DEG analysis using DESeq2 (RNA-seq)

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

Loading input files

```
inf <- fs::path(wd, "data_preproc", "readcount", "count_rna_exon", "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_rna zt0_2_rna zt12_1_rna zt12_2_rna zt18_1_rna zt18_2_rna
AT1G01010        20        30         44         43         35         40
AT1G01020        81        82         81         66         91         88
AT1G03987         7         7          0          1          4          2
AT1G01030       167       205         55         62        121        119
AT1G01040       566       633        384        396        738        734
AT1G03993        87       111         68         68        130        134
          zt21_1_rna zt21_2_rna zt3_1_rna zt3_2_rna zt6_1_rna zt6_2_rna
AT1G01010         39         23        31        27        32        22
```

```
AT1G01020        109        109        72        80        66        79
AT1G03987          1          7         9         5         3         6
AT1G01030         88         87       114       121        93        81
AT1G01040        620        606       489       532       454       439
AT1G03993        105         84        89       100        73        63
```

Prepare column data

```
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

```
# 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_rna.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
```

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

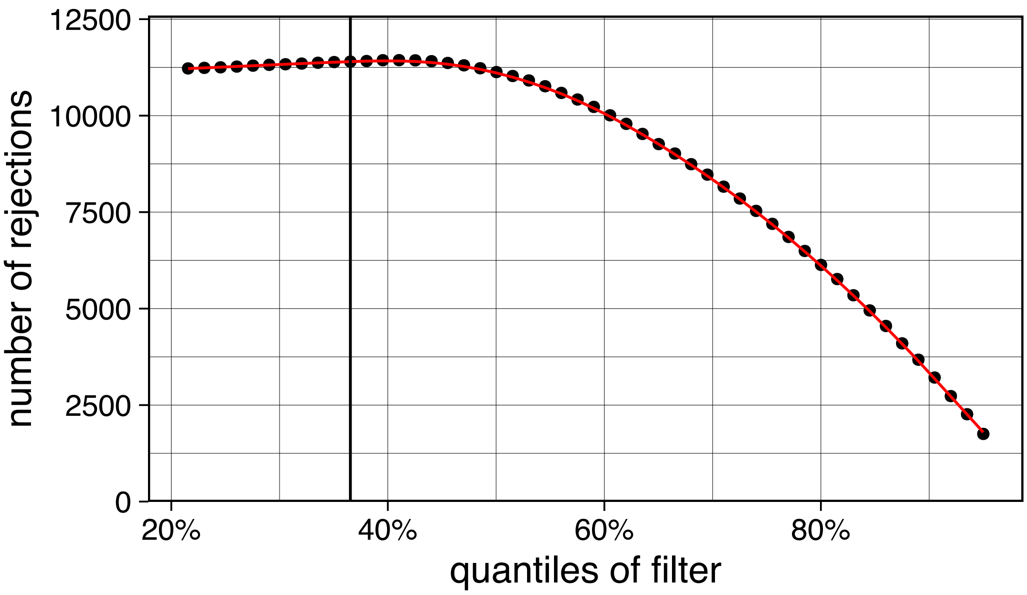
```

out of 28961 with nonzero total read count
adjusted p-value < 0.01
LFC > 0 (up)       : 5636, 19%
LFC < 0 (down)     : 5764, 20%
outliers [1]       : 0, 0%
low counts [2]     : 5534, 19%
(mean count < 1)
```

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

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

```
36.54063%
 1.429563
```

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



```
# 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: 14,840
Columns: 32
$ Geneid             <chr> "AT1G03987", "AT1G01030", "AT1G01040", "AT1G03993", …
$ baseMean           <dbl> 4.454530e+00, 1.107441e+02, 5.502269e+02, 9.254411e+…
$ baseVar            <dbl> 9.506573e+00, 2.215501e+03, 1.594485e+04, 5.495213e+…
$ allZero            <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispGeneEst        <dbl> 0.00000001, 0.00000001, 0.00000001, 0.00000001, 0.00…
$ dispGeneIter       <dbl> 1, 15, 2, 1, 10, 4, 27, 4, 24, 15, 21, 8, 32, 10, 1,…
$ dispFit            <dbl> 0.481918224, 0.021089472, 0.005663564, 0.024887637, …
$ dispersion         <dbl> 0.390547005, 0.014402018, 0.003769101, 0.018127449, …
$ dispIter           <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 9, 9, 9, 9, 9, 9, 9, …
$ dispOutlier        <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FAL…
$ dispMAP           <dbl> 0.390547005, 0.014402018, 0.003769101, 0.018127449, …
$ Intercept         <dbl> 2.9062791, 7.6322722, 9.3233630, 6.7218400, 8.763140…
$ zt_zt3_vs_zt0     <dbl> -0.08329908, -0.73898621, -0.31078112, -0.14273780, …
$ zt_zt6_vs_zt0     <dbl> -0.64890363, -1.10622565, -0.43724847, -0.55131358, …
$ zt_zt12_vs_zt0    <dbl> -4.0939117, -1.9504505, -0.9048266, -0.8233145, -0.4…
$ zt_zt18_vs_zt0    <dbl> -1.4007260, -0.8025715, 0.1230619, 0.2455744, 0.2500…
$ zt_zt21_vs_zt0    <dbl> -0.87184776, -1.13903564, -0.02148886, -0.11645286, …
$ SE_Intercept      <dbl> 0.74505188, 0.14354634, 0.07524626, 0.17152605, 0.08…
$ SE_zt_zt3_vs_zt0  <dbl> 1.05371452, 0.21083125, 0.10780912, 0.24351029, 0.12…
$ SE_zt_zt6_vs_zt0  <dbl> 1.0918847, 0.2180846, 0.1091581, 0.2521860, 0.126994…
$ SE_zt_zt12_vs_zt0 <dbl> 1.7410252, 0.2310367, 0.1106915, 0.2521712, 0.121877…
$ SE_zt_zt18_vs_zt0 <dbl> 1.14410585, 0.21039553, 0.10487346, 0.23700078, 0.11…
$ SE_zt_zt21_vs_zt0 <dbl> 1.10592612, 0.21791422, 0.10621667, 0.24349504, 0.11…
$ LRTStatistic      <dbl> 9.0865803, 80.9487830, 115.9936426, 24.1201404, 178.…
$ LRTPvalue         <dbl> 1.056602e-01, 5.312793e-16, 2.212859e-23, 2.058696e-…
$ 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, 2, 2, 2, 2, 3, 2, 3, 3, 4, 3, 2, 3, 3, 2, 6, 2, 2…
$ deviance          <dbl> 51.032277, 89.629214, 111.987250, 89.703534, 106.606…
$ maxCooks          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
$ padj              <dbl> 1.704910e-01, 2.855998e-15, 1.583332e-22, 5.215465e-…
$ l2fc_amp          <dbl> 4.010613, 1.211464, 1.027888, 1.068889, 1.250259, 2.…
```

```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_rna <- tbl_out$AGI[!is.na(tbl_out$padj)]
AGI_filtered_rna %>% saveRDS(path_out("AGI_filtered_rna.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
```