Lindsay Reisman & Zunqiu Wang
Transcriptomics : Project Proposal

1. Select an experiment for your final project: **GSE179320**

2. Write a page describing the experiment:
   - Bulk RNA seq was performed from PMs(Promyelocytes), MCs(Myelocytes), NPs(Neutrphils) using Quanten RNeasy Plus Micro kit.

   a. Why is this experiment important to you?
      - Reticular Dysgenesis (RD) is a rare but destructive form of severe combined immunodeficiency, characterized by a maturation arrest of the myeloid and lymphoid lineages paired with sensorineural hearing loss. To find the expression pattern of genes will bring new insights on biomarker selections in the context of therapy.

   b. What are you hoping to discover after analyzing the data?
      - From analyzing this data, it would be interesting to understand why AK2 is depleted and how that affects myelopoiesis and its metabolic pathway. Detecting any pattern from the cells would be telling in how they affect any downstream processes they are involved in.

   c. How many different factors are present in the experiment and how many levels in each factor?
      - The dataset has 2 factors (condition and cell) and within each factor, there are 2 levels in condition(AAVS1 CRSPR, AK2 CRISPR) and 3 levels in cell (Promyelocyte, Myelocyte, Neutrophil).

   d. How many biological replicates are available for each sample?
      - 3 replicates of each cell.

   e. How are you going to analyze the dataset?
      - Using Deseq2 to statistically analyze the differentially expressed cells in relation to the condition. Our plan is to also use GO term enrichment. Using a lot of visualization such as PCA, heat maps and hierarchical clustering to learn more about the cells and the relationships to each other in both conditions. Using KEGG and GSEA to learn more about the the pathways that are involved.

   f. What statistical methods are you going to use ?
      - **DESeq2**: **Normalization steps:** including computation of geometric means, each gene count divided by this mean, getting median of these ratios as size factor for that sample thus correcting RNA composition bias

- **design formula:** involves multiple factor analysis in negative binomial generalized linear model.
- **Wald test**: due to unreliableness from a small number of replicates, the shrunken estimate of dispersion and LFC is implemented and then is divided by its standard error, resulting in a z-statistic, which is compared to a standard normal distribution.
- **FDR**: correcting multiple hypothesis testing using BH method
- **GO term**: Hypergeometric testing
- **GSEA**: based on association of ranked gene list and phenotype of interest to compute permutation of p value of category based test statistics. Then, the upper and lower tail of probabilities are calculated by comparing original category based test statistics to permutation test statistics of that category. The result is to decide whether genes in that category are spread evenly throughout the ranked list.

2. You don’t have to run any commands just explain your approach.
    - Our approach is using the above mentioned statistical methods to analyze how cells and conditions respectively alter gene expressions in the pathways involved in cell development and metabolism.