Published online 18 March 2017

Nucleic Acids Research, 2017, Vol. 37, No. 12 ***1–6***
doi:10.1093/nar/gkn000

HiC Contact Map Comaprison Using Graphlet Approach

Behnam Rasoolian [1,*], **Debswapna Bhattacharya** [1 *]

[1] Auburn University

ABSTRACT

In this study, we investigated dissimilarities between normal cells and leukemic cells in terms of their three dimensional conformation. We first thresholed HiC data corresponding to one normal cell line and four leukemic cells lines. We then used the thresholded data to extract the first 72 orbits for all loci in each contact map. We then measured pairwise graphlet distances as well as pairwise graphlet distribution correlations for each pair of cells and compared them using statistical methods. Our results show that normal-cancer pairs have significantly higher dinstances from each other than cancer-cancer pairs. Also, we concluded that for certain orbits, cancer-cancer pairs demonstrate higher correlations than normal-cancer pairs in terms of orbit frequency distributions.

INTRODUCTION

This reasearch is an effort to compare the three-dimensional structure of cell genomes. To the best of our knowledge little research has dealt with quantitative comparison of spatial conformation of chromosomes. Ideally, it is desirable to compare 3D structures of cell in order to make such comparisons. However, the main challenge in this regard is that 3D structure of a cell is not readily available. In order to find dissimilarities in the 3D structure of chromosomes, we used HiC dataset. The HiC method, which was developed by (1), captures interactions between chromosomal fragments in kilobase resolution. Based on HiC data, a *contact map* or *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 have found statistically significant differences between interaction matrices of normal cells and cancerous ones.

Recently, HiC data has been mostyly used in efforts to predict the 3D conformation of chromosomes using HiC data. (2, 3, 4, 5, 6, 7, 8, 9, 10). These efforts, although different in approach, usually translate interaction frequencies in contact maps as an inverse measure of distance. For example, (2) developed a method, Chromosome3D which uses distance geometry simulated annealing (DGSA) in order to reconstruct genome structure. These reconstructed 3D structures could be used in order to compare normal and malignant cell lines. However, the main obstacle to doing so is that the process of reconstruction is computationally expensive. Almost all reconstruction methods rely on either deterministic or heuristic optimization approaches in order to find the best confomation that matches contact maps. Also, there is no real way of verifying whether the reconstruction results are accurate. The only method of verifying such results is Fluorescence in situ hybridizaiton (FISH) which can capture the actual 3D configuration of chromosomes. (11, 12, 13) However, this method can only be used locally and cannot map the whole structure of the chromosomes.

Recently, graphlet comparison has emerged as a novel method for comparing large networks in order to find local similarities in them. Authors of (14) provide a new measure of protein-protein interaction (PPI) network comparison based on 73 constraints. The method is used in order to compare two large PPI networks in order to detect similarities.

In (15) the authors provide heuristics to compare two nodes in a graph based on signature vectors, which are 73-dimensional vectors $\mathbf{s}^T = [s_0, s_2, ..., s_{72}]$ where s_i denotes the number of nodes in the network that are part of an orbit i. They concluded that proteins with similar surroundings perform similar functions.

In (16), 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.

In this paper, we applied the same approach of graphlet distance and graphlet distribution comparison on HiC data in order investigate genome spatial confomations differences between normal and leukemic cells.

Notations In this paper, matrices and vectors are represented using bold capital and bold small letters respectively. Matrix

*Tel: +1 334 5212814; Email: bzr0014@auburn.edu

rows and columns are represented by a *dot* notation. For example, the ith row of matrix M is denoted by $M_{i.}$ and its jth column is represented by $M_{.j}$.

We denote the set of all contact maps in cell line K with $\mathbb{C}^K$. If no particular cell line is addressed, the subscripts are dropped. Any arbitrary member of $\mathbb{C}$ is denoted by $\mathbf{C}_{ij}$, where i and j $(j \geq i)$ represent the two chromosomes involved. In human cells this set contains a total of 276 contact maps, 23 of which are intra-chromosomal and the rest are inter-chromosomal. For ease of representations, intra-chromosomal contact maps are distinguished by a single superscript, so we have $\mathbf{C}_{i,i} = \mathbf{C}_i$.

We denote the number of loci in a chromosome i by N_i. The set of all loci involved in contact map C_{ij} is denoted by $\mathbb{V}_{ij}$. In intra-chromosomal contact maps, $\mathbb{V}_{i,i}$ containts only the loci of that particular chromosome ($|\mathbb{V}_i| = N_i$), while in inter-chromosomal contact maps $\mathbb{V}_{ij}$ contains the loci in the both of chromosomes involved ($\mathbb{V}_{ij} = \mathbb{V}_i \cup \mathbb{V}_j$).

MATERIALS AND METHODS

We re-used Leukemic Hi-C libraries created in (18) These libraries were sequenced for cases of primary human B-acute lymphoblastic leukemia (B-ALL or ALL), the MHH-CALL-4 B-ALL cell line (CALL4), and the follicular lymphoma cell-line (RL). As in (18), We used normal B-cell line (GM068990) from (1) for our comparisons. We created contact maps of resolution 500kb and normalized it using the `iced` package in python developed by (19). We also performed and extra normalization by performing network deconvolution (20) in order to remove the effect of indirect interactions.

Thresholding contact maps

In order to be able to extract graphlets, HiC contact maps should be modeled as unweighted graphs where the nodes represent the loci and an edge between two nodes represent a *significant* interaction between the loci. This can be achieved by thresholding the contact maps. The result of the thresholding procedure is a binary matrix which also can serve as an adjacency matrix for an unweighted, undirected graph. The graph can then be used for orbit extraction.

When thresholding contact maps, it is necessary to make sure that both global and local features are maintained. We could consider thresholding the contact maps by simply setting values above a fixed value to one and the rest to zero; However, in practice, this method resulted in graphs that capture the local structure of the contact maps poorly. This is because intensities follow an exponential distribution with a mean close to zero with a few very larges values that correspond to interactions along or close to the main diagonal of the contact maps. Thus, picking relatively large numbers would result in ignoring interactions that are far from the main diagonal while picking small values will lead to capturing too many (insignificant) interactions.

To the best of our knowledge, little work has dealt with the task of thresholding HiC contact maps. There has been some statistical approaches developed on similar data in other fields. For example, authors of (21) developed Statistical Network (SPN) analysis where the choice of thresholding value is made by statistical inference. This method, although very robust, works within the framework of design of experiments where the same network can be extracted for different individuals under different treatments. Thus a relatively large set of different contact maps need to be available in order for this method to be applicable towards our end.

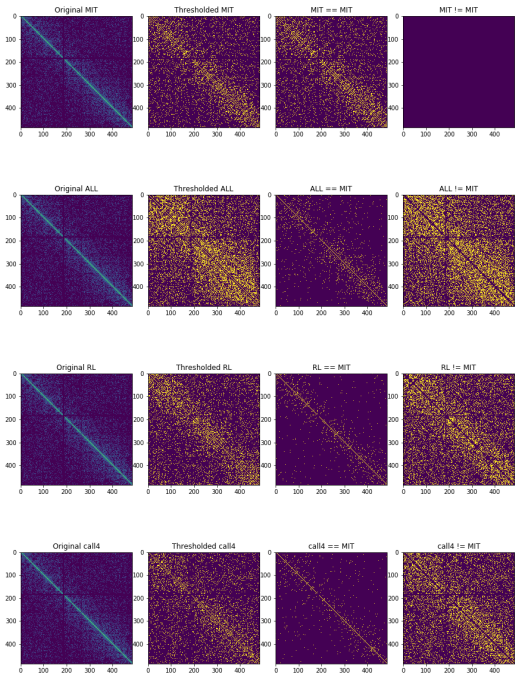


Figure 1. Result of thresholding interchromosomal contact map of chromosome 1 using a kernels of size 5×5 for all cell lines. The first row shows the thresholded maps. Second and third rows demonstrate pair-wise similarities and differences between contact maps respectively.

Instead, in order to threshold the matrix so that both global and local patterns are captured, we borrowed the concept of *adaptive thresholding* from image processing context. In this method, in order to be set, a pixel should have an intensity larger than the average of non-zero intensities in its *neighborhood*. The neighborhood is defined by a sliding kernel that passes through the contact map with the pixel at its middle at each step. Figure 1 demonstrates result of this thresholding approach for intra-chromosomal contact maps of chromosome 1. Refer to supplementary material for all 23 interchromosomal thresholding results.

Orbit Extraction

Once the thresholded contact maps are obtained, graphlets and orbits can be extracted. We used the `orca` package in `R` programming language to extract the graphlets. As a result of graphlet extraction, For each loci in each contact map, a *signature vector* of size 73 is created. Thus for each cell line, we would have 276 *signature matrices* of size $|V^{ij}| \times 73$, where V^{ij} is the number of loci involved in contact map

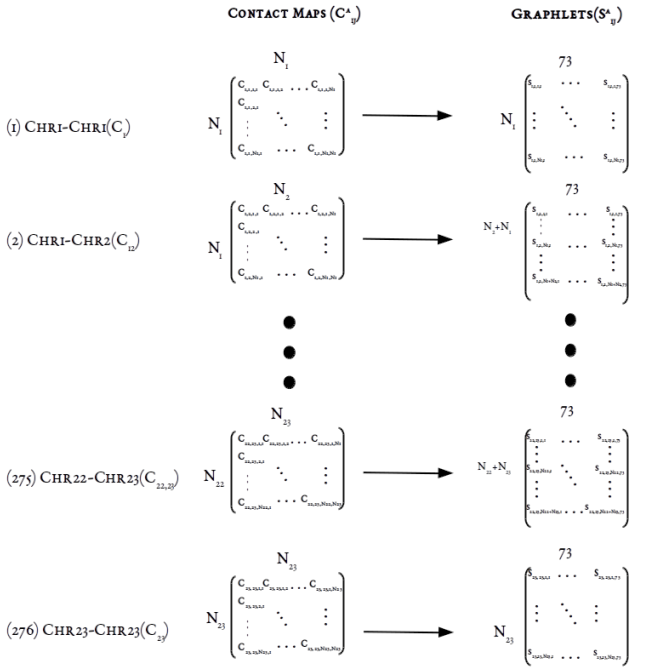


Figure 2. Graphlet extraction for the four cell lines. For each loci in each contact map between chromosomes i and j, the signature vectors of length 73 are extracted, resulting in a *signature matrix* of size $|V^{ij}| \times 73$, where V^{ij} is the number of loci involved.

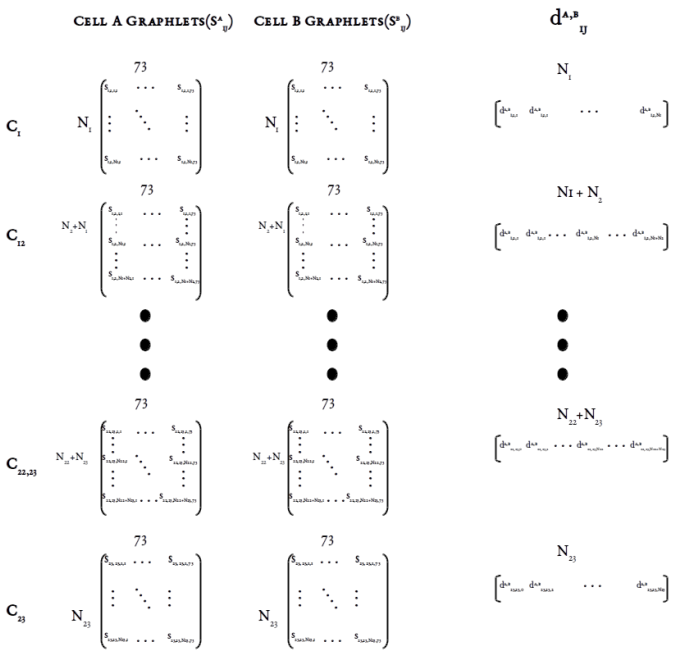


Figure 3. Calculating pair-wise loci distances. For each loci (row) in each contact map in MIT cell line, its distance is calculated based on equation **1** with the corresponding loci in leukemic cells. The result of this process is a *signature distance vector* of size $|V^{ij}| = N_i + N_j$ for each contact map.

between chromosomes i and j. Figure 2 illustrates the process and results of signature matrix extraction schematically.

For a particular $\mathbf{C}_{ij}$, we denote $\mathbf{S}_{ij}$ as its *signature matrix*. Each cell S_{ijlo} in $\mathbf{S}_{ij}$ captures how many times loci l in $\mathbf{C}_{ij}$ occured as part of orbit o.

We consider two measures of *difference* when comparing contact map graphlets across cell lines. The first measure is *signature distance vectors* between each contact map of two cell lines. For a pair of cells A and B, let $\mathbf{S}^A_{ij}$ and $\mathbf{S}^B_{ij}$ be their signature matrices. The *signature distance* of contact map $\mathbf{C}_{i,j}$ between A and B is denoted by $\mathbf{d}^{A,B}_{ij}$. $\mathbf{d}^{A,B}_{ij}$ is a vector of size $|V_{i,j}|$ and its elements $d^{A,B}_{i,j,l}$ are calculated using the following formula from (14):

$$d^{A,B}_{i,j,l} = \frac{1}{73}\sqrt{\sum_{o=0}^{72} t^2_{lo}} \qquad (\mathbf{1})$$

where elements of $t_{i,j,l,o}$ is the distance between each loci (row) l in $\mathbf{S}^A$ and the the same loci in $\mathbf{S}^B$ for orbit o as is calculated as below:

$$t_{lo} = w_o \times \frac{log(S^A_{ijlo}+1) - log(S^B_{ijlo}+1)}{log(max(S^A_{ijlo}, S^B_{ijlo})+2)} \qquad (\mathbf{2})$$

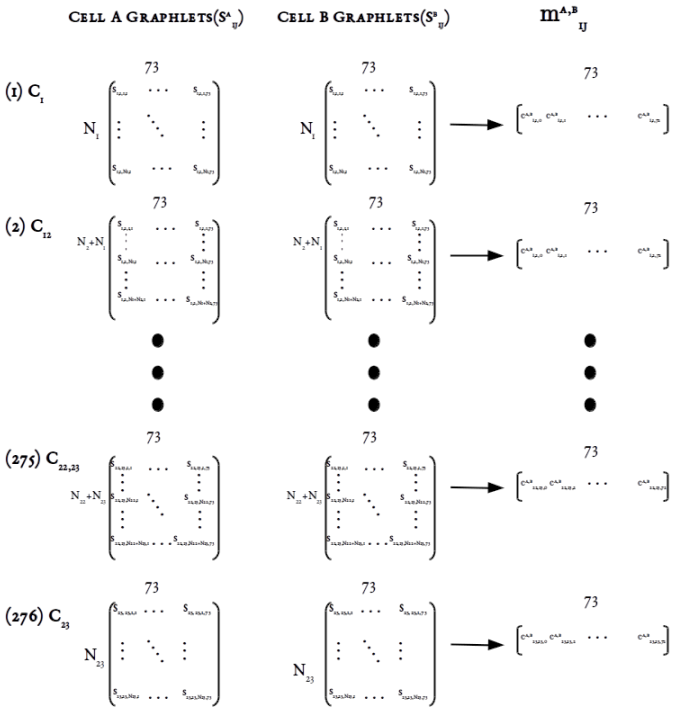


Figure 4. Calulating pair-wise orbit correlations. For each orbit (column) in each contact map in MIT cell line, its correlation with the same orbit in the same contact map in leukemic cells is calculated. The result of this process is a *signature correlation* vector of size 73 which captures how similar frequencies of two orbits are.

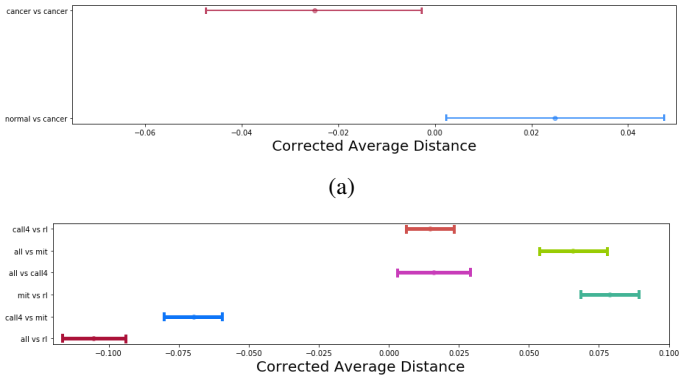


Figure 5. Average Pair-wise graphlet signature difference over all 276 contact maps: Each point on a graph is the result of averaging all the distances across all loci of all contact maps for each pair of cells. A 0.1 standard deviation error bar is also plotted for each point. ($\bar{\mathbf{d}}^{A,B}$).

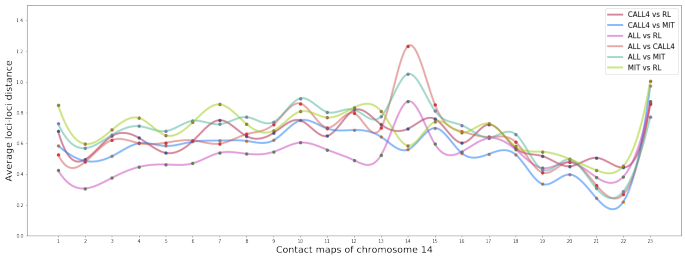


Figure 6. Pair-wise graphlet signature distances for all contact maps of chromosome 14. Warm colors as used for cancer-cancer pairs and cold colors are used for normal-cancer pairs. As can be seen cancer-cancer pairs tend to be close to each other than to the normal cell. This is specially true for the first 13 inter-chromosomal contact maps.

This process is illustrated in Figure 3. Using this distance measure, we can quantify how two loci are close to each other in terms of local neighborhood between the two contact maps.

The second measure of comparison that we use captures how similar two orbits are in terms of their count frequencies across loci between two contact maps. Each column in S_{ij} can provide information regarding the *frequency distribution* of orbits throughout the contact map C_{ij}. We can find how similar these distributions are to each other using correlation measures. These correlations are denoted by $\mathbf{m}_{i,j}^{A,B}$ and can be calculate using any plausible correlation measure. In this study, for each contact map, we calculated similarity between orbit distributions using Pearson's r correlation, which is computationally efficient. However, pearson's r might not be able to capture non-functional relationships between distributions. As a result, we also used Maximal Information Coefficient (MIC) (22) in order to compare correlations. MIC calculates mutual information (MI) between two distributions, but utilizes dynamic programming in order adjust bin sizes and numbers in order to achieve highest MI. MIC values between two variables fall between 0 and 1, with 0 meaning the two variables are completely independent and 1 meaning one is dependant on the other. We used both Pearson's r and MIC in order to compare orbit frequencies. Although results from both approaches were more or less consistent, MIC showed higher robustness than Pearson's r method.

If MIC is used as correlation measure, each element of $\mathbf{c}$ is calculated as below:

$$m_{ijo}^{A,B} = MIC(\mathbf{S}_{ij.o}^{A}, \mathbf{S}_{ij.o}^{B}) \quad (3)$$

Alternatively, if we use Pearson criterion we would have:

$$m_{ijo}^{A,B} = Pearson(\mathbf{S}_{ij.o}^{A}, \mathbf{S}_{ij.o}^{B}) \quad (4)$$

RESULTS AND DISCUSSIONS

Contact Map Orbit Vector Distance

By comparing signature distance vectors, one can find how contact maps differ from each other in terms of local structure. Contact maps can serve as measures of spatial proximity between loci. Graphlets capture certain patterns of interaction, or in other words, spatial neighborhood for each loci. Thus, if signature vectors of two loci are close, it can be inferred that they have similar spatial neighborhood.

We can compare pairs of contact maps in terms of their closeness to each other. As an example, in figure 6, all pairs of cells are compared to each other in terms of their distance for contact maps involving chromosome 14. We can se that for the first 13 inter-chromosomal contact maps, ALL and RL cell lines are closer to each other than to other cell lines.

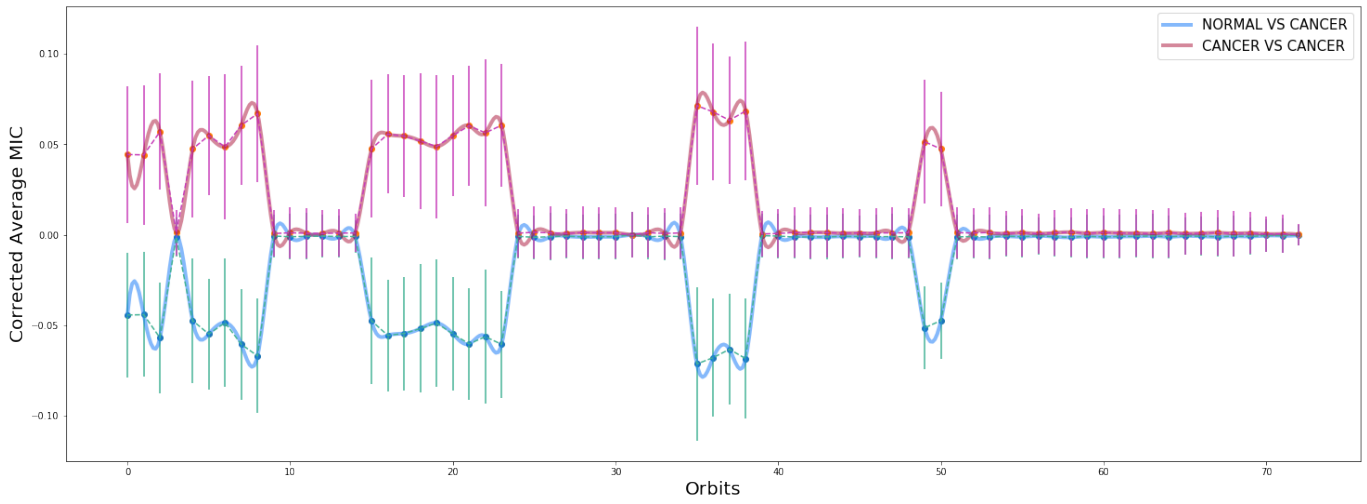
We performed one-way ANOVA statistical test to see if there are significant difference between cancer-normal and cancer-cancer pairs. We found that the difference statistically significant difference. $(F(1,1654) = 20.49, p < 0.0001)$. As illustrated in figure 5a, we can see that normal-cancer pairs have higher distance from each other than cancer-cancer pairs.

We then continued to investigate each pair separately to see if there is any significant difference between them. Again, our statistical tests (ANOVA) showed significant difference between pairs of cells. The results are shown in figure 5b. We can see that ALL-RL pair are closest to each other while ALL-MIT and MIT-RL are most distant. We found statistcally significant difference for difference between individual pairs except for ALL-CALL4 and CALL4-RL as well as MIT-RL and ALL-MIT. The results of thse tests can be found in supplementary material.

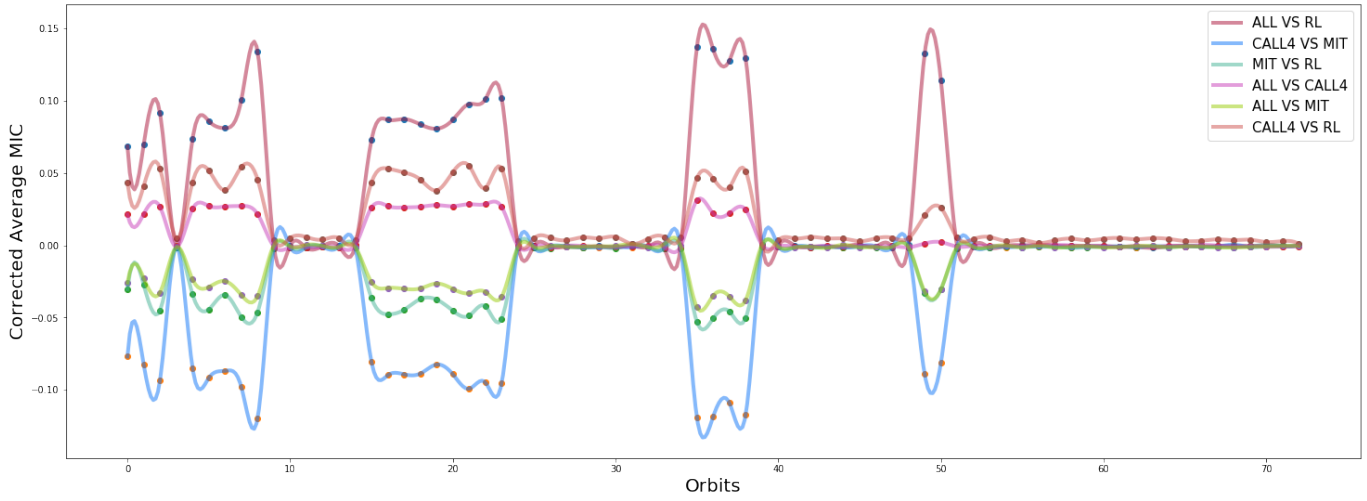
The results in figure 5b is also in keeping with what we see in figure 6. For example, as mentioned earlier, our results show that on average, ALL and RL cell lines are closer to each other than to other cell lines, which is also the case in figure 6 for majority of contact maps.

MIC Comparison

In addition to comparing cells in terms of their orbit distances, we can compare them by measuring how often certain graphlets occur in their contact maps. By doing so, we measure the frequency distribution of the spatial structures represented by orbits in each contact map. In order to see how closely such structures are distributed, we can compare



(a)



(b)

Figure 7. Pair-wise average contact map orbit correlations for all contact maps: ($\bar{\mathfrak{m}}^{A,B}_{i,j} \quad \forall i,j \in \{1...23\} \quad \& \quad j \geq i$: average along the red *vertical* arrow in figure 4). These values are calculated by averaging over pairwise correlations of orbits of $\mathbb{Q}$ in a contact map. **Pair-wise average orbit correlations:** In figure 7b, each point in the graph is the result of averaging pair-wise orbit correlations over all contact maps ($\frac{1}{276}\sum_{i=0}^{23}\sum_{j=i}^{23} m^{A,B}_{i,j,o} \quad \forall o \in \{0,1,...,72\}$: average along the red *horizontal* arrows in figure 4). Counts for certain orbits are always zero in inter-chromosomal maps, leading to average value close to zero in Figure 7b. In this figure, cancer cells data points are depicted in warm colors while normal cells are depicted in cold colors in increased contrast. As can be seen orbit distributions are more similar to each other for cancer cells.

contact maps by calculating the correlation between their orbit distributions. A higher correlation for certain orbits would mean higher similarity in terms of that particular spatial structure between the loci involved.

Before going on with the results, it is worth mentioning that interchromosomal thresholded contact maps represent a bipartide graph with the loci from each chromosome on one side. Due to this bipartide nature of the graphs in inter-chromosomal maps, count of certain orbits is always 0, resulting in a correlation values of 0 for them as well. You can see the bias in figure 7 where average correlations of orbits $\mathbb{Q} = \{3,9,10\text{-}14,20\text{-}34,39\text{-}48,51\text{-}72\}$ are close to zero. In fact all correlations corresponding to these orbits are 0 except for the ones between the same chromosomes.

We caclulated pair-wise MIC values for each orbit in each of the 276 contact maps from MIT, ALL, RL, and CALL4 data separately. We found statistically significant difference between cancer-cancer and normal-cancer correlations. ($F(73,1582) = 6.29$, $p < 0.00001$, Wilk's $\Lambda = 0.775$) The difference is also illustrated in figure 7a, which plots corrected mean difference of MIC values for cance-cancer and normal-cancer correlations. We performed statistical test to see if there is a significant difference correlation between individual pairs of cells. Figure 7b demonstrate such difference. Average correlation over all contac maps for normal-cancer pairs are smaller than cancer-cancer pairs. This is corroborated by the results of statical test which verify that such difference is in

fact signifcant.($F(365,7882) = 3.91$, $p < 0.00001$, Wilk's $\Lambda = 0456$) Figures 7a and 7b both show more details about this difference in correlation. As can be observed, for obits in $\mathbb{Q}$, normal correlations are smaller than cancer correlations, while for the rest of the orbits, there is no difference.

Figure 7 demonstrates that certain orbits of Leukemic cells have higher correlation to each other than to the normal MIT cell. In fact our statistical analysis shows that *for orbits NOT in $\mathbb{Q}$, intra-leukemic orbit correlations are significantly higher than leukemic-normal orbit correlations*. This implies there are significant differences between normal and leukemic cells in terms of their local structure.

RESOURCES

Hi-C Datasets:

1. Code base for this article
2. Datasets including cancerous cells
3. Original Datasets

REFERENCES

1. 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.
2. 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.
3. William Noble, Zhi-jun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, and C Anthony Blau. A three-dimensional model of the yeast genome. In *International Conference on Research in Computational Molecular Biology*, pages 320–320. Springer, 2011.
4. Mathieu Rousseau, James Fraser, Maria A Ferraiuolo, Josée Dostie, and Mathieu Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC bioinformatics*, 12(1):414, 2011.
5. Ming Hu, Ke Deng, Zhaohui Qin, Jesse Dixon, Siddarth Selvaraj, Jennifer Fang, Bing Ren, and Jun S Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS computational biology*, 9(1):e1002893, 2013.
6. Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert. A statistical approach for inferring the 3d structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.
7. Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. *Nucleic acids research*, 42(7):e52–e52, 2014.
8. ZhiZhuo Zhang, Guoliang Li, Kim-Chuan Toh, and Wing-Kin Sung. 3d chromosome modeling with semi-definite programming and hi-c data. *Journal of computational biology*, 20(11):831–846, 2013.
9. Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3d genome reconstruction from chromosomal contacts. *Nature methods*, 11(11):1141, 2014.
10. Davide Baù, Amartya Sanyal, Bryan R Lajoie, Emidio Capriotti, Meg Byron, Jeanne B Lawrence, Job Dekker, and Marc A Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature Structural and Molecular Biology*, 18(1):107, 2011.
11. Susan M Gasser. Visualizing chromatin dynamics in interphase nuclei. *Science*, 296(5572):1412–1416, 2002.
12. Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2):104, 2007.
13. Julio Mateos-Langerak, Manfred Bohn, Wim de Leeuw, Osdilly Giromus, Erik MM Manders, Pernette J Verschure, Mireille HG Indemans, Hinco J Gierman, Dieter W Heermann, Roel Van Driel, et al. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences*, 106(10):3812–3817, 2009.
14. Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.
15. Tijana Milenkoviæ and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257, 2008.
16. 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.
17. 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.
18. 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.
19. Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. Hic-pro: an optimized and flexible pipeline for hi-c data processing. *Genome biology*, 16(1):259, 2015.
20. 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.
21. 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.
22. David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.