wind: wORKFLOW FOR PiRNAs AnD BEYONd

Computational workflow for Differential Expression analysis of selected samples from the TCGA regarding Breast Cancer

Constantinos Yeles (Konstantinos Geles)

Thu Oct 29 2020

Contents

Introduction

Following the data_exploration_salmon_fc worklow in most cases we want to perform differential expression (DE) analysis. We follow instructions from various packages utilized for DE with the objects resulted from the previous workflow.

Load libraries

```
suppressPackageStartupMessages({
  library('tidyverse')
  library('edgeR')
  library('DESeq2')
})
```

Add date of the analysis

We use it as an identifier for the folder and generally the analysis

```
todate <- format(Sys.time(), "%d_%b_%Y")
```

Make the directory for the results of the DE analysis

```
my_basename <- "Breast_Cancer_TCGA"
my_exp <- "BRCA"
my_exp_sal <- "salmon"
my_exp_fc <- "featureCounts"
dat_path <-  str_glue("{my_basename}/DEA_{my_exp}_{todate}/{c(my_exp_sal,my_exp_fc)}") %>%
  as.list %>%
  set_names("salmon","featureCounts")
dat_path %>% map(~dir.create(str_glue("./{.x}"), recursive = TRUE))
```

1. Import the normalized files

```
list_norm_dgls <- list.files(path = my_basename, pattern = "list_norm_dgls.+rds",
  recursive = TRUE, full.names = TRUE)

# load salmon normalized files
salmon_norm <- list_norm_dgls %>%
  unlist %>%
  str_detect("salmon") %>%
  list_norm_dgls[.] %>%
  read_rds()
# load featurecounts normalized files
fc_norm <- list_norm_dgls %>%
  unlist %>%
  str_detect("featureCounts") %>%
  list_norm_dgls[.] %>%
  read_rds()
```

2. Extract normalized objects

We will work with TMM normalization and TMM voom with quality weights

```
salmon_edgR_TMM <- salmon_norm[["TMM"]]
salmon_vm_QW_TMM <- salmon_norm[["voomQW_TMM"]]
fc_edgR_TMM <- fc_norm[["TMM"]]
fc_vm_QW_TMM <- fc_norm[["voomQW_TMM"]]
```

3. Create the design matrix

If we load the voom object we can extract the design matrix otherwise we can create it again from the dgl object.

```
#1 voom object
design <- salmon_vm_QW_TMM$design
#or dgl object
targets <- salmon_edgR_TMM$samples
design <- model.matrix(~0 + targets$group ,data = targets)
colnames(design) <- colnames(design) %>%
  str_remove(".+\\$group")
```

4. EdgeR

Perform the analysis with edgeR TMM normalization for both salmon and featurecounts

```
# design ----
con_mat <- makeContrasts(
  tumor_vs_normal = Primary_Solid_Tumor - Solid_Tissue_Normal,
  levels = design)

## salmon ----
salmon_edgR_TMM <- estimateDisp(salmon_edgR_TMM, design = design, robust=TRUE)
salmon_edgR_TMM <- glmQLFit(salmon_edgR_TMM, design, robust = TRUE)

DE_salmon_edgR <- con_mat %>%
  colnames() %>%
  set_names() %>%
 map(~glmQLFTest(salmon_edgR_TMM, contrast = con_mat[,.x]) %>%
  topTags(n = nrow(.), adjust.method = "BH", sort.by = "PValue", p.value = 1) %>%
   .$table %>%
  as_tibble(rownames = "smallRNA") %>%
  write_tsv(str_glue("{dat_path[['salmon']]}/DE_salmon_edgR_TMM_{.x}.txt"))
 )

hist(DE_salmon_edgR[[1]]$PValue, breaks = 0:20/20,
     col = "grey50", border = "white")

salmon_edgeR_TMM_p <- DE_salmon_edgR[[1]] %>%
  mutate(salmon_edgeR = if_else(
    FDR >= 0.05, 0, if_else(
      logFC > 0, 1, -1
    )
  )) %>%
  select(smallRNA , salmon_edgeR )
```

```
## featureCounts ----
fc_edgR_TMM <- estimateDisp(fc_edgR_TMM, design = design, robust=TRUE)
fc_edgR_TMM <- glmQLFit(fc_edgR_TMM, design, robust = TRUE)

DE_FC_edgR <- con_mat %>% colnames() %>% set_names() %>%
 map(~glmQLFTest(fc_edgR_TMM, contrast = con_mat[,.x]) %>%
  topTags(n = nrow(.), adjust.method = "BH", sort.by = "PValue", p.value = 1) %>%
   .$table %>%
  as_tibble(rownames = "smallRNA") %>%
  write_tsv(str_glue("{dat_path[['featureCounts']]}/DE_fc_edgR_TMM_{.x}.txt"))
 )

hist(DE_FC_edgR[[1]]$PValue, breaks = 0:20/20,
     col = "grey50", border = "white")

fc_edgeR_TMM_p <- DE_FC_edgR[[1]] %>%
  mutate(fc_edgeR = if_else(
    FDR >= 0.05, 0, if_else(
      logFC > 0, 1, -1
    )
  )) %>%
  select(smallRNA , fc_edgeR )

# venn diagram for salmon/fc edgeR -----
results <-  salmon_edgeR_TMM_p %>%
  inner_join(fc_edgeR_TMM_p) %>%  select(-smallRNA)

pdf(str_glue("{dat_path}/venn_diagram_DE_salmon_fC_edgeR_miR_450a_e.pdf"))
vennDiagram(results,
    include=c("up", "down"),
    counts.col=c("red", "blue"),
    circle.col = c("red", "blue", "green3"))
dev.off()
```

5. Limma

```
# design ----

## salmon ----
salmon_vm_QW_TMM <- lmFit(salmon_vm_QW_TMM, design = design)
salmon_vm_QW_TMM <- contrasts.fit(salmon_vm_QW_TMM, con_mat)
salmon_vm_QW_TMM <- eBayes(salmon_vm_QW_TMM, robust = TRUE)

salmon_DES <- con_mat %>% colnames() %>% set_names() %>%
 map(~salmon_vm_QW_TMM %>% topTable(., coef = .x,
                            confint = TRUE,
                            number = nrow(.),
                            adjust.method = "fdr",
                            sort.by = "p") %>%
  as_tibble(rownames = "smallRNA") %>%
  rename_at(vars(logFC:B), list(~str_c(.,"_",!!quo(.x)))) %>%
```

```
  write_tsv(str_glue("{dat_path[['salmon']]}/DE_salmon_vm_QW_TMM_{.x}.txt"))
 )

hist(salmon_DES[[1]] %>% select(starts_with("P.Value")) %>% deframe(),
     breaks = 0:20/20,
     col = "grey50", border = "white")

salmon_vm_QW_TMM_p <- salmon_DES[[1]] %>%
  mutate(salmon_voomQ = if_else(
    adj.P.Val_tumor_vs_normal >= 0.05, 0, if_else(
      logFC_tumor_vs_normal > 0, 1, -1
    )
  )) %>%
  select(smallRNA , salmon_voomQ )

## featureCounts ----
fc_vm_QW_TMM <- lmFit(fc_vm_QW_TMM, design = design)
fc_vm_QW_TMM <- contrasts.fit(fc_vm_QW_TMM, con_mat)
fc_vm_QW_TMM <- eBayes(fc_vm_QW_TMM, robust = TRUE)

fc_DES <- con_mat %>% colnames() %>% set_names() %>%
 map(~fc_vm_QW_TMM %>% topTable(., coef = .x,
                                confint = TRUE,
                                number = nrow(.),
                                adjust.method = "fdr",
                                sort.by = "p") %>%
  as_tibble(rownames = "smallRNA") %>%
  rename_at(vars(logFC:B), list(~str_c(.,"_",!!quo(.x)))) %>%
  write_tsv(str_glue("{dat_path[['featureCounts']]}/DE_fc_vm_QW_TMM_{.x}.txt"))
 )

hist(fc_DES[[1]] %>% select(starts_with("P.Value")) %>% deframe(),
     breaks = 0:20/20,
     col = "grey50", border = "white")

fc_vm_QW_TMM_p <- fc_DES[[1]] %>%
  mutate(fc_voomQ = if_else(
    adj.P.Val_tumor_vs_normal >= 0.05, 0, if_else(
      logFC_tumor_vs_normal > 0, 1, -1
    )
  )) %>%
  select(smallRNA , fc_voomQ )

# venn diagram for salmon/fc limma -----
nc_RNA_categories <- plyranges::read_gff2("../genome_transc_human/ncRNA_transcripts_100bp_RNA_Central_p
  as_tibble() %>%
  select(gene_id, gene_type) %>%
  distinct(gene_id, .keep_all = TRUE)

results <-  salmon_vm_QW_TMM_p %>%
  inner_join(fc_vm_QW_TMM_p) %>%  select(-smallRNA)

pdf(str_glue("{dat_path}/venn_diagram_DE_salmon_fC_limma__miR_450a_e.pdf"))
```

```
vennDiagram(results,
    include=c("up", "down"),
    counts.col=c("red", "blue"),
    circle.col = c("red", "blue", "green3"))
dev.off()
# join both results ----
identical(fc_DES %>% names(), salmon_DES %>% names)

map2(fc_DES, salmon_DES, ~.x %>%
       select_at(vars(starts_with(c("smallRNA","logFC",
                                    "P.Value","adj.P.Val")))) %>%
       rename_at(vars(!matches("smallRNA")), list(~str_c(.,"_FC"))) %>%
       full_join(.y%>%
       select_at(vars(starts_with(c("smallRNA","logFC",
                                    "P.Value","adj.P.Val")))) %>%
       rename_at(vars(!matches("smallRNA")), list(~str_c(.,"_salmon"))))
     ) %>%
  purrr::reduce(full_join) %>%
  inner_join(nc_RNA_categories, by = c("smallRNA" = "gene_id")) %>%
  write_tsv(str_glue("{str_remove(dat_path[1],'/salmon|/featureCounts')}/all_comparisons_voom_QW_TMM_sa
```

6. Compare the DE results with public results

```
#create the file with names and rnacentral ids -----
all_comp <- read_tsv(str_glue("{str_remove(dat_path[1],'/salmon|/featureCounts')}/all_comparisons_voom_

smallRNAs_gtf <- plyranges::read_gff2("../genome_transc_human/ncRNA_transcripts_100bp_RNA_Central_piRNA
  as_tibble() %>%
  select(gene_id, gene_type) %>%
  distinct(gene_id, .keep_all = TRUE) %>%
  dplyr::rename("smallRNA" = gene_id)

all_comp <- all_comp %>%
  left_join(smallRNAs_gtf) %>%
  separate(col = smallRNA,
           into = c("RNAcentral_id", "GR"),
           sep = "_GR_")

RNACentral_ids <- data.table::fread("../genome_transc_human/rnacentral_ids_hg38.txt")

smallRNAs_gtf %>%
  as_tibble() %>%
  left_join(RNACentral_ids) %>%
  mutate( is_correct = if_else(RNA_type == gene_type,true = T,F)) %>%
  filter(!is.na(GR), is.na(is_correct))

all_comp %>%
  left_join(RNACentral_ids) %>%
   write_tsv(str_glue("{str_remove(dat_path[1],'/salmon|/featureCounts')}/all_comparisons_with_gene_nam

# cross the results created with the public results -----
```

```
all_comp <- "GSE129076/DEA_br_cancer_12_Sep_2020/all_comparisons_with_gene_names.txt" %>% vroom
public_res <- "GSE129076/Public_results_GSE129076.csv" %>%
  vroom

rnacentral_ids <- vroom::vroom("../genome_transc_human/id_mapping.tsv.gz",
                         col_names = c("RNAcentral_id",
                                       "Database",
                                       "external_id",
                                       "NCBI_taxon_id",
                                       "RNA_type",
                                       "gene_name"),
                         delim = "\t") %>%
  filter(NCBI_taxon_id == "9606") %>%
  mutate(gene_name = tolower(gene_name))

public_res %>%
  left_join(rnacentral_ids, by = c("smallRNA" = "gene_name")) %>%
  distinct(RNAcentral_id, .keep_all = T) %>%
  right_join(all_comp) %>%
  mutate(across(.cols = starts_with("logFC_"),
                gtools::logratio2foldchange,
                .names = "{str_remove({col}, pattern='log')}")) %>%
  vroom_write("GSE129076/all_comparisons_public_res.txt")

rnacentral_ids %>%
  distinct(RNAcentral_id, gene_name,.keep_all = T) %>%
  right_join(all_comp) %>%
  mutate(across(.cols = starts_with("logFC_"),
                gtools::logratio2foldchange,
                .names = "{str_remove({col}, pattern='log')}")) %>%
  vroom::vroom_write("GSE129076/all_comparisons_public_names.txt")

res_ids <- public_res$smallRNA %>% str_c(collapse = "|")

all_comp %>% mutate(across(.cols = starts_with("logFC_"),gtools::logratio2foldchange,.names = "{str_reme
```

7. Find predicted targets

```
suppressPackageStartupMessages(library(plyranges))
# load gtf of smallRNAs
smallRNAs_gtf <- read_gff2("../genome_transc_human/ncRNA_transcripts_100bp_RNA_Central_piRNAbank_hg38.g
  as_tibble() %>%
  select(gene_id, gene_type) %>%
  distinct(gene_id, .keep_all = TRUE) %>%
  filter(gene_type == "piRNA")
#load targets
targets_all <- read_tsv("../genome_transc_human/human_data/piRNA_predicted_Targets.txt")

# targets DE union
targets_DEs_keep <- all_comp %>% unite(piRNA_id, c(RNAcentral_id, GR),sep = "_GR_") %>% filter(gene_type
```

```
  write_tsv("Breast_Cancer_TCGA/DEA_BRCA_06_Oct_2020/DE_targets_predicted.txt")
```

8. Make a heatmap of differential expressed piRNAs

i. Load the libraries

```
library(wesanderson)
library(ComplexHeatmap)
library(circlize)
```

ii. Load data

```
# load the piRNAs log fold changes
piRNAs_DE <- list.files(pattern = "all_comparisons_voom_QW_TMM") %>%
  read_tsv() %>%
  filter(gene_type == "piRNA",
         across(.cols = contains("adj.P.Val"),
                .fns = ~ .x< 0.05)) %>%
  dplyr::select(smallRNA, contains("logFC"))


# load the piRNA expression matrix----
## featurecounts
fc_list <- "list_norm_dgls_featureCounts.rds" %>%
  read_rds() %>%
  .[["TMM"]]

fc_cpm <- fc_list  %>%
  cpm(log =TRUE) %>%
  .[rownames(.) %in% piRNAs_DE$smallRNA,]

## salmon
salmon_list <- "list_norm_dgls_salmon.rds" %>%
  read_rds() %>%
  .[["TMM"]]

salmon_cpm <- salmon_list  %>%
  cpm(log =TRUE) %>%
  .[rownames(.) %in% piRNAs_DE$smallRNA,]

# make the matrices for the heatmap -----
FC_mat_1 <- fc_cpm %>%
  t() %>% scale() %>% t()
FC_mat_1 %>% dim()
FC_mat_1 %>% head()
hist(FC_mat_1)

salmon_mat_1 <- salmon_cpm %>%
  t() %>% scale() %>% t()
```

```r
salmon_mat_1 %>% dim()
salmon_mat_1 %>% head()
hist(salmon_mat_1)

# logFCS
lfc_piRNAs_DE <- piRNAs_DE %>%
  column_to_rownames("smallRNA") %>%
  as.matrix()

lfc_piRNAs_DE %>% dim()
lfc_piRNAs_DE %>% head()
hist(lfc_piRNAs_DE)

lfc_piRNAs_DE <- lfc_piRNAs_DE[rownames(FC_mat_1) ,]

colnames(FC_mat_1) <- colnames(FC_mat_1) %>% str_remove("-13_mirna_gdc_realn")

# add the Annotation ----
#expression
ha_1 <- HeatmapAnnotation(Group = fc_list$samples$group,
                          annotation_name_side = "left",
                          col = list(Group = fc_list$colours %>%
                                       set_names(fc_list$samples$group)))

# lFCS
ha_1_LFCs <- HeatmapAnnotation(Method = c("FeatureCounts", "salmon"),
                               col = list(Method = wes_palettes$Moonrise2[c(1,4)] %>%
                          set_names("FeatureCounts", "salmon")))

## Colours of heatmap -----
#expression
f_1 <- colorRamp2(c(round(quantile(FC_mat_1, probs = 0.25)),
                    median(FC_mat_1),
                    round(quantile(FC_mat_1, probs = 0.75))),
                  c("blue", "black", "yellow"))

# lFCS
f_1_LFCs <- colorRamp2(c(quantile(lfc_piRNAs_DE, probs = 0.25) %>% round,
                    mean(lfc_piRNAs_DE),
                    quantile(lfc_piRNAs_DE, probs = 0.75) %>% round),
                    c("forestgreen", "black", "red"))

## Heatmaps ------
ht_1 <- Heatmap(matrix = FC_mat_1, #data
        top_annotation = ha_1, #annot
        col = f_1, #colors data
        show_row_dend = TRUE,
        show_row_names = FALSE,
        show_column_names = FALSE,
        name = "z-score equivalent expression",
        clustering_distance_columns = "spearman",
        clustering_method_columns =  "ward.D2",
        clustering_method_rows = "ward.D2",
```

```
        clustering_distance_rows = "spearman",
        row_dend_reorder = TRUE
)
rownames(lfc_piRNAs_DE) <- lfc_piRNAs_DE %>%
  rownames() %>%
  str_remove("_GR_.+")
ht_1_lFCs <- Heatmap(matrix = lfc_piRNAs_DE, #data
        top_annotation = ha_1_LFCs, #annot
        col = f_1_LFCs, #colors data
        show_row_dend = FALSE,
        show_row_names = TRUE,
        show_column_names = FALSE,
        name = "Log Fold Change",
        clustering_distance_columns = "spearman",
        clustering_method_columns =  "ward.D2",
        clustering_method_rows = "ward.D2",
        clustering_distance_rows = "spearman",
        row_dend_reorder = TRUE
)

draw(ht_1+ht_1_lFCs,column_title = str_glue("Heatmap of {nrow(FC_mat_1)} DE piRNAs"),
     merge_legend = TRUE)

tiff("BRCA_tumour_vs_normal_heatmap_FC.tiff",
     compression = "none", height = 10, width = 14,  units = 'in', res = 300)
draw(ht_1+ht_1_lFCs,
     column_title = str_glue("Heatmap of {nrow(FC_mat_1)} DE piRNAs"),
     merge_legend = TRUE)
dev.off()
```