HiC Data Analysis

Amal Agarwal, Alex Chen, Lingzhou Xue, and Yu Zhang

Outline

1. Introduction
2. Methodology
3. Results
4. Summary and Future Work

HiC is a chromosome conformation capture technique

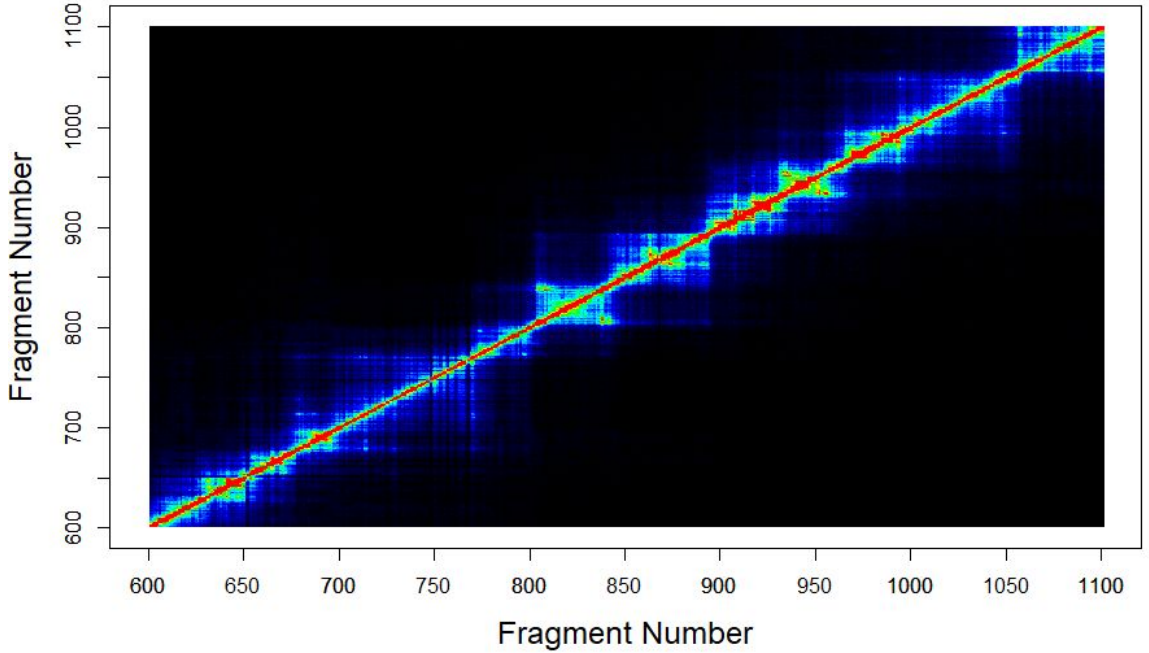
- Used to analyze interactions within a chromosome (Lieberman-Aiden et al., 2009)
- Quantifies the interactions between all possible pairs of fragments
- ~3 billion base pairs(b.p.) in the human genome
- Gene locations(fragments) consists of ~10K b.p. (Berkum et al., 2010)
- Expensive chromosome capture technique

Data Structure

- HiC
  - Two Human Cell types: “Gm12878”(Normal) and “K562”(Cancer)
  - For each cell, 22 normal and an ‘X’ chromosome
  - An intensity matrix for each chromosome could have different granularities. For e.g. 10K, 40K
  - ~600M. Elements in the HiC matrix per chromosome
- Epigenetic data set
  - 36 covariates

Goals

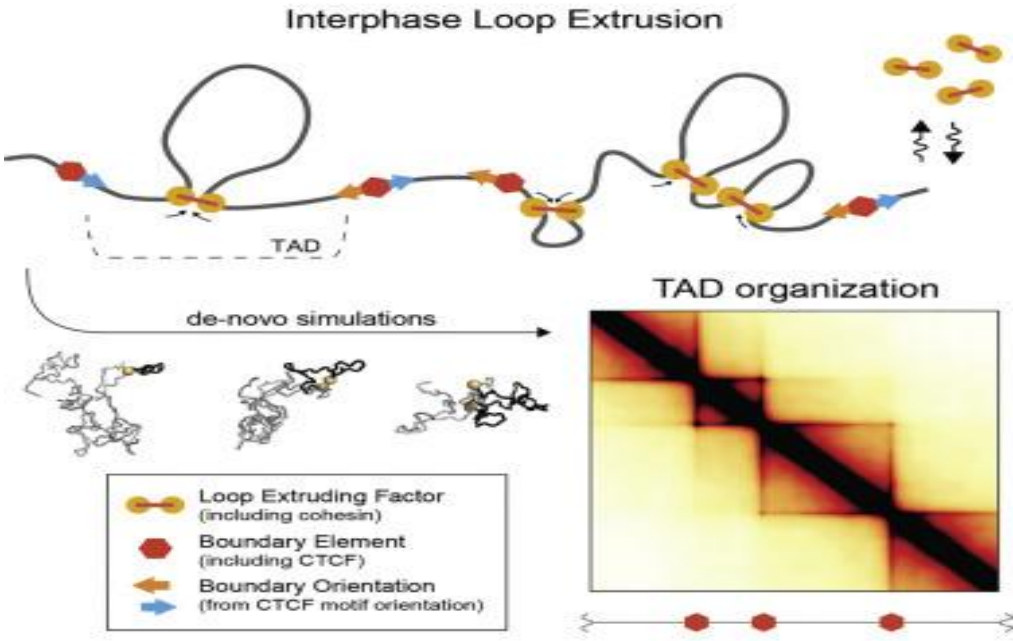
- Using HiC data to fit a model to uncover important gene locations
- Building models over epigenetic data to predict important gene locations



Visual representation of the intensity matrix from the Gm12878 cell chromosome 1 from fragment numbers: 601 to 1100

Challenges

- Methodological
  - Capturing true Topologically Associating Domains (TADs; Dixon et al., 2012, Nora et al., 2012)
- Computing
  - Processing large amounts of data in an adequate amount of time
  - Unable to model the entire chromosome in one go



A visual showing how the attraction between two gene locations in figure A causes the box shapes in figure B (Fudenberg et al. 2016).

Data Preprocessing

1. Summed each index and removed indices with less than 100 signal
2. Used a bandwidth parameter to define which gene locations to use as predictors
    a. A bandwidth of 200 for index 1, would use gene locations up to gene location 200
3. Design matrix created indicating 1 if gene location is within bandwidth, 0 otherwise.
4. HiC matrix transformed to a dataframe of intensities and design matrix

Penalized Lasso Regression Model

- Shrinkage estimations are popular for high dimensional data (Tibshirani 1996)
- Important genes after shrinkage should reveal all the TAD boundaries
- Distance from the main diagonal, | i - j |, could explain intensity variation

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

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

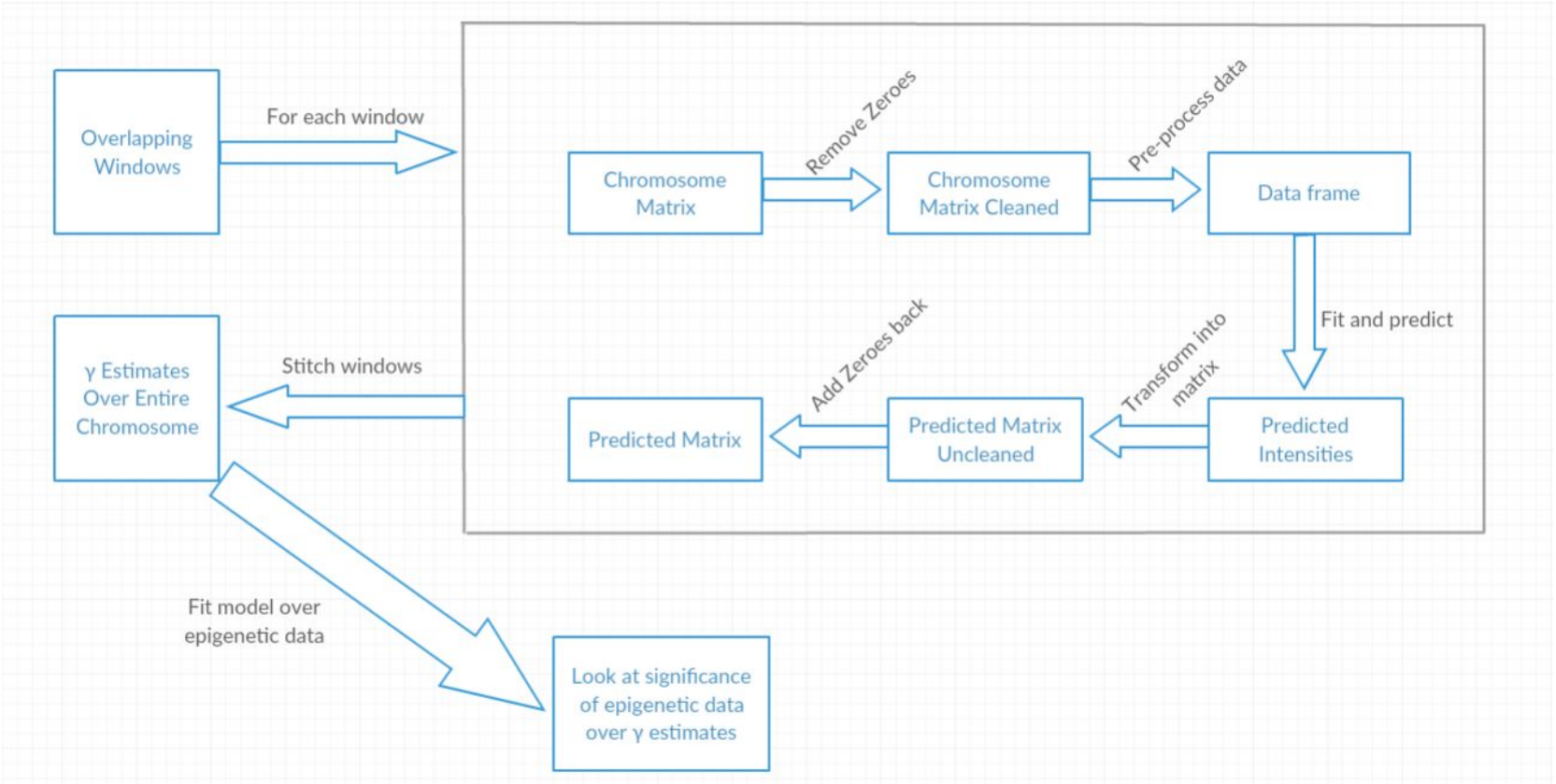
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Where N is the number of fragments in the chromosome, (i , j) are row and col. Indices in HiC matrix

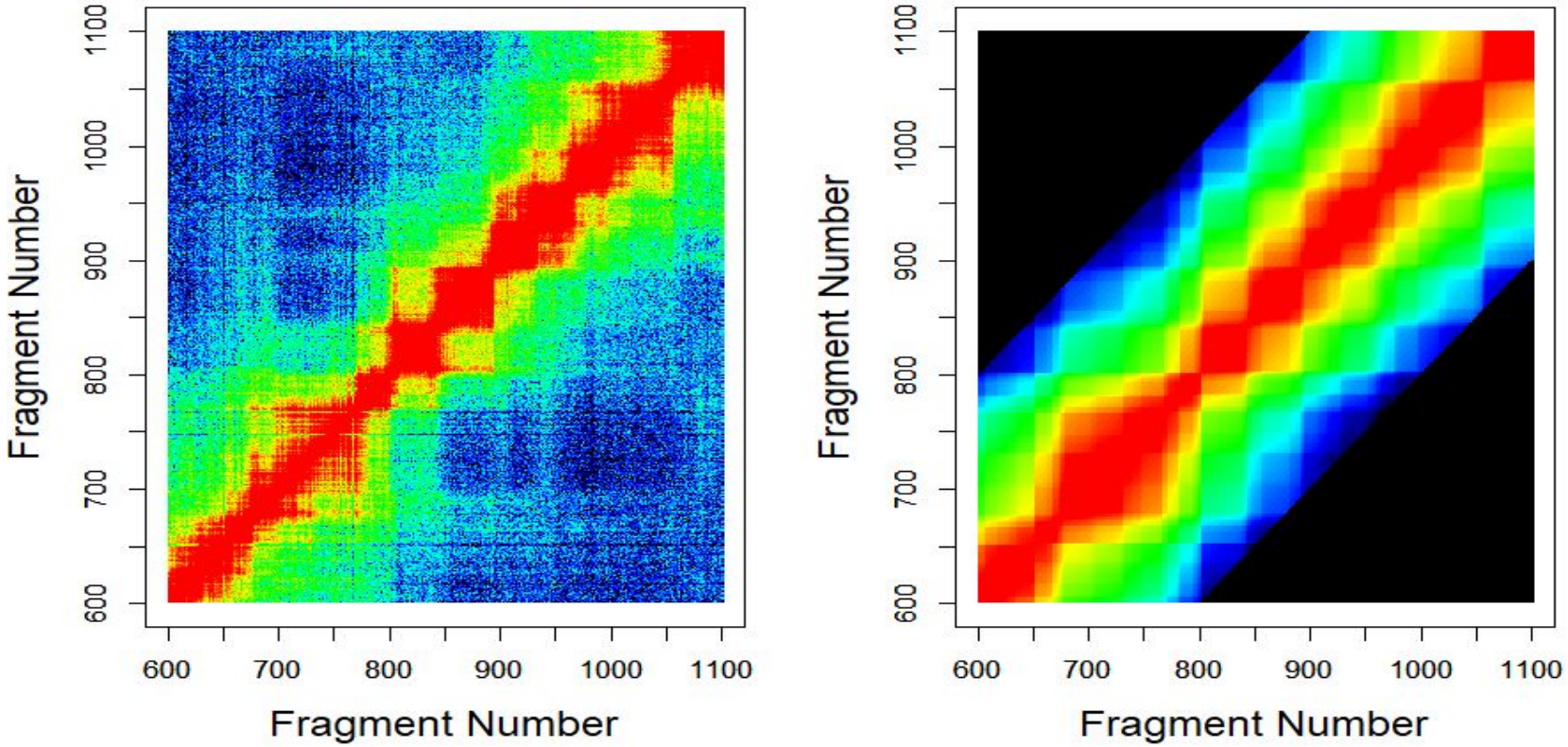
Penalized Lasso estimation procedure

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

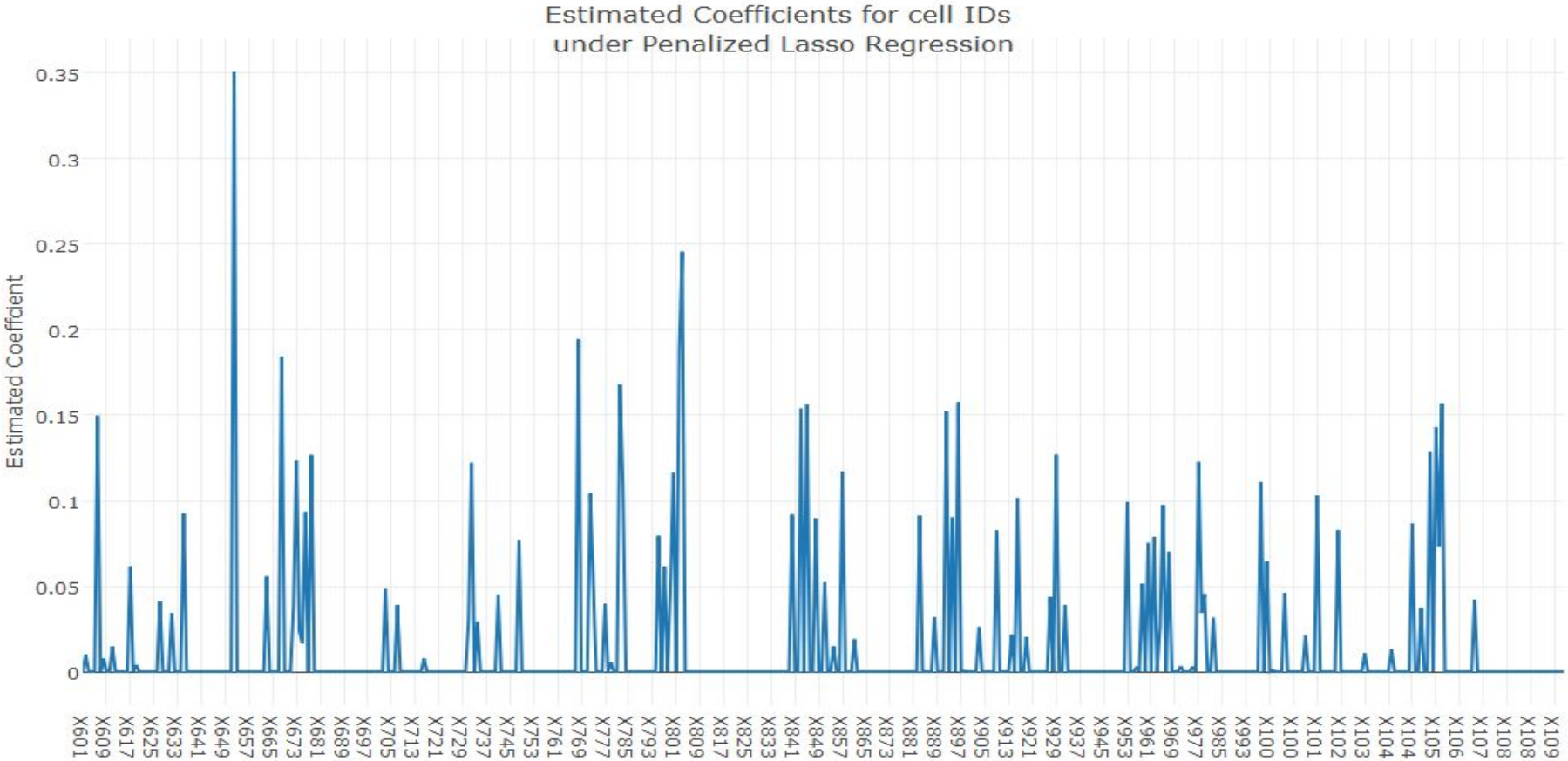
Current Method



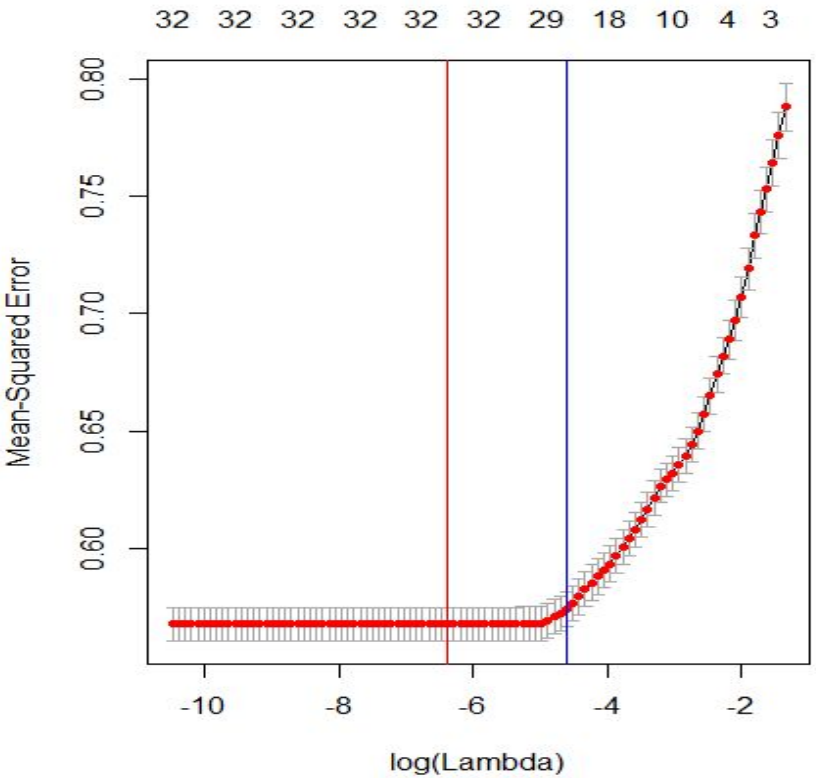
HiC matrices original and predicted on Gm12878 chromosome 1, window (601 - 1100)

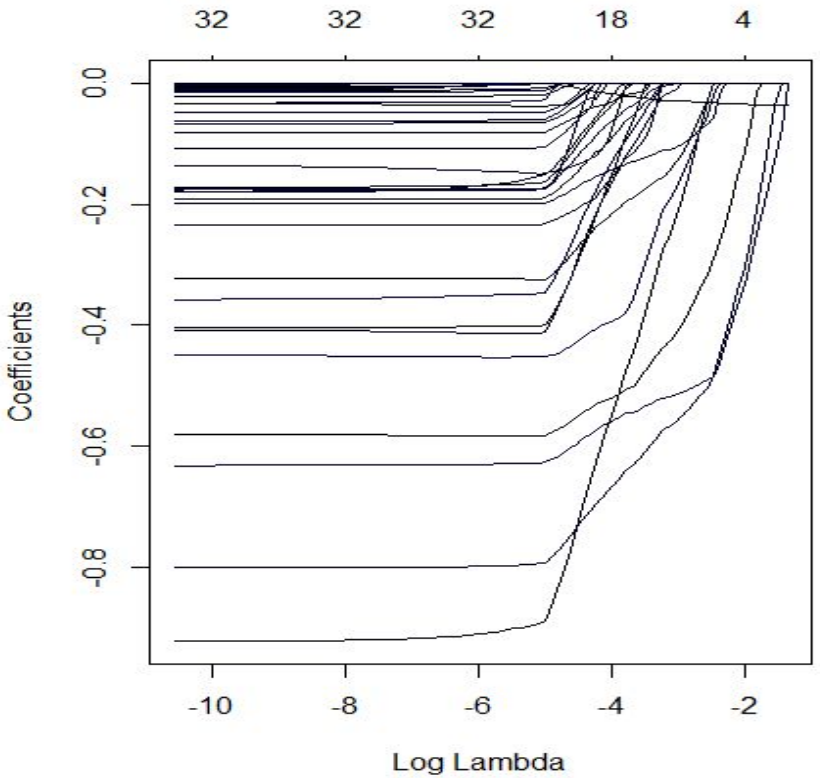


(Left) original chromosome intensity matrix of cell Gm12878 chromosome 1 and (Right) the predicted matrix using our penalized Lasso model

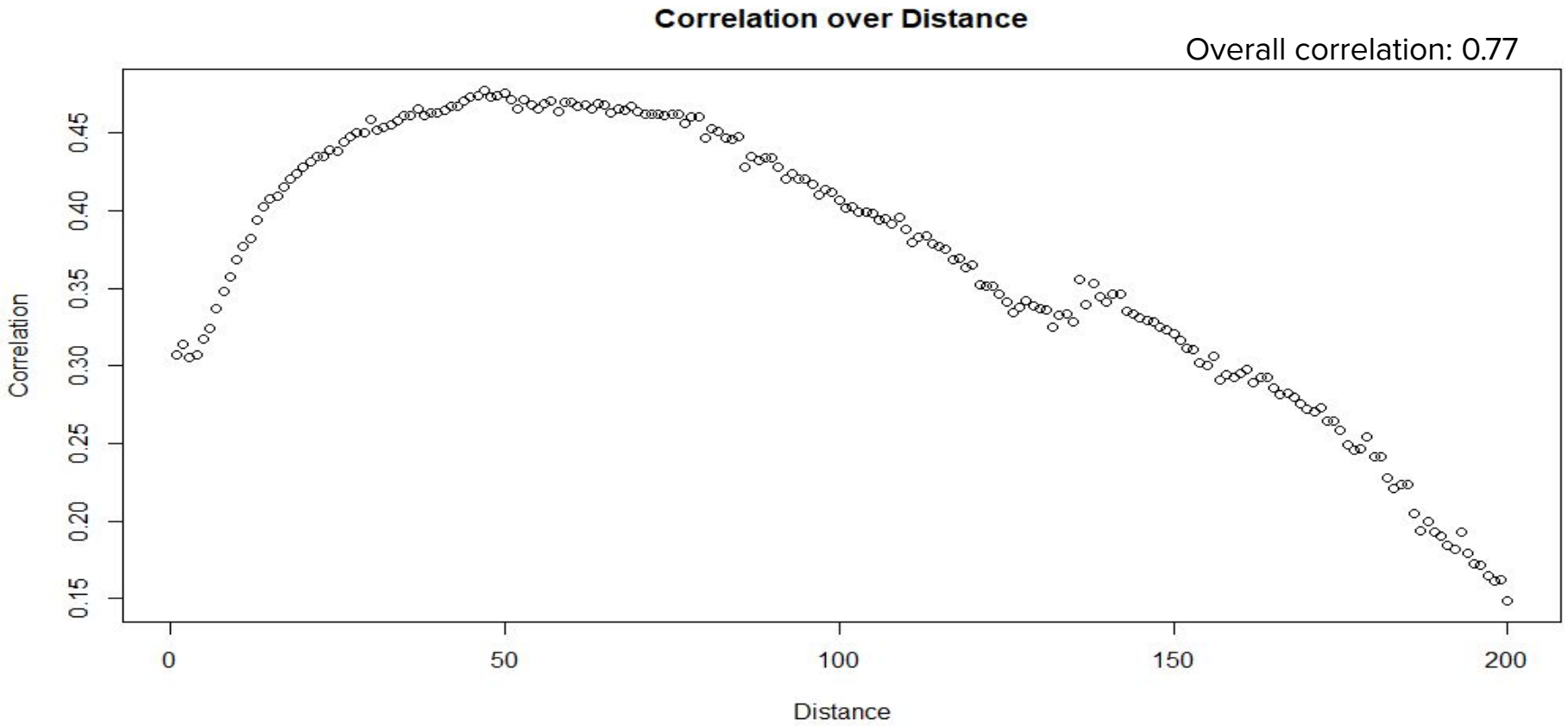


A plot of the gamma coefficients in cell Gm12878 chromosome 1 where large spikes correspond to TAD boundaries found in the previous slide. Coefficients were translated to positive for readability.

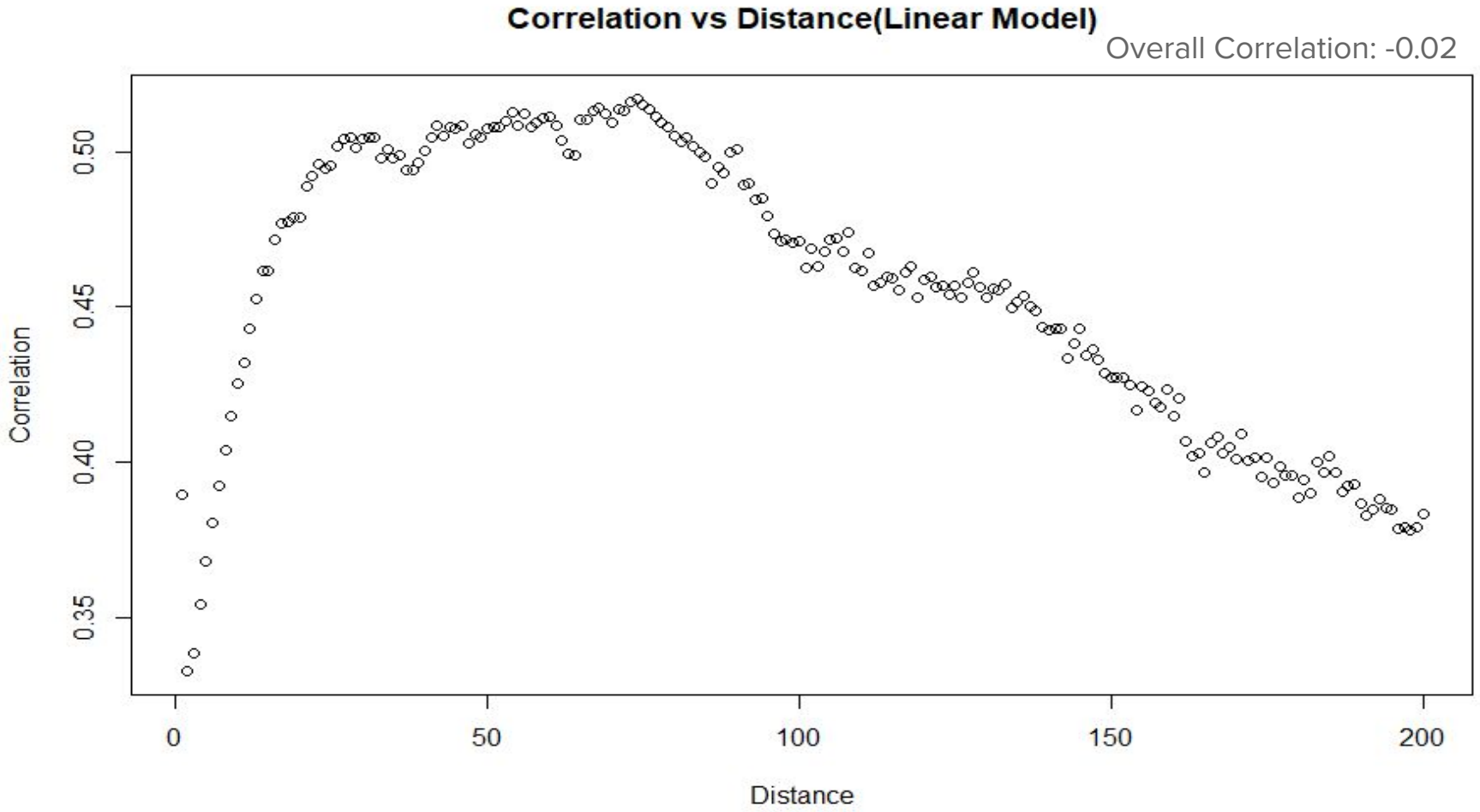




(Left) Graph of optimal λ values vs mean squared error. Blue line indicates λ_{min}. Red line indicates λ_{1se}. (Right) Regularization path



Pearson Correlation varying with distance from main diagonal over predicted vs. original matrix in the Gm12878 cell for chromosome 1


Correlation vs Distance(Linear Model)
Overall Correlation: -0.02
Correlation
0.35
0.40
0.45
0.50
0
50
100
150
200
Distance

Summary and Future Work

So far

- Pre-processed data
- Penalized Lasso Model captures TAD boundaries

Upcoming

- Create a baseline linear regression model without shrinkage for HiC data
- Experiment with window sizes, overlap parameter and different sub-sampling procedures
- Mixture Models taking into account orientation and fragments
- Checking consistency over replicates
- Comparing normal and cancer cells