Vignette___Pathway_Enrichment_Analysis

Christina Hillig

3/15/2021

Introduction

The following section describes the pathway enrichment analysis of up-regulated genes in the IL17A transcript positive and negative group in Leukocytes in the ST data set.

Initialisation

Load required R and Bioconductor packages:

```
# R packages
rlibs <- c("dplyr", "gtools", "hash", "kableExtra", "knitr",
           "stringr", "tibble", "xlsx", "hash", "Hmisc")
invisible(lapply(rlibs, require, character.only = TRUE))
# Bioconductor packages
bioclibs <- c("ReactomePA", "pathview",  "enrichplot", "org.Hs.eg.db", "DOSE")
invisible(lapply(bioclibs, require, character.only = TRUE))

getwd()
```

```
## [1] "/Users/christina.hillig/PycharmProjects/ST_Immune_publication/Publication_analysis/r_scripts/pa
```

```
environment()
```

```
## <environment: R_GlobalEnv>
```

```
# Source R-scripts
source(file.path("..", "..", "r_scripts", "pathway_analysis", 'init.R'))
source(file.path("..", "..", "r_scripts", "pathway_analysis", 'load_data.R'))
source(file.path("..", "..", "r_scripts", "pathway_analysis",'utils.R'))
# R-script to import immune Pathways which shall be highlighted in the publication
source(file.path("..", "..", "r_scripts", "pathway_analysis", 'pathway_lists.R'))
# R-script to plot the Pathway enrichment result
source(file.path("..", "..", "r_scripts", "pathway_analysis", "plot_pathways.R"))
```

Define input directory.

```
## Input
# Input directory
# TODO change to relative path
input.folder <- file.path("..", "..", "input", "pathway_analysis")
# Date
date <- "2021-02-01"
# Data set
dataset.type <- 'Leukocytes'
```

```
# Sequencing technique
seq.technique <- "ST"
# Comparison
cytokine <- 'IL17A'
comparison <- 'IL17A_vs_Others'
plot_cytokine <- TRUE

# Used design function:
design.function <- "cdr_patient_annotation_cyto"

# General input directory
input.dir <- get_inputdir(input.folder = input.folder, date.file = date,
                         dataset.type = dataset.type, seq.technique = seq.technique,
                         comparison = comparison, design.function = design.function,
                         genename=cytokine)

# print(input.dir)
```

Determine cut-parameters to identify differential expressed genes and enriched pathways.

```
# Cut parameter
l2fc.factor <- 1
fdr.value <- 0.05
pval.cut <- 0.05
p.value <- 0.05
minGSSize <- 10
# Multi-test method Benjamini-Hochberg (BH)
multitest.method <-  "BH"
```

```
# Plot Parameters
show_dotplot_categories <- 15

width_img <- 16
height_img <- 8
```

Create output directory.

```
## Output
output.folder <- file.path("..", "..")
# General output directory
output.dir <- get_outputdir(output.folder = output.folder, dataset.type = dataset.type,
                            seq.technique = seq.technique, genename=cytokine)
# print(output.dir)
```

Load DGE analysis results and prepare dataframe for further analysis steps.

```
# 1. Get all DGE .csv files in subfolders
all_filenames <- list.files(path = input.dir, pattern = ("*.csv|*.xlsx"), recursive = TRUE)
# 2. remove metaData from list
dge_filename <- all_filenames[!grepl("metaData*", all_filenames)]
# 3. Load DGE Analysis result file with colnames:
# "X" "gene_symbol" "gene_name" "entrezid" "pval" "padj" "log2fc" "hkg"
df.dge_res <- load_files(path_name_file = file.path(input.dir,  dge_filename))

# 4. Filter data for duplicates, unwanted columns
df.dge_res <- filter_data(df.data = df.dge_res, signature_gene = cytokine,
```

```r
                          plot_signaturecytokine=TRUE)
print.data.frame(df.dge_res[1:10, ], digits = 4)
```

```
##        gene_symbol                                         gene_name entrezid
## IL17A        IL17A                                   interleukin 17A     3605
## ACP7          ACP7 acid phosphatase 7, tartrate resistant (putative)   390928
## FCHSD1      FCHSD1                      FCH and double SH3 domains 1    89848
## HEPHL1      HEPHL1                                 hephaestin like 1   341208
## TGM3          TGM3                                transglutaminase 3     7053
## LCN2          LCN2                                       lipocalin 2     3934
## GM2A          GM2A                         GM2 ganglioside activator     2760
## ATP1B1      ATP1B1         ATPase Na+/K+ transporting subunit beta 1      481
## SPRR2D      SPRR2D                     small proline rich protein 2D     6703
## GBA            GBA                           glucosylceramidase beta     2629
##             pval      padj  log2fc hkg
## IL17A  6.079e-90 9.480e-86 -37.690   n
## ACP7   1.147e-41 8.942e-38  -2.360   n
## FCHSD1 6.524e-35 3.392e-31  -1.937   n
## HEPHL1 1.834e-34 7.151e-31  -2.437   n
## TGM3   1.052e-33 3.281e-30  -1.707   n
## LCN2   4.130e-33 1.074e-29  -2.607   n
## GM2A   1.821e-32 4.056e-29  -1.225   n
## ATP1B1 7.710e-32 1.503e-28  -1.934   n
## SPRR2D 2.260e-31 3.916e-28  -2.251   n
## GBA    4.877e-30 7.606e-27  -1.481   n
```

Pathway enrichment analysis

The pathway enrichemnt analysis included the follwoing steps: 1. Identify significant and background genes 1. Run PA analysis 1. Plot Save PA analysis results

In a first step, we convert the gene symbol to entrezID. That is needed to use the later the ReactomePA::enrichPathway function. In best case use always the esemblID.

```r
df.dge_res <- rename_genetoentrezid(df.dge_results = df.dge_res)
```

Get significantly DEx genes and background genes

Now we can start with idetifying differential expressed genes with our manually set cut parameters.

```r
# I. Define significant DEx genes and background genes
# I.a) sort genes into groups belonging either to reference or test (control) condition
df.ref_degenes <- get_significantgenes(df.dge_results = df.dge_res, p_value = p.value,
                              lfc_factor = -l2fc.factor, op = `<`)
df.ref <- df.ref_degenes[[1]]
# I. Rank genes based on their fold change
ranked_genes.ref <- do_rank_genes(df.dge_results = df.ref)

df.ctrl_degenes <- get_significantgenes(df.dge_results = df.dge_res, p_value = p.value,
                                       lfc_factor = l2fc.factor, op = `>`)
df.ctrl <- df.ctrl_degenes[[1]]

# I. Rank  genes based on their fold change
ranked_genes.ctrl <- do_rank_genes(df.dge_results = df.ctrl)
```

```r
# I.b) Background genes are all genes from our (sup-) data set
bg_genes <- as.character(df.dge_res$entrezid)
```

Identify enriched pathways using ReactomePA

```r
# II.a) Find enriched Pathways for reference condition
reactome_object.ref <- ReactomePA::enrichPathway(
  gene = df.ref_degenes[[2]], # a vector of entrezID
  universe = bg_genes, organism = 'human',
  qvalueCutoff = fdr.value, pvalueCutoff = pval.cut,  pAdjustMethod = multitest.method,
  minGSSize = minGSSize, maxGSSize = 500, readable = T)

# II.b) Find enriched Pathways for control condition
reactome_object.ctrl <- ReactomePA::enrichPathway(
  gene = df.ctrl_degenes[[2]], # a vector of entrezID
  universe = bg_genes, organism = 'human',
  qvalueCutoff = fdr.value, pvalueCutoff = pval.cut, pAdjustMethod = multitest.method,
  minGSSize = minGSSize, maxGSSize = 500, readable = T)
```

In order to plot gene names instead of entrezIDs they have to be converted. For this use the function "*DOSE::setReadable*" to convert entrezIDs to gene symbol.

```r
################### ---> convert gene ID to Symbol <--- ###################
reactome.ctrl <- setreadable_pa(paenrich_object = reactome_object.ctrl)
print("Pathways associated with cytokine-negative group")
```

```
## [1] "Pathways associated with cytokine-negative group"
```

```r
print.data.frame(reactome.ctrl[1:3, ], digits = 4)
```

```
##                           ID                         Description GeneRatio
## R-HSA-6805567 R-HSA-6805567                      Keratinization    28/339
## R-HSA-6809371 R-HSA-6809371 Formation of the cornified envelope    28/339
## NA                      <NA>                                <NA>      <NA>
##                BgRatio    pvalue  p.adjust    qvalue
## R-HSA-6805567 104/8810 1.021e-16 3.262e-14 3.182e-14
## R-HSA-6809371 104/8810 1.021e-16 3.262e-14 3.182e-14
## NA                <NA>        NA        NA        NA
##
## R-HSA-6805567 FLG/CASP14/KRT10/KRT2/LCE1C/LCE1E/LCE1A/LCE1B/LCE6A/RPTN/KRT73/LCE2C/LCE2A/LCE2B/LCE2D/
## R-HSA-6809371 FLG/CASP14/KRT10/KRT2/LCE1C/LCE1E/LCE1A/LCE1B/LCE6A/RPTN/KRT73/LCE2C/LCE2A/LCE2B/LCE2D/
## NA
##               Count
## R-HSA-6805567    28
## R-HSA-6809371    28
## NA               NA
```

```r
reactome.ref <- setreadable_pa(paenrich_object = reactome_object.ref)
print("Pathways associated with cytokine-positive group")
```

```
## [1] "Pathways associated with cytokine-positive group"
```

```r
print.data.frame(reactome.ref[1:3, ], digits = 4)
```

```
##                           ID                                Description
## R-HSA-449147   R-HSA-449147                  Signaling by Interleukins
## R-HSA-6783783 R-HSA-6783783                   Interleukin-10 signaling
## R-HSA-6785807 R-HSA-6785807 Interleukin-4 and Interleukin-13 signaling
##               GeneRatio  BgRatio    pvalue  p.adjust    qvalue
## R-HSA-449147     40/219 419/8810 9.467e-14 5.472e-11 5.003e-11
## R-HSA-6783783    13/219  43/8810 1.837e-11 5.308e-09 4.853e-09
## R-HSA-6785807    17/219 101/8810 3.585e-10 6.906e-08 6.314e-08
##
## R-HSA-449147  IL17A/LCN2/IL19/IL1RN/IL36RN/IL36G/SOD2/SHC1/CXCL1/IL17F/CXCL8/LYN/NOS2/IL20/IL1B/CCL3/
## R-HSA-6783783
## R-HSA-6785807
##               Count
## R-HSA-449147     40
## R-HSA-6783783    13
## R-HSA-6785807    17
```

Save and Plot Pathwyas

```r
###############################################################################
################### ---> Save results to csv file <--- #####################
###############################################################################
# Attention:
# ctrl (= negative log2FC) and ref (= positive log2FC) are switched for Immune publication
save_enrichobject_as_csv(paenrich_object = reactome.ctrl, condition = 'Cytoneg',
                         pa_database = 'REACTOME', output_path = output.dir)
save_enrichobject_as_csv(paenrich_object = reactome.ref, condition = 'Cytopos',
                         pa_database = 'REACTOME', output_path = output.dir)
```

Visualise Pathways in a cnet- and dotplot.

```r
# III.a) Plot variables
# select pathways or Enriched gene sets manually
publication_pas <- pathwaysofinterest()
if (seq.technique == 'SC')
{
  pas_publication <- grep(paste('sc', cytokine, sep = "_"), hash::keys(publication_pas),
                         value = TRUE)
} else
{
  pas_publication <- grep(paste('st', cytokine, sep = "_"), hash::keys(publication_pas),
                         value = TRUE)
}
show_categories <- publication_pas[[pas_publication]]
```

First, we plot the enriched pathways of the *IL17A*-positive group and save the plots as .pdf.

```r
# III.b) Reference Condition
# If a gene is associated with two or more enriched PAs
# but less than those are shown than this results in a bug
# ==> the log2FC of that gene is not correctly shown
if (!is.null(nrow(reactome.ref)))
{
  if (nrow(reactome.ref) > 1 & any(show_categories %in% reactome.ref$Description))
  {
    # Cnetplots to visualise enriched pathways
    fig.pathways.REACTOME(reactome_res = reactome.ref,
                          entrezid_log2fc = ranked_genes.ref,
                          showCategories = show_categories,
                          output.dir = output.dir,
                          title = "Cytopos_REACTOME_Pathway_Enrichment_Analysis.pdf",
                          width = width_img, height = height_img)

    # Dotplot to visualise enriched pathways
    fig.pathway.dotplot(pathway_res = reactome.ref,
                        showCategories = show_dotplot_categories,
                        method = 'REACTOME',
                        output.dir = output.dir,
                        title = "Cytopos_REACTOME_dotplot.pdf",
                        width = width_img, height = height_img)
  }
}
```

```
## Warning in all(entrezid_log2fc): wandle Argument des Typs 'double' nach boolesch

## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.

## pdf
##   2
```

Second, we visualise the enriched pathways in the *IL17A*-negative group.

```
# # III.c) Control Condition
if (!is.null(nrow(reactome.ctrl)))
{
  if (nrow(reactome.ctrl) > 1 & any(show_categories %in% reactome.ctrl$Description))
  {
    # Cnetplots to visualise enriched pathways
    fig.pathways.REACTOME(reactome_res = reactome.ctrl,
                          entrezid_log2fc = ranked_genes.ctrl,
                          showCategories = show_categories,
                          width = width_img, height = height_img,
                          output.dir = output.dir,
                          title = "Cytoneg_REACTOME_Pathway_Enrichment_Analysis.pdf")

    # Dotplot to visualise enriched pathways
    fig.pathway.dotplot(pathway_res = reactome.ctrl,
                        showCategories = show_dotplot_categories,
                        method = 'REACTOME',
                        width = width_img, height = height_img,
                        output.dir = output.dir,
                        title = "Cytoneg_REACTOME_dotplot.pdf")
  }
}
```

```
## Warning in all(entrezid_log2fc): wandle Argument des Typs 'double' nach boolesch

## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.

## pdf
##   2
```

References

Cite used Bioconductor packages:

```
citation("ReactomePA")
```

```
##
## Please cite G. Yu (2015) for using ReactomePA. In addition, please cite
## G. Yu (2012) when using compareCluster in clusterProfiler package, G.
## Yu (2015) when applying enrichment analysis to NGS data by using
## ChIPseeker
##
##   Guangchuang Yu, Qing-Yu He. ReactomePA: an R/Bioconductor package for
##   reactome pathway analysis and visualization. Molecular BioSystems
##   2016, 12(2):477-479
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization},
##     author = {Guangchuang Yu and Qing-Yu He},
##     journal = {Molecular BioSystems},
##     year = {2016},
##     volume = {12},
##     number = {12},
##     pages = {477-479},
##     pmid = {26661513},
##     url = {http://pubs.rsc.org/en/Content/ArticleLanding/2015/MB/C5MB00663E},
##     doi = {10.1039/C5MB00663E},
##   }
```

```
citation("DOSE")
```

```
##
## Please cite G. Yu (2015) for using DOSE. In addition, please cite G. Yu
## (2012) when using compareCluster in clusterProfiler package, G. Yu
## (2015) when applying enrichment analysis to NGS data by using
## ChIPseeker and G. Yu (2010) when using GOSemSim for GO semantic
## similarity analysis
##
##   Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, Qing-Yu He. DOSE: an
##   R/Bioconductor package for Disease Ontology Semantic and Enrichment
##   analysis. Bioinformatics 2015 31(4):608-609
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis},
##     author = {Guangchuang Yu and Li-Gen Wang and Guang-Rong Yan and Qing-Yu He},
##     journal = {Bioinformatics},
##     year = {2015},
##     volume = {31},
##     number = {4},
##     pages = {608-609},
##     url = {http://bioinformatics.oxfordjournals.org/content/31/4/608},
##     doi = {10.1093/bioinformatics/btu684},
##   }
```

```r
citation("org.Hs.eg.db")
```

```
## Warning in citation("org.Hs.eg.db"): no date field in DESCRIPTION file of
## package 'org.Hs.eg.db'
```

```
##
## To cite package 'org.Hs.eg.db' in publications use:
##
##   Marc Carlson (2020). org.Hs.eg.db: Genome wide annotation for Human.
##   R package version 3.11.4.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {org.Hs.eg.db: Genome wide annotation for Human},
##     author = {Marc Carlson},
##     year = {2020},
##     note = {R package version 3.11.4},
##   }
##
## ATTENTION: This citation information has been auto-generated from the
## package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```

```r
citation("enrichplot")
```

```
##
## To cite package 'enrichplot' in publications use:
##
##   Guangchuang Yu (2020). enrichplot: Visualization of Functional
##   Enrichment Result. R package version 1.8.1.
##   https://github.com/GuangchuangYu/enrichplot
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {enrichplot: Visualization of Functional Enrichment Result},
##     author = {Guangchuang Yu},
##     year = {2020},
##     note = {R package version 1.8.1},
##     url = {https://github.com/GuangchuangYu/enrichplot},
##   }
```

```r
citation("pathview")
```

```
##
## To cite pathview:
##
##   Luo, W. and Brouwer C., Pathview: an R/Bioconductor package for
##   pathway-based data integration and visualization. Bioinformatics,
##   2013, 29(14): 1830-1831, doi: 10.1093/bioinformatics/btt285
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     author = {{Luo} and {Weijun} and {Brouwer} and {Cory}},
##     title = {Pathview: an R/Bioconductor package for pathway-based data integration and visualization
##     journal = {Bioinformatics},
##     year = {2013},
##     doi = {10.1093/bioinformatics/btt285},
##     volume = {29},
##     number = {14},
##     pages = {1830-1831},
##   }
##
## This free open-source software implements academic research by the
## authors. Its development took a large amount of extra time and effort.
## If you use it, please support the project by citing the listed journal
## articles.
```

```
sessionInfo()
```

```
## R version 4.0.3 Patched (2020-10-23 r79366)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] C/UTF-8/C/C/C/C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] cowplot_1.1.1        DOSE_3.14.0          org.Hs.eg.db_3.11.4
##  [4] AnnotationDbi_1.50.3 IRanges_2.24.0       S4Vectors_0.28.0
##  [7] Biobase_2.50.0       BiocGenerics_0.36.0  enrichplot_1.8.1
## [10] pathview_1.28.1      ReactomePA_1.32.0    Hmisc_4.5-0
## [13] ggplot2_3.3.3        Formula_1.2-4        survival_3.2-10
## [16] lattice_0.20-41      xlsx_0.6.5           tibble_3.1.0
## [19] stringr_1.4.0        knitr_1.31           kableExtra_1.3.4
## [22] hash_2.2.6.1         gtools_3.8.2         dplyr_1.0.5
##
## loaded via a namespace (and not attached):
##   [1] backports_1.2.1     fastmatch_1.1-0     systemfonts_1.0.1
##   [4] plyr_1.8.6          igraph_1.2.6        splines_4.0.3
##   [7] BiocParallel_1.24.1 urltools_1.7.3      digest_0.6.27
##  [10] htmltools_0.5.1.1   GOSemSim_2.14.2     viridis_0.5.1
##  [13] GO.db_3.11.4        fansi_0.4.2         magrittr_2.0.1
##  [16] checkmate_2.0.0     memoise_2.0.0       cluster_2.1.1
##  [19] Biostrings_2.56.0   graphlayouts_0.7.1  svglite_2.0.0
##  [22] prettyunits_1.1.1   jpeg_0.1-8.1        colorspace_2.0-0
##  [25] blob_1.2.1          rvest_1.0.0         rappdirs_0.3.3
##  [28] ggrepel_0.9.1       xfun_0.22           crayon_1.4.1
##  [31] RCurl_1.98-1.3      jsonlite_1.7.2      graph_1.66.0
##  [34] scatterpie_0.1.5    glue_1.4.2          polyclip_1.10-0
##  [37] gtable_0.3.0        zlibbioc_1.36.0     XVector_0.30.0
##  [40] webshot_0.5.2       graphite_1.34.0     Rgraphviz_2.32.0
##  [43] scales_1.1.1        DBI_1.1.1           Rcpp_1.0.6
##  [46] viridisLite_0.3.0   progress_1.2.2      htmlTable_2.1.0
##  [49] gridGraphics_0.5-1  foreign_0.8-81      bit_4.0.4
##  [52] reactome.db_1.70.0  europepmc_0.4       htmlwidgets_1.5.3
##  [55] httr_1.4.2          fgsea_1.14.0        RColorBrewer_1.1-2
##  [58] ellipsis_0.3.1      pkgconfig_2.0.3     XML_3.99-0.6
##  [61] rJava_0.9-13        farver_2.1.0        nnet_7.3-15
##  [64] utf8_1.2.1          labeling_0.4.2      ggplotify_0.0.5
##  [67] tidyselect_1.1.0    rlang_0.4.10        reshape2_1.4.4
##  [70] munsell_0.5.0       tools_4.0.3         cachem_1.0.4
##  [73] generics_0.1.0      RSQLite_2.2.5       ggridges_0.5.3
##  [76] evaluate_0.14       fastmap_1.1.0       yaml_2.2.1
##  [79] bit64_4.0.5         tidygraph_1.2.0     purrr_0.3.4
##  [82] KEGGREST_1.28.0     ggraph_2.0.5        KEGGgraph_1.48.0
##  [85] DO.db_2.9           xml2_1.3.2          compiler_4.0.3
##  [88] rstudioapi_0.13     png_0.1-7           tweenr_1.0.2
##  [91] stringi_1.5.3       Matrix_1.3-2        vctrs_0.3.7
##  [94] pillar_1.5.1        lifecycle_1.0.0     BiocManager_1.30.12
##  [97] triebeard_0.3.0     data.table_1.14.0   bitops_1.0-6
## [100] qvalue_2.22.0       R6_2.5.0            latticeExtra_0.6-29
## [103] gridExtra_2.3       MASS_7.3-53.1       assertthat_0.2.1
## [106] xlsxjars_0.6.1      withr_2.4.1         hms_1.0.0
## [109] grid_4.0.3          rpart_4.1-15        tidyr_1.1.3
## [112] rmarkdown_2.7       rvcheck_0.1.8       ggforce_0.3.3
## [115] base64enc_0.1-3
```