4DNvestigator: a toolbox for the analysis of time series Hi-C and RNA-seq data

Scott Ronquist[1], Sijia Liu[2], Michael Perlman[3], and Indika Rajapakse[1,4]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA 02142, USA
[3]Department of Statistics, University of Washington, Seattle, WA 98195, USA
[4]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA

December 30, 2019

Abstract

The combined analysis of genome structure and function over time, and how these changes affect cellular phenotype, is referred to as the 4D Nucleome (4DN). 4DN analysis is necessary to fully understand how a cell operates, but 4DN analysis tools are currently underdeveloped. We present the "4DNvestigator", a user-friendly toolbox for the analysis of time-series genome structure, measured by genome-wide chromosome conformation capture (Hi-C), and genome function, measured by RNA sequencing (RNA-seq).

Availability: https://github.com/scotronq/4DNvestigator

1 Introduction

The genome is a dynamical system where changes in genome structure and function over time affect the cell's phenotype. This dynamical relationship between genome structure, transcriptional landscapes, and cellular phenotypes is referred to as the 4D Nucleome (4DN) [1]. The 4DN is an emerging field of study, wherein a significant investment is being made by research institutes to enhance our understanding of how changes in nuclear organization affects normal development and disease states [2].

To analyze the 4DN, genome-wide chromosome conformation capture (Hi-C) and RNA sequencing (RNA-seq) are often used to observe genome structure and function, respectively. As high throughput sequencing costs decline, the availability and volume of Hi-C and RNA-seq data sets is expected to increase. With this increase in data, the development of methods to properly analyze these data sets is imperative. Furthermore, analyzing these high-dimensional data sets is non-trivial and requires a high-level understanding of biology, mathematics, and computer science.

To aid researchers in the analysis of 4DN data, we present the 4DNvestigator. The 4DNvestigator is a MATLAB toolbox that loads time series Hi-C and RNA-seq data, extracts important structural and functional features, and is compatible with established Hi-C and RNA-seq processing programs and formats. This toolbox includes both established and novel 4DN data analysis methods, and displays important findings in simple visualizations. The 4DNvestigator takes a network-based approach to this analysis. This approach allows for the computation of network centrality and network entropy, both of which have been shown to reflect biological phenomena [3, 4]. Additionally, we present a novel statistical method for comparing two or more Hi-C matrices. We compare this method against established Hi-C comparison methods within this text. We believe adoption by the community at large of analytical tools such as the one described in our

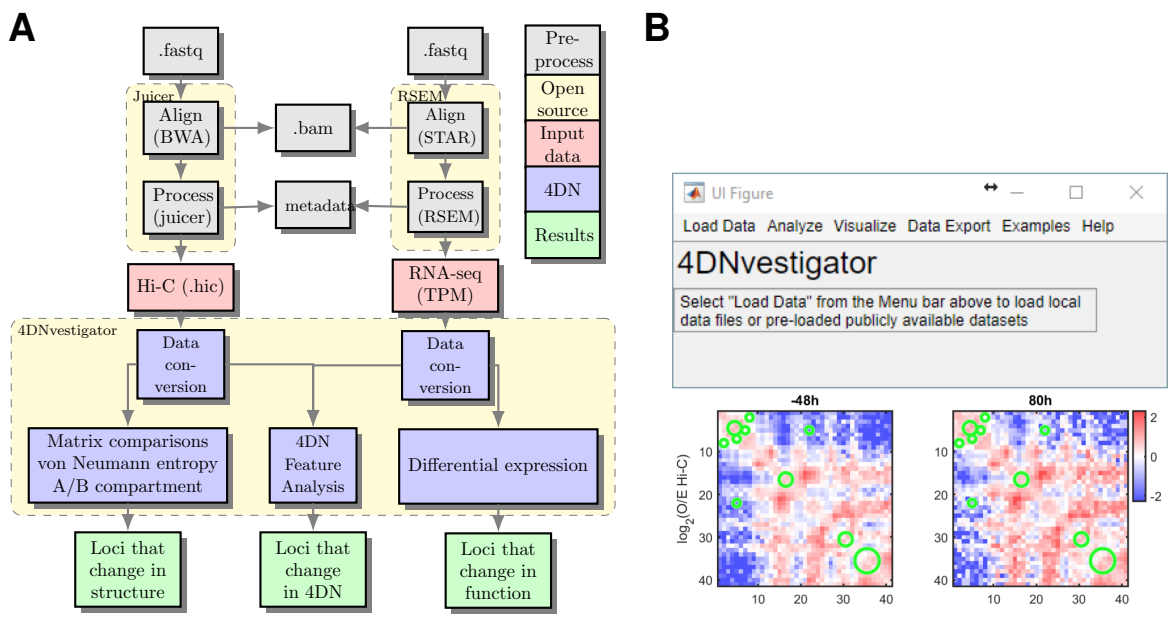

Figure 1: Overview of the 4DNvestigator. (**A**) Data processing pipeline for the 4DNvestigator. Within this diagram, 4DN refers to the 4DNvestigator. (**B**) Overview of the 4DNvestigator GUI, with a comparison between time series Hi-C samples at a 4Mb region surrounding the gene locus *MYH1*. Data is obtained from Liu *et al.*, time points -48, 8, 80h [3]. The procedure described by Larntz and Perlman (See section "LP method for comparing Hi-C matrices") detects regions that are most significantly different between samples, and rejects the null hypothesis that these matrices are equal. Green circles highlight regions with the largest difference between samples. Example codes to recalculate these results are included within the 4DNvestigator.

paper is critical to advance the field of 4DN research.

2 MATERIALS AND METHODS

An overview of the 4DNvestigator workflow is depicted in Figure 1A. The 4DNvestigator takes processed Hi-C and RNA-seq data as input, along with a metadata file which describes the sample and time point for each input Hi-C/RNA-seq file (See See S1 text). A graphical user interface (GUI) is provided for ease of use (Figure 1B). A number of novel methods for analyzing 4DN data are included within the 4DNvestigator and are described below. A list of accepted input file formats is provided at https://github.com/scotronq/4DNvestigator.

2.1 4DN feature analyzer

The "4DN feature analyzer" quantifies and visualizes how much a genomic region changes in structure and function over time. To achieve this, we adopt a network point of view, where genomic regions are nodes and interactions between regions are edges. Edge weights are set relative to the number of contacts between regions in Hi-C. With this network point of view, we can quantify the importance of each node using network centrality. Network centrality identifies nodes that play an influential topological role in the network [5]. There are a number of different centrality measures, each of which identifies a particular type of nodal influence. For example, *degree centrality* measures the connectedness of a node locally, calculated by the

summation of edge weights for all edges connected to the node. In contrast, *betweenness centrality* is a more global connectedness measure, which counts the number of times a given node is on the shortest path between all other pairs of nodes. Centrality measures derived from Hi-C have been shown to have biological meaning, reflecting the chromatin accessibility of each region and predicting A/B compartment switch locations [3]. In the 4DN feature analyzer, centrality measures and gene expression (RNA-seq) define the state of each genomic region at each time point. The 4DN feature analyzer quantifies and visualizes how each genomic region changes in this state over time. Genomic regions can be defined at many scales, such as gene-level, 100 kb-level, or Topologically Associating Domain (TAD)-level.

Hi-C matrices are balanced using the Knight-Ruiz (KR) algorithm and are observed over expected (O/E) normalized using Juicer Tools [6]. We define Hi-C matrices of this form as $\mathbf{A}^{(m)}$, where m denotes the time point (or sample). The following network centrality measures are extracted from $\mathbf{A}^{(m)}$: degree, eigenvector, betweenness, and closeness [3]. The centrality measures are combined with the RNA-seq expression for the corresponding genomic regions to form a "feature" matrix that defines the state of each genomic region at each time point. RNA-seq data is measured by the $\log_2$ transformation of Transcripts Per Million (TPM). For regions containing more than one gene, the mean $\log_2$(TPM) is computed. We define RNA-seq data of this form as $\mathbf{r}^{(m)}$. The z-score for each feature is computed to normalize the data and to put features on the same relative scale. Feature matrices for each time point are then projected to a common low dimensional space (2D or 3D) via PCA, Laplacian Eigenmaps, or t-SNE [7, 8]. This allows the genomic regions to be visualized in 2D. Genomic regions that move significantly within this low dimensional space also change the most in structure and function over time. Although Hi-C matrices are often highly structured, with a majority of contacts occurring between neighboring genomic regions, this analysis highlights regions that fluctuate in structure and function relative to the other genomic regions included in the analysis. Points that correspond to the same genomic region at different time points are fit with a Minimum Volume Ellipsoid (MVE). The MVE volume quantifies the "4DN variance" of each genomic region. A Minimum Area Ellipse (MAE) is calculated if the points lie on a plane. Methods to perform this analysis are described in Algorithm 1, following methods outlined in Liu *et al.* [3].

Algorithm 1: 4DN feature analyzer

Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n\times n}$, and RNA-seq vectors $\mathbf{r}^{(m)} \in \mathbb{R}^{n\times 1}$, $m = 1, \ldots, k$

Output: low dimensional space $\mathbf{Y}^{(m)}$, and 4DN variance $\mathbf{v}$

1 Compute degree, eigenvector, betweenness, and closeness centrality of $\mathbf{A}^{(m)}$, and define as $\mathbf{b}^{(m)}_{deg}$, $\mathbf{b}^{(m)}_{eig}$, $\mathbf{b}^{(m)}_{bet}$, $\mathbf{b}^{(m)}_{close}$, respectively, where each $\mathbf{b}^{(m)} \in \mathbb{R}^{n\times 1}$

2 Form the feature matrices $\mathbf{X}^{(m)} = [\mathbf{b}^{(m)}_{deg}, \mathbf{b}^{(m)}_{eig}, \mathbf{b}^{(m)}_{bet}, \mathbf{b}^{(m)}_{close}, \mathbf{r}^{(m)}]$, where $\mathbf{X}^{(m)} \in \mathbb{R}^{n\times 5}$

3 Normalize the columns of $\mathbf{X}^{(m)}$

4 Compute the common low dimensional space $\mathbf{Y}^{(m)}$ using PCA, Laplacian Eigenmaps, or t-SNE

5 Compute the 4DN variance for each genomic region $\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^{n\times 1}$

Return: $\mathbf{Y}^{(m)}$ and $\mathbf{v}$

2.2 Network entropy

Entropy measures the order within a system, where higher entropy corresponds to more disorder [9]. Here, we apply this measure to Hi-C data to quantify the order in chromatin structure. Biologically, genomic regions with high entropy likely correlate with high proportions of euchromatin, as euchromatin is more structurally permissive than heterochromatin [4]. Furthermore, entropy can be used to quantify stemness, since cells with high pluripotency are less defined in their chromatin structure [10]. Since Hi-C is a multivariate analysis measurement (each contact coincidence involves two variables, the two loci), we use multivariate entropy, Von Neumann Entropy (VNE). The algorithm to compute VNE is given in Algorithm 2.

2.3 LP method for comparing Hi-C matrices

A number of methods have been developed in the Hi-C research community to compare Hi-C matrices. These methods can be broadly grouped into 2 categories: Hi-C reproducibility methods (e.g. HiCRep,

Algorithm 2: VNE Computation

Input: Hi-C matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$

Output: VNE

1 Compute the correlation matrix $\mathbf{C} = \text{corr}(\log_2(\mathbf{A}))$

2 Compute the eigendecomposition of $\mathbf{C}$, where $\lambda_1 \leq \lambda_i \leq \lambda_n$ are the eigenvalues of $\mathbf{C}$

3 Normalize the eigenvalues: $\bar{\lambda}_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}$

4 Compute VNE: $\text{VNE} = -\sum_{i=1}^{n} \bar{\lambda}_i \ln(\bar{\lambda}_i)$

Return: VNE

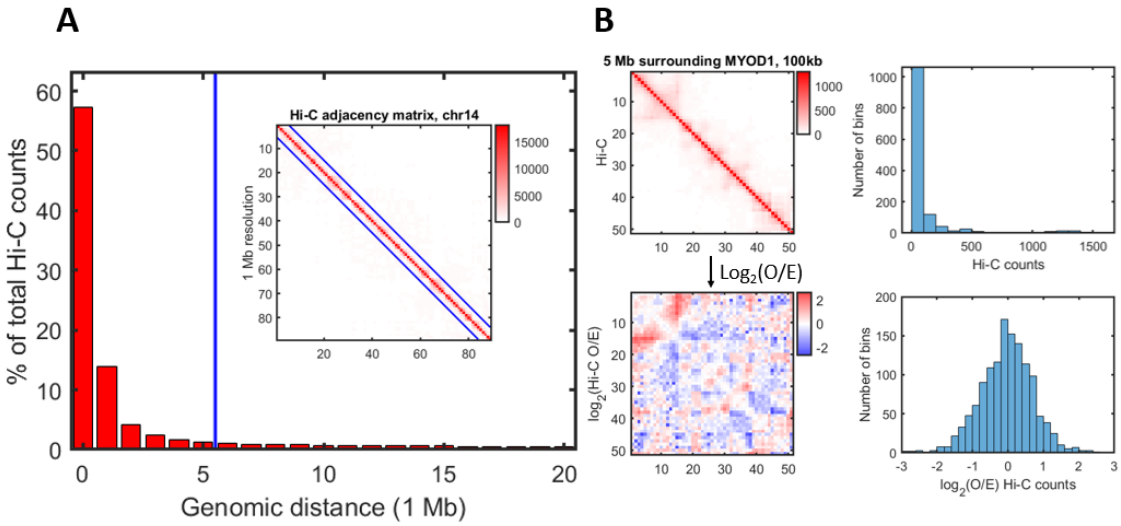

Figure 2: Hi-C normalization for the LP method. **(A)** Percentage of total Hi-C contacts by genomic distance. Inset figure shows a typical Hi-C intra-chromosome adjacency matrix at 1 Mb resolution. The blue line denotes where the genomic distance exceeds 5 Mb. **(B)** A 5Mb Hi-C matrix (100 kb resolution, top left) and a histogram of the counts within the matrix (top right). When the $\log_2$(O/E) transformation is computed on the Hi-C matrix (bottom left), elements within the matrix are normally distributed (bottom right).

HiC-spector), and Differential Chromatin Interactions (DCI) methods (e.g. SELFISH, HiCCompare, FIND, diffHic) [11, 12, 13, 14, 15, 16]. Hi-C reproducibility methods are useful for assessing the equality of Hi-C matrices genome-wide, and detecting technical bias between samples. DCI methods determine which loci have a significantly different number of Hi-C contacts between samples. All methods listed above are comparisons between only two matrices.

Here we pose a different, but related, question: Is the chromatin structure, within a genomic region, equivalent between samples? To answer this question, we take a multivariate statistical approach. Elements within Hi-C matrices have been shown to be normally distributed when the $\log_2$ transformation is applied to $\mathbf{A}^{(m)}$ (See Figure 2B) [17]. For data of this form, a method for testing the equality of correlation matrices was proposed by Larntz and Perlman in 1988 [18]. Here, we apply this technique to Hi-C correlation matrices to determine the statistical difference between multiple Hi-C samples at genomic regions of interest. Furthermore, we extend this method to determine where the matrices are most significantly different between samples. We refer to this method herein as the "LP" method. We recommend that the LP method is performed at Hi-C resolutions $\leq$ 100 kb and that regions do not extend >5 Mb, as signal (counts) often become sparse as the genomic distance between loci increases (See Figure 2A). The full algorithm is given in Algorithm 3. Algorithm 3 outputs a P-value for the equality of the Hi-C matrices and a

matrix $\mathbf{S}$ where the largest values in $\mathbf{S}$ correspond to the genomic regions that are most different between samples. An example of this analysis is shown in Figure 1B, which can be recreated using codes given at https://github.com/scotronq/4DNvestigator.

Algorithm 3: LP method

Input: Hi-C matrices $\mathbf{A}^{(m)} \in \mathbb{R}^{n\times n}$, $m = 1, \ldots, k$

Output: P-value p, and test statistic $\mathbf{S}$

1 Compute the correlation matrix $\mathbf{C}^{(m)} = \text{corr}(\log_2(\mathbf{A}^{(m)}))$. Define the corresponding population correlation matrices as $\mathbf{P}^{(m)}$

2 Define the null hypothesis

$$H_0 : \mathbf{P}^{(1)} = \cdots = \mathbf{P}^{(k)}$$

3 Compute the Fisher z-transformation $\mathbf{Z}^{(m)}$. Elements in $\mathbf{Z}^{(m)}$ and $\mathbf{C}^{(m)}$ are denoted as $z_{ij}^{(m)}$ and $c_{ij}^{(m)}$, respectively, and

$$z_{ij}^{(m)} = \frac{1}{2}\ln\left[\frac{1+c_{ij}^{(m)}}{1-c_{ij}^{(m)}}\right]$$

4 Form $\mathbf{S}$, where elements in $\mathbf{S}$ are denoted as s_{ij} and

$$s_{ij} = (n-3)\sum_{m=1}^{k}(z_{ij}^{(m)} - \bar{z}_{ij})^2,\ \bar{z}_{ij} = k^{-1}\sum_{m=1}^{k} z_{ij}^{(m)}$$

5 Calculate the test statistic

$$T = \max_{1\le i<j\le n} s_{ij}$$

6 Reject H_0 at level α if $T > \chi^2_{k-1,\epsilon(\alpha)}$, where $\chi^2_{k-1,\epsilon(\alpha)}$ is the chi-squared distribution with $k-1$ degrees of freedom and $\epsilon(\alpha) = (1-\alpha)^{2/n(n-1)}$ is the Šidák correction

7 Calculate p, the level α at which $T > \chi^2_{k-1,\epsilon(\alpha)}$

Return: p and $\mathbf{S}$

2.4 Chromatin partitioning and differential expression

The 4DNvestigator includes a suite of previously developed Hi-C and RNA-seq analysis methods. Hi-C A/B compartments can be extracted using previously defined methods [19, 20]. Regions that change compartments between samples are automatically identified. The 4DNvestigator also utilizes developed MATLAB scripts for differential gene expression that follow methods outlined in Anders *et al.* [21]. The 4DNvestigator takes results obtained from the methods described above and identifies regions that change in both Hi-C and RNA-seq.

The first method for determining A/B compartments is based on the Fielder vector of the Hi-C matrix, following methods outlined in Chen *et al.* [22]. The second method for determining A/B compartments is based on the first principal component of the Hi-C matrix, following methods outlined in Lieberman-Aiden *et al.* [20]. The A/B compartment partitioning for every chromosome and sample is plotted in a figure, where regions that have a large change in sign and magnitude are highlighted.

2.5 Simulated Hi-C data

Simulated Hi-C data was created to compare the LP method against alternative Hi-C comparison methods. Two distinct simulated data sets were created to have: (1) changes in chromatin loop structure and (2) changes in chromatin compartment structure. Chromatin loops and chromatin compartments are two features that have been used to characterize Hi-C structure [23, 20]. Both simulated data sets are created by perturbing data from real Hi-C matrices: a 400 kb region (10 kb resolution) is used for the chromatin loop data set and a 2Mb region (50 kb resolution) is used for the chromatin compartment data set.

One additional simulated matrix was also created by adding a small amount of random noise to the 400 kb region (10 kb resolution) Hi-C matrix. A random number sampled from a normal distribution, $\mathcal{N}(0,\ 0.05)$,

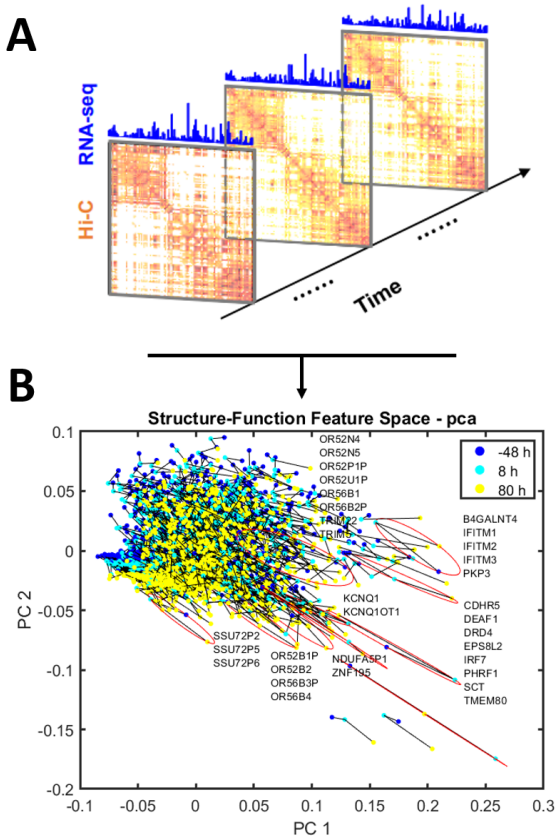

Figure 3: 4DN feature analyzer. (A) Time series Hi-C and RNA-seq data. (B) 4DN feature analyzer example. Points correspond to genomic loci (100 kb). Loci that change the most in structure and function (top 10% of bins, default setting) are highlighted with red minimum volume ellipses, and genes within these regions are listed within the figure.

is added to each element in $\log_2(\mathbf{A})$. This matrix is used to determine how robust Hi-C matrix comparison methods are to small amounts of noise.

For each simulated data set, 10 matrices were created that are incrementally more divergent from the original Hi-C matrix. For changes in loop structure, counts were added to a specific off-diagonal region, following a 2D gaussian distribution ($\sigma = 1$), to model a chromatin loop structure. For changes in compartment structure, counts aligned to a specific genomic region (bin) were changed to decrease the correlation coefficient between the specified bin in the simulated Hi-C $\log_2(\mathbf{A})$ matrix and the specified bin in the original Hi-C $\log_2(\mathbf{A})$ matrix. These changes reflect what would be observed if the specified bin was changing its compartment structure. Methods to recreate the simulated Hi-C matrices are provided within the 4DNvestigator.

3 RESULTS

3.1 4DN feature analyzer

We demonstrate how the 4DN feature analyzer makes use of time series information by analyzing time series Hi-C and RNA-seq data collected on human dermal fibroblasts that are being reprogrammed to the muscle lineage, mediated by the addition of master transcription factor MYOD1 [3]. Three time points are analyzed here, corresponding to 100 kb resolution Hi-C and RNA-seq data on samples collected: 48 h prior to MYOD1 addition ("-48 h"; control), 8 h post-MYOD1 addition, and 80 h post-MYOD1 addition. Figure 3 shows the

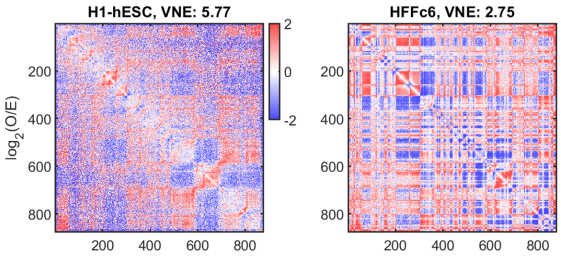

Figure 4: VNE difference between cell types. The VNE for an undifferentiated cell type (embryonic stem cell, H1-hESC) is much higher than the more differentiated cell type (fibroblast, HFFc6). Intra-chromosome $\log_2(\mathbf{A})$ Hi-C matrices are shown at 100 kb resolution for chromosome 14.

results from the 4DN feature analyzer for chromosome 11. Each time point is labeled with a unique color to show how the regions change in 4DN over time. Regions that move the most in this low dimensional space are highlighted with red ellipses, and the genes within these regions are shown within the figure. Data and code to perform this analysis are provided within the toolbox "examples4DNvestigator.m".

3.2 Network entropy

To demonstrate how VNE can be used to quantify disorder in chromatin structure, we have calculated the VNE of two distinct cell types: a differentiated cell type, HFFc6, and an undifferentiated cell type, H1-hESC. It is clear from the Hi-C matrices that the differentiated cell type is more ordered in its chromatin structure (See Figure 4). VNE reflects this observation, as the VNE for the differentiated cell type HFFc6 is much lower (2.87) than the undifferentiated cell type H1-hESC (5.74). This analysis is expanded with additional samples and cell types in Supporting Information (See S2 text and S1 figure).

3.3 Hi-C matrix comparison

Within the 4DNvestigator we present a novel statistical method for detecting differences in Hi-C matrices, the LP method. Specifically, we test the null hypothesis that the Hi-C correlation matrices, derived from different samples, are equivalent (See Section LP method for comparing Hi-C matrices). In order to assess the LP method's ability to detect differences in Hi-C matrices between samples, we have compared the LP method against alternate Hi-C comparison methods: HiCRep, HiC-spector, and SELFISH [11, 12, 13]. HiCRep, HiC-spector are Hi-C reproducibility metrics, while SELFISH is a DCI method that was recently shown to outperform all prior methods for DCI detection. We note that there are no methods, that we are aware of, that are a direct comparison with our method. All of the alternate methods we are comparing here were designed to address related, but fundamentally different, questions: Hi-C reproducibility metrics for genome-wide comparisons, and DCI for loci interaction differences (not overall structural differences within a region). Nevertheless, we compare them here to highlight where our method is advantageous.

There are many ways in which Hi-C matrices can be different between samples. Two well established Hi-C features are loops and compartments [23, 20]. To assess how the LP method performs when these features are different between samples, we have created simulated Hi-C data sets with incremental changes in these features. We then determined the point at which the LP method, as well as alternate methods, detect differences between the matrices (See Section Simulated Hi-C data).

For Hi-C matrices with changes in loop structure as the only differences between samples, the LP method does not detect that the matrices are different until the loop interaction contained six times the mean number of contacts for loci at the given genomic distance. Simulated matrices for this analysis are displayed in Figure 5A. In this situation, SELFISH does the best job of detecting differences between samples. HiC-spector shows a consistent decrease in its reproducibility score, while HiCRep performs similar to the LP method. We note that the LP method, as well as HiCRep, relies on changes in the correlation between samples, and thus changes to a small number of elements within the Hi-C matrices do not affect the measurement significantly.

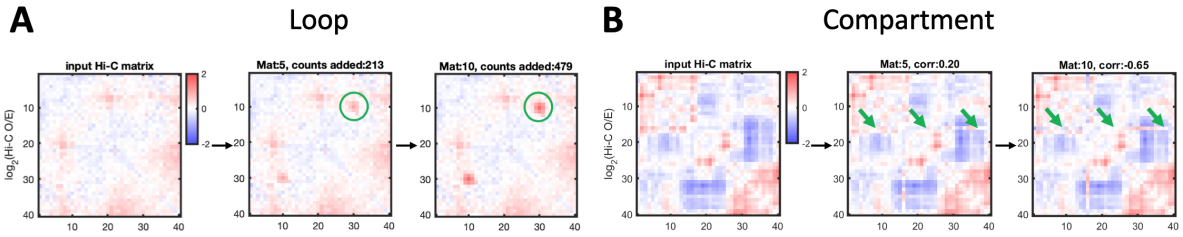

Figure 5: Simulated Hi-C data. **(A)** Simulated Hi-C matrices for loop changes (See Table 1). The original Hi-C matrix is shown on the left, with matrices that are increasingly more divergent from left to right. Green circles indicate where the matrix has been perturbed to change the chromatin loop structure. **(B)** Simulated Hi-C matrices for compartment changes (See Table 1). The original Hi-C matrix is shown on the left, with matrices that are increasingly more divergent from left to right. Green arrows indicate where the matrix has been perturbed to change the chromatin compartment structure.

Table 1: Hi-C matrix comparison methods. "Counts added" refers to the number of counts added to the simulated Hi-C matrix, at the specified location (See Section "Simulated Hi-C data"). "Correlation" refers to the correlation between the selected column in the original Hi-C matrix, and the same column in the simulated Hi-C matrix (See Section "Simulated Hi-C data"). We note here that each method outputs a different unit, as specified within the table. The lowest P-value output from SELFISH is given here. This table is related to Figure 5.

	Counts added					Correlation					
Method	53	160	266	372	479	1 (noise)	0.87	0.57	0.2	-0.26	-0.65
LP (P-value)	1.0	1.0	0.99	0.41	0.045	1.0	1.0	2.1E-03	2.5E-07	1.7E-13	0.0
SELFISH (P-value)	0.0	0.0	0.0	0.0	0.0	0.02	2.9E-05	1.4E-08	1.3E-10	3.5E-13	2.1E-07
HiCRep (SCC)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.99	0.98	0.97	0.96
HiC-spector (Q)	0.99	0.98	0.96	0.94	0.92	0.39	0.43	0.33	0.48	0.44	0.50

For Hi-C matrices with changes in the compartment structure of a single bin (genomic region) as the only differences between samples, the LP method performs very well. Simulated matrices for this analysis are displayed in Figure 5B. SELFISH detects differences between all samples, but also detects differences when only a small amount of noise is added to the original matrix. The LP method is robust to noise and shows a trend consistent with the amount of change created. The HiC-spector reproducibility measurement shows no clear trend as the matrices diverge, while HiCRep begins to detect differences between the matrices much later than the LP method. Compartment changes are often observed in time series Hi-C matrices as cells transition between cell states, either through reprogramming or differentiation [24].

4 DISCUSSION

We believe the 4DN research community will find the 4DNvestigator suite of tools useful for the analysis time series Hi-C and RNA-seq data. The 4DN feature analyzer methods have already been used to detect important 4DN changes in times series Hi-C and RNA-seq data sets [3]. Providing the standardized code here within the 4DNvestigator will allow a wider audience of researchers to apply this technique to their own

work. Methods to systematically quantify stemness and plasticity via VNE will help identify cells in specific states. Such methods may be especially useful in the context of cancer, as identifying and targeting cells with high stemness is an attractive clinical strategy [25].

The LP method for comparing Hi-C matrices will be extremely valuable to the Hi-C research community as it fulfills a number of unmet needs. The LP method can be used to analyze a range of Hi-C resolutions, from 5 kb resolution to 100 kb chromatin compartment changes, and all resolutions in between. Many Hi-C analysis techniques were developed to detect specific biological phenomena within Hi-C matrices, such as loops or compartments, and are therefore less robust for detecting overall changes in chromatin structure. Also, the LP method outputs a p-value for the simple statistical test that entire chromatin structure is different between the samples being compared, not just specific chromatin interactions, making the LP technique more robust. Additionally, the LP method is built for the analysis of multiple Hi-C samples, since there is no limit on k for the algorithm input (See Algorithm 3). This is useful for the analysis of time series data, which often include more than two time points.

5 CONCLUSION

The 4DNvestigator provides methods to analyze time series Hi-C and RNA-seq data in a rigorous yet automated manner. The combined analysis of network centrality and RNA-seq over time can be easily performed using the 4DN feature analyzer. This analysis outputs a simple 2D plot representing how the genome changes in 4DN, with large changes highlighted. Network entropy is a simple metric which can be used to characterize the structural disorder in a region. Finally, the 4DNvestigator introduces a novel statistical method for comparing Hi-C matrices, the LP method. The LP method is distinct from established Hi-C matrix comparison methods, as it takes a statistical approach to test for matrix equality, and allows for the comparison of ≥ 2 matrices simultaneously. The 4DNvestigator can be applied to any time series Hi-C and RNA-seq data set, and many methods are applicable to non-time series data samples as well.

6 DATA AVAILABILITY

Written in MATLAB and available as MATLAB scripts on https://github.com/scotronq/4DNvestigator.

7 SUPPORTING INFORMATION

S1 text: Data pre-processing
S2 text: VNE analysis expanded
S1 figure: VNE difference between cell types. VNE difference between cell types. Hi-C matrices are shown in order of VNE, left to right. The more pluripotent cell lines (hESC, HUVEC) have a much higher VNE than the more differentiated cell lines (IMR90, hHFF6c, GM12878). Intra-chromosome $\log_2(\mathbf{A})$ Hi-C matrices are shown at 100 kb resolution for chromosome 14.

8 ACKNOWLEDGEMENTS

We would like to thank Alnawaz Rehemtulla, Stephen Lindsly, Can Chen, and Gabrielle Dotson for feedback on the manuscript and assistance in developing toolbox codes.

9 FUNDING

This work is supported, in part, under Air Force Office of Scientific Research Award No: FA9550-18-1-0028, DARPA Deep-Purple, FunCC Programs, and the Smale Institute.

9.0.1 Conflict of interest statement.

None declared.

References

[1] Ried, T. and Rajapakse, I. (2017) The 4D Nucleome. *Methods,* **123**, 1–2.

[2] Dekker, J., Belmont, A. S., Guttman, M., Leshyk, V. O., Lis, J. T., Lomvardas, S., Mirny, L. A., O'Shea, C. C., Park, P. J., Ren, B., et al. (2017) The 4D nucleome project. *Nature,* **549**(7671), 219–226.

[3] Liu, S., Chen, H., Ronquist, S., Seaman, L., Ceglia, N., Meixner, W., Chen, P.-Y., Higgins, G., Baldi, P., Smale, S., et al. (2018) Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming. *iScience,* **6**, 232–246.

[4] Macarthur, B. D. and Lemischka, I. R. (2013) Statistical mechanics of pluripotency. *Cell,* **154**(3), 484–489.

[5] Newman, M. (2010) Networks: an introduction, Oxford university press, .

[6] Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., and Aiden, E. L. (2016) Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems,* **3**(1), 95–98.

[7] Belkin, M. and Niyogi, P. (2001) Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Nips,* **14**, 585–591.

[8] Van Der Maaten, L. and Hinton, G. E. (2008) Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research,* **9**, 2579–2605.

[9] Cover, T. M. and Thomas, J. A. (2012) Elements of information theory, John Wiley & Sons, .

[10] Meshorer, E. and Misteli, T. (2006) Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews Molecular cell biology,* **7**(7), 540.

[11] Yang, T., Zhang, F., Yardimci, G. G., Song, F., Hardison, R. C., Noble, W. S., Yue, F., and Li, Q. (2017) HiCRep: assessing the reproducibility of Hi-C data using a stratum- adjusted correlation coefficient. *Genome Research,* **27**(11), 1939–1949.

[12] Yan, K. K., Yardlmcl, G. G., Yan, C., Noble, W. S., and Gerstein, M. (2017) HiC-spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. *Bioinformatics,* **33**(14), 2199–2201.

[13] Ardakany, A. R., Ay, F., and Lonardi, S. (2019) Selfish: Discovery of Differential Chromatin Interactions via a Self-Similarity Measure. *bioRxiv,* p. 540708.

[14] Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018) FIND: DifFerential chromatin INteractions Detection using a spatial Poisson process. *Genome Research,* **28**(3), 412–422.

[15] Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018) HiCcompare: An R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics,* **19**(1), 13–16.

[16] Lun, A. T. and Smyth, G. K. (2015) diffHic: A Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics,* **16**(1), 1–11.

[17] Calandrelli, R., Wu, Q., Guan, J., and Zhong, S. (2018) GITAR: An Open Source Tool for Analysis and Visualization of Hi-C Data. *Genomics, Proteomics and Bioinformatics,* **16**(5), 365–372.

[18] Larntz, K. and Perlman, M. D. (1988) A simple test for equality of correlation matrices. *Statistical Decision Theory and Related Topics IV,* pp. 289–298.

[19] Chen, H., Chen, J., Muir, L. a., Ronquist, S., Meixner, W., Ljungman, M., Ried, T., Smale, S., and Rajapakse, I. (2015) Functional organization of the human 4D Nucleome. *Proceedings of the National Academy of Sciences,* **112**(26), 8002–8007.

[20] Lieberman-aiden, E., Berkum, N. L. V., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science,* **326**, 289–294.

[21] Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology,* **11**(R106), 1–12.

[22] Chen, J., Hero, A., and Rajapakse, I. (2016) Spectral Identification of Topological Domains. *Bioinformatics,* **32**(14), 2151–2158.

[23] Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell,* **159**(7), 1665–1680.

[24] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2, 2015) Chromatin architecture reorganization during stem cell differentiation. *Nature,* **518**(7539), 331–336.

[25] Tang, C., Ang, B. T., and Pervaiz, S. (2007) Cancer stem cell: target for anti-cancer therapy. *The FASEB Journal,* **21**(14), 3777–3785.