DEG analysis using DESeq2 (Ribo-seq, mORF)

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")
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_central_cds_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
AT1G01010         59         31          82          56          65          45
AT1G01020         74         56          35          32          55          50
AT1G01030        125         61          47          50          63          57
AT1G01040        136         55         334          77         175         177
AT1G01050       1123        597         616         314        1233        1276
AT1G01060       4257       2356           0           2         122         132
          zt21_1_ribo zt21_2_ribo zt3_1_ribo zt3_2_ribo zt6_1_ribo zt6_2_ribo
```

```
AT1G01010          31          41         48         25         19         28
AT1G01020          46          56         43         40         18         35
AT1G01030          43          74         55         45         20         38
AT1G01040         123         139        160         54        154         90
AT1G01050         887        1111        414        312        314        263
AT1G01060        1618        2120        168        102         12         18
```

Prepare column data

```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)
)
```

LRT

```r
# Construct DESeq2::DESeqDataSet-class object
dds <-
  ngsmisc::ds2_rcdf_to_dds(
    rcdf = rcdf,
    coldata = coldata,
    design = ~ zt
  )

# Extract scaling factor to normalise read-count data
sf_default <- ngsmisc::ds2_dds_get_sizefactor(dds)
saveRDS(sf_default, path_out("sf_default_ribo.rds"))

# Test by the DESeq2::nbinomLRT() function
dds <-
  dds %>%
  ngsmisc::ds2_dds_set_sizefactor(sf_default) %>%
  ngsmisc::ds2_dds_estimate_disp() %>%
  ngsmisc::ds2_dds_test_nbinomLRT()
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```r
# Check the results
ddr <- DESeq2::results(dds, alpha = .01)
ddr %>% DESeq2::summary()
```

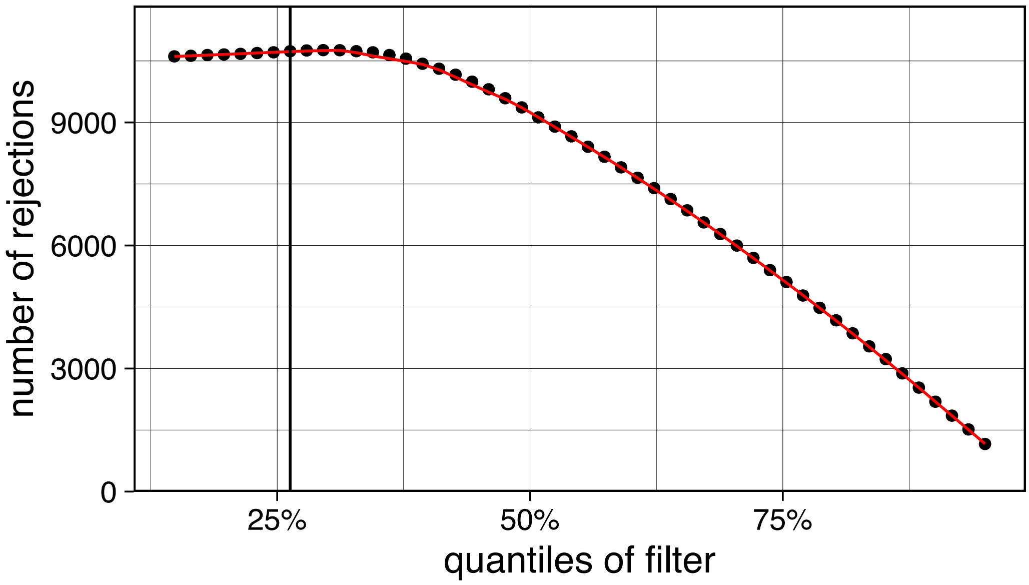
```

out of 23535 with nonzero total read count
adjusted p-value < 0.01
LFC > 0 (up)       : 5356, 23%
LFC < 0 (down)     : 5380, 23%
outliers [1]       : 0, 0%
low counts [2]     : 3165, 13%
(mean count < 1)
```

```
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

```r
# Independent filtering
ddr@metadata$filterThreshold
```

```
26.27765%
 1.440742
```

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



```r
# Extract data from DESeqDataSet-class object and write it to the csv file.
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: 13,163
Columns: 32
$ Geneid            <chr> "AT1G01010", "AT1G01050", "AT1G01060", "AT1G01070", …
$ baseMean          <dbl> 4.172239e+01, 6.590611e+02, 8.127792e+02, 1.398334e+…
$ baseVar           <dbl> 1.450791e+02, 9.568168e+04, 1.333375e+06, 7.761707e+…
$ allZero           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispGeneEst       <dbl> 0.000000010, 0.010398680, 0.006560301, 0.006383818, …
$ dispGeneIter      <dbl> 1, 4, 5, 23, 31, 4, 29, 1, 31, 5, 29, 34, 18, 30, 4,…
$ dispFit           <dbl> 0.06678309, 0.01773848, 0.01711159, 0.03004638, 0.38…
$ dispersion        <dbl> 0.041167649, 0.014156165, 0.012398164, 0.021714659, …
$ dispIter          <dbl> 7, 5, 11, 11, 7, 10, 7, 11, 10, 7, 8, 11, 8, 11, 11,…
$ dispOutlier       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispMAP           <dbl> 0.041167649, 0.014156165, 0.012398164, 0.021714659, …
$ Intercept         <dbl> 5.287117, 9.536073, 11.483244, 8.156520, 2.296236, 7…
$ zt_zt3_vs_zt0     <dbl> -0.01272721, -0.92368964, -4.31293835, -1.58757170, …
$ zt_zt6_vs_zt0     <dbl> -0.372868914, -0.993101074, -7.213721727, -2.5218906…
$ zt_zt12_vs_zt0    <dbl> 0.63515270, -0.89799044, -11.67018981, -2.40246690, …
$ zt_zt18_vs_zt0    <dbl> 0.329877795, 0.593716013, -4.657615861, -0.791865078…
$ zt_zt21_vs_zt0    <dbl> -0.22011588, 0.32666295, -0.71976503, -0.57452122, -…
$ SE_Intercept      <dbl> 0.25887493, 0.12657231, 0.11505461, 0.16106784, 0.64…
$ SE_zt_zt3_vs_zt0  <dbl> 0.3731092, 0.1834988, 0.1844778, 0.2460151, 0.907447…
$ SE_zt_zt6_vs_zt0  <dbl> 0.3928358, 0.1853350, 0.3092105, 0.2748170, 0.879119…
$ SE_zt_zt12_vs_zt0 <dbl> 0.3541161, 0.1820212, 1.0315639, 0.2566482, 0.917498…
$ SE_zt_zt18_vs_zt0 <dbl> 0.3588798, 0.1777136, 0.1852925, 0.2327656, 0.911796…
$ SE_zt_zt21_vs_zt0 <dbl> 0.3728172, 0.1783387, 0.1634128, 0.2316199, 0.924428…
$ LRTStatistic      <dbl> 10.3433921, 152.0159328, 2365.0146161, 155.6008864, …
$ LRTPvalue         <dbl> 6.606993e-02, 4.969975e-31, 0.000000e+00, 8.568006e-…
$ fullBetaConv      <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
$ reducedBetaConv   <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE…
$ betaIter          <dbl> 3, 2, 4, 3, 3, 3, 3, 2, 2, 3, 17, 3, 2, 2, 3, 3, 14,…
$ deviance          <dbl> 81.46520, 131.13599, 101.58850, 99.63640, 54.80464, …
$ maxCooks          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ padj              <dbl> 1.051917e-01, 7.168221e-30, 0.000000e+00, 1.287258e-…
$ l2fc_amp          <dbl> 1.008022, 1.586817, 10.950425, 1.947369, 1.159886, 1…
```

```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"))

# Output AGI code analysed after the Independent filtering
AGI_filtered_ribo <- tbl_out$AGI[!is.na(tbl_out$padj)]
AGI_filtered_ribo %>% saveRDS(path_out("AGI_filtered_ribo.rds"))
```

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] farver_2.1.1                generics_0.1.3
[41] IRanges_2.30.1              cachem_1.0.6
[43] withr_2.5.0                 SummarizedExperiment_1.26.1
[45] BiocGenerics_0.42.0         cli_3.6.0
[47] survival_3.3-1              crayon_1.5.2
[49] memoise_2.0.1               evaluate_0.20
[51] fs_1.5.2                    fansi_1.0.3
[53] forcats_1.0.0               tools_4.2.1
[55] hms_1.1.3                   lifecycle_1.0.3
[57] matrixStats_0.62.0          stringr_1.5.0
[59] S4Vectors_0.34.0            locfit_1.5-9.6
[61] munsell_0.5.0               DelayedArray_0.22.0
[63] Biostrings_2.64.1           AnnotationDbi_1.58.0
[65] compiler_4.2.1              GenomeInfoDb_1.32.4
[67] rlang_1.1.0                 grid_4.2.1
[69] RCurl_1.98-1.9              rstudioapi_0.14
[71] ngsmisc_0.4.0               labeling_0.4.2
[73] bitops_1.0-7                rmarkdown_2.24
[75] gtable_0.3.1                codetools_0.2-18
[77] DBI_1.1.3                   R6_2.5.1
[79] knitr_1.42                  dplyr_1.1.1
[81] fastmap_1.1.0               bit_4.0.5
[83] utf8_1.2.2                  rprojroot_2.0.3
[85] readr_2.1.4                 stringi_1.7.12
[87] parallel_4.2.1              Rcpp_1.0.11
[89] geneplotter_1.74.0          png_0.1-7
[91] vctrs_0.6.1                 tidyselect_1.2.0
[93] xfun_0.40
```