1 Purpose of this research

In this study, we plan to find dissimilarities between normal cells and cancerous cells, through investigating HiC contact maps. We suspect that there are systematic differences between how chromosomes are structured between normal cells and cancerous cells. we Ideally, it is desirable to compare 3D structures of cell in order to make such comparisons. However, the main challenge that we face is that 3D structure of a cell is not readily available. Based on [1], fluorescence in situ hybridizaiton (FISH) is used for investigating 3D configuration of chromosomes. However, this method can only be used locally and cannot map the whole structure of the chromosomes. In orther to find dissimilarities in the 3D structure of chromosomes, we used HiC dataset. The HiC method, which was developed by [2], captures interactions between chromosomal fragments in kilobase resolution. Based on HiC data, an *interaction frequency (IF)* matrix can be developed between *loci* at a desired resolution. A cell IF_{ij} in an interaction frequency matrix captures the number of interaction detected in HiC dataset between locus i and locus j in the genome. An interaction matrix can be used to develop both inter- and intra-chromosomal interaction matrices. We believe differences in interaction matrices can be found between normal cells and cancerous ones.

2 Genetics and Genomics

What is genetics? studies heredity. Offspring resembles its parents becasue they inherit their *genes*.

Gene: Sets of instructions that determine the traits of an organism. A sequence of DNA. A DNA thread builds chromosomes. Each organism has a unique number of chromosomes: Humans: 46 chromosomes. You get 23 from your mother and 23 from your father.

Homologous chromosomes: A set of two chromosomes consisting of genes that correspond to the same traits, one coming from mother and the other coming from father. These corresponding genes are called *alleles*. Alleles are either *dominant* or *recessive*.

Phenotype vs. Genotype: Phenotype is a visible physical trait that is caused partially by genes and partially environmental. These traits include eye of hair color, height etc. *genotype* is the gense that correspond to a particular phenotype. Their presence is necessary but not enough for the existence of phenotype.

Homozygous vs. Heterozygous alleles: There are 3 combinations of genotypes: tt, tT (or Tt) and TT. If the two alleles are the same then the person is Homozygous (dominant or recessive) in the corresponding trait. Otherwise, the person is Heterozygous in that trait.

Nucleotide: The monomer units that comprise DNAs. There are 4 types of nucleotides: (C, G, A, and T)

DNA: Macro-molecules that provide recipes for creating proteins.

Gene Expression: `https://en.wikipedia.org/wiki/Gene_expression`

Authors of [3] provide good insight into structure a nucleus of a eukaryotic cell:

> "The cell of a eukaryotic species forms a multi-granularity genome structure (e.g., nucleosome, chromatin fiber, chromatin cluster, chromosome, and genome) in order to compactly store a very long genomic DNA sequence in its small nucleus. A nucleosome is a basic unit consisting of 145-147 base pairs of DNA wrapped around a protein complex (histone octamer). Tens of nucleosomes are further collapsed into a larger dense structural unit chromatin fiber - of several kilobase (Kb) pairs. Multiple chromatin fibers form a large module of megabase pairs (Mb) DNA, which may be referred to as domains, globules, gene loci, or chromatin clusters in different contexts. A number of chromatin clusters then fold into a large independent physical structure - chromosome, which occupies a physical space in nucleus often being referred to as chromosome territory. One or more chromosomes interact to constitute the dynamic three-dimensional (3D) conformation of the entire genome of a cell."

The following is from this Youtube video:

Base: Each pair of nucleotides in the DNA are called a base. A kilo-base resolution is a resolution in which each cell in the matrix corresponds to 1000 pairs of nucleotides in DNA. If you unfold the DNA inside one of your cells, it would measure 2 meters end to end. How is it folded up withing a nucleus which is only 6 micorns wide?

Procedure:

1. Freeze the DNA in place.
2. Cut the genome in tiny pieces. Mark the ends using Biotin, and glue them together into diffused pieces of DNA. These diffused pieces is made up of two bits of the genome that are spatial neighbors.
3. Using DNA sequencing, the two parts of the diffused DNA are identified and a dataset is created where each cell corresponds to a pair.

Frequent types of folds:

1. **Loop:**

   When two parts of the genome that are far from each other sequentially are attached.

   The length is $200Kb$.

   Often seen that the gene that occures at one of the ends of the loop is turned on.

   A protein called CTCF is always present at the loops. Of the four possible directions that the two CTCF can take, in a loop, the two CTCF alwas point to one another (towards the loop).

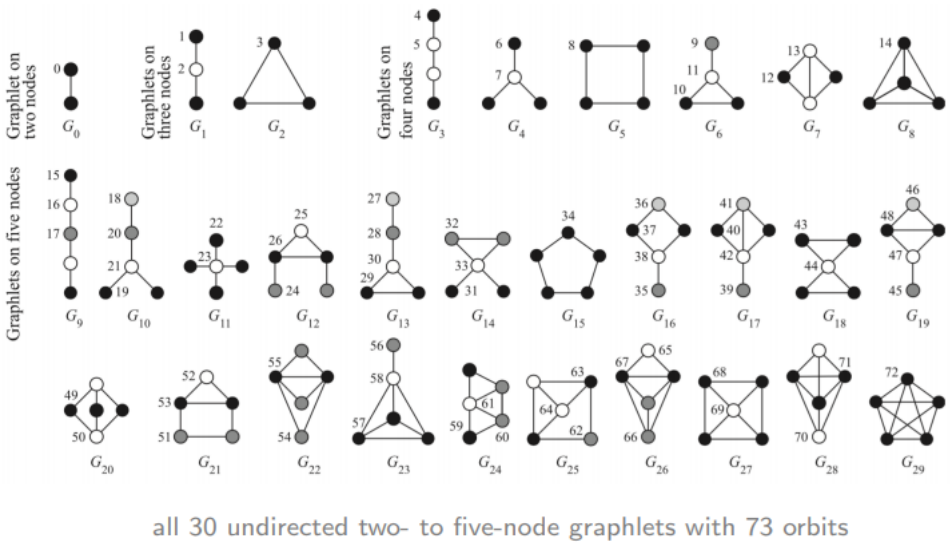


Figure 1: Graphlets and Orbits

2. **Contact domains:**

   Can happen in or out of a loop.

   Typical length: $200Kb$

   Domains with same *marks* tend to be located in the same place inside the nucleus.

3. **Nuclear sub-compartments:**

   Spatial neighborhoods where domains with the same *flavor* are present.

Many of the folding principles that were present in human cells were also present in mouse cells at corresponding genomic positions.

Speculation: Like genes, folds are preserved across species.

each cell in the body has the same gene but different cell types have different 3D structure (folding types).

3 Graph-theoretic concepts

Definitions: (from [4])

- **Fragment:** A connected subgraph.
- **Motifs:** Fragments that occur with a frequency much higher than that occuring in a randomly generated graph.
- **Graphlets:** An arbitrary, induced fragment. An edge is the only two-node graphlet.

- **Induced graphs:** Given a graph $G(V, E)$ and $S \subseteq V$, then $G'(S, E')$ is a graphlet iff $E' = \{(u, v)|u, v \in V$ and $(u, v) \in E \rightarrow (u, v) \in E'\}$
- **Orbits:** Set of all nodes in a graphlet that can be swapped with each other while not changing the graph.

Concepts

- **Global vs. local network comparison**:

  Global is inappropriate for incomplete networks.
- **Problem of Motifs:** They don't take into account infrequent and average subnetworks.
- **GDD: graphlet degree distribution**

4 Literature Review

4.1 Hi-C

Authors of [5] developed MOGEN.

4.2 Graphlets

Graphlet comparison is a novel method used to compare large networks in order to find local similarities in them. Authors of [6] provide a new measure of PPI network comparison based on 73 constraints. This is used in order to compare two large networks in order to detect similarities.

[7] provide heuristics to compare two nodes based on some feature (or signature) vectors, which is a 73-dimensional vector $\mathbf{s}^T = [s_0, s_2, ..., s_{72}]$ where s_i denotes the number of nodes in the network that are part of an orbit i.

Important Result: Proteins with similar surroundings perform similar functions.

In [8], the same author investigates cancer-causing genes to find similarities in their signatures. After clustering the genes based on *signature similarity* criteria, some clusters contain a lot of cancerous genes. They use 4 different clustering methods with varying parameters to cluster the proteins. They then predict the cancer-relatedness of a protein i using an enrichment criteria $\frac{k}{|C_i|}$ where C_i is the cluster where protein i belongs and k is the number of cancer-causing proteins in C_i and $|C_i|$ is the size of C_i.

Implementations of algorithms of extracting graphlets:

- GraphCrunch: `http://www0.cs.ucl.ac.uk/staff/natasa/graphcrunch2/usage.html`

- PGD: `http://nesreenahmed.com/graphlets/`
- ORCA: Graphlet and orbit counting algorithm

  `https://CRAN.R-project.org/package=orca`

  This package is in R. In order to install it, type `install.packages("orca")`.

The authors of [9] generalized the idea of graphlets to ordered graphs were the nodes are labeled in ascending order. These graphlets are illustrated in Figure 2. As can be viewed, there are a total of 14 orbits for graphlets of size 2 and 3 since the label of graphlets is also included in toplogy. In the new definition, d_v^i denotes the number of orbit i touches node v. Each node, is then assigned a vector of length 14 [1] $(d_v^1, d_v^2, ..., d_v^{14})$ and similarity of two nodes in two contact maps can be compared by how geometrically close their corresponding vectors are.

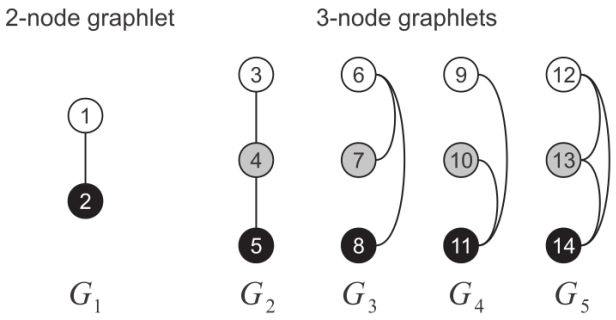


Figure 2: The five 2-and 3-node ordered graphlets and the corresponding 14 automorphism orbits. The ordering of the graphlet nodes within each graphlet $G_i, \quad i \in \{1, ..., 5\}$ is represented by their colors: white nodes $<$ gray nodes $<$ black nodes

4.3 Balanced Network Deconvolution and Residual Networks

Proposed by [10]. They developed a method that eliminated indirect effect from a weighted graph. Their method assumes that the observed graph G_{obs} is a summation of its direct graph G_{dir} and some indirect terms as follows:

$$G_{obs} = G_{dir} + G_{dir}^2 + G_{dir} + ... \tag{1}$$

They assume also that both G_{obs} and G_{dir} can be eigen-decomposed and they have the same eigen-vectors, that is

$$G_{obs} = X\Sigma_{obs}X^T \tag{2}$$

[1] number of orbits in graphlets of size 2 and 3

where

$$\Sigma_{obs} = \begin{pmatrix} \lambda_1^{obs} & 0 & \cdots & 0 \\ 0 & \lambda_2^{obs} & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n^{obs} \end{pmatrix} \tag{3}$$

and

$$G_{dir} = X\Sigma_{dir}X^T \tag{4}$$

where

$$\Sigma_{dir} = \begin{pmatrix} \lambda_1^{dir} & 0 & \cdots & 0 \\ 0 & \lambda_2^{dir} & & \\ \vdots & & \ddots & \\ 0 & & & \lambda_n^{dir} \end{pmatrix} \tag{5}$$

so

$$G_{obs} = X\Sigma_{obs}X^T = X(\Sigma_{dir} + \Sigma_{dir}^2 + ...)X^T \tag{6}$$

They also assume that eigen-values of the direct network are all between -1 and 1, i.e.

$$-1 < \lambda_i^{dir} < 1 \quad \forall 1 \leq i \leq n \tag{7}$$

Thus

$$\Sigma_{obs} = \Sigma_{dir} + \Sigma_{dir}^2 + ... \tag{8}$$

$$\lambda_i^{obs} = \sum_{j=1}^{\infty} \lambda_{ij}^{dir} \qquad \forall i = 1...n \tag{9}$$

$$\lambda_i^{obs} = \frac{\lambda_i^{dir}}{1 - \lambda_i^{dir}} \tag{10}$$

$$\lambda_i^{dir} = \frac{\lambda_i^{obs}}{1 + \lambda_i^{obs}} \tag{11}$$

Source codes:

- Python: `https://github.com/gidonro/Network-Deconvolution`
- Matlab: `http://compbio.mit.edu/nd/`

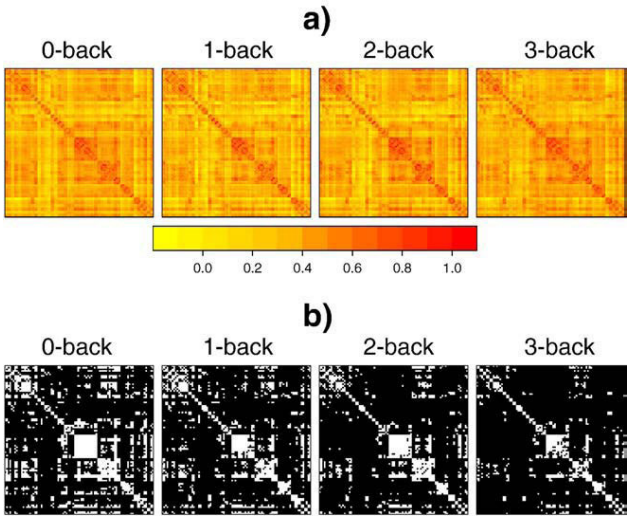


Figure 3

Authors of [11] proposed ...

4.4 Weighted graph comparison

`https://www.cs.cmu.edu/~jingx/docs/DBreport.pdf`

`https://en.wikipedia.org/wiki/Gromov%E2%80%93Hausdorff_convergence`

Authors of [12] propose a graphlet decomposition method which considers situations where we observe an undirected weighted network, encoded by a symmetric adjacency matrix with integer entries and diagonal elements equal to zero. The term graphlet here is UNRELATED to the Przulj graphlet.

Authors of [13] developed Statistical Network (SPN) analysis where the choice of thresholding value is made by statistical inference. This method works within the framework of design of experiments where the same network can be extracted for different individuals under different treatments. The effect of those treatments can then be infered using this method. In [13] for example, they studied neuron connectivity networks among 43 subjects for 4 different memory tasks (0-back, 1-back, 2-back and 3-back), and obtained the following thresholded networks. The results of which can be viewed in Figure 3.

Their method can be very usefull if a large population of information from the same chromosomes exist.

Distance between two weighted graphs can be measured based on [14]:

$$d(G_1, G_2) = |\sum_{w \in V_1} w - \sum_{v \in V_2} v|$$

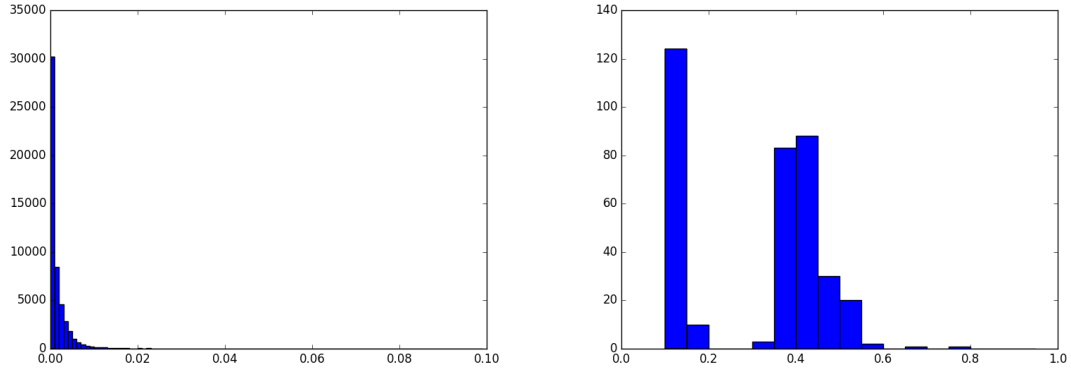
Exhaustive enumeration of all graphlets being prohibitively expensive, [15] introduce two theoretically grounded speedup schemes, one based on sampling and the second one specifically designed for bounded degree graphs

5 Problems and Questions

1. Can we convert larger Hi-C data sets by combining data from several loci? This way we can create a DOE of different individuals.
2. How can we verify our results? Do we have any information on similarities that we already know exist?
3. how do we get inter-chromosomal graphs? I was not able to find them.
4. Is there indirect depencecies in the Hi-graphs? How are they caused?
5. How to run the algorithms on such huge matrices?
6. What are random graph models?
7. As far as I get, graphlet distribution is a means of finding similarities in graphs. What we are trying to do is to find dissimilarities. is it the right track to go?

6 Experiments and Observations

I have changed the network deconvolution code where it deals with scaling factor. The rationale behind it was not clear to me so I replaced it with my own understanding of it which might not be correct.
The intensity values in each data set follows an exponential distribution, that is the majority of the intensities lie close to the minimum value while some intensities are absurdly large. Because of this, thresholding intensity values based on quantiles (percentiles) cannot be helpful since first and third quartiles are very close to each other but very far from the maximum value.



(a) **(b)**

Figure 4: 4a is a histogram of intensities form "`chr1_1mb_matrix`", in the interval of $[0, 0.1]$ and bin size of 0.005. 4b is a histogram of the same data but for interval of $[0.1, 1]$ and bin size of 0.01. As can be seen, the two histograms are orders of magnitude different in terms of frequency.

Each row and column in the graphs represents a loci on the chromosome, the next row (column) represents the adjacent loci on the same chromosome, thus the graph cannot be scrambled, that is, it is constructed as a stack of interaction intensities of a locus with all other loci on the chromosome. You cannot just swap two rows in the graph and have the same information since the order is important. The only valid case is when last rows are removed and stacked on top of the graph.

Taking a logarithm can be very useful in this case. As can be viewed in Figures 5 and 6, you can see that the data for all chromosomes look normal.

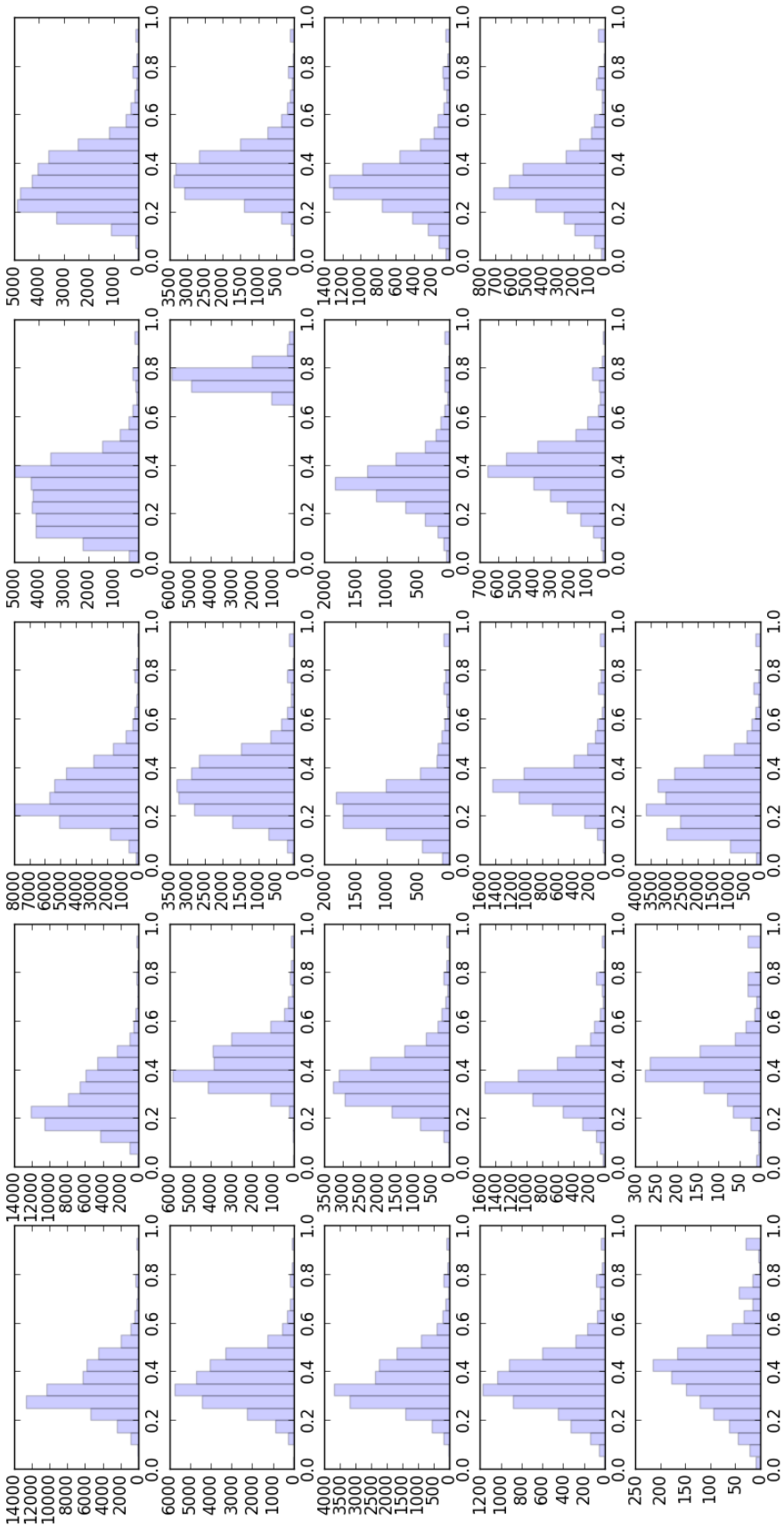


Figure 5: The histogram of the logarithm of the data for all 23 chromosomes. The data looks normal now.

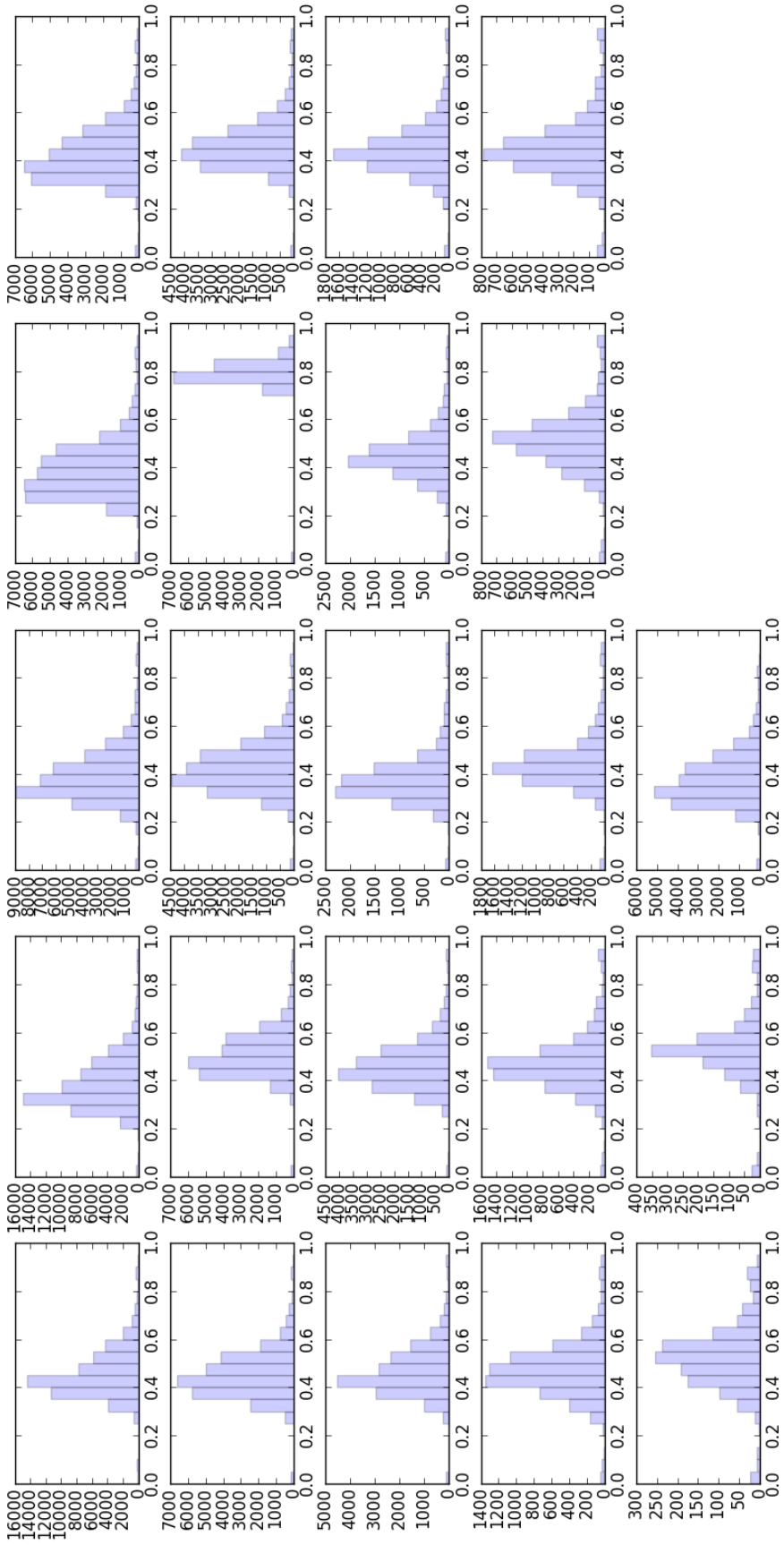
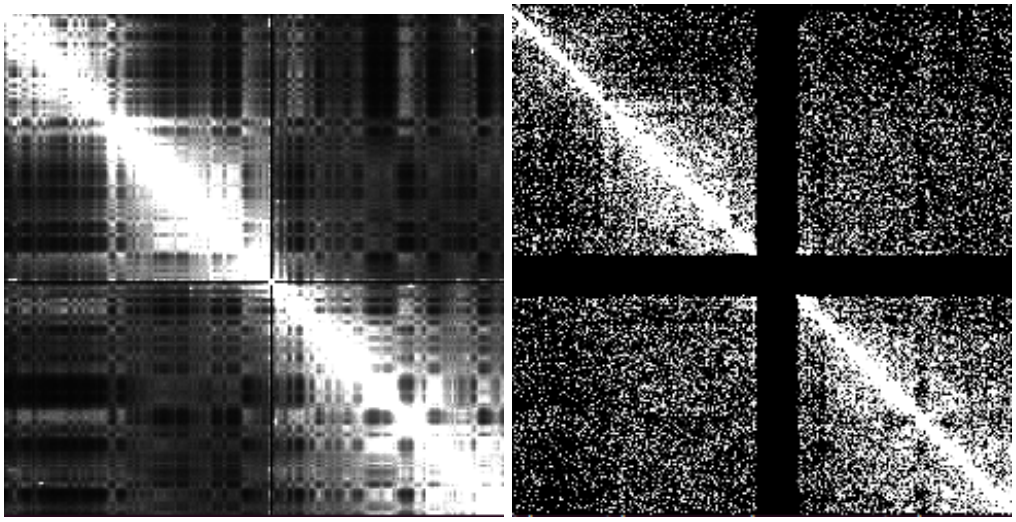


Figure 6: The histogram of the deconvoluted data in Figure 5.

We decided to lock our data set on this dataset, since it includes Hi-C data for both normal test cells and 3 various cancerous cells. This allows for the sort of comparison we are looking for in this study.

Cleaning the data requires some step that I'm not familiar yet. For example a normalized Hi-C heatmap for chromosome 1 should look like something like 7a, which can be found here, but what I have managed to get from raw data is 7b, which clearly is not normalized.



(a) **(b)**

Figure 7: As can be seen, 7b is not as smooth as 7a. The black area in 7b account for 19 rows and columns, which is exactly the difference between the number of columns in 7b and 7a.

Based on [3], two normalization methods can be implemented on a contact matrix. The first of is Sequential Component Normalization (SCN [16]) and the other is simple Pearson's correlation.

In SCN, each column of the contact matrix is divied by its norm, then its rows are divided by theirs norms. This process continues until the contact matrix is symmetric again.

In Pearson's correlation, a matrix C is produced where C_{ij} is the pearson correlation of rows i and j.

I have come up with what I call the *pyramid algorithm* where I simply apply a pyramid down and then pyramic up on contact matrices to fill the empty spaces between. This algorithm is readily available in OpenCV [2].

The results of the normalizations on chrmosome 1 of a normal cell is presented in Figure 8.

[2] www.opencv.org

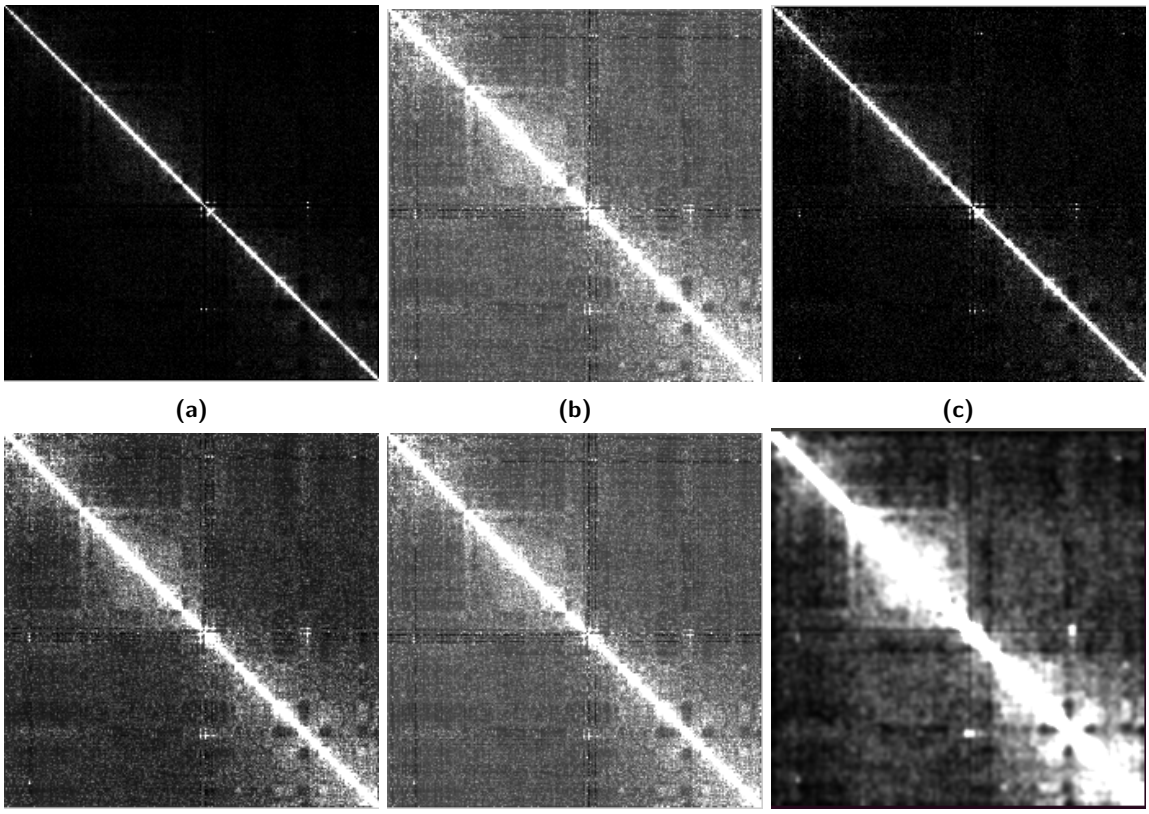


Figure 8: Results of different normalizing methods: 8a is the original matrix of 7b with the 19 empty rows and columns removed. 8b demonstrates the result of applying pearson normalization to the 8a. In the same manner 8c demonstrates the result of applying scn to 8a. 8d shows the result of applying pearson to 8c and 8e illustrates result of applying scn to 8b. 8f is the result of applying my pyramid method to the image which makes it somehow smooth and as I will describe later, will significantly improve histogram of logarithm matrix.

Figure 9 illustrates the result of histograms of logarithms of figures that result from pearson normalization or chromosomes 1 through 22 and chromosome X (the last one).

Figure 10 illustrates the result of histograms of logarithms of figures that result from applying pyramid normalization on the contact matrices of chromosomes 1 through 22 and chromosome X (the last one).

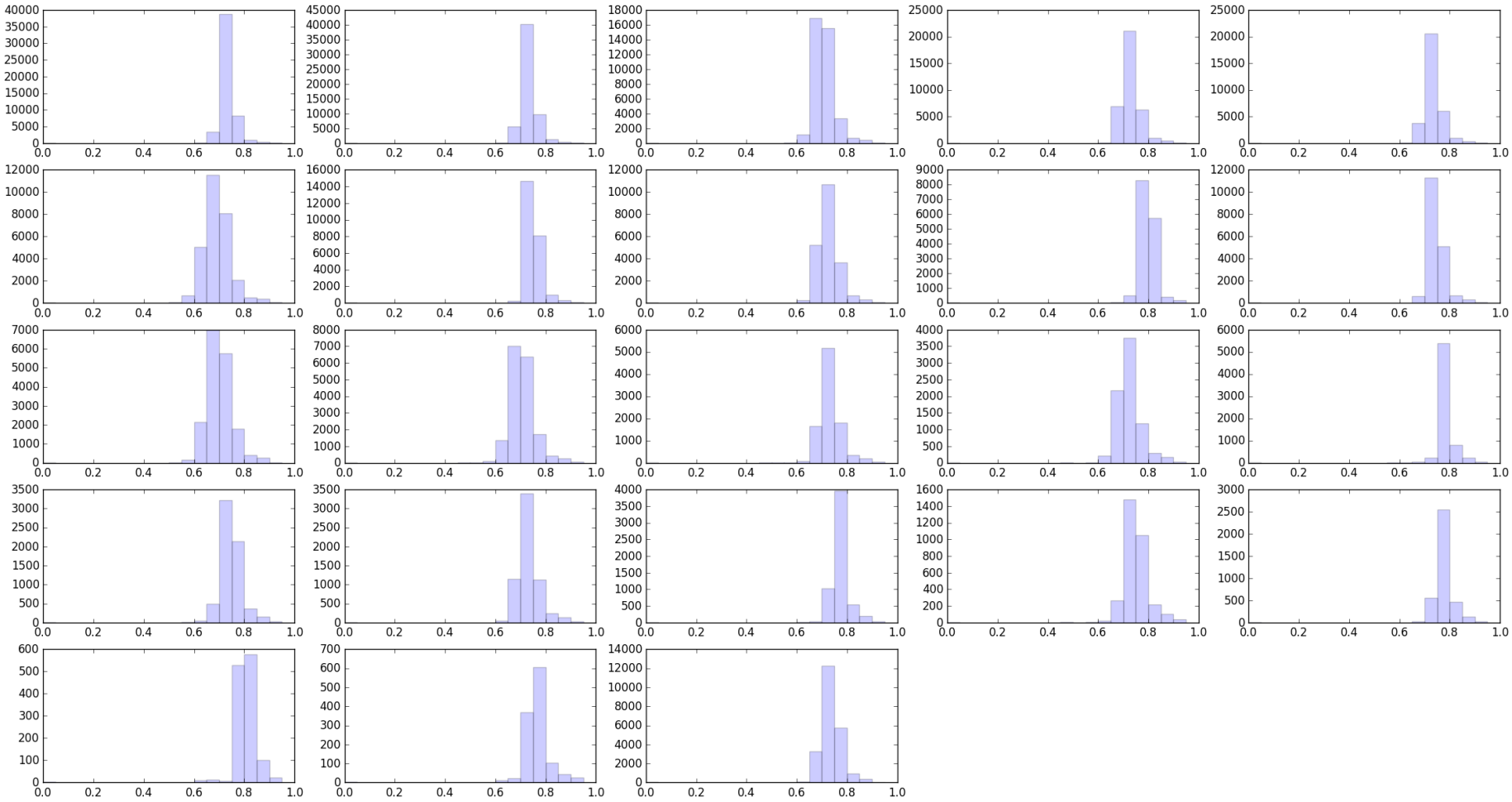


Figure 9

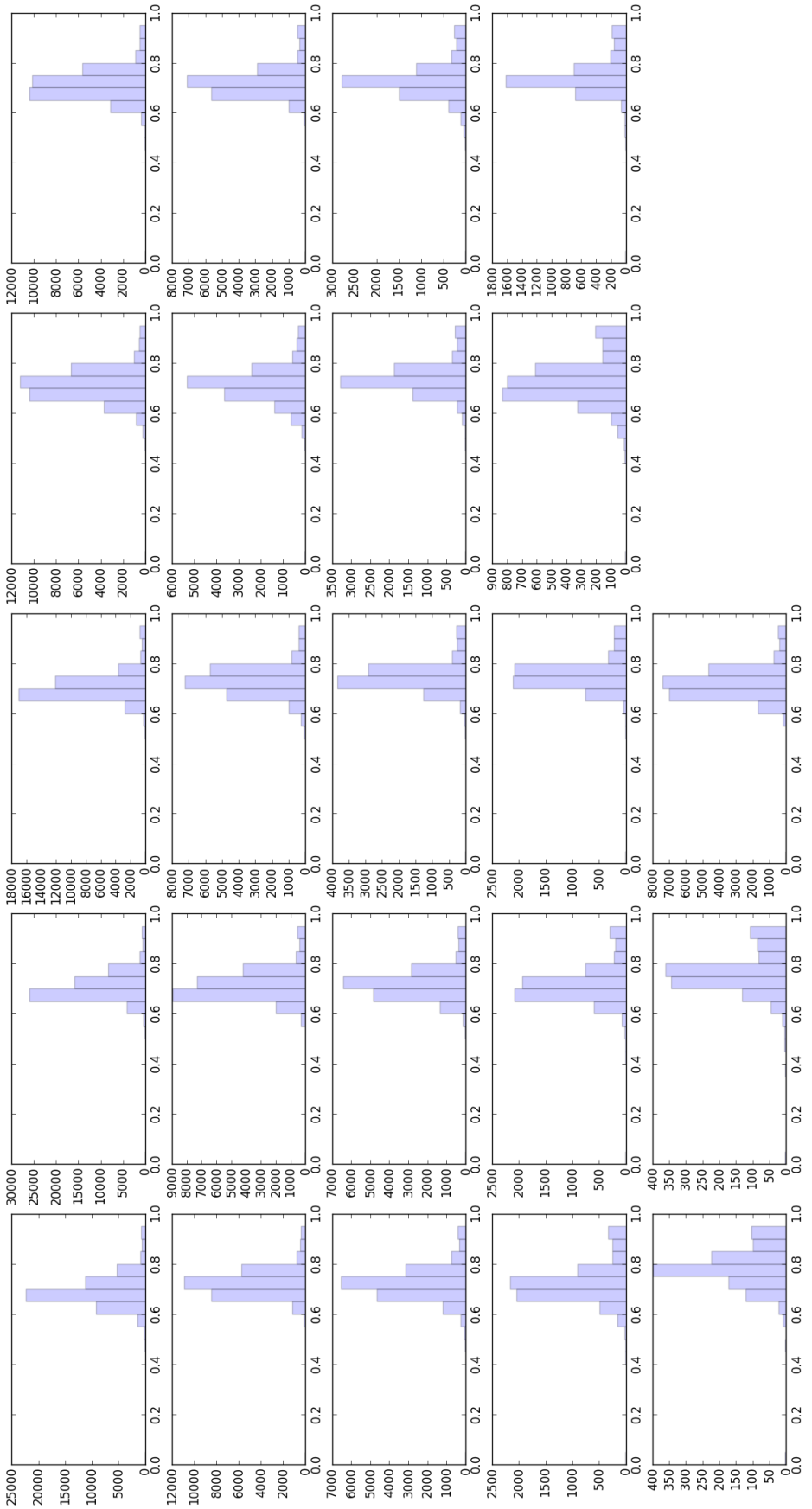


Figure 10

ALL and MIT cell comparison

One idea is the vary the threshold limits on the two indices to find the optimal set of thresholds (t_{ALL}, t_{MIT}), where the difference between the two pixels are minimum. The remaining pixels that are still different are the ones that are definitely correct. In this case, the objective function should be properly penalized in order to prevent the thresholds from getting too high or too low.

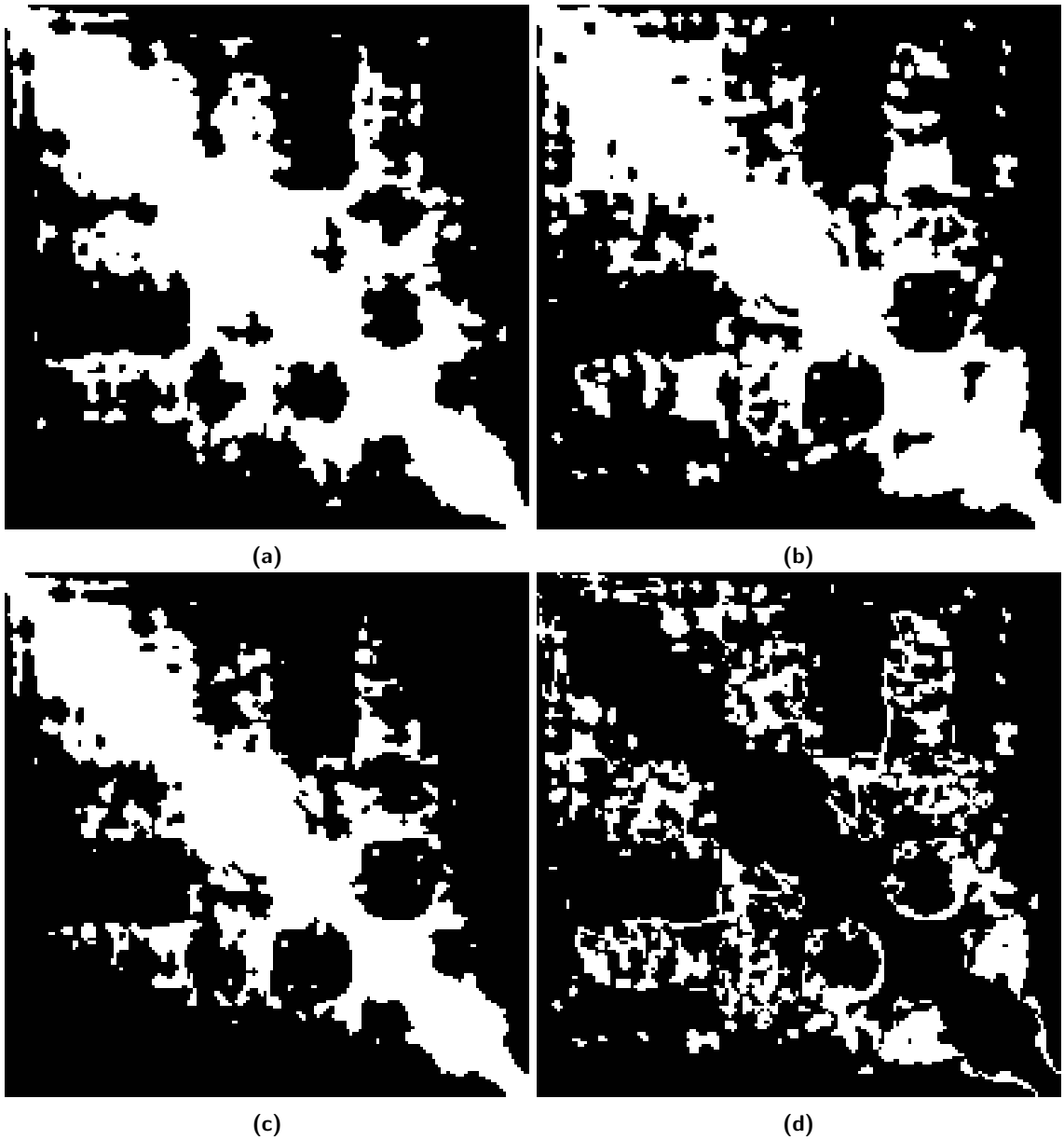


Figure 11: 11a is HiC contact matrix of chromosome 14 of ALL cell and 11b is the contact matrix of the same chromosome in MIT cells. Both matrices are thresholded based on their 60th percentile. 11c is the common pixels and 11d is the pixels where they are different.

7 Resources

Publications related to Hi-C:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2858594/
2. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0058793
3. http://nar.oxfordjournals.org/content/42/7/e52.full
4. http://bioinformatics.oxfordjournals.org/content/early/2015/12/31/bioinformatics.btv754.abstract?keytype=ref&ijkey=A97WhKqBiEIcuzd
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4417147/
6. http://www.pnas.org/content/112/47/E6456.full
7. http://www.pnas.org/content/113/12/E1663.full

Hi-C Datasets:

1. Original Datasets: https://bcm.app.box.com/v/aidenlab/folder/11234760671
2. Including cancerous cells: http://sysbio.rnet.missouri.edu/T0510/tmp_download/link_to_download_genome_data/
3. Chromosome3D project: http://sysbio.rnet.missouri.edu/bdm_download/chromosome3d/

Contact Matrix Analysis:

1. https://omictools.com/contact-matrix-normalization-category
2. http://hifive.docs.taylorlab.org/en/latest/

Labs working on 3D Human Genome:

1. http://mirnylab.mit.edu
2. http://dostielab.biochem.mcgill.ca
3. http://www.aidenlab.org/
4. http://web.cmb.usc.edu/people/alber/index.htm
5. http://calla.rnet.missouri.edu/cheng/nsf_career.html

Resources related to Graphlet:

1. https://en.wikipedia.org/wiki/Graphlets
2. https://academic.oup.com/bioinformatics/article/23/2/e177/202080/Biological-network-comparison-usi
3. http://www0.cs.ucl.ac.uk/staff/N.Przulj/index.html

References

[1] Badri Adhikari, Tuan Trieu, and Jianlin Cheng. Chromosome3d: reconstructing three-dimensional chromosomal structures from hi-c interaction frequency data using distance geometry simulated annealing. *BMC genomics*, 17(1):886, 2016.

[2] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289–293, 2009.

[3] Zheng Wang, Renzhi Cao, Kristen Taylor, Aaron Briley, Charles Caldwell, and Jianlin Cheng. The properties of genome conformation and spatial gene interaction and regulation networks of normal and malignant human cell types. *PloS one*, 8(3):e58793, 2013.

[4] J. Gross. Automorphisms. http://www.cs.columbia.edu/ cs6204/files/Lec5-Automorphisms.pdf, April 2010.

[5] Tuan Trieu and Jianlin Cheng. Mogen: a tool for reconstructing 3d models of genomes from chromosomal conformation capturing data. *Bioinformatics*, 32(9):1286–1292, 2015.

[6] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[7] Tijana Milenkoviæ and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257, 2008.

[8] Tijana Milenković, Vesna Memišević, Anand K Ganesan, and Nataša Pržulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of the Royal Society Interface*, 7(44):423–437, 2010.

[9] Pietro Di Lena, Piero Fariselli, Luciano Margara, Marco Vassura, and Rita Casadio. Fast overlapping of protein contact maps by alignment of eigenvectors. *Bioinformatics*, 26(18):2250–2258, 2010.

[10] Soheil Feizi, Daniel Marbach, Muriel Médard, and Manolis Kellis. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology*, 31(8):726–733, 2013.

[11] Hai-Ping Sun, Yan Huang, Xiao-Fan Wang, Yang Zhang, and Hong-Bin Shen. Improving accuracy of protein contact prediction using balanced network deconvolution. *Proteins: Structure, Function, and Bioinformatics*, 83(3):485–496, 2015.

[12] Hossein Azari Soufiani and Edo Airoldi. Graphlet decomposition of a weighted network. In *Artificial Intelligence and Statistics*, pages 54–63, 2012.

[13] Cedric E Ginestet and Andrew Simmons. Statistical parametric network analysis of functional connectivity dynamics during a working memory task. *Neuroimage*, 55(2):688–704, 2011.

[14] Hamed Daneshpajouh, Hamid Reza Daneshpajouh, and Farzad Didehvar. A metric on the space of weighted graphs. *arXiv preprint arXiv:0906.2558*, 2009.

[15] Nino Shervashidze, SVN Vishwanathan, Tobias Petri, Kurt Mehlhorn, and Karsten Borgwardt. Efficient graphlet kernels for large graph comparison. In *Artificial Intelligence and Statistics*, pages 488–495, 2009.

[16] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):436, 2012.