Bioinformatic Analysis of Cellular Proteomic Data

STA130 Course Project

Mahe Chen and Matthew Yu

December 8, 2022

Project Description:

- The purpose of this project is to apply statistics in bioinformatic research work to help fight cancer!
- Our data is based on advances of single cell analysis in the fields of Flow Cytometry and cellular proteomic processes in the fields of Mass Spectrometry.
- Based on the above analysis, we will measure the multivariate landscape of proteomic activity for a single cell in any experimental condition for any cell type at scale.
- We will explore how to direct deleterious cellular states to transition into non-deleterious states by analyzing the typical cellular homeostatis of healthy and deliterious cells and the phenotypical transformation of cellular proteomic homeostatsis over time in response to different experimental conditions.

Our Plan

- By recognizing correlations and patterns between protein makeup(most specifically our outcome proteins) of a cell under some condition, we can classify the cell type and label the conditions as a possible causation
- Analyze protein levels of cells over time and under different treatment conditions
- Protein structure data to recognize patterns of healthy cell states turning deleterious, allow us to interfere with treatment as soon as possible
- Measure protein structure with Mass Spectrometry and Flow Cytometry Each cell, under some condition and at some time, measures levels of 22 AP-1 transcription factors and 4 outcome phenotype proteins.
- Destroy cells to take measurements
- Split cells into groups and take measurements from each group

Project Information

- The grid below illustrates what melanoma cells(which our skin cancer cells can spread to the rest of the body) can appear as during their differentiation states, meaning they are changing their function type from stem cell to specific cell.

```
##                    MiTFg NGFR SOX10 AXL
## Undifferentiated   Low   Low  Low   High
## Neural crest-like  Low   High High  High
## Transitory         High  High High  Low
## Melanocytic        High  High Low   Low
```

- Some background information regarding our data set:
  - We are going to narrow our focus to the 4 phenotype protein levels
  - Measurements are taken at time intervals following administration of treatment
    - 0.5, 2, 6, 15, 24, 72, 120h
    - Drugs include Vem or Vem and Tram
    - Doses measured in micrometress

Project Information

- Cellular Phenotype
  - Undifferentiated are like stem cells, these tissues do not have a specialised function yet (Not yet a "mature" cell). Melanoma cells that appear as this are challenging to diagnose and rapidly grow and spread
  - Neural crest-like melanoma cells behave very similarly to neural crest cells(which are normal healthy cells used in body regeneration) during their early stages of spreading and invasiveness, and thus can be challenging to detect. Derived from various tissue or stem cells, unlike neural crest cells.
  - Transitory cells - critical white blood cell (immune cell) that targets and kills antigens and cancer cells
  - Melanocytic cells give melanin, or tanning color, and is where melanoma cancer cells can develop.
- Phenotype Proteins:
  - MiTFg - involved with melanocyte protein and crucial for many cells' function
  - NGFR - nerve growth factor receptor
  - SOX10 - important in the development of several things, most relevantly formation of melanocytes
  - AXL - pushes cells to divide, can get out of hand in case of cancer, while less of it promotes apoptosis(cells kill themselves)

Project Objectives

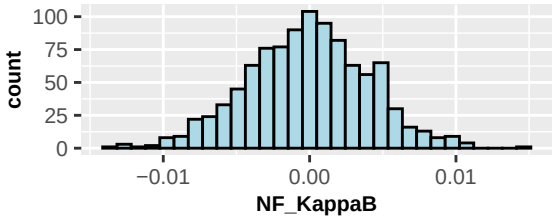
Our project will be able to answer the following questions:

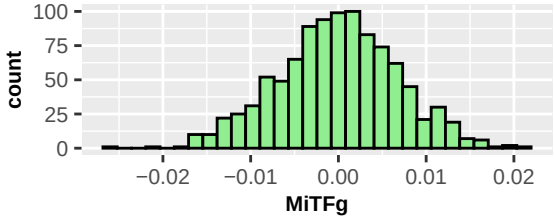
1. Do phenotype protein levels in experimental condition 'x' change over time 't'?
   - Method: Two Sample Hypothesis Testing
2. Do phenotype protein levels at a time 't' change between experimental conditions x1 and x2?
   - Method: Two Sample Hypothesis Testing
3. At a time of 0.5h with a dose of 0uM, what are the relationships between different proteins?
   - Method: Confidence Intervals
4. Can we predict cellular phenotype outcomes (Y) values/states from transcription factors?
   - Method: Regression and Correlation
5. Can we determine resulting cellular phenotype depending on levels of 4 phenotype proteins?
   - Method: Regression and Correlation

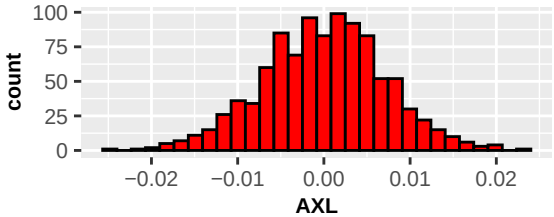
Analysis

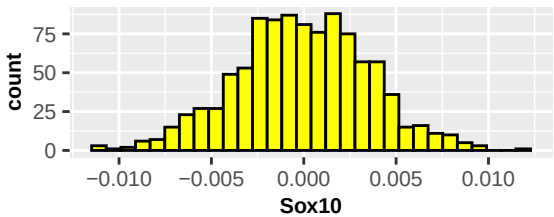
Formula: $H_0 : p_1 = p_2 \implies H_0 : p_1 - p_2 = 0$

Do proteins in experimental conditions Drug(Vem) and Dose(1uM) change over time(0.5h-120h)?









Explaination

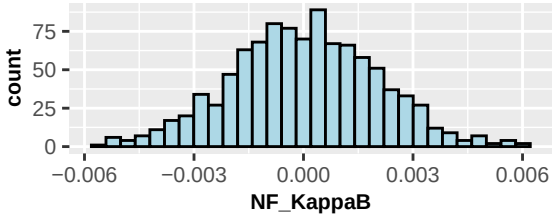
Formula: $H_0 : p_1 = p_2 \implies H_0 : p_1 - p_2 = 0$

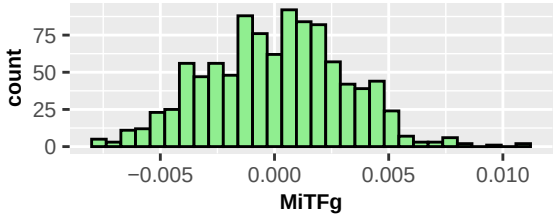
- Do proteins in experimental conditions Drug(Vem) and Dose(1uM) change over time(0.5h-120h)?
  1. We first assume the two populations are the same
  2. Then, use the permutation test in which we shuffle the two samples to determine if there's a difference in means based on the two populations.
  3. Use observed test statistics x1-x2 based on n1 and n2 sample, simulate the Sampling Distribution assuming the NULL Hypothesis is TRUE.
     - Note: the test statistics will be the difference between the statistics of the two groups.
     - Null hypothesis is a statement of equivalence.
     - Re-sample/shuffle both samples and calculate the test statistic each time.
- Conclusion:
  - The Two Hypothesis Test based on NF_Kappa_B, MiTFg, AXL and Sox10 in experimental conditions Drug(Vem) and Dose(1uM) at timepoint 0.5h and 120h showed p-values of 0, observed statistics of 0.24(NF_KappaB), 0.48(MiTFg), 0.44(AXL), 0.14(Sox10). A p-value of 0 means there is no chance of observing results at least as extreme. We will set an a-significance level of 0.05(also the probability of a Type I error of rejecting a true H0) and we reject H0 at the a-significance level since $0 < 0.05$.

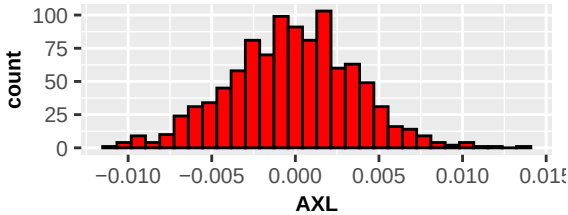
Analysis

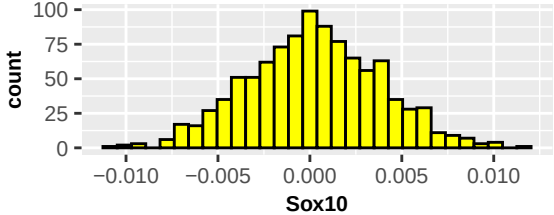
Formula: Formula: $H_0 : p_1 = p_2 \implies H_0 : p_1 - p_2 = 0$

Are protein levels at time 2h different between experimental conditions Vem and Vem+Tram?









Explaination

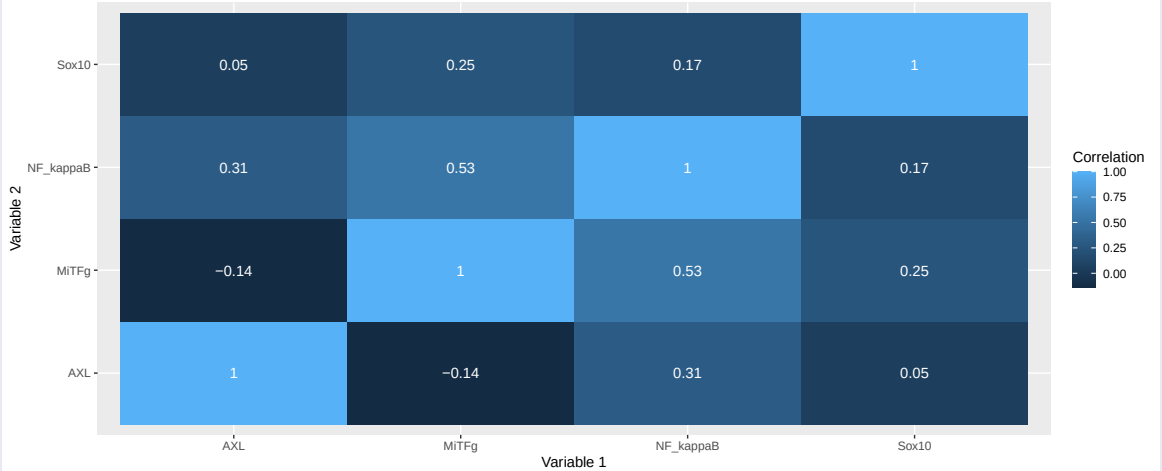
Formula: Formula: $H_0 : p_1 = p_2 \implies H_0 : p_1 - p_2 = 0$

- Are protein levels at time 2h different between experimental conditions Vem and Vem+Tram
  - The method we are using is the same as the previous one.
- Conclusion:
  - The Two Hypothesis Test is based on NF_Kappa_B, MiTFg, AXL and Sox10 in experimental conditions Drug(Vem) and Dose(0uM) and Drug(Vem + Tram) and Dose(0uM) at timepoint 2h. It showed a p-value of 0 for NF_Kappa_B, AXL and Sox10, and a p-value of 0.74 for MiTFg. The observed statistics are -0.01(NF_Kappa_B), 0.001(MiTFg), -0.02(AXL), -0.1(Sox10). A p-value of 0 means there is no chance of observing results at least as extreme. A p-value of 0.72 means there is 72% chance of observing results at least as extreme. We will set an a-significance level of 0.05(probability of a Type I error of rejecting a true H0) and we reject H0 at the a-significance level since $0 < 0.05$, failed to reject H0 at the a-significance level since $0.05 < 0.72$.

Correlation Matrix

Formula: $H_0 : \mu = m_0 \quad | \quad x_i \sim N(\mu, \sigma) \quad | \quad \$H\_0:$

At time 120h in experimental condition Vem and 1uM, what is the relationship between different proteins?

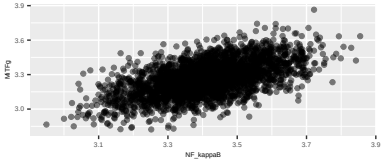


Linear Association

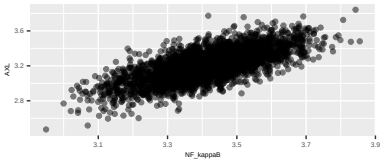
Formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{(n-1)s_{xy}}{\sqrt{(n-1)s_x^2(n-1)s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{\mathrm{Cov}(x,y)}{\mathrm{SD}(x)\mathrm{SD}(y)}$$

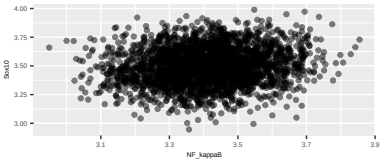
Correlation: 0.57



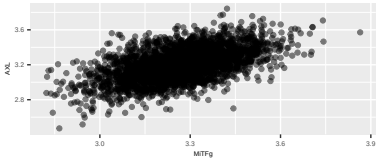
Correlation: 0.71



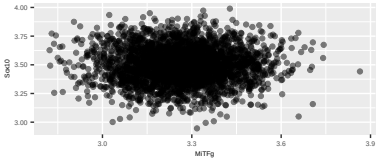
Correlation: 0.14



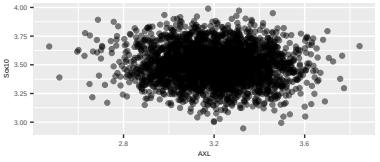
Correlation: 0.54



Correlation: −0.01



Correlation: −0.07



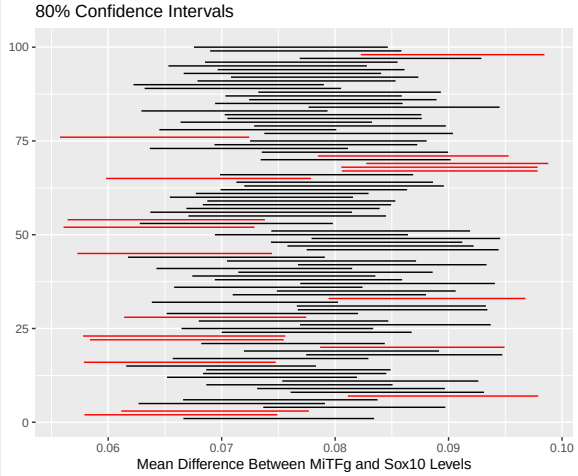
Corelation Explained:

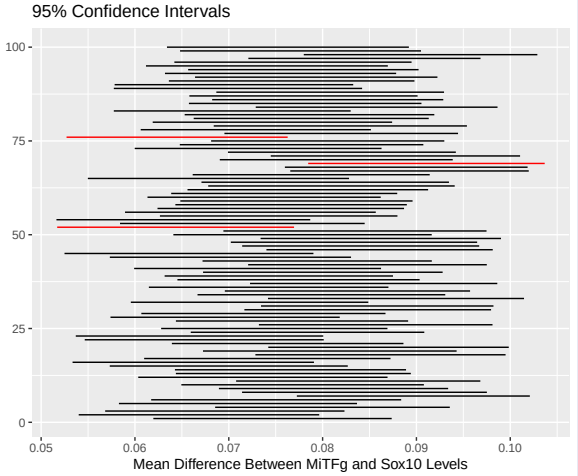
At time 120h in experimental condition Vem and 1uM, what is the relationship between different proteins?

- Our ideas
  - The denominator scales the numerator so that the total $-1 \leq r \leq 1$ always
  - r measures *linear association*, with $r > 0$ *positive* and $r < 0$ means *negative*
- As we can see from the graph:
  1. NF_kappa_B is somewhat associated with MiTFg. (Positive i.e. Increasing Correlation of 0.57)
  2. NF_kappa_B is strongly associated with AXL. (Positive i.e. Increasing Correlation of 0.71)
  3. NF_kappa_B is weakly associated with Sox10. (Positive i.e. Increasing Correlation of 0.14)
  4. MiTFg is somewhat associated with Sox10. (Positive i.e. Increasing Correlation of 0.54)
  5. MiTFg is very weakly associated with AXL. (Negative i.e. Decreasing Correlation of -0.01)
  6. AXL is very weakly associated with Sox10. (Negative i.e. Decreasing Correlation of -0.07)

Bootstraping and Confidence Intervals

Formula: $\bar{x} \pm z \frac{s}{\sqrt{n}}$





Confidence Intervals Explained

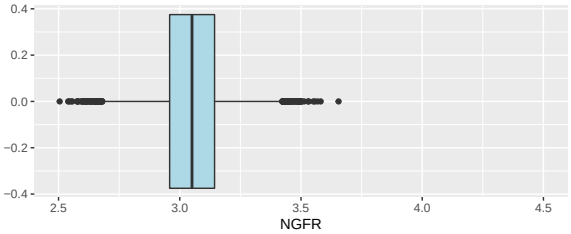
Relationships Between Different Proteins at time of 0.5h with dose 0uM

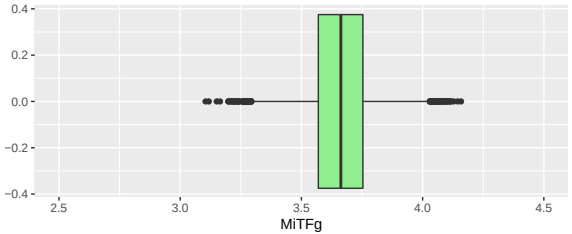
- With confidence intervals, we can estimate many parameter values between our phenotype proteins. For example, if we were to keep all other factors constant(so in this case we are choosing a time point of 0.5h, drug dose of 0uM), we could estimate the difference between means of phenotype proteins MiTFg and Sox10.
- We performed bootstrapping by taking 2 samples, one for each phenotype protein level, from the population with our condition/time restrictions.
- Then we resample(with replacement) from each sample, and calculate the difference between the means of the 2 phenotype protein levels.
- As our example we used the last bar for our 95% confidence interval and we can say we are 95% confident that the interval (0.06317026, 0.08802395) will contain the true difference between the means. This is a pretty tight interval with high confidence, so this is an accurate and appropriate estimation for the true parameter.
- I've simulated this bootstrapping 100 times to demonstrate the idea of 95% confidence(which are signified by the black intervals, making up the majority) and we could theoretically repeat this process between different proteins, conditions, and confidence intervals to get a better understanding of the relationships between protein levels

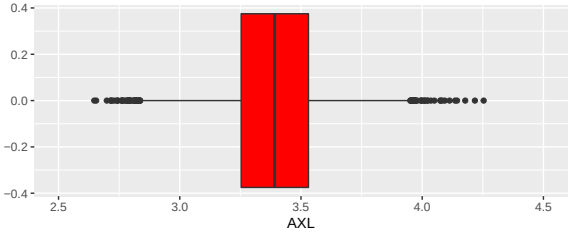
Predicting Cellular Phenotype Outcomes From Transcription Factors

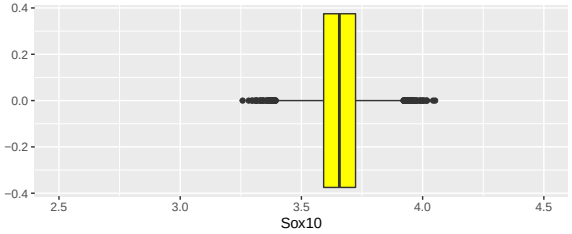
- Here we got R to calculate the pearson correlation coefficient between all 26 proteins(4 phenotype and 22 transcriptions) with conditions of Vem drug, 15 hours after administration, and dose of 3.16 uM.
- Keeping in concise, the correlation between 2 proteins is calculated by Cov(x,y) $\div$ SD(x)SD(y), where we divide the covariance(squared difference in values) by the standard deviation in order to standardize the value
- We also decided to classify the phenotype values into "High" and "Low" categories (continuous data $\rightarrow$ discrete data) to fit a classification model on it
- The graphs below depict the distribution of the value of phenotype proteins in the cell (which follows normal distribution)
- The distribution for phenotypes across repetitions is generally similar, therefore we can use the data across repetitions to classify any values for the phenotypes above the median as "high" and below the median as "low"

An Overview of Means of the indicator proteins









Continuing with Multivariate Linear Regression and Classification

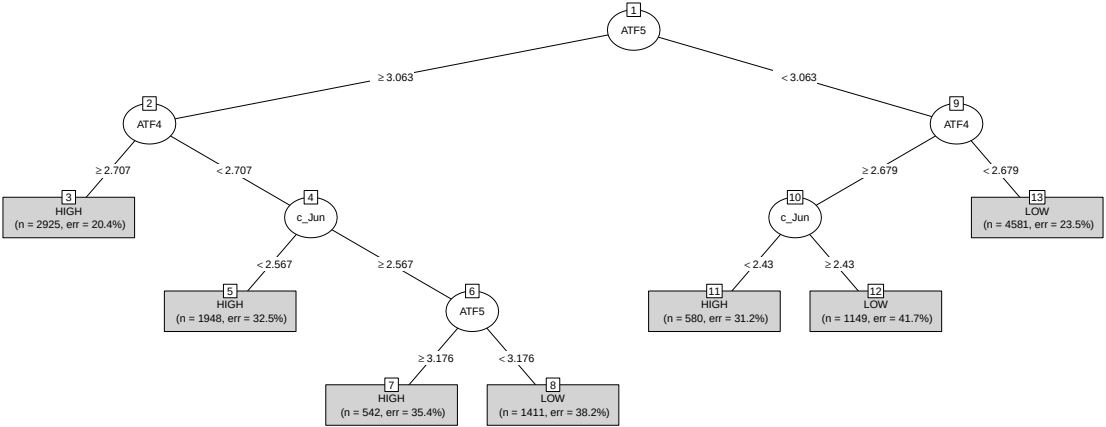
- Utilizing the same conditions as above, we can create multivariate linear regression models using all available proteins as our explanatory variables. In real life, it may be costly to have to record and calculate data for all 25 protein levels just to predict 1 protein, but we are already provided with full data on all 26 protein levels. Since we are only predicting the phenotype values we don't need to worry about multicollinearity.
- Another concern that may arise with this many explanatory variables is overfitting, but we perform an 80-20 train split test(save 80% of data to build our models on, and test them out on the new 20% of the rest of data), then calculate the root mean square error(RMSE) to estimate the spread of our error. And the results show us the RMSE is quite low for the linear models of all 4 phenotype proteins. NGFR has the best model at about 0.0795 RMSE and AXL was the predictive model at about 0.137 RMSE

```
## NGFR RMSE: 0.08 | MiTFg RMSE: 0.09 | AXL RMSE: 0.14 | Sox10 RMSE: 0.08
```

Classification

- Below we built a classification tree based on all 25 explanatory variables as the covariates for MiTFG levels, we have decided to use all 25 seeing as we have access to all 25, however, this may have diminishing returns. It is not too complicated due to the stopping rules that don't allow trees to split if there is not a significant increase in classification. Similarly, we can repeat this for all the other proteins.
- The calculated accuracy(proportion we got correct), precision(proportion of ones we identified as high that are correct), sensitivity(proportion of actually high levels we identified as high), and specificity(proportion of actually low we identified as low) are all in the 0.61 to 0.77 range, indicating our tree is a solid predictor of these values
- Finally, we used to trained classification model to predict the levels of the phenotypes (High or Low) and used those values to classify the types of cells (Undifferentiated, Transitory, etc)

Decision Tree for MiTFg



Accuracy: 0.71 | Precision: 0.67 | Sensitivity: 0.74 | Specificity: 0.68

Conclusions Visualized

- The answer to Question 4 is yes. We can see that the low RMSE values for each phenotype and the fairly high accuracy of the classification model implies that our model fits the data well. For example, classifying a cell as melanocytic may suggest cells in early stages of cancer.

```
## # A tibble: 3,284 x 5
##    NGFR_level_hat MiTFg_level_hat AXL_level_hat Sox10_level_hat `Type of cell`
##    <fct>          <fct>           <fct>         <fct>           <chr>
##  1 LOW            LOW             HIGH          LOW             Undifferentiated
##  2 LOW            LOW             LOW           LOW             <NA>
##  3 HIGH           LOW             LOW           LOW             <NA>
##  4 LOW            LOW             LOW           LOW             <NA>
##  5 HIGH           HIGH            HIGH          HIGH            <NA>
##  6 HIGH           LOW             HIGH          HIGH            Neural crest-li~
##  7 HIGH           HIGH            HIGH          HIGH            <NA>
##  8 HIGH           LOW             LOW           LOW             <NA>
##  9 LOW            LOW             HIGH          LOW             Undifferentiated
## 10 LOW            LOW             HIGH          LOW             Undifferentiated
## # ... with 3,274
```

Project Acknowledgements

- Our ideas come from Dr. Scott Schwartz who is a seasoned professional at the University of Toronto worked in Integrative Biology, Nutrition and Complex Disease, and Next Generation Sequencing labs. Link: https://github.com/pointOfive
- Our data come from the article (Refer to the below link) finding that the "AP-1 transcription factor network" (i.e., the relative distributions and dependency relationships of transcription factors) are predictive of "cellular plasticity in melanoma" (i.e., how easily changeable the phenotype are melanoma cell lines) Link: https://www.biorxiv.org/content/10.1101/2021.12.06.471514v1.full

Project References

- NCI Dictionaries, National Cancer Institute
  - https://www.cancer.gov/publications/dictionaries
- The Neural Crest and Cancer: A Developmental Spin on Melanoma, National Library of Medicine
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3809092
- What Is Melanoma Skin Cancer?
  - https://www.cancer.org/cancer/melanoma-skin-cancer/about/what-is-melanoma.html
- AP-1 transcription factor network explains diverse patterns of cellular plasticity in melanoma
  - https://www.biorxiv.org/content/10.1101/2021.12.06.471514v1.full
- MITF gene
  - https://medlineplus.gov/genetics/gene/mitf
- SOX10 gene
  - https://medlineplus.gov/genetics/gene/sox10/
- Nerve Growth Factor Receptor
  - https://www.sciencedirect.com/topics/medicine-and-dentistry/nerve-growth-factor-receptor
- AXL receptor tyrosine kinase as a promising anti-cancer approach
  - https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-019-1090-3