Hi-C Chromosome conformation capture prediction under LASSO

Alex Chen, Zack Vliet

Overview

1. Background
    a. Motivation
    b. Data
2. Challenges
3. Related Work
4. Model
5. Results
6. Future Work
7. Conclusion

Background

- Chromosome Conformation Capture (often abbreviated 3C) is a set of methods to determine the spatial organization of chromatin in a cell
    - Understanding spatial organization of chromatin is necessary to find patterns in interactions between genes
- Topologically Associated Domains (TADs) are areas where DNA physically interact with each other and are most known for regulating gene expression
- 3C (one-to-one), 4C (one-to-all), 5c (many-to-many), Hi-C (all-to-all)

Motivation

1. Accurately modeling chromatin within cells
    a. Being able to accurately model chromosomes and discovering structural features of chromosomes
2. Viewing patterns of interactions between chromatins and TADs to predict disruptions in cell regulation and replication
    a. Ability to predict potentially cancerous cells and capturing TAD boundaries accurately
3. Performing Hi-C analysis of a chromosome is expensive

Data

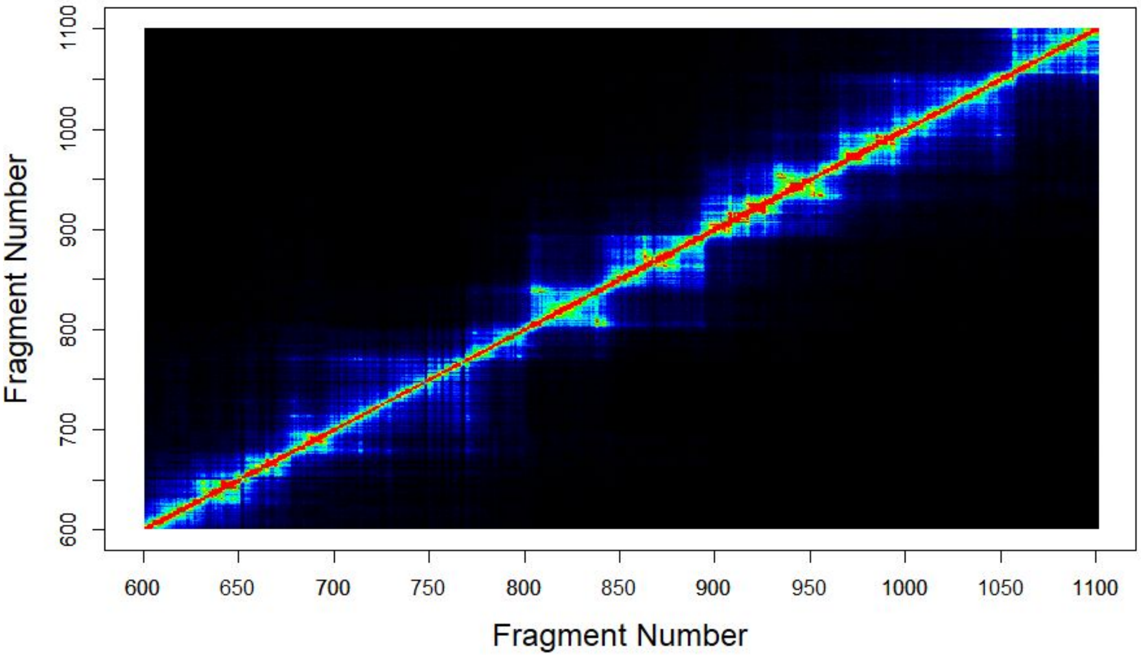
2 cells

- Gm12878 (Normal)
- K562 (Cancer)

Intensity matrix per chromosome

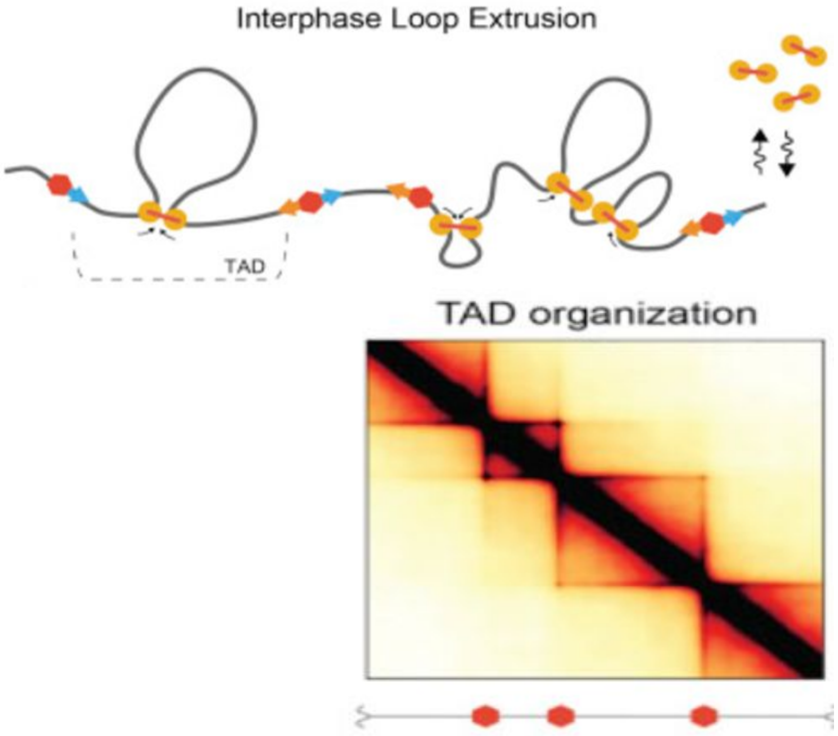
Elements closer to diagonal have higher values

Epigenetic data set with 36 covariates



Challenges

1. Computational
    a. Modeling all data at once is unfeasible (600 million elements in each chromosome; 4.6gb per chromosome)
2. Statistical
    a. Discovery of TAD (Topologically Associating Domains) boundaries with accuracy
    b. Acceptable accuracy with sparse number of predictors
    c. Relatively generalizable model (Cancer vs non-cancer)



Research Questions

1. Can we model the entire chromosome with high accuracy?

2. Is our model capturing TAD boundaries?

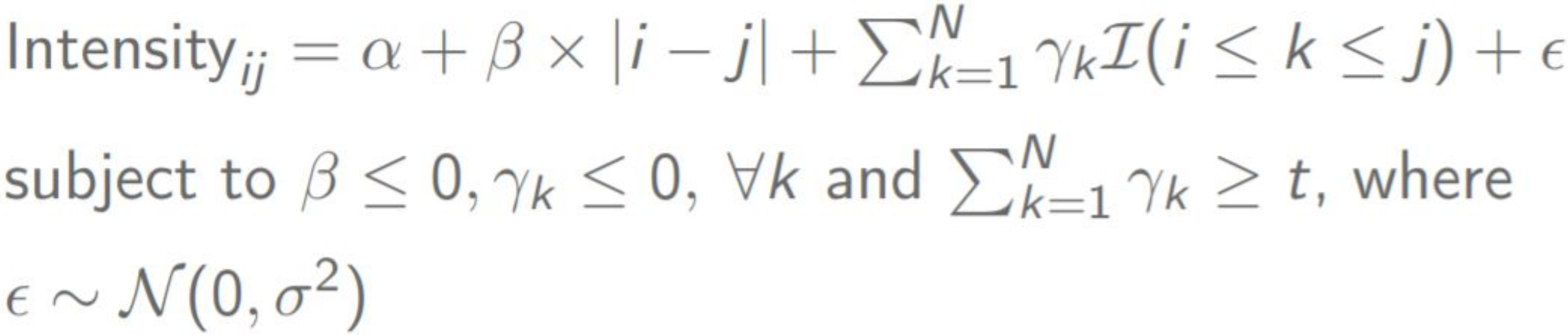
3. Are there notable differences between cancer and normal cells?

Related Work

1. Predicting High-order Chromatin Interactions from Human Genomic Sequence using Deep Neural Networks (Peng 2017)
   a. Calculated probability of interaction

2. Predicting the spatial organization of chromosomes using epigenetic data (Mourad 2015)
   a. Used Bayesian additive regression trees (BART) to find TAD boundaries

Model



$$\text{Intensity}_{ij} = \alpha + \beta \times |i - j| + \sum_{k=1}^{N} \gamma_k \mathcal{I}(i \leq k \leq j) + \epsilon$$

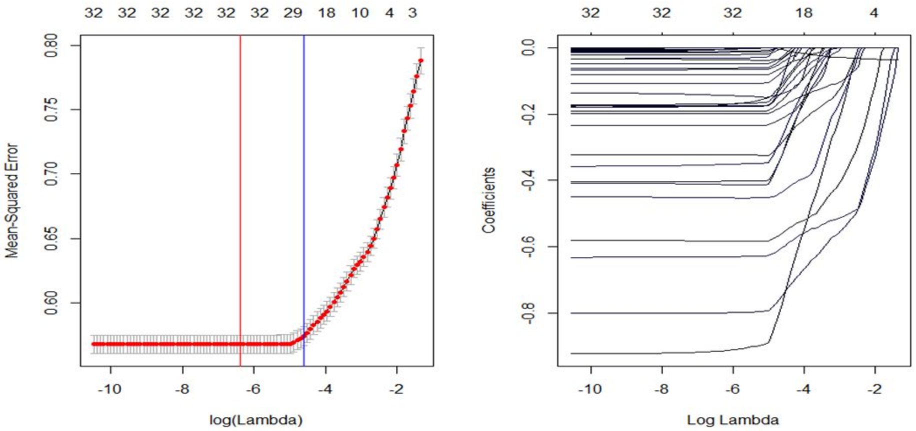
$$\text{subject to } \beta \leq 0, \gamma_k \leq 0, \ \forall k \text{ and } \sum_{k=1}^{N} \gamma_k \geq t, \text{ where}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Incorporate term for distance from the diagonal
- Evaluate important indices (Which locations have high intensity)

Parameter Optimization

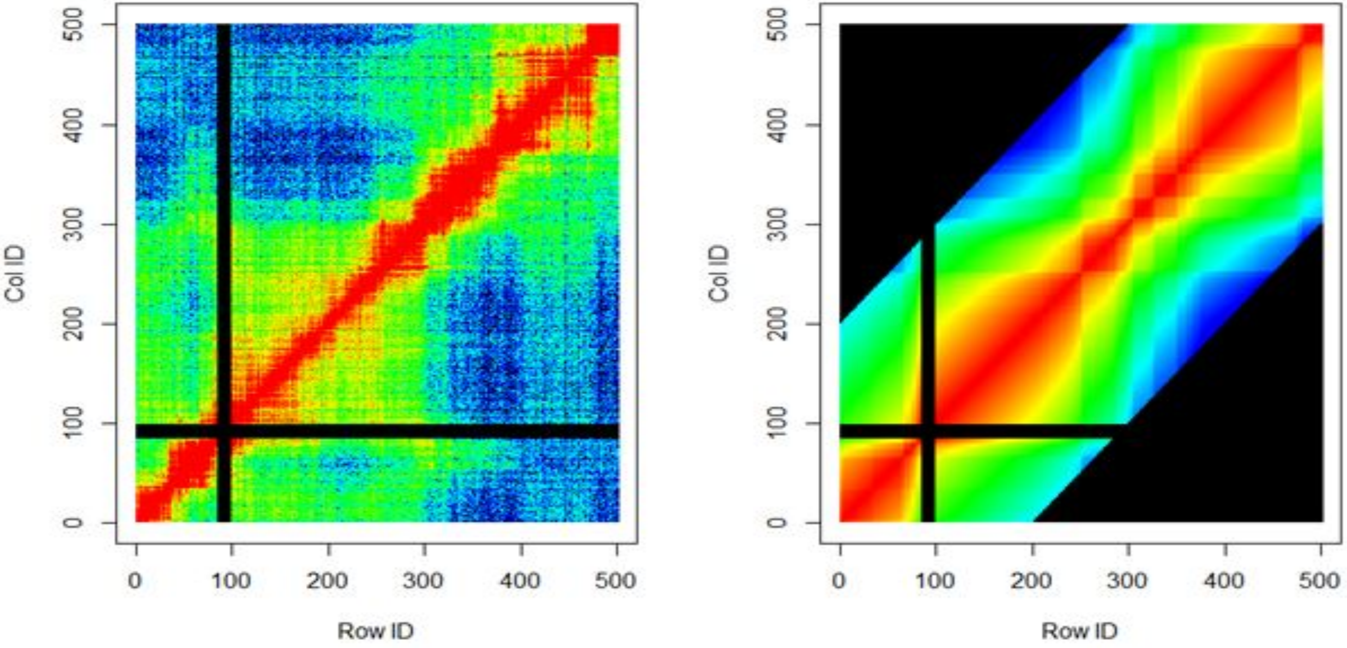
1. Fit the model over the data only penalizing fragment parameters
2. 10-fold Cross-validation to choose tuning parameter λ
3. Select λ_{1se} to impose higher penalty and more shrinkage



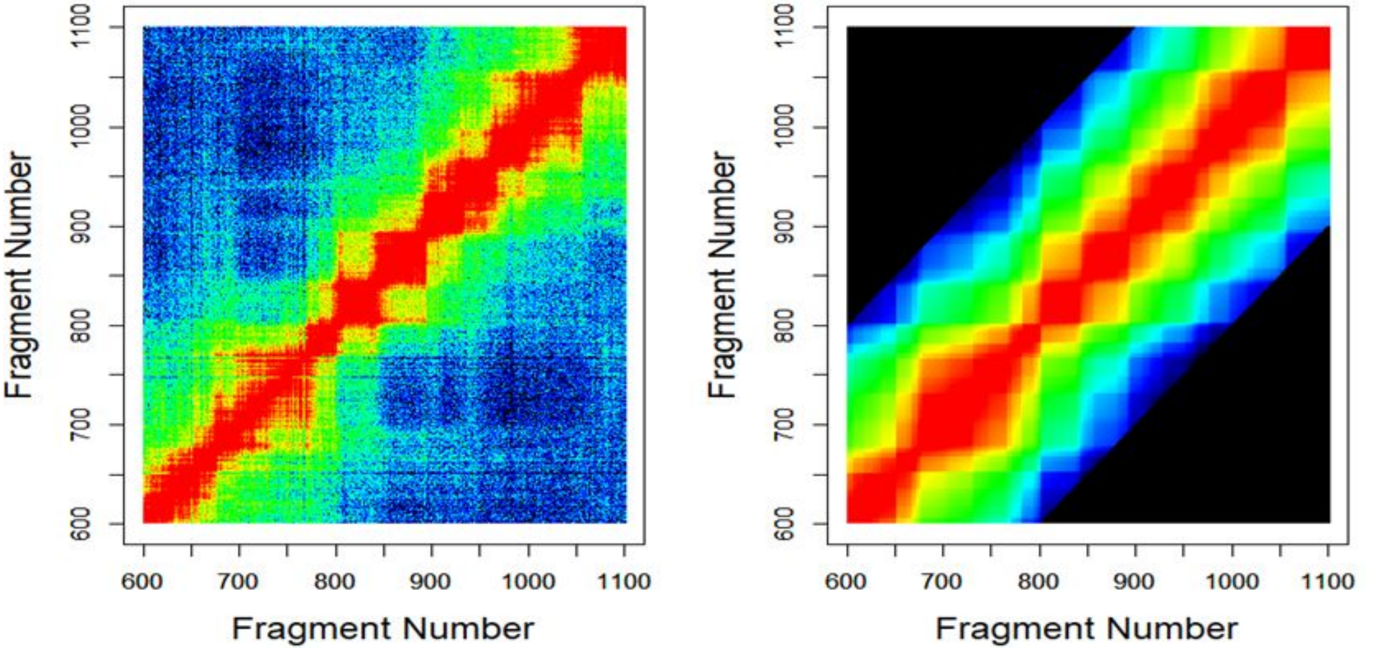
Procedure

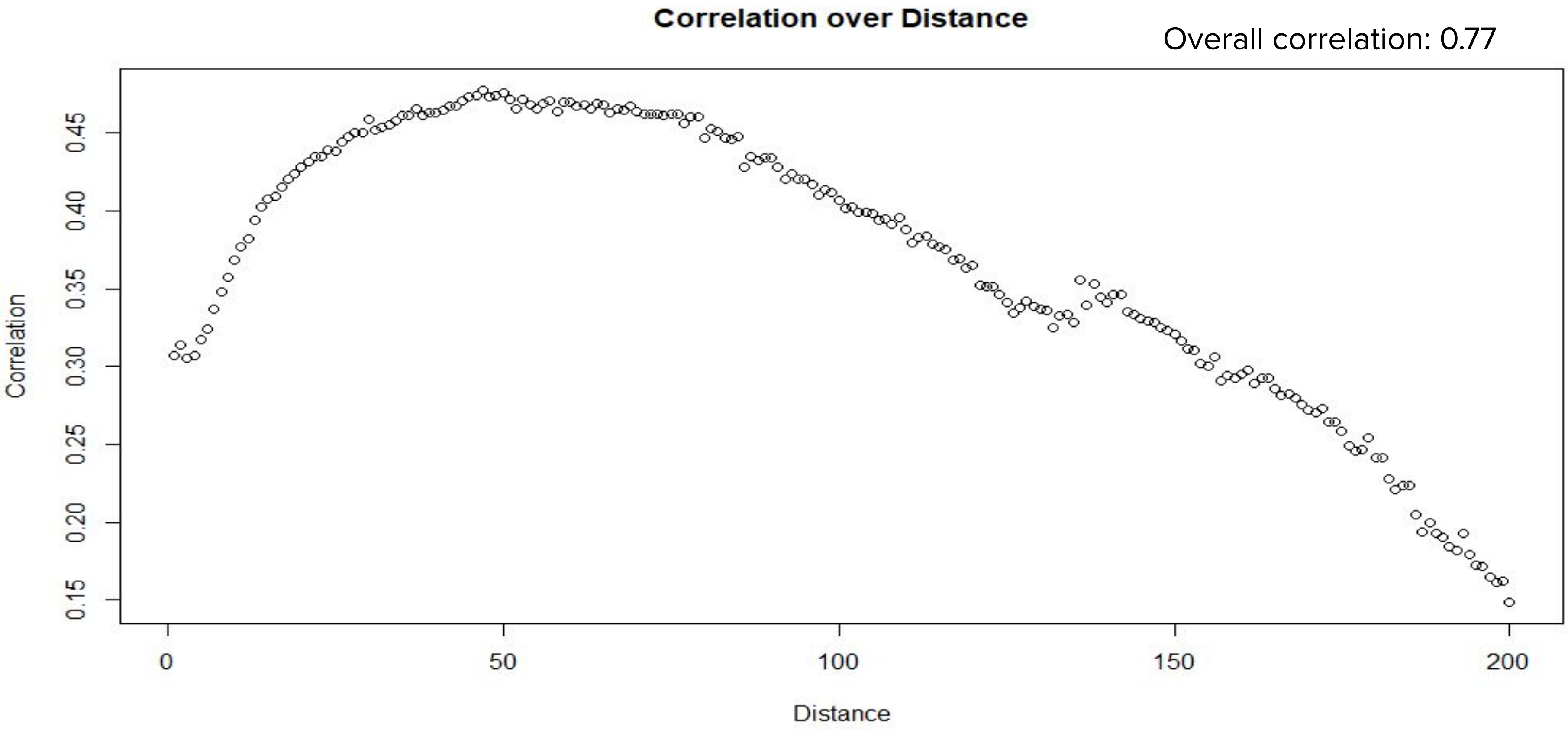
1. Take a slice of the entire chromosome (500 indices)
2. Removed indices that contained very little signal (less than 100)
3. Use a maximum range of 200 indices to use as predictors for each index
   a. Helps with computation times
   b. Create a design matrix to indicate which indices to use
4. Model data and find coefficients
5. Move slice 300 indices
6. Take average of overlapping coefficients

Results (alpha = 0.9)

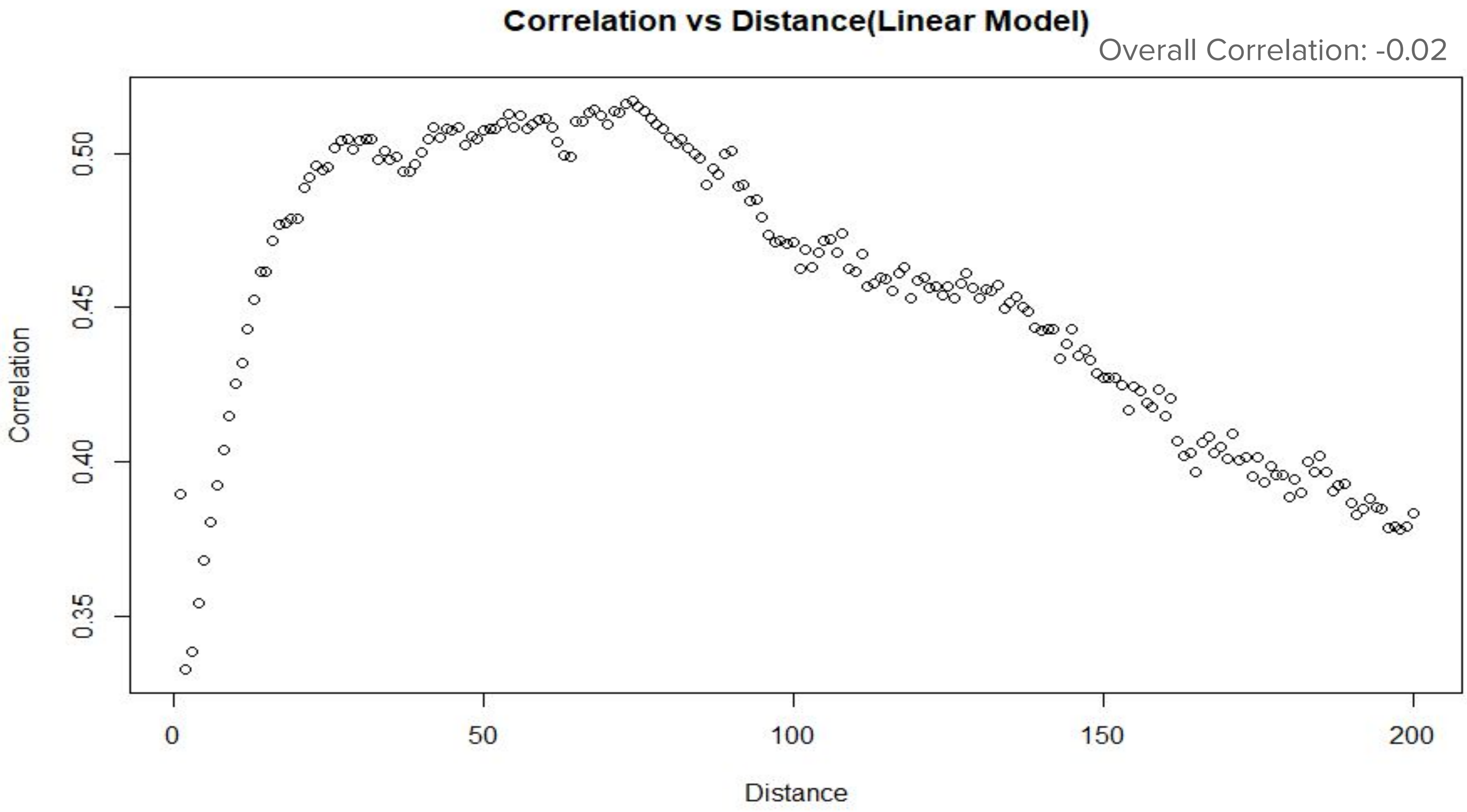


Results (alpha = 0.5)

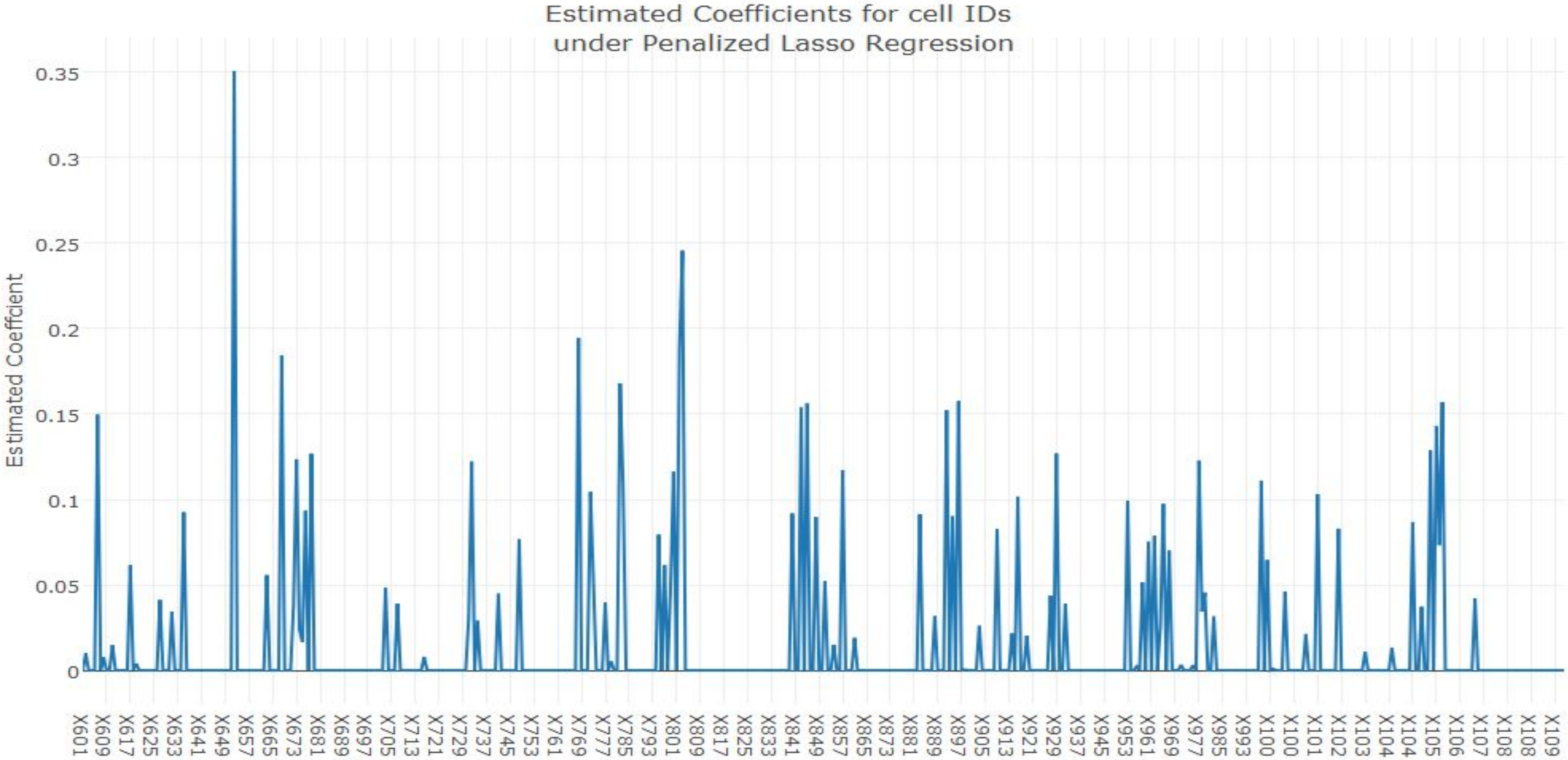




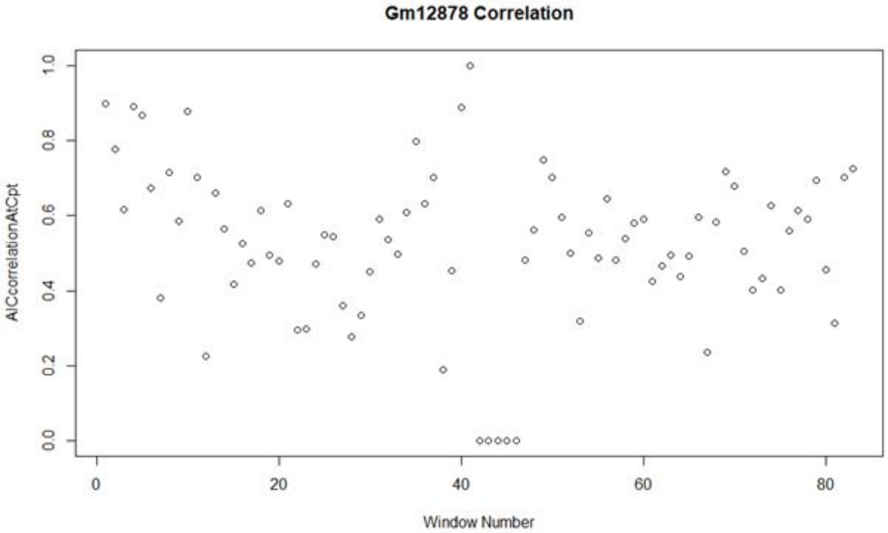
Overall correlation: 0.77

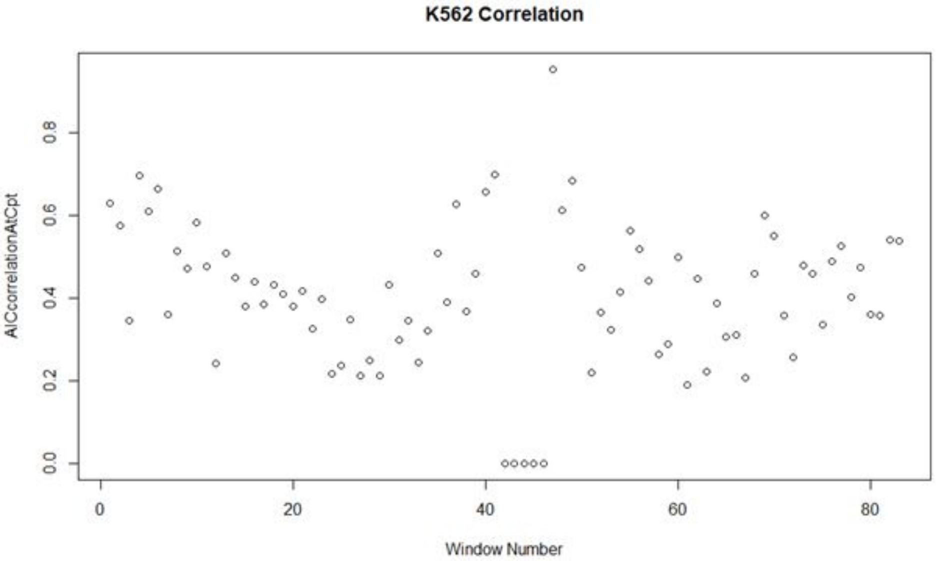


Overall Correlation: -0.02


Estimated Coefficients for cell IDs
under Penalized Lasso Regression
Estimated Coeffcient
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
X601
X609
X617
X625
X633
X641
X649
X657
X665
X673
X681
X689
X697
X705
X713
X721
X729
X737
X745
X753
X761
X769
X777
X785
X793
X801
X809
X817
X825
X833
X841
X849
X857
X865
X873
X881
X889
X897
X905
X913
X921
X929
X937
X945
X953
X961
X969
X977
X985
X993
X100
X100
X101
X102
X103
X104
X104
X105
X106
X107
X108
X108
X109

Correlations among windows of both types of cell





https://aclheexn.shinyapps.io/Vis_App/

Future Work

- Try different size of slices and size of maximum range of indices to predict
- Create a model using epigenetic data to predict locations of TAD boundaries
- Try other methods for comparison to our method
- Look for significant differences between cancer and normal cells for coefficients
- Try method on other chromosome numbers

Conclusion

1. We can use penalized LASSO to obtain TAD boundaries

2. There is an inverse relationship between the correlation we obtain and the ability to get accurate TAD boundaries

3. Performance of our regression still performs worse than state of the art machine learning techniques

Questions?