Deconvolution of HiC Data With ATAC-seq Data Simulation Report

Ziqiao Ma
2020/2/16

1 Brief Background

Single cell HiC is a widely applied sequencing technique useful regarding investigations of 3D chromatin conformation inside the nucleus. Yet, it remains challenging to deconvolute tissue specific HiC data into fractions corresponding to cell types. Our goal is to deconvolute the HiC data from an islet tissue into HiC data specific to α and β cells, and late the other 3 minority type (δ for example).

For simulation purpose, we verify the deconvolutability on h1HESC and HFF:

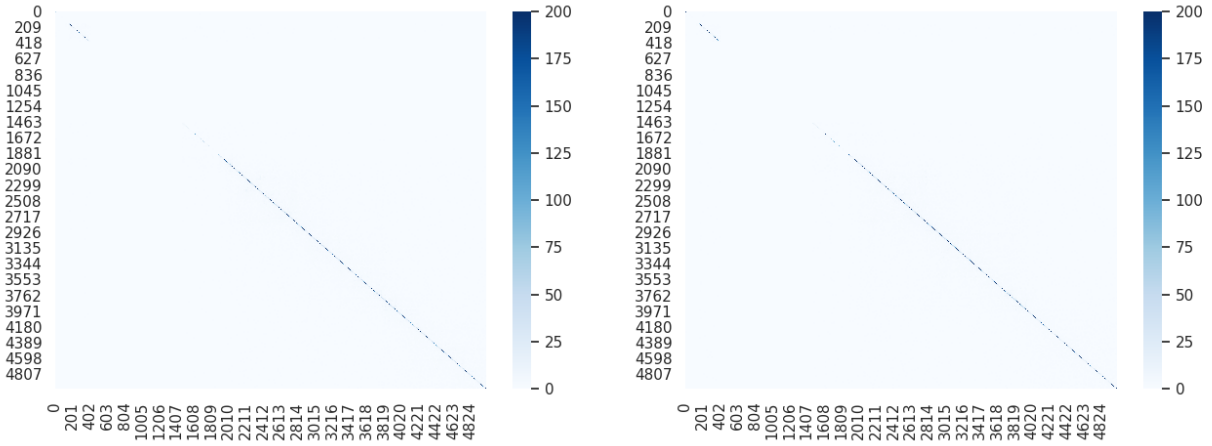
- Obtain HiC data of both cell types, plot the heatmap for ground truth;
- Mix data by adding the contact matrix in ratio of $1:1$ to simulate issue specific HiC data;
- Decolvolute the mixed HiC data based on cell type specific ATAC-seq data of both types;
- plot the obtained heatmap and evaluate the performance based on HiCRep Score.

2 Brief Algorithm

For simulation, we choose chr21, ChromStart = 1e7, ChromEnd = 2e7.

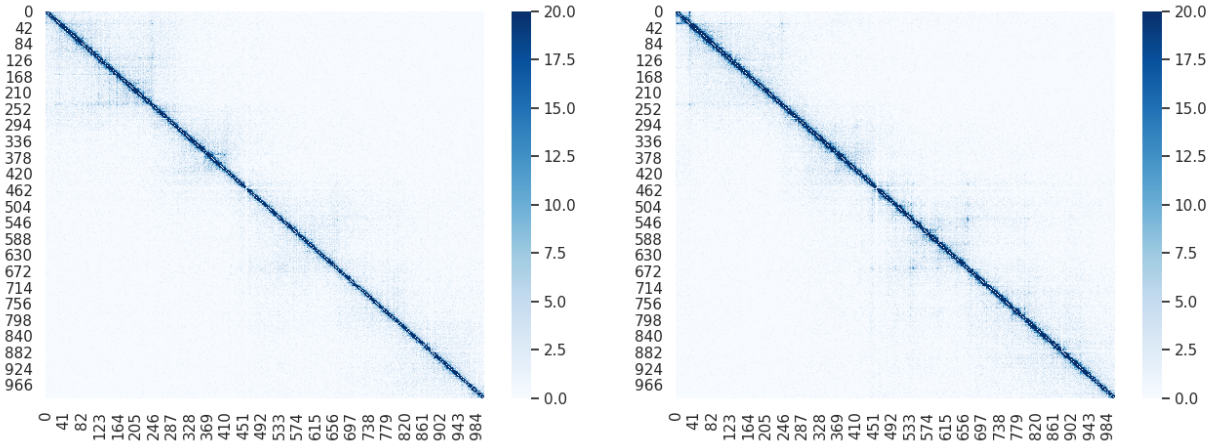
- Lift over reference genome. The HiC data are obtained with hg38 while the bed files of h1HESC and HFF are obtained wrt hg19. This is down with `LiftOver`, from `http://genome.ucsc.edu/cgi-bin/hgLiftOver/`.
- Dump HiC data in 2kbp of specified region with `juicer_tools` (see `https://github.com/aidenlab/juicer/wiki/Data-Extraction`), and read out all open regions in the studied interval from the bed files.
- Process the HiC data into contact matrices, and plot the heatmap with `juice box`. In this simulation, for simplicity, I plot with `seaborn.heatmap`.

The plots are as follows.



(a) h1ESC HiC Plot, in region $[1e7, 2e7]$ (b) HFF HiC Plot, in region $[1e7, 2e7]$

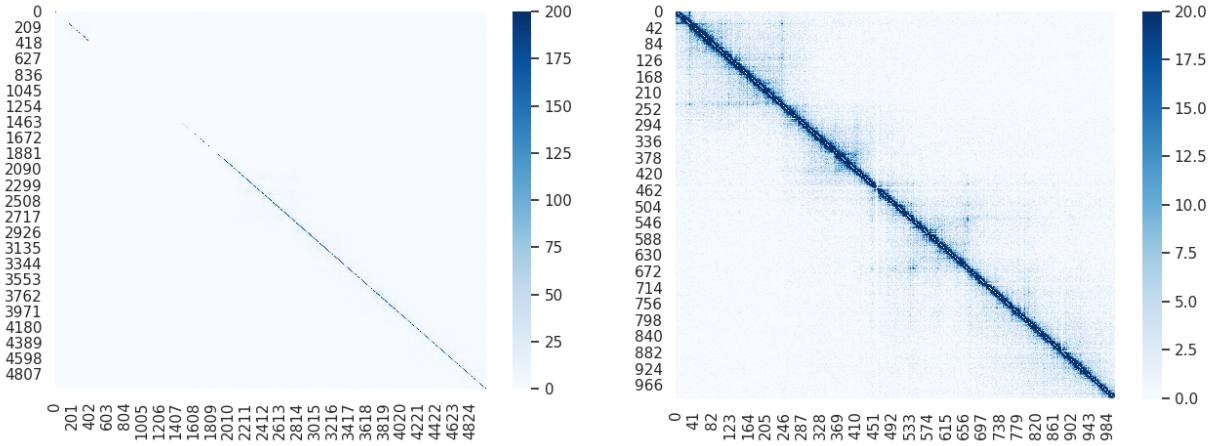
Figure 1: HiC plot over $[1e7, 2e7]$



(a) h1ESC HiC Plot, in region $[1.5e7, 1.7e7]$ (b) HFF HiC Plot, in region $[1.5e7, 1.7e7]$

Figure 2: HiC plot over $[1.5e7, 1.7e7]$

- Add up the contact matrix and make the mixed plot.



(a) Mixed HiC Plot, in region $[1e7, 2e7]$

(b) Mixed HiC Plot, in region $[1.5e7, 1.7e7]$

Figure 3: Mixed plots

- Calculate the open rate in each HiC interval. In this case:

$$\texttt{Open rate} = \frac{\texttt{length of open region}}{2000}$$

- Calculate the joint activate rate for each pair of positions on the contact map (each element in the matrix) defined by

$$\texttt{activate rate} = \texttt{open rate}_1 \cdot \texttt{open rate}_2 \cdot \texttt{decay}$$

where

$$\texttt{decay} = \exp(-\frac{\texttt{position}_1 - \texttt{position}_2}{2000})$$

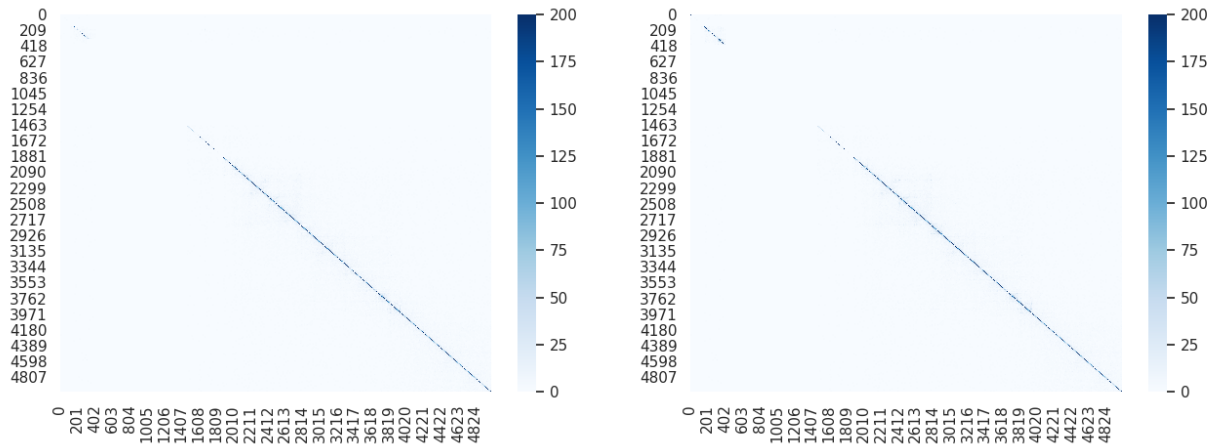
- Calculate the proportion of type i cell in each HiC matrix element by

$$\texttt{proportion} = \frac{\texttt{activate rate}_i}{\sum_{i=1}^{n} \texttt{activate rate}_k}$$

Set each proportion evenly when all activate rates are zero.

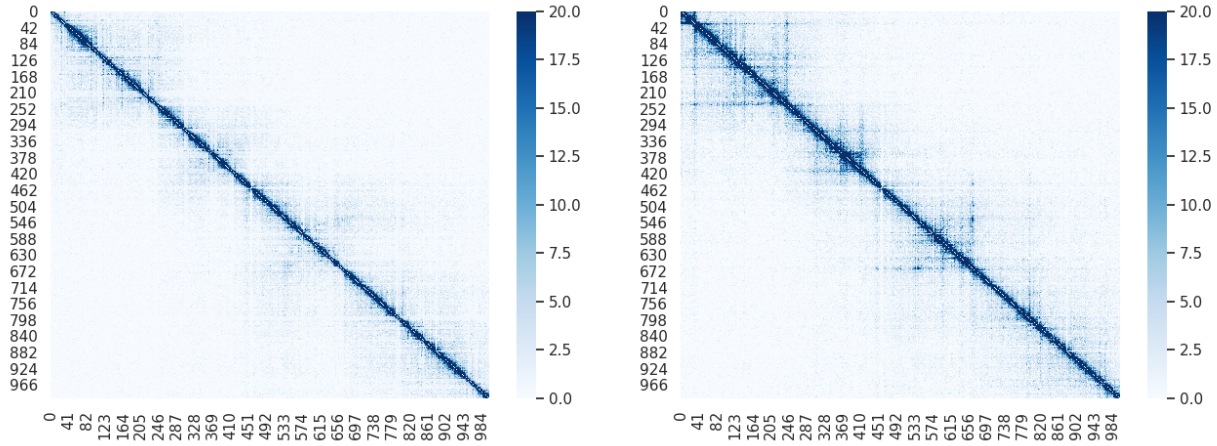
- Element-wise multiply the mixed HiC matrix with each proportion matrix to obtain the deconvoluted matrix.

- Plot the deconvoluted matrix and calculate the HiCRep score of each with original ground truth.



(a) Deconvoluted h1ESC HiC Plot, in region $[1e7, 2e7]$ (b) Deconvoluted HFF HiC Plot, in region $[1e7, 2e7]$

Figure 4: Deconvoluted HiC plot over $[1e7, 2e7]$



(a) Deconvoluted h1ESC HiC Plot, in region $[1e7, 2e7]$ (b) Deconvoluted HFF HiC Plot, in region $[1e7, 2e7]$

Figure 5: Deconvoluted HiC plot over $[1e7, 2e7]$

- Simulation output:

```
Data Loading Succeed!
The HiCRep score between 2 original HiC datasets is  0.9030655377293668
Heatmaps generated.
Mixed HiC heatmap generated (rate = 1:1).
Deconvolution Succeed!
The HiCRep score of hESC HiC deconvolution is  0.847168467644803
The HiCRep score of HFF HiC deconvolution is  0.8274473149221404
The HiCRep score between 2 deconvoluted datasets is  0.5766062583356114
Heatmaps generated.
```

3 Brief Discussion

Achieved so far:

- Deconvolution with HiCRep score over 0.8;
- Maintain special regions of each type in HiC Plots.

Problematic so far:

- Accuracy. The heatmaps seem reasonable only for large value. For small values, the algorithm is not accurate in many regions;
- Scattering. In true contact matrix, large values tend to gather around the diagonal. However, in my algorithm, the large values scatter around the diagonal, causing the color of some regions denser than it should be.

Possible improvements: Right now I have only access to some bed files, and the only valuable information is the open regions. The peak, peak value, variance of each open region signal are missing because I have no access to HFF `BigWig` file, (or to be better, `.bam` files of both data to call peak). If I could have access to signal information, I may try some non-supervised machine learning methods, or some numerical methods in Fourier domain.