Documentation of the workflow for the H5N1 project

The following pipeline is used to analyze all data in our paper and generate the figures.

The recommended way to run the pipeline is using our **Singularity** container. For example, if you are using Singularity on NIAID HPC, follow these steps to start the container

$ module load Singularity

$ cd /hpcdata/sg/sg\_data/singularity/test

$ singularity shell -B \  
 /hpcdata/sg/sg\_data/CHI/PROJECTS/H5N1/PAPER/:/var/workflow1 \  
 h5n1\_image\_180410.img

Then you can execute R from the container with R command.

If Singularity is **not available**, it is possible to use standalone R.

Our pipeline requires the following R packages:

# CRAN:

install.packages(c("circlize", "cluster", "data.table", "ggdendro", "gplots",   
"gridExtra", "lazyeval", "plyr", "pROC", "rlang", "tidyverse",   
"tmod", "WGCNA"))

# Bioconductor 3.6:

source("http://bioconductor.org/biocLite.R")

biocLite()

biocLite(c("affy", "Biobase", "ComplexHeatmap", "genefilter", "GEOquery", "mygene"))

In the beginning of each R session run

setwd("/var/workflow1")

source("SCRIPTS/0\_initialize.r")

to change the working directory, initialize required variables and load commonly used libraries. This script creates PROJECT\_DIR variable pointing to the location of the workflow files. By defaults it is set to /var/workflow1 directory visible from the container.

# Gene Expression PBMC data processing

First, the CEL files were processed with Power Tools RNA-sketch algorithm.

source("SCRIPTS/MA/processing\_pbmc/apt.config.r")

source("SCRIPTS/MA/processing\_pbmc/apt.call.r")

Quantile normalization implement in RMA-sketch algorithm uses random subset of data to decrease memory usage. Because of this at different runs the algorithm may generates slightly different results. For reproducibility we recommend using **precomputed APT output** located in DATA\_PROCESSED/Microarrays/PBMC/apt\_summary.txt.

**Create ExpressionSet from APT output**

source("SCRIPTS/MA/processing\_pbmc/eset.config.r")

source("SCRIPTS/MA/processing\_pbmc/eset.call.r")

**Probesets to genes mapping**

Input data:

HuGene-2\_1-st-v1.na35.hg19.transcript.csv file downloaded from Affymetrix web site

Convert the table to probeset-gene mapping

source("SCRIPTS/MA/annotation/affy\_hugene-2\_1-st\_annotation.r")

Select the best probeset for a gene

source("SCRIPTS/MA/annotation/generate\_ps2gene\_map.config.r")

source("SCRIPTS/MA/annotation/generate\_ps2gene\_map.call.r")

**Data post processing**

We found that two samples were switched. This is to correct it.

source("SCRIPTS/MA/filtering\_pbmc/switch.samples/switch.samples.call.r")

Apply different filtering to samples and genes. The probesets mapped to genes.

source("SCRIPTS/MA/filtering\_pbmc/samples.clean\_genes.all/filtering.r")

source("SCRIPTS/MA/filtering\_pbmc/samples.clean\_genes.iqr/filtering.r")

source("SCRIPTS/MA/filtering\_pbmc/samples.all\_genes.all/filtering.r")

source("SCRIPTS/MA/filtering\_pbmc/samples.all\_genes.iqr/filtering.r")

Calculate fold change from day 0

source("SCRIPTS/MA/calculate\_d0\_fc/calculate\_d0\_fc\_pbmc.r")

# Pattern discovery in post-vaccination profiles of gene expression

Profiles clustering with DIANA

source("SCRIPTS/MA/pattern\_discovery/pattern\_discovery\_161109.r")

Cut the dendrogram tree at different levels and detect stable clusters

source("SCRIPTS/MA/pattern\_discovery/patterns\_cutTree\_stable.r")

Filter the patterns

source("SCRIPTS/MA/pattern\_discovery/pattern\_filter.r")

Summarize patterns stats

source("SCRIPTS/MA/pattern\_discovery/patterns\_stats.r")

Figure 2B - patterns profile plot

source("SCRIPTS/MA/pattern\_discovery/plot\_patterns.r")

Expand list of pattern signature genes

source("SCRIPTS/MA/pattern\_discovery/pattern\_filter\_expanded.r")

Compute correlations and clean up the genes

source("SCRIPTS/MA/pattern\_discovery/pattern\_expanded\_genes\_cor.r")

source("SCRIPTS/MA/pattern\_discovery/pattern\_expanded\_genes\_clean.r")

Compute subject scores for each pattern

source("SCRIPTS/MA/pattern\_discovery/patterns\_to\_subjects.r")

Output the table of genes (with annotations) for each pattern

source("SCRIPTS/MA/pattern\_discovery/pattern\_genes\_output.r")

BTM enrichment in patterns genes. Figure 2C

source("SCRIPTS/MA/pattern\_discovery/pattern\_BTM\_enrichment.r")

Add data for subject s10 and update the score matrix

source("SCRIPTS/MA/pattern\_discovery/s10\_peaks\_assessment.r")

source("SCRIPTS/MA/pattern\_discovery/pattern\_scores\_in\_samples\_GE\_incl.s10.r")

# Pattern discovery in post-vaccination profiles of flow cytometry data

Figure 2D (use Julian’s data)

Figure 2E (use Julian’s data)

Supplemental Figure 1D

Add data for subject s10

Figure 2F (partially use Julian’s data)

Supplemental Figure 2B

# Emory data analysis blindly predicting adjuvant status

Supplemental Figure 2C

Supplemental Figure 2D

Supplemental Figure 2E

# Find signature for adjuvant status prediction

Julian’s scripts for elastic net models

Figure 3B

Figure 3C

Figure 3E

Figure 3F and 3G

# Gene Expression PAXgene data processing

**Data post processing**

We found that two samples were switched. This is to correct it.

source("SCRIPTS/MA/filtering\_pax/switch.samples/switch.samples.call.r")

Apply different filtering to samples and genes. The probesets mapped to genes.

source("SCRIPTS/MA/filtering\_pax/filtering.r")

Calculate fold change from day 0

source("SCRIPTS/MA/calculate\_d0\_fc/calculate\_d0\_fc\_pax.r")

# Baseline data analysis

Preparing PBMC day 0 samples

source("SCRIPTS/MA/baseline\_pbmc/d0\_filter.r")

WGCNA clustering of PBMC samples

source("SCRIPTS/MA/baseline\_pbmc/d0\_wgcna.r")

source("SCRIPTS/MA/baseline\_pbmc/d0\_wgcna\_output.r")

BTM enrichment analysis of data from PBMC samples

source("SCRIPTS/MA/baseline\_pbmc/d0\_wgcna\_BTM\_enrichment.r")

Preparing whole blood (PAXgene) day 0 samples

source("SCRIPTS/MA/baseline\_pax/d0\_filter\_pax.r")

WGCNA clustering of whole blood samples

source("SCRIPTS/MA/baseline\_pax/d0\_wgcna.r")

source("SCRIPTS/MA/baseline\_pax/d0\_wgcna\_output.r")

BTM enrichment analysis of data from whole blood samples

source("SCRIPTS/MA/baseline\_pax/d0\_wgcna\_BTM\_enrichment.r")

FIgure 4A. Combining BTM enrichment results from PBMC and whole blood samples

source("SCRIPTS/MA/baseline/plot\_BTM\_pbmc\_pax.r")

Julian’s scripts for elastic net models

Figure 4B

Figure 5A

Figure 5B

# SOMAscan data analysis

Julian’s scripts for elastic net models

Figure 6A

Figure 6B

# Tfh cells data analysis

Figure 7A

Figure 7B

Figure 7C

Figure 7D

Figure 7E