1. Starting with a **raw count array** (genes are rows, columns are samples (or patients) with first column being gene name)
2. Gene names are appended using bioMart(differentiate pseudo genes, protein coding genes, miRNA, lncRNA etc).
3. Filter out genes whose expression dominates the expression of all genes[reported to be bias to the RNAseq data], create heatmap for sanity check of outliers.
4. Create **normalized array**: we create a DESeq2 data set and do size factors normalization.
5. We then transform the normalized data using log2(1+p) or whatever transform you'd like (DESeq2 has the built-in function to do the variance stabilization)
6. Using the raw count array, we then filter genes using Fred's filtering (>90% zero OR maxcount < 10 OR 95th quantile count < 10 AND > 50% zero) and pseudogenes based on gene annotation
7. Then the genes are clustered using a fast correlation table (can set cor max) and a graph, and all genes but one are removed from each cluster (we can do this after obtaining the final data frame for easier tracking; need to revise the clustering algorithm)
8. Perform PCA and clustering analysis(t-SNE/UMAP) on normalized array to identify potential batch effect.
9. Using linear regression function(algorithm to be determined) to regress out batch effect while keep the signals from other important factors, Perform PCA and clustering analysis after batch effect removal. This leads to the third **batch free array**.
10. For both normalized and batch free array, keep genes in raw count array.
11. Other kinds of explorative analysis with RNAseq data to be added. (e.g. creating topology structure between protein coding genes, miRNA, and lncRNA)