Using Sequence Determinants to Predict CRISPRa Ricin Susceptibility

Jenny Yang

Summer 2018

sgRNA

- used to direct Cas9 in binding DNA at specific target sequence (protospacer adjacent motif - PAM)
    - 5’ - NGG - 3’
- depending on design specifications, Cas9 can be “programmed” to cleave host’s genome at virtually any position
- key step in implementing CRISPR genetic screens is selecting sgRNAs that mediate high Cas9 activity

Previous Work

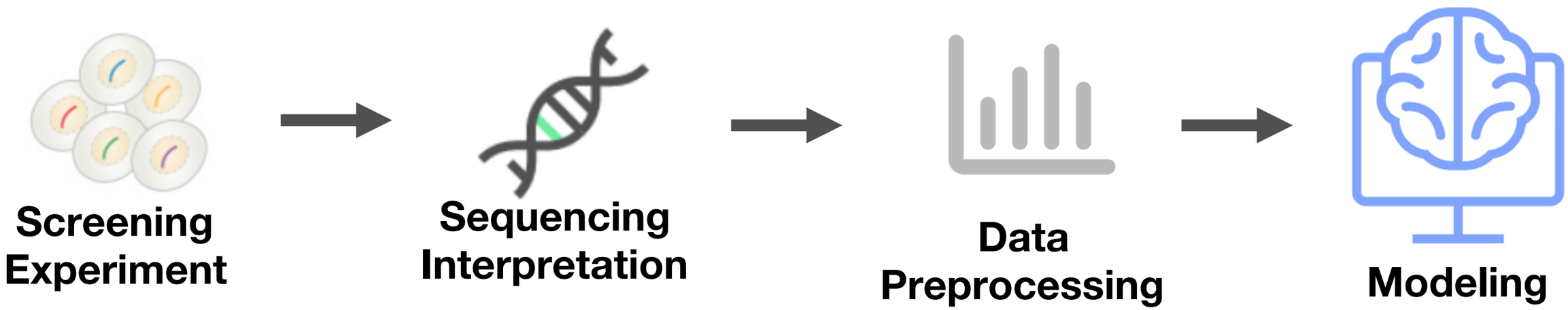
- nucleosome occupancy, sequence features, etc. can influence Cas9 activity

- sensitivity to toxins can provide insights into complex pathway mapping

- having a quantitative model incorporating these features can help predict highly active sgRNAs for CRISPRi and CRISPRa

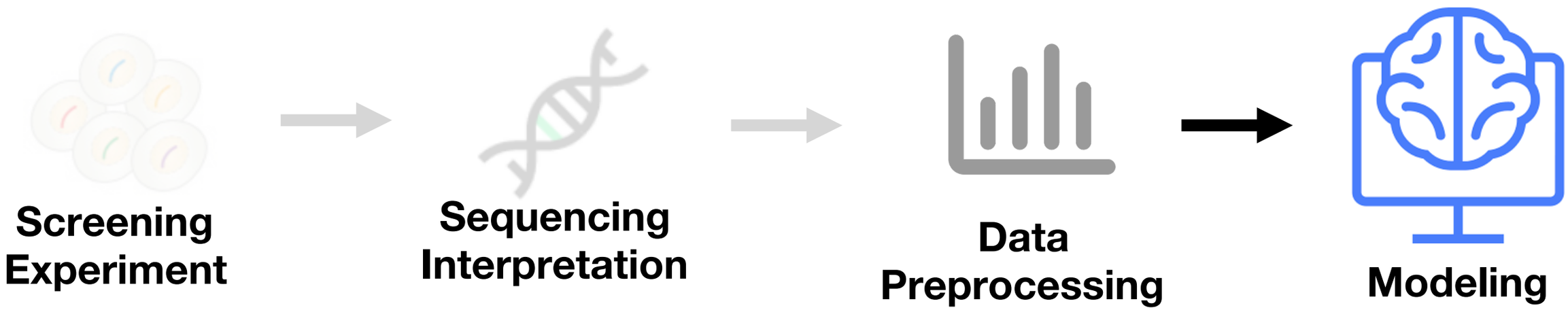
Our Goal

Create a predictive model using sequence determinants to predict CRISPRa ricin susceptibility



Our Goal

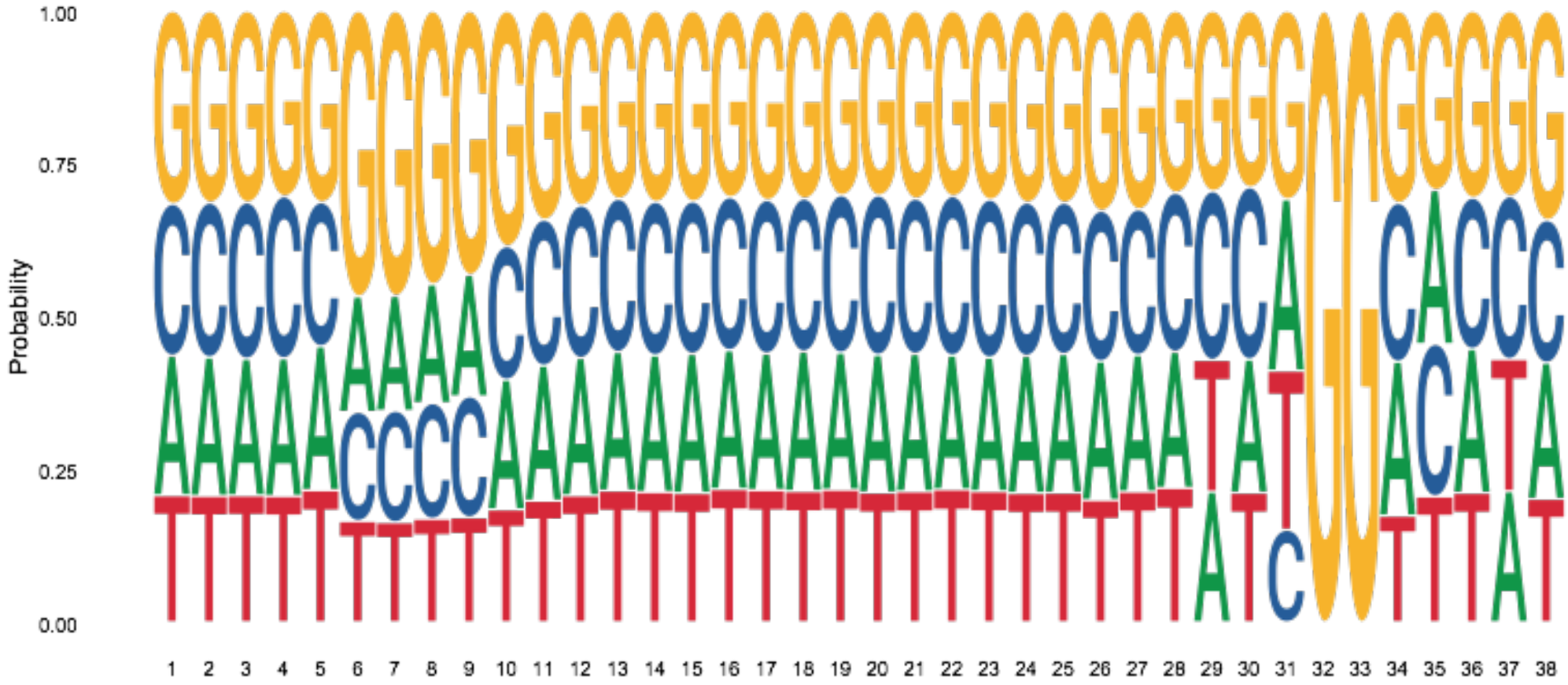
Create a predictive model using sequence determinants to predict CRISPRa ricin susceptibility



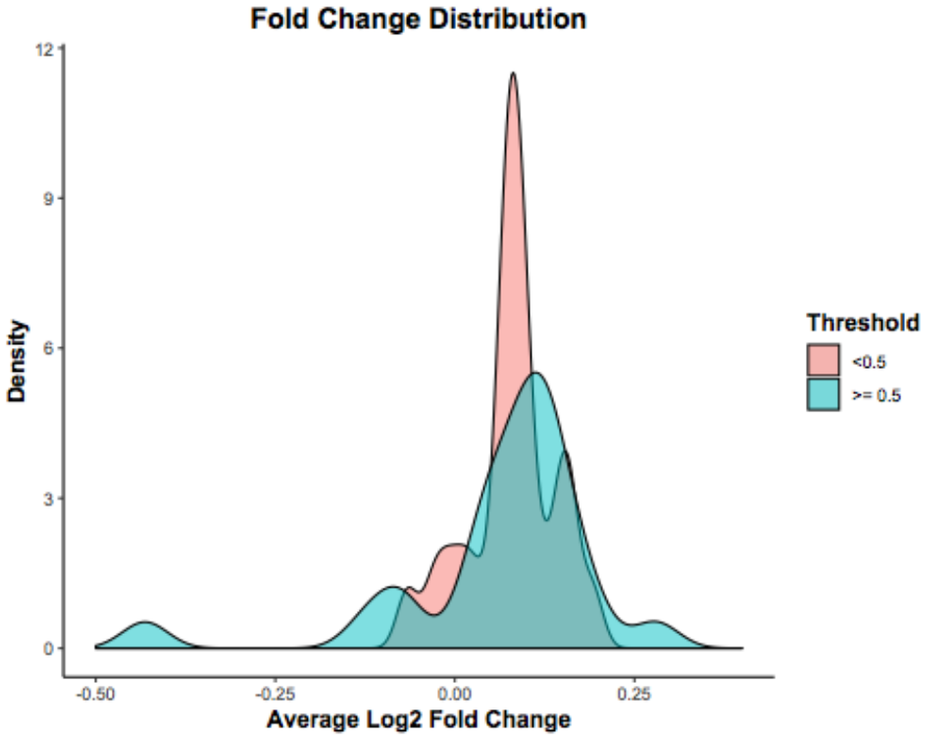
Calculating Features

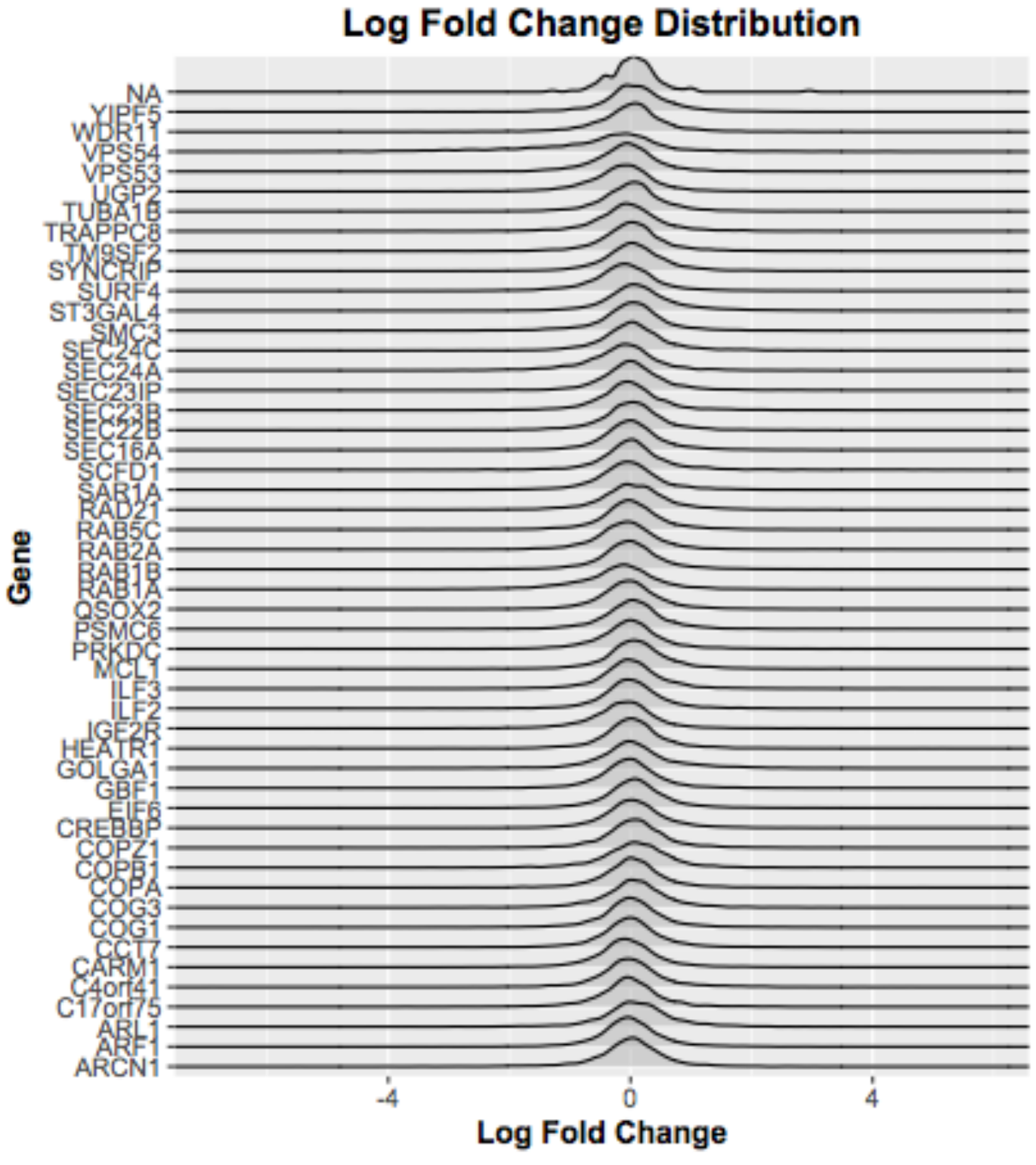
- used CRISPRa ricin tiling data from Gilbert et al. (2014)
- 48 genes chosen, known to modulate ricin sensitivity (Bassik et al., 2013; Gilbert et al., 2014)

- for sequence:
    - width of 38 bases
    - PAM located at positions 31, 32, 33



- want to predict activity:
  - activity calculated by log fold change of standard conditions vs. ricin treated expression levels
  - normalized about 0





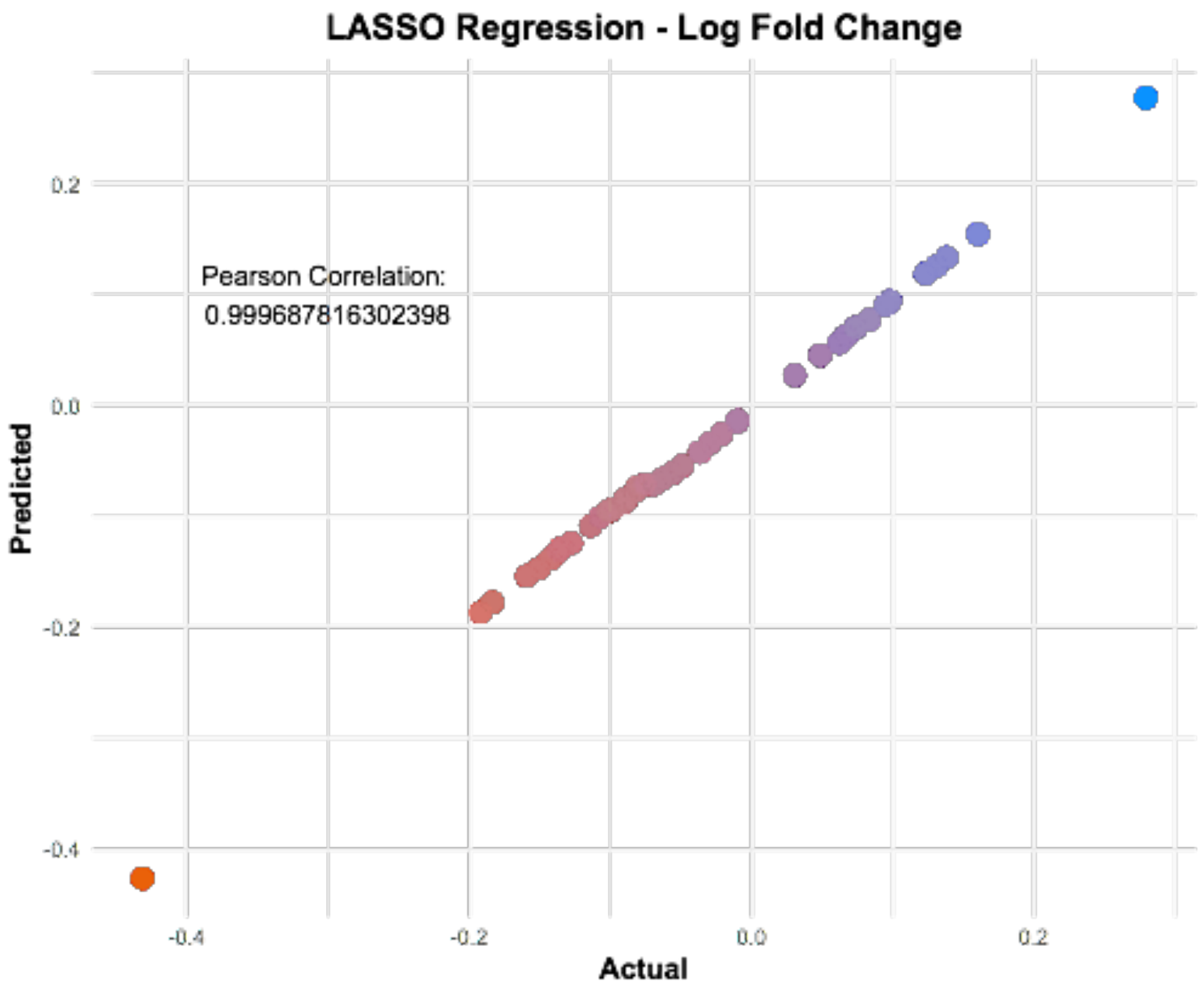
Models

- **LASSO**
    - shrinkage and variable selection method for linear regression models

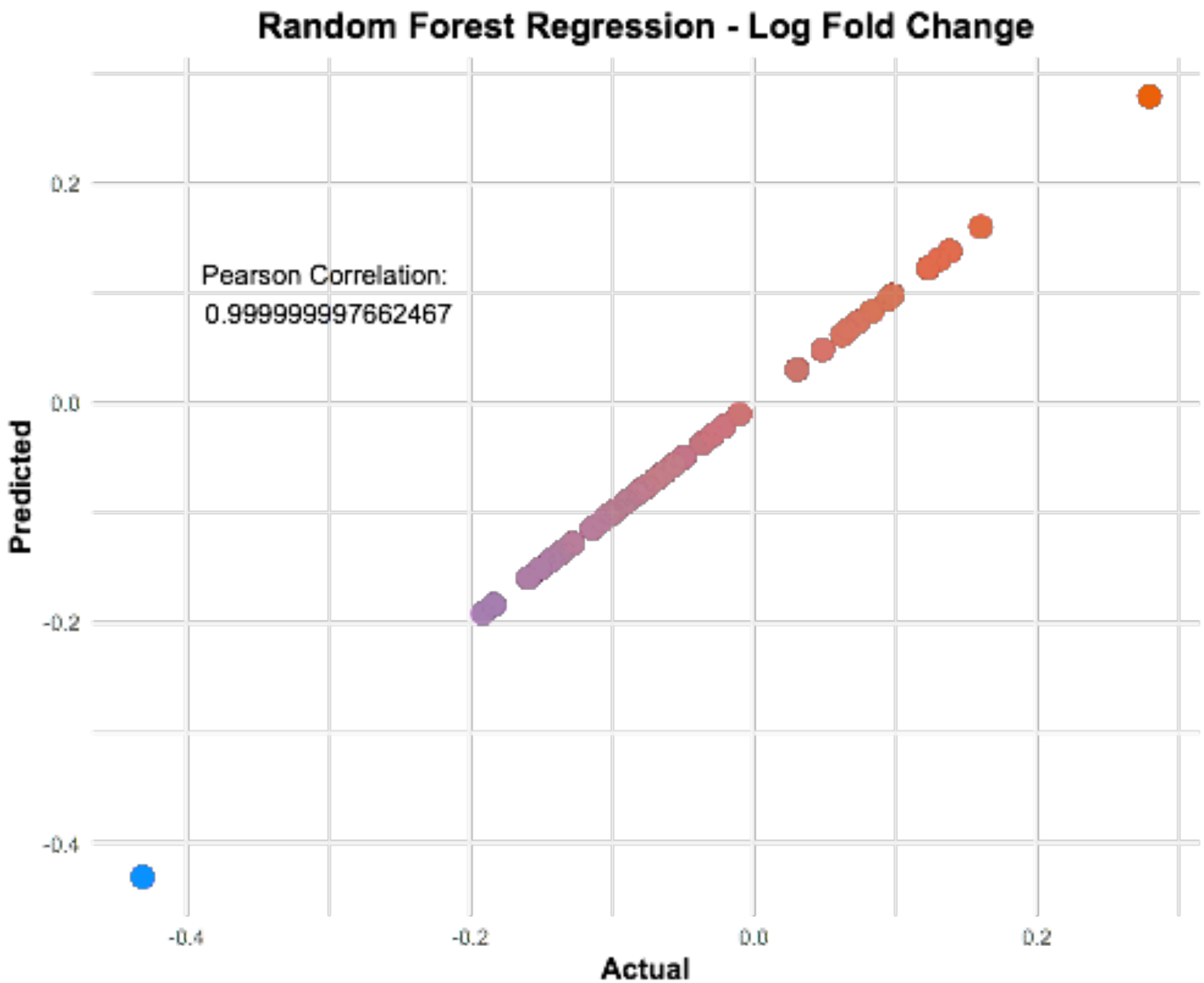
- **Random Forest Regression**
    - builds decision trees and merges them together for prediction

- **Support Vector Machine**
    - looks for hyperplane in N-dimensional space that distinctly classifies data points

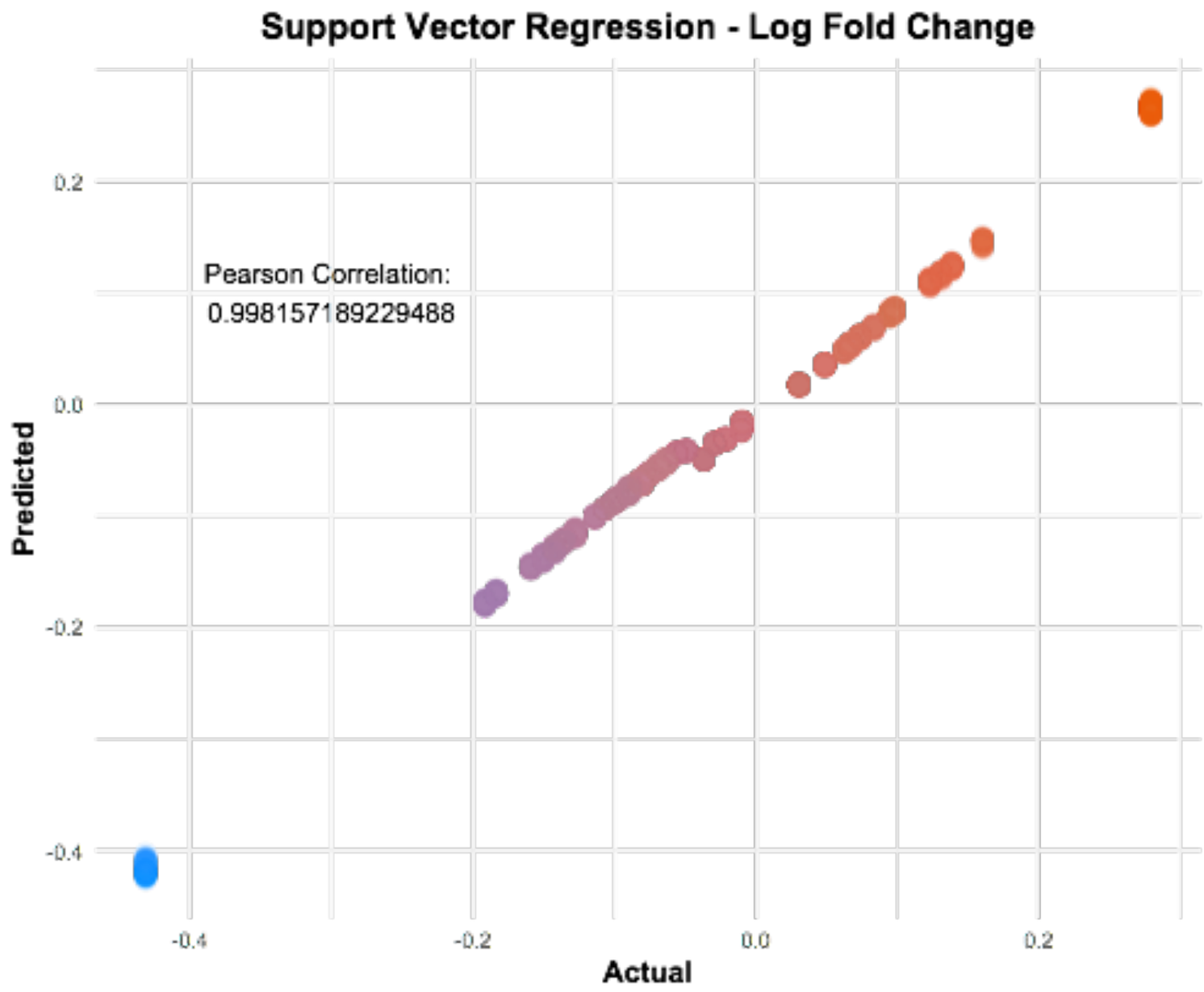
Lasso Regression



Random Forest Regression



Support Vector Machine



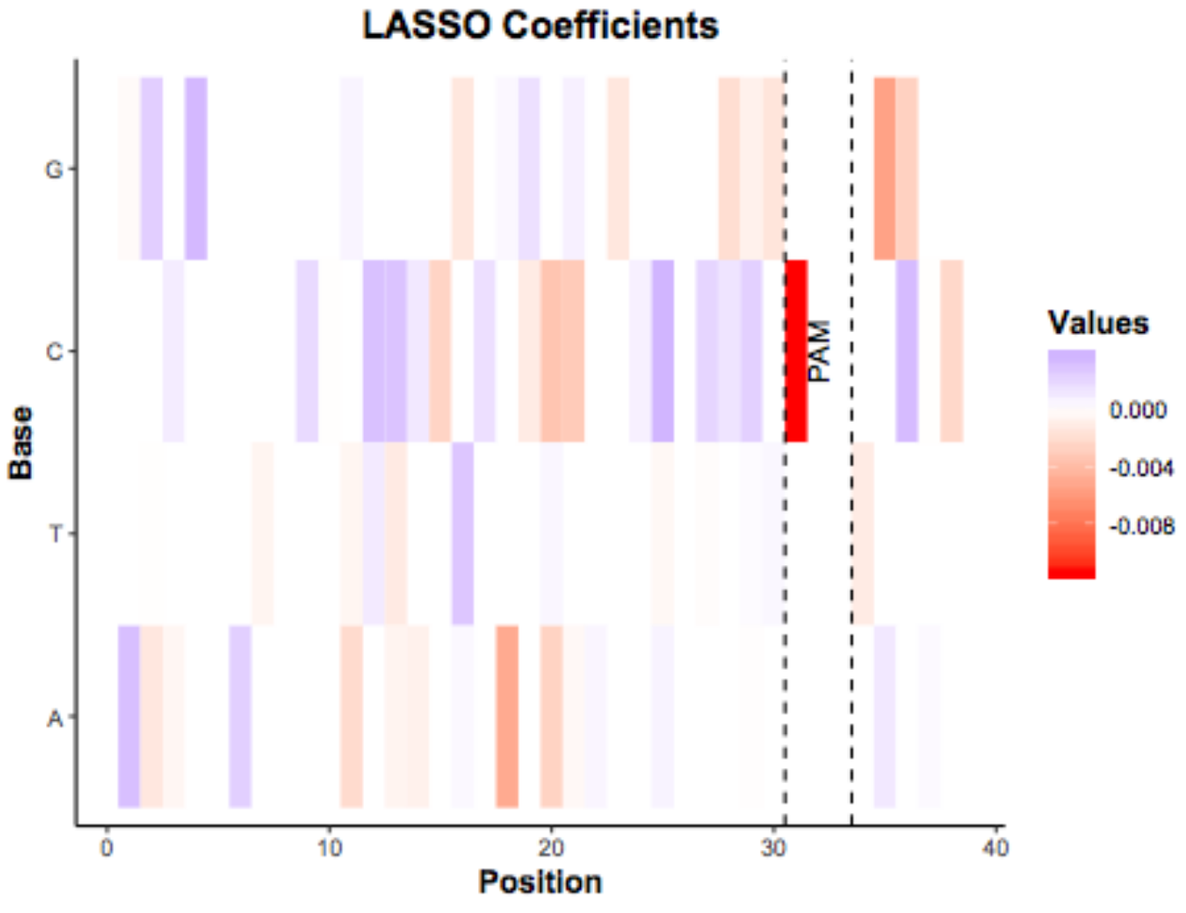
Model Comparison

	LASSO Regression	Random Forest Regression	Support Vector Regression
Mean Squared Error	1.8813×10^{-5}	8.2483×10^{-11}	1.6140×10^{-4}
Pearson Correlation Coefficient	0.99969	0.99999	0.99815

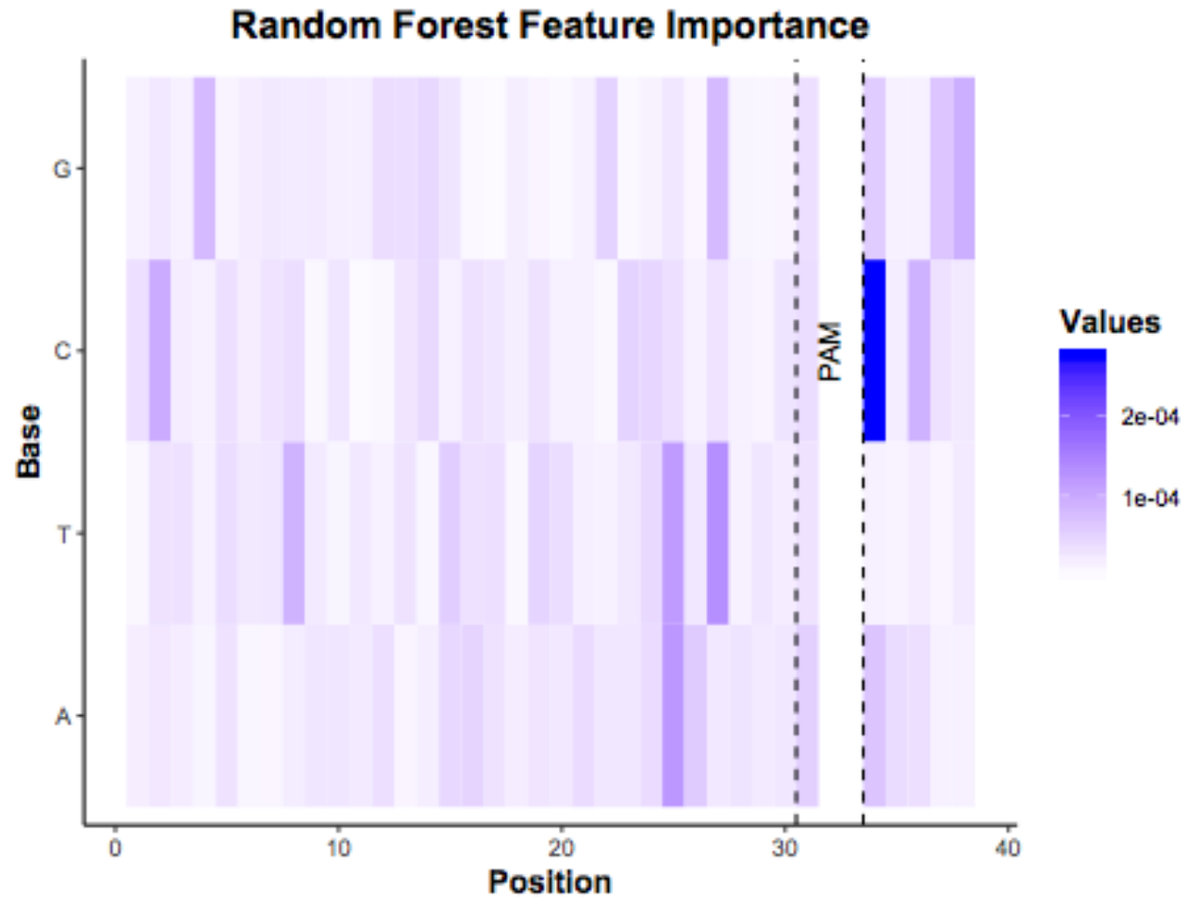
Model Comparison

	LASSO Regression	Random Forest Regression	Support Vector Regression
Mean Squared Error	1.8813×10^{-5}	8.2483×10^{-11}	1.6140×10^{-4}
Pearson Correlation Coefficient	0.99969	0.99999	0.99815

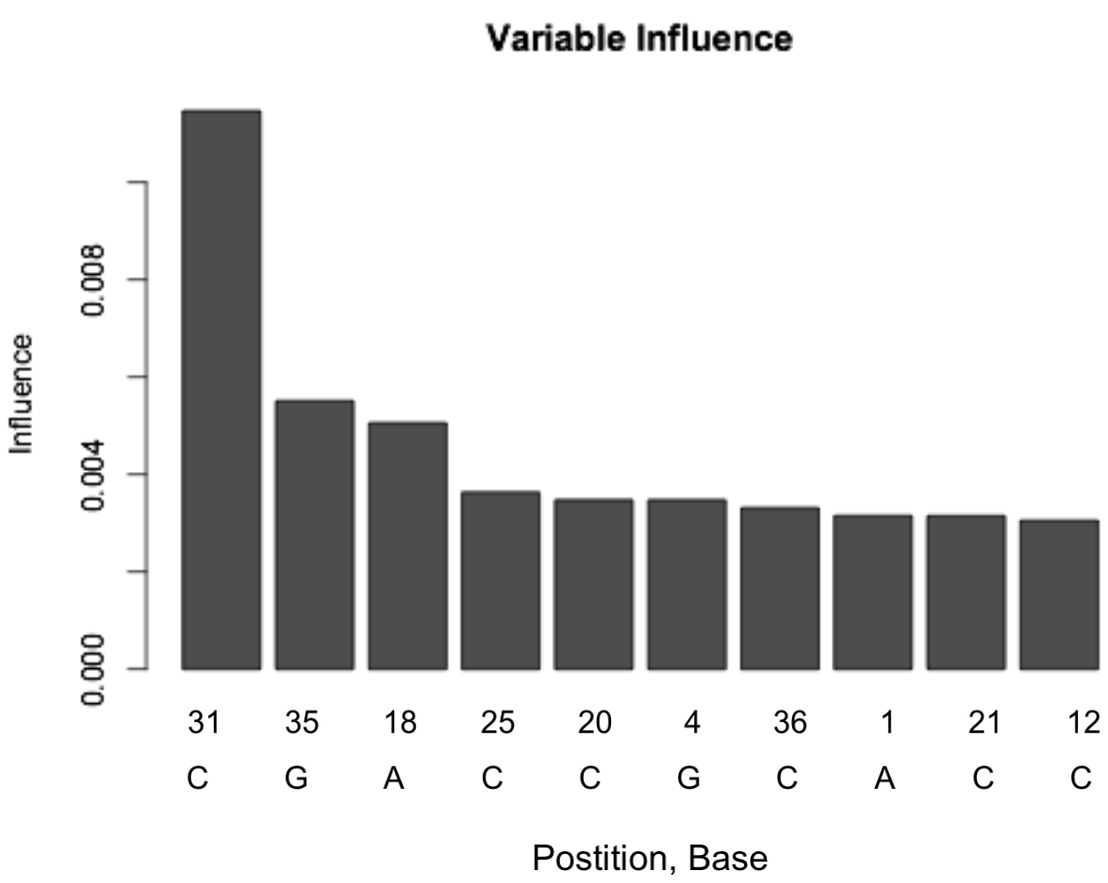
Lasso Regression



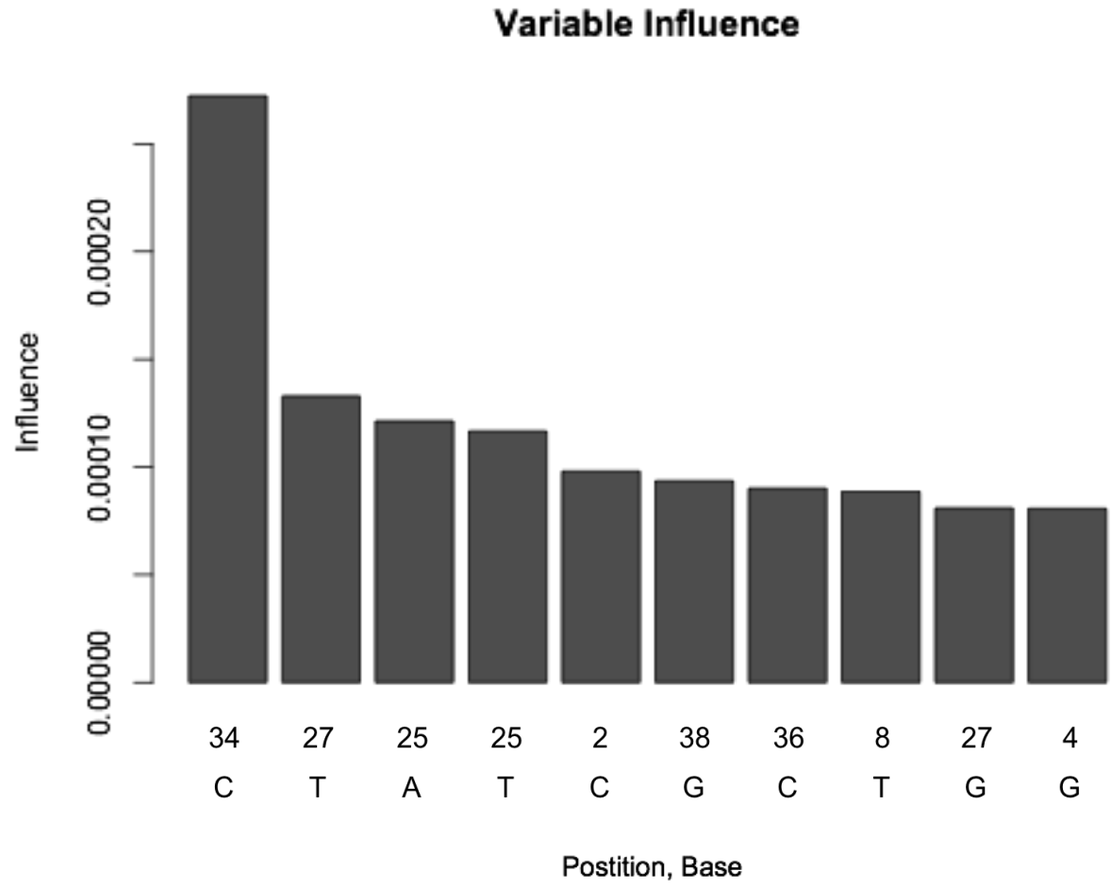
Random Forest Regression



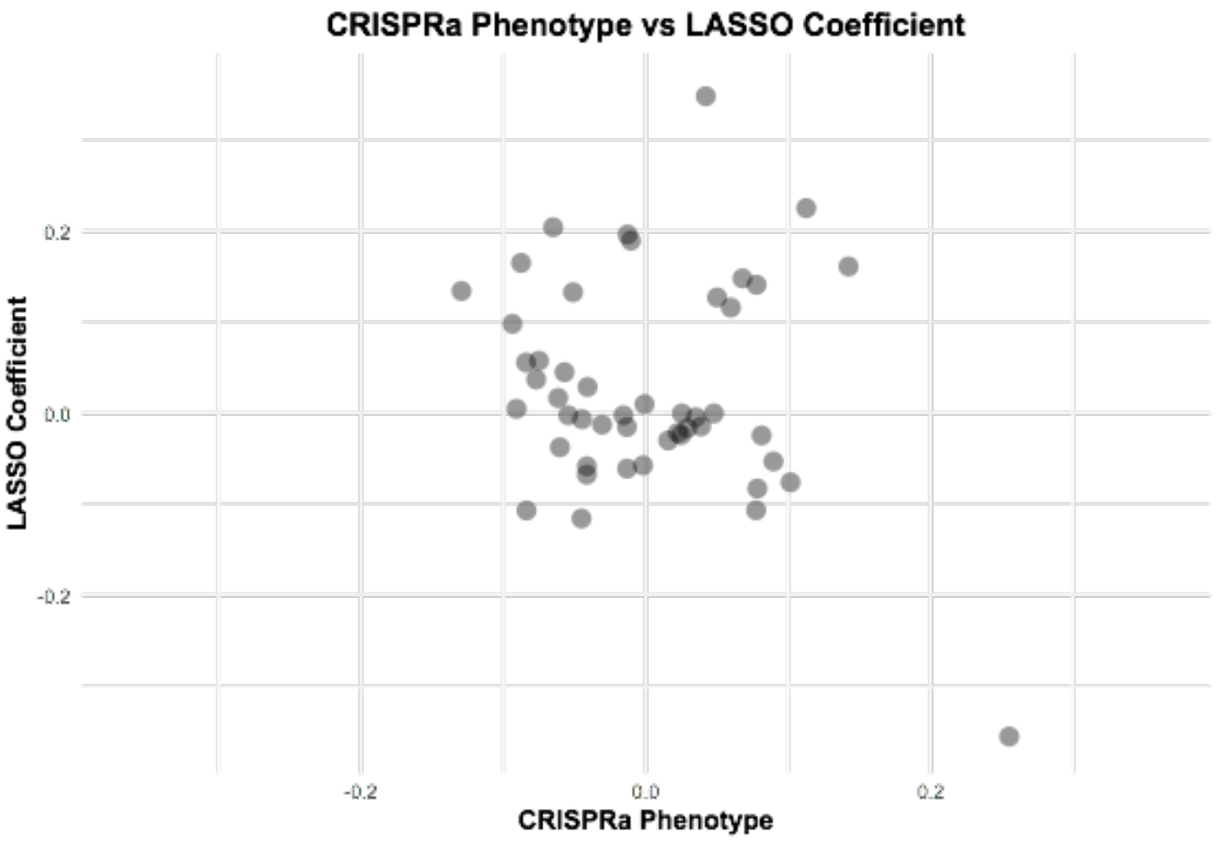
Lasso Regression



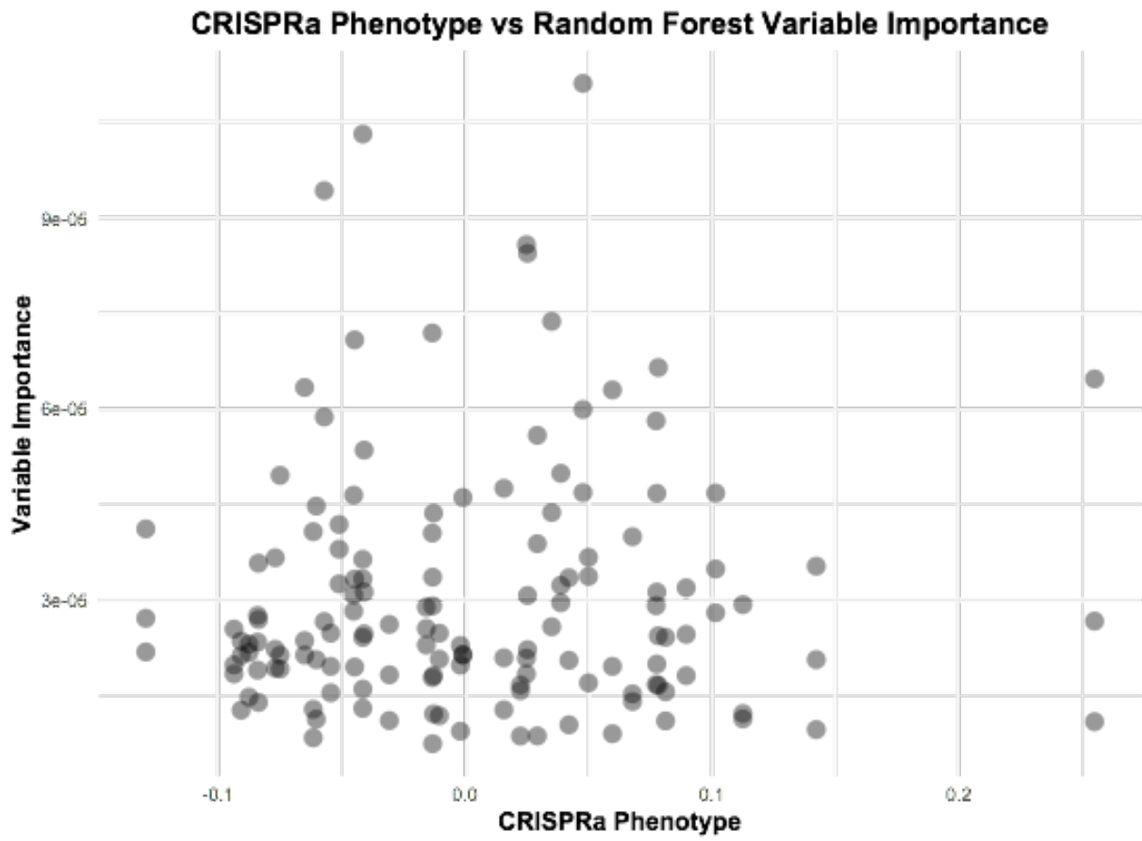
Random Forest Regression



Lasso Regression



Random Forest Regression



Considerations

- not enough structured data points
- models encode correlation, not causation or ontological relationships
- each narrow application needs to be specially trained

Next Steps

- further investigate layers of model
    - clusters of features that form the nodes
    - nodes with the strongest output signal
        - branching paths of random forest
- add more types of data (epigenetic, environmental, etc.)
- try other models (autoencoder, etc.)

Acknowledgements

- Timothy Daley
- Stanley Qi

Thank you Qi lab!

- Virginia Diaz