Statistical Analyisis on the AP-1 transcription factor network on Melanoma cells

Will Kukkamalla, Tianliang Bai, Zeyuan Zhuang, Caullin Seale

Dec 3, 2022

Introduction

In this presentation we will be analyzing the effect of AP-1 transcription factors on the cellular plasticity in melanoma cells. We will use multiple statistical methods to observe the relations between the transcription factors and the phenotype indicators.

Objectives

What is “good” homeostasis?

What is “bad” homeostasis?

how can “bad” cellular homeostasis be changed to be good?

Data Summary

We will be analyzing all 22 transcription factors and the effects on the phenotype indicator Sox10 and NGFR. This is because Sox10 is responsible for marking malignant melanoma cells and NGFR is a tumor suppressor.

Controls: Rep = 1, drug id = 1, dose id = 1

```
## Rows: 30,893
## Columns: 11
## $ Rep           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ dose_id       <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2
## $ drug_id       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1
## $ Phospho_c_Fos <dbl> 2.247442, 2.732822, 2.515971, 2.485299, 2.323357,
## $ NGFR          <dbl> 2.951467, 3.143596, 3.197096, 2.978923, 2.983139,
## $ Sox10         <dbl> 3.618069, 3.941809, 3.766320, 3.678794, 3.742037,
## $ Phospho_p38   <dbl> 3.016619, 3.128048, 3.135749, 2.920509, 2.869370,
## $ ATF6          <dbl> 2.667389, 2.913234, 2.888233, 2.715590, 2.847456,
## $ JunB          <dbl> 3.569666, 3.924559, 3.406373, 3.626526, 3.462569,
```

Homeostasis

We have decided to define good homeostasis as when the cellular phenotype is undifferentiated or melanocytic.

We define bad homeostasis as when the cellular phenotype is transitory or Neural-crest-like.

Thus to turn “bad” cellular homeostasis to “good” homeostasis the phenotype must go from transitory or Neural-crest-like to undifferentiated or melanocytic to do this we must find out what transcription factors best lend to increacing the precense of the NFGR protein.

Statistical Methods: Two Sample Hypothesis Test

We decided to do a 2 sample hypothesis test on 22 transcription factors. Our variables: $\alpha = 0.01$, $\mu_1 =$ the mean protein level in x transcription factor at time 0.5h, $\mu_2 =$ the mean protein level in x transcription factor at time 120h

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

This will be a two tailed test to determine whether or not the transcription factors protein level changes over time.

Statistical Methods: Regression

We decided to create multiple linear regression models to predict the phenotypical outcomes from the transcription factors.

The reason we did Regression was to answer the question:Can we predict cellular phenotypical outcomes values/states from transcription factors?

We took the 9 smallest differences from the observed and simulated values from the transcription factors over time and made multiple models to decide if at least one had a RMSE value of under 0.3.

Statisitical Methods: Classification Trees

We decided to create classification trees to answer the question:At time in experimental condition , what TF are most predictive of cellular values/states

In order the determine what the best transcription factor is for determining the cellular phenotype we have decided to create classification trees.

We took all the transcription factors and made a classification tree in order at different times to see which transcription factor is the best predictor for each time.

To determine the best transcription factor we took the one being used as the first split.

Statistical Methods: Correlation

After using the Classification trees to determine the most predictive Transcription factor for NGFR and Sox10 we then created models to analyze the correlation between the most predictive transcription factor and the NGFR and Sox10.

The purpose of which is to answer the question at different timepoints how does the transcription factor correlate with the protein level of the NGFR and Sox10.

Since NGFR aids in tumor suppression and sox10 marks malignant melanoma cells answering this will give us insight on how we can turn “bad” homeostasis into “good” homeostasis.

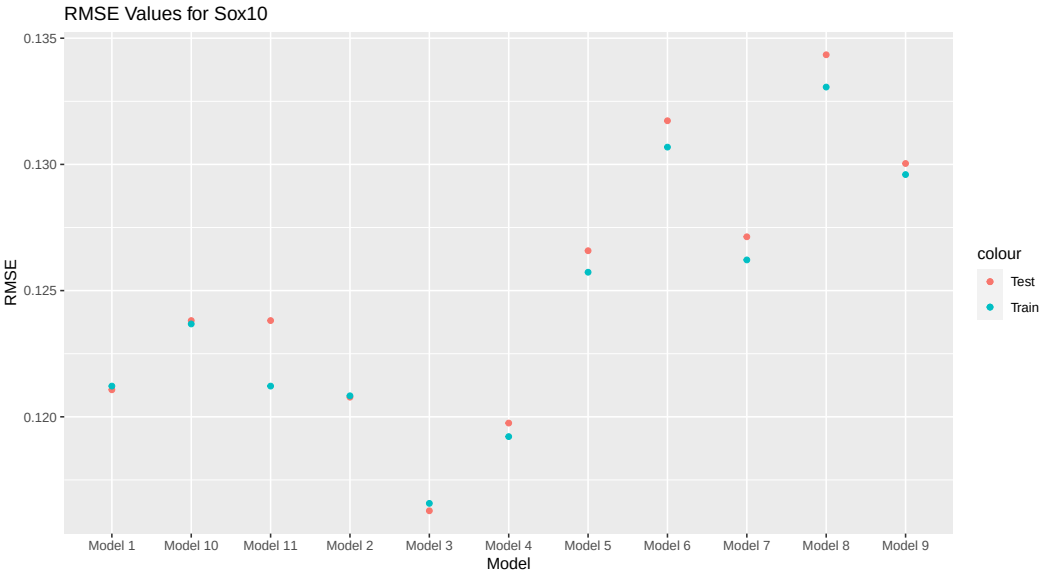
Results of Two Sample Hypothesis tests

P-value for each Transcription Factor

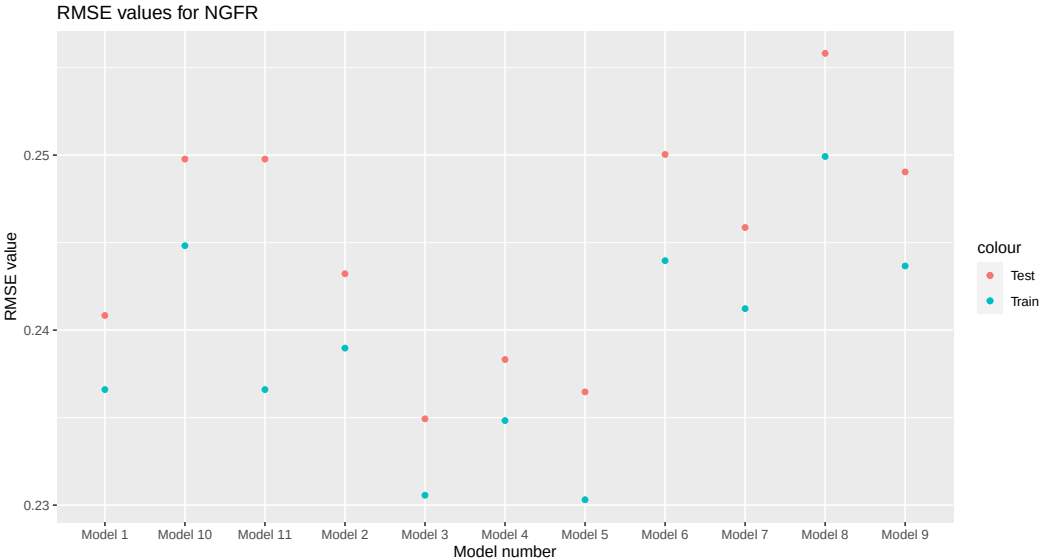
```
## # A tibble: 10 x 2
##    Name          P_value
##    <chr>           <dbl>
##  1 ATF2                0
##  2 ATF3                0
##  3 ATF4                0
##  4 ATF5                0
##  5 Phospho_ATF1        0
##  6 Phospho_ATF2        0
##  7 Phospho_ATF4        0
##  8 ATF6                0
##  9 JunB                0
## 10 c_Jun               0
```

```
## # A tibble: 12 x 2
##    Name          P_value
##    <chr>           <dbl>
##  1 NF_kappaB           0
##  2 Fra1                0
##  3 Fra2                0
##  4 c_Fos               0
##  5 Ki_67               0
##  6 Phospho_Fra1        0
##  7 Phospho_c_Fos       0
##  8 Phospho_p38         0
##  9 JunD                0
## 10 Phospho_S6          0
## 11 Phospho_c_Jun       0
## 12 Phospho_Erk1        0
```

Results of Linear regression models: Sox10



Results fo Linear regression models: NGFR



Classification Tree Results Sox10

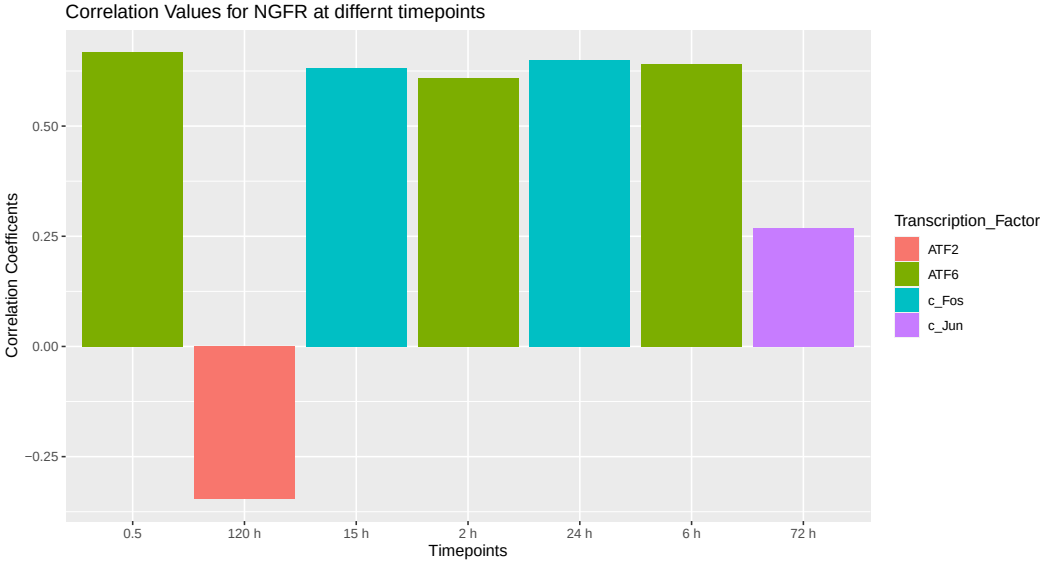
```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
```

```
## # A tibble: 7 x 2
##   timepoints Most_important_transcription_factor
##   <chr>      <chr>
## 1 0.5 h      ATF4
## 2 2 h        ATF4
## 3 6 h        ATF4
## 4 15 h       ATF4
## 5 24 h       c_Fos
## 6 72 h       c_Jun
## 7 120 h      Fra1
```

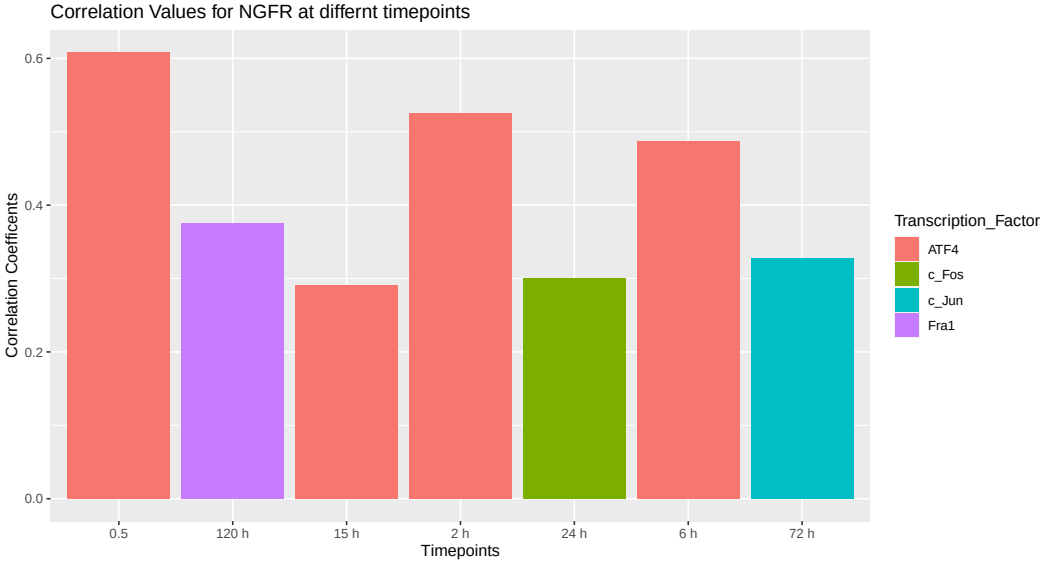
Classification Tree Results NGRF

```
## # A tibble: 7 x 2
##   timepoints Most_important_transcription_factor
##   <chr>      <chr>
## 1 0.5 h      ATF6
## 2 2 h        ATF6
## 3 6 h        ATF6
## 4 15 h       c_Fos
## 5 24 h       c_Fos
## 6 72 h       c_Jun
## 7 120 h      ATF2
```

Results from Correlation NGFR



Results from Correlation Sox10



Conclusion of the Two sample Hypothesis tests

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

Since the p-values are all 0 we must reject the null hypothesis. Thus we can conclude that there is a difference between the protein values over time.

Limitations: There is a possibility that we could have encountered a type one error by falsely rejecting the null hypothesis thus meaning the the transcription factors do not change over time.

Conclusion of the Regression anaylisis

Since all the models have an RMSE values under 0.3 we can conclude that we can predict the phenotype value for Sox10 and NGFR from the transcription factors.

Bigger picture: This means that it makes it easier to determine whether the there is a high presence of the Sox10 and NGFR phenotype indicator from the transcription factors. From this we can conclude whether a cell has is likely to be a malignant melanoma cell in either the transitory or nueral-crest-like phenotypes or the good homeostasis phenotypes.

Conclusion of the Classification tree

We can see that what at the different times what the most predictive transcription factor are.

This is important because the transcription factors are responsible for turning on the genes to then create these phenotype proteins thus from each timepoint we can estimate the best transcription factor that will either turn on the genes for Sox10 or NFGR thus making it easy to mark and or suppress tumors.

Thus we have begun to answer the final question on how to change the homeostasis from good to bad.

Conclusion of the Correlation

From the table of r values and r squared values we can see that there is a positive correlation between all of the transcription factors except ATF2

What this means is that as we increase the presence of these transcription factors in these certain time points it is likely that we increase the presence of the NGFR and Sox10 protein,and in the case for Timepoint 120h if we decreace ATF2 we see an positive correlation with NGFR

The reason that we believe that there is a possibility that it is causation and not correlation is that Transcription factors regulate gene expression

This means that the increase in a particular Transcription factor might increase mRNA levels thus resulting in more of a specific protein

Thus increasing the Transcription factors that increace NGFR will result in tumor suppression thus turning our “bad” homeostasis into “good” homeostasis

Limitations: It is possible that there may be some unseen confounder that results in the correlation