**Project Proposal Title**: “To Tea or Not to Tea”

**Group members:** Brandon Thong & Dawn Layman

**Topic and Research Question**:

* Weekly comparison of Zonal and Direct topical EGCG application for Skin Scarring.
* Does Zonal Priming or Direct Treatment with EGCG improve outcomes for scar-free or minimized scar healing, and can such improvements be captured and identified via differential gene expression analysis?

**Hypothesis:**

EGCG has been attributed to having anti-inflammatory, anti-angiogenic, and antioxidant properties that promote wound healing. This project investigates the presumed effect of “Direct Application” of topical EGCG compared with “Zonal Priming” on improved outcome of scar healing, measured by differential gene expression.

**Data sources:**

NCBI GEO is a database/repository for publicly available datasets, this dataset associated with our project is assigned the accession number: GSE124161

And we downloaded the associated data files:

* GSE124161\_readcount.txt
* GSE124161\_series\_matrix.txt

Paper:

* “A Double-Blind, Randomized Trial Shows the Role of Zonal Priming and Direct Topical Application of Epigallocatechin-3-Gallate in the Modulation of Cutaneous Scarring in Human Skin.”
  + Journal of Investigative Dermatology (2019) 139, 1680e1690; doi:10.1016/j.jid.2019.01.030

**Methods:**

We would like to reproduce key figures in the paper and compare the results from our analysis to the results achieved in the paper. The following computational strategy will be utilized to achieve this goal.

Sequencing data is based on reads that map to a reference and is quantifiable as “counts''. Counts are classified as discrete data, which means values are determined by whole number integers that are confined to specific values. Additionally, sequencing data is not normally distributed, therefore certain statistical analyses need to be employed in order to determine differential expression and statistical significance.

Differential Gene Expression (DGE) will be performed utilizing DESeq2, an R package based on negative binomial distribution, which is ideal in the processing of non-negative count data. The P-adjust and fold change cutoffs for the selection of genes for further analysis will be selected according to paper parameters utilizing Benjamini-Hochberg approach, for controlling the False Discovery Rate, selecting a Corrected p-value 0.005 and Log2FC = 1. Data selection will utilize the “subset()” function to select the differentially expressed genes (DEG’s) that meet the selection criteria.

Clustering the data will be performed utilizing course methods, as directions in the clustering within the paper were vague. Clustering of DEG data will utilize distance as a function of correlation (1- cor), and subsequent clustering utilizing hclust(distance, method = “average”). The quality of clustering will be evaluated through the generation of a silhouette plot utilizing the “cluster” package, and sample clustering will be evaluated via PCA plot utilizing the function prcomp() with ggplot() to plot the data.

Once clustering and cluster analysis is performed, the DEG’s will be further evaluated visually by generatingHeatmaps for overall visualization of gene and sample clustering, utilizing the “pheatmap” package; generating Volcano Plots to to visualize biologically statistically significant genes.

Enrichment analysis will be performed to uncover and assign biologically relevant processes to our DEG’s through GO-term enrichment utilizing Bioconductor packages “Gostats”, “GO.db”, “Category” and “org.Hs.eg.db”; and creating a GOHyperGParams object to create a list of GO-terms, categorized by: BP = Biological Process, CC= Cellular Component, and MF = Molecular Function.

**Expected Results:**

After performing our differential gene expression analysis, we expect to see downregulation in mast cell, angiogenic, and inflammatory genes, as well as an upregulation of antioxidant genes. Furthermore, when compared to the original paper, we expect these findings to be very similar, but not necessarily identical. This is due to potential differences that may arise due to R package and database (KEGG/GO) updates as the original experiment was performed in 2019 on GRCh37/hg19 with the package versions available at the time.

**Potential Problems and Solutions:**

Potential problems we may encounter include not being able to perfectly replicate the plots and graphs provided in the original paper, due to incomplete knowledge about how their data was clustered and how their graphics were generated. To address this, we will perform the analysis with various parameters to ensure the accuracy of our results, even if they may not match the exact appearance of the original paper's plots and graphs. We also may encounter some differences in our DGE and GO-term enrichment analysis due to updates in packages and databases. However, we do not expect these to alter our results significantly, and any differences that do exist will be noted and explained in our findings.