scAPAmod: profiling alternative polyadenylation modalities in single cells from single-cell RNA-seq data

Xiaohui Wu, Hongjuan Fu

2021-05-20

Contents

1 Preparations **1**

1.1 PAC data of mouse sperm cells 1

1.2 Preprocess 2

2 Analyses of APA dynamics **3**

2.1 Identifying modalities in 3' UTR 3

2.2 Identifying modalities in non 3' UTR 3

2.3 Research APA preferences 5

3 Statistics of modalities **6**

3.1 Statistics on modalities distribution of different cell types 6

3.2 Distribution of usage modalties at different stages of differentiation 7

3.3 Changes in the modalities of different cell differentiation stages 8

3.4 Visualize the distribution of PA expression according to the components 9

3.5 GO analysis 10

4 Session Information **11**

1 Preparations

1.1 PAC data of mouse sperm cells

```
library(movAPA, warn.conflicts = FALSE, quietly=TRUE)
data(scPACds)
head(scPACds@counts[1:2,1:5])
       AAACCTGAGAGGGCTT AAACCTGAGCTTATCG AAACCTGCATACGCCG AAACCTGGTTGAGTTC
PA3443                0                0                0                0
PA3446                0                0                0                0
       AAACCTGTCAACGAAA
PA3443                0
PA3446                0
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"
```

1.2 Preprocess

```
library(scAPAmod, warn.conflicts = FALSE, quietly=TRUE)
# 3' UTR
index <- which(scPACds@anno$ftr == "3UTR")
UTR_gene <- scPACds@anno$gene[index]
UTR_chr <- scPACds@anno$chr[index]
UTR_strand <- scPACds@anno$strand[index]
UTR_coord <- scPACds@anno$coord[index]
UTR_ftr_start <- scPACds@anno$ftr_start[index]
UTR_ftr_end <- scPACds@anno$ftr_end[index]
UTR_three_UTR_length <- scPACds@anno$three_UTR_length[index]
UTR_anno <- data.frame(chr = as.character(UTR_chr),
                       strand = as.character(UTR_strand),
                       coord =as.integer(UTR_coord),
                       gene = as.character(UTR_gene),
```

```
                        ftr_start = as.integer(UTR_ftr_start),
                        ftr_end = as.integer(UTR_ftr_end),
                        three_UTR_length = as.integer(UTR_three_UTR_length))
UTR_counts <-  scPACds@counts[,index]
ct1 <- which(scPACds@colData$celltype[index] == "SC")
result1 <- extrPairPA(UTR_counts[,ct1],
                      as.character(UTR_anno$gene),UTR_anno)
8 PACs
# non 3' UTR
ct <- which(scPACds@colData$celltype == "SC")
results <- exnon3UTRPA(scPACds@counts[,ct],
                        scPACds@anno$gene, scPACds@anno, scPACds@anno$ftr,
                        gn = 1, cn = 1)
114 PACs
```

2 Analyses of APA dynamics

2.1 Identifying modalities in 3' UTR

```
mod <- getMod(result1$PUI)
mod$modalities
[1] "Multimodal" "Unimodal"   "Unimodal"   "Bimodal"
```

if you want to see the modalities directly in 3' UTR, you can use *UTRmod*

```
scUTRds <- get3UTRAPAds(scPACds)
mod <- UTRmod(scUTRds,ct1)
6 PACs
mod$modalities
[1] "Unimodal" "Unimodal" "Unimodal"
```

2.2 Identifying modalities in non 3' UTR

```
nonmod <- getMod(results$PUI)
nonmod$modalities
 [1] "Unimodal"   "Unimodal"   "Bimodal"    "Multimodal" "Bimodal"
 [6] "Bimodal"    "Multimodal" "Bimodal"    "Multimodal" "Unimodal"
[11] "Multimodal" "Bimodal"    "Bimodal"    "Unimodal"   "Multimodal"
[16] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[21] "Multimodal" "Bimodal"    "Bimodal"    "Multimodal" "Multimodal"
[26] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[31] "Multimodal" "Multimodal" "Bimodal"    "Multimodal" "Unimodal"
[36] "Unimodal"   "Multimodal" "Multimodal" "Bimodal"    "Multimodal"
```

```
[41] "Multimodal" "Bimodal"    "Multimodal" "Multimodal" "Multimodal"
[46] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[51] "Multimodal" "Multimodal" "Bimodal"    "Bimodal"    "Multimodal"
[56] "Multimodal" "Bimodal"    "Multimodal" "Multimodal" "Multimodal"
[61] "Bimodal"    "Bimodal"    "Bimodal"    "Multimodal" "Multimodal"
[66] "Bimodal"    "Multimodal" "Bimodal"    "Multimodal" "Multimodal"
[71] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[76] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Bimodal"
[81] "Multimodal" "Bimodal"
# use chi-square test to test Bimodal
ind <- which(nonmod$modalities == "Bimodal")
bigene <- results$gene[ind]
label <- lapply(c(1:length(nonmod$results)), function(y){
  la <- nonmod$results[[y]][[2]][["cluster_labels"]]
  if (length(which(is.na(results$PUI[y,])==TRUE))>0) {
    dat.tmp <- results$PUI[y,][-which(is.na(results$PUI[y,]))]
  }else{
    dat.tmp <- results$PUI[y,]
  }
  names(la) <- names(dat.tmp)
  return(la)})
bilabel <- label[ind]

# if(length(which(is.na(bigene)))>0){
#   id <- which(is.na(bigene))
#   bigene <- bigene[-id]
#   bilabel <- bilabel[-id]}
pval1 <- chisqtest(results$filter.data, results$gene, bigene, bilabel)
pval1 <- p.adjust(pval1, method = "BH")
# use KS test to test Bimodal
pval2 <- KStest(results$PUI, results$gene, results$ftr, bigene, bilabel)
pval2 <- p.adjust(pval2, method = "BH")
```

if you want to see the modalities directly in non 3' UTR, you can use *nonUTRmod*

```
nonmod <- nonUTRmod(scPACds,ct,gn = 1, cn = 1)
114 PACs
nonmod$modalities
 [1] "Unimodal"   "Unimodal"   "Bimodal"    "Multimodal" "Bimodal"
 [6] "Bimodal"    "Multimodal" "Bimodal"    "Multimodal" "Unimodal"
[11] "Multimodal" "Bimodal"    "Bimodal"    "Unimodal"   "Multimodal"
[16] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[21] "Multimodal" "Bimodal"    "Bimodal"    "Multimodal" "Multimodal"
[26] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[31] "Multimodal" "Multimodal" "Bimodal"    "Multimodal" "Unimodal"
[36] "Unimodal"   "Multimodal" "Multimodal" "Bimodal"    "Multimodal"
[41] "Multimodal" "Bimodal"    "Multimodal" "Multimodal" "Multimodal"
```

```
[46] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[51] "Multimodal" "Multimodal" "Bimodal"    "Bimodal"    "Multimodal"
[56] "Multimodal" "Bimodal"    "Multimodal" "Multimodal" "Multimodal"
[61] "Bimodal"    "Bimodal"    "Bimodal"    "Multimodal" "Multimodal"
[66] "Bimodal"    "Multimodal" "Bimodal"    "Multimodal" "Multimodal"
[71] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Multimodal"
[76] "Multimodal" "Multimodal" "Multimodal" "Multimodal" "Bimodal"
[81] "Multimodal" "Bimodal"
```

2.3 Research APA preferences

There are two types of APA preferences, one is the major PA, and the other is the minor PA. Here, the ratio value of the APA data is calculated, and the largest is extracted as the major PA, and the smallest is the minor PA. Analyze the modalities of APA usage with different preferences and the distribution of APA usage modalities in different regions.

```
# the major PA
mmod <- MAMIMod(scPACds,"SC","MajorPA")
677 PACs
mmod$modalities
[1] "Multimodal" "Multimodal" "Bimodal"    "Multimodal"
# the minor PA
mimod <- MAMIMod(scPACds,"SC","MinorPA")
677 PACs
mimod$modalities
[1] "Unimodal"   "Multimodal" "Bimodal"    "Unimodal"
```

If you want to see the ratio value specifically, you can use *exMajorPA*, or if you want to directly model the recognition modalities, you can use *getMMod* which is a bit different from PUI data modeling.

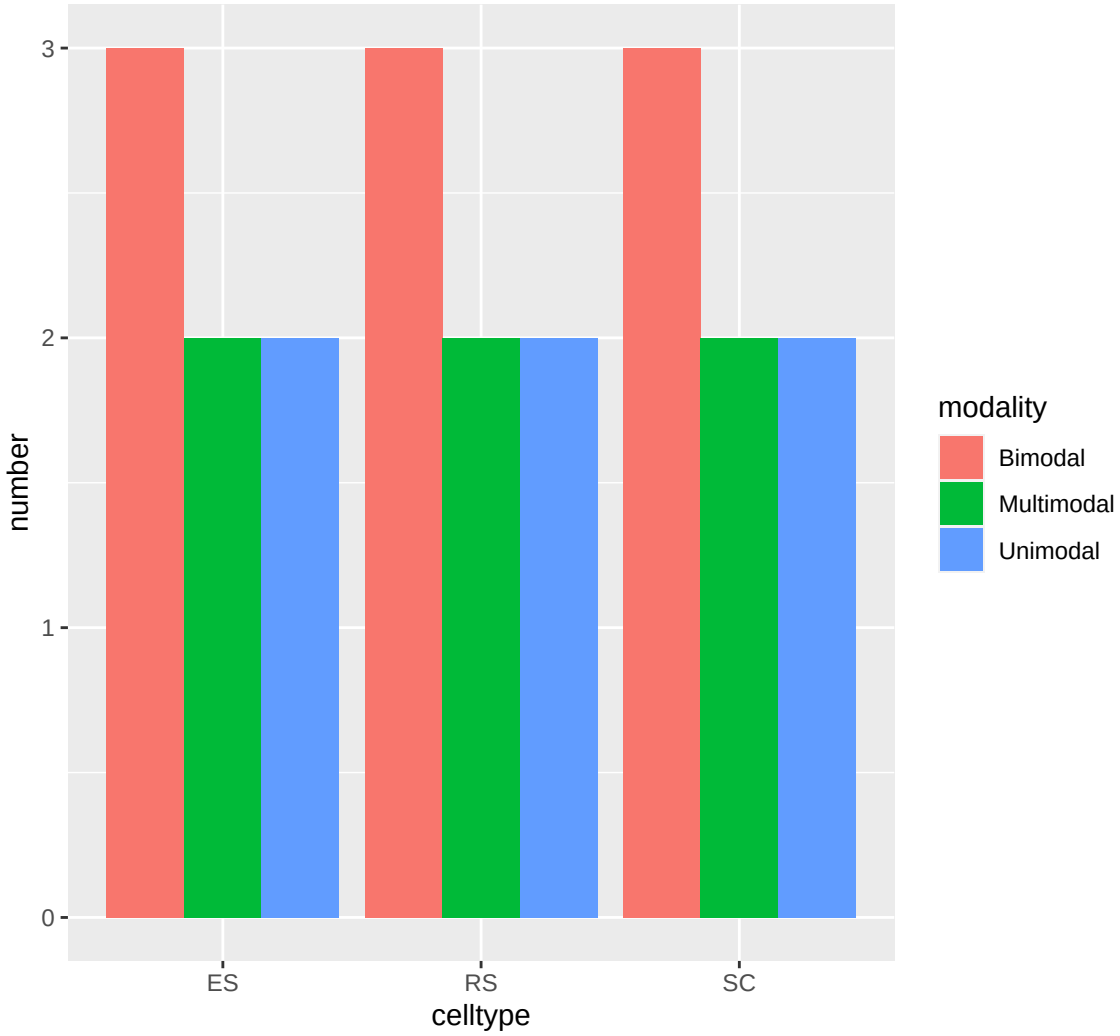
```
# the major PA and minor PA
mresult <- exMajorPA(scPACds,"SC")
677 PACs
# modalities of majorPA
mmod <- getMMod(mresult$PAmax,"PAmax")
mmod$modalities
[1] "Multimodal" "Multimodal" "Bimodal"    "Multimodal"
# modalities of minorPA
mimod <- getMMod(mresult$PAmin,"PAmin")
mimod$modalities
[1] "Unimodal"   "Multimodal" "Bimodal"    "Unimodal"
```

3 Statistics of modalities

3.1 Statistics on modalities distribution of different cell types

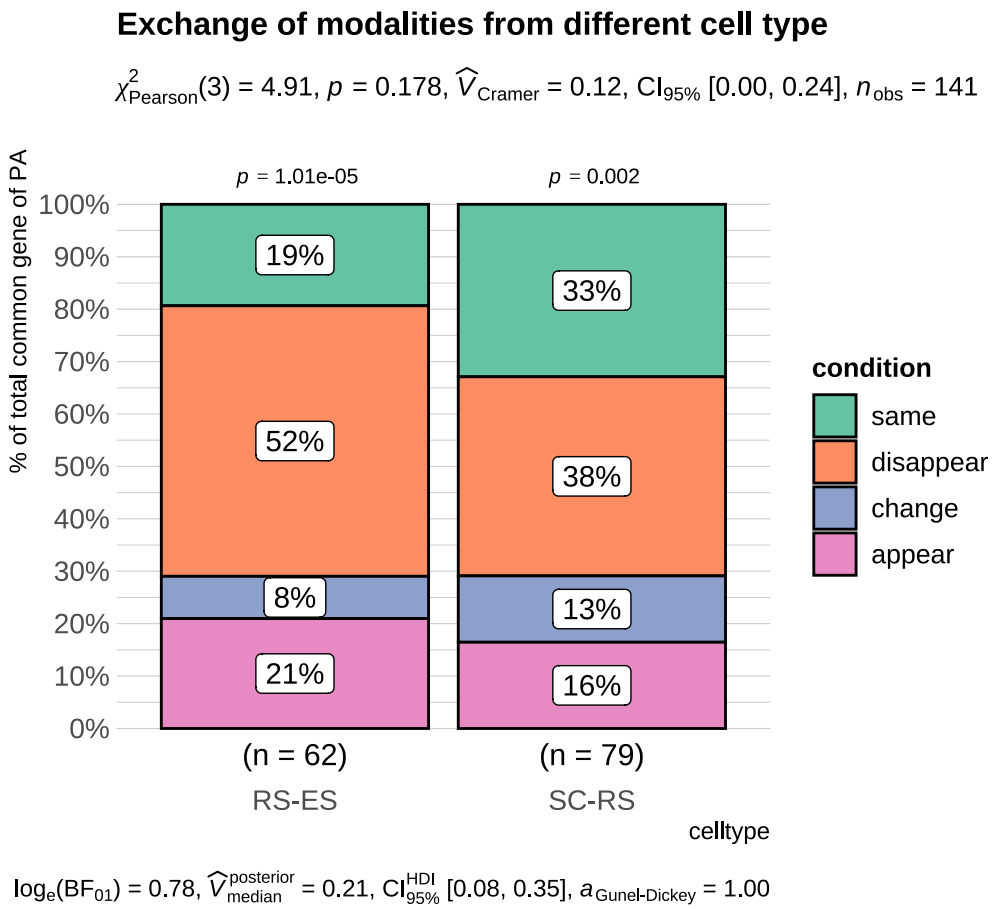
```
# cell type of RS
ct2 <- which(scPACds@colData$celltype[index] == "RS")
result2 <- extrPairPA(UTR_counts[,ct2],
                      as.character(UTR_anno$gene),UTR_anno)
4 PACs
mod2 <- getMod(result2$PUI)
mod2$modalities
[1] "Bimodal" "Bimodal"
# cell type of ES
ct3 <- which(scPACds@colData$celltype[index] == "ES")
result3 <- extrPairPA(UTR_counts[,ct3],
                      as.character(UTR_anno$gene),UTR_anno)
4 PACs
mod3 <- getMod(result3$PUI)
mod3$modalities
[1] "Unimodal" "Unimodal"
```

```
# set the cell type
celltype <- c(rep("SC", 3), rep("RS", 3), rep("ES", 3))
data <- data.frame(celltype)
data$modality <- c("Bimodal","Multimodal","Unimodal")
data$number <- c(table(mod$modalities),table(mod2$modalities),table(mod3$modalities))
ggplot(data, aes(x=celltype, y=number)) +
  ggplot2::geom_bar(stat = "identity", position = "dodge", aes(fill=modality))
```



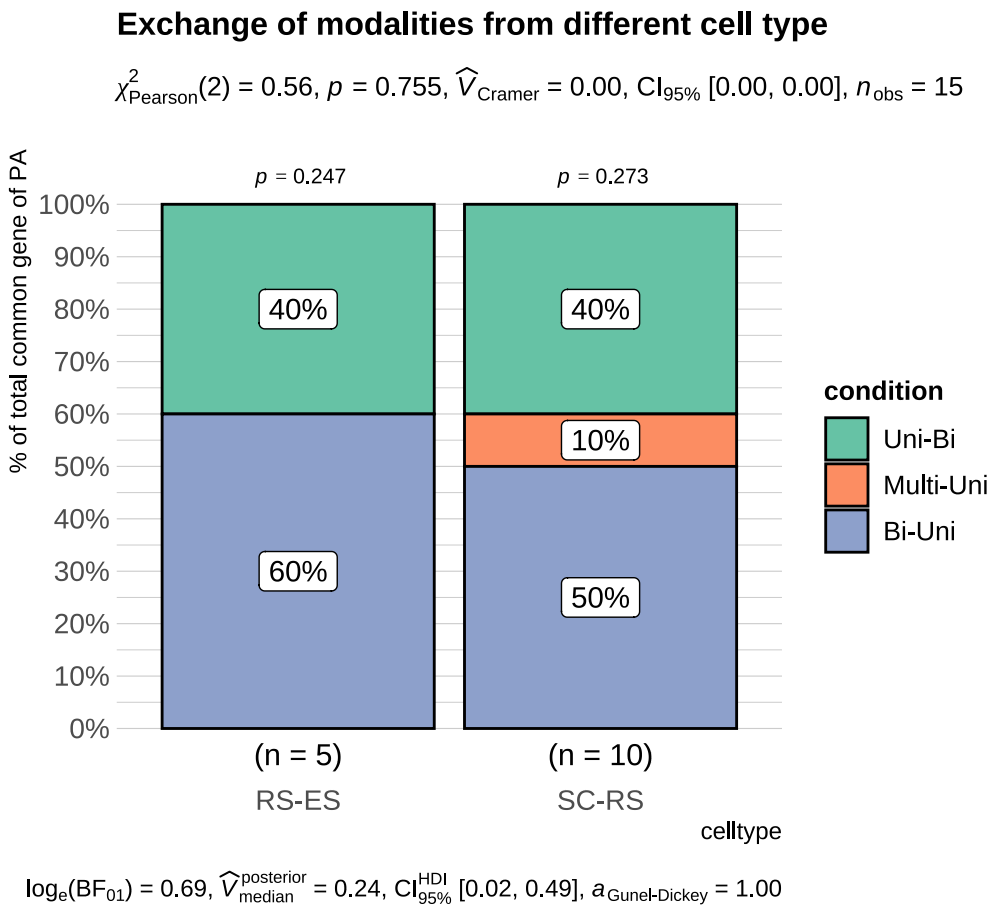
3.2 Distribution of usage modalties at different stages of differentiation

```
library(ggplot2)
library(ggstatsplot)
#extrafont::loadfonts()
data("tUTRModalChange")
ggstatsplot::ggbarstats(data = staChange, x = condition, y = celltype,
                        title = "Exchange of modalities from different cell type",
                        ylab = "% of total common gene of PA",
                        # ggstatsplot.layer = FALSE,
                        sampling.plan = "jointMulti",
                        ggtheme = hrbrthemes::theme_ipsum_pub(),
                        legend.title = "condition", messages = F, palette = "Set2")
```



3.3 Changes in the modalities of different cell differentiation stages

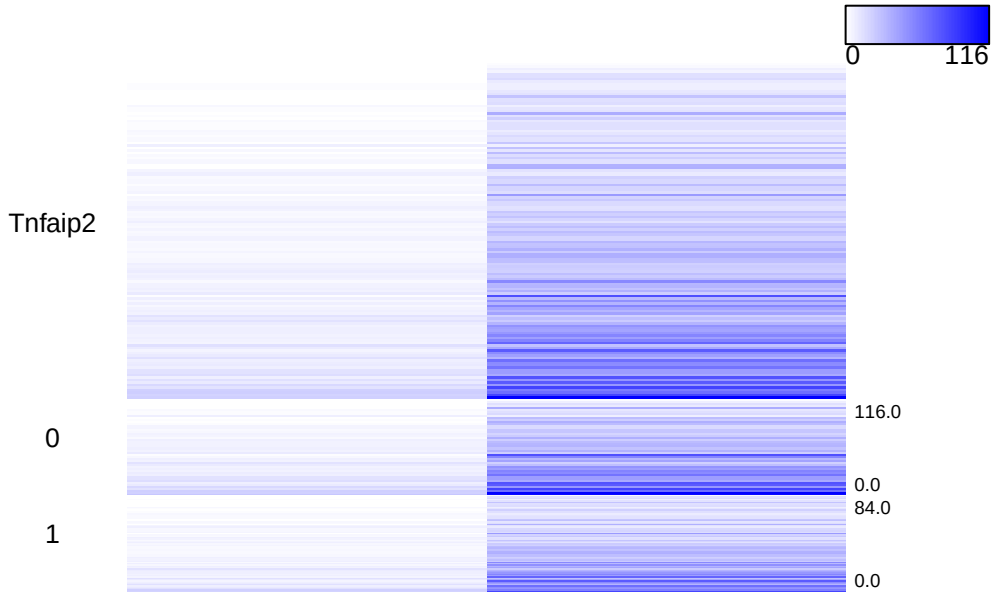
```
library(extrafont)
data("tUTRModalChangeDetail")
ggstatsplot::ggbarstats(data = detChange, x = condition, y = celltype,
                        title = "Exchange of modalities from different cell type",
                        ylab = "% of total common gene of PA",
                        # ggstatsplot.layer = FALSE,
                        sampling.plan = "jointMulti",
                        ggtheme = hrbrthemes::theme_ipsum_pub(),
                        legend.title = "condition", messages = F, palette = "Set2")
```



3.4 Visualize the distribution of PA expression according to the components

```
# heatmap
library(grid)
library(org.Mm.eg.db)
PUI <- result1$PUI
tUTR.pair.cd.tmp <- result1$filter.data
tUTR.gene <- result1$gene
tUTR.gene.pui <- rownames(PUI)
which(mod$modalities == "Bimodal")
integer(0)
id2 <- tUTR.gene.pui[3]
genename <- select(org.Mm.eg.db, keys = id2,
                   columns = c("SYMBOL","ENTREZID","GENENAME"),
                   keytype = "ENSEMBL")
tUTR.gene <- select(org.Mm.eg.db, keys = tUTR.gene,
```

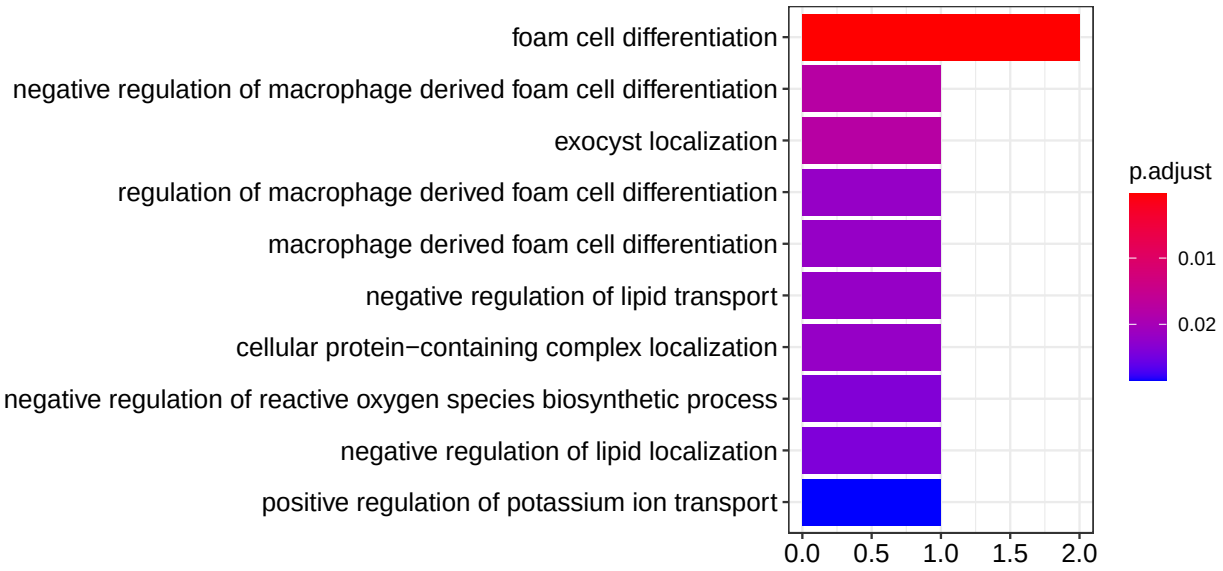
```
                    columns = c("SYMBOL","ENTREZID","GENENAME"),
                    keytype = "ENSEMBL")
plotGenePACount3(org.Mm.eg.db, genename$ENTREZID, tUTR.pair.cd.tmp,
                 tUTR.gene$ENTREZID,label[[3]])
There was a problem when running diffusion map. Trying PCA instead...
The standard deviations of PC1: 1.230856
```



3.5 GO analysis

```
library(clusterProfiler)
scego <- enrichGO(OrgDb="org.Mm.eg.db", gene = rownames(result1$PUI),
                  keyType = "ENSEMBL", ont = "ALL", pAdjustMethod = "BH",
                  pvalueCutoff = 0.05,qvalueCutoff = 0.05, readable= TRUE)

barplot(scego,showCategory = 10)
```



4 Session Information

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

```
sessionInfo()
R version 3.6.0 (2019-04-26)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19042)

Matrix products: default

locale:
[1] LC_COLLATE=Chinese (Simplified)_China.936
[2] LC_CTYPE=Chinese (Simplified)_China.936
[3] LC_MONETARY=Chinese (Simplified)_China.936
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_China.936

attached base packages:
 [1] grid      stats4    parallel  stats     graphics  grDevices utils
 [8] datasets  methods   base

other attached packages:
 [1] clusterProfiler_3.14.3      org.Mm.eg.db_3.10.0
 [3] extrafont_0.17              ggstatsplot_0.6.6
 [5] scAPAmod_0.1.0              ClusterR_1.2.2
 [7] gtools_3.8.2                movAPA_0.1.0
```

```
 [9] DEXSeq_1.32.0                DESeq2_1.26.0
[11] SummarizedExperiment_1.16.1 DelayedArray_0.12.3
[13] BiocParallel_1.20.1         matrixStats_0.57.0
[15] GenomicFeatures_1.38.2      AnnotationDbi_1.48.0
[17] Biobase_2.46.0              ggbio_1.34.0
[19] BSgenome_1.54.0             rtracklayer_1.46.0
[21] Biostrings_2.54.0           XVector_0.26.0
[23] ggplot2_3.3.2               data.table_1.13.2
[25] RColorBrewer_1.1-2          GenomicRanges_1.38.0
[27] GenomeInfoDb_1.22.1         IRanges_2.20.2
[29] S4Vectors_0.24.4            BiocGenerics_0.32.0
[31] reshape2_1.4.4              dplyr_1.0.2

loaded via a namespace (and not attached):
  [1] Hmisc_4.4-1               class_7.3-17
  [3] ps_1.4.0                  Rsamtools_2.2.3
  [5] lmtest_0.9-38             crayon_1.3.4
  [7] V8_3.4.0                  MASS_7.3-53
  [9] PMCMRplus_1.7.1           nlme_3.1-150
 [11] backports_1.1.10          metafor_2.4-0
 [13] ggcorrplot_0.1.3          GOSemSim_2.12.1
 [15] rlang_0.4.8               readxl_1.3.1
 [17] performance_0.6.0         extrafontdb_1.0
 [19] nloptr_1.2.2.2            callr_3.5.1
 [21] bit64_4.0.5               loo_2.4.0
 [23] glue_1.4.2                rstan_2.21.2
 [25] processx_3.4.4            DOSE_3.12.0
 [27] haven_2.3.1               tidyselect_1.1.0
 [29] rio_0.5.16                XML_3.99-0.3
 [31] tidyr_1.1.2               zoo_1.8-8
 [33] sysfonts_0.8.3            SuppDists_1.1-9.5
 [35] GenomicAlignments_1.22.1  mc2d_0.1-18
 [37] xtable_1.8-4              MatrixModels_0.4-1
 [39] magrittr_1.5              evaluate_0.14
 [41] gdtools_0.2.2             cli_2.1.0
 [43] zlibbioc_1.32.0           hwriter_1.3.2
 [45] rstudioapi_0.11           miniUI_0.1.1.1
 [47] rpart_4.1-15              fastmatch_1.1-0
 [49] ensembldb_2.10.2          shiny_1.5.0
 [51] xfun_0.22                 askpass_1.1
 [53] parameters_0.10.0         inline_0.3.17
 [55] pkgbuild_1.1.0            cluster_2.1.0
 [57] bridgesampling_1.0-0      tidygraph_1.2.0
 [59] WRS2_1.1-0                tibble_3.0.4
 [61] expm_0.999-5              Brobdingnag_1.2-6
 [63] ggrepel_0.8.2             biovizBase_1.34.1
 [65] png_0.1-7                 reshape_0.8.8
```

```
 [67] zeallot_0.1.0             ez_4.4-0
 [69] showtextdb_3.0            withr_2.3.0
 [71] rcompanion_2.3.26         ggforce_0.3.2
 [73] bitops_1.0-7              RBGL_1.62.1
 [75] plyr_1.8.6                cellranger_1.1.0
 [77] AnnotationFilter_1.10.0   e1071_1.7-4
 [79] coda_0.19-4               RcppParallel_5.0.2
 [81] pillar_1.4.6              Rmpfr_0.8-1
 [83] multcomp_1.4-15           europepmc_0.4
 [85] paletteer_1.2.0.9000      vctrs_0.3.4
 [87] ellipsis_0.3.1            generics_0.0.2
 [89] nortest_1.0-4             urltools_1.7.3
 [91] tools_3.6.0               foreign_0.8-71
 [93] tweenr_1.0.1              munsell_0.5.0
 [95] fgsea_1.12.0              fastmap_1.0.1
 [97] compiler_3.6.0            abind_1.4-5
 [99] httpuv_1.5.4              DescTools_0.99.39
[101] ggExtra_0.9               GenomeInfoDbData_1.2.2
[103] gridExtra_2.3             lattice_0.20-41
[105] later_1.1.0.1             prismatic_0.2.0
[107] BiocFileCache_1.10.2      jsonlite_1.7.1
[109] GGally_2.0.0              scales_1.1.1
[111] gld_2.6.2                 graph_1.64.0
[113] hrbrthemes_0.8.0          pbapply_1.4-3
[115] carData_3.0-4             genefilter_1.68.0
[117] lazyeval_0.2.2            promises_1.1.1
[119] car_3.0-10                BWStest_0.2.2
[121] tidyBF_0.4.0              latticeExtra_0.6-29
[123] metaBMA_0.6.5             effectsize_0.4.1
[125] pairwiseComparisons_3.1.1 checkmate_2.0.0
[127] rmarkdown_2.8             openxlsx_4.2.3
[129] sandwich_3.0-0            cowplot_1.1.0
[131] statmod_1.4.35            ipmisc_5.0.1
[133] forcats_0.5.0             dichromat_2.0-0
[135] logspline_2.1.16          igraph_1.2.6
[137] survival_3.2-7            numDeriv_2016.8-1.1
[139] yaml_2.2.1                systemfonts_0.3.2
[141] metaplus_0.7-11           rstantools_2.1.1
[143] htmltools_0.5.0           memoise_1.1.0
[145] VariantAnnotation_1.32.0  fastGHQuad_1.0
[147] modeltools_0.2-23         locfit_1.5-9.4
[149] graphlayouts_0.7.1        viridisLite_0.3.0
[151] gmp_0.6-1                 digest_0.6.27
[153] assertthat_0.2.1          mime_0.9
[155] rappdirs_0.3.1            Rttf2pt1_1.3.8
[157] bayestestR_0.8.0          RSQLite_2.2.1
[159] LaplacesDemon_16.1.4      Exact_2.1
```

```
[161] blob_1.2.1                splines_3.6.0
[163] Formula_1.2-4             labeling_0.4.2
[165] rematch2_2.1.2            showtext_0.9-2
[167] OrganismDbi_1.28.0        ProtGenerics_1.18.0
[169] RCurl_1.98-1.2            hms_0.5.3
[171] colorspace_1.4-1          base64enc_0.1-3
[173] BiocManager_1.30.10       libcoin_1.0-6
[175] nnet_7.3-14               Rcpp_1.0.5
[177] coin_1.3-1                mvtnorm_1.1-1
[179] enrichplot_1.6.1          multcompView_0.1-8
[181] fansi_0.4.1               R6_2.4.1
[183] ggridges_0.5.2            lifecycle_0.2.0
[185] EMT_1.1                   StanHeaders_2.21.0-6
[187] rootSolve_1.8.2.1         statsExpressions_0.6.1
[189] zip_2.1.1                 BayesFactor_0.9.12-4.2
[191] curl_4.3                  ggsignif_0.6.0
[193] minqa_1.2.4               DO.db_2.9
[195] Matrix_1.2-18             qvalue_2.18.0
[197] TH.data_1.0-10            stringr_1.4.0
[199] htmlwidgets_1.5.2         triebeard_0.3.0
[201] polyclip_1.10-0           biomaRt_2.42.1
[203] purrr_0.3.4               gridGraphics_0.5-0
[205] mgcv_1.8-33               openssl_1.4.3
[207] insight_0.11.0            lmom_2.8
[209] htmlTable_2.1.0           bdsmatrix_1.3-4
[211] codetools_0.2-16          GO.db_3.10.0
[213] prettyunits_1.1.1         dbplyr_1.4.4
[215] correlation_0.5.0         gtable_0.3.0
[217] DBI_1.1.0                 httr_1.4.2
[219] stringi_1.4.6             kSamples_1.2-9
[221] progress_1.2.2            farver_2.0.3
[223] viridis_0.5.1             annotate_1.64.0
[225] xml2_1.3.2                rvcheck_0.1.8
[227] bbmle_1.0.23.1            boot_1.3-25
[229] lme4_1.1-26               geneplotter_1.64.0
[231] ggplotify_0.0.5           bit_4.0.4
[233] jpeg_0.1-8.1              ggraph_2.0.4
[235] pkgconfig_2.0.3           knitr_1.30
```