Statistics in Genomics
Final Project
Group Member: Zhanmo Ni and Tanxin Liu

Introduction

We studied part of the experiment from Luperchio, Boukas et al.Leveraging the Mendelian disorders of the epigenetic machinery to systematically map functional epigenetic variation" (1). Specific epigenetic changes may causally mediate phenotypes through the alteration states. Recently, Mendelian Disorders of the Epigenetic Machinery (MDEMs) caused by coding variants disrupting to loss-of-function variation(2), can shed light on the causal relationship between epigenetic/transcriptome variation and diseases(3). Kabuki syndrome type 1 (KS 1) and type 2 (KS2) are widely studied MDEMs, caused by haploinsufficiency in histone methyltransferase *KMT2B* and *KDM6A* (1). Rubinstein-Taybi type 1 (RT or RT1) shares phenotypes with KS but is clinically distinct (caused by haploinsufficiency in histone acetyltransferase *CREBBP*) (1).

As part of the experiment, we compared wild-type mice to mice with a loss-of-function variant in the gene *CREBBP* causing Rubenstein-Taybi's syndrome (RT or RT1). We would like to test whether evidence that (1) a set of genes/loci are differentially expressed between B-cell vs T-cell within wild-type mice. (2) are differentially expressed genes between mutant and wild-type within B-cells.

Methods

A main task in the analysis of the count data from RNA-seq is the detection of differentially expressed genes among B cells versus T cells and mutant type vs. wild type. We had 26 samples with different genotypes and cell types. (Exploratory analyses include visualizing the sample difference by principal components). Linear regression models are performed to compare the counts of genes between different samples using DESeq2, accounting for depth.We plotted MA-plot to show the log2 fold changes attributable to the difference(differences between mutant versus wild-type (reference) within B cells; difference between T cells versus B cells (reference)) over the mean of normalized counts. In addition, we also performed surrogate variable analysis(SVA) to account for potential confounders, such as batch effects.

Results

T cells vs. B cells within wildtype:

We ran the differential expression analysis comparing gene expressions in the B cells and T cells among the 14 wild type samples, which includes seven samples for each cell type. After filtering for genes with mean counts less than one over the 14 wild type samples, 16,646

genes were left in our analysis. We used an FDR < 0.1 as a threshold to account for multiple testing and we identified that 11,309 genes were differentially expressed in the T cells compared to B cells. Among the 11,309 genes, 5,284 genes were expressed more in B cells and only 11 of them had a log fold change (base of 2) larger than 5. On the other hand, within the 6,025 genes expressed more in T cells, 487 had a log fold change larger than 5 and expression of ENSMUSG00000040498 (mean of normalized counts = 146) had a log fold change larger than 10, which is in the immunoglobulin superfamily (Figure.1). The p-values were distributed uniformly except the left tail (Figure.2).

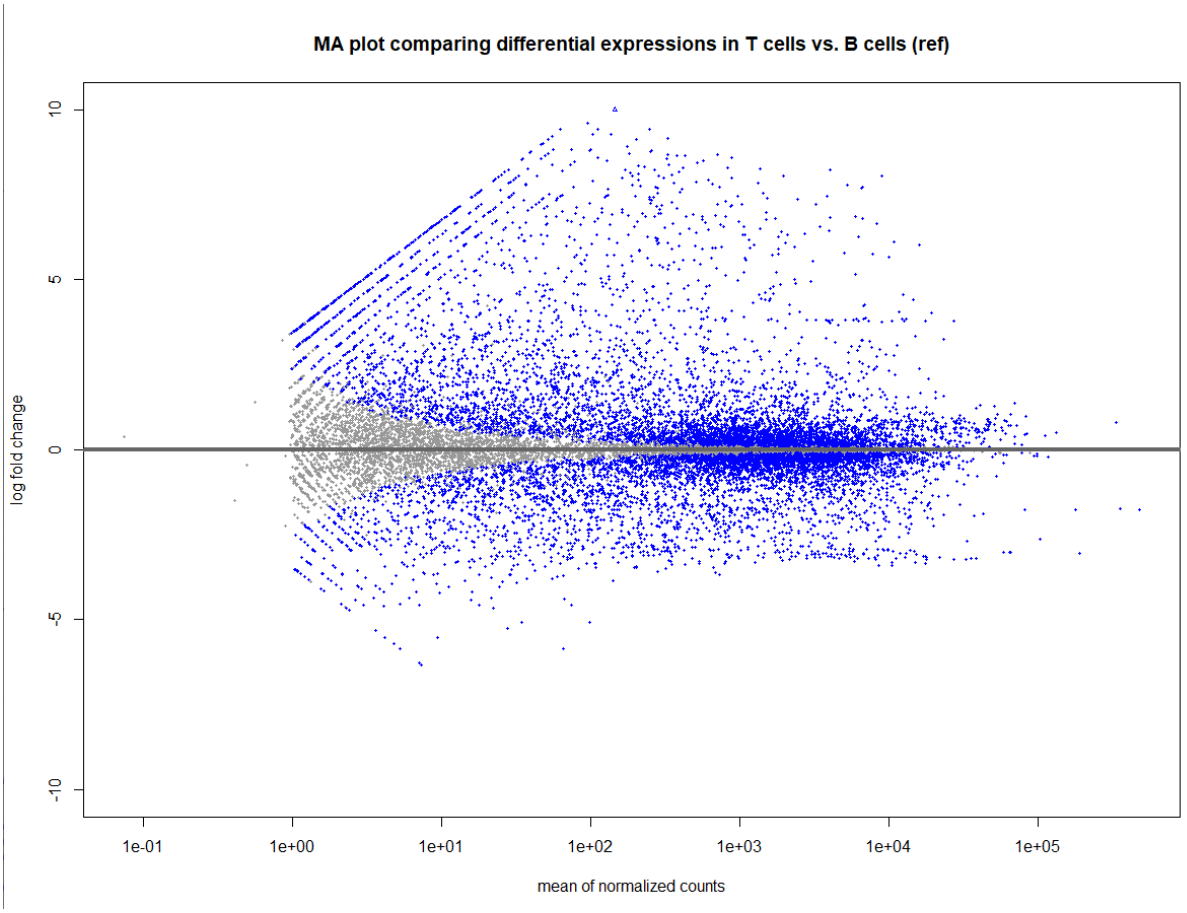

Figure 1. MA plot comparing differential expressions in T cells vs. B cells (ref).
(Points will be colored blue if the adjusted p value is less than 0.1. Points which fall out of the window are plotted as open triangles pointing either up or down)

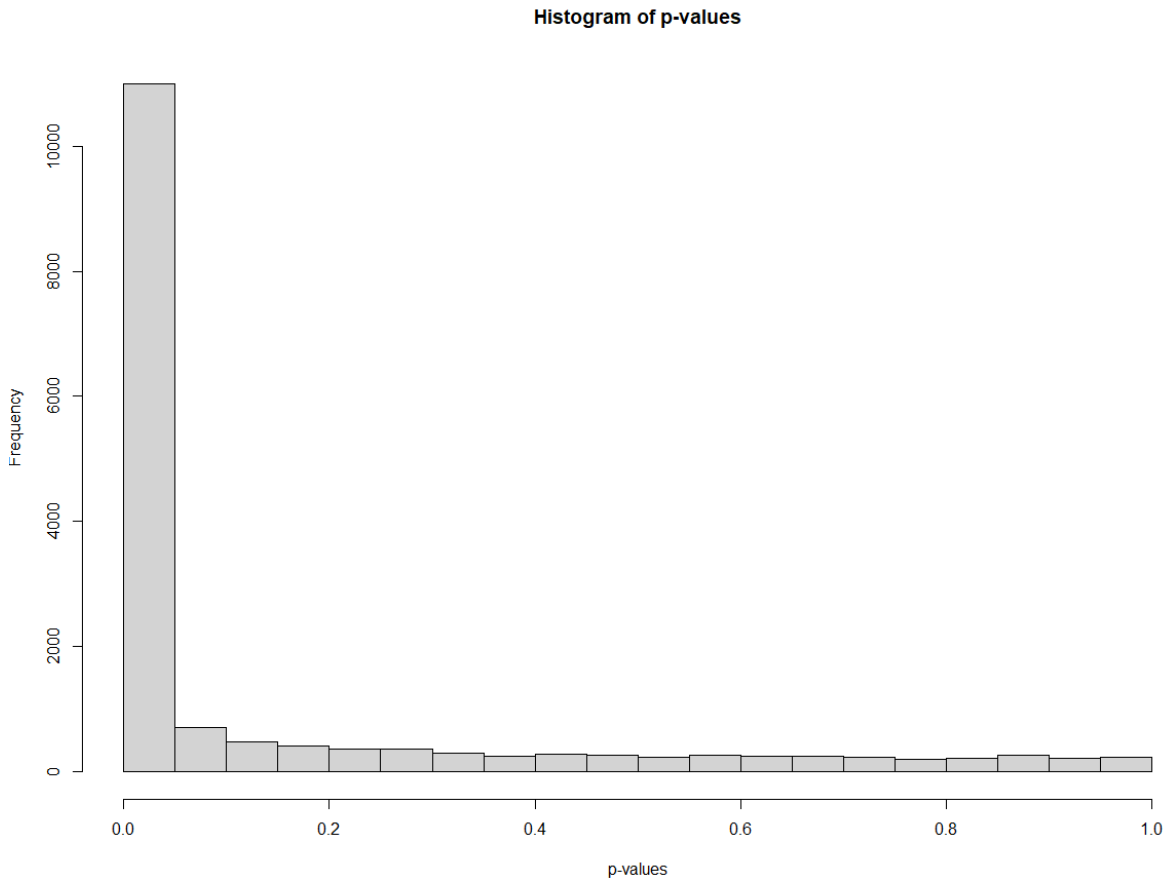

Figure 2. Histogram of p-values of differential expression analysis comparing T cells and B cells.

According to the following SVA, after adjusting for three surrogate variables, 11,098 genes were expressed significantly differentially (FDR < 0.1). 5,178 genes were expressed more in B cells and 5,920 genes were expressed more in T cells. We still noticed the pattern that there were more genes with greater magnitude of expression change (log2 fold change > 5) among those expressed more in T cells (Figure 3.). Besides ENSMUSG00000040498, three more genes, ENSMUSG00000015709, ENSMUSG00000096417 and ENSMUSG00000079733, have a log2 fold change greater than 10, but all of them have very limited base mean. Overall, the gene expression of T cells and B cells among wildtype samples are very different, with many gene expressions much higher in T cells.

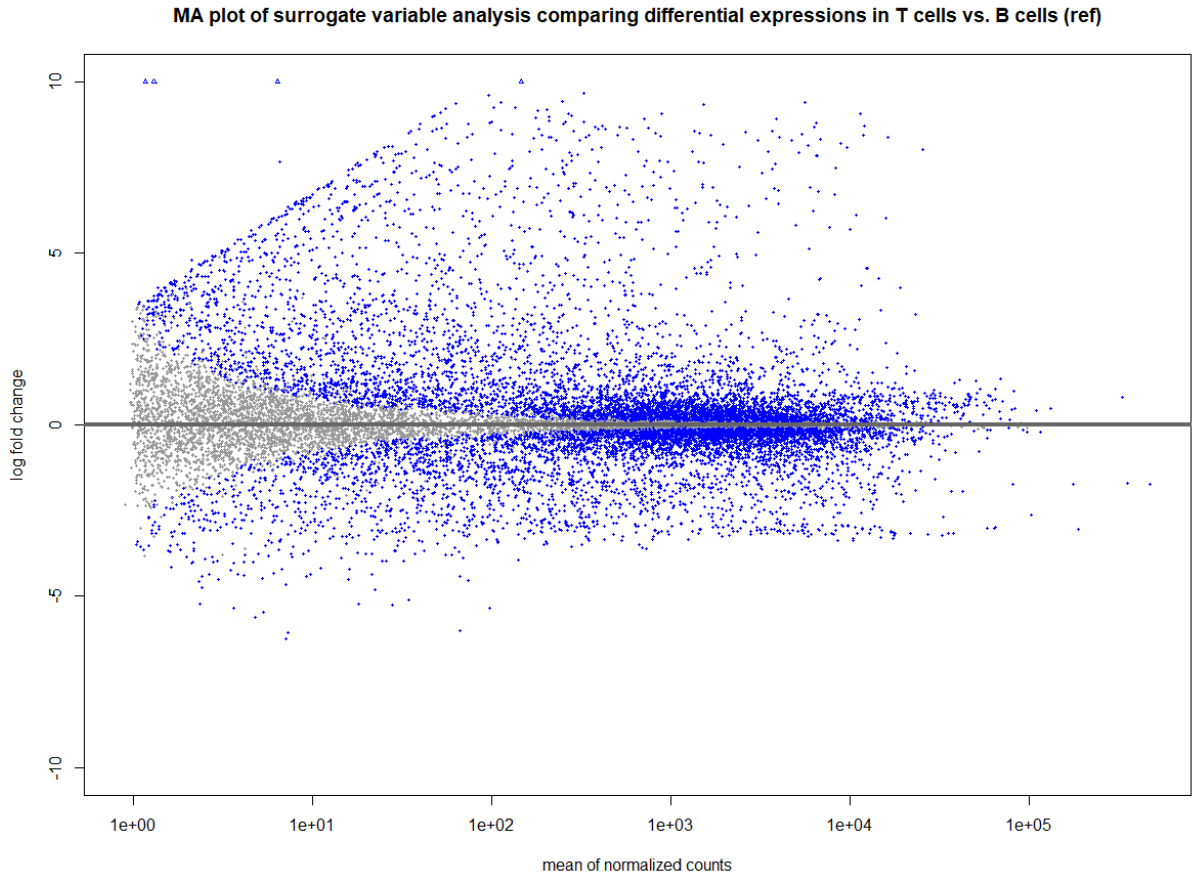

Figure 3. MA plot of SVA comparing differential expression in T cells vs. B cells in wild type samples

According to the principal component analysis, the first principal component explained 99% of the variance, which is uncommon. However, the results are acceptable considering that the genes were expressed very differentially in the two cell types (Figure 4.).

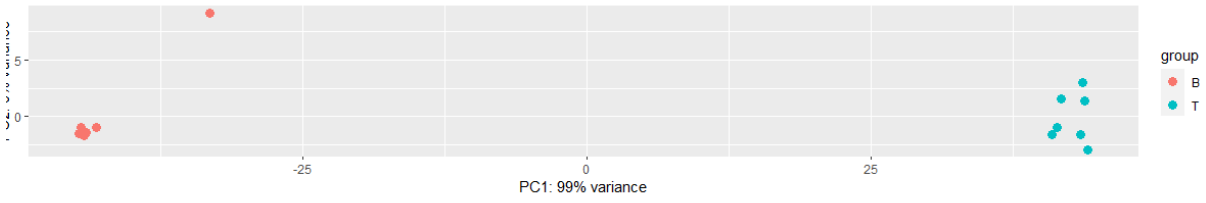

Figure 4. PC plots of the differential expression in T cells vs. B cells in wild type samples

Mutant vs. wild-type within B cells:

We followed with the differential expression analysis comparing gene expressions in mutant and wild-type cells among the 12 B cells, which include six samples for each cell type. After filtering for genes with mean counts less than one over the 12 B cell samples, 16,121 genes were left in our analysis. We used FDR <0.1 as a threshold to account for multiple testing.

We identified 2,149 genes that were differentially expressed in the mutant cells compared to wild-type. Among the 2,149 genes, 976 were expressed more in mutant cells and only 24 of them had a log fold change (base of 2) larger than 2. (None of them are larger than 5).ENSMUSG00000030516 (mean of normalized counts = 12.6) had a log fold change larger than 4 (log fold change= 4.15988), which is the tight junction protein 1 (Figure.5) .On the other hand, within the 1,173 genes expressed more in wild-type cells, 3 had a log fold change larger than 2 (ENSMUSG00000061414, ENSMUSG00000063430, ENSMUSG00000069972). Expression of ENSMUSG00000069972 (mean of normalized counts = 26.8) had a log fold change larger than 5, which is the ribosomal protein S13, pseudogene 2 (Figure.5). The p-values were distributed uniformly except the left tail (Figure.6).

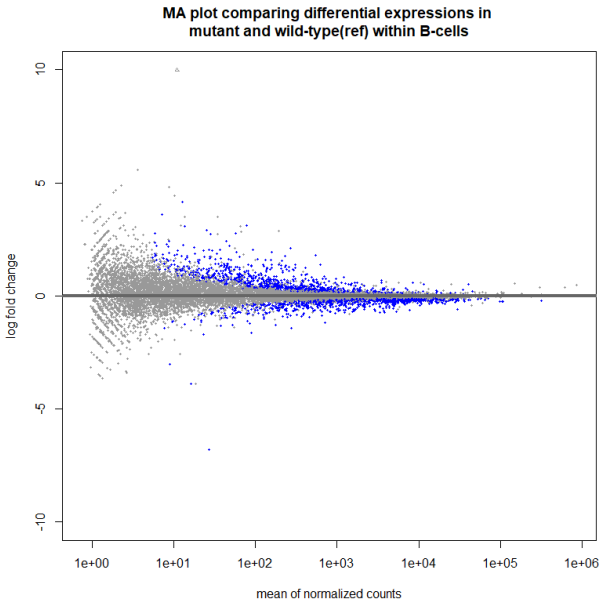

Figure 5. MA plot comparing differential expression in mutant vs. wild-type in B cells

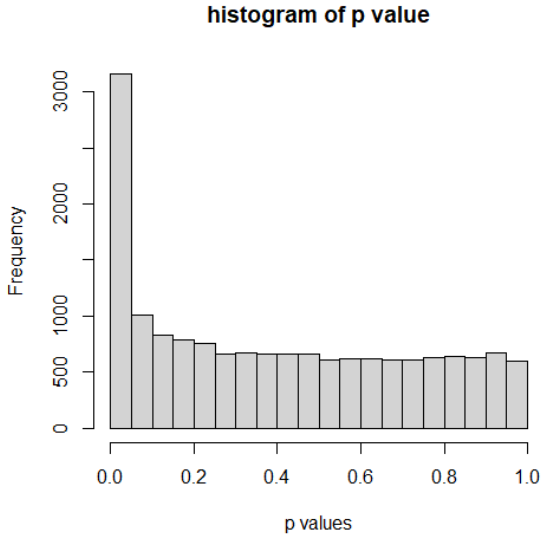

Figure 6. Histogram of p-values of differential expression analysis comparing mutant and wildtype within B cells.

According to the following SVA, after adjusting for two surrogate variables, 1,844 genes were expressed significantly differentially (FDR < 0.1). 840 genes were expressed more in mutant cells and 1004 genes were expressed more in wildtype cells. We still notice the pattern that there were more genes with greater magnitude of expression change (log2 fold change >2) among those expressed more in mutant cells than wildtype cells (mutant : 23 vs wildtype: 4) (Figure 7.) Among the 840 genes expressed more in mutant cells, ENSMUSG00000083596 (mean of normalized counts = 16.6) had a log fold change larger than 5 (log fold change: 12.36894).

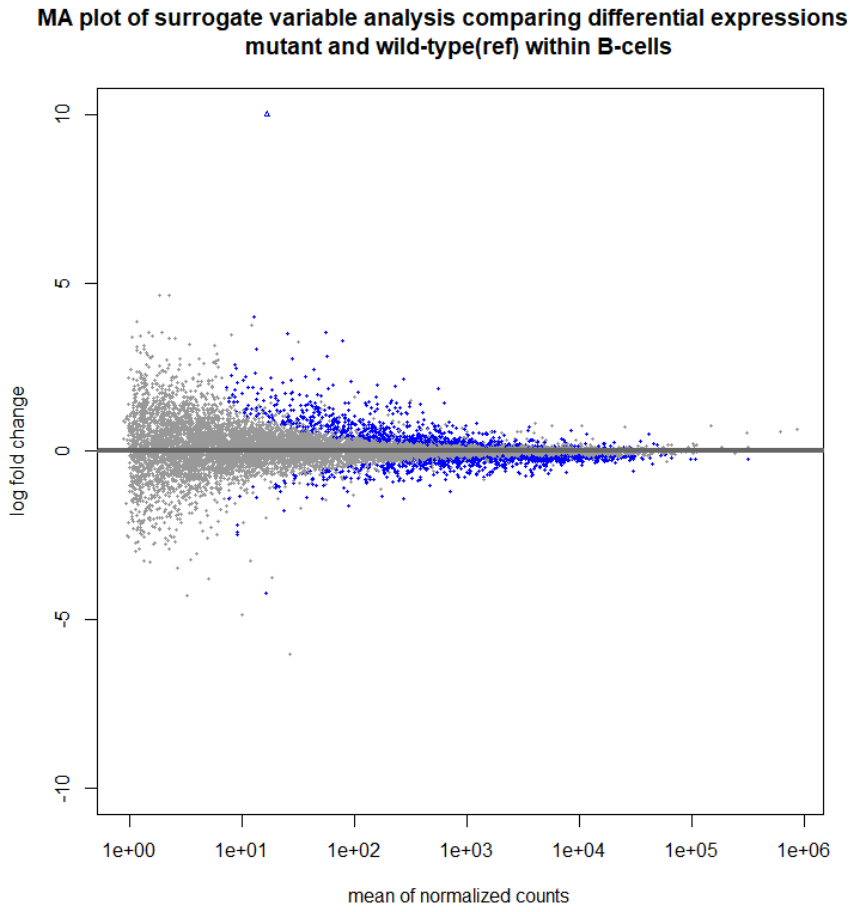

Figure 7. MA plot of SVA comparing differential expressions in mutant and wild-type in B cells

Principal component analysis shows the clustering of the samples based on gene expression and several genes that contribute most to the clustering. The accessibility signal separates the mutant genotypes from their wild-type littermates. The first principal component explained 27% of variance and the second principal component explained 20% of variance (Figure 8.).

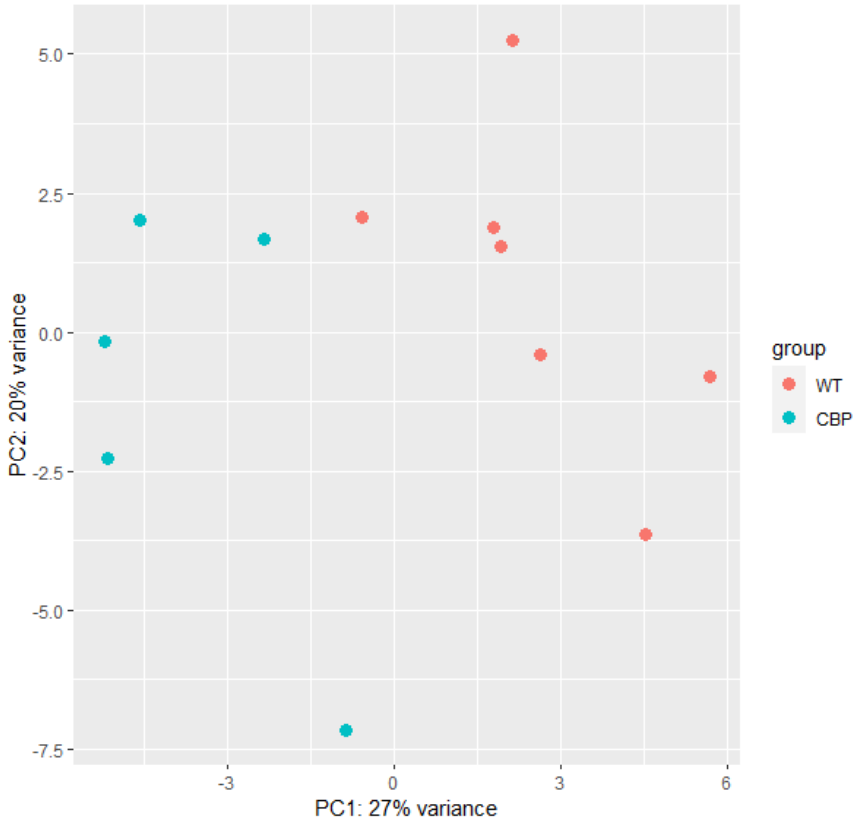

Figure 8. PC plots of the differential expression in mutant vs. wild type samples in B cells

Conclusions

To conclude, we can tell that cell type differences are much larger than mutant-wildtype differences, in terms of both magnitudes of changes and number of genes significantly differentially expressed. More than 3,000 thousand genes expressed differentially comparing different T cells and B cells within wildtype samples with a stringent FDR of 1*10^-20. However, with the same threshold, only 19 genes were identified comparing mutant and wildtype B cell samples (Figure 9.).Comparing wild-type mice to mice with a loss-of-function variant in the gene *CREBBP* causing RT syndrome, there is a greater amount of differentially expressed genes between mutant and wild-type within B-cells as compared with the amount of differentially expressed genes B-cell vs T-cell within wild-type mice.

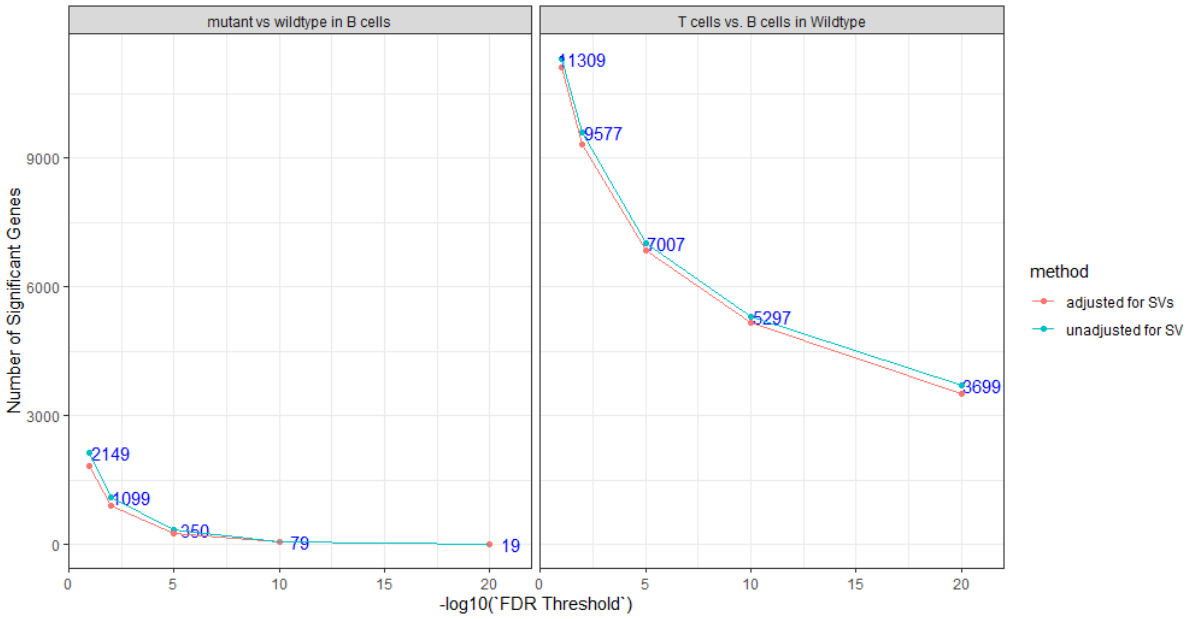

Figure 9. Number of Genes differentially expressed with different FDR threshold

References:

1. Luperchio TR, Boukas L, Zhang L, et al. Leveraging the Mendelian disorders of the epigenetic machinery to systematically map functional epigenetic variation. *Elife*. 2021;10:e65884. Published 2021 Aug 31. doi:10.7554/eLife.65884

2. Fahrner, Jill A., and Hans T. Bjornsson. "Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects." *Human Molecular Genetics* 28.R2 (2019): R254-R264.

3. Boukas, Leandros et al. “Coexpression patterns define epigenetic regulators associated with neurological dysfunction.” *Genome research* vol. 29,4 (2019): 532-542. doi:10.1101/gr.239442.118