SCEPTRE vs SEURAT Plots

2023-03-08

```
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)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(ggplot2)
```

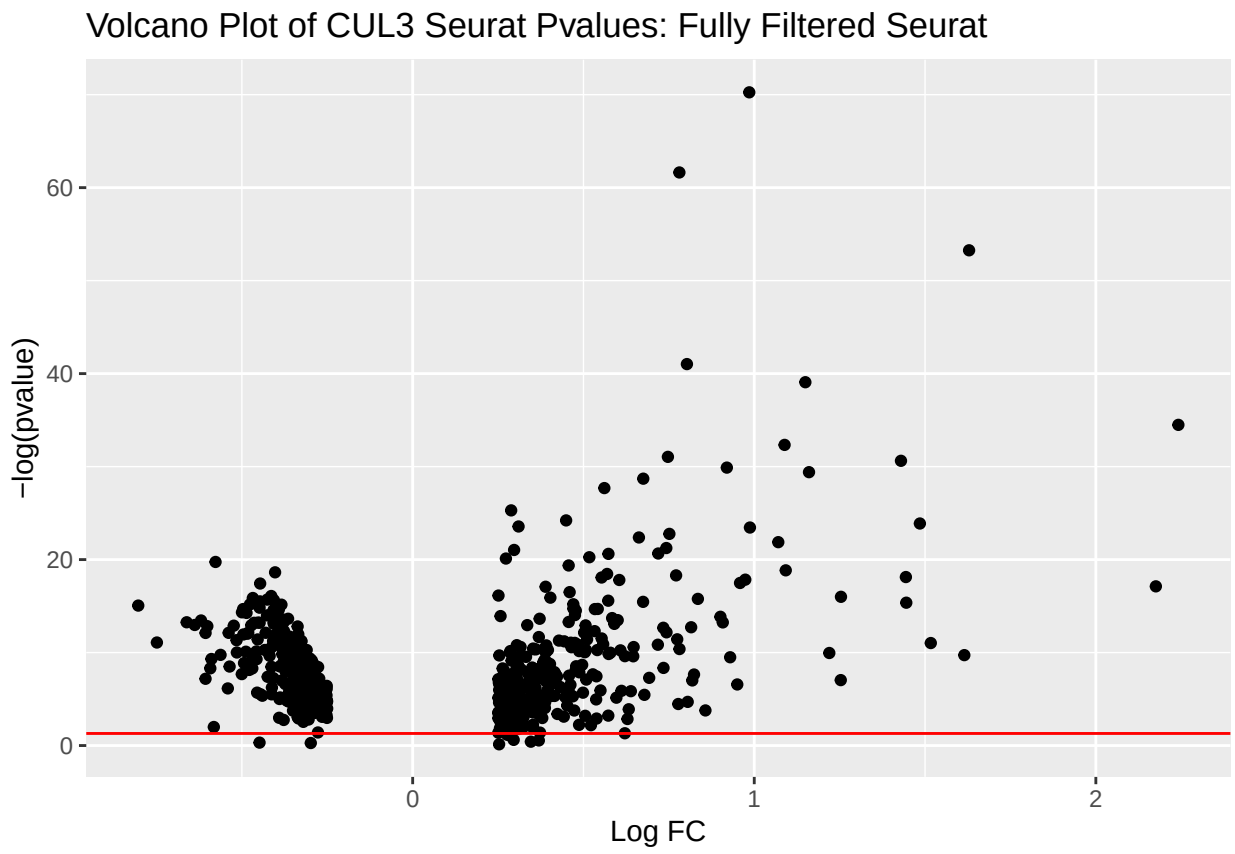
Getting Results From SCEPTRE Analysis

```
#using absolute paths to download results since files exist on github
code_dir = .get_config_path("LOCAL_CODE_DIR")
data.dir = paste0(code_dir,"/sceptre2-manuscript/writeups/papalexi_analysis/")
gene_path = paste0(data.dir,
                   'sceptre_CUL3_and_PDL1_mrna_results_with_effect_size.rds')
protein_path = paste0(data.dir,'sceptre_protein_results_with_effect_size.rds')
seurat_path = paste0(data.dir,'papalexi_results_seurat.rds')
seurat_CUL3_path = paste0(data.dir,'seurat_CUL3_results.rds')
seurat_CUL3_path_all = paste0(data.dir,'seurat_CUL3_results_all.rds')
seurat_CUL3_path_no_filter = paste0(data.dir,'seurat_CUL3_results_no_filter.rds')
#Note that sceptre results have columns pvalue, grna, target
#get sceptre perturbation on PDL1 mrna results
gene_result = readRDS(gene_path)
gene_result$log_fold_change = signif(gene_result$log_fold_change, digits=2)
#get sceptre perturbation on protein results
protein_result = readRDS(protein_path)
protein_result$log_fold_change = signif(protein_result$log_fold_change,digits=2)
#get seurat DE results. Columns 1,2,and 6 correspond to pvalue, effect size
#and perturbation
seurat_result = readRDS(seurat_path)
#change seurat to numeric
seurat_result$p_val = as.numeric(seurat_result$p_val)
seurat_result$avg_log2FC = as.numeric(seurat_result$avg_log2FC)

#change of base for seurat logfc
seurat_result$avg_log2FC = seurat_result$avg_log2FC * log(2,base = exp(1))
#round
#seurat_result$p_val = signif(seurat_result$p_val, digits=2)
#seurat_result$avg_log2FC = signif(seurat_result$avg_log2FC, digits=2)

#get seurat CUL3
seurat_CUL3_result = readRDS(seurat_CUL3_path)
seurat_CUL3_result_all = readRDS(seurat_CUL3_path_all)
seurat_CUL3_result_no_filter = readRDS(seurat_CUL3_path_no_filter)
```

Volcano Plots

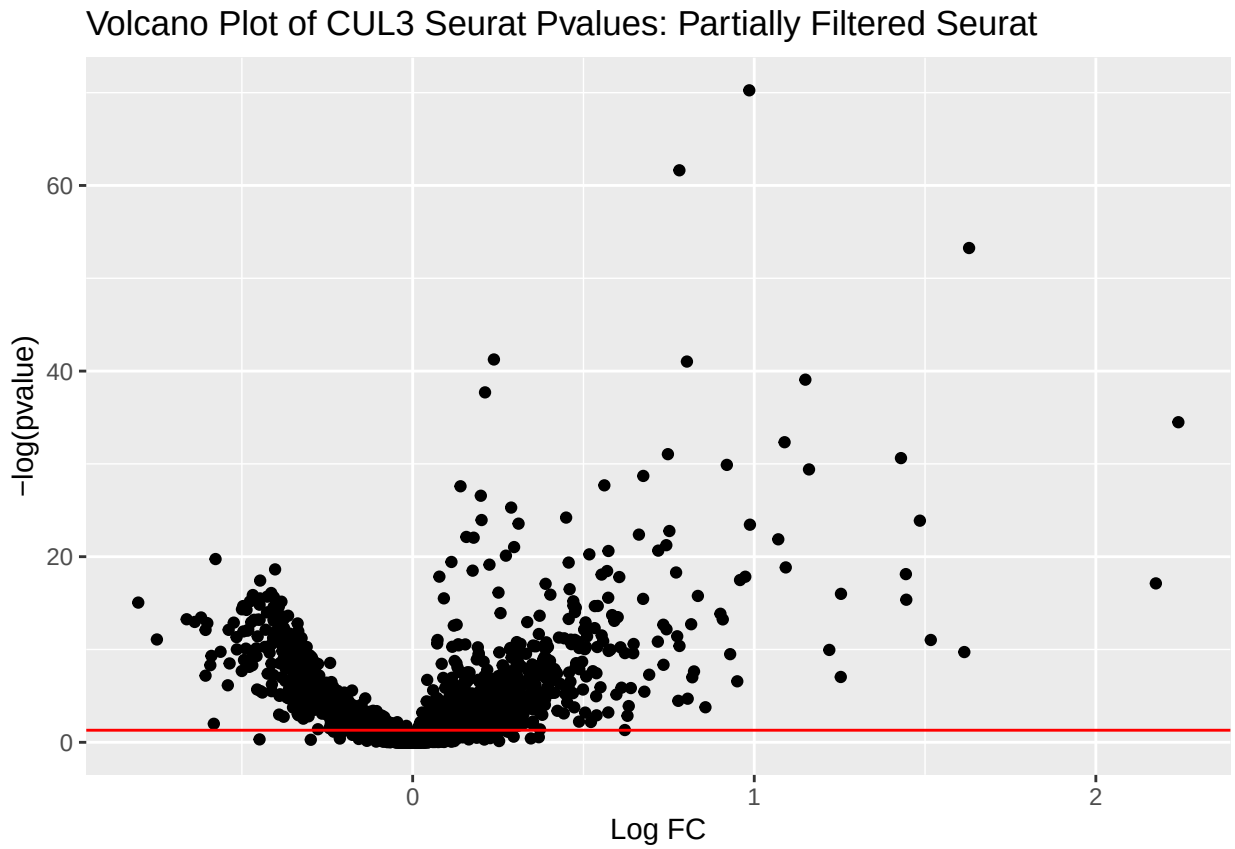
For Seurat Results With Full Filtering (min.pct = 0.1, min logfc = 0.25)

```
#get seurat data
seurat_data = subset(seurat_CUL3_result,select = c(p_val,avg_logFC))
seurat_data$p_val = -log(seurat_data$p_val,base = 10)
ggplot(seurat_data,aes(x = avg_logFC,y = p_val)) + geom_point() +
  ggtitle('Volcano Plot of CUL3 Seurat Pvalues: Fully Filtered Seurat') +
  labs(y = '-log(pvalue)',x = 'Log FC')+
  geom_hline(yintercept=-log(0.05,base = 10),color = 'red')
```



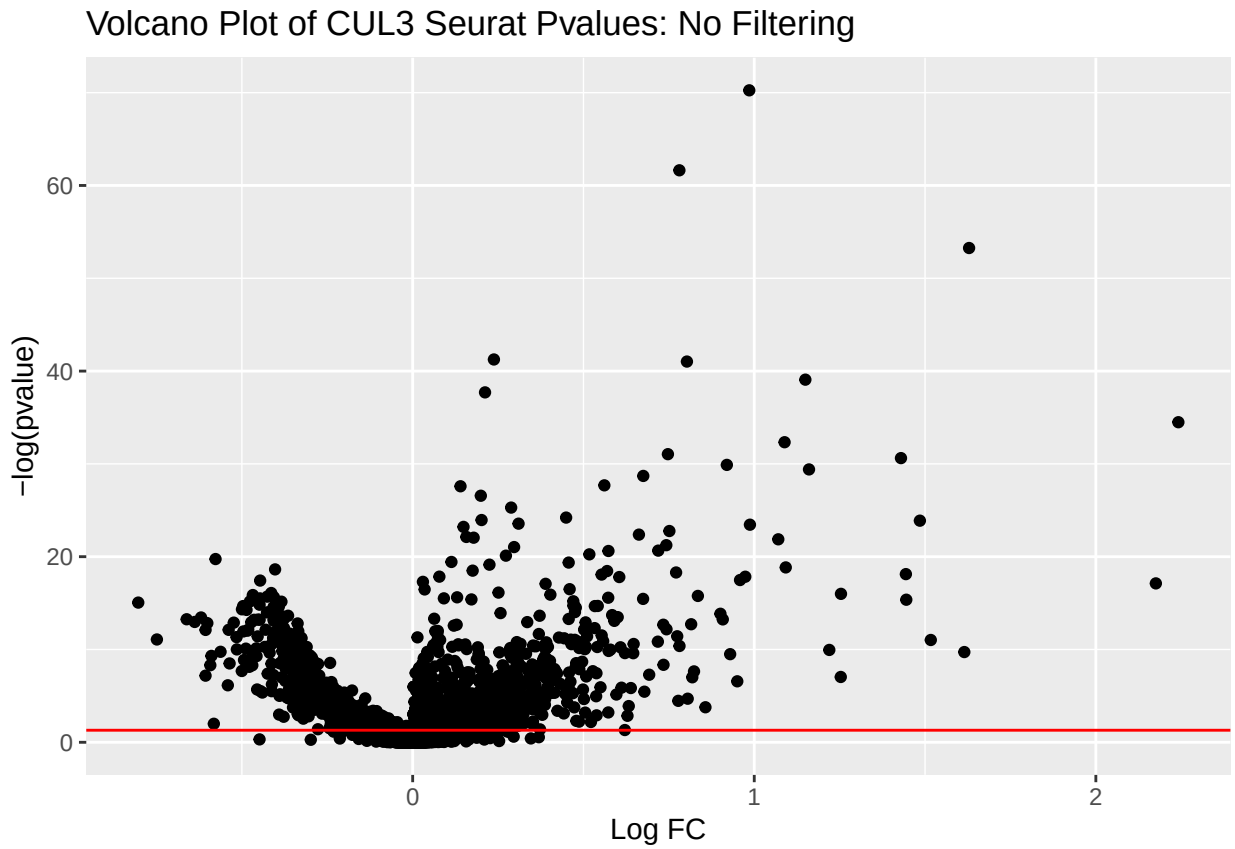
For Seurat Results With Percentage Filtering (min.pct = 0.1)

```
#get seurat data
seurat_data = subset(seurat_CUL3_result_all,select = c(p_val,avg_logFC))
seurat_data$p_val = -log(seurat_data$p_val,base = 10)
ggplot(seurat_data,aes(x = avg_logFC,y = p_val)) + geom_point() +
  ggtitle('Volcano Plot of CUL3 Seurat Pvalues: Partially Filtered Seurat') +
  labs(y = '-log(pvalue)',x = 'Log FC')+
  geom_hline(yintercept=-log(0.05,base = 10),color = 'red')
```



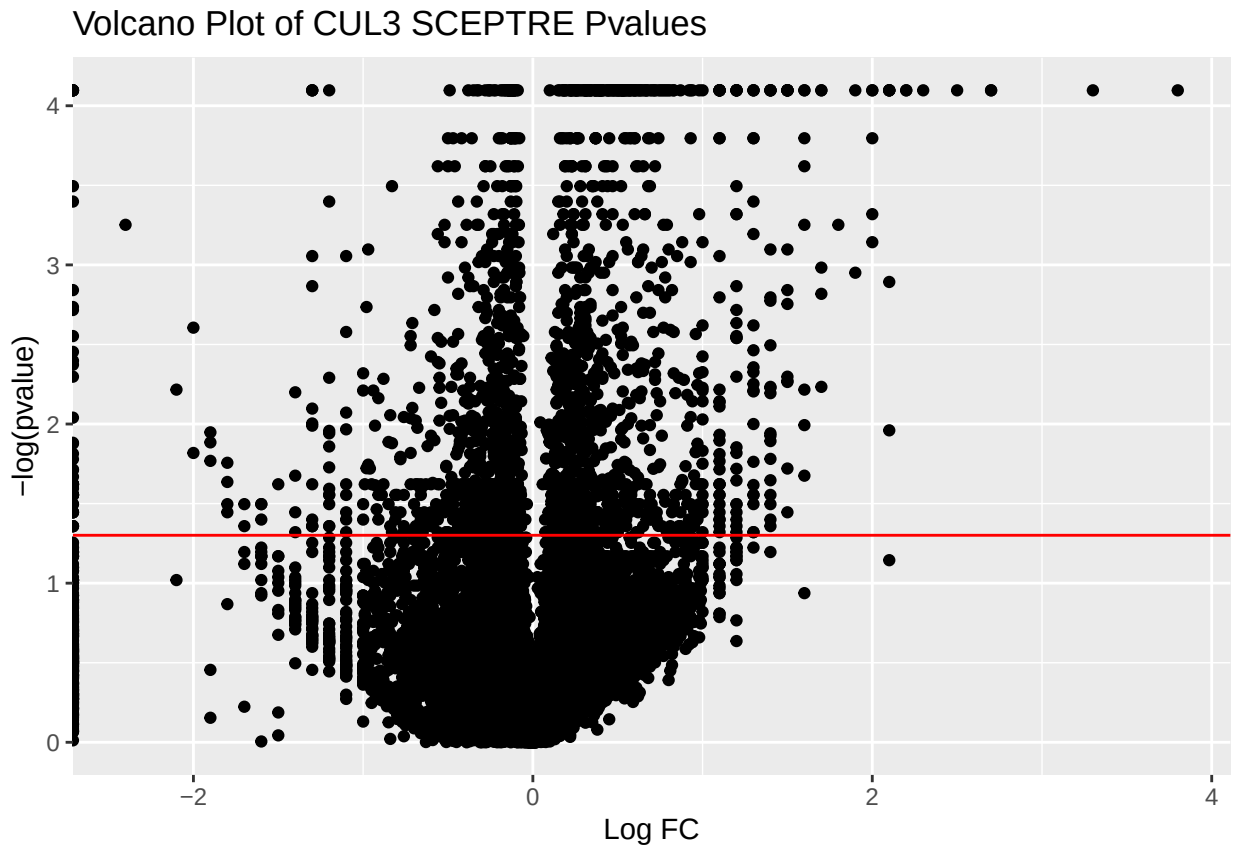
For Seurat Results Without Filtering

```
#get seurat data
seurat_data = subset(seurat_CUL3_result_no_filter,select = c(p_val,avg_logFC))
seurat_data$p_val = -log(seurat_data$p_val,base = 10)
seurat_data = subset(seurat_data,is.na(avg_logFC) == F & is.na(p_val) == F)
ggplot(seurat_data,aes(x = avg_logFC,y = p_val)) + geom_point() +
  ggtitle('Volcano Plot of CUL3 Seurat Pvalues: No Filtering') +
  labs(y = '-log(pvalue)',x = 'Log FC')+
  geom_hline(yintercept=-log(0.05,base = 10),color = 'red')
```



For SCEPTRE Results

```
#get seurat data
sceptre_data = subset(gene_result,select = c(p_value,log_fold_change))
#one point has log fold change that is NA
sceptre_data = subset(sceptre_data,is.na(log_fold_change) == F)
#take negative log
sceptre_data$p_value = -log(sceptre_data$p_value,base = 10)
#volcano plot
ggplot(sceptre_data,aes(x = log_fold_change,y = p_value)) + geom_point() +
  ggtitle('Volcano Plot of CUL3 SCEPTRE Pvalues') +
  labs(y = '-log(pvalue)',x = 'Log FC')+
  geom_hline(yintercept=-log(0.05,base = 10),color = 'red')
```



Identity Plots

Pvalue Plots and Log Fold Change Identity Plots: CUL3 vs All Genes: SCEPTRE vs Unfiltered Seurat

```
#match sceptre and seurat
sceptre_CUL3 = subset(gene_result,grna_group == "CUL3")
sceptre_CUL3 = subset(sceptre_CUL3,select =
                        c(response_id,p_value,log_fold_change))[-1,]
sceptre_CUL3 = subset(sceptre_CUL3,
                      response_id %in% rownames(seurat_CUL3_result_no_filter))
rownames(sceptre_CUL3) = unfactor(sceptre_CUL3$response_id)
seurat_CUL3_temp = seurat_CUL3_result_no_filter[
  rownames(seurat_CUL3_result_no_filter)%in%sceptre_CUL3$response_id,]

seurat_CUL3_temp = seurat_CUL3_temp[rownames(sceptre_CUL3),]
mean(rownames(sceptre_CUL3)==rownames(seurat_CUL3_temp))
```

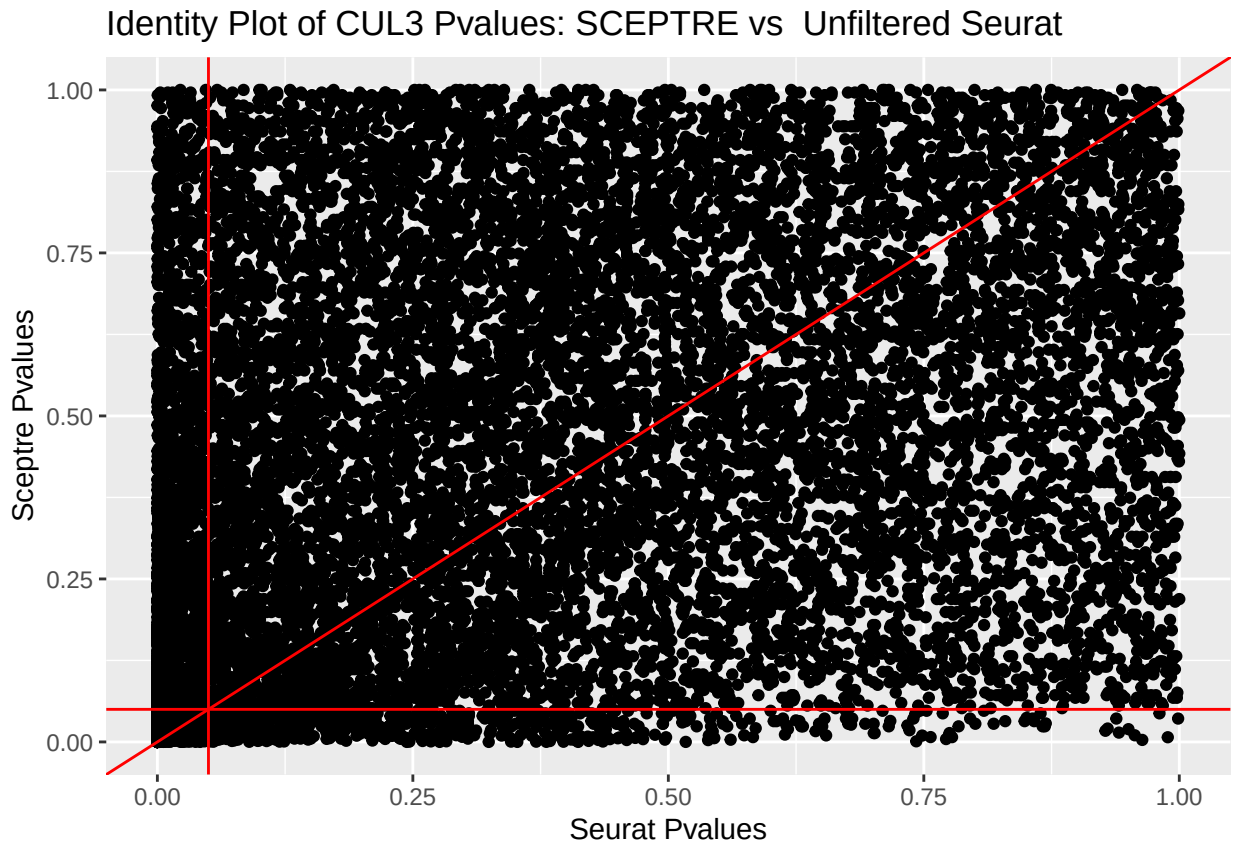
```
## [1] 1
```

```
#get pvalue and logFC for seurat data
seurat_CUL3_temp = subset(seurat_CUL3_temp,select = c(p_val,avg_logFC))
```

```
#get pvalue and log fold change for sceptre data
sceptre_CUL3 = subset(sceptre_CUL3,select =
                        c(p_value,log_fold_change))
```

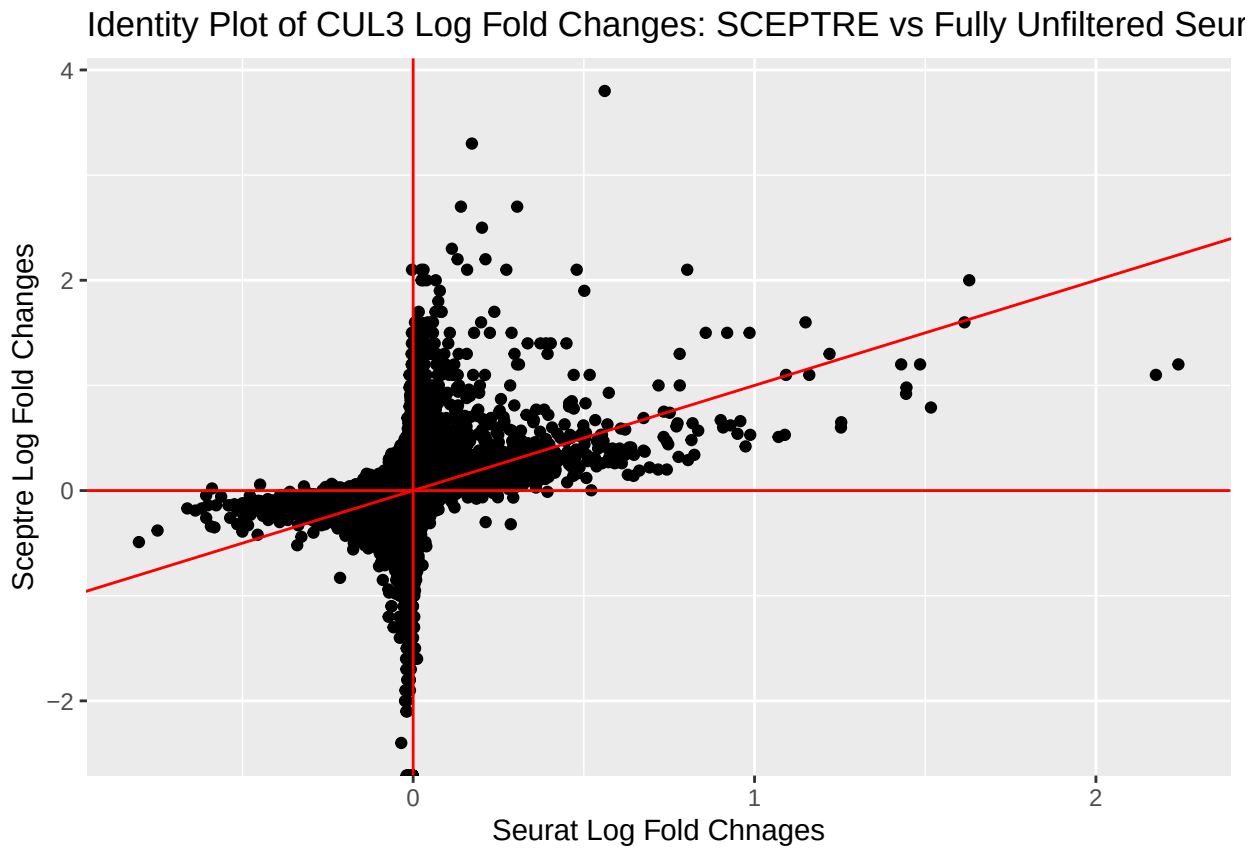
```
#merge pvalues
Pval = data.frame(seuratP = seurat_CUL3_temp$p_val,
                  sceptreP = sceptre_CUL3$p_value)
#merge log FC
LogFC = data.frame(seuratLogFC = seurat_CUL3_temp$avg_logFC,
                   sceptreLogFC = sceptre_CUL3$log_fold_change)
```

```
#volcano plot
ggplot(Pval,aes(x = seuratP,y = sceptreP)) + geom_point() +
  ggtitle('Identity Plot of CUL3 Pvalues: SCEPTRE vs  Unfiltered Seurat') +
  labs(y = 'Sceptre Pvalues',x = 'Seurat Pvalues')+
  geom_abline(slope=1, intercept = 0,color = 'red')+
  geom_vline(xintercept = 0.05,color = 'red')+
  geom_hline(yintercept = 0.05, color = 'red')
```




```
#volcano plot
ggplot(LogFC,aes(x = seuratLogFC,y = sceptreLogFC)) + geom_point() +
  ggtitle('Identity Plot of CUL3 Log Fold Changes: SCEPTRE vs Fully Unfiltered Seurat') +
  labs(y = 'Sceptre Log Fold Changes',x = 'Seurat Log Fold Chnages')+
  geom_abline(slope=1, intercept = 0,color = 'red')+
```

```
  geom_vline(xintercept = 0,color = 'red')+
  geom_hline(yintercept = 0, color = 'red')
```



Pvalue Plots and Log Fold Change Identity Plots: CUL3 vs All Genes: SCEPTRE vs Fully Filtered Seurat

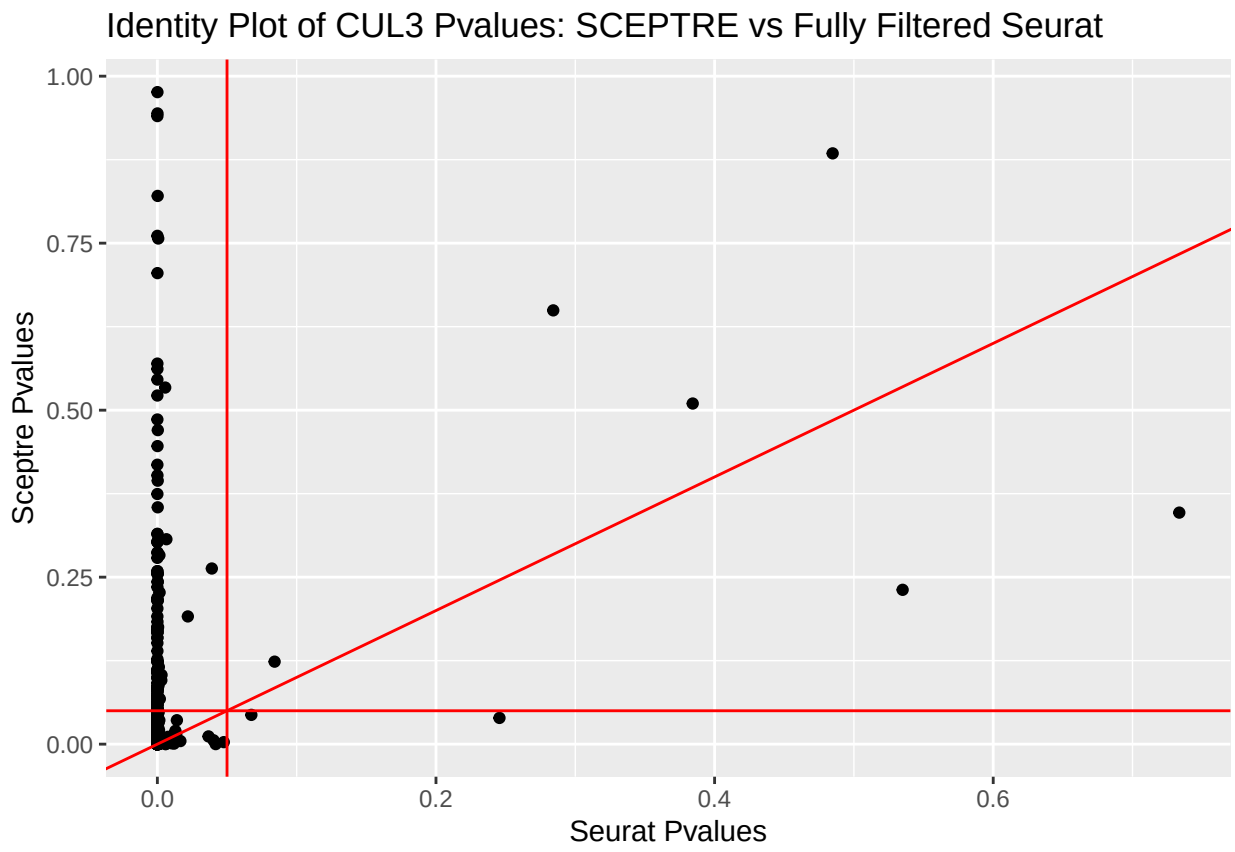
```
#match sceptre and seurat
sceptre_CUL3 = subset(gene_result,grna_group == "CUL3")
sceptre_CUL3 = subset(sceptre_CUL3,select =
                        c(response_id,p_value,log_fold_change))[-1,]
sceptre_CUL3 = subset(sceptre_CUL3,
                      response_id %in% rownames(seurat_CUL3_result))
rownames(sceptre_CUL3) = unfactor(sceptre_CUL3$response_id)
seurat_CUL3_temp = seurat_CUL3_result[
  rownames(seurat_CUL3_result)%in%sceptre_CUL3$response_id,]

seurat_CUL3_temp = seurat_CUL3_temp[rownames(sceptre_CUL3),]
mean(rownames(sceptre_CUL3)==rownames(seurat_CUL3_temp))
```

```
## [1] 1
```

```
#get pvalue and logFC for seurat data
seurat_CUL3_temp = subset(seurat_CUL3_temp,select = c(p_val,avg_logFC))
#get pvalue and log fold change for sceptre data
sceptre_CUL3 = subset(sceptre_CUL3,select =
                        c(p_value,log_fold_change))
```

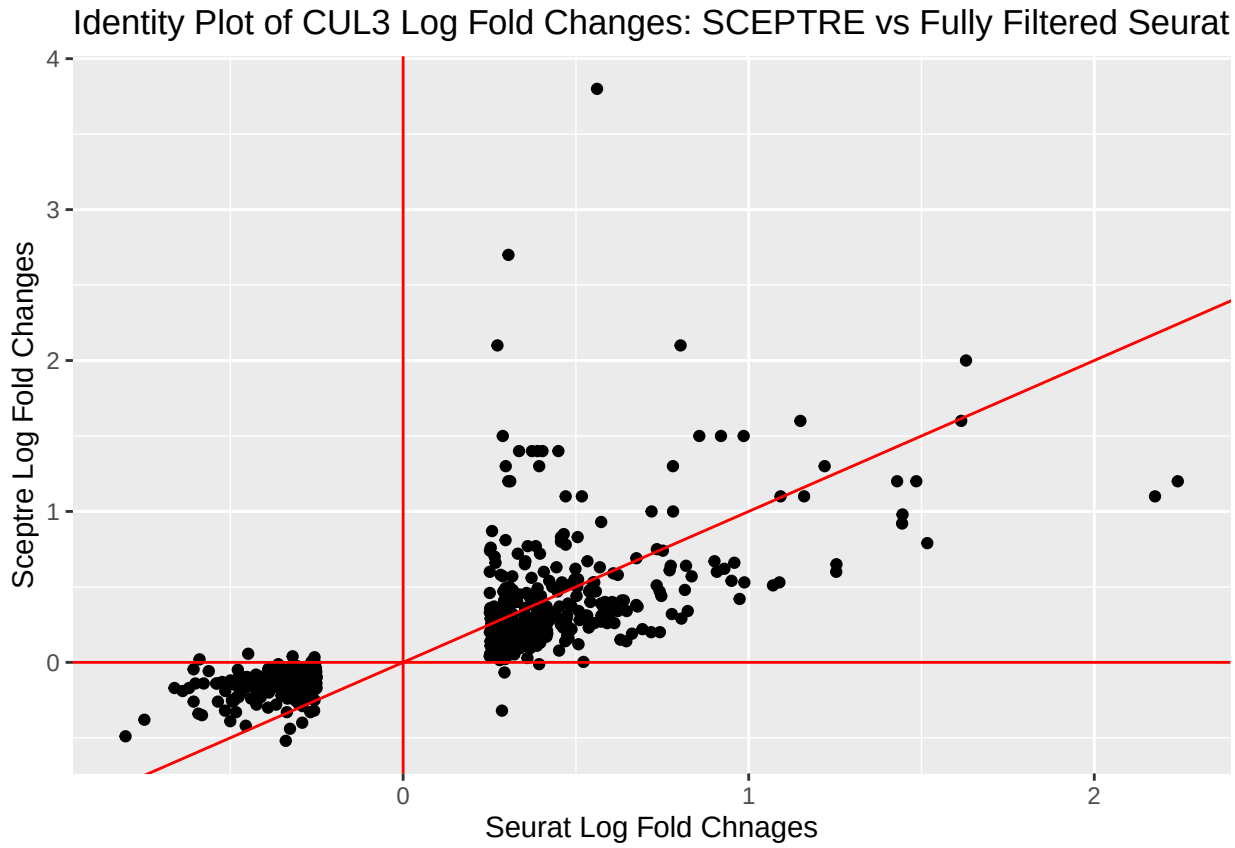
```
#merge pvalues
Pval = data.frame(seuratP = seurat_CUL3_temp$p_val,
                  sceptreP = sceptre_CUL3$p_value)
#merge log FC
LogFC = data.frame(seuratLogFC = seurat_CUL3_temp$avg_logFC,
                   sceptreLogFC = sceptre_CUL3$log_fold_change)
```

```
#volcano plot
ggplot(Pval,aes(x = seuratP,y = sceptreP)) + geom_point() +
  ggtitle('Identity Plot of CUL3 Pvalues: SCEPTRE vs Fully Filtered Seurat') +
  labs(y = 'Sceptre Pvalues',x = 'Seurat Pvalues')+
  geom_abline(slope=1, intercept = 0,color = 'red')+
  geom_vline(xintercept = 0.05,color = 'red')+
  geom_hline(yintercept = 0.05, color = 'red')
```




```
#volcano plot
ggplot(LogFC,aes(x = seuratLogFC,y = sceptreLogFC)) + geom_point() +
  ggtitle('Identity Plot of CUL3 Log Fold Changes: SCEPTRE vs Fully Filtered Seurat') +
  labs(y = 'Sceptre Log Fold Changes',x = 'Seurat Log Fold Chnages')+
```

```
  geom_abline(slope=1, intercept = 0,color = 'red')+
  geom_vline(xintercept = 0,color = 'red')+
  geom_hline(yintercept = 0, color = 'red')
```



Pvalue Plots and Log Fold Change Identity Plots: SCEPTRE vs Seurat on Protein

```
#match sceptre and seurat
sceptre_protein = protein_result
seurat_protein = subset(seurat_result,Target %in% c("PDL1","CD86","PDL2",
                                                    "CD366"))
```

```
#match perturbations for CD86
seurat_CD86 = seurat_protein[seurat_protein$Target == "CD86",]
sceptre_CD86 = sceptre_protein[sceptre_protein$response_id == 'CD86']
sceptre_CD86$grna_group = unfactor(sceptre_CD86$grna_group)
sceptre_CD86 = subset(sceptre_CD86,grna_group%in%seurat_CD86$PRTB)
sceptre_CD86 = sceptre_CD86[order(sceptre_CD86$grna_group),]
seurat_CD86 = seurat_CD86[order(seurat_CD86$PRTB),]
```

```
#match perturbations for PDL1
seurat_PDL1 = seurat_protein[seurat_protein$Target == "PDL1",]
```

```
sceptre_PDL1 = sceptre_protein[sceptre_protein$response_id == 'PDL1']
sceptre_PDL1$grna_group = unfactor(sceptre_PDL1$grna_group)
sceptre_PDL1 = subset(sceptre_PDL1,grna_group%in%seurat_PDL1$PRTB)
sceptre_PDL1 = sceptre_PDL1[order(sceptre_PDL1$grna_group),]
seurat_PDL1 = seurat_PDL1[order(seurat_PDL1$PRTB),]
```

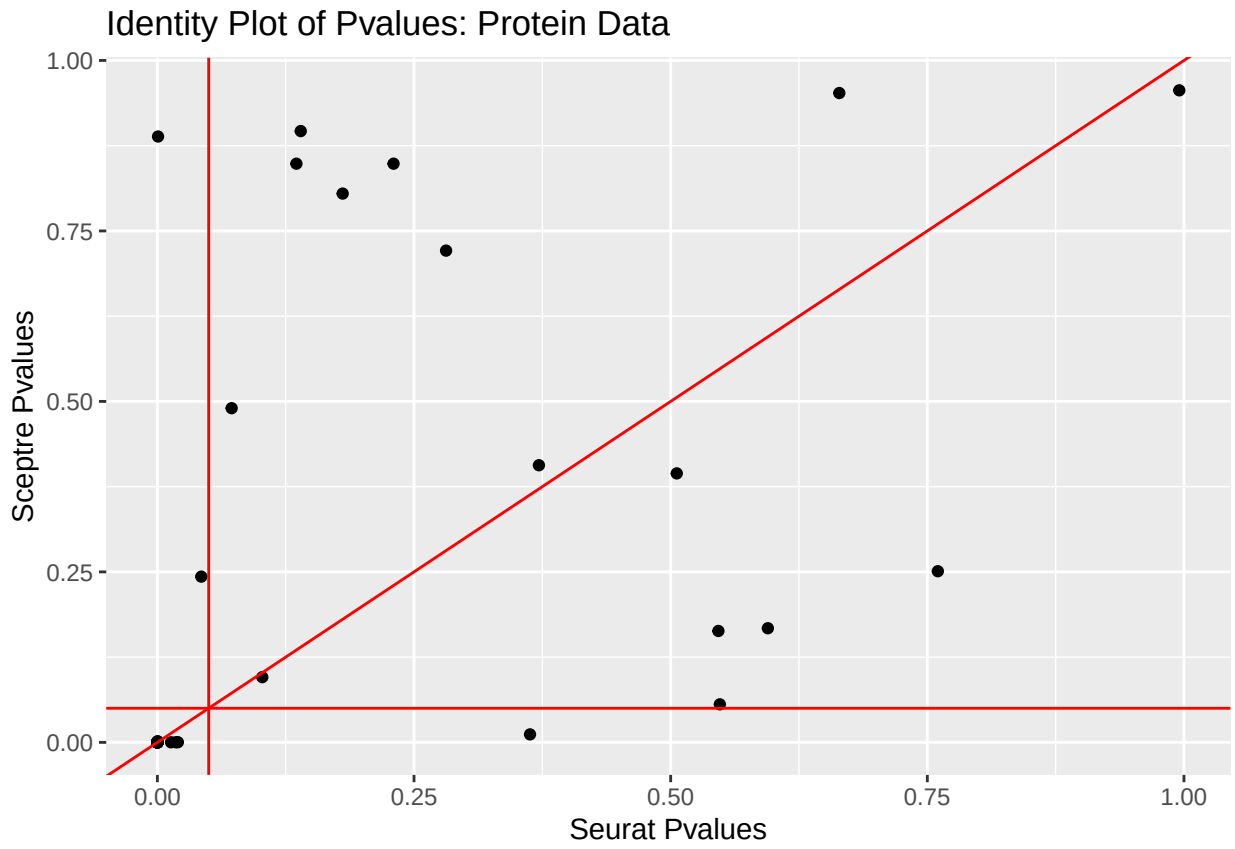
```
#match perturbations for PDL2
seurat_PDL2 = seurat_protein[seurat_protein$Target == "PDL2",]
sceptre_PDL2 = sceptre_protein[sceptre_protein$response_id == 'PDL2']
sceptre_PDL2$grna_group = unfactor(sceptre_PDL2$grna_group)
sceptre_PDL2 = subset(sceptre_PDL2,grna_group%in%seurat_PDL2$PRTB)
sceptre_PDL2 = sceptre_PDL2[order(sceptre_PDL2$grna_group),]
seurat_PDL2 = seurat_PDL2[order(seurat_PDL2$PRTB),]
```

```
#match perturbations for CD366
seurat_CD366 = seurat_protein[seurat_protein$Target == "CD366",]
sceptre_CD366 = sceptre_protein[sceptre_protein$response_id == 'CD366']
sceptre_CD366$grna_group = unfactor(sceptre_CD366$grna_group)
sceptre_CD366 = subset(sceptre_CD366,grna_group%in%seurat_CD366$PRTB)
sceptre_CD366 = sceptre_CD366[order(sceptre_CD366$grna_group),]
seurat_CD366 = seurat_CD366[order(seurat_CD366$PRTB),]
```

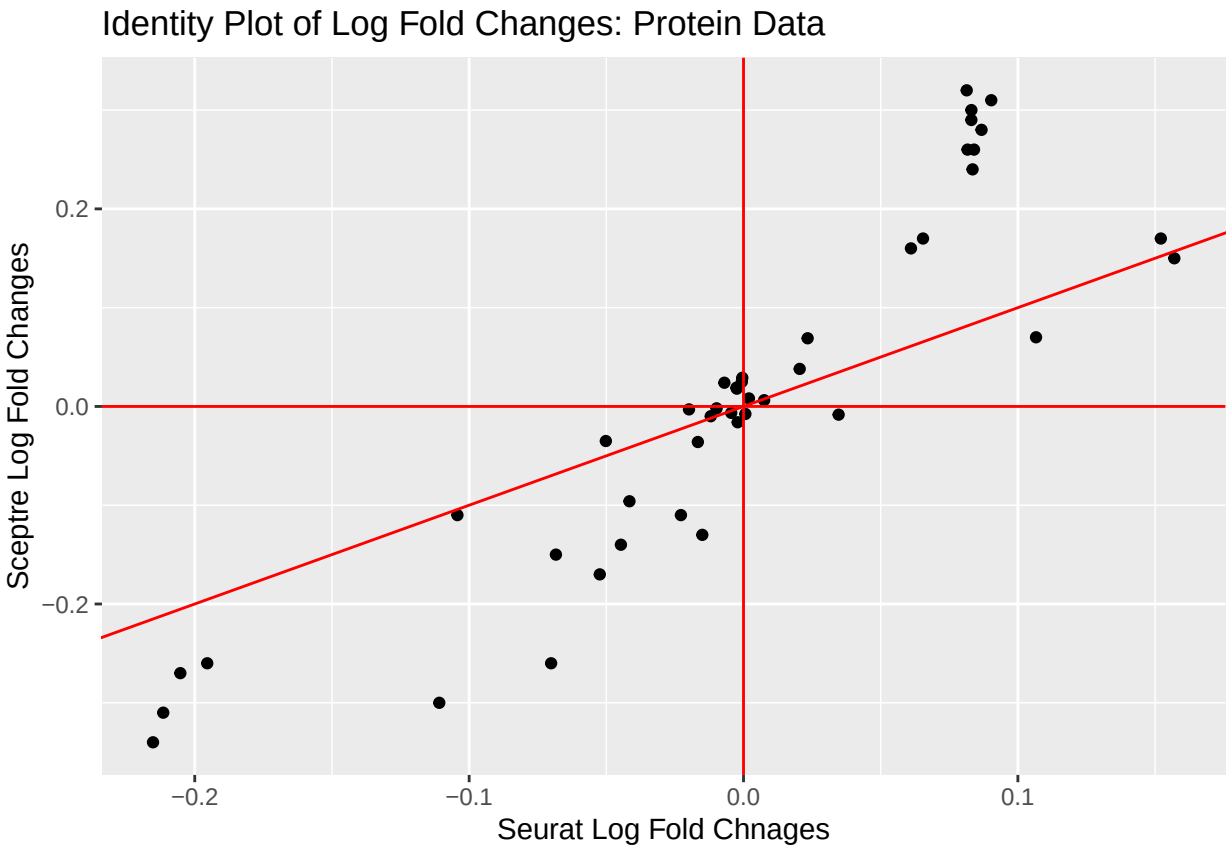
```
#get all data
seurat_protein_all = rbind(seurat_PDL1,seurat_PDL2,seurat_CD86,seurat_CD366)
sceptre_protein_all = rbind(sceptre_PDL1,sceptre_PDL2,sceptre_CD86,
                            sceptre_CD366)
```

```
#merge pvalues
Pval = data.frame(seuratP = seurat_protein_all$p_val,
                  sceptreP = sceptre_protein_all$p_value)
#merge log FC
LogFC = data.frame(seuratLogFC = seurat_protein_all$avg_log2FC,
                   sceptreLogFC = sceptre_protein_all$log_fold_change)
```

```
#volcano plot
ggplot(Pval,aes(x = seuratP,y = sceptreP)) + geom_point() +
  ggtitle('Identity Plot of Pvalues: Protein Data') +
  labs(y = 'Sceptre Pvalues',x = 'Seurat Pvalues')+
  geom_abline(slope=1, intercept = 0,color = 'red')+
  geom_vline(xintercept = 0.05,color = 'red')+
  geom_hline(yintercept = 0.05, color = 'red')
```




```
#volcano plot
ggplot(LogFC,aes(x = seuratLogFC,y = sceptreLogFC)) + geom_point() +
  ggtitle('Identity Plot of Log Fold Changes: Protein Data') +
  labs(y = 'Sceptre Log Fold Changes',x = 'Seurat Log Fold Chnages')+
  geom_abline(slope=1, intercept = 0,color = 'red')+
  geom_vline(xintercept = 0,color = 'red')+
  geom_hline(yintercept = 0, color = 'red')
```



Summary

Volcano Plots

It is apparent that Seurat's Pvalues and Log Fold change estiamtes are more correlated than that of SCEPTRE. I think the reason that the Seurat Pvalue and Seurat Log Fold change correaltion weakens as you remove filters is because the genes included seem to be all have significant pvalues and small log fold changes.

Identity Plots

Pvalues

There seems to be almost no agreement between SCEPTRE and Seurat pvalues. This can be seen by the almost uniform Identity plot of Seurat Pvalues vs SCEPTRE Pvalues when there is no filtering. When there is filtering, the Seurat pvalues seem to be much smaller than the SCEPTRE pvalues.

Log Fold Change

Log Fold change estimates seem to be in agreement generally as far as sign goes. With no filtering, there are a decent amount of estimates whose signs differ. This is seen by some points being in the third quadrant. However, when applying the Seurat Filters, these points disappear, suggesting that seurat estimates for genes with small changes may be driven by technical factors. We see that SCEPTRE's estimates tend to be

smaller in magnitude than Seurat's estimates. This can be seen by most points being above the 45 degree line in the first quadrant and below it in the third quadrant.

Next Steps

There is a big discepancy between Seurat and SCEPTRE Pvalue estimation. This should be investigated further before continuing any further with comparisons between the two methods.