Papalexi Analysis Sceptre

2023-02-16

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(httr)
library(rlist)
library(jsonlite)
```

```
##
## Attaching package: 'jsonlite'
##
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
library(varhandle)
library(stringi)
```

Getting Results From SCEPTRE Analysis

```
#using absolute paths to download results since files exist on github
data.dir = '/Users/kmason/sceptre2-manuscript/writeups/papalexi_analysis/'
gene_path = paste0(data.dir,'gene_result_PDL1_mrna.rds')
protein_path = paste0(data.dir,'protein_result.rds')
seurat_path = paste0(data.dir,'papalexi_results_seurat.rds')
gene_result = readRDS(gene_path)
protein_result = readRDS(protein_path)
seurat_result = readRDS(seurat_path)
```

Adjusting Pvalues

```
#See which gene perturbations are associated with PDL1 protein expression

#get pvalues from sceptre
P_adj = protein_result[,1]
#unlist pvalues
P_adj = unlist(P_adj)
#some pvalues are negative so take absolute value
P_adj = abs(P_adj)
#make numeric
P_adj = as.numeric(P_adj)
#perform BH procedure
P_adj = p.adjust(P_adj,method = 'BH')

#replace results matrix pvalues with adjusted pvalues
protein_adjusted= cbind(P_adj,protein_result[,c(2,3)])
```

Papalexi et al Comparison

Papalexi notes that there have been numerous studies showing that exposure to IFN-gamma induces PDL1 expression. Core components of the the IFN-gamma response include IRF1,JAK,STAT, and the IFN-gamma receptors themselves.We can think of these as positive controls that when knocked out, result in underexpression of PDL1.

PDL1 Protein Results

The authors state that they found 8 perturbations to be significantly associated with PDL1 protein expression. These 8 perturbations are BRD4,MYC,CUL3,IRF1,STAT1,IFNGR1,IFNGR2,and JAK2. The last 5 are positive controls identified in the paper.

```
#filter to just look at PDL1 pvalues
protein_PDL1 = protein_adjusted[which(protein_adjusted[,3] == 'PDL1')]
#get significant perturbations
sig_genes = subset(protein_PDL1,P_adj < 0.05)$grna_group
#unfactor
sig_genes = unfactor(sig_genes)
A = 'Significant perturbations for the expression of PDL1 protein are:\n '
cat(paste0(A,paste(sig_genes,collapse = ',')))
```

```
## Significant perturbations for the expression of PDL1 protein are:
##  IRF1,BRD4,CUL3,CMTM6,CD86,IFNGR1,IFNGR2,JAK2,MYC,PDCD1LG2,STAT1,STAT3
```

In addition to the 8 perturbations that papalexi et al found significant, SCEPTRE finds 4 more (CMTM6,CD86,PDCD1LG2,STAT3).

```
#get sceptre results on PDL1 protein
protein_PDL1_raw = protein_result[which(protein_result[,3] == 'PDL1'),c(1,2)]
#get seurat results on PDL1 protein
seurat_PDL1_raw = seurat_result[which(seurat_result$Target == 'PDL1'),c(1,2,6)]
#remove perturbations that aren't in both analyses
```

```
for(target in protein_PDL1_raw$grna_group){
  if (target %in% seurat_PDL1_raw$PRTB == F){
    seurat_PDL1_raw = rbind(seurat_PDL1_raw,c(NA,NA,target))
  }
}
protein_PDL1_raw = subset(protein_PDL1_raw,grna_group %in% seurat_PDL1_raw$PRTB)
#reorder rows of sceptre analysis
protein_PDL1_raw = protein_PDL1_raw[match(seurat_PDL1_raw$PRTB,
                                          protein_PDL1_raw$grna_group)]
protein_PDL1_raw$grna_group = unfactor(protein_PDL1_raw$grna_group)
#bind them
combined_results_PDL1 = data.frame(protein_PDL1_raw$p_value,
                                   seurat_PDL1_raw$p_val,
                                   seurat_PDL1_raw$avg_log2FC,
                                protein_PDL1_raw$grna_group)
colnames(combined_results_PDL1) = c("SCEPTRE Pvalues","Seurat Pvalues",
                                    'Seurat Log Change',
                                    "Perturbation")
#reorder columns
combined_results_PDL1 = combined_results_PDL1 [,c(4,1,2,3)]
#order by pvalue
combined_results_PDL1 = combined_results_PDL1[
  order(combined_results_PDL1$`SCEPTRE Pvalues`),]
combined_results_PDL1
```

```
##    Perturbation SCEPTRE Pvalues        Seurat Pvalues    Seurat Log Change
## 1          BRD4      -0.0000008  5.03704648791047e-29    0.219431300674301
## 15         CD86      -0.0000008                  <NA>                 <NA>
## 3        IFNGR1       0.0000008 4.80639672997527e-303   -0.296123832691995
## 4        IFNGR2       0.0000008 3.09522669645794e-294    -0.30510460394117
## 5          IRF1       0.0000008  1.57616101780465e-70   -0.150428359042648
## 6          JAK2       0.0000008 1.32288915682211e-275   -0.310481551659606
## 8         STAT1       0.0000008 3.98304971401922e-124   -0.281972717049105
## 12        CMTM6       0.0000008                  <NA>                 <NA>
## 20     PDCD1LG2       0.0000032                  <NA>                 <NA>
## 7           MYC       0.0000152  1.63893775436278e-08    0.226520041834825
## 2          CUL3       0.0011208   5.5042967921981e-11    0.153812138600912
## 22        STAT3       0.0193600                  <NA>                 <NA>
## 19       NFKBIA       0.1090400                  <NA>                 <NA>
## 21       POU2F2       0.1220800                  <NA>                 <NA>
## 9         STAT2       0.1687200    0.0427852117568571   -0.017221256835477
## 16         ETV7       0.4840000                  <NA>                 <NA>
## 14         CAV1       0.5182400                  <NA>                 <NA>
## 17         IRF7       0.5506400                  <NA>                 <NA>
## 18       MARCH8       0.5634400                  <NA>                 <NA>
## 25       UBE2L6       0.8733600                  <NA>                 <NA>
## 23       STAT5A       0.8840000                  <NA>                 <NA>
## 13         ATF2       0.8876800                  <NA>                 <NA>
## 24     TNFRSF14       0.9540800                  <NA>                 <NA>
## 11         SPI1       0.9576000     0.139657481517574  -0.0723864697078291
## 10        SMAD4       0.9995200     0.664283575269382 -0.00645627045399122
```

```
#View(combined_results_PDL1)
```

Other Protein Results

The authors state that "Importantly, perturbation of these eight genes did not result in appreciable shifts in CD86 or PD-L2 protein expression,suggesting that these regulatory effects are specific to PD-L1'. I will now check to see if this is true when using SCEPTRE.

CD86

```
#get sceptre results on CD86 protein
protein_CD86_raw = protein_result[which(protein_result[,3] == 'CD86'),c(1,2)]
#get seurat results on CD86 protein
seurat_CD86_raw = seurat_result[which(seurat_result$Target == 'CD86'),c(1,2,6)]
#remove perturbations that aren't in both analyses
for(target in protein_CD86_raw$grna_group){
  if (target %in% seurat_CD86_raw$PRTB == F){
    seurat_CD86_raw = rbind(seurat_CD86_raw,c(NA,NA,target))
  }
}
protein_CD86_raw = subset(protein_CD86_raw,grna_group %in% seurat_CD86_raw$PRTB)
#reorder rows of sceptre analysis
protein_CD86_raw = protein_CD86_raw[match(seurat_CD86_raw$PRTB,
                                          protein_CD86_raw$grna_group)]
protein_CD86_raw$grna_group = unfactor(protein_CD86_raw$grna_group)
#bind them
combined_results_CD86 = data.frame(protein_CD86_raw$p_value,
                                   seurat_CD86_raw$p_val,
                                   seurat_CD86_raw$avg_log2FC,
                                protein_CD86_raw$grna_group)
colnames(combined_results_CD86) = c("SCEPTRE Pvalues","Seurat Pvalues",
                                    'Seurat Log Change',
                                    "Perturbation")
#reorder columns
combined_results_CD86 = combined_results_CD86[,c(4,1,2,3)]
#order by pvalue
combined_results_CD86 = combined_results_CD86[
  order(combined_results_CD86$`SCEPTRE Pvalues`),]
combined_results_CD86
```

```
##    Perturbation SCEPTRE Pvalues       Seurat Pvalues   Seurat Log Change
## 3         IFNGR1      -0.0000008 2.60239767540458e-53   0.121201786637612
## 4         IFNGR2      -0.0000008 1.30126716030559e-50   0.119823683194634
## 5           IRF1      -0.0000008 1.42717472959289e-26  0.0943809179985599
## 6           JAK2      -0.0000008 1.55211033346226e-42   0.117354219064871
## 8          STAT1      -0.0000008 6.69452748138104e-29   0.125148310062355
## 12         CMTM6      -0.0000008                  <NA>                <NA>
## 1           BRD4       0.0000008  2.5243801374878e-18  -0.159965674501846
## 15          CD86       0.0000008                  <NA>                <NA>
## 7            MYC       0.0001648     0.01994394800964  -0.101149843612133
```

```
## 21       POU2F2       0.0191200                <NA>                <NA>
## 22        STAT3       0.0456800                <NA>                <NA>
## 18       MARCH8       0.1648000                <NA>                <NA>
## 20     PDCD1LG2       0.2644000                <NA>                <NA>
## 16         ETV7       0.2649600                <NA>                <NA>
## 9         STAT2       0.4392800   0.0723538653474799  0.0108483032483386
## 25       UBE2L6       0.5301600                <NA>                <NA>
## 14         CAV1       0.6592000                <NA>                <NA>
## 24     TNFRSF14       0.7219200                <NA>                <NA>
## 11         SPI1       0.7258400    0.281215388551183  0.0337233024451993
## 13         ATF2       0.7275200                <NA>                <NA>
## 23       STAT5A       0.7884000                <NA>                <NA>
## 19       NFKBIA       0.8016000                <NA>                <NA>
## 10        SMAD4       0.8167200    0.230063901155693 -0.0141535510881986
## 2          CUL3       0.8398400    0.180497845816223  0.0500153972640598
## 17         IRF7       0.8580800                <NA>                <NA>
```

```
#View(combined_results_CD86)
```

It seems as though these regulatory effects are not specific to PDL1. This is also true when using seurat. The positive controls for PDL1 (IRF1,IFNGR1,etc...) also seem to regulate expression of CD86.

PDL2

```
#get sceptre results on PDL2 protein
protein_PDL2_raw = protein_result[which(protein_result[,3] == 'PDL2'),c(1,2)]
#get seurat results on PDL2 protein
seurat_PDL2_raw = seurat_result[which(seurat_result$Target == 'PDL2'),c(1,2,6)]
#remove perturbations that aren't in both analyses
for(target in protein_PDL2_raw$grna_group){
  if (target %in% seurat_PDL2_raw$PRTB == F){
    seurat_PDL2_raw = rbind(seurat_PDL2_raw,c(NA,NA,target))
  }
}
protein_PDL2_raw = subset(protein_PDL2_raw,grna_group %in% seurat_PDL2_raw$PRTB)
#reorder rows of sceptre analysis
protein_PDL2_raw = protein_PDL2_raw[match(seurat_PDL2_raw$PRTB,
                                          protein_PDL2_raw$grna_group)]
protein_PDL2_raw$grna_group = unfactor(protein_PDL2_raw$grna_group)
#bind them
combined_results_PDL2 = data.frame(protein_PDL2_raw$p_value,
                                   seurat_PDL2_raw$p_val,
                                   seurat_PDL2_raw$avg_log2FC,
                                protein_PDL2_raw$grna_group)
colnames(combined_results_PDL2) = c("SCEPTRE Pvalues","Seurat Pvalues",
                                    'Seurat Log Change',
                                    "Perturbation")
#reorder columns
combined_results_PDL2 = combined_results_PDL2[,c(4,1,2,3)]
#order by pvalue
combined_results_PDL2 = combined_results_PDL2[
  order(combined_results_PDL2$`SCEPTRE Pvalues`),]
```

```
combined_results_PDL2
```

```
##    Perturbation SCEPTRE Pvalues       Seurat Pvalues     Seurat Log Change
## 12        CMTM6      -0.0000008                 <NA>                  <NA>
## 15         CD86      -0.0000008                 <NA>                  <NA>
## 5          IRF1       0.0000008 1.90888948648785e-23   -0.0599705742955019
## 20     PDCD1LG2       0.0000008                 <NA>                  <NA>
## 1          BRD4       0.0000016   0.0133055581996283    -0.032916899117855
## 2          CUL3       0.0000016 4.83442777171049e-06   -0.0644568893179793
## 7           MYC       0.0014168 0.000420069274583692   -0.0986338568170708
## 19       NFKBIA       0.0017120                 <NA>                  <NA>
## 9         STAT2       0.0393600    0.547932718592113  -0.00350818641405315
## 25       UBE2L6       0.0824000                 <NA>                  <NA>
## 14         CAV1       0.0962400                 <NA>                  <NA>
## 3        IFNGR1       0.1015200    0.102204529166128   -0.0100674365044353
## 10        SMAD4       0.1274400    0.594672103673449 -0.000877054816040834
## 4        IFNGR2       0.1477600    0.546498208410352 -0.000650280686811056
## 24     TNFRSF14       0.1812800                 <NA>                  <NA>
## 17         IRF7       0.1894400                 <NA>                  <NA>
## 6          JAK2       0.2396800     0.76024657053491  -0.00369575712521553
## 23       STAT5A       0.2475200                 <NA>                  <NA>
## 22        STAT3       0.3740800                 <NA>                  <NA>
## 11         SPI1       0.4065600    0.505944202945339   -0.0240017445637708
## 21       POU2F2       0.5328000                 <NA>                  <NA>
## 18       MARCH8       0.5420000                 <NA>                  <NA>
## 8         STAT1       0.8042400 0.000640778131322823   -0.0286697132262335
## 16         ETV7       0.9152000                 <NA>                  <NA>
## 13         ATF2       0.9703200                 <NA>                  <NA>
```

```
#View(combined_results_PDL2)
```

For PDL2, the results are more or less in line with what paplexi reports. IRF1 seems to be the only gene that regulates PDL1 that also regulates PDL2.

mRNA Results

I will now analyze whether or not the perturbations affect PDL1 mRNA expression. I will compare results on the normalized and un-normalized assays when using seurat.

Raw Data (Not Normalized)

```
#get sceptre results on PDL1 gene
gene_PDL1_raw = gene_result[which(gene_result$response_id == 'CD274'),c(1,2)]
#get seurat results on PDL1 gene
seurat_PDL1_raw = seurat_result[which(seurat_result$Target == 'PDL1_raw'),c(1,2,6)]
#remove perturbations that aren't in both analyses
for(target in gene_PDL1_raw$grna_group){
  if (target %in% seurat_PDL1_raw$PRTB == F){
```

```
    seurat_PDL1_raw = rbind(seurat_PDL1_raw,c(NA,NA,target))
  }
}
gene_PDL1_raw = subset(gene_PDL1_raw,grna_group %in% seurat_PDL1_raw$PRTB)
#reorder rows of sceptre analysis
gene_PDL1_raw = gene_PDL1_raw[match(seurat_PDL1_raw$PRTB,
                                    gene_PDL1_raw$grna_group)]
gene_PDL1_raw$grna_group = unfactor(gene_PDL1_raw$grna_group)
#bind them
combined_results_PDL1_raw = data.frame(gene_PDL1_raw$p_value,
                                seurat_PDL1_raw$p_val,
                                seurat_PDL1_raw$avg_log2FC,
                              gene_PDL1_raw$grna_group)
colnames(combined_results_PDL1_raw) = c("SCEPTRE Pvalues","Seurat Pvalues",
                                 'Seurat Log Change',
                                 "Perturbation")
#reorder columns
combined_results_PDL1_raw = combined_results_PDL1_raw[,c(4,1,2,3)]
#order by pvalue
combined_results_PDL1_raw = combined_results_PDL1_raw[
  order(combined_results_PDL1_raw$`SCEPTRE Pvalues`),]

combined_results_PDL1_raw
```

```
##    Perturbation SCEPTRE Pvalues       Seurat Pvalues   Seurat Log Change
## 3         IFNGR1       0.0000008 1.25296504251056e-79  -0.494261700858603
## 4         IFNGR2       0.0000008 3.88431749918983e-81  -0.520836585016837
## 6           JAK2       0.0000008 5.95716063506082e-75  -0.525395039368531
## 8          STAT1       0.0000008 9.63147084806935e-34  -0.521101575491434
## 5           IRF1       0.0000096  4.3730380264779e-07  -0.139955861389527
## 2           CUL3       0.0000224 1.42396508597443e-11    0.69845248884795
## 7            MYC       0.0000264 8.34387552747722e-10    1.13889698296579
## 16          ETV7       0.0420000                 <NA>                <NA>
## 22         STAT3       0.1553600                 <NA>                <NA>
## 18        MARCH8       0.1897600                 <NA>                <NA>
## 20     PDCD1LG2       0.1946400                 <NA>                <NA>
## 19       NFKBIA       0.2524800                 <NA>                <NA>
## 11          SPI1       0.2531200    0.897122650969822  0.0461432151045903
## 24     TNFRSF14       0.3843200                 <NA>                <NA>
## 23       STAT5A       0.4168000                 <NA>                <NA>
## 21       POU2F2       0.5276800                 <NA>                <NA>
## 9          STAT2       0.5440800    0.612742711587726  0.0283387109126413
## 25       UBE2L6       0.5622400                 <NA>                <NA>
## 14          CAV1       0.6335200                 <NA>                <NA>
## 10         SMAD4       0.7674400    0.155064135787154 -0.0795434529236683
## 17          IRF7       0.7689600                 <NA>                <NA>
## 13          ATF2       0.8680800                 <NA>                <NA>
## 1           BRD4       0.8944800 1.67914463783413e-05     0.3875987237838
## 15          CD86       0.9194400                 <NA>                <NA>
## 12         CMTM6       0.9832800                 <NA>                <NA>
```

```
#View(combined_results_PDL1_raw)
```

We see that the positive controls indeed correspond to the strongest signals in the data. Interestingly, BRD4 perturbation does not affect gene expression of PDL1 while CUL3 does, consistent with the paper.

Normalized Data

```
#get sceptre results on PDL1 gene
gene_PDL1_raw = gene_result[which(gene_result$response_id == 'CD274'),c(1,2)]
#get seurat results on PDL1 gene
seurat_PDL1_raw = seurat_result[which(seurat_result$Target == 'PDL1_norm'),c(1,2,6)]
#remove perturbations that aren't in both analyses
for(target in gene_PDL1_raw$grna_group){
  if (target %in% seurat_PDL1_raw$PRTB == F){
    seurat_PDL1_raw = rbind(seurat_PDL1_raw,c(NA,NA,target))
  }
}
gene_PDL1_raw = subset(gene_PDL1_raw,grna_group %in% seurat_PDL1_raw$PRTB)
#reorder rows of sceptre analysis
gene_PDL1_raw = gene_PDL1_raw[match(seurat_PDL1_raw$PRTB,
                                    gene_PDL1_raw$grna_group)]
gene_PDL1_raw$grna_group = unfactor(gene_PDL1_raw$grna_group)
#bind them
combined_results_PDL1_norm = data.frame(gene_PDL1_raw$p_value,
                                  seurat_PDL1_raw$p_val,
                                  seurat_PDL1_raw$avg_log2FC,
                                gene_PDL1_raw$grna_group)
colnames(combined_results_PDL1_norm) = c("SCEPTRE Pvalues","Seurat Pvalues",
                                   'Seurat Log Change',
                                   "Perturbation")
#reorder columns
combined_results_PDL1_norm = combined_results_PDL1_norm[,c(4,1,2,3)]
#order by pvalue
combined_results_PDL1_norm = combined_results_PDL1_norm[
  order(combined_results_PDL1_norm$`SCEPTRE Pvalues`),]

combined_results_PDL1_norm
```

```
##    Perturbation SCEPTRE Pvalues       Seurat Pvalues   Seurat Log Change
## 3         IFNGR1       0.0000008   0.0021481529064697    0.14397023308446
## 4         IFNGR2       0.0000008 0.000117559232115827   0.157532182584959
## 6           JAK2       0.0000008 7.82875254483344e-05   0.162901269113791
## 8          STAT1       0.0000008    0.154518816172577   0.141813012643277
## 5           IRF1       0.0000096 6.96789008866567e-26   0.201020757255128
## 2           CUL3       0.0000224    0.961040653388516 -0.0646596565806011
## 7            MYC       0.0000264 7.03181042117195e-05  -0.577955564998542
## 16          ETV7       0.0420000                 <NA>                <NA>
## 22         STAT3       0.1553600                 <NA>                <NA>
## 18        MARCH8       0.1897600                 <NA>                <NA>
## 20     PDCD1LG2       0.1946400                 <NA>                <NA>
## 19        NFKBIA       0.2524800                 <NA>                <NA>
```

```
## 11         SPI1       0.2531200    0.746342790425717  -0.109037226731291
## 24     TNFRSF14       0.3843200                 <NA>                <NA>
## 23       STAT5A       0.4168000                 <NA>                <NA>
## 21       POU2F2       0.5276800                 <NA>                <NA>
## 9         STAT2       0.5440800    0.945810922573788  -0.011569776135782
## 25       UBE2L6       0.5622400                 <NA>                <NA>
## 14         CAV1       0.6335200                 <NA>                <NA>
## 10        SMAD4       0.7674400    0.486325602565747 -0.0178206665165305
## 17         IRF7       0.7689600                 <NA>                <NA>
## 13         ATF2       0.8680800                 <NA>                <NA>
## 1          BRD4       0.8944800    0.123111224845803  -0.160949138158263
## 15         CD86       0.9194400                 <NA>                <NA>
## 12        CMTM6       0.9832800                 <NA>                <NA>
```

```
#View(combined_results_PDL1_norm)
```

If we use the normalized data, we see that while the positive controls are still the strongest signal in the data, the direction of over/under expression is now flipped. From this I have two possible conclusions. Therefore I reasonably certain that they used the un-normalized data to perform all DE analyses. This is because if they used the normalized data, all their results would contradict the known controls. it is possible that their implementation of neighborhood adjustment is wrong in the sense that they are subtracting in the wrong direction (a-b instead of b-a). However, even if this is the case, the pvalues are much larger which casts doubt to if they would still be significant after pvalue adjustment.
If they did not use the normalized data, this begs the question as to if their results are driven by technical factors rather than true biological signal.