DEG analysis using DESeq2 (Ribo-seq, uORF)

Toshihiro Arae

General directory setting

```
wd <- here::here()
shared <- fs::path(fs::path_dir(wd), "shared")
```

Loading packages

```
library(magrittr)
library(ggplot2)
```

Load common R scripts

```
#source(fs::path(wd, "script_r", "MISC.R"))
#source(fs::path(here::here(), "script_r", "MISC_PALETTE.R"))
```

Directory setting

```
dir_output <- fs::path("analysis", "deseq2_ribo_uorf")
path_out <- function(...) fs::path(wd, dir_output, ...)
fs::dir_create(path_out())
```

Loading input files

```
inf <- fs::path(wd, "data_preproc", "readcount", "count_ribo_uorf_psite", "count_by_gene.csv")
tbl_input <- readr::read_csv(inf, show_col_types = FALSE)

tbl_count <-
  tbl_input %>%
  dplyr::select(-Length, -dplyr::matches("^tpm_"))
```

Data pre-processing

Convert tibble to data.frame

```
rcdf <-
  tbl_count %>%
  ngsmisc::ds2_tbl_to_rcdf()
head(rcdf)
```

```
          zt0_1_ribo zt0_2_ribo zt12_1_ribo zt12_2_ribo zt18_1_ribo zt18_2_ribo
AT1G01060       1158        909           0           3          31          29
AT1G01210         15          7           3           5          21          15
AT1G01230         12          3          25          14          10           6
AT1G01240         27         21           9          10           9          11
AT1G01260         12         11          16          10          13           7
AT1G01370          1          1           2           4           1           0
          zt21_1_ribo zt21_2_ribo zt3_1_ribo zt3_2_ribo zt6_1_ribo zt6_2_ribo
AT1G01060         432        580         101         63          7         11
```

```
AT1G01210        8        9        7        4        5        5
AT1G01230        5        7       21       17       18       15
AT1G01240       19       20       32       22       24       16
AT1G01260       11        7       11        8       14       13
AT1G01370        2        0        5        2        2        3
```

```r
zt_lev <- paste0("zt", c(0, 3, 6, 12, 18, 21))
coldata <- data.frame(
  zt =
    colnames(rcdf) %>%
    stringr::str_extract("zt\\d+") %>%
    forcats::fct_relevel(zt_lev)
)

dds <-
  ngsmisc::ds2_rcdf_to_dds(
    rcdf = rcdf,
    coldata = coldata,
    design = ~ zt
  )

sf_default_ribo <-
  readRDS(fs::path(wd, "analysis", "deseq2_ribo", "sf_default_ribo.rds"))

dds <-
  dds %>%
  ngsmisc::ds2_dds_set_sizefactor(sf_default_ribo) %>%
  ngsmisc::ds2_dds_estimate_disp() %>%
  ngsmisc::ds2_dds_test_nbinomLRT()
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```r
ddr <- DESeq2::results(dds, alpha = .01, independentFiltering = FALSE)
ddr %>% DESeq2::summary()
```

```
out of 4778 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 486, 10%
LFC < 0 (down)     : 464, 9.7%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

```r
ddr@metadata$filterThreshold
```

```
NULL
```

```
ddr %>% ngsmisc::ds2_ddr_plot_independent_filtering()
```


number of rejections (y-axis)

quantiles of filter (x-axis)

```
tbl_out <-
  dds %>%
  ngsmisc::ds2_dds_to_tbl() %>%
  dplyr::rowwise() %>%
  dplyr::mutate(
    l2fc_amp =
      range(dplyr::c_across(zt_zt3_vs_zt0:zt_zt21_vs_zt0)) %>%
      {.[2] - .[1]}
  ) %>%
  dplyr::ungroup()
tbl_out %>% dplyr::filter(abs(l2fc_amp) >= 1) %>% dplyr::glimpse()
```

```
Rows: 3,419
Columns: 32
$ Geneid        <chr> "AT1G01060", "AT1G01210", "AT1G01230", "AT1G01240", …
$ baseMean      <dbl> 254.3188852, 8.2480697, 12.9148785, 18.7547461, 11.1…
$ baseVar       <dbl> 1.296034e+05, 1.876696e+01, 5.479936e+01, 7.473174e+…
$ allZero       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispGeneEst   <dbl> 0.009389061, 0.000000010, 0.000000010, 0.000000010, …
$ dispGeneIter  <dbl> 5, 31, 1, 1, 7, 2, 5, 4, 7, 32, 1, 16, 1, 5, 2, 1, 5…
$ dispFit       <dbl> 0.02346224, 0.24518855, 0.16238186, 0.11681076, 0.18…
$ dispersion    <dbl> 0.01964044, 0.17349947, 0.11988888, 0.08523613, 0.13…
$ dispIter      <int> 9, 9, 9, 9, 9, 9, 9, 10, 9, 9, 9, 7, 9, 9, 7, 9, 9, …
$ dispOutlier   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispMAP       <dbl> 0.01964044, 0.17349947, 0.11988888, 0.08523613, 0.13…
$ Intercept     <dbl> 9.8577232, 3.2433710, 2.6643408, 4.4196322, 3.372020…
$ zt_zt3_vs_zt0 <dbl> -3.40473784, -0.69063722, 1.69722337, 0.43648712, -0…
$ zt_zt6_vs_zt0 <dbl> -6.3253066, -0.5561038, 1.7495931, 0.2757544, 0.7501…
$ zt_zt12_vs_zt0 <dbl> -9.453640801, -1.376113207, 1.413896100, -1.32652875…
$ zt_zt18_vs_zt0 <dbl> -5.11422020, 0.76219159, 0.17131661, -1.26057730, -0…
$ zt_zt21_vs_zt0 <dbl> -0.98050370, -0.25234447, -0.18277325, -0.22771928, …
$ SE_Intercept  <dbl> 0.1465333, 0.5293893, 0.5218892, 0.3645109, 0.479569…
$ SE_zt_zt3_vs_zt0 <dbl> 0.2342737, 0.8086539, 0.6726997, 0.5107713, 0.692817…
$ SE_zt_zt6_vs_zt0 <dbl> 0.3971225, 0.8178920, 0.6782665, 0.5228094, 0.667808…
$ SE_zt_zt12_vs_zt0 <dbl> 0.8552808, 0.8457843, 0.6729154, 0.5749511, 0.671981…
$ SE_zt_zt18_vs_zt0 <dbl> 0.2767735, 0.7201881, 0.7261655, 0.5706237, 0.687848…
$ SE_zt_zt21_vs_zt0 <dbl> 0.2097477, 0.7639684, 0.7561215, 0.5244241, 0.695184…
```

```
$ LRTStatistic    <dbl> 1124.9996859, 8.6008579, 18.1421911, 17.4727015, 3.0…
$ LRTPvalue       <dbl> 5.153795e-241, 1.260834e-01, 2.773379e-03, 3.685305e…
$ fullBetaConv    <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
$ reducedBetaConv <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
$ betaIter        <dbl> 5, 3, 3, 3, 3, 6, 3, 8, 2, 3, 8, 10, 3, 4, 13, 5, 3,…
$ deviance        <dbl> 89.32033, 58.36542, 63.79390, 69.07509, 63.59909, 40…
$ maxCooks        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ padj            <dbl> 1.937312e-237, 3.213204e-01, 1.560648e-02, 1.984679e…
$ l2fc_amp        <dbl> 8.473137, 2.138305, 1.932366, 1.763016, 1.028085, 3.…
```

```r
coefs <-
  c("intercept", "zt3", "zt6", "zt12", "zt18", "zt21") %>%
  paste0("coef_", .)
ses <-
  c("intercept", "zt3", "zt6", "zt12", "zt18", "zt21") %>%
  paste0("se_", .)
colnames(tbl_out) <-
  c("AGI", "baseMean", "baseVar", "allZero",
    "dispGeneEst", "dispGeneIter", "dispFit",
    "dispersion", "dispIter", "dispOutlier", "dispMAP",
    coefs, ses, "stat_LRT", "pvalue", "fullBetaConv", "reducedBetaConv",
    "betaIter", "deviance", "maxCooks", "padj", "l2fc_amp")
readr::write_csv(tbl_out, path_out("deg_all.csv"))
```

Sessioninfo

```r
sessionInfo()
```

```
R version 4.2.1 (2022-06-23)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Ventura 13.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

other attached packages:
[1] ggplot2_3.4.2  magrittr_2.0.3

loaded via a namespace (and not attached):
 [1] MatrixGenerics_1.8.1      Biobase_2.56.0
 [3] httr_1.4.5                splines_4.2.1
 [5] bit64_4.0.5               vroom_1.6.0
 [7] jsonlite_1.8.4            here_1.0.1
 [9] BiocManager_1.30.18       stats4_4.2.1
[11] blob_1.2.3                renv_1.0.3
[13] GenomeInfoDbData_1.2.8    yaml_2.3.6
[15] pillar_1.9.0              RSQLite_2.2.18
[17] lattice_0.20-45           glue_1.6.2
[19] digest_0.6.31             RColorBrewer_1.1-3
[21] GenomicRanges_1.48.0      XVector_0.36.0
[23] colorspace_2.0-3          htmltools_0.5.3
[25] Matrix_1.6-4              DESeq2_1.36.0
```

```
[27] XML_3.99-0.11            pkgconfig_2.0.3
[29] genefilter_1.78.0        zlibbioc_1.42.0
[31] purrr_1.0.1              xtable_1.8-4
[33] scales_1.2.1             tzdb_0.3.0
[35] BiocParallel_1.30.4      tibble_3.2.1
[37] annotate_1.74.0          KEGGREST_1.36.3
[39] generics_0.1.3           IRanges_2.30.1
[41] cachem_1.0.6             withr_2.5.0
[43] SummarizedExperiment_1.26.1 BiocGenerics_0.42.0
[45] cli_3.6.0                survival_3.3-1
[47] crayon_1.5.2             memoise_2.0.1
[49] evaluate_0.20            fs_1.5.2
[51] fansi_1.0.3              forcats_1.0.0
[53] tools_4.2.1              hms_1.1.3
[55] lifecycle_1.0.3          matrixStats_0.62.0
[57] stringr_1.5.0            S4Vectors_0.34.0
[59] locfit_1.5-9.6           munsell_0.5.0
[61] DelayedArray_0.22.0      Biostrings_2.64.1
[63] AnnotationDbi_1.58.0     compiler_4.2.1
[65] GenomeInfoDb_1.32.4      rlang_1.1.0
[67] grid_4.2.1               RCurl_1.98-1.9
[69] rstudioapi_0.14          ngsmisc_0.4.0
[71] bitops_1.0-7             rmarkdown_2.24
[73] gtable_0.3.1             codetools_0.2-18
[75] DBI_1.1.3                R6_2.5.1
[77] knitr_1.42               dplyr_1.1.1
[79] fastmap_1.1.0            bit_4.0.5
[81] utf8_1.2.2               rprojroot_2.0.3
[83] readr_2.1.4              stringi_1.7.12
[85] parallel_4.2.1           Rcpp_1.0.11
[87] geneplotter_1.74.0       png_0.1-7
[89] vctrs_0.6.1              tidyselect_1.2.0
[91] xfun_0.40
```