# Analysis of Proteins and Transcription Factors

## A close examination of protein levels and their correlation among various experimental conditions

Chris (Jechang) Oh, Olivia (Ximei) Lin, Joseph (Cheng Peng), and Ethan Chan

December 8, 2022

What is the data?

1. Data observations consist of 22 AP-1 transcription factors (causes) and 4 phenotype proteins (outcomes), resulting in a total of 26 variables
2. Moreover, the data also contains 4 additional experimental conditions: drug ID, dosage ID, time, and repetitions

# Objectives

We will examine and analyze the following questions:

**Q1.** Do protein levels in experimental condition $x$ change over time $t$?

**Q2.** Are protein levels at time t different between experimental conditions $x_1$ and $x_2$?

**Q3.** At time $t$ in experimental condition $x$, what is the relationship between different proteins?

**Q4.** (Meta Analysis) What patterns can we observe from the results of the questions themselves?

# Q1 Data Summary

**Purpose of Data**: To examine and analyze how protein levels in experimental condition x change over time t.

- The experimental conditions we chose to analyse were the tests using drug ID 1 (Vem), with dosage IDs 1 (0uM), 3 (0.316uM), and 5 (3.16uM).

```
MiTFg_data <- data %>%
  filter((dose_id == 1 | dose_id == 3 | dose_id == 5) & drug_id == 1) %>%
  select(MiTFg, timepoint_id, Timepoint, drug_id, Drugs, dose_id,
  Doses, time)
```

- The relevant data for the MiTFg protein was filtered and stored in a variable called 'MiTFg_data'. In addition, an extra variable, `time`, was mutated from the original data to represent the timepoint as a numeric value.

# Q1 Statistical Method: Two Sample Hypothesis Test

To understand whether or not protein levels at two different time points were **significantly** different, we employed a 2-sample hypothesis test on MiTFg. Before the test, we set an alpha-significance level of $\alpha = 0.05$.

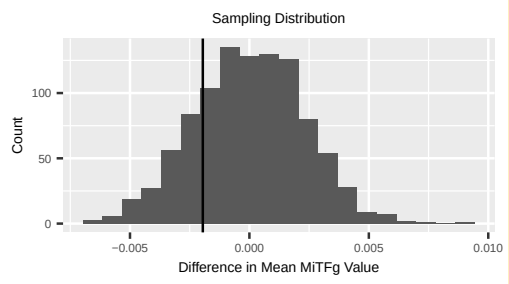
## Null Hypothesis

- $H_0$: Protein levels in experimental condition do not change over time from time point t.

## Alternative Hypothesis

- $H_1$: Protein levels in experimental condition change over time from time point t.

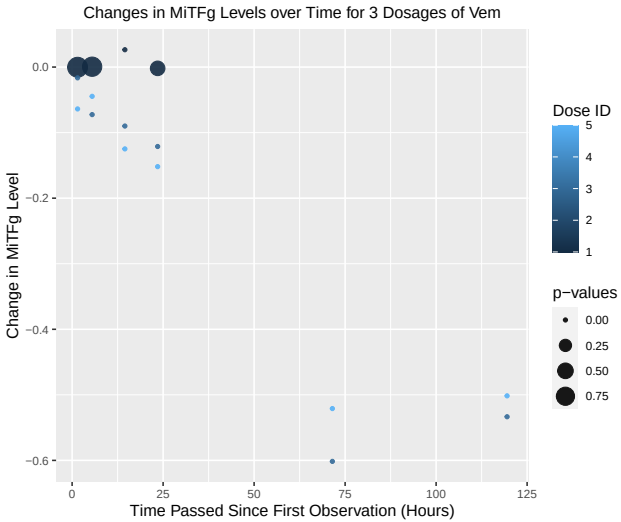
## Example of our Two Sample Hypothesis Test (Dose ID 1, timepoints 1 and 5)




```
## [1] -0.001951233
```

# Q1 Results

**The testing was repeated for all time points and our chosen experimental conditions:**



# Q1 Results

1. We observe that the p-values are significantly *above* the alpha-significance level for when the dose ID is 1, and the time passed is still on the lower end, meaning we **fail to reject the null hypothesis**, and conclude that there was no significant change in MiTFg levels. However, we note that for when the dose ID is 3 or 5, or the time passed is greater, the p-values are *below* the alpha-significance level, meaning we **reject the null hypothesis** and conclude that there was a significant change in MiTFg level for that experimental condition across time t.
2. As time grows, and as dosage increases, the change in MiTFg also increases. Furthermore, we note that these are statistically significant changes, as the p-value is near 0, so it seems like dosage and time are factors that have a strong negative effect on the level of MiTFg, which will be useful knowledge in affecting the balance of proteins and cellular homeostasis.

# Q2 Data Summary

**Purpose of Data**: To examine and analyze whether protein levels at time t are different between experimental conditions $x_1$ and $x_2$.

- The data was cleaned to focus specifically on the MiTFg protein. First, the data was filtered at time point 0.5h, and the experimental conditions were drug ID 1, dose ID 2 ($x_1$), and drug ID 1, dose ID 3 ($x_2$).

```
MiTFg_data <- data %>%
  filter(timepoint_id == 1) %>%
  select(MiTFg, timepoint_id, Timepoint, drug_id, Drugs, dose_id, Doses)

MiTFg_data <- data %>%
  filter(drug_id == 1 & dose_id == 2 | drug_id == 1 & dose_id == 3) %>%
  mutate(exp_cond = case_when(
  drug_id == 1 & dose_id == 2 ~ "x1",
  drug_id == 1 & dose_id == 3 ~ "x2"
))
```

# Q2 Statistical Methods

With a significance level of $\alpha = 0.05$, we will run a two sample mean hypothesis test to examine the difference in MiTFg levels under the two different experimental conditions.

## Null Hypothesis

- There is no mean difference at time point 0.5h of protein levels in condition $x_1$ and condition $x_2$

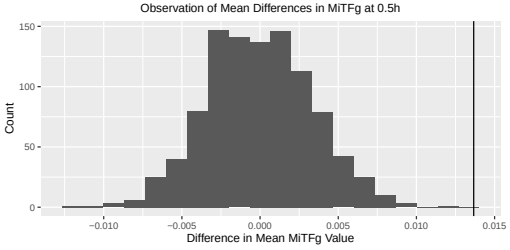
## Alternative Hypothesis

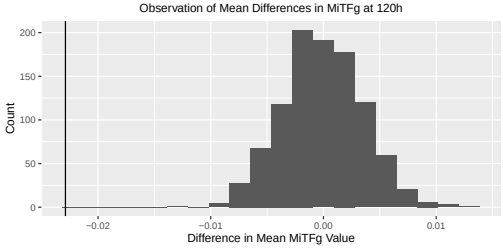
- There is a mean difference at time point 0.5h of protein levels in condition $x_1$ and condition $x_2$

# Q2 Statistical Methods

Assuming the null hypothesis is true, we calculate the **test statistic** and simulate a sampling distribution at time point 0.5h (left). Then, we generate another sampling distribution with the same process at time point 120h (right), with the experimental conditions listed above

- A comparison between the first and last time points within a drug would show the greatest differentiation, and allow more data to be analyzed. The second distribution has similar null and alternative hypotheses as the first.




```
## [1] 0.01366422
```

```
## [1] -0.02289173
```

# Q2 Results

The table summarizes the computed p-values, along with additional data computed from different experimental conditions not shown above.

**$x_1$: drug ID = 1, dose ID = 2. $x_2$: drug ID = 1, dose ID = 3**

| Conditions | Time Point | P-value |
|---|---|---|
| $x_1$ vs. $x_2$ | 0.5 h | $0 < \alpha$ |
| $x_1$ vs. $x_2$ | 120 h | $0 < \alpha$ |

**$x_1$: drug ID = 1, dose ID = 4. $x_2$: drug ID = 2, dose ID = 4**

| Conditions | Time Point | P-value |
|---|---|---|
| $x_1$ vs. $x_2$ | 0.5 h | $0 < \alpha$ |
| $x_1$ vs. $x_2$ | 120 h | $0 < \alpha$ |

# Q2 Results

- Among all observations, the p-value is computed to be either 0 or extremely close to 0. This means we **reject the null hypothesis** and conclude that there is a mean difference in protein levels between the experimental conditions tested.
- Since there is a mean difference in MiTFg levels between the experimental conditions tested, this means that cellular homeostasis is impacted by this change as the balance of protein cells changes. Because of this, it is possible that non-deleterious cellular states could transition into deleterious states or vice versa. Thus, based on the results, the experimental conditions could be adjusted to maintain homeostasis.

# Q3 Data Summary

**Purpose of Data**: To find the relationship between the 4 different proteins at time t in experimental condition x.

- Observed proteins at time point 5, drug ID 2, and dose ID 3

```
file %>%
  filter(timepoint_id == 5 & drug_id == 2 & dose_id == 3) -> new_file
```

- A portion of the data is extracted according to the above experimental conditions and stored in a variable called 'new_file'. In context, the experimental conditions are values that are filtered in the R code, and the 4 proteins are MiTFg, AXL, NGFR, and Sox10.

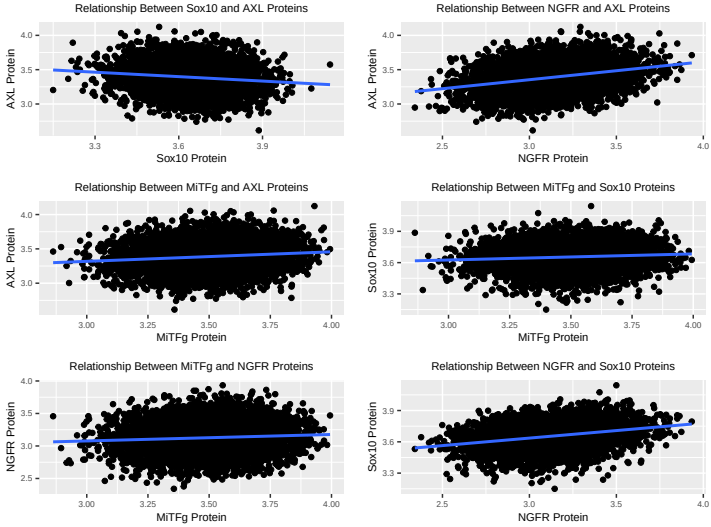
**Without bootstrapping**

- Each protein was paired with another protein. Then, at the experimental conditions chosen, the two proteins were regressed as a scatter plot to observe their relationship, correlation coefficient, the slope of the linear model, and its corresponding p-value.

| Protein Pair | Correlation Coefficient | Slope | P-value |
|---|---|---|---|
| Sox10/AXL | -0.127 | -0.08 | 1.08e-27 |
| NGFR/AXL | 0.302 | 0.35 | 1.09e-153 |
| MiTFg/AXL | 0.121 | 0.11 | 4.00e-25 |
| MiTFg/Sox10 | 0.086 | 0.13 | 2.42e-13 |
| MiTFg/NGFR | 0.075 | 0.06 | 1.39e-10 |
| Sox10/NGFR | 0.281 | 0.55 | 2.50e-132 |

# Q3 Statistical Methods

**6 graphs were produced to display the relationship between different proteins.**



**With Bootstrapping**

- Afterwards, each protein pair sample was bootstrapped, with sample sizes of 500 over 1000 iterations. With these bootstrap samples, a **95% confidence interval** was produced to calculate each correlation coefficient.

| Protein Pair | Confidence Interval (95%) |
|---|---|
| Sox10/AXL | (-0.1987, -0.0565) |
| NGFR/AXL | (0.2374, 0.3673) |
| MiTFg/AXL | (0.0498, 0.1869) |
| MiTFg/Sox10 | (0.0170, 0.1522) |
| MiTFg/NGFR | (0.0061, 0.1468) |
| Sox10/NGFR | (0.2374, 0.3673) |

# Q3 Results

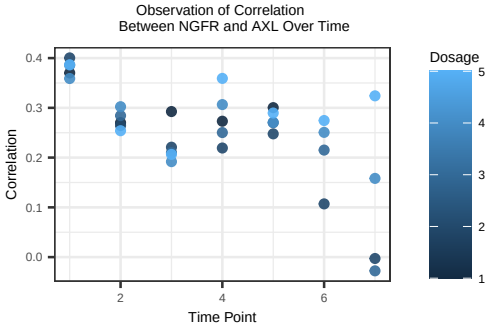
1. Based on the results of samples that were not bootstrapped, it appears all of the proteins have a weak correlation at the experimental conditions. Among the proteins, NGFR and AXL have the strongest correlation of 0.302, whereas MiTFg and NGFR have the weakest correlation of 0.075. The **null hypothesis** for these set of results would have been that there is *no relationship* between each protein pair. However, the p-values are extremely close to 0, which allows us **reject** the null hypothesis, concluding that there is a relationship between each protein pair.
2. The bootstrapped samples should be studied more carefully, as they have been sampled 1000 times to accurately approximate the population. All confidence intervals tend to be centered around 0, which supports the linear regression models in the previous slides, where the linear model appears to be widely scattered, without any strong correlation. Thus, we have 95% confidence the true correlation coefficient parameter is contained in the computed interval.

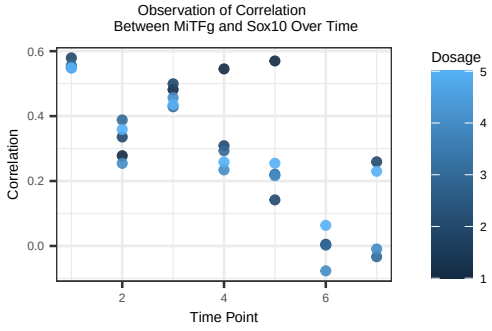
# Meta Question

**How do the observed correlations of NGFR and AXL (one pair), and MiTFg and Sox10 (one pair) evolve over time under the different experimental conditions?**

- In addition to time as the independent variable and the correlation coefficient as the dependent variable, the drug dosage condition was added as an *extra variable to consider different experimental conditions*.
- The graphs on the next slide shows correlation as a function of time, with dosage as an extra variable, whose level is indicated by colour. The dosage levels range from 1 to 5, which covers all levels in the original data.

# Meta Question



# Overall Results

**Q1/Q2** - Through these statistical analyses, we conclude that both an increase in time as well as dosage produce greater changes in the level of the outcome proteins, as well as more statistically significant ones.

**Q3/META** - We also conclude that as time passes, the correlation between MiTFg and Sox10, as well as NGFR and AXL, decreases. At lower dosages, the correlation seems to decrease at a faster rate, than compared to higher dosages.

**Real-World Application** - As we have identified factors that seem to create statistically significant change in the levels of MiTFg, whether that be through time or dosage, this means that we are capable of influencing cellular homeostasis as the balance between the proteins will have changed. Additionally, we know that if we want to influence Sox10, we should not influence MiTFg, since as time passes, the correlation reaches 0, indicating any changes in MiTFg will not be proportionally reflected in Sox10. Through these results, we may be able to transition non-deleterious cellular states into deleterious states or vice versa, perhaps allowing us to play a role in cellular homeostasis.

# Limitations and Shortcomings

1. As the data used were cleaned according to the purpose of our analysis, the results concluded from our observations may not hold true for the rest of the data that was unused. Therefore, our observations about the possibility of intervening in cellular homeostasis can only be generalized to the proteins covered by our data.
2. The overall data was sampled from a batch of cells, instead of a single cell. Therefore, the same cell was not measured over and over across time. Since the experimental conditions were applied to groups of cells, we must assume that these groups of cells were split to be small enough to represent each individual cell undergoing cellular homeostasis.

# Thank you