RNA-Seq analysis assignment write-up
Xinyi Zhang CID: 02020037

Q.1 (word limit: 300 words; 25%)
For what purposes can a PCA result be used in RNA-seq data analysis? Explain two different uses of PCA outputs.

The PCA result can be applied in summarizing the resulting pattern. It allows a reduction in the number of dimensions required to be analyzed and thus reduces computational complexity. By just analyzing the first two to three principal components (PC), the sample plot PCA plots shows the sample distribution and presentation of data structure. Therefore, similarities and differences between the samples can be analyzed. If samples from different experimental groups are separated in the plot, this can give a hint that there are variations in gene expression in different experimental groups. Also, gene loading plot can provide information about the position of each gene or each sample in the space. By comparing the sample plot and the gene loading, the position of a cluster in the space can be identified. In addition, it is vital to check the scree plot to see how many percentages of the variations are captured in each PC. However, the first three PCs usually contain the most variations.

PCA can also contribute to the quality control of the experiment. The outliers can be identified by checking the sample plot. If one gene or sample deviates from the area where other genes or samples from the same experimental group stay, that gene or sample is likely to be an outlier, and further analysis, such as hierarchical clustering can be used to investigate the reasons. The potential outliers might indicate fundamental biological mechanisms but can also be caused by contamination or other mishandling during the experiment. In addition, PCA analysis can monitor the batch effect. If genes or samples from different groups cluster together but do not conform to biological principles, it indicates the batch effect. PCA can therefore detect and flag potential non-biological mistakes in the experiment. Thus, the PCA can ensure that the following analysis is based on a reliable dataset.

Q. 2 (word limit: 1000 words including 2a-c; 50%)
2a) What was your own hypothesis to analyze the thymus dataset?
2b) Which experimental groups were compared to address your hypothesis?
2c) Describe your finding. Explain your idea about how your data analysis (and follow-up experiments) can address your hypothesis. (You do not have to paste your data in your answer, but you are allowed to do so).

Pdcd-1 gene, or programmed cell death protein 1, is an essential gene in the negative selection of T cell development (Arasanz et al., 2017). Although the negative selection is often found in single positive cells expressing either CD4 or CD8 co-receptor on the surface, some double positive (DP) cells can also experience negative selection.

DP cells turn blue when affected by the TCR signal and the timer protein. Then part of the cell gradually turns red, making the DP cells blue-red color because of cell differentiation. When the TCR signal fades, the DP cells become red.

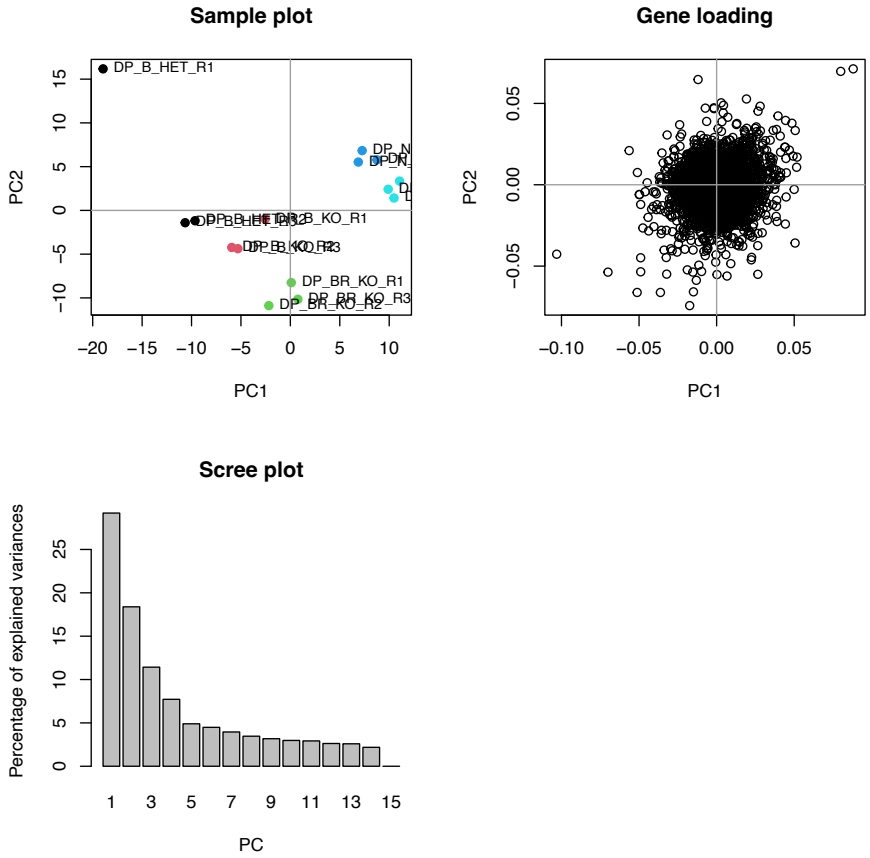

Figure 1: PCA analysis by prcomp. Sample plot shows the variation of sample clusters; gene loading shows relationship between sample and genes; scree plot shows percentage of variation that is explained by principal components.

Principal Component Analysis (PCA) shows the relationship between the samples. From Figure 1, it can be observed that DP_BR_KO and DP_B_KO clusters are at different sample space, indicating differences in their gene expression profile. Therefore, the experimental groups compared in this report will be blue+ (B) and blue+ red+ (BR) in the double positive (DP) cells with BIM knocked-out (KO) (Target: DP_BR_KO; Reference: DP_B_KO).

The hypothesis I used for the thymus dataset is that the gene *Pdcd-1*, which causes the negative selection of genes, undergoes gene expression changes when the fluorescent timer protein tagged colonies change from blue+ (B) to blue+ red+ (BR) population with their BIM being knocked out.

The follow-up analysis will be based on identifying the *Pdcd-1* using differentially

expressed genes (DEGs) identification and further analysis, along with possible activated pathways detection.

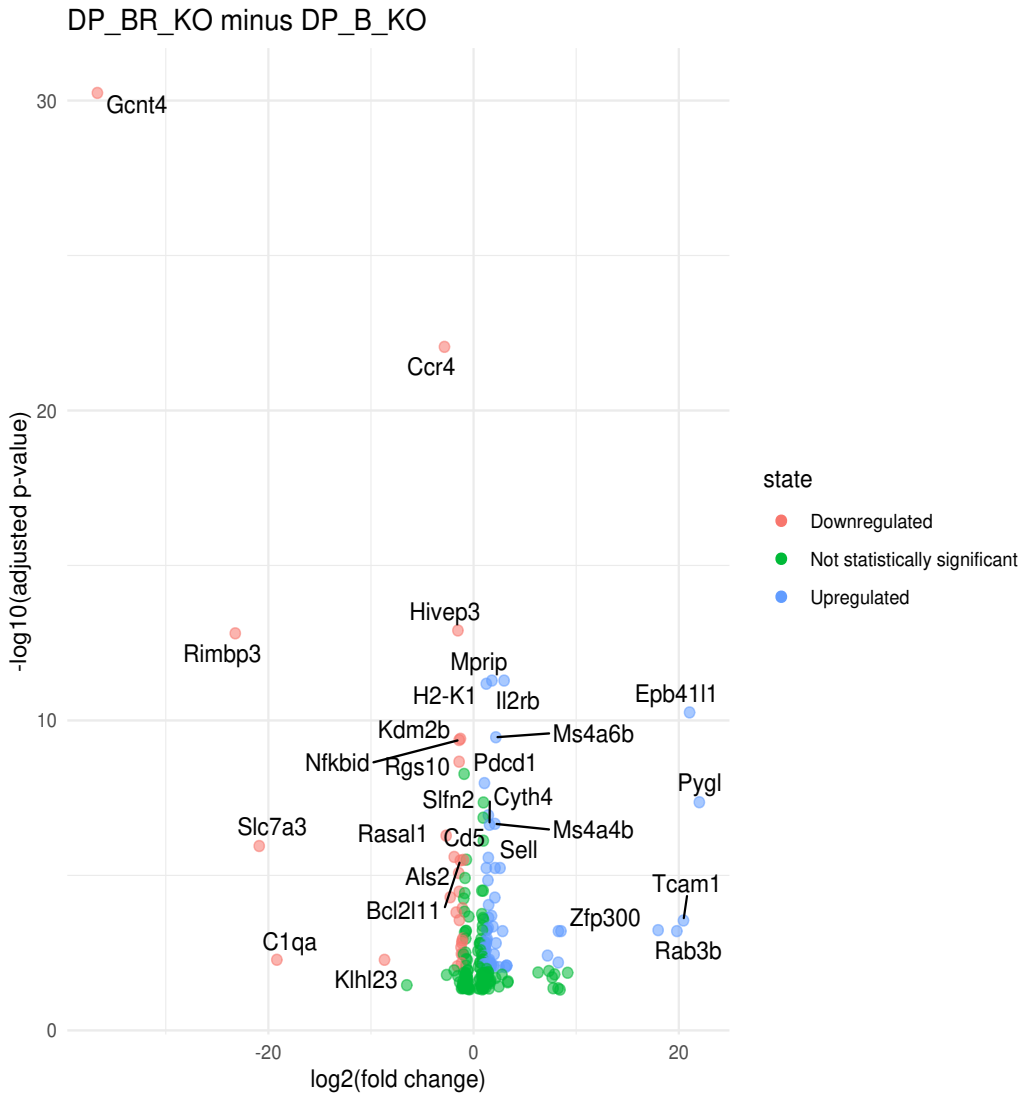

Figure 2: The volcano plot comparing DP_BR_KO and DP_B_KO samples. Genes with log2 fold change >1 or <-1 and adjusted p-value < 0.01 are significant. According to the state, *Pdcd-1* is being upregulated.

First, the volcano plot can show the changes in the gene differentiation level, comparing the reference to the target. Figure 2 shows that part of the genes is upregulated, and part is downregulated, while some are not statistically significant. It can be observed that the gene *Pdcd1* is being upregulated in DP_BR_KO cells compared to DP_B_KO cells, with log2 fold change >1 and -log10 adjusted p-value at around 8. Thus, it is certain that *Pdcd1* is part of the DEGs.

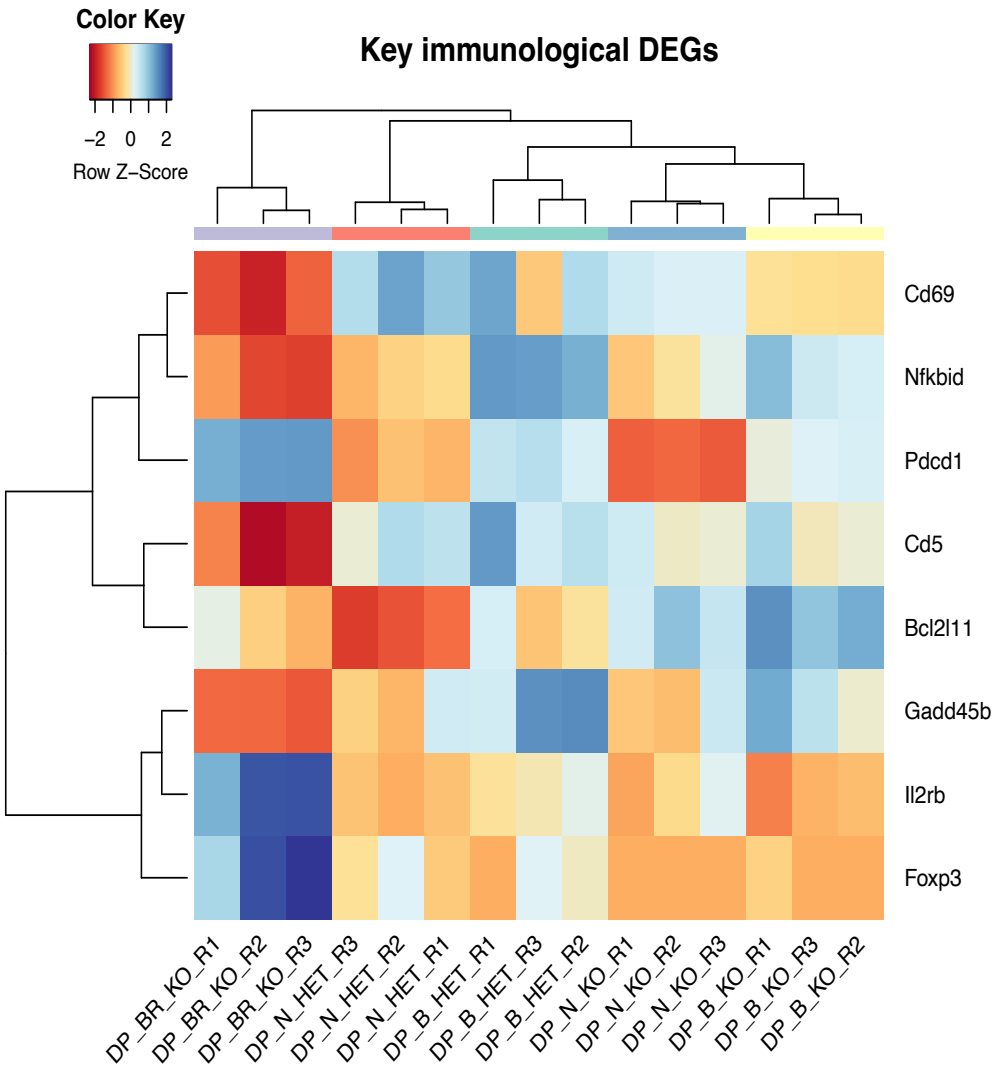

Figure 3: The hierarchical clustering and heatmap of the characteristic immunological DEGs in DP cells. *Pdcd-1* is expressed at different level. The higher the expression, the darker is the blue color. The lower the expression, the darker is the red color.

Plotting the heatmap is another way to verify if the *Pdcd-1* is upregulated during the transition from the reference to the target. As shown in Figure 3, DP_B_KO cells generally show a baby blue color, while DP_BR_KO cells have a darker blue color. According to the color key, *Pdcd-1* is more highly expressed in DP_BR_KO than DP_B_KO. Thus, it can be confirmed that *Pdcd-1* is upregulated as it differentiates, responding to the TCR signal.

In order to give a clearer representation of the degree of upregulation of *Pdcd-1*, the boxplot can display the expression level more quantitatively. As shown in Figure 4, the expression level of *Pdcd-1* increases from about 6000 to more than 10000 when comparing DP_BR_KO to DP_B_KO—the expression level increases by at least 60%. Also, the expression level does not vary too much within either DP_BR_KO or DP_B_KO, suggesting that the expression of *Pdcd-1* is relatively constant in each sample. In this case, *Pdcd-1* is clearly to have been upregulated after differentiation.

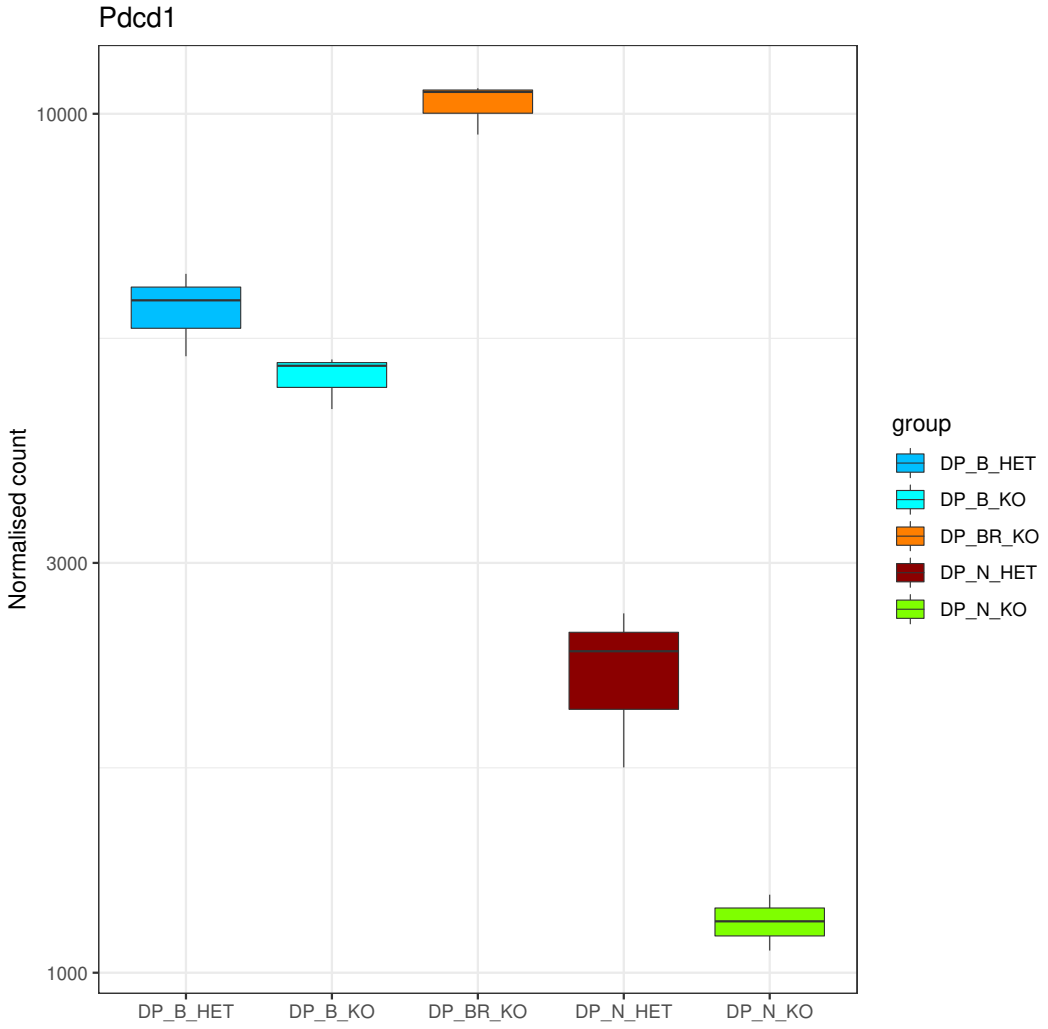

Figure 4: The boxplot of *Pdcd-1* expression level represented by normalized count in DP population. The expression level of *Pdcd-1* is higher in DP_BR_KO cells.

Next, pathway analysis by Gene Set Enrichment Analysis (GSEA) is used to see if any related pathway is being activated. However, as shown in Figure 5, no significant pathway is used by the DP_BR_KO cells. According to Figure 2, although some genes are more or less significantly expressed, the cooperation between these genes does not change much as the genes differentiate. This suggests no remarkable difference in the pathway activated when transitioning from DP_B_KO to DP_BR_KO when the TCR signal is sustained.

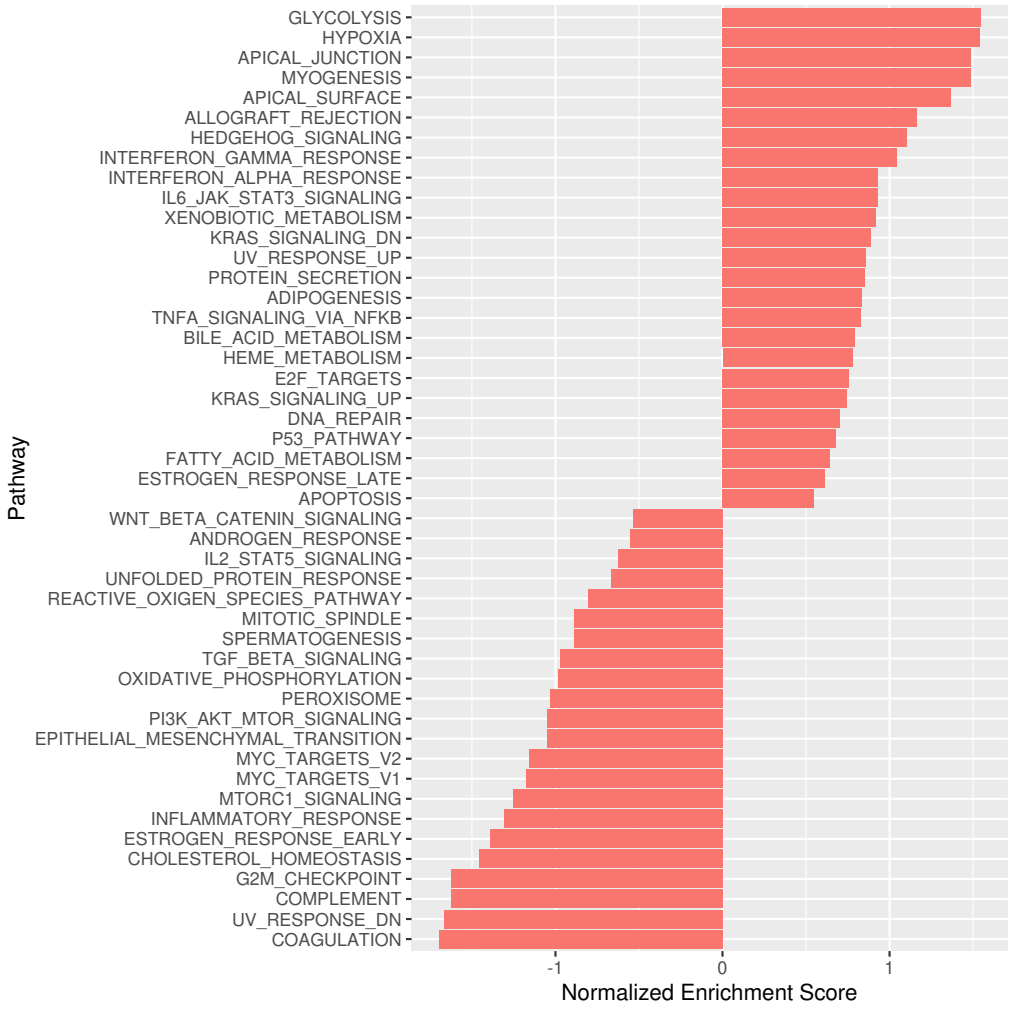

Figure 5: The pathways activated when transitioning from DP_B_KO to DP_BR_KO cells. The diagram shows normalized enrichment score and p-value. Pathways with p-value<0.05 is regarded as significant. The pathways here are all at the same color which means that there are no significant pathways functioning.

In order to see if no significant activation of pathways is a general trend in the DP_KO cells population or if it is a specific case for the transition between DP_B_KO to DP_BR_KO cells only, another GSEA is done to see whether there are significant pathway changes when DP_N_KO cells change into DP_B_KO cells when the cells just start to receive the TCR signal, shown in Figure 6. However, there is still no sign of noticeable pathway activation changes.

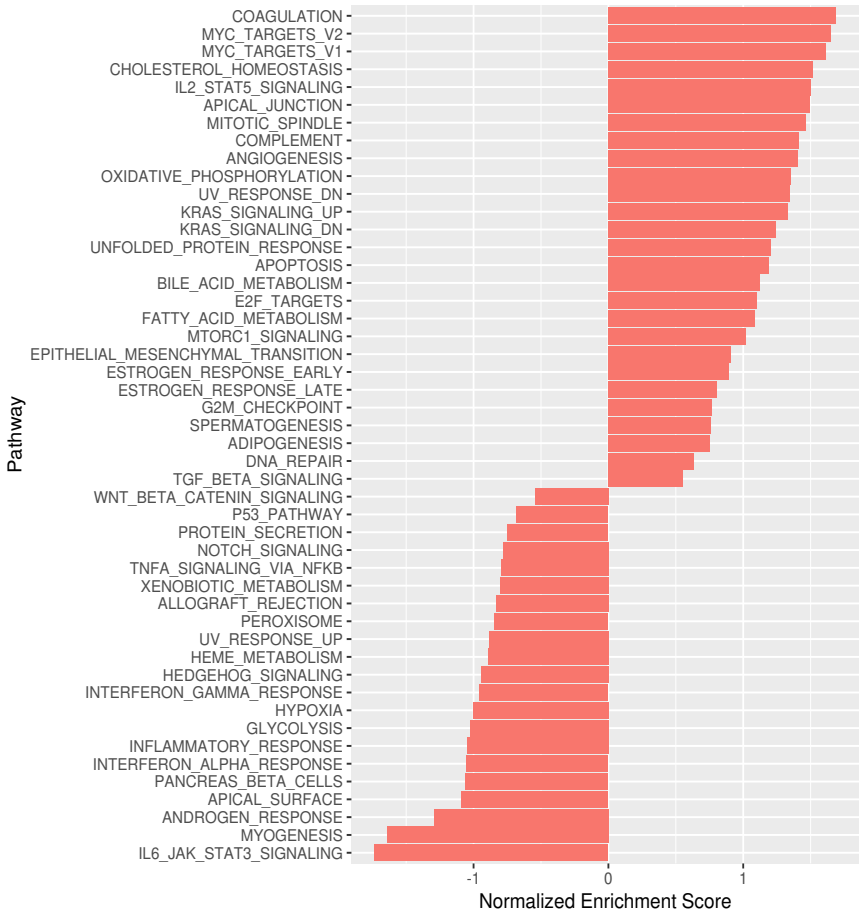

Figure 6: The pathways activated when transitioning from DP_N_KO to DP_B_KO cells. The pathways here are all at the same color which means that there are no significant pathways functioning.

Relate to the gene *Pdcd-1*, the gene is supposed to be part of the apoptosis pathway. However, there are no notable changes in its activation from DP_N_KO to DP_B_KO or from DP_B_KO to DP_BR_KO. This might be explained by the fact that *Pdcd-1* is highly expressed in the transition between double negative to double positive cells (Shi et al., 2013). Thus, it is possible that the level of *Pdcd-1* still rises when the DP cells experience further differentiation but do not cause significant changes in the negative selection pathway. At the same time, other apoptosis genes might not be actively differentiating. Therefore, GSEA cannot capture significant increase in the apoptosis pathway activation.

Although the data show that the expression level of *Pdcd-1* increases and the DP cells actively differentiates as the TCR signal persists, data analysis does not always reveal the actual situation of the gene expression and the pathway function. Thus, further experiments should be done to see if *Pdcd-1* is essential to the negative selection process in T cell development. For example, blocking antibodies or using gene silencing technique can be applied to impede the *Pdcd-1* activity (Arasanz et al., 2017) and see

if some of the T cells will persist in being activated instead of going through programmed cell death so that more T cells pass the negative selection than expected.

Q. 3 (word limit: 400 words; 25%)
Discuss why are biologists' views needed for RNA-seq experiments and data analysis? Why can't this be done in an automated manner by computational scientists only (at least currently)?

The RNA-sequencing (RNA-seq) experiments need biologists to decide on the sample being detected, the experimental environment, and the handling process of the samples, such as RNA isolation, cDNA generation, and fragment production. Biologists possess knowledge of these procedures.

Regarding data interpretation, biologists can bring in some professional views about the biological context behind the RNA-seq results. They know the role genes play in the process and which part of the information is more critical. They can interpret the sequencing data from a biological point of view. For example, pathway analysis is often involved in RNA-seq analysis. Biologists know the meaning behind the pathways, and they can provide more professional interpretations. Also, sometimes the data itself does not represent the actual scenario. For example, even though pathway analysis by GSEA marked some pathways as insignificant, biologists with expertise in the field know whether it is reliable. They know if further analysis is needed to draw a conclusion. The conclusion summarized from the experiment will, therefore, not only contain an interpretation of the numbers and diagrams but is also connected to biological mechanisms.

Also, since biologists are the ones who handle all the samples, they know the details of the experiments and are more likely to identify potential mistakes. Outliers might be produced in the process, and biologists know which of them are regarded as expected, which are caused by technical mistakes, or whether they suggest anything significant. However, computational scientists can only analyze the data based on their statistical knowledge, which is insufficient to understand the whole picture.

One important reason for data analysis is for further experimental design. Knowledge of gene expression, gene regulation, and pathway function can help biologists formulate further hypotheses and design follow-up experiments to validate new findings. It is difficult for computational scientists alone to find the intrinsic relationship between genes or between pathways and come up with good experimental ideas. Only biologists know which techniques are helpful to use in the following experiment.

Reference
Arasanz, H., Gato-Cañas, M., Zuazo, M., Ibañez-Vea, M., Breckpot, K., Kochan, G. & Escors, D. (2017) PD1 signal transduction pathways in T cells. *Oncotarget*. 8 (31),

51936–51945. doi:10.18632/oncotarget.17232.

Shi, L., Chen, S., Yang, L. & Li, Y. (2013) The role of PD-1 and PD-L1 in T-cell immune suppression in patients with hematological malignancies. *Journal of Hematology & Oncology*. 6 (1), 74. doi:10.1186/1756-8722-6-74.