4DNvestigator: Time Series Hi-C and RNA-seq Data Analysis Toolbox

Stephen Lindsly[1], Scott Ronquist[1], Sijia Liu[2], Samuel Dilworth[3], Michael Perlman[4], and Indika Rajapakse[1,5]

[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]iReprogram, Ann Arbor, MI 48105, USA
[4]Department of Statistics, University of Washington, Seattle, WA 98195, USA
[5]Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA

November 9, 2020

Abstract

Motivation: Comprehensive investigation of genome structure and function dynamics underlying cell phenotype produces vast amounts of high-dimensional, multilayered data. New methods are required to organize these data into an informative framework. Here we provide a comprehensive data analysis toolbox guided by theory of networks.

Results: 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). Our toolbox provides network-based data analysis tools such as single-layer/multilayer network centrality and network entropy to characterize the 4D Nucleome. The toolbox also contains statistical methods for comparing genomic structure at multiple genomic scales.

Availability: https://github.com/lindsly/4DNvestigator
Contact: indikar@umich.edu

Introduction

The genome is a dynamical system where changes in genome structure and function over time affect cell phenotype. The relationship between genome structure, gene transcription, and cellular phenotypes is referred to as the 4D Nucleome (4DN) (*1, 2*). To analyze the 4DN, genome-wide chromosome conformation capture (Hi-C) and RNA sequencing (RNA-seq) are used to observe genome structure and function, respectively. The availability and volume of Hi-C and RNA-seq data is expected to increase as high throughput sequencing costs decline, thus the development of methods to analyze these data is imperative.

The relationship of genome structure and function has been studied previously (*3–5*), yet comprehensive and accessible tools for 4DN analysis are underdeveloped. Unlike the aforementioned literature, 4DNvestigator provides a unified toolbox that loads time series Hi-C and RNA-seq data, extracts important structural and functional features, and conducts both established and novel 4DN data analysis methods. The 4DNvestigator also includes network-based approaches, such as von Neumann network entropy which provides an efficient way to track the amount of uncertainty of the dynamic genome, and multilayer network centrality which elucidates structural changes.

Methods

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 and RNA-seq file. The 4DNvestigator workflow is depicted in Fig. 1A, and a Getting Started document is provided to guide the user through the main functionalities of the 4DNvestigator.

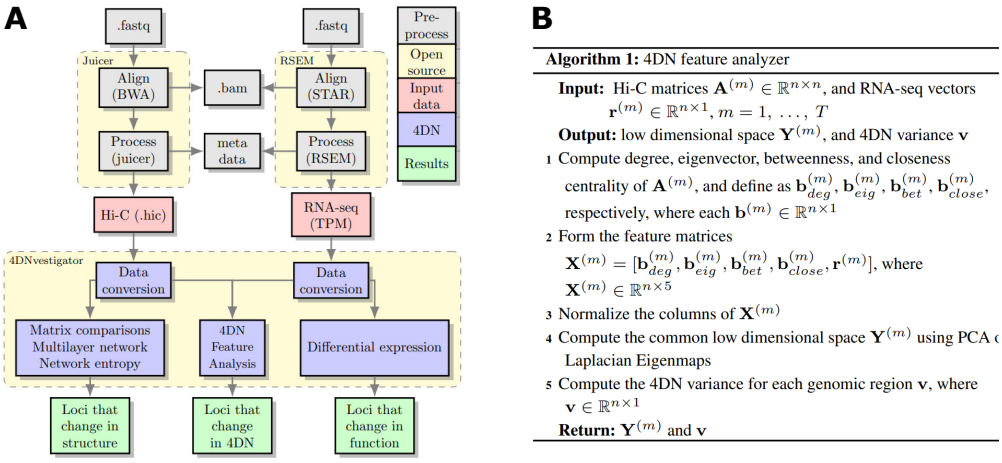

Figure 1: Overview of the 4DNvestigator. (**A**) Data processing pipeline for the 4DNvestigator. Here, 4DN refers to the 4DNvestigator. (**B**) Algorithm of the 4DN Feature Analyzer.

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 genomic regions are edges. We can then quantify the importance of each node using network centrality (*6*), which has been shown to reflect the dynamics of genome structure over time (*5*). Centrality measures and gene expression (RNA-seq) are then integrated to define the structure-function state of each genomic region at each time point (Fig. 1B).

Multilayer Networks

In addition to combining genome features over time, the 4DNvestigator enables modeling of dynamic genome structure as a multilayer network, where the inter-layers correspond to the snapshots of the network at different time points. Nodal degree is an important network centrality measure, but in a multilayer network, one needs to know the degree within a single layer and how the degree is distributed across different layers. In 4DNvestigator, we use the *multiplex participation coefficient* to evaluate the heterogeneity of nodal degrees in dynamic biological networks.

Network Entropy

Entropy measures the order within a system, where higher entropy corresponds to more disorder (*7*). The 4DNvestigator applies 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 and (*8, 9*) 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).

Additional 4DNvestigator Tools

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 (*3, 11*). Regions that change compartments between samples are automatically identified. The 4DNvestigator also utilizes developed MATLAB scripts for differential gene expression using established methods (*12*). In addition, we include a method for testing the equality of correlation matrices proposed by Larntz and Perlman (*13*). For more details on all methods, data, and analysis validation, see Supplementary Materials.

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. Multilayer network centrality and network entropy can also be applied to characterize the structural disorder in a region over multiple samples or time points.

Acknowledgements

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

Funding

This work is supported, in part, under Air Force Office of Scientific Research Award No: FA9550-18-1-0028 and Lifelong Learning Machines (L2M) program from DARPA/MTO.

References

1. Job Dekker, Andrew S. Belmont, Mitchell Guttman, Victor O. Leshyk, John T. Lis, Stavros Lomvardas, Leonid A. Mirny, Clodagh C. O'Shea, Peter J. Park, Bing Ren, Joan C. Ritland Politz, Jay Shendure, and Sheng Zhong. The 4D nucleome project. *Nature*, 549(7671):219–226, 2017.

2. Thomas Ried and Indika Rajapakse. The 4D Nucleome. *Methods*, 123:1–2, 2017.

3. 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, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S. Lander, and Job Dekker. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326:289–294, 2009.

4. Jesse R. Dixon, Inkyung Jung, Siddarth Selvaraj, Yin Shen, Jessica E. Antosiewicz-Bourget, Ah Young Lee, Zhen Ye, Audrey Kim, Nisha Rajagopal, Wei Xie, Yarui Diao, Jing Liang, Huimin Zhao, Victor V. Lobanenkov, Joseph R. Ecker, James A. Thomson, and Bing Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, 2 2015.

5. Sijia Liu, Haiming Chen, Scott Ronquist, Laura Seaman, Nicholas Ceglia, Walter Meixner, Pin-Yu Chen, Gerald Higgins, Pierre Baldi, Steve Smale, Alfred Hero, Lindsey A. Muir, and Indika Rajapakse. Genome Architecture Mediates Transcriptional Control of Human Myogenic Reprogramming. *iScience*, 6:232–246, 2018.

6. Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

7. Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

8. Ben D Macarthur and Ihor R Lemischka. Statistical mechanics of pluripotency. *Cell*, 154(3):484–489, 2013.

9. I. Rajapakse, M. Groudine, and M. Mesbahi. What can systems theory of networks offer to biology? *PLoS computational biology*, 8(6):e1002543, 2012.

10. Eran Meshorer and Tom Misteli. Chromatin in pluripotent embryonic stem cells and differentiation. *Nature reviews Molecular cell biology*, 7(7):540, 2006.

11. Jie Chen, Alfred Hero, and Indika Rajapakse. Spectral Identification of Topological Domains. *Bioinformatics*, 32(14):2151–2158, 2016.

12. Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(R106):1–12, 2010.

13. Kinley Larntz and Michael D. Perlman. A simple test for equality of correlation matrices. *Statistical Decision Theory and Related Topics IV*, pages 289–298, 1988.