Genome-wide flame feature detection pipeline for Hi-C chromatin conformation maps

Shiv Khandelwal

Sameer Abraham and Martin Falk, Mentors

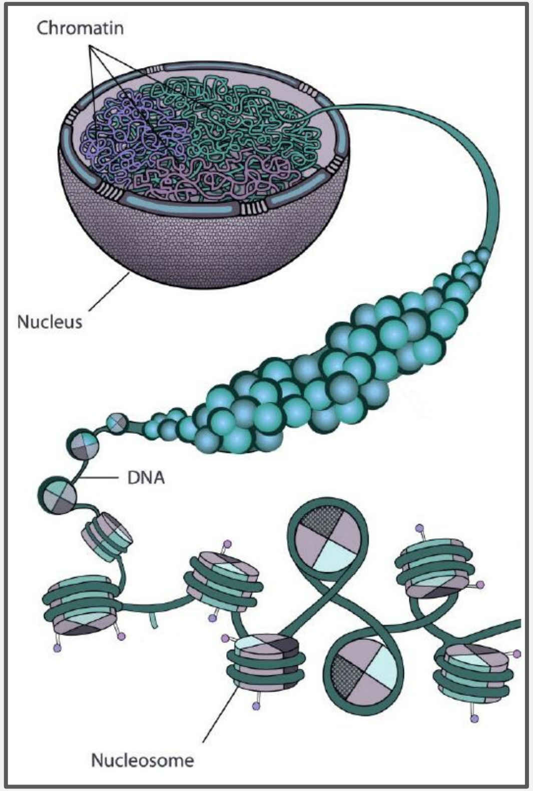
MIT PRIMES Conference

October 20th, 2019

Mirny Lab

Chromatin

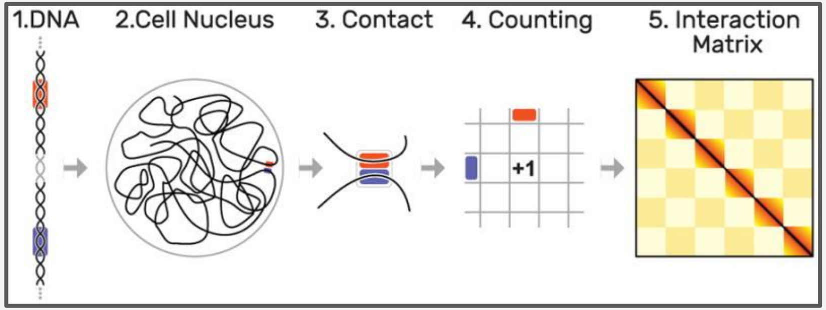
- Complex of DNA coiled around histone proteins
- Efficiently packages 2 m genome into 4-8 µm nucleus
- Preserves structure and sequence



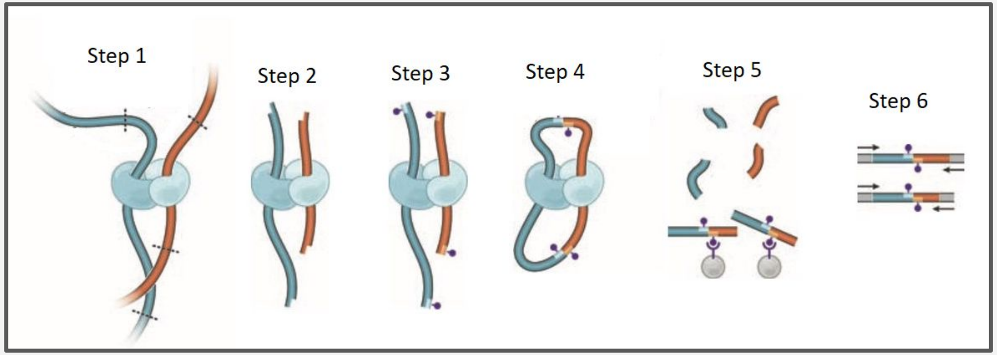
Background

Hi-C chromatin conformation heatmaps

- Genome-wide interaction maps
- Darker index indicates higher interaction between those two genomic loci
- Computed over ensemble average of over 10^7 cells
- Symmetric about the diagonal
- Checkerboarding pattern due to phase separation



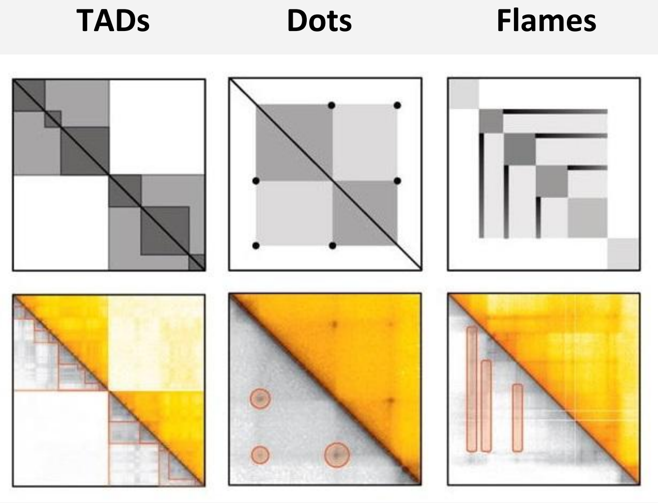
Gathering Hi-C data



1. Crosslink DNA
2. Cut with restriction enzyme
3. Fill and mark ends with biotin
4. Re-ligate
5. Purify and Sheer DNA
6. Sequence using paired-ends

Hi-C visual features

Definitions



Topologically Associating Domains (TADs)

- contiguous regions of increased contact frequency
- appear as relatively insulated squares

Dots

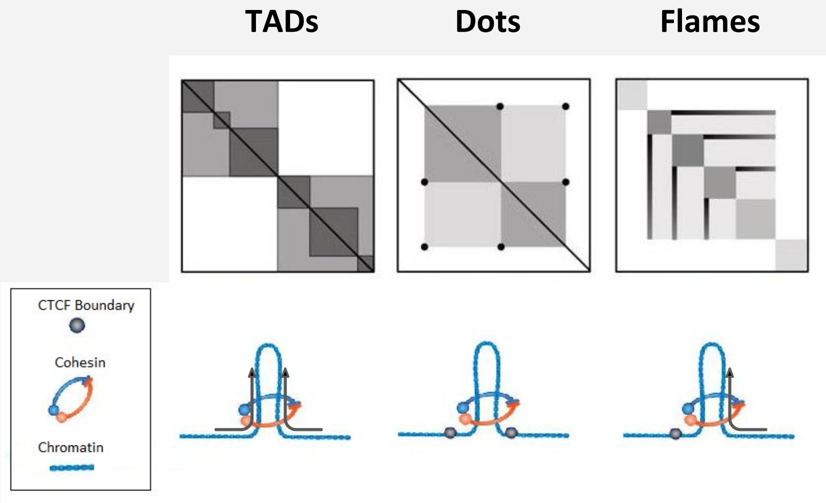
- Small circular regions of increased contact frequency
- Appear on corners of TADs

Flames

- Horizontal or Vertical linear regions of increased contact frequency
- Occasionally appear on border of TADs

Hi-C visual features

Loop Extrusion Mechanisms



Topologically Associating Domains (TADs)

- Cohesin is not blocked by CTCF on either side of chromatin fiber
- Loop is extruding through both sides of chromatin fiber

Dots

- Cohesin is blocked by CTCF on both sides of chromatin fiber
- Loop is temporarily immobile

Flames

- Cohesin is blocked by CTCF on one side of chromatin fiber
- Loop is extruding through one side of chromatin fiber

Research Question

How can we computationally locate and demarcate all flames within a Hi-C map?

Observations

- Classic line detection algorithms look for high variation between adjacent pixels
- Hi-C maps contain too much noise for classic line detection algorithms

⬇

Custom image processing pipeline

Methods

Observed Over Expected Normalization

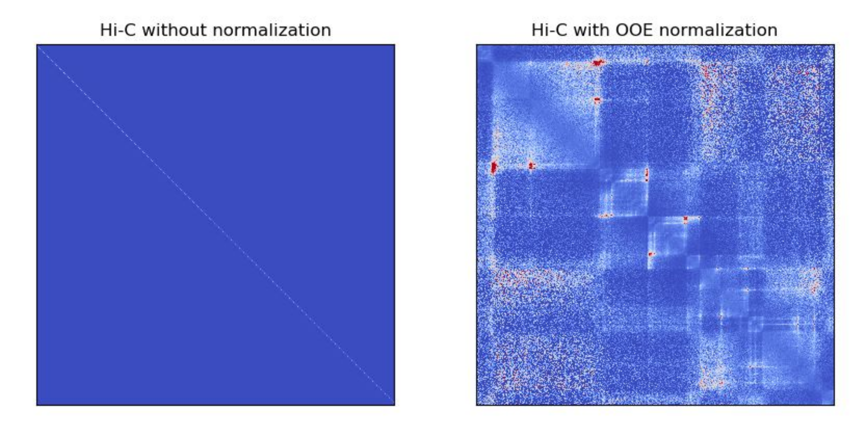
Step 1

Motivation

- Visual features are lost because relative intensity of main diagonal is so strong
- Need to counteract distance decay along main diagonal

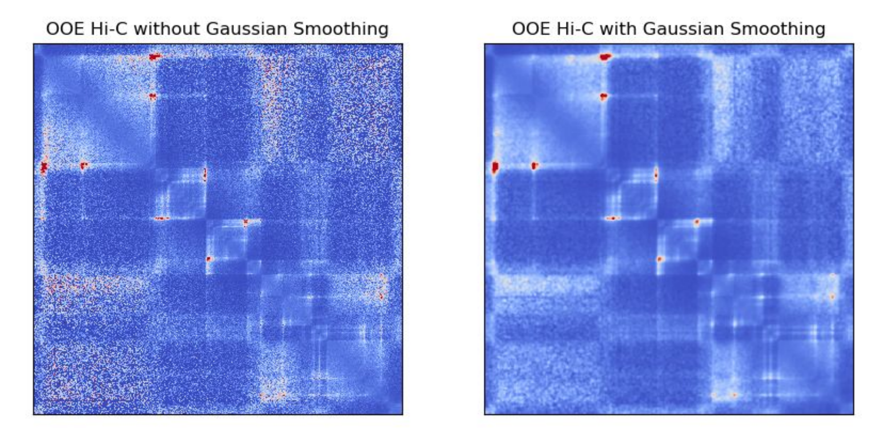
Mechanism

- Compute mean along each adjacent diagonal
- Divide values along diagonal by computed mean



Gaussian Smoothing

Step 2



Motivation

- Need to reduce noise to accentuate boundaries of flames
- Apply Gaussian filter (a type of low-pass filter) to smooth over entire map, reducing noise
- Preserves edges better than mean filter

Mechanism

- Normally distributed kernel computed by:

$$\boxed{G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}}$$

  - x is the x-coordinate
  - y is the y-coordinate
  - σ is the standard deviation of the distribution (optimized at 1.5)

- Kernel convolves around Hi-C map
- Outputs “weighted average” of each pixel’s neighborhood, average weighted towards central pixels

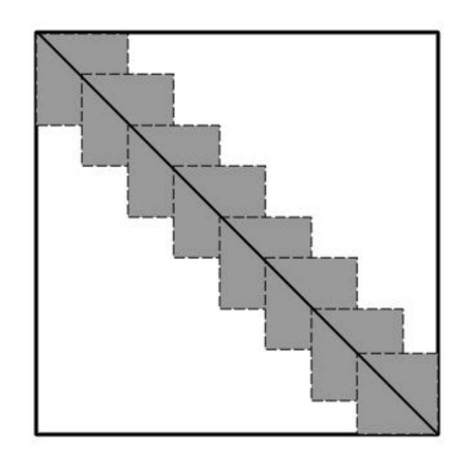
Diagonal Slicing

Step 3

Motivation

- The scale of a single Hi-C maps is measured in megabases (1,000,000 bases), which is too large
- One pass through the entire genome fails to pick up smaller features
- Looking at sub-regions in close proximity to main diagonal increases efficiency and accuracy

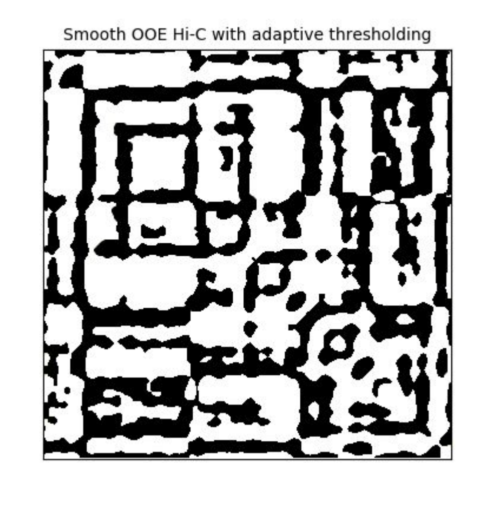
Sub-regions indicated by gray squares



Methods

Adaptive Thresholding

Step 4



Motivation

- Need to separate visual features from surrounding data in the Hi-C Map
- Accomplished by binarization with respect to a threshold

Mechanism

- Computes binary thresholded mask image based on local pixel neighborhood
- Threshold value:
    - Weighted mean for local pixel neighborhood of radius 20
    - Subtracted by a constant (optimized to 0.01)

Methods

Skeletonization

Step 5

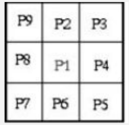
Motivation

- Reduce binary Hi-C to its structural skeleton to isolate flames

Mechanism

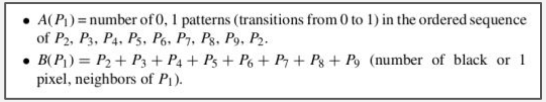
- Applying Zhang-Suen Thinning Algorithm to binary Hi-C map:

8 pixel local neighborhood:



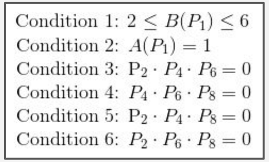
P9	P2	P3
P8	P1	P4
P7	P6	P5

Equations:

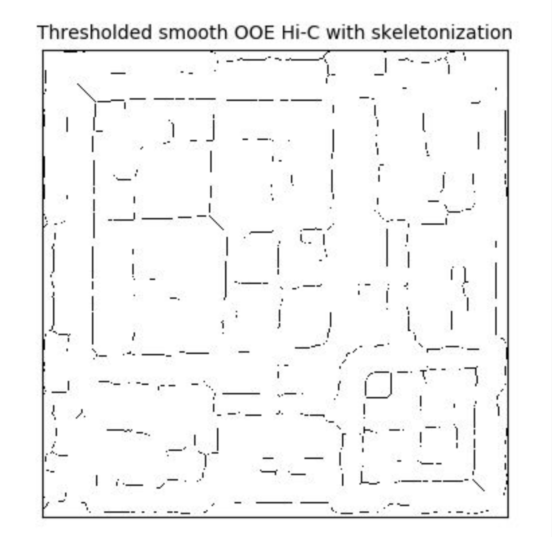


- $A(P_1)$ = number of 0, 1 patterns (transitions from 0 to 1) in the ordered sequence of $P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_2$.
- $B(P_1) = P_2 + P_3 + P_4 + P_5 + P_6 + P_7 + P_8 + P_9$ (number of black or 1 pixel, neighbors of P_1).

Conditions to turn P_1 from black to white:



Condition 1: $2 \leq B(P_1) \leq 6$
Condition 2: $A(P_1) = 1$
Condition 3: $\mathrm{P}_2 \cdot P_4 \cdot P_6 = 0$
Condition 4: $P_4 \cdot P_6 \cdot P_8 = 0$
Condition 5: $\mathrm{P}_2 \cdot P_4 \cdot P_8 = 0$
Condition 6: $P_2 \cdot P_6 \cdot P_8 = 0$



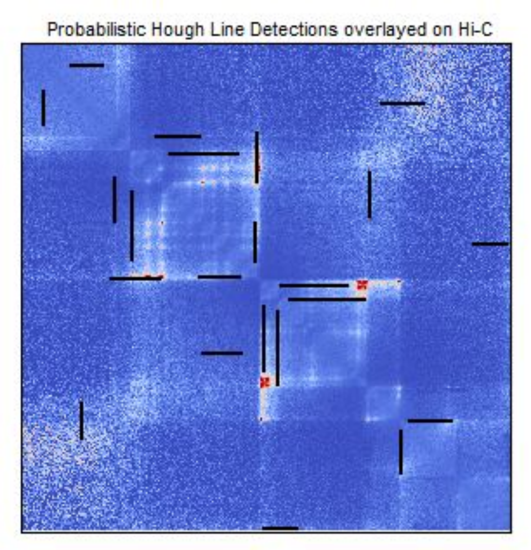
Thresholded smooth OOE Hi-C with skeletonization

Probabilistic Hough Transform

Step 6

Motivation

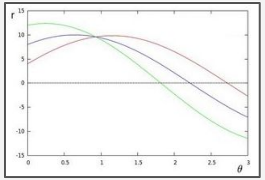
- Need algorithm to demarcate flames from skeleton binary Hi-C
- Should be computationally efficient given size of Hi-c maps



Probabilistic Hough Line Detections overlayed on Hi-C

Mechanism

- Standard Hough Transform:
  - Represent line in polar form: $\boxed{r = x\cos\theta + y\sin\theta}$
  - Define family of lines going through point (x_0, y_0): $\boxed{r_\theta = x_0 \cdot \cos\theta + y_0 \cdot \sin\theta}$
  - Example: Family of lines going through points (8, 6), (4, 9), (12, 3) in Θ-r plane:



  - Line containing all three points defined by intersection r = 9.6, Θ = 0.925
  - Hough Transform tracks intersections between curves of every point using discrete accumulator matrix
  - If number of intersections greater than predefined threshold, line is detected with (Θ, r) at intersection point
- Probabilistic Hough Transform:
  - Uses randomly selected subset of all points in image for increased efficiency

Line Length Thresholding

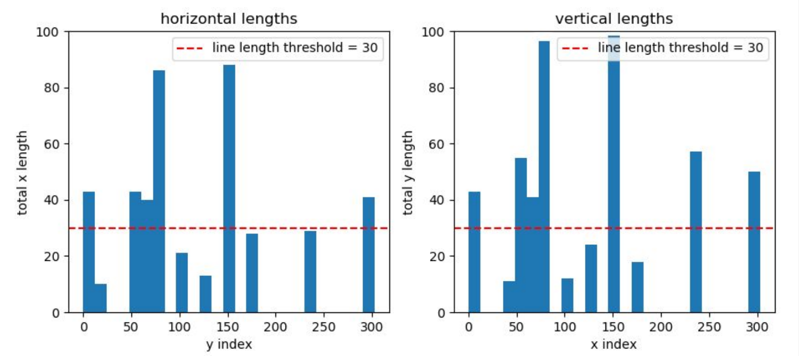
Step 7

Motivation

- Need method to reduce false positives among detected lines
- Along any horizontal or vertical loci, the larger the total length of detected lines, the higher the probability of flame existence

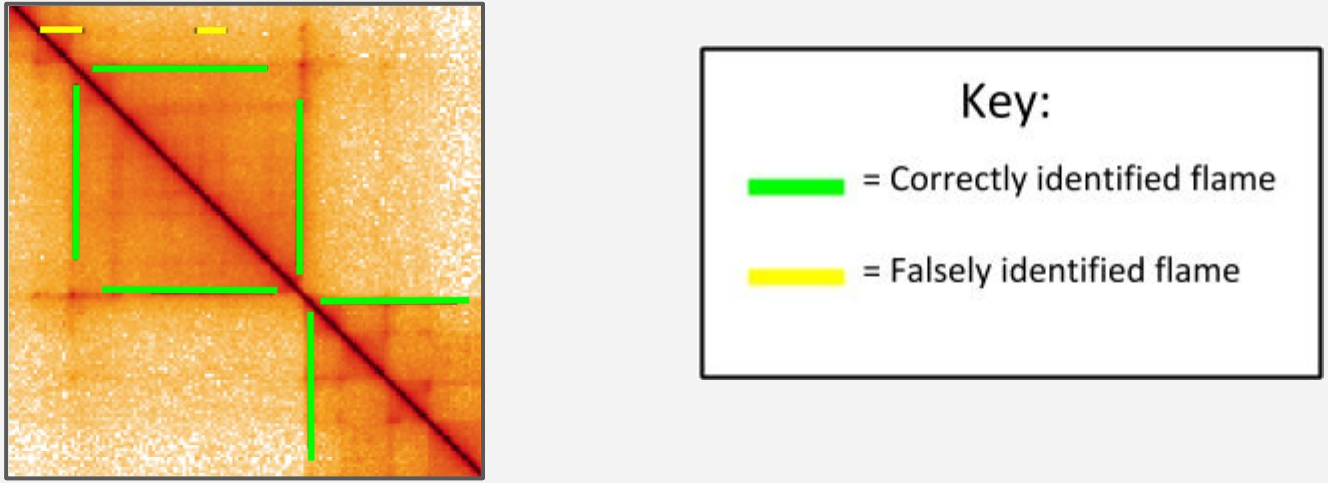
Mechanism

- Compute total detected line length along all horizontal and vertical loci
- Thresholded at length of 30, according to histogram
- Any horizontal or vertical loci with total detected line length greater than 30 considered a flame



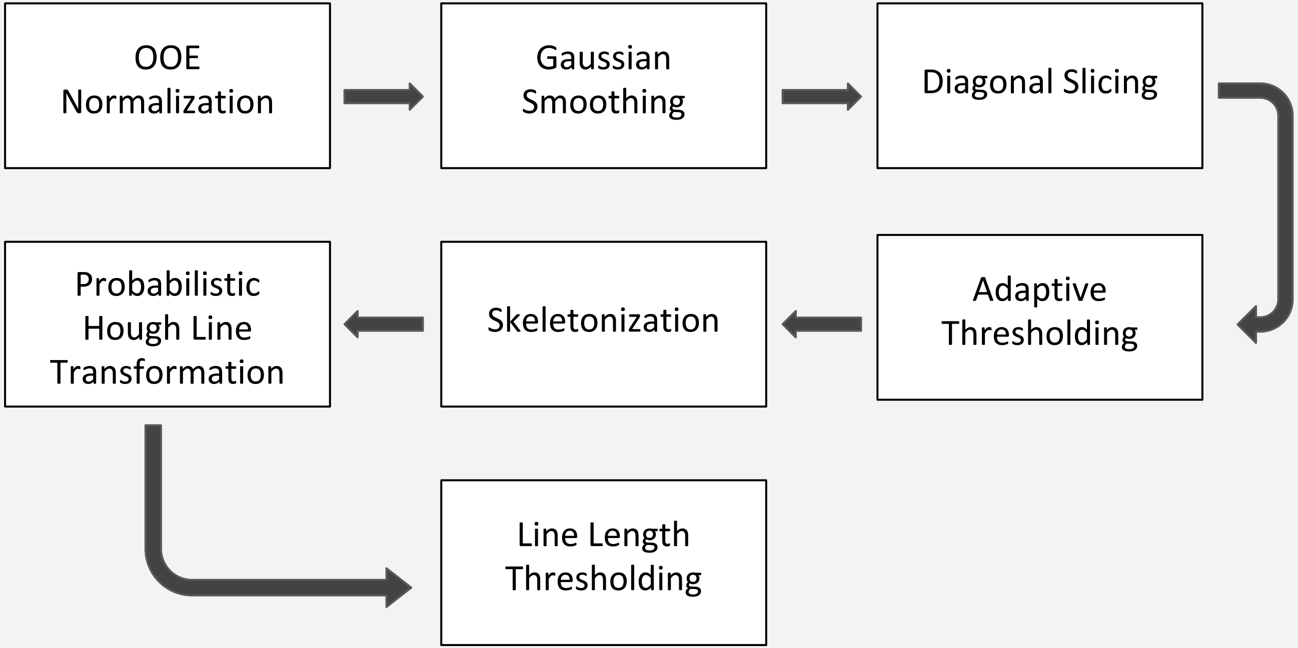
Genome-wide flame demarcation on HiGlass viewing platform

Sample taken from HiGlass:



Conclusion

We created a complete image processing pipeline to delineate flames within Hi-C maps:



Future Research

1. Reduce false positives
   - Further optimize parameters across entire pipeline

2. Biological implications of flames
   - Analysis of flames on a larger scale is now possible

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome.

2. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. (2017). Chromatin organization by an interplay of loop extrusion and compartmental segregation.

3. Gedraite, Estevao & Hadad, M. (2011). Investigation on the effect of a Gaussian Blur in image filtering and segmentation. 393-396.

4. Dey, Nilanjan & Dutta, Saurab & Dey, Goutami & Chakraborty, Sayan & Ray, Ruben & Roy, Payel. (2014). Adaptive thresholding: A comparative study. 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies.

5. T. Y. Zhang and C. Y. Suen. (1984) A fast parallel algorithm for thinning digital patterns. Communications of the ACM, March 1984, Volume 27, Number 3.

6. Kiryati, N., Eldar, Y., & Bruckstein, A. M. (1991). A probabilistic Hough transform. Pattern Recognition.

Acknowledgments

I would like to thank:

- My mentors, Sameer Abraham and Martin Falk
- The Mirny Lab
- The MIT-PRIMES program
- My Parents

Questions?