Mapping cancer cell line dependencies to patient populations

Sinead Dunphy, Alyson Freeman, Kristina Garber

Introduction and Motivation

Cancer is one of the leading health burdens with close to 20 million newly-diagnosed cases and nearly 10 million deaths worldwide in 2020. It arises when normal cells of the body undergo cellular changes and begin to proliferate uncontrollably. Cancers are caused by a combination of eternal agents (for example - UV radiation, components of tobacco smoke, or certain pathogens) and internal genetic factors.

Cancer is a broad term encompassing a number of subtypes affecting different organs of the body, with breast and lung cancers being the most prevalent. Even within a cancer subtype, there can be a diverse range of cancer presentations. Histology, which is the characterisation of the cancer cells under a microscope, is employed to define different types of cancer.

While treatments for cancer have progressed hugely in recent decades, improvements are still required. Most standardised chemotherapies simply target rapidly proliferating cells, meaning normal cells with naturally high proliferation (such as the intestinal epithelium) are also affected, leading to significant side effects for the patient. Furthermore, cancers can develop resistance to treatments.

Targeted cancer therapies rely of the identification of components, such as genetic markers or proteins, which are specific to the cancer cell. Targeted therapies thus can induce less side effects by being more precise in their action and can bring new hope against treatment-resistant cancers.

The aim of this project is to investigate datasets of genetic factors in cancer cell lines and in samples from human cancer patients to try and identify novel targets which could serve as future therapeutic avenues. Cancer cell lines are cells which have been derived from human cancers and which are grown *in vitro* (outside the body) for laboratory-based cancer research.

Key Question: Can we identify specific genetic dependencies in a relevant patient cancer histological type?

Data Sources: DepMap

Cancer cells have genetic mutations and gene expression differences which differentiate them from normal cells. Some of these differences may be critical for the survival of the cancer cell. If this is the case, the cancer cell line is considered to be dependent on that gene. Characterising genetic dependencies of different cancers is an important part of identifying new therapeutic targets.

Researchers at the Broad Institute have embarked on an ambitious effort to profile almost one thousand cancer cell lines which they call Project DepMap. The aim is to define a Cancer **Dep**endency **Map** to better understand cancer genetic dependencies, identify genetic targets for cancer therapies and determine the subsets of patients for whom these therapies could be most effective. Project DepMap includes CRISPR screening data to measure genetic dependencies, gene expression data, and sample metadata for each cell line. CRISPR gene editing is a genetic engineering technique which allows a specific sequence of DNA within a cell (such as a gene) to be targeted and “cut-out”. The survival of the cell is then measured to determine if that cell is dependent on that gene.

The Genetic Dependency data and the Cellular Models metadata were joined using the unique DepMap cell line IDs. The Expression data and Genetic Dependency data were joined using the gene names (also called symbols).

As part of their commitment to open science, Project DepMap makes all their datasets available to the public. We will be using a number of datasets from Project DepMap downloaded directly from the website (DepMap Data Downloads).

DepMap Datasets

Genetic Dependency - CRISPR

Format: CSV **Size: 340MB**

In this dataset, CRISPR gene editing has been employed on over 1000 cancer cell lines, cutting out over 17000 genes and determining if each cancer cell line is dependent on each gene. The index contains unique DepMap IDs for each cell line, the columns are individual genes, and the dependency is measured in floating point numbers.

Cellular Models - Expression

Format: CSV **Size: 428MB**

This dataset provides details on the extent of expression of 19177 genes in over 1300 cancer cell lines, as determined by RNASeq which is a technique to measure how much of the gene’s product (RNA) is present in cells. The index contains unique DepMap IDs for each cell line, the columns are individual genes, and the expression levels of each gene are given in floating point numbers.

Cellular Models – Cell Line Sample Info

Format: CSV **Size: 394MB**

This data contains the metadata for the cell lines including unique DepMap ID, cell line name, lineage (e.g. if it came from ovary cancer, lung cancer, lymphoma, etc.) whether it is a primary cancer or metastasis, and the cancer subtype.

Data Sources: TCGA

The Cancer Genome Atlas Program (TCGA) was an effort between the National Cancer Institute and the National Human Genome Research Institute to characterize over 20,000 human tumor samples. The goal of this effort was to enable researchers and clinicians to better diagnose, treat, and predict cancer.

The main dataset that was used in this analysis was from the Pan-Cancer Project encompassing over 11,000 patient samples from the 33 most prevalent types of cancer. The gene expression of the samples was measured via RNAseq. We downloaded the data from the UCSC Xena repository (Downloads). The Patient Sample Metadata containing the details for each sample was downloaded from the National Cancer Institute Genomic Data Commons (PDF) and joined to the Gene Expression dataframe via the unique sample IDs.

By comparing the list of gene names in the DepMap dataset and the TCGA datasets, it was clear that were differences in the naming conventions used. Therefore, we used the HUGO Gene Nomenclature Committee Database (text file) to update the gene names in the TCGA to Approved Symbols which more closely match the names used in the DepMap data. This was done by making a dataframe with one column of Approved Names and another column of all aliases and then merging to the Gene Expression dataframe.

TCGA Datasets

Gene Expression

Format: TSV **Size: 1.88GB**

The Gene Expression file contains the expression levels of close to 20,000 genes for each of the over 11,000 patient samples. The index is unique sample IDs, the columns are individual genes and the gene expression of each gene per sample is given as a floating point number.

Patient Sample Metadata

Format: Excel **Size: 2.9MB**

The Patient Sample Metadata contains information about the patient and cancer including unique sample ID, gender, age, tumor location, primary or metastasis, histological type, stage, and treatment outcome.

Gene Names

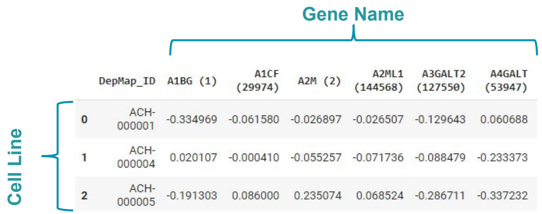
Format: TXT **Size: 1MB**

A text file was generated from the HUGO Gene Nomenclature Committee Database using their custom downloads feature. It was then opened in Excel and saved as a CSV. The downloaded file contains Approved Symbols, Previous Symbols, and Alias Symbols.

Initial Plan and Data Manipulation

Our initial plan was to start with the DepMap Genetic Dependency CRISPR dataset and identify genes with interesting dependency profiles for further investigation within the TCGA datasets.

The Genetic Dependency CRISPR dataset profiles the effect of knocking out 17646 genes in 990 cancer cell lines. Negative numbers indicate cell death was observed when the gene was knocked out. By convention, a score of -1 or lower is used to distinguish that the removal of the gene is lethal to the cell line.


Gene Name
Cell Line

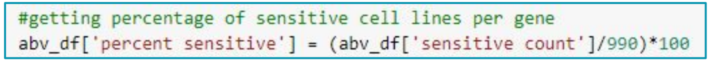
	DepMap_ID	A1BG (1)	A1CF (29974)	A2M (2)	A2ML1 (144568)	A3GALT2 (127550)	A4GALT (53947)
0	ACH-000001	-0.334969	-0.061580	-0.026897	-0.026507	-0.129643	0.060688
1	ACH-000004	0.020107	-0.000410	-0.055257	-0.071736	-0.088479	-0.233373
2	ACH-000005	-0.191303	0.086000	0.235074	0.068524	-0.286711	-0.337232

We first wanted to disregard genes which were lethal to all cell lines (“panlethal”). The rational being that if a gene were essential to all cell types, targeting it in a patient would lead to toxicities because that gene would be necessary for the survival of normal cells and tissues as well. We also wanted to eliminate genes which had no effect on any cell lines (“insensitive”) because targeting that gene would not inhibit cancer growth.

To eliminate the panlethal and insensitive genes we:

1. Transposed the dataframe so genes were in the rows
2. Determined the minimum and maximum score for each row
3. Dropped rows where maximum was < -1 as this indicates no cell lines survived knockout of the gene (i.e. panlethal, all sensitive)
4. Dropped rows where minimum was > -1 as this indicates no cell lines were killed by the knockout of that gene (i.e. all insensitive)

We then wanted to select genes which had a good mix of sensitive and insensitive cell lines. We did this by creating a column with the calculated percentage of sensitive cell lines per gene

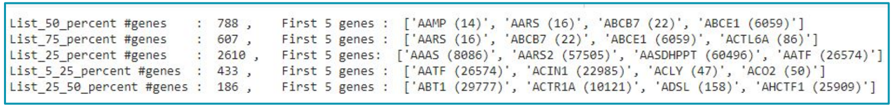


```
#getting percentage of sensitive cell lines per gene
abv_df['percent sensitive'] = (abv_df['sensitive count']/990)*100
```

We next wanted to examine different profiles of gene dependencies that were somewhere in between panlethal and insensitive as we thought these would be the most interesting for following up as potential targeted therapies.

These groups were defined as:

1. Over 25% of cell lines were sensitive
2. Over 50% of cell lines were sensitive
3. Over 75% of cell lines were sensitive
4. Between 5 – 25% of cell lines were sensitive
5. Between 25 – 50% of cell lines were sensitive

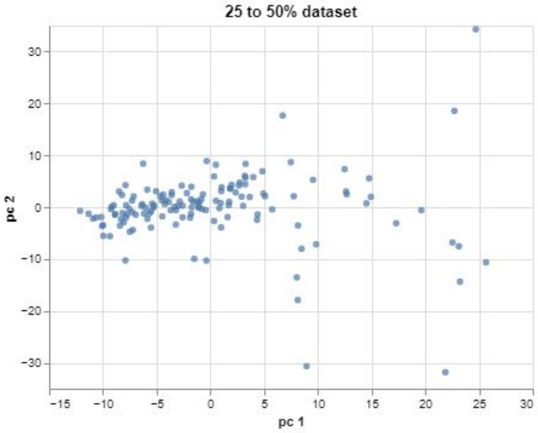


```
List_50_percent #genes    :  788 ,    First 5 genes :  ['AAMP (14)', 'AARS (16)', 'ABCB7 (22)', 'ABCE1 (6059)']
List_75_percent #genes    :  607 ,    First 5 genes :  ['AARS (16)', 'ABCB7 (22)', 'ABCE1 (6059)', 'ACTL6A (86)']
List_25_percent #genes    :  2610 ,   First 5 genes:  ['AAAS (8086)', 'AARS2 (57505)', 'AASDHPPT (60496)', 'AATF (26574)']
List_5_25_percent #genes  :  433 ,    First 5 genes :  ['AATF (26574)', 'ACIN1 (22985)', 'ACLY (47)', 'ACO2 (50)']
List_25_50_percent #genes :  186 ,    First 5 genes :  ['ABT1 (29777)', 'ACTR1A (10121)', 'ADSL (158)', 'AHCTF1 (25909)']
```

Using PCA to Visualize Genetic Dependency and Patient Samples

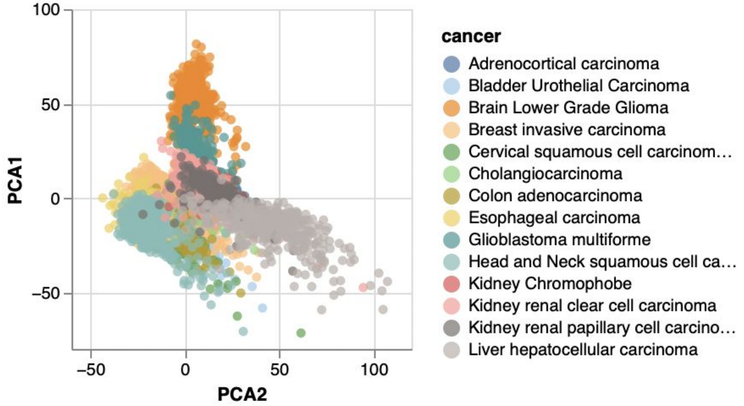
Principal Component Analysis (PCA) is a method of analysis that reduces the dimensionality of the data while trying to preserve variance. To do this, the data is projected to a lower dimensional space using Singular Value Decomposition. Performing PCA on the CRISPR dataset allows us to explore the data by visualizing the principal components on a scatterplot in order to potentially identify clusters.

We ran PCA on the 5 subsets of genes previously described in order to identify interesting patterns of gene dependencies using PCA from sklearn and visualized them using Altair. These PCA's were unable to show distinct clusters in order to identify genes that could potential targets for cancer treatment. The PCA on the subset of genes with 25 to 50% cell line sensitivity showed some genes that were outliers, but no distinct clusters. We decided to then run PCA on the TCGA dataset in order to identify any clustering within the patient tumor samples that could potentially be targeted in similar ways.



Because we did not see well defined clusters on the DepMap Dependency data, we revised our analysis plan. We chose to use PCA to identify interesting clusters in the primary patient data from TCGA and find the features that best predicted those clusters so that we could map them back to the DepMap CRISPR data. Then we would identify which CRISPR dependencies were most closely related to those specific clusters in the patients.

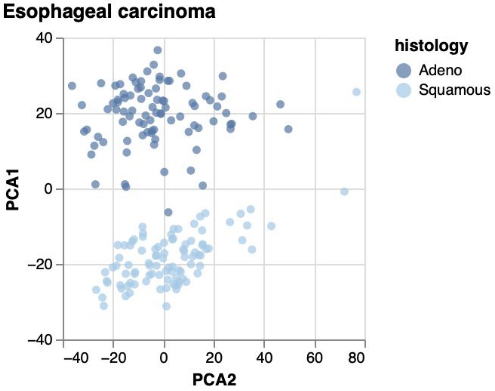
For the PCA on the TCGA expression data, we dropped the missing values using pandas dropna(), calculated the variance of each gene across all samples, and made plots for the first five principal components of the top 5000 variant genes against each other. Because Altair has a limit of 5000 rows, we visualized the first 5000 cancers when sorted by cancer type. These clusters were colored by cancer type (lineage) to better visualize the clusters. Unfortunately, many of the clusters overlapped with each other and they were hard to differentiate in this space, as shown below.



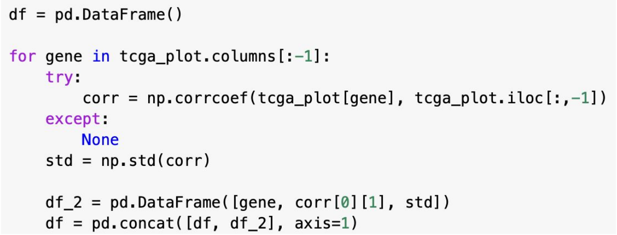
Characterizing Patient Data using Histology and Gene Expression

Because the PCA plots of all of the patient data together did not yield clear clusters for us to better characterize, we decided to perform individual PCA analyses for each cancer type and color the samples by histological subtypes to see if this produced better defined clusters.

We again performed the PCA clustering using the RNAseq expression data, as in the previous slide. One visualization that stood out was the clustering of the adenocarcinoma and squamous cell carcinoma types in esophageal cancer. As shown in the figure below, there were two clear clusters.

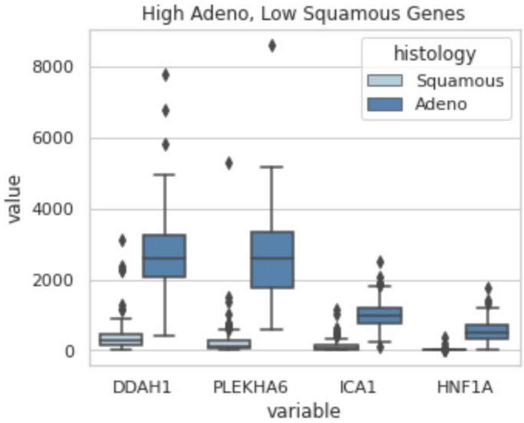


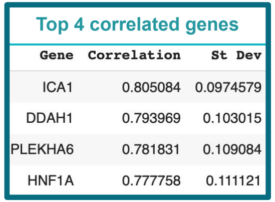
Next, we encoded the histology types (squamous was 0, adenocarcinoma was 1) and the correlation between histology and expression of each gene in the TCGA Gene Expression dataset was measured using numpy correlation function. Some genes were present more than once and had different values, so in order to skip those we used a try and except to only calculate the correlation if the gene was present one time.



```
df = pd.DataFrame()

for gene in tcga_plot.columns[:-1]:
    try:
        corr = np.corrcoef(tcga_plot[gene], tcga_plot.iloc[:,-1])
    except:
        None
    std = np.std(corr)

    df_2 = pd.DataFrame([gene, corr[0][1], std])
    df = pd.concat([df, df_2], axis=1)
```

The top four genes that correlated with the esophageal histology types were ICA1, PLEKHA6, DDAH1 and HNF1A (shown below). Because ICA1 had the best correlation with esophageal histology type at 0.81, we chose to use that in the next part of the analysis. Of note, we also looked for negative correlations but the magnitude of the positive ones were larger.





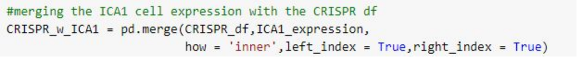
Top 4 correlated genes

Gene	Correlation	St Dev
ICA1	0.805084	0.0974579
DDAH1	0.793969	0.103015
PLEKHA6	0.781831	0.109084
HNF1A	0.777758	0.111121

Linking the Datasets – Return to DepMap

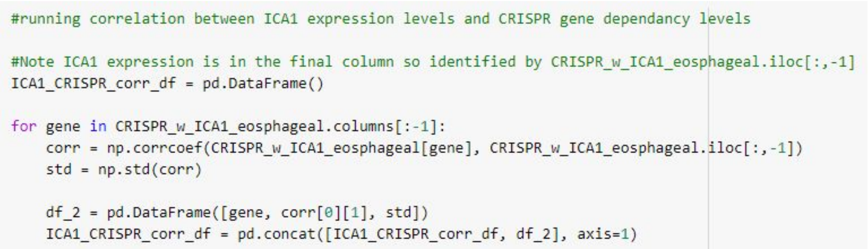
Following the identification that esophageal cancer subtypes showed the most distinct clustering, and that the gene ICA1 correlated best with this division between the subtypes per the primary TCGA dataset, we returned to investigate esophageal cancers and ICA1 in the cancer cell line DepMap database. We wanted to determine if there were any notable correlations between ICA1 expression levels and genetic dependency of any gene, with a focus on esophageal cancer cell lines.

First, expression levels of ICA1 in all cell lines was extracted from the Cellular Models – Expression dataset. ICA1 expression levels were merged into the Genetic Dependency CRISPR dataset so that now we had a column with ICA1 expression in all cell lines alongside the genetic dependency for all genes for these cell lines.



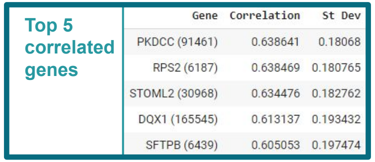
```
#merging the ICA1 cell expression with the CRISPR df
CRISPR_w_ICA1 = pd.merge(CRISPR_df,ICA1_expression,
                         how = 'inner',left_index = True,right_index = True)
```

Next, we determined the correlation coefficient between ICA1 expression and genetic dependency of each gene and sorted the results to find the highest correlations to identify genes of interest.



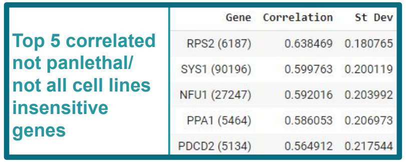
```
#running correlation between ICA1 expression levels and CRISPR gene dependancy levels

#Note ICA1 expression is in the final column so identified by CRISPR_w_ICA1_eosphageal.iloc[:,-1]
ICA1_CRISPR_corr_df = pd.DataFrame()

for gene in CRISPR_w_ICA1_eosphageal.columns[:-1]:
    corr = np.corrcoef(CRISPR_w_ICA1_eosphageal[gene], CRISPR_w_ICA1_eosphageal.iloc[:,-1])
    std = np.std(corr)

    df_2 = pd.DataFrame([gene, corr[0][1], std])
    ICA1_CRISPR_corr_df = pd.concat([ICA1_CRISPR_corr_df, df_2], axis=1)
```

We moved forward with the five genes where we observed the greatest positive correlation between genetic dependency and ICA1 expression. We also performed the correlation between ICA1 expression and the selection of genes we had previously identified as not panlethal/not all cell lines insensitive. The top result was RPS2, which was identified already above, but we included the second result, SYS1, in our visualisations. Furthermore, we included AP2M1 which exhibited negative correlation.



Top 5 correlated genes

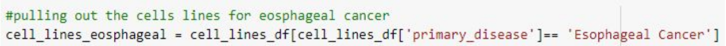
Gene	Correlation	St Dev
PKDCC (91461)	0.638641	0.18068
RPS2 (6187)	0.638469	0.180765
STOML2 (30968)	0.634476	0.182762
DQX1 (165545)	0.613137	0.193432
SFTPB (6439)	0.605053	0.197474



Top 5 correlated not panlethal/ not all cell lines insensitive genes

Gene	Correlation	St Dev
RPS2 (6187)	0.638469	0.180765
SYS1 (90196)	0.599763	0.200119
NFU1 (27247)	0.592016	0.203992
PPA1 (5464)	0.586053	0.206973
PDCD2 (5134)	0.564912	0.217544

Cell lines derived from esophageal cancers were identified in the Cellular Models – Cell Line Sample Info dataset. There were 28 such cell lines – 22 from the squamous subtype and 6 from the adenocarcinoma subtype.



```
#pulling out the cells lines for eosphageal cancer
cell_lines_eosphageal = cell_lines_df[cell_lines_df['primary_disease']== 'Esophageal Cancer']
```

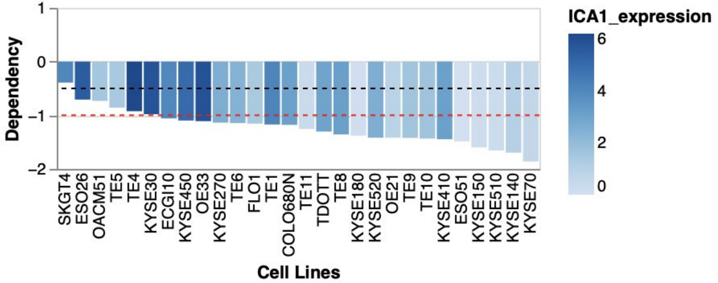
We wanted to investigate the dependence of the eosphageal cell lines on the selected genes of interest. To do this, we graphed waterfall plots of the genetic dependency of the 28 cell lines for each of the 6 chosen genes. Waterfall plots are a type of bar chart with one categorical axis and one numerical axis. The data can be sorted in descending order so that it's easy to see trends of the variable being plotted. For our plots, the cell lines are the categorical variable and were plotted on the x-axis. The genetic dependency was the numerical variable on the y-axis. We coloured each bar by the underlying esophageal cancer subtype (squamous or adenocarcinoma) or ICA1 expression levels, and included cut-off lines to show where the genetic dependency was -0.5 and -1 (the cut-off showing if removing the gene is lethal to the cell line)

Visualizing ICA1 Expression, Histology Types, and Gene Dependencies

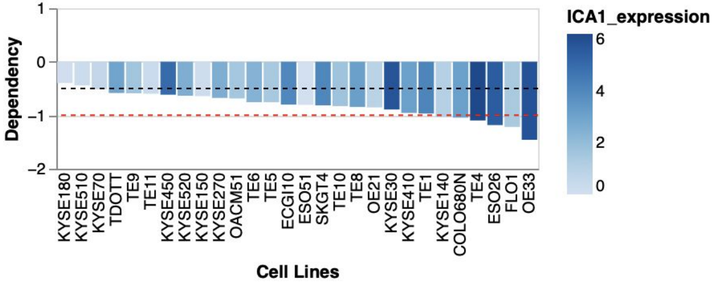
Of the genes investigated, 3 genes (PKDCC, STOML2 and DQX1) did not met the -0.5 cut-off for genetic dependency in any of the eosphageal cancer cell lines. This indicates that targeting these genes would have minimal impact as a therapeutic avenue. One gene, RPS2, crossed the lethality threshold of -1 in all cell lines, meaning that it was lethal to all types of esophageal cancer and therefore not a promising target for a specific targeted therapy against a certain subtype. For SFTPB, a single cell line breached the -0.5 cut-off, but no notable pattern was seen for the two cancer subtypes.

The most interesting genetic dependency patterns observed were for SYS1 and AP2M1, where there was a spread of cell lines over and under the -1 threshold. SYS1 dependency correlated best with low ICA1 expression and squamous subtype whereas AP2M1 dependency correlated with high ICA1 expression and adenocarcinoma subtype. The plots below were generated using Altair and either colored by ICA1 expression or histology type to demonstrate this.

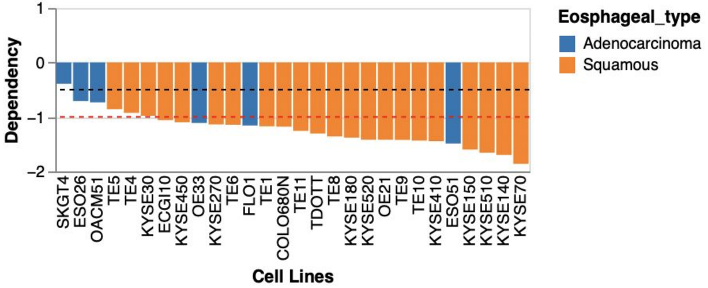
SYS1 Dependency



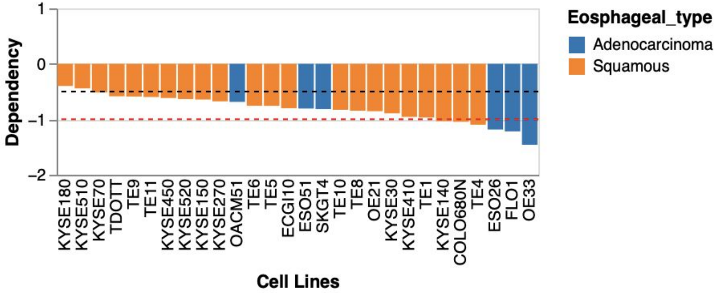
AP2M1 Dependency



SYS1 Dependency



AP2M1 Dependency



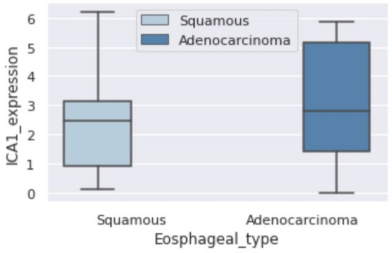
Learnings and Next Steps

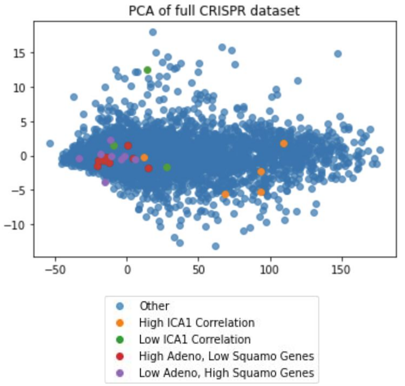
In the project, we encountered a number of challenges and achieved some interesting findings which would be worth further exploration.

Our first major challenge was the lack of clear clustering per the PCA on the DepMap data. We overcame this by reversing our approach and deciding to use the primary patient TCGA dataset as our starting point and linking back to the DepMap cell line data instead. We also faced issues in running the PCAs due to our vast dataset but were able to address this by changing our plotting tool from Altair to Matplotlib. Overall, this project taught us the importance of being adaptable and researching different tools and techniques.

The aim of our project was to identify genetic dependency markers within cancer histological subtypes. Our analysis flagged esophageal cancer subtypes as displaying notable differences in gene expression patterns and highlighted ICA1 as a gene of interest in the primary patient cell dataset. Further investigation identified two genes, SYS1 and AP2M1, as exhibiting interesting genetic dependency patterns within the esophageal cancer cell lines.

We have made some preliminary additional investigations into ICA1 in the cell line (DepMap data) and found it’s expression does not distinguish the esophageal subtypes as well as it does in the patient samples (TCGA data). As shown in the boxplot, there is higher expression of ICA1 in the adenocarcinomas but it is not very pronounced and the correlation was only 0.14 as compared to 0.81 in the TCGA data. Furthermore, when we highlighted our genes of interest from the TCGA dataset in a PCA from the full CRISPR dataset, we found that genes that were ‘high in esophagus adenocarcinoma, low in esophagus squamous’ and ‘low in esophagus adenocarcinoma, high in esophagus squamous’ tended to be more similar to each other than genes that had high ICA1 correlation.





There are many possible reasons for the differences between the patient and cell line data. First, it should be noted that there are only 6 adenocarcinoma cell lines out of the 28 total esophageal cell lines in the DepMap data. Perhaps if the sample number was larger, a better correlation would be seen. Additionally, it is possible that ICA1 is not a robust feature of histological type *in vitro*, or that the cell lines in the DepMap dataset are not very representative of actual patient samples.

Given the observed differences, the potential next step could be to compare our cancer cell lines to the primary patient samples to determine which actually most closely align in terms of overall gene expression, and then select the cell lines that best represent the patient samples so that any learnings from the *in vitro* systems would have a better chance at success in real world settings. We could then establish if any trends with SYS1 and AP2M1 persist, or if different genes are flagged as warranting further investigation.

Statement of work

Collaboration (all)

- Google Colab notebooks for code sharing
- Meet regularly to discuss progress and challenges
- Write and edit final report

DepMap data

- Clean and join files in dataset (**Sinéad**)
- Perform summary statistics to identify gene targets (**Sinéad**)
- Perform PCA to cluster genes (**Kristina**)
- Create waterfall plot to illustrate cell line dependency on the genes (**Sinéad** and **Alyson**)
- Perform PCA to visualize cell line clusters following analysis (**Kristina**)

TCGA data

- Clean, manipulate, and join files in dataset (**Alyson**)
- Perform PCA to cluster patient samples (**Alyson**)
- Perform correlation analysis and visualizations between gene expression and histology (**Kristina** and **Alyson**)

References

- "Cancer." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/cancer.
- Galarnyk, Michael. "PCA Using Python (Scikit-Learn)." Medium, Towards Data Science, 4 Dec. 2017, https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- "Genomic Data Commons Data Portal." National Cancer Institute, https://portal.gdc.cancer.gov/
- Kobak, D., Berens, P. The art of using t-SNE for single-cell transcriptomics. Nat Commun 10, 5416 (2019). https://doi.org/10.1038/s41467-019-13056-x
- Noorbakhsh, J., Vazquez, F. & McFarland, J.M. Bridging the gap between cancer cell line models and tumours using gene expression data. Br J Cancer 125, 311–312 (2021). https://doi.org/10.1038/s41416-021-01359-0
- "Personalized Medicine: Redefining Cancer Treatment." Kaggle, www.kaggle.com/c/msk-redefining-cancer-treatment/overview/description.
- "Targeted Cancer Therapies Fact Sheet." National Cancer Institute, 15 Sept. 2021, www.cancer.gov/about-cancer/treatment/types/targeted-therapies/targeted-therapies-fact-sheet.