell types

To make statistics of PACs among cell types, first we pool cells of the same cell type.

```
scPACdsCt=subsetPACds(scPACds, group='celltype', pool=TRUE)
```

Make statistics of PAC distributions in each cell type.

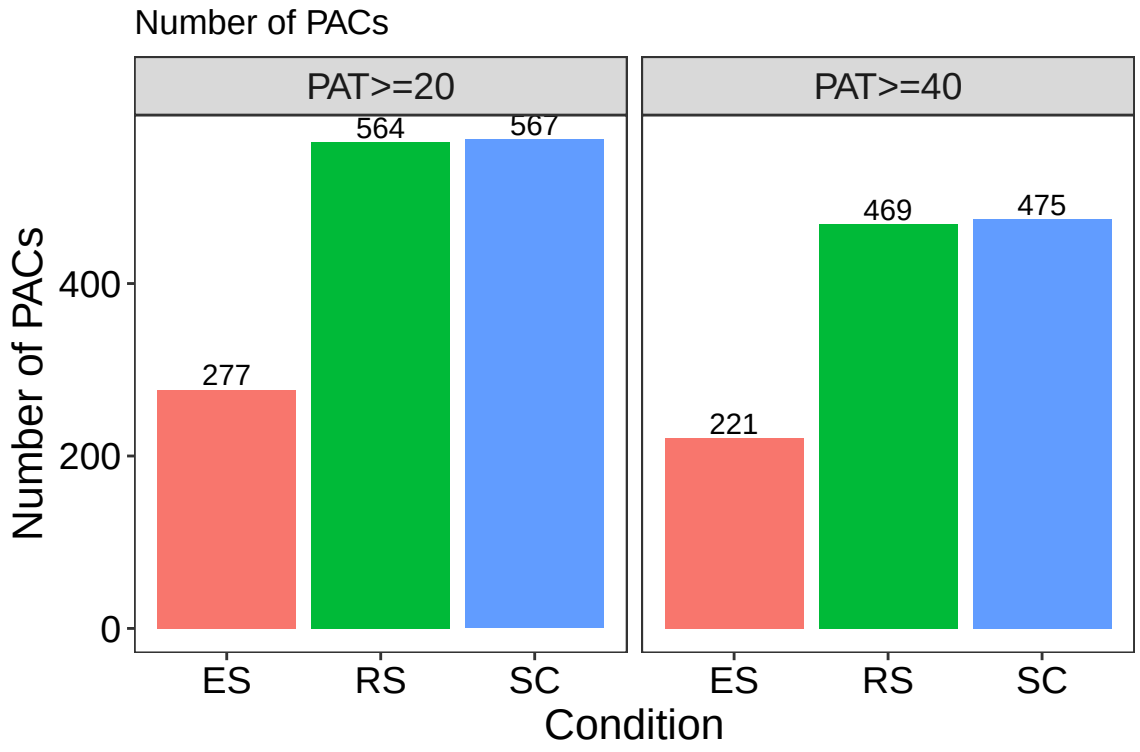
```
scPACdsCtStat=movStat(scPACdsCt, minPAT=c(20, 40), ofilePrefix=NULL)
```

Statistical results of PACs with total read counts >=20.

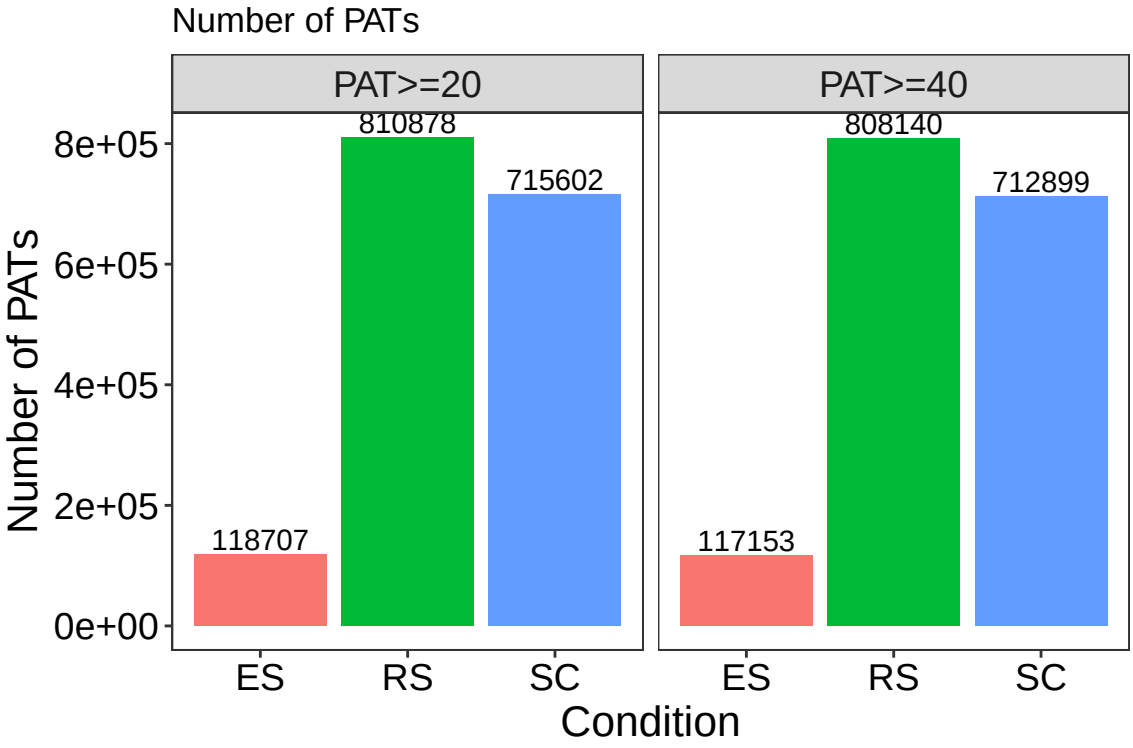
```
scPACdsCtStat$pat20
#>       nPAC    nPAT nGene nAPAgene APAextent 3UTR_nPAT CDS_nPAT intergenic_nPAT
#> SC     567  715602   354      142 0.4011299    706741      989             1816
#> RS     564  810878   366      134 0.3661202    795569      586             2513
#> ES     277  118707   225       45 0.2000000    116090        0              835
#> total  669 1645187   403      167 0.4143921   1618400     1575             5164
#>       intron_nPAT 3UTR_nPAC CDS_nPAC intergenic_nPAC intron_nPAC
#> SC           6056       534        2              13          18
#> RS          12210       534        1              11          18
#> ES           1782       266        0               3           8
#> total       20048       626        2              19          22
```

Plot statistical results by various barplots. Results showed that there are more PACs expressed in RS and SC than in ES.

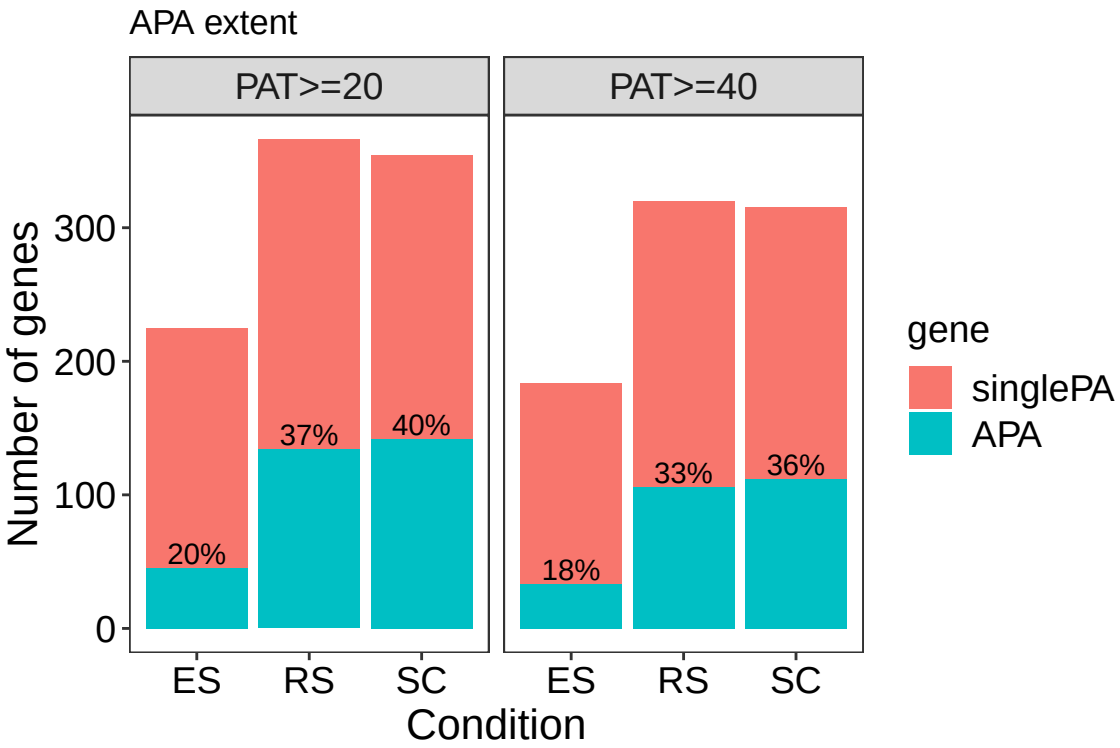
```
plotPACdsStat(scPACdsCtStat, pdfFile=NULL)
```



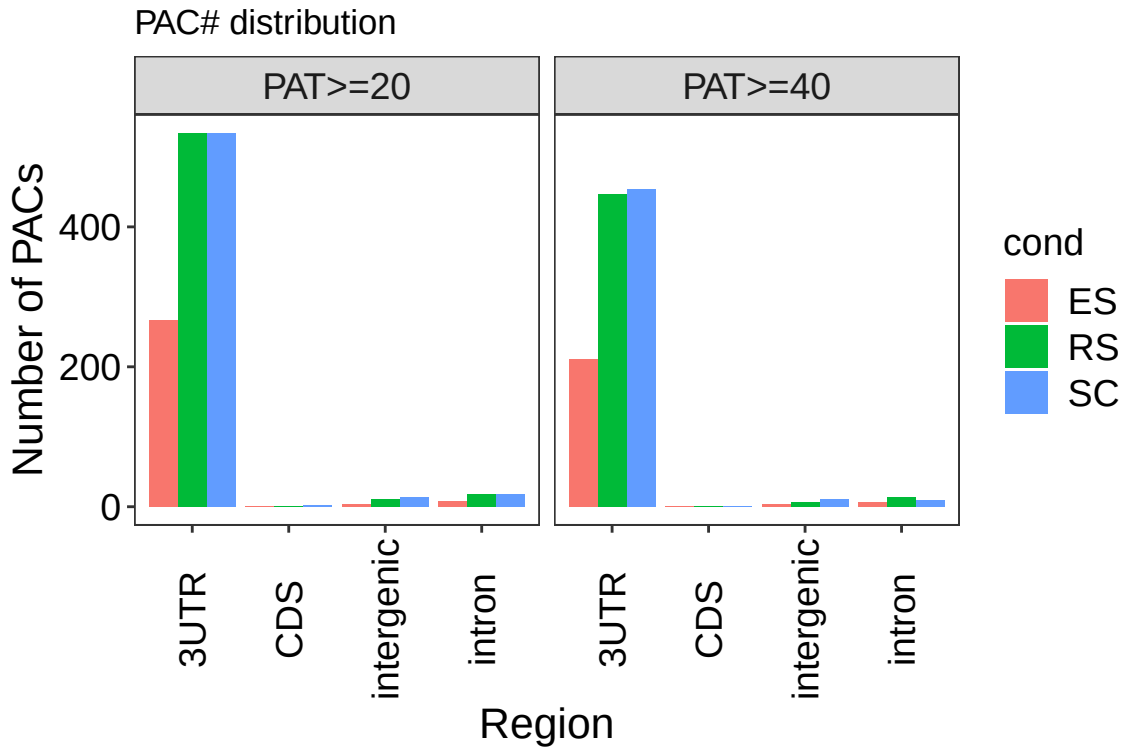
```
#> Plot Number of PACs
```



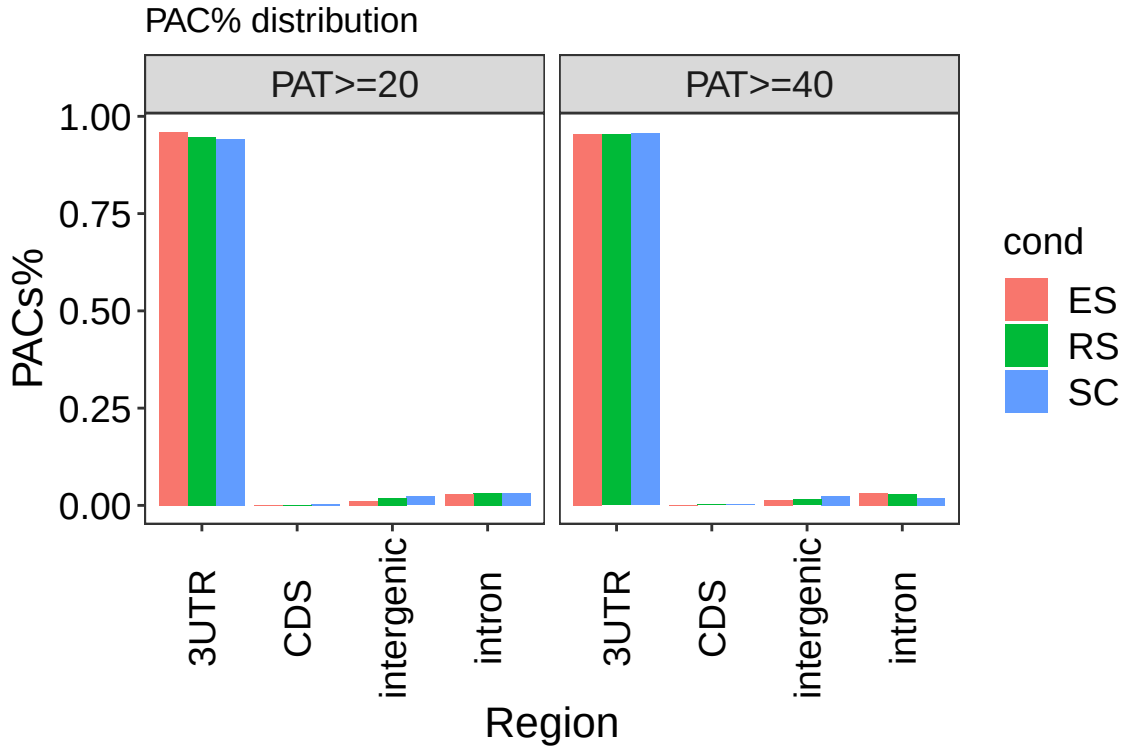

```
#> Plot Number of PATs
```



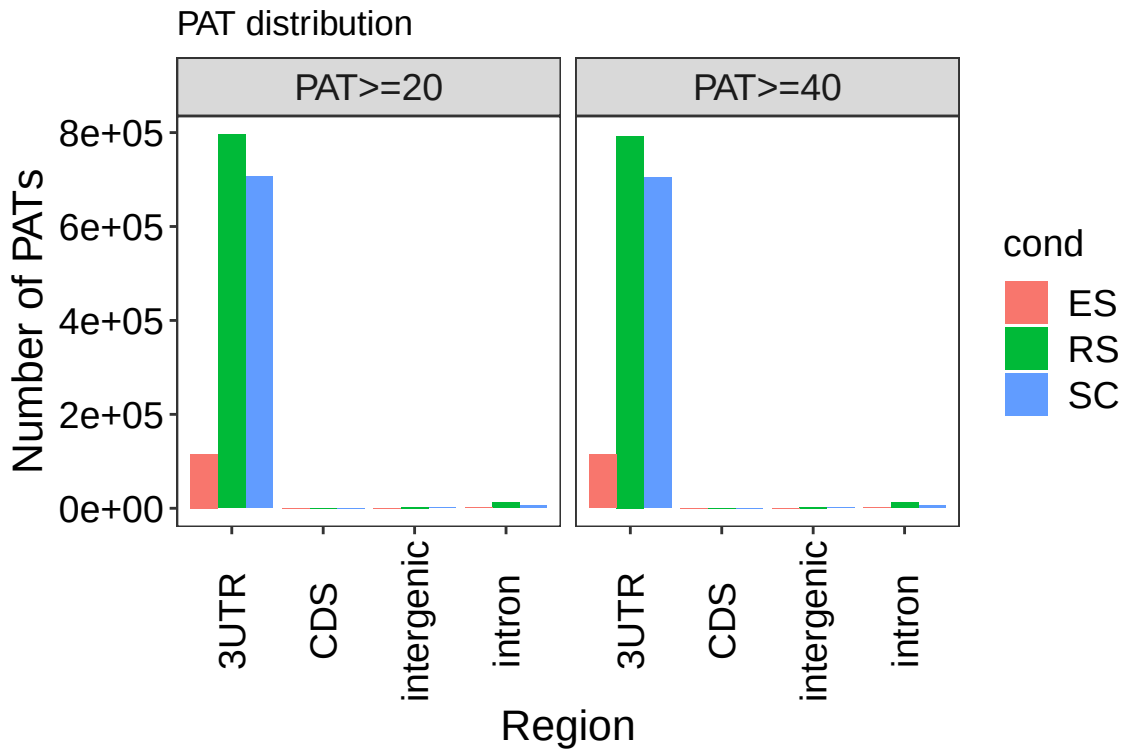
```
#> Plot APA extent
```



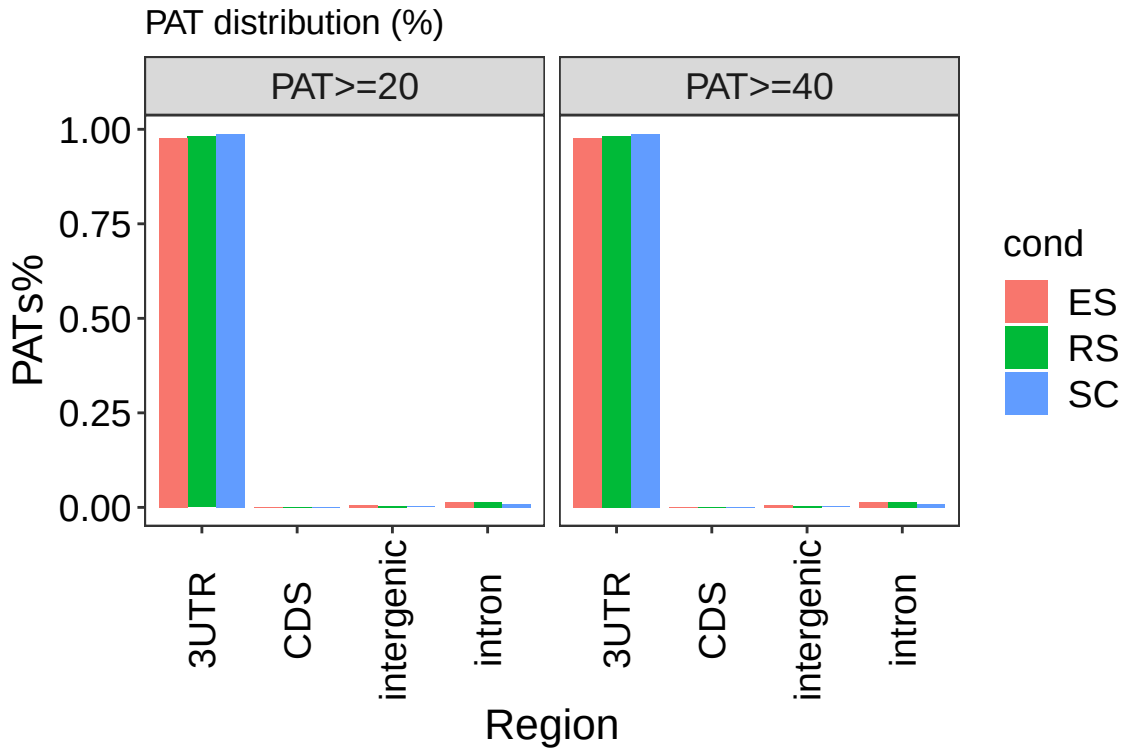

```
#> Plot PAC# distribution
```



```
#> Plot PAC% distribution
```




```
#> Plot PAT# distribution
```



```
#> Plot PAT% distribution
```

4.2 PAC distributions in single cells

Make statistics for PAT and PAC distributions in each cell, using PAT cutoffs 1 and 5.

```
scPACdsStat=movStat(scPACds, minPAT=c(1,5), ofilePrefix='scPACds.stat')
#> >>> scPACds.stat.pat1.stat
#> >>> scPACds.stat.pat5.stat
```

Statistics of pooled data

```
scPACdsStat$pat1['total',]
#>       nPAC    nPAT nGene nAPAgene APAextent 3UTR_nPAT CDS_nPAT intergenic_nPAT
#> total  771 1650337   436      197 0.4518349   1622965     1618            5442
#>       intron_nPAT 3UTR_nPAC CDS_nPAC intergenic_nPAC intron_nPAC
#> total       20312       714        3              24          30
```

Summary of PAC# in each cell, ranging from 56 PACs per cell to 354 PACs per cell.

```
summary(scPACdsStat$pat1$nPAC[1:(nrow(scPACdsStat$pat1)-1)])
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>      56     128     147     154     169     354
```

Summary of PAT# (read count) in each cell, ranging from 154 PACs per cell to 5712 PACs per cell.

```
summary(scPACdsStat$pat1$nPAT[1:(nrow(scPACdsStat$pat1)-1)])
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   154.0   503.2   644.5   808.2   847.8  5712.0
```

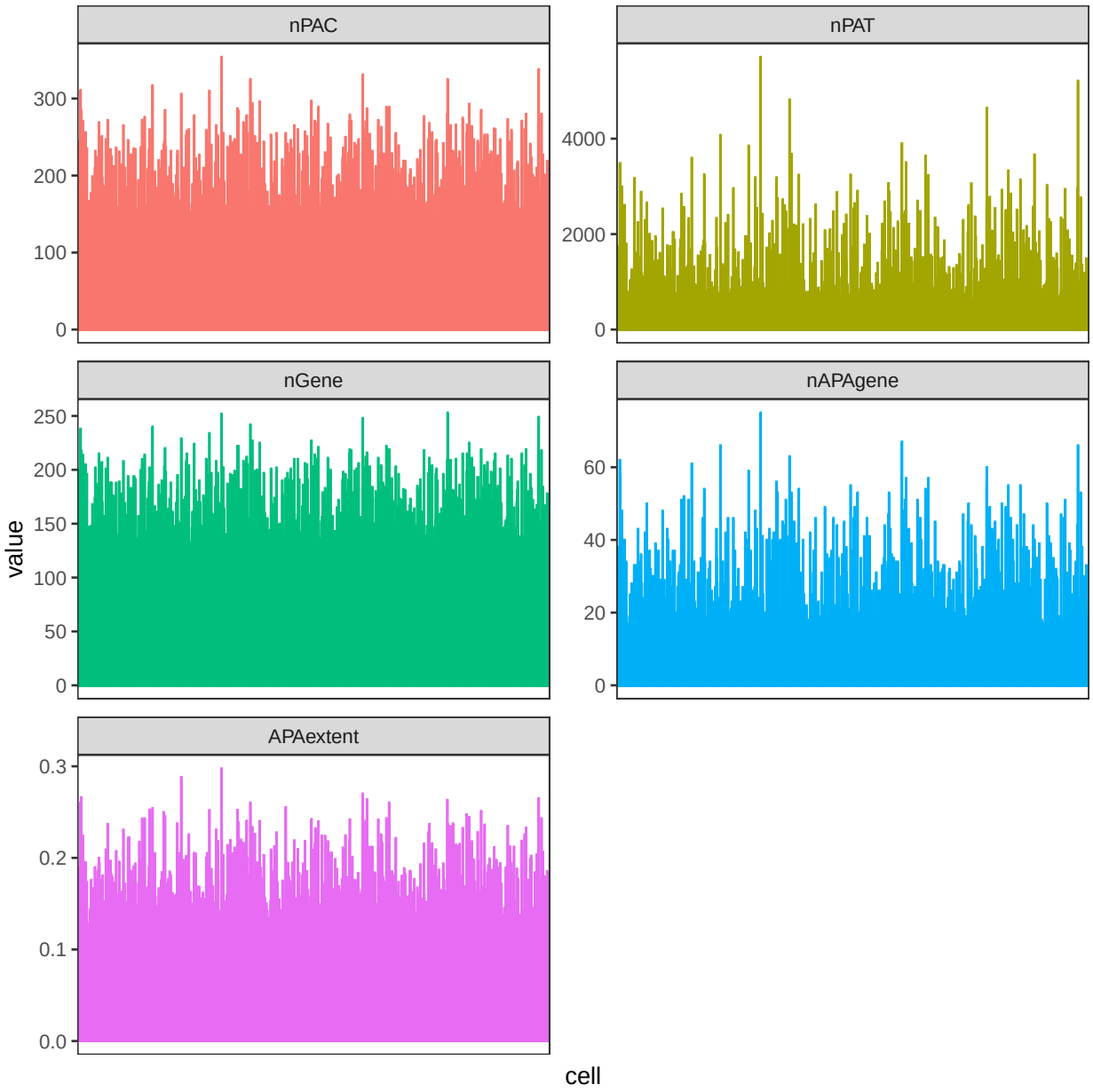
Here we plot barplots showing distributions of PACs and PATs among cells. First we create the data for plot using all PACs (PAT cutoff=1), and remove the 'total' line.

```
d=scPACdsStat$pat1[, c('nPAC','nPAT','nGene','nAPAgene','APAextent')]
d$cell=rownames(d)
d=d[1:(nrow(d)-1), ]
d=reshape2::melt(d)
```

Plot barplots to Show the distribution of PAT#, PAT#, gene#, APA gene#, and APA gene%.

```
library(ggplot2, quietly = TRUE)
sp <- ggplot(data=d, aes(x=cell, y=value, color=variable)) +
  geom_bar(stat='identity', width=1.1)
sp = sp + theme_bw() +  guides(color=FALSE) +
```

```
  theme(axis.ticks.x = element_blank(), axis.text.x = element_blank(),
        panel.grid.minor=element_blank(),
        panel.grid.major=element_blank())
sp + facet_wrap( ~ variable, ncol=2, scales="free")
```



5 Analyses of APA dynamics

5.1 Detecting DE PACs

movAPA provides the function *movDEPAC* to identify DE PACs between samples. Three strategies were utilized: (i) using DESeq2 with replicates; (ii) using DEXseq with replicates; (iii) using chi-

squared test without replicates ("chisq"). The strategy of chi-squared test was used in the study on single cell APA for detecting differential usage of PACs among cells (Shulman and Elkon, 2019). For single cell data, we highly recommand the chisq method because it is much faster than the other two methods.

Note: DE detection should be performed in caution, because different methods would have significant and different impact on the DE results!

Detecting DE PACs using chisq method for genes with total counts>=50.

```
DEqPAC=movDEPAC(scPACdsCt, method='chisq', group='celltype',
                minSumPAT=50, chisqPadjust=TRUE)
```

Make statistics of the DE PAC result from the chisq method. Here the value column of DEqPAC is 1-pvalue_of_the_gene. So using padjThd=0.05 and valueThd=0.95 means filtering DE PACs with adjusted pvalue of PAC <0.05 and adjusted pvalue of gene <0.05.

```
stat=movStat(object=DEqPAC, padjThd=0.05, valueThd=0.95)
#> All cond pairs in heat@colData, get de01 and deNum
stat$nsig
#>       sig.num
#> SC.RS     167
#> SC.ES     140
#> RS.ES      85
head(stat$ovp)
#>          pair n1.sig.num n2.sig.num novp.sig.num
#> 1 SC.RS-SC.ES        167        140          115
#> 2 SC.RS-RS.ES        167         85           65
#> 3 SC.ES-RS.ES        140         85           78
head(stat$siglist[[1]])
#> [1] "PA11112" "PA11113" "PA11288" "PA11291" "PA11375" "PA11467"
```

Output full list of DE PACs.

```
sel=movSelect(aMovRes=DEqPAC, condpair='SC.RS',
              padjThd=0.05, valueThd=0.95,
              out='full', PACds=scPACdsCt)
head(sel, n=2)
#>        PA   chr strand    coord    peakID  ftr gene_type ftr_start  ftr_end
#> 1 PA11112 chr12      + 86963789 peak11112 3UTR      <NA>  86962668 86965362
#> 2 PA11113 chr12      + 86965333 peak11113 3UTR      <NA>  86962668 86965362
#>                 gene gene_start gene_end gene_stop_codon upstream_id
#> 1 ENSMUSG00000034157   86947343 86965362        86962667        <NA>
#> 2 ENSMUSG00000034157   86947343 86965362        86962667        <NA>
#>   upstream_start upstream_end downstream_id downstream_start downstream_end
#> 1             NA           NA          <NA>              NA             NA
#> 2             NA           NA          <NA>              NA             NA
#>   three_UTR_length three_extend toStop  SC  RS ES         padj value
```

```
#> 1             1122           NA   1122 507 573 14 6.218416e-08     1
#> 2             2666           NA   2666 586 269  7 3.042999e-10     1
```

For all DE PACs in all condtions pairs, examine the DE status of each PAC.

```
head(stat$tf01)
#>          SC.RS SC.ES RS.ES
#> PA11105     0     0     1
#> PA11112     1     0     0
#> PA11113     1     0     0
#> PA11167     0     0     1
#> PA11169     0     1     1
#> PA11288     1     1     1
## Output stat results into files: "DEqPAC.plots.pdf" and 'DEqPAC.stat'.
## outputHeatStat(heatStats=stat, ostatfile='DEqPAC.stat', plotPre='DEqPAC')
```

Visualize a DE PAC in gene ENSMUSG00000019969 by *movViz*.

First, we examine all PACs in this gene. There are two 3'UTR PACs (PA2503 and PA2504).

```
gene='ENSMUSG00000019969'
gp=scPACds[scPACds@anno$gene==gene, ]
cbind(gp@anno$ftr, Matrix::rowSums(gp@counts))
#>        [,1]   [,2]
#> PA2503 "3UTR" "2510"
#> PA2504 "3UTR" "3773"
```

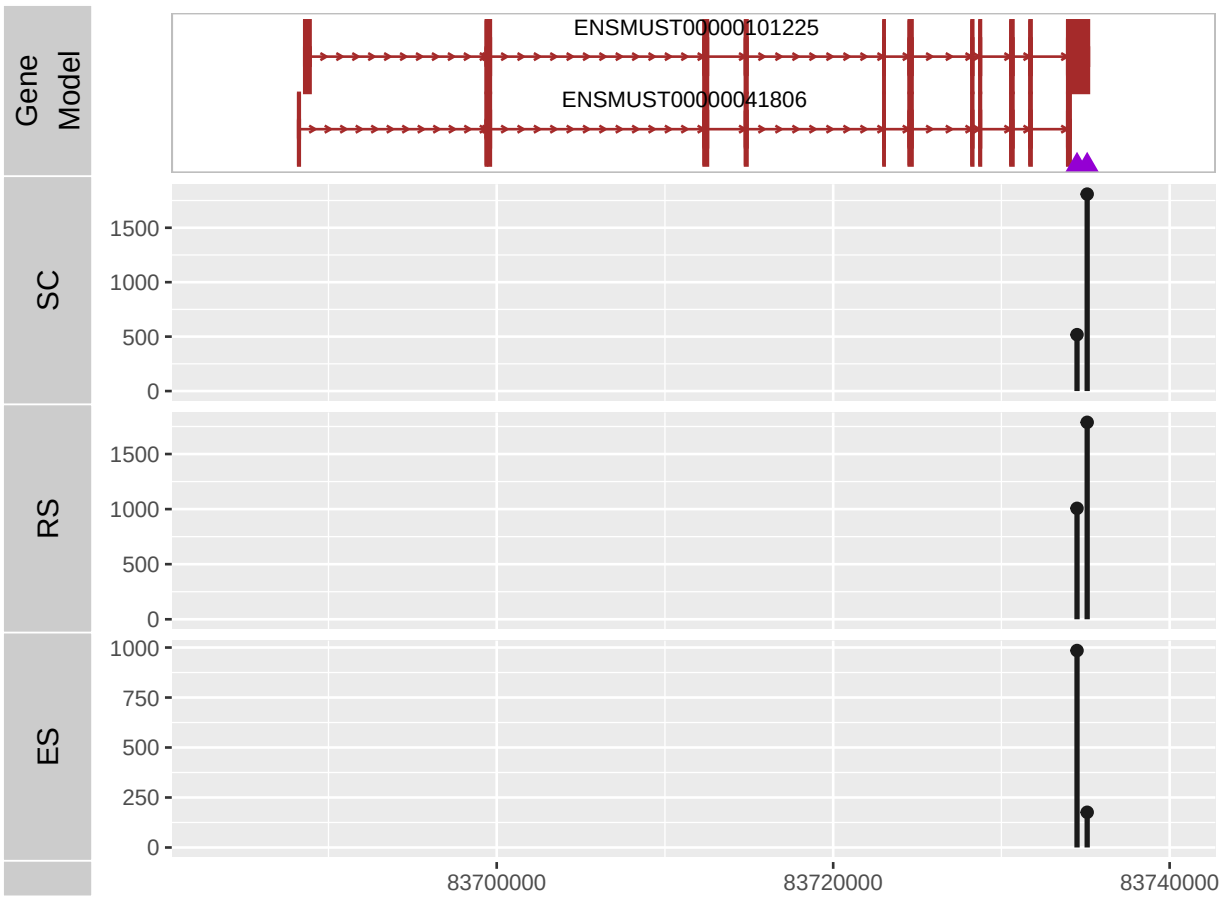
5.2 Plot DE PACs

Plot all PACs in this gene. Here we used scPACdsCt instead of scPACds to plot the total expression levels of PACs in a cell type. But, first we need to prepare the genome annotation.

```
library(TxDb.Mmusculus.UCSC.mm10.ensGene)
#> Loading required package: GenomicFeatures
#> Loading required package: BiocGenerics
#>
#> Attaching package: 'BiocGenerics'
#> The following object is masked from 'package:movAPA':
#>
#>     rbind
#> The following objects are masked from 'package:stats':
#>
#>     IQR, mad, sd, var, xtabs
#> The following objects are masked from 'package:base':
#>
#>     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
```

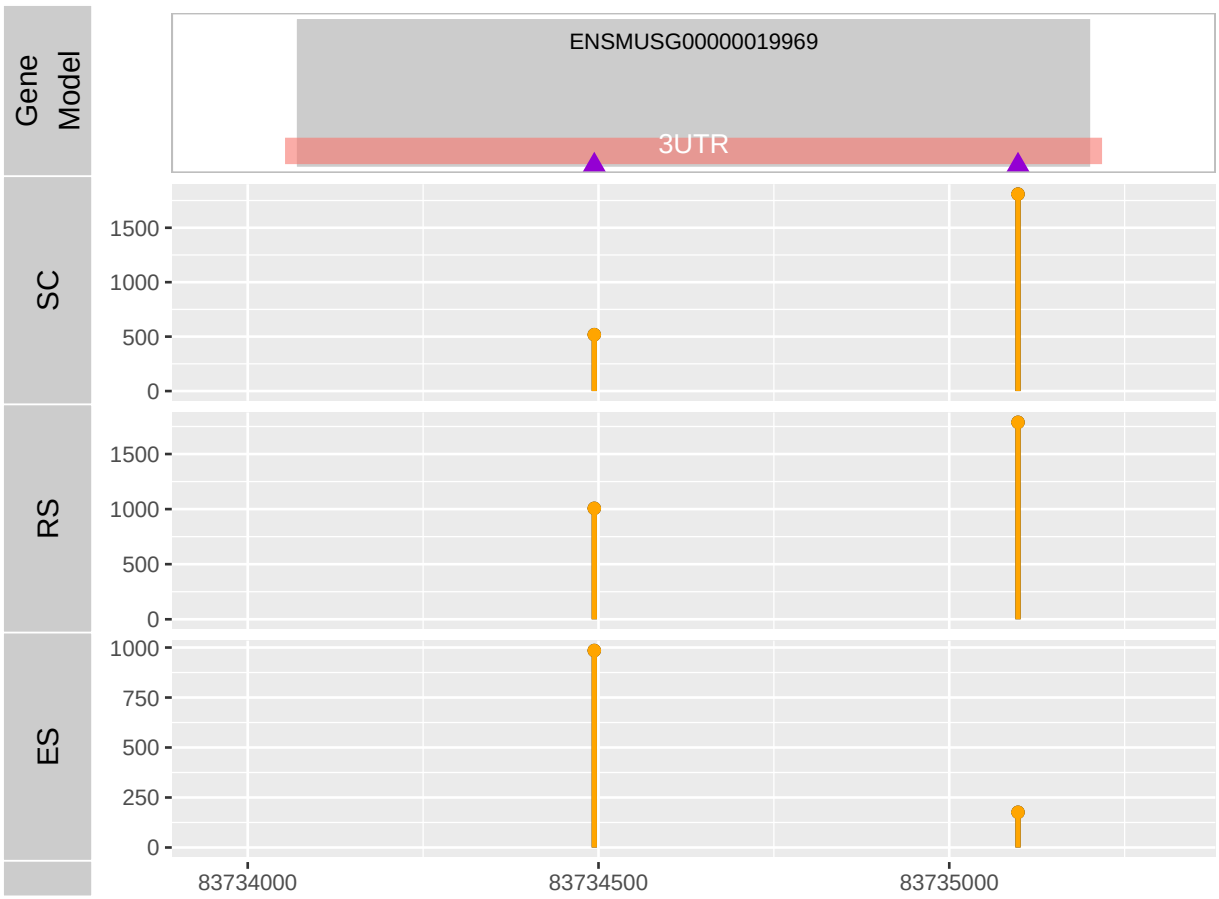
```
#>     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
#>     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
#>     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
#>     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
#>     table, tapply, union, unique, unsplit, which.max, which.min
#> Loading required package: S4Vectors
#> Loading required package: stats4
#>
#> Attaching package: 'S4Vectors'
#> The following objects are masked from 'package:base':
#>
#>     expand.grid, I, unname
#> Loading required package: IRanges
#>
#> Attaching package: 'IRanges'
#> The following object is masked from 'package:grDevices':
#>
#>     windows
#> Loading required package: GenomeInfoDb
#> Loading required package: GenomicRanges
#> Loading required package: AnnotationDbi
#> Loading required package: Biobase
#> Welcome to Bioconductor
#>
#>     Vignettes contain introductory material; view with
#>     'browseVignettes()'. To cite Bioconductor, see
#>     'citation("Biobase")', and for packages 'citation("pkgname")'.
txdbmm10 <- TxDb.Mmusculus.UCSC.mm10.ensGene
gff=parseGenomeAnnotation(txdbmm10)
```

```
movViz(object=scPACdsCt, gene=gene, txdb=gff, group='celltype')
```



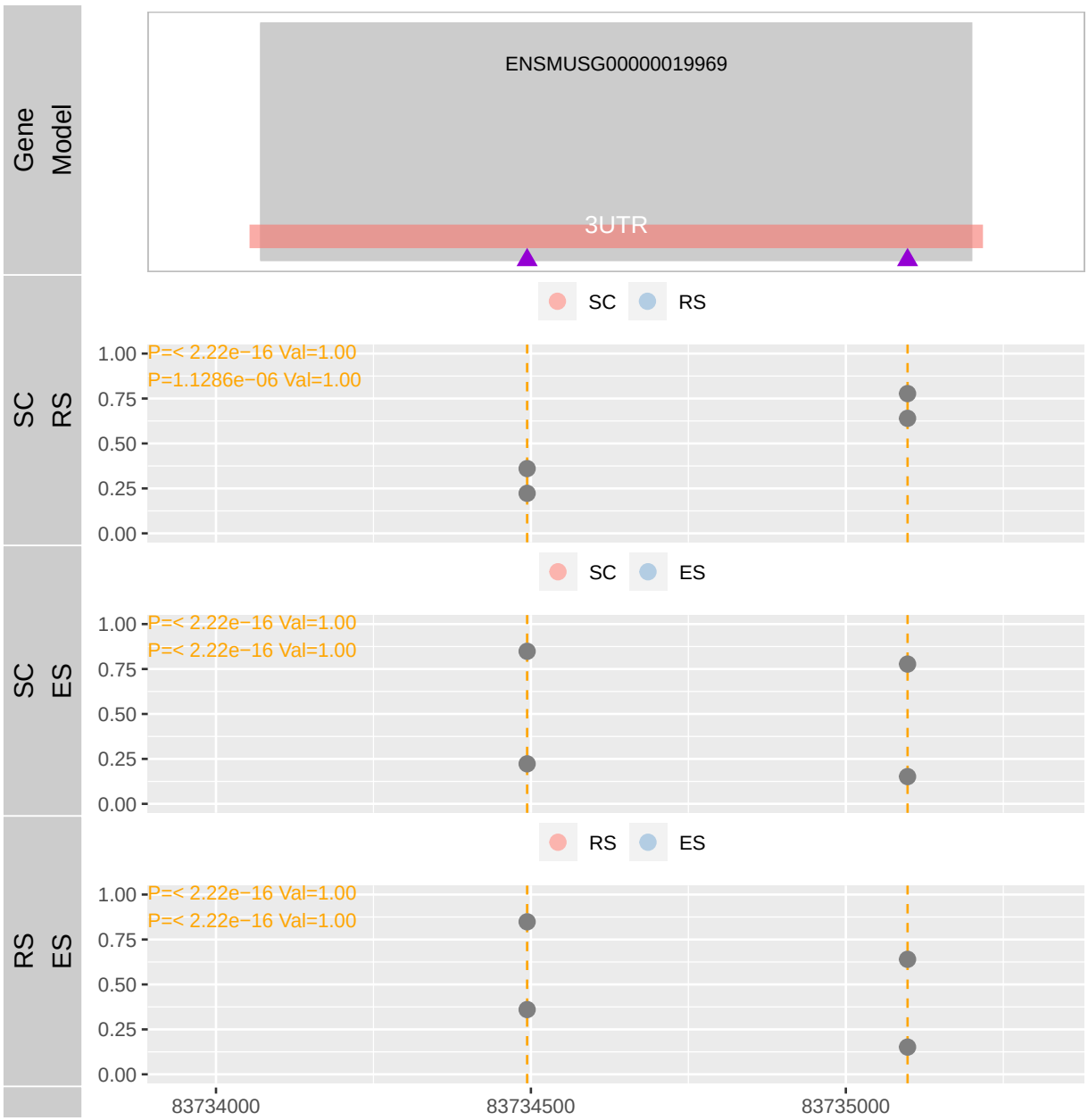
Visualize PACs of this gene in individual cell types. In the plot, the Y-axis is read count, the scale of which is different among conditions. Only show DE PACs with padj < padjThd and show 3'UTR region instead of the gene model.

```
movViz(object=DEqPAC, gene=gene, txdb=NULL,
       PACds=scPACdsCt,
       collapseConds=FALSE, padjThd=0.01,
       showRatio=FALSE, showAllPA=FALSE)
```



Show condition pairs in individual tracks. If padjThd is given, then the DE PACs (padj < padjThd) will be highlighted (dashed yellow line).

```
movViz(object=DEqPAC, gene=gene, txdb=NULL, PACds=scPACdsCt, collapseConds=T,
       padjThd=0.01, showPV=TRUE, showRatio=TRUE, showAllPA=T)
```



5.3 Fisher's exact test for APA genes

Here we detect genes with dynamic APA usages among cell types using Fisher's exact test. This is similar to the method (*test_apa*) used in (Shulman and Elkon, 2019). The Fisher's exact test is performed for genes with at least two 3'UTR PACs.

```
sw=movAPAswitch(PACds=scPACds, group='celltype',
                avgPACtag=0, avgGeneTag=0,
                only3UTR=TRUE, mergeReps='pool',
                aMovDEPACRes=NULL, DEPAC.padjThd=NULL, nDEPAC=0,
                mindist=0, fisherThd=0.05, logFCThd=0, cross=FALSE,
```

```
                selectOne='fisherPV')
#> SC.RS
#> SC.ES
#> RS.ES

head(sw@fullList$SC.RS, n=2)
#>                 gene nPAC geneTag1 geneTag2 avgUTRlen1 avgUTRlen2     fisherPV
#> 1 ENSMUSG00000002996    2       41      139   822.1707   429.2518 4.229427e-07
#> 2 ENSMUSG00000007411    2       74      807   317.4865   212.8761 5.571850e-24
#>      logFC change    PA1    PA2 dist nDEPA nSwitchPair                PAs1
#> 1 -2.82639     -1  PA535  PA536  856     0           2    PA535=8;PA536=33
#> 2 -4.47032     -1 PA1207 PA1208  248     0           1 PA1207=40;PA1208=34
#>                    PAs2
#> 1    PA535=91;PA536=48
#> 2 PA1207=778;PA1208=29
## Filter results with padj<0.05.
swstat=movStat(object=sw, padjThd=0.05, valueThd=0)
#> All cond pairs in heat@colData, get de01 and deNum
## Number of significant switching genes between cell types.
swstat$nsig
#>       sig.num
#> SC.RS     125
#> SC.ES     106
#> RS.ES      79
head(swstat$siglist$SC.RS)
#> [1] "ENSMUSG00000002996" "ENSMUSG00000007411" "ENSMUSG00000014905"
#> [4] "ENSMUSG00000017843" "ENSMUSG00000019969" "ENSMUSG00000020561"
```

5.4 Detecting 3'UTR switching genes

This is similar to the above Fisher's exact test, while it is stricter. The switching criteria: at least 1 DEqPAC; fisher's test of the two PACs pvalue<fisherThd; logFC of the two PACs>=logFCThd.

If more than one switching pair was found in a gene, only the pair with the smallest Fisher's test's pvalue (selectOne='fisherPV') was returned.

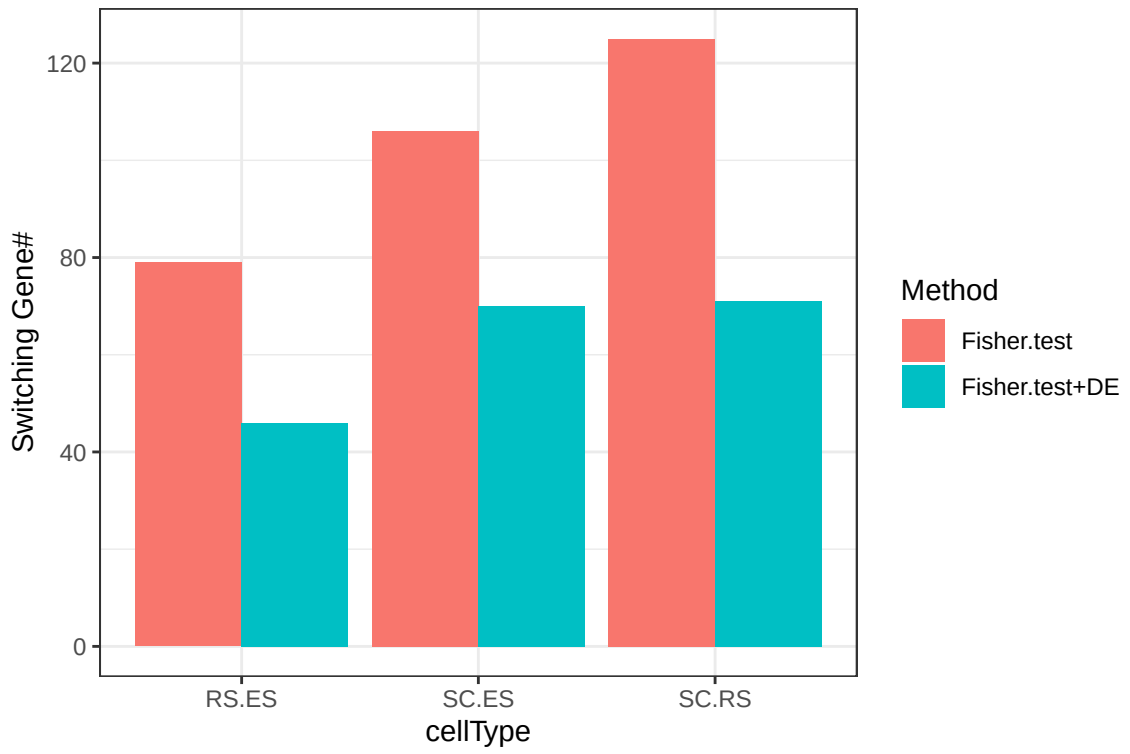
```
swDEq=movAPAswitch(PACds=scPACds, group='celltype',
                   avgPACtag=0, avgGeneTag=0,
                   only3UTR=TRUE, mergeReps='pool',
                   aMovDEPACRes=DEqPAC, DEPAC.padjThd=0.05, nDEPAC=1,
                   mindist=50, fisherThd=0.05, logFCThd=1,
                   cross=FALSE, selectOne='fisherPV')
#> SC.RS
#> SC.ES
#> RS.ES
```

Get 3'UTR switching results.

```
swDEqStat=movStat(object=swDEq, padjThd=0.01, valueThd=1, upThd=NULL, dnThd=NULL)
#> All cond pairs in heat@colData, get de01 and deNum
swDEqStat$nsig
#>       sig.num
#> SC.RS      71
#> SC.ES      70
#> RS.ES      46
```

Compare results from the two 3'UTR switching methods.

```
nsig=as.data.frame(cbind(swstat$nsig, swDEqStat$nsig))
colnames(nsig)=c('Fisher.test', 'Fisher.test+DE')
nsig$cellType=rownames(nsig)
nsig=reshape2::melt(nsig, variable.name='Method', value.name="SwitchingEvents")
ggplot(data=nsig, aes(x=cellType, y=SwitchingEvents, fill=Method)) +
  geom_bar(stat="identity", position=position_dodge()) +
  ylab("Switching Gene#") + theme_bw()
```



Plot heatmap to view switching status of each gene among all cell types. First convert the *movRes* object to a heatmap object.

```
heat=movRes2heatmapResults(swDEq)
```

Filter switching genes.

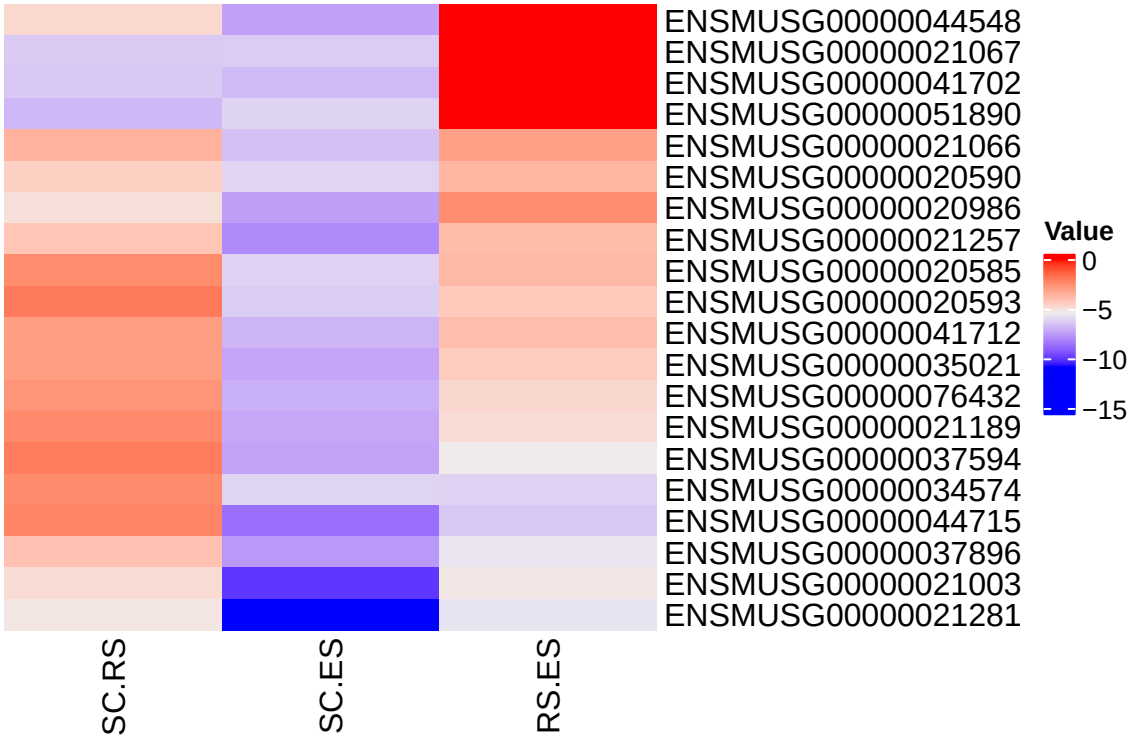
```
heat=subsetHeatmap(heat, padjThd=0.001, valueThd=2)
nrow(heat@value)
#> [1] 79
```

Select top 20 genes for the plot.

```
heat@value=heat@value[rowSums(is.na(heat@value))==0, ]
heat@value=heat@value[order(rowMeans(abs(heat@value)), decreasing =T ), ]
nrow(heat@value)
#> [1] 79
```

From the heatmap, we can see gene ENSMUSG00000021281 is shorter from SC to RS (value=-5), from SC to ES (value=-14), and from RS to ES (value=-5). This means that the length of 3'UTR of this gene is ES < RS < SC, which is consistent with the 3UTR shortening during sperm cell differentiation found in the previous study.

```
plotHeatmap(heat@value[1:20, ],  show_rownames=TRUE, plotPre=NULL)
```




```
heat@value['ENSMUSG00000021281',]
#>                          SC.RS     SC.ES     RS.ES
#> ENSMUSG00000021281 -5.161861 -10.87117 -5.709308
```

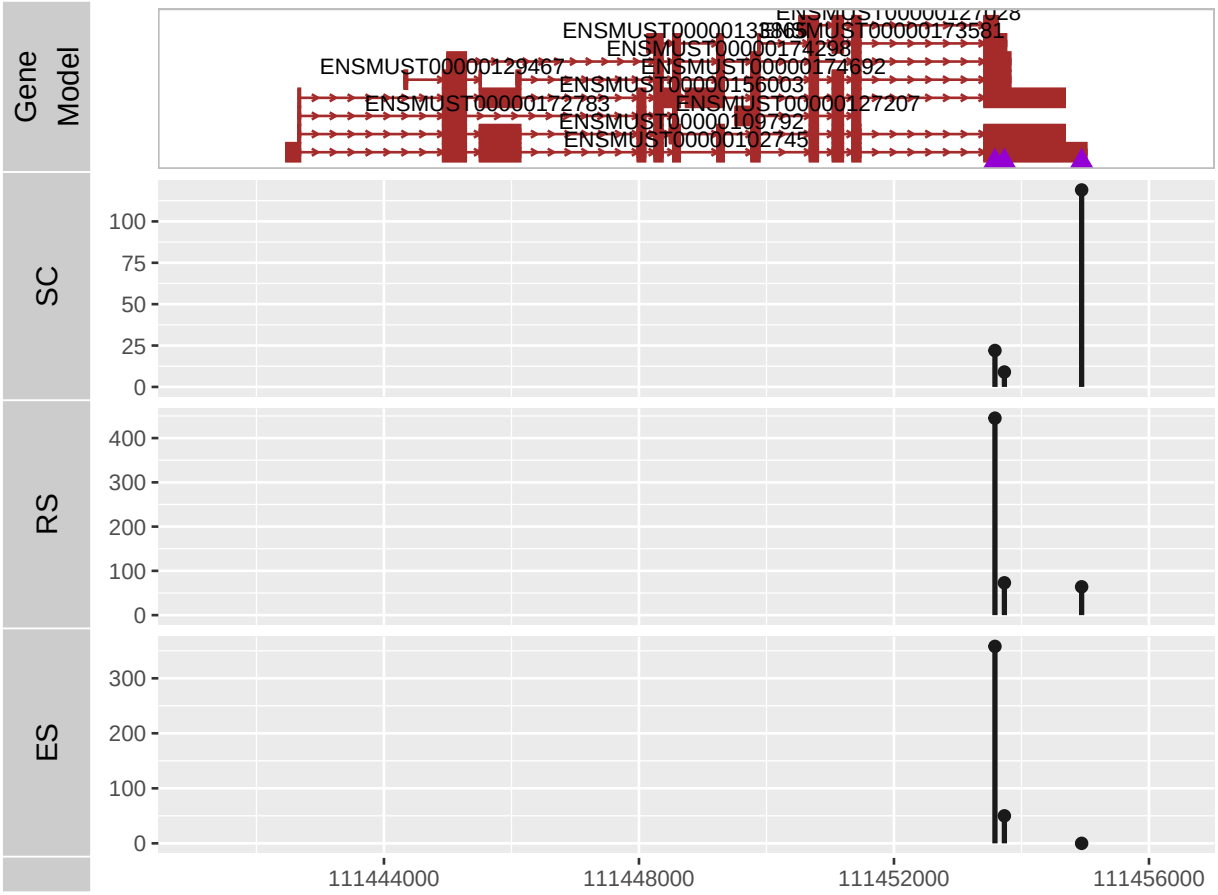
Get the APA switching list between SC and RS for this gene. This gene has three pairs of switching APA sites between SC and RS. But because we set selectOne='fisherPV' in the movAPAswitch function, only the pair with smallest pvalue was returned.

```
swDEq@fullList$SC.RS[swDEq@fullList$SC.RS$gene=='ENSMUSG00000021281',]
#>                  gene nPAC geneTag1 geneTag2 avgUTRlen1 avgUTRlen2     fisherPV
#> 36 ENSMUSG00000021281    2      141      509   1169.801    193.002 8.388516e-59
#>        logFC change    PA1    PA2 dist nDEPA nSwitchPair                  PAs1
#> 36 -5.161861     -1 PA3541 PA3543 1361     2           3 PA3541=22;PA3543=119
#>                    PAs2
#> 36 PA3541=445;PA3543=64
```

5.5 Visualization of 3'UTR switching genes

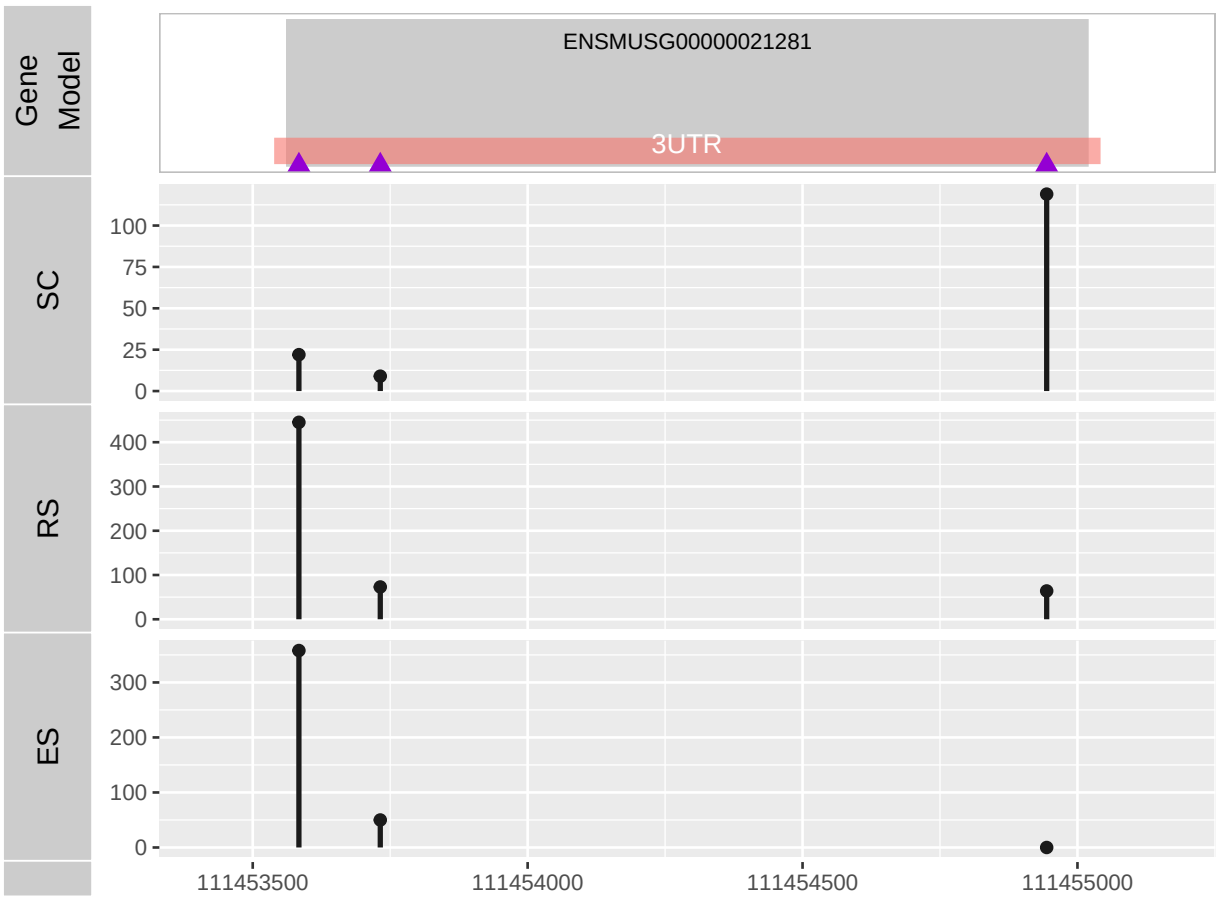
Use *movViz* to show this gene which has three 3'UTR PACs.

```
gene='ENSMUSG00000021281'
movViz(object=swDEq, gene=gene, txdb=gff, PACds=scPACdsCt)
```



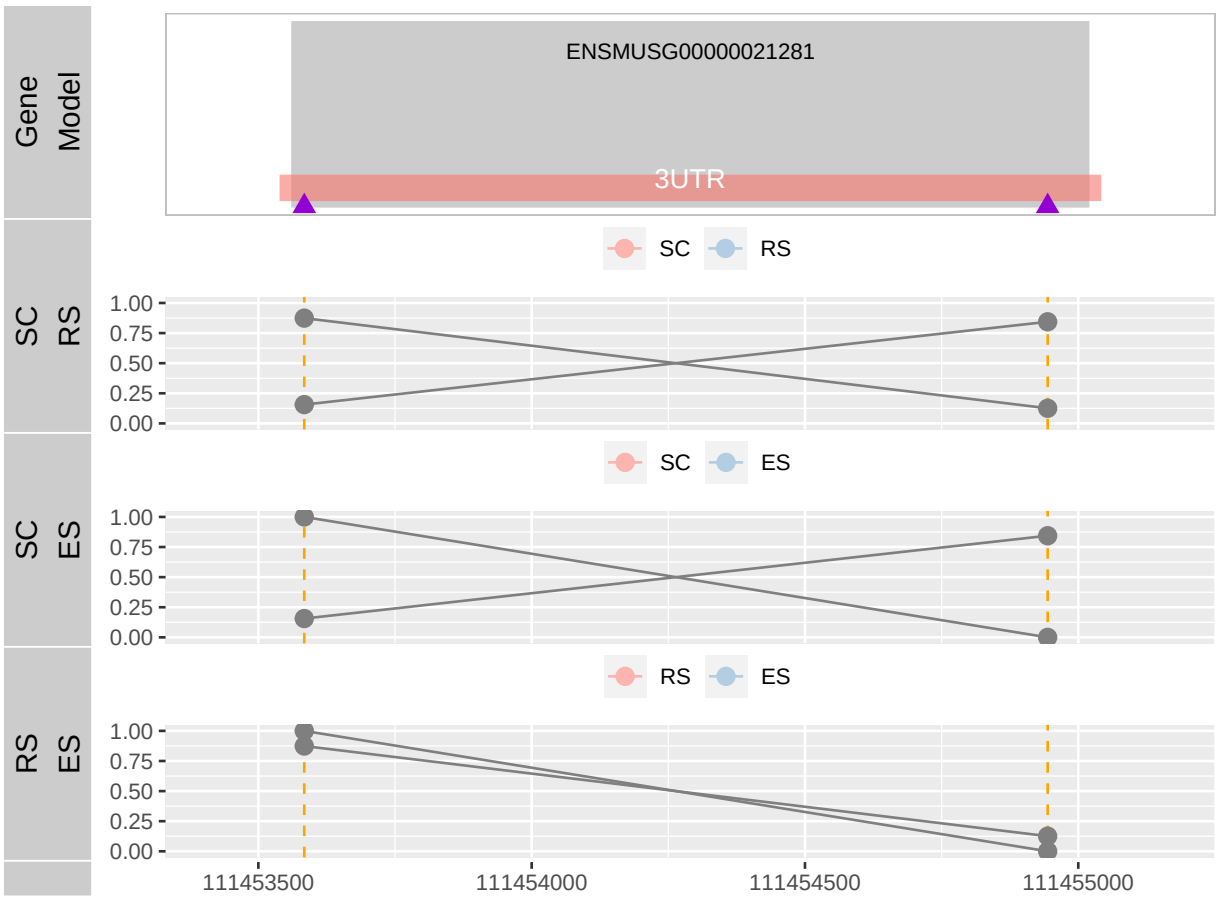
Just show the 3'UTR region.

```
movViz(object=swDEq, gene=gene, txdb=NULL, PACds=scPACdsCt)
```



Show only PACs involved in the 3'UTR switching.

```
movViz(object=swDEq, gene=gene, txdb=NULL, PACds=scPACdsCt, collapseConds=TRUE,
       conds=NULL, highlightConds=NULL, showRatio=TRUE, linkPAs=TRUE,
       padjThd=0.01, showAllPA=FALSE, showPV=FALSE)
```



5.6 Proximal PAC's GPI index

Here we calculate GPI index of each APA gene for each cell. GPI of a gene is the "geo"" score of the proximal poly(A) site. The "geo" metric measures the usage of a poly(A) site by the geometric mean, which was used for measuring poly(A) site usage in single cells (Shulman et al, 2019). First, filter 3'UTR's proximal and distal PACs.

```
ds=get3UTRAPAds(scPACds, sortPA=TRUE, choose2PA='PD')
gpi=movAPAindex(ds, method="GPI")
head(gpi[1:5, 1:5])
gpi=gpi[rowSums(is.na(gpi))==0, ]
```

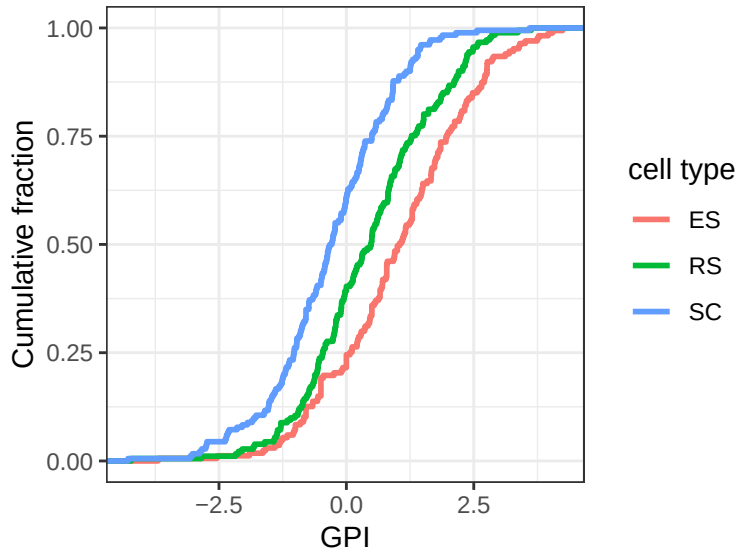
Calculate GPI for each cell type.

```
ds2=subsetPACds(scPACdsCt, group='celltype', pool=TRUE)
ds2=get3UTRAPAds(ds2, sortPA=TRUE, choose2PA='PD')
gpi2=movAPAindex(ds2, method="GPI")
summary(gpi2)
#>        SC                  RS                  ES
#>  Min.   :-4.2849   Min.   :-4.2247   Min.   :-3.69508
#>  1st Qu.:-1.0248   1st Qu.:-0.4604   1st Qu.: 0.09118
```

```
#>  Median :-0.3051   Median : 0.4860   Median : 1.04373
#>  Mean   :-0.3137   Mean   : 0.4510   Mean   : 1.00735
#>  3rd Qu.: 0.5000   3rd Qu.: 1.2804   3rd Qu.: 1.96527
#>  Max.   : 3.5931   Max.   : 3.6259   Max.   : 4.24392
#>  NA's   :1                           NA's   :14
```

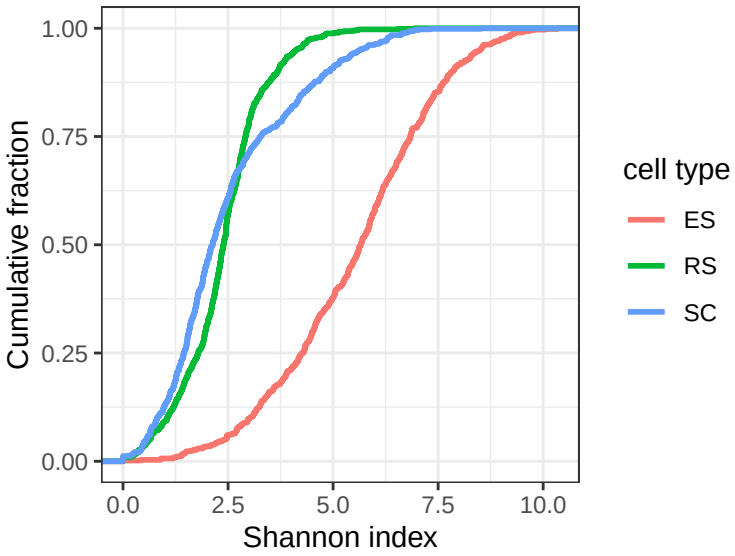
The plot of the distribution of GPI values is similar to the Fig. 3D of Shulman et al, 2019.

```
plotCummPAindex(PAindex=gpi2, groupName='cell type', xlab='GPI')
```



The plot of the distribution of Shannon index (tissue-specificity) for each PAC.

```
shan=movPAindex(scPACdsCt, method="shan")
#> Using count for Shannon.
#> Tissue-specific PAC's H_cutoff (mean-2*sd):  0.2826073
#> Tissue-specific PAC's Q_cutoff (mean-2*sd):  0.2489532
#> Tissue-specific PAC# (H<H_cutoff):  33
#> Tissue-specific PAC# (Q<Q_cutoff):  24
#> Constitutive PAC's H_cutoff (mean+2*sd):  1.585916
#> Constitutive PAC's Q_cutoff (mean+2*sd):  2.726938
#> Constitutive PACs (H>H_cutoff):  0
#> Constitutive PACs (Q>Q_cutoff):  8
plotCummPAindex(PAindex=shan[, -c(1:3)],
                groupName='cell type', xlab='Shannon index')
```



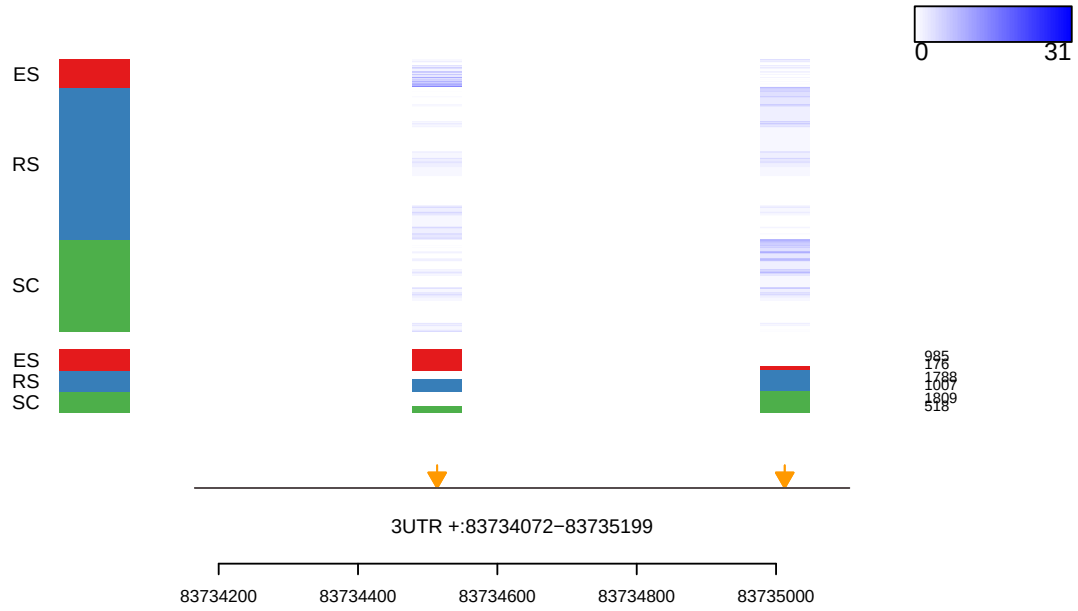
6 Visualize PACs in single cells

We can use *movVizSC* function which utilizes the R packages millefy (Ozaki et al., 2020) for the single-cell plot.

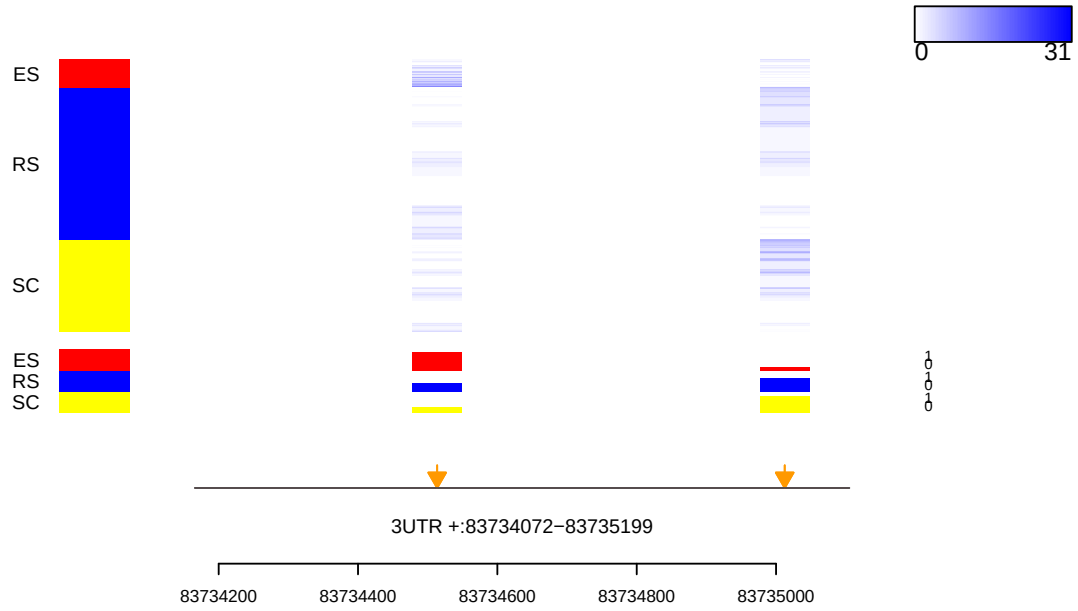
However, we recommend use the *vizAPA* package for more elegant plots.

```
gene='ENSMUSG00000019969'

movVizSC(scPACds, gene, cellGroupName='celltype', txdb=NULL,
         cellGroupColors=NULL, showRatio = F)
#> [1] "Begin Plot: 2023-10-10 16:48:17"
#> There was a problem when running diffusion map. Trying PCA instead...
#> The standard deviations of PC1: 1.091806
```




```
## Plot the expression levels of cell types as ratios, and specify colors for cell types.
movVizSC(scPACds, gene, cellGroupName='celltype', txdb=NULL,
         cellGroupColors=c(SC="yellow", ES="red",RS="blue"), showRatio = T)
#> [1] "Begin Plot: 2023-10-10 16:48:19"
#> There was a problem when running diffusion map. Trying PCA instead...
#> The standard deviations of PC1: 1.091806
```



7 Session Information

The session information records the versions of all the packages used in the generation of the present document.

```
sessionInfo()
#> R version 4.2.2 (2022-10-31 ucrt)
#> Platform: x86_64-w64-mingw32/x64 (64-bit)
#> Running under: Windows 10 x64 (build 22621)
#>
#> Matrix products: default
#>
#> locale:
#> [1] LC_COLLATE=Chinese (Simplified)_China.utf8
#> [2] LC_CTYPE=Chinese (Simplified)_China.utf8
#> [3] LC_MONETARY=Chinese (Simplified)_China.utf8
#> [4] LC_NUMERIC=C
#> [5] LC_TIME=Chinese (Simplified)_China.utf8
#>
#> attached base packages:
#> [1] stats4    stats     graphics  grDevices utils     datasets  methods
```

```
#> [8] base
#>
#> other attached packages:
#>  [1] TxDb.Mmusculus.UCSC.mm10.ensGene_3.4.0
#>  [2] GenomicFeatures_1.50.2
#>  [3] AnnotationDbi_1.60.0
#>  [4] Biobase_2.58.0
#>  [5] GenomicRanges_1.50.1
#>  [6] GenomeInfoDb_1.34.9
#>  [7] IRanges_2.32.0
#>  [8] S4Vectors_0.36.0
#>  [9] BiocGenerics_0.44.0
#> [10] ggplot2_3.4.0
#> [11] movAPA_0.2.0
#>
#> loaded via a namespace (and not attached):
#>   [1] utf8_1.2.2                  tidyselect_1.2.0
#>   [3] RSQLite_2.2.18              htmlwidgets_1.5.4
#>   [5] grid_4.2.2                  ranger_0.14.1
#>   [7] BiocParallel_1.32.1         munsell_0.5.0
#>   [9] destiny_3.12.0              codetools_0.2-18
#>  [11] interp_1.1-3                withr_2.5.0
#>  [13] colorspace_2.0-3            filelock_1.0.2
#>  [15] OrganismDbi_1.40.0          highr_0.9
#>  [17] knitr_1.41                  rstudioapi_0.14
#>  [19] SingleCellExperiment_1.20.0 robustbase_0.95-0
#>  [21] vcd_1.4-11                  VIM_6.2.2
#>  [23] TTR_0.24.3                  MatrixGenerics_1.10.0
#>  [25] labeling_0.4.2              GenomeInfoDbData_1.2.9
#>  [27] bit64_4.0.5                 farver_2.1.1
#>  [29] vctrs_0.5.1                 generics_0.1.3
#>  [31] xfun_0.35                   ggthemes_4.2.4
#>  [33] biovizBase_1.46.0           BiocFileCache_2.6.0
#>  [35] R6_2.5.1                    doParallel_1.0.17
#>  [37] clue_0.3-63                 RcppEigen_0.3.3.9.3
#>  [39] locfit_1.5-9.6              AnnotationFilter_1.22.0
#>  [41] bitops_1.0-7                cachem_1.0.6
#>  [43] reshape_0.8.9               DelayedArray_0.24.0
#>  [45] assertthat_0.2.1            BiocIO_1.8.0
#>  [47] scales_1.2.1                nnet_7.3-18
#>  [49] gtable_0.3.1                Cairo_1.6-0
#>  [51] ggbio_1.46.0                ensembldb_2.22.0
#>  [53] rlang_1.0.6                 scatterplot3d_0.3-43
#>  [55] GlobalOptions_0.1.2         splines_4.2.2
#>  [57] rtracklayer_1.58.0          lazyeval_0.2.2
#>  [59] hexbin_1.28.3               dichromat_2.0-0.1
#>  [61] checkmate_2.1.0             BiocManager_1.30.19
```

```
#>  [63] yaml_2.3.6                  reshape2_1.4.4
#>  [65] abind_1.4-5                 backports_1.4.1
#>  [67] Hmisc_5.0-0                 RBGL_1.74.0
#>  [69] tools_4.2.2                 ellipsis_0.3.2
#>  [71] RColorBrewer_1.1-3          proxy_0.4-27
#>  [73] Rcpp_1.0.9                  plyr_1.8.8
#>  [75] base64enc_0.1-3             progress_1.2.2
#>  [77] zlibbioc_1.44.0             purrr_0.3.5
#>  [79] RCurl_1.98-1.9              prettyunits_1.1.1
#>  [81] rpart_4.1.19                deldir_1.0-6
#>  [83] GetoptLong_1.0.5            zoo_1.8-11
#>  [85] SummarizedExperiment_1.28.0 cluster_2.1.4
#>  [87] tinytex_0.43                magrittr_2.0.3
#>  [89] data.table_1.14.6           RSpectra_0.16-1
#>  [91] magick_2.7.3                circlize_0.4.15
#>  [93] lmtest_0.9-40               pcaMethods_1.90.0
#>  [95] ProtGenerics_1.30.0         matrixStats_0.63.0
#>  [97] hms_1.1.2                   evaluate_0.18
#>  [99] smoother_1.1                XML_3.99-0.12
#> [101] jpeg_0.1-10                 gridExtra_2.3
#> [103] shape_1.4.6                 compiler_4.2.2
#> [105] biomaRt_2.54.0              tibble_3.1.8
#> [107] crayon_1.5.2                htmltools_0.5.3
#> [109] Formula_1.2-4               tidyr_1.2.1
#> [111] DBI_1.1.3                   dbplyr_2.2.1
#> [113] ComplexHeatmap_2.14.0       MASS_7.3-58.1
#> [115] rappdirs_0.3.3              boot_1.3-28
#> [117] Matrix_1.5-3                car_3.1-1
#> [119] cli_3.4.1                   parallel_4.2.2
#> [121] pkgconfig_2.0.3             GenomicAlignments_1.34.0
#> [123] foreign_0.8-83              laeken_0.5.2
#> [125] sp_1.5-1                    xml2_1.3.3
#> [127] foreach_1.5.2               XVector_0.38.0
#> [129] stringr_1.4.1               VariantAnnotation_1.44.0
#> [131] digest_0.6.30               graph_1.76.0
#> [133] Biostrings_2.66.0           rmarkdown_2.18
#> [135] htmlTable_2.4.1             edgeR_3.40.0
#> [137] restfulr_0.0.15             curl_4.3.3
#> [139] ggplot.multistats_1.0.0     Rsamtools_2.14.0
#> [141] rjson_0.2.21                lifecycle_1.0.3
#> [143] carData_3.0-5               limma_3.54.0
#> [145] BSgenome_1.66.2             fansi_1.0.3
#> [147] pillar_1.8.1                lattice_0.20-45
#> [149] GGally_2.1.2                KEGGREST_1.38.0
#> [151] fastmap_1.1.0               httr_1.4.4
#> [153] DEoptimR_1.0-11             survival_3.4-0
#> [155] xts_0.13.0                  glue_1.6.2
```

```
#> [157] png_0.1-7                   iterators_1.0.14
#> [159] bit_4.0.5                   class_7.3-20
#> [161] stringi_1.7.8               blob_1.2.3
#> [163] RcppHNSW_0.4.1              latticeExtra_0.6-30
#> [165] memoise_2.0.1               dplyr_1.0.10
#> [167] irlba_2.3.5.1               e1071_1.7-13
```

8 References

[1] Shulman, E.D. and Elkon, R. (2019) Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. Nucleic Acids Res., 47, 10027-10039.

[2] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. Genome Biol., 11, 2010-2011.

[3] Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26, 139-140.

[4] Ozaki, H., Hayashi, T., Umeda, M. and Nikaido, I. (2020) Millefy: visualizing cell-to-cell heterogeneity in read coverage of single-cell RNA sequencing datasets. BMC Genomics, 21, 177.