Progress Report

Charlene Harasym, Jackson Chen & Kira Li

Introduction:

Lung Squamous Cell Carcinoma is the most common lung cancer in men, often associated with smoking. It starts in squamous cells near the bronchi and is often found in the central part of the lung. Our analysis of this cancer uses the Lung Squamous Cell Carcinoma TCGA PanCancer dataset, particularly the files containing mutations, RNA-Seq, and clinical data. To gain a basic understanding of the population, we will analyze and visualize the clinical data. Using the mutation data, after filtering, we will cluster based on the mutations of each patient. These clusters can then be linked to gene expression data to perform differential expression and further, pathway analysis with the strongest upregulated and downregulated genes. Finally, survival analysis of each cluster may reveal differences between the clusters that characterize them as different subtypes of Lung Squamous Cell Carcinoma.

Project Goal:

Through analysis of mutation, gene expression and clinical data, the goal is to identify the major subtypes of Lung Squamous Cell Carcinoma. Each subtype will have some common mutations and highly expressed genes that characterize it, which is expected to impact the survival analysis. The identification of different subtypes can help the development of better-targeted treatments or therapies with the goal of improving overall survival.

Analysis Plan:

Describing the Population

Using data from the "clinical_patient_data" file, we will perform descriptive statistics and create plots to help visualize our patient set. We will use columns "AGE", "SEX", "AJCC_PATHOLOGIC_TUMOR_STAGE", "RACE", "OS_MONTHS", "RADIATION_THERAPY", and "NEW_TUMOR_EVENT_AFTER_INITIAL_TREATMENT".

Filtering Mutations Data

We will assess and select various methods of filtering the "data_mutations_extended" data to achieve a smaller dataset that contains only mutations that are likely to have had a significant impact on the development and progression of the cancer. The variables that we are going to filter by are "Consequence", "IMPACT", "SIFT", PolyPhen", "BIOTYPE", "t-depth", and VAF and read count through the use of "t_ref_count" and" t_alt_count". When applicable, we will analyze the variable through graphing and pick threshold values to base our filtering on. We will ensure that we do not filter by too many variables to avoid over-constraining our data.

Clustering on Filtered Mutations Data

We will then take the filtered mutations data and create a binary matrix of patients vs genes that could contain a mutation. This boolean matrix will then be used to complete k-means clustering. We will perform clustering for k values of 1-10 and calculate the total within-cluster sum of squares for each k. We will then plot the curve of each sum of squares and look for a bend in the curve to indicate the number of clusters that should be used to determine our final clusters. Then descriptive statistics on our clusters will be generated using the same variables as step 1. The variable "PERSON_NEOPLASM_CANCER_STATUS" will also be included to determine whether the patient is tumor-free or not at the end of the study.

Differential Expression and Pathway Analysis

We will perform differential expression analysis on the "RNA_Seq" data to find the differences in gene expression within the previously defined clusters. We will first filter the "RNA_Seq" data to remove any genes that have counts of 0 or 1 for all patients. Then, we will sort the RNA_Seq data according to the previously defined clusters. An appropriate p-value, false discovery rate threshold, log2 fold threshold will be chosen and differential expression analysis will be performed. We will then identify the strongest

up regulation and down regulation genes and perform pathway analysis on these genes using the gage function.

Survival Analysis
We will first link the clusters to the "clinical_patient_data" and create a new data frame that includes the variables of importance for each patient that has been classified into a cluster, and the cluster that they are assigned to. The variables of importance are "PATIENT_ID", "OS_STATUS", "OS_MONTHS" and "DAYS_LAST_FOLLOWUP". We will use these to perform survival analysis and create Kaplan-Meier Plots. The significance of the difference between the curves will be tested using the p-value to ensure statistically significant curves have been created.

Findings Analysis
After all of the analyses have been performed in R, we will gather our findings and draw conclusions on how the different subtypes of Lung Squamous Cell Carcinoma vary in gene expression, and how these subtypes relate to the survival of the patient.

Challenges:

There are a few challenges that we expect to encounter during the execution of our analysis plan. One is dealing with variability in the completeness of data between all patients, specifically in the filtering of the mutation data. A second challenge that could arise is not being able to determine significant clusters due to data under/overfitting. Another will be identifying and accounting for confounding variables within survival analysis. Finally, there is the potential to have failure in obtaining statistically significant results at any stage in our analysis, which could lead to difficulties in drawing conclusions.

Timeline and Duties:

All group members researched the stages of analysis individually before meeting to determine the best course of action. We met and decided to execute the following plan.

Task	**Deadline**	**Jackson**	**Kira**	**Charlene**
Progress Report	November 19	Wrote challenges and timeline and duties	Wrote introduction and project goal	Wrote analysis plan and edited
Computational Analysis Stage 1	November 23	Complete filtering and clustering on mutations data	N/A	N/A
Computational Analysis Stage 2	December 1	N/A	Complete population description and survival analysis	Complete differential expression analysis and pathway analysis
Planning	December 2	Meet to discuss findings and plan the discussion section of the paper		
First Draft of Paper	December 6	Write the discussion and conclusion sections	Write the abstract and introduction sections	Write the methods and results sections
Filming of Presentation	December 8	Complete discussion and conclusion slides	Complete abstract and introduction slides	Complete methods and results slides
Final Draft	December 9	Edit and review the paper		