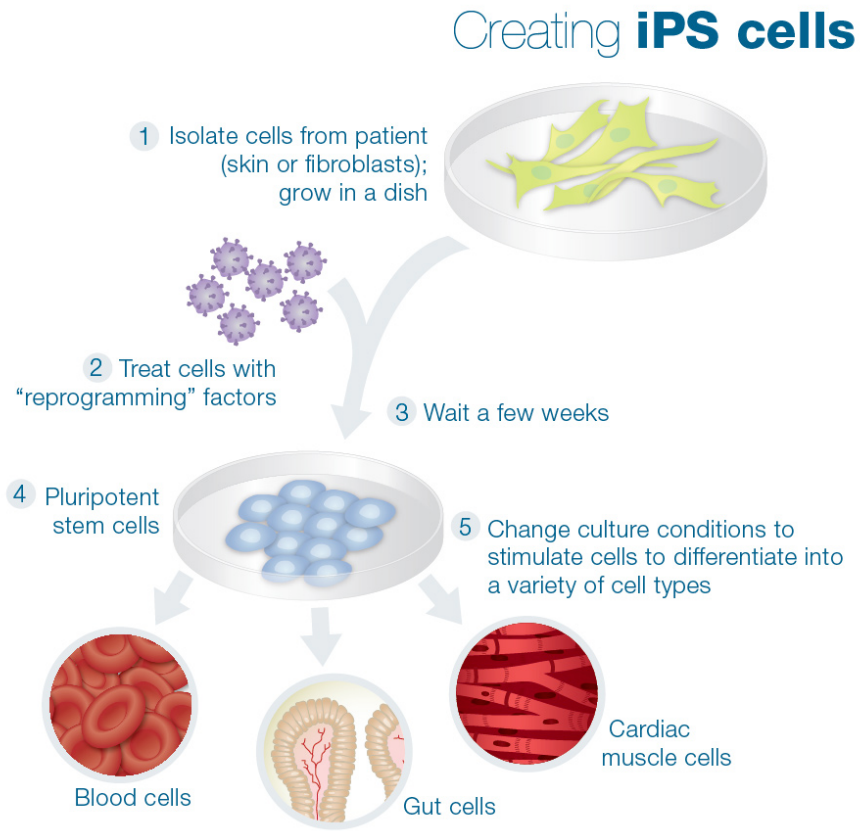
Gaussian Processes to Predict Gene Expression for Induced Pluripotent Stem Cells Differentiation

NIRMAL KRISHNAN

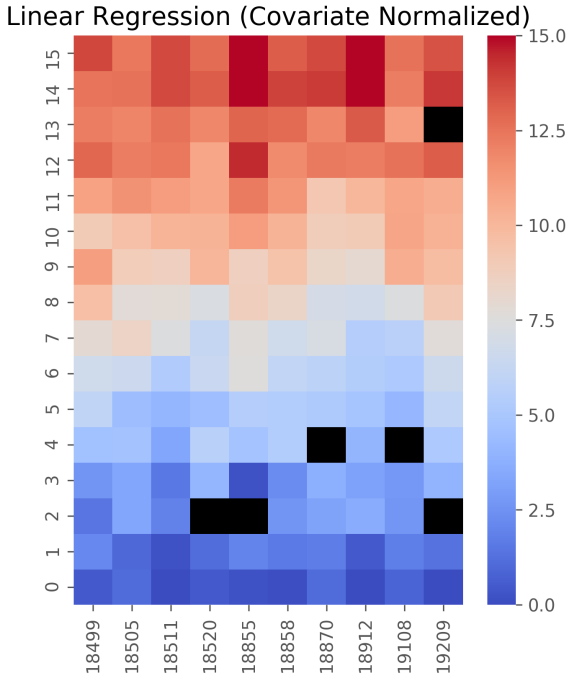
Introduction- Data

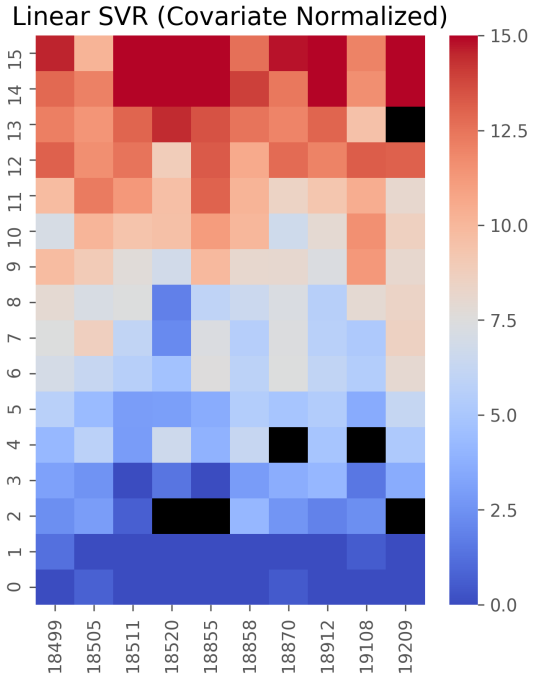
- 10 Induced Pluripotent Stem Cells (iPSCs) differentiated into cardiomyocyte cells
- 16 day process between iPSC to fully differentiated, beating cardiomyocyte
- Gene expression from RNA sequencing recorded at each time step
- 154 total samples
- 15,794 genes



Introduction- Prior Work / Goals for this Project

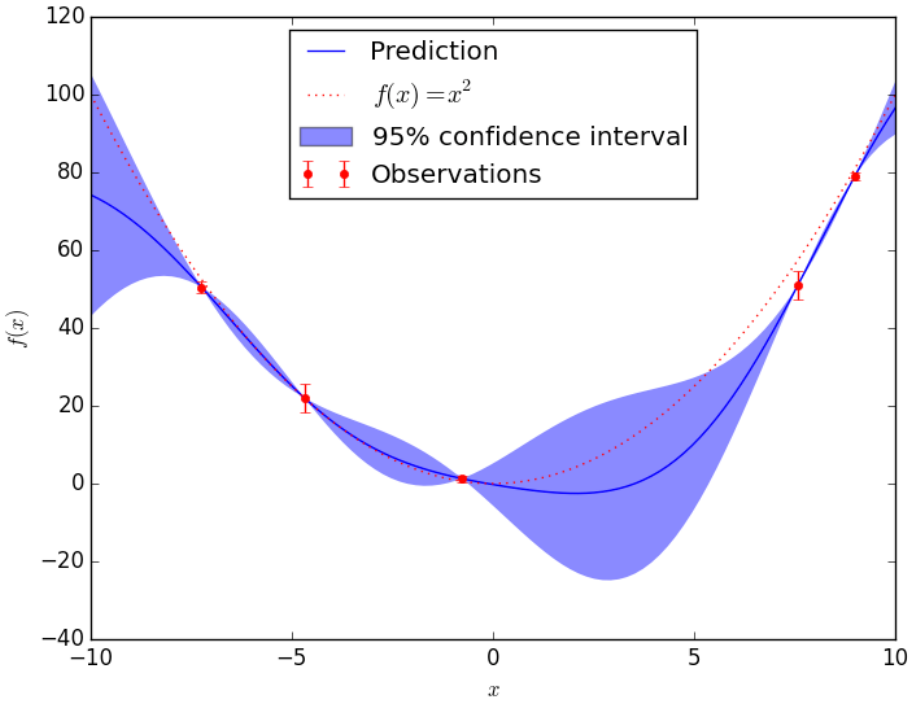
- Battle/Gilad lab interested in understanding stages of differentiation using gene expression results
- Prior work: predicting time using gene expression (“pseudotime” prediction)
- Useful for understanding the differentiation process, predicting how quickly a cell will become a cardiomyocyte
- Instead of modeling time, this project focuses on modeling gene expression
- Goal: model gene expression of iPSC cells using Gaussian Processes and compare predictive accuracy to traditional models





Methods- Gaussian Processes

- Gaussian Processes (GP) have quickly gained acclaim in scientific community for flexibility and predictive accuracy
- Every point in some continuous input space is associated with a normally distributed random variable
- Parameterized by mean and kernel function
- Mean function usually set to 0 if there is no informative prior. Mean function heavily outweigh effects of kernel function and GPs can be fully fitted using just the kernel function.



Methods- Choosing Kernel Function

- Radial-basis function (RBF), also known as squared exponential kernel
  - Infinitely differentiable
  - Very smooth

$$k(x_i, x_j) = \exp\left(-\frac{1}{2} d(x_i/l, x_j/l)^2\right)$$

- Matern kernel
  - Generalization of RBF
  - Extra parameter ν - can control smoothness
  - Allows for adaptability of true underlying functional relation

$$k(x_i, x_j) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\gamma\sqrt{2\nu} d(x_i/l, x_j/l)\right)^{\nu} K_\nu \left(\gamma\sqrt{2\nu} d(x_i/l, x_j/l)\right),$$

- Exp-Sine-Squared kernel
  - Allows for modeling periodic functions

$$k(x_i, x_j) = \exp\left(-2\left(\sin(\pi/p * d(x_i, x_j))/l\right)^2\right)$$

- Standard non-informative priors set for all kernel functions

Method- Data Preprocessing

- Data is quantile normalized
- SVA run on the gene expression matrix—21 latent factors unrelated to time
- Using 21 factors, trained linear model to predict expression for each gene

$$Y = X\,B$$

  - X -> 21 Latent Factors for each sample
  - Y -> Expression of a single gene for all samples
  - B -> Learned Coefficients

- Subtracted predicted expression from the original expression matrix—effectively “regressing out” covariates

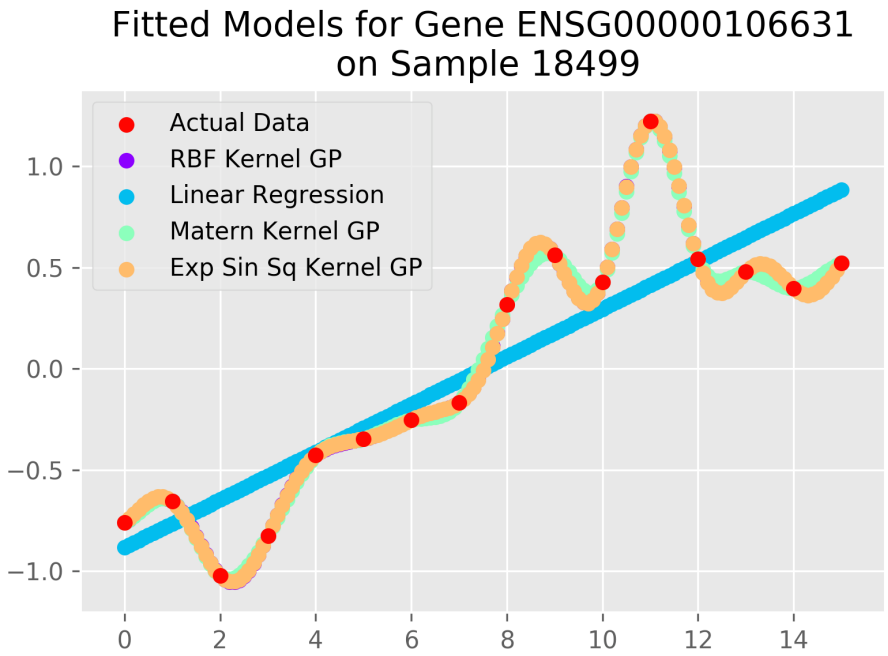
Methods- Training Models to Predict Expression

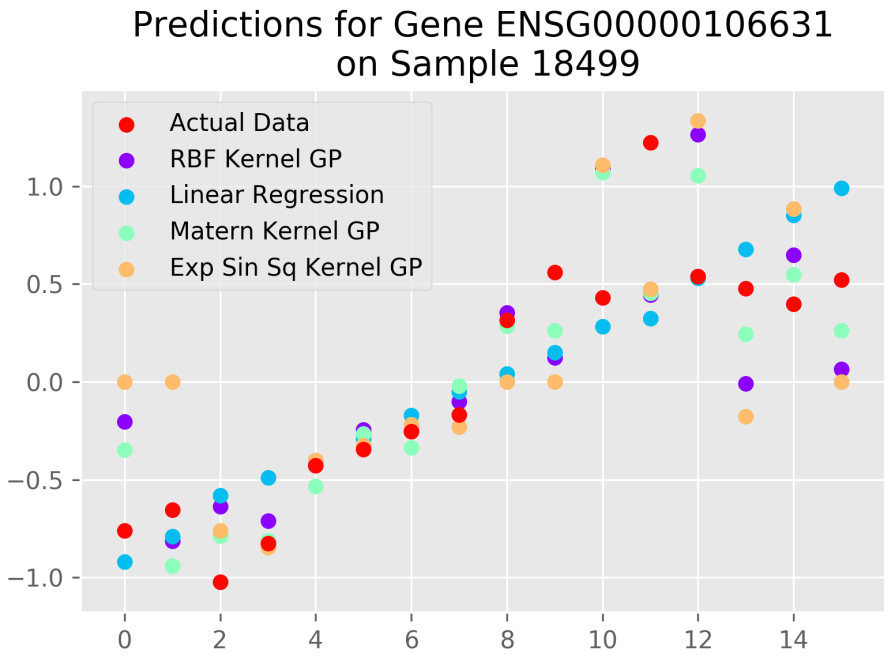
- Selected top 100 genes with the most variance across samples from original expression matrix (due to extensive training time)
- Baseline model: Linear Regression
- Gaussian process model kernels: RBF, Matern, Exp-Sine-Squared

For each gene:

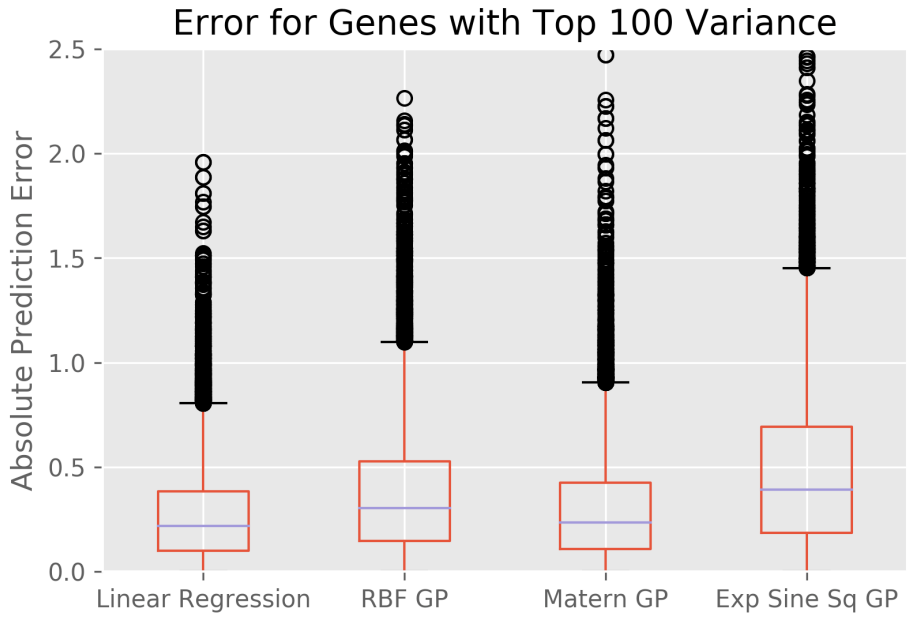
- For each cell line:
  - For each time step:
    - Fit model using all data (on this gene and cell line) except for current time step
      - Train model hyper-parameters using:
        - X-> time [0: 15], except current time step
        - Y-> gene expression
    - Predict on current time step and record absolute prediction error

Results- Fully Fitted Model vs Predictions





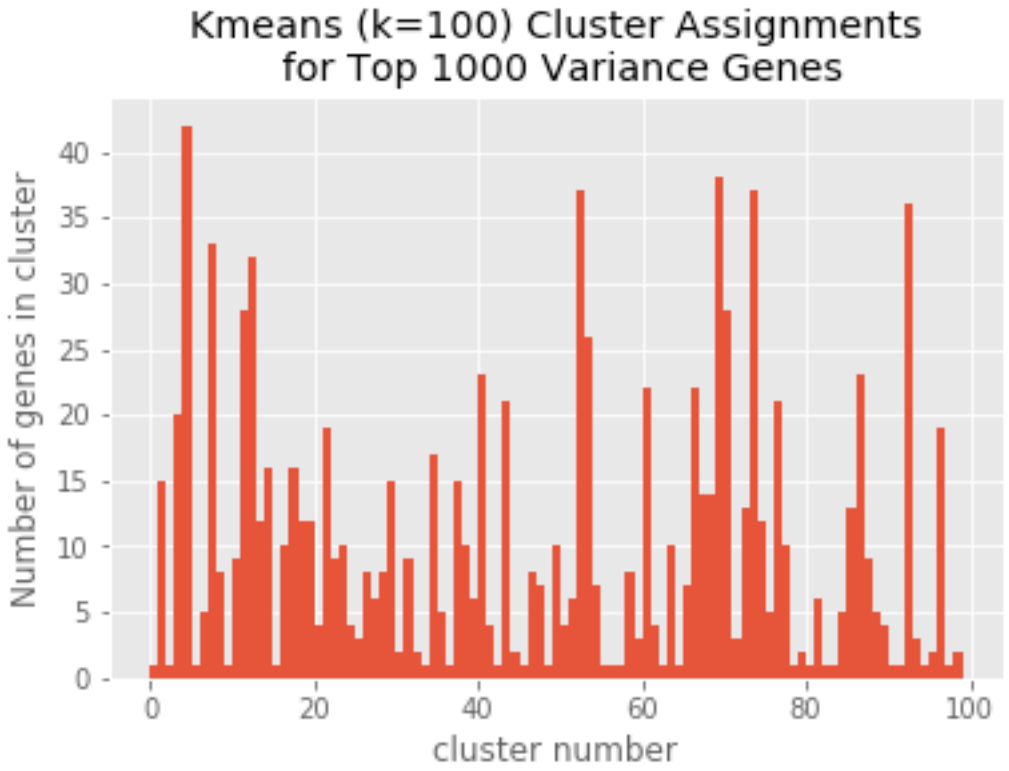
Results- Prediction Error Across Models



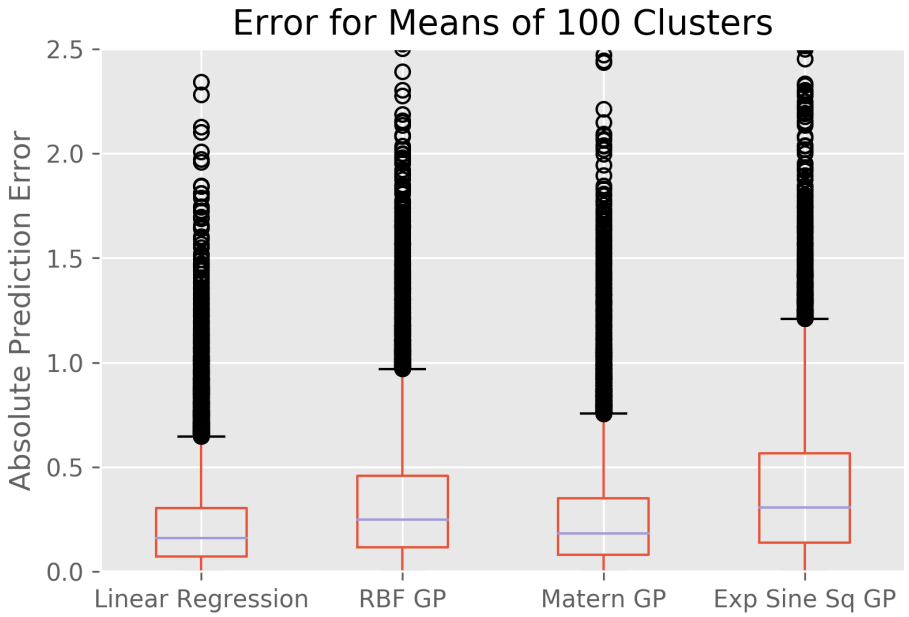
Model 1	Model 2	Rank Sums Test Statistic	p-value
Linear Regression	RBF GP	-35.31	**2.90E-73**
Linear Regression	Matern GP	-9.30	**1.29E-20**
Linear Regression	Exp-Sine-Sq GP	-61.62	**0**

Results- Filtering Out Noise

- Theory: individual gene expression measurements for genes tend to be noisy
- Potential solution:
  - Cluster genes using kmeans on expression across samples
  - Calculate mean vector of expression in each cluster, effectively filtering out noisy genes
  - Re-run analysis on mean vectors of each cluster
- Implementation:
  - Retrieved genes with top 1000 variance now
  - Cluster using k=100 and calculate mean vectors



Results- Prediction Error Across Models for Cluster Means



Model 1	Model 2	Rank Sums Test Statistic	p-value
Linear Regression	RBF GP	-35.17	**4.40E-271**
Linear Regression	Matern GP	-9.80	**6.00E-23**
Linear Regression	Exp-Sine-Sq GP	-49.55	**0**

Conclusions

- In iPSC differentiation, gene expression is best modeled by simple methods like linear regression
- Gaussian Processes tend to overfit in this domain—they are overly flexible, resulting in poor predictive accuracies
- Of the kernel functions, the Matern function best models this process
- Further research:
  - Using low-order polynomial-basis functions and comparing the results with models
  - Sparse Gaussian Process Approximation using Pseudo-Inputs
    - Selects M datapoints where M << N which best generalize the actual data
    - Combats overfitting and could potentially improve results
    - Also interesting in understanding differentiation process, selecting time points that are more important could lend itself to better understanding differentiation process