HiC DOC: detecting and comparing genomic compartments

Cyril Kurylo, Sylvain Foissac, Matthias Zytnicki

HiC DOC: detecting and comparing genomic compartments

Cyril Kurylo[1], Sylvain Foissac[1] and Matthias Zytnicki[2]

[1] GenPhySE (INRAE), 24 Chemin de Borde Rouge, 31320, Auzeville Tolosane, France
[2] MIAT (INRAE), 24 Chemin de Borde Rouge, 31320, Auzeville Tolosane, France

Corresponding Author: **cyril.kurylo@inrae.fr**

Genomic compartmentalization is a biological factor affecting cell functionality. Different compartments can be observed in the nucleus of eukaryotic cells, grouping genomic regions into clusters. Active compartments are usually associated with open chromatin and gene expression while inactive compartments are usually associated with closed chromatin and gene repression [1]. Analysis of data produced by the Hi-C protocol reveals compartmentalization of chromatin in the nucleus, which can vary as a tissue develops. Today, existing methods to detect genomic compartmentalization are limited in at least one of the following ways: detecting compartments qualitatively with no confidence measure, ignoring experimental biases, and/or dismissing replicate variability.

We propose an improvement over existing methodology to detect compartments and compare compartmentalization between conditions. First, we properly correct the diverse biases inherent to Hi-C data, using cyclic loess normalization [2] to reduce technical biases and Knight-Ruiz matrix balancing [3] to mitigate biological biases. Then, we correct interaction counts with a loess regression to clearly expose the compartmentalization information captured by the data. Finally, we use an unsupervised learning method, constrained K-means [4], to computationally detect compartments from the normalized data. This method enables us to produce quantitative “concordance” values for each genomic region in each replicate, supporting our compartment predictions. Finally, we use these concordance values for differential analysis of compartmentalization between conditions. From their distributions, we obtain p-values revealing the significance of each predicted compartment change.

The method is implemented in an R package available at *github.com/mzytnicki/HiCDOC*, and is applied to Hi-C data originating from fetal pig muscles. Our data consists of three biological replicates at 90 days of gestation and three biological replicates at 110 days of gestation [5]. The detected compartment changes open a way towards a better understanding of neonatal mortality affecting piglets.

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, Richard Sandstrom, Bradley Bernstein, Michael 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):289–293, 2009.

[2] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments. *Bioinformatics*, (35):2916–2923, 2019.

[3] Philip A Knight, and Daniel Ruiz. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* (33): 1029–1047, 2012.

[4] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained K-means Clustering with Background Knowledge. In *18th International Conference on Machine Learning*, pages 577–584. ICML, 2001.

[5] Maria Marti-Marimon, Hervé Acloque, Matthias Zytnicki, Sarah Djebali Quelen, Nathalie Villa-Vialaneix, Ole Madsen, Yvette Lahbib Mansais, Diane Esquerre, Florence Mompart, Laurence Liaubet, Martien Groenen, Martine Bouissou-Matet Yerle, and Sylvain Foissac. Characterization of 3D genomic interactions in fetal pig muscle. In *36th International Society for Animal Genetics Conference*, page 43. ISAG, 2017.