**Foundations of Programming, Data Analytics, and Machine Learning in Python ENG BF550: Fall 2021**

PROJECT 2

**PROJECT REPORT:**

**Single-cell mRNA sequencing identifies**

**subclonal heterogeneity in anti-cancer drug**

**responses of lung adenocarcinoma cells.**

SUBMITTED TO:

ILIJA DUKOVSKI

SUBMITTED BY:

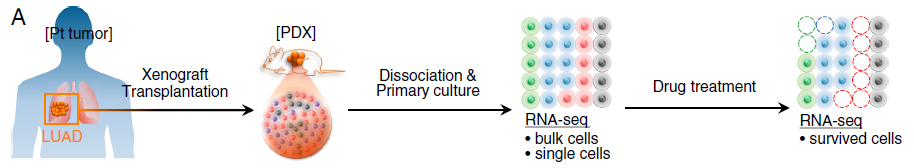
NIKITA TOMAR

**Introduction:**

Intra-tumoral genetic and functional heterogeneity correlates with cancer clinical prognoses. However, the mechanisms by which intra-tumoral heterogeneity impacts therapeutic outcome remain poorly understood. RNA sequencing (RNA-seq) of single tumor cells can provide comprehensive information about gene expression and single-nucleotide variations in individual tumor cells, which may allow for the translation of heterogeneous tumor cell functional responses into customized anti-cancer treatments.

In the research study, 34 patient-derived xenograft (PDX) tumor cells were isolated from a lung adenocarcinoma patient tumor xenograft. Individual tumor cells were subjected to single cell RNA-sequence for gene expression profiling and expressed mutation profiling. Fifty tumor-specific single-nucleotide variations, including KRASG12D, which is a key player in the receptor tyrosine kinase (RTK)-RAS-mitogen-activated protein kinase (MAPK) pathway, genetic alterations in which cause LUAD(Lung Adenocarcinoma), were observed to be heterogeneous in individual PDX cells. Semi-supervised clustering, based on KRASG12D mutant expression and a risk score representing expression of 69 lung adenocarcinoma-prognostic genes, classified PDX cells into four groups. PDX cells that survived in vitro anti-cancer drug treatment displayed transcriptome signatures consistent with the group characterized by KRASG12D and low risk score.

As a result of the research study, single-cell RNA-sequencing on viable PDX cells identified a candidate tumor cell subgroup associated with anti-cancer drug resistance that were masked in bulk tumor analyses. Thus, single-cell RNA-sequencing is a powerful approach for identifying unique tumor cell specific gene expression profiles which could facilitate the development of optimized clinical anti-cancer strategies. Transcriptome profiling was also performed on single PDX cells from a LUAD patient to elucidate the molecular mechanisms and underlying genomic characteristics of tumor cell resistance to anti-cancer drug treatments.



**Fig**: Schematic representation of experiments. A portion of a LUAD patient tumor (*Pt tumor*) was propagated by xenograft transplantation in humanized immunocompromised female NOG mice. PDX cells (*PDX*) were dissociated and cultured from xenograft tumors, and subjected to drug screening.

**Data:**

* + **De novo method:** Tumor cell-enriched PDX cells (LC-PT-45) were analyzed by single-cell RNA-seq using the Fluidigm C1™ autoprep system with SMART-seq.
  + **Data:** This dataset consisted of the sequenced transcriptomes from one batch of 34 lung adenocarcinoma enriched patient-derived xenograft (PDX) cells (LC-PT-45), 49, 50 human lung cancer cell line cells (H358), and a second batch of 43 lung adenocarcinoma enriched PDX tumor cells (LC-PT-45-Re).

**Analysis 1 – Multiple Regression Analysis**

Our dataset of LUAD cells has essentially two types of samples. First is the genomic expression for all single cells combined or ‘bulk sample’ and the second is individual single cell samples. The idea was to find out the degree of transcriptome heterog-eneity. This was evaluated by multiple regression analysis of different sized pools (n = 5, 15, 25, 34/35, 50; randomly selected by permutation × 1000) of single cell transcriptomes to the bulk sample. (Fig. 1C). Multiple regression analysis estimated how many single cells hypothetically accounted for the pooled cell fraction. Single-cell samples were randomly chosen with the given number and the adjusted R2 were determined 1000 times with permutation.  
In our dataset we have 3 types of PDX cells – H358 (50 single cells), LC-PT-45 (34 single cells) and LC-PT-45-Re (43 single cells). The modeling demonstrated that five H358 or PDX individual cells represented >70 % of the gene expression of the whole population. When averaging increased numbers of cells, the single cell data approximated the bulk up to 85 %, suggesting that the single cell data are consistent with the bulk data. We repeated the single cell isolation and RNA-seq using 43 additional PDX cells and obtained comparable results that were highly correlated with the first data set. Fig. 1A shows that the bulk cells and the individual cancer cells of a type are highly correlated which means we can continue with our analysis of an individual cell as against the whole bulk cell. Fig. 1A is a scatter plot of gene expression log¬¬¬¬2¬(TPM+1). Fig. 1B shows the inter-correlation between each single cell of the four PDX cells – H358, LC-PT-45, LC-PT-45-Re. We see that the LC-PT-45 and reversed LT-PC-45-Re have a broader distribution of inter-correlation than H358 suggesting that the PDX cells (LC-PT-45 and LC-PT-45-Re) show a wider spectrum of heterogeneity.

**Code Description 1 -** *Please refer to analysis\_1.ipynb for this purpose*

This code description is for Fig. 1C plot. The code for this analysis has 3 sections.

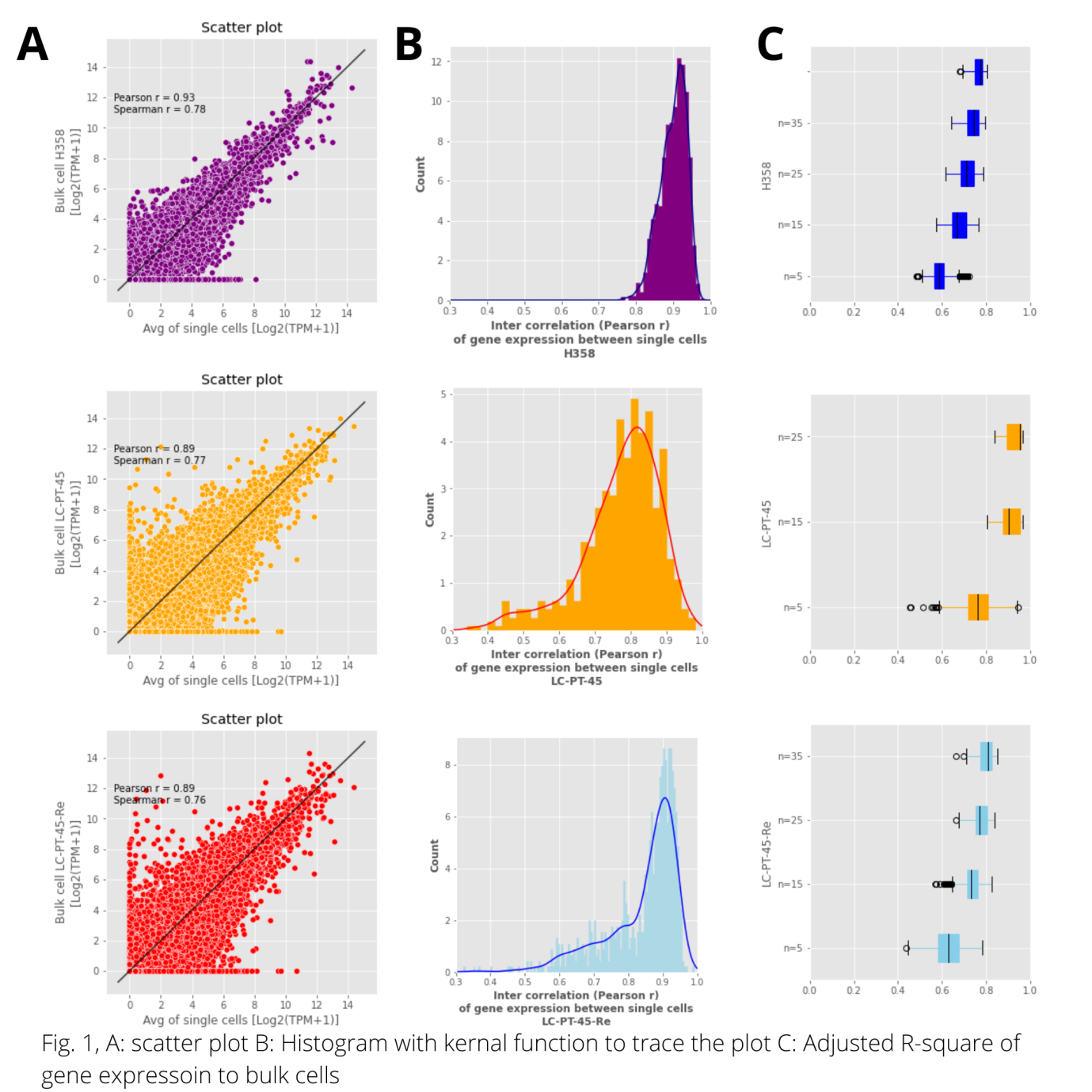
Firstly I defined a function. Here I first created a linear regression analysis model. Then I used the set of single cells data as a multi variable ‘X’ and the bulk cell data as ‘y’ and fit into this model to get a multivariable regression analysis model. I used this model to compute the R2 value and the adjusted R2 values. With the following formulas :

R2:

Adjusted R2 :

Then, I defined a function that would calculate adjusted R2 values for a set of n = 5, 15, 25, 3 cells selected randomly for 1000 times. I did this for each of the 3 types of PDX cells and plotted box plots to get an idea of how consistent the single cells are with their respective bulk cell samples. I saw that for smaller sample sizes the bulk cells and the selected single cells show lower values for R2 (<0.6) for all the 3 types of cells thereby indicating that the individual cells have more varied genetic information that their bulk cells. As I increased the value of n (randomly selected n cells from the total single cells of a particular type) I saw that the single cells show high consistency with the bulk cells (>0.8) which is an expected behavior.

For Fig. 1A the I took the average of gene expression of all sing cells (Log2(TPM+1)) and plotted this average against the gene expression of bulk cell (Log2(TPM+1)). The code does this simple mathematic operations and plots a scatter plot – def showscatterplots

For Fig. 1B I have coded the functionality under def showhistplot. I performed Pearsons correlation on between each of the single cells of a type. The result is a symmetrical matrix with inter-correlation data of the columns. The diagonals have a value of 1 since it represents the correlation of a cell with itself. I used this information to plot a histogram and a kernel function to trace the top of histogram. This plot gave me a visibility on how vibrant the single cells are among themselves. A broader distribution meant a more heterogeneity. 

**Fig.1** **Intra-tumoral heterogeneity of PDX cells.**

a) Scatter plots of the average gene expression of single cells (H358, n = 50; LC-PT-45, n = 34; LC-PT-45-Re, n = 43) compared with those of the corresponding bulk cells (~1 × 105 cells). Black dotted lines are x = y lines with correlation coefficients (Pearson r and Spearman r) for linear fit. TPM (transcripts per million).

b) Inter-correlation (Pearson r) between gene expression in single cells. Density plots were constructed with a kernel function fitting over the histograms.

c) Explanatory power (adjusted R-square) of gene expression in various numbers of single cells relative to the bulk cells was determined by multiple regression analysis with randomly selected cell numbers with permutation (×1000).

**Analysis 2 – Principal Component Analysis**

Analysis 2 – Principal Component Analysis

The idea is to identify heterogeneity in the individual single cells of LUAD. For this purpose samples were taken from PDX cells. The aim was to find a relation between these 4 cells. Therefore to identify these subgroups, with possible phenotypic implications in the PDX cells, we utilized the expression profiles of 69 genes related to the clinical prognosis of LUAD patients as multivariate markers to compute a risk score. Risk Score were calculated my doing a sum over multiplication of these 69 gene beta-coefficient and the corresponding gene expression in each of the single cell. Moreover, a higher RS was significantly associated with the KRAS mutation in the LUAD patient population. Interestingly, individual PDX cells were calculated to have a wide RS distribution. Eighteen out of the 34 PDX cells or 21 out of 43 of the replicate samples were determined to be high-RS. Altogether, the PDX cells were grouped using the information of the KRAS mutation and RS. This resulted into classifying the PDX cells into four groups: group 1, no KRAS G12D (KRAS WT)/low RS (n = 3); group 2, KRAS G12D/low RS (n = 25); group 3, no KRAS G12D (KRAS WT)/high RS (n = 3); and group 4, KRAS G12D/high RS (n = 35). The distinct gene expression signatures among the four groups were visualized by a principal component analysis (PCA) plot using genes exclusively expressed by each group. (Fig. 1 A) I used the first three principal components to perform a K-means clustering so that these cells can be distinguished into clusters. For this first I created an elbow plot for k values in the range 1 to 7 and found the optimal cluster number k to be 2. Then, I used this cluster number to perform a K-means clustering on the first 3 components of PCA and plotted the results (Fig. 1B). The two clusters represent high and low risk scores. As a supplementary plot I also created a boxplot to express KRAS mutation in each of the PDX cells as a log function Log¬¬2¬(TPM+1). Here we can see that the mutation exists in all cells.

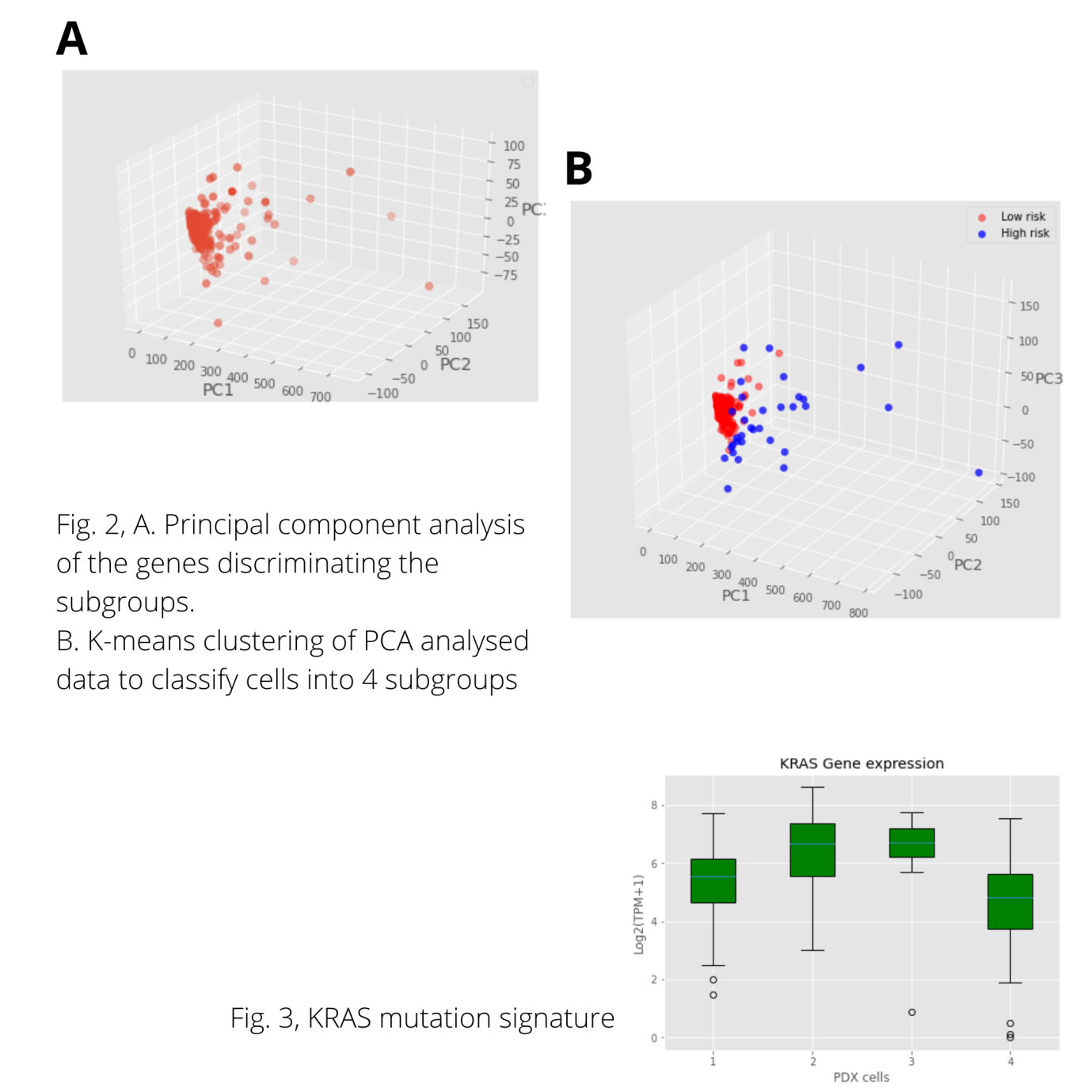
Code description 2 - Please refer to analysis\_2.ipynb for this purpose.

Firstly I created a pandas DataFrame of the data which has the genomic expressions for each type of LUAD single cells and their corresponding bulk sample cell. I then used the expression profiles of 69 genes to calculate Risk Scores for each of the single cells using this formula –

These risk scores along with the information of whether the single cell have ‘KRAS’ mutation expression in it were used to group the 197 single cells into four groups. These four groups were as follows – 2 groups with low RS and 2 groups with high RS. Each of these 2 group had one which had KRAS mutation expressed and the other had not. In total 66 cells out of 197 were grouped in such manner.

From the dataframe we filtered out the columns corresponding to these 66 PDX single cells. For the PCA analysis we filtered out genomic expressions which were mutually exclusive to these groups. A total of ~6k rows of genomic expressions were filtered out from 50k genomic expressions. Then I used this set of processed data to run a PCA analysis. For the PCA analysis, first I scaled these values to standardscalar which meant the mean of this data became 0 and the standard deviation became 1. This scaled data is then fit into a PCA model. To find relevant number of principal components to use for the plot I decided to plot the variance as a function of principal component and choose the top three of them. These 3 components were further used to do a 3D-scatterplot hence completing our goal of discriminating the subgroups.

**Result 2:**



**Fig. 2 and 3: Identification of PDX cell subclones using single-cell RNA-seq data.**

**Fig 2:** Principal component analysis of the genes discriminating the subgroups. Ellipsoids were generated with standard deviations around each group.

**Fig3:** Comparative features among the classified single cell subgroups. KRAS gene expression (Log2 ratio of transcripts per million + 1). Gene set signature scores (computed by gene set variation analysis) corresponding to the KRAS over-expression signature.

**Conclusion:**

The unexpected results indicate that (1) tumor cells with activated KRAS signatures were drug targets, but

the KRAS mutation itself was not a target, and (2) the actual tumor population responsible for drug resistance

might be masked by dominant genomic characteristics within a bulk population. In this study, the cells that survived the effective treatments retained the KRAS mutation but seemed to stay in a dormant state without

activating KRAS signaling. Interestingly, the molecular signatures of this group indicated upregulation of genes involved in the ion channel transport and P-type ATPases, which might play key roles in drug resistance. Whether this potentially drug-resistant population is indeed a pre-existing tumor subclone or dynamically

changes gene expression signatures in response to drug treatments needs to be addressed by future studies.

**References:**

* **1.Research paper : Kim, Kyu-Tae et al. “Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells.” *Genome biology* vol. 16,1 127. 19 Jun. 2015, doi:10.1186/s13059-015-0692-3**