HiCEnterprise: Identifying long range chromosomal contacts in HiC data

Hania Kranas[†*]

Irina Tuszynska[†]

Bartek Wilczynski[⋆]

Contact: bartek@mimuw.edu.pl
Faculty of Mathematics, Informatics and Mechanics,
University of Warsaw, Poland

[†] These authors contributed equally to this work.
[⋆] To whom correspondence should be addressed.
[*] Current address: Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain

May 23, 2019

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27753v1 | CC BY 4.0 Open Access | rec: 23 May 2019, publ: 23 May 2019

Abstract

Motivation: Computational analysis of chromosomal capture data is currently gaining popularity with the rapid advance in experimental techniques providing access to growing body of data. An important problem in this area is the identification of long-range contacts between distinct chromatin regions. Such loops were shown to exist at different scales, either mediating interactions between enhancers and promoters or providing much longer interactions between functionally interacting distant chromosome domains. A proper statistical analysis is crucial for accurate identification of such interactions from experimental data.

Results: We present HiCEnterprise, a software tool for identification of long-range chromatin contacts. It implements three different statistical tests for identification of significant contacts at different scales as well as necessary functions for input, output and visualization of chromosome contact matrices.

Availability: The software and its documentation is available at: `https://github.com/regulomics/HiCEnterprise`

Contact: `bartek@mimuw.edu.pl`

Supplementary information: Supplementary data are available at the website `http://regulomics.mimuw.edu.pl/wp/hicenterprise`

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27753v1 | CC BY 4.0 Open Access | rec: 23 May 2019, publ: 23 May 2019

Chromosomes in eukaryotic cells are very complex polymers that function in a very tightly packed 3D environment of the cell nucleus (Sazer and Schiessel, 2018). The packing of chromosomes is at the same time dynamic and visibly different between cells in the same population, yet its conformation is proven to be non-random to allow for efficient activation and repression of subsets of genes defined by the dynamic epigenetic state of the cell (Spector, 2003). Scientists have been interested in studying the rules governing the chromatin structure and its dynamics for a long time, however we were mostly limited to theoretical studies based on very limited imaging data until the development of the Hi-C technique (Lieberman-Aiden *et al.*, 2009). Since then, the body of data from Hi-C experiments is quickly growing, allowing us to answer more questions related to chromosome structure and its relation to gene regulation. In particular, the question of identifying chromosomal contacts have been studied both on the level of enhancer-promoter interactions (Won *et al.*, 2016) as well as domain-to-domain interactions (Niskanen *et al.*, 2017). Since more researchers are interested in identifying such contacts, it may be helpful for them to have a readily available implementation of previously published statistical methods.

HiCEnterprise is a package consisting of two types of contact analyses - between regions and between domains. The first part of the package can be used for identification of long-range contacts between small (1-3 bin) regions. It is an implementation of the method for identification of point interactions and creating interaction profiles based on Hi-C data as introduced by Won *et al.* (2016). As a Hi-C cis-contact map should be symmetrical, interaction profile for a region located in a particular bin is obtained by extracting intensities only horizontally, from the left and right of this bins positions on the diagonal. Significant contacts between bins are identified as enrichments under background distribution (fitted Weibull distribution matched by chromosome and distance (Won *et al.*, 2016)). FDR (False Discovery Rate) calculated with a Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) is used to correct for multiple testing.

It may be difficult to distinguish between true interactions and noise, so the program allows to provide additional evidence for the validity of such interactions by including analysis of biological replicates. If several maps (replicates) are provided on input, interactions are considered significant only if their FDR value is above the selected threshold in all replicates. In the example interaction profile plot

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27753v1 | CC BY 4.0 Open Access | rec: 23 May 2019, publ: 23 May 2019

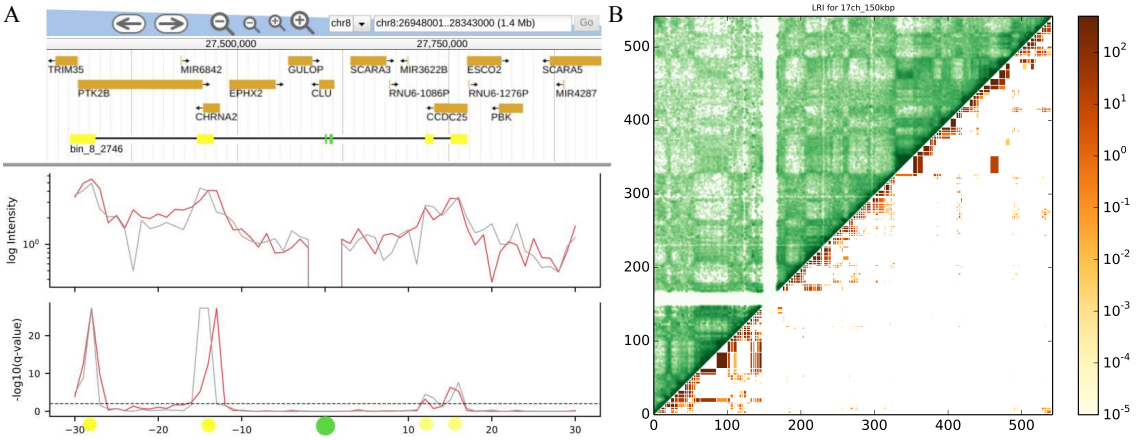

Figure 1: Visualization of long-range enhancer-promoter contacts (a) and domain-domain contacts (b).

(Figure 1a) one can see the analysis for two Fetal Brain maps (Won *et al.*, 2016) at 10kb resolution, at 50 bin distance each way. Figure 1a shows, top to bottom: a Jbrowse (Skinner *et al.*, 2009) screenshot with genes and interaction profile representation (initial regions in green, contact predictions in yellow); interaction profile plot with intensities (weighted by distance) and $-log_{10}$ of FDR corrected P-values (Q-values) with a threshold set at 0.01. The results shown are consistent with the results from the original paper (Won *et al.*, 2016).

To run the point interactions analysis it is required that the user provides at least one Hi-C chromosome map in numpy format and BED file with coordinates of regions to extract interaction profiles. Since this mode will be likely frequently used to scan for enhancer related interactions, we have provided also a mode, where the user gives an EnhancerAtlas (Gao *et al.*, 2016) FASTA file as input. It is possible to plot the results either with matplotlib (Hunter, 2007) or rpy2 (Gautier, 2012). Output files with contacting regions are available in three formats: txt, BED-like and GFF-like. Remapping between assemblies is possible for BED and GFF files (pyliftover package is required).

The other option of HiCEnterprise is a calculation of long-range interactions (LRI) significance scores for pairs of topological domains in Hi-C contact maps as used in (Niskanen *et al.*, 2017). In this mode the user needs to define borders of topological domains obtained using external software, (e.g. HOMER (Heinz *et al.*, 2010); see (Dali and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27753v1 | CC BY 4.0 Open Access | rec: 23 May 2019, publ: 23 May 2019

Blanchette, 2017) for more examples). As the first step of this method, the new matrix M (with the shape of NxN where N is the number of domains in the chromosome considered) is calculated. M[i,j] represents the total number of Hi-C contacts calculated for the pair of domains i and j. Next, for each pair of domains in the new matrix, a P-value is calculated based on the hypergeometric, Poisson or negative binomial test. Our software calculates the parameters of the chosen distribution based on the data observed in the actual Hi-C matrix and calculates a p-value for enrichment under the null model. These p-values are again converted to FDR Q-values to account for multiple hypothesis testing.

HiCEnterprise domains mode can be run with different options. Some of them are necessary to run the program: the chromosome number, the Hi-C map, the resolution of Hi-C map, the file with domains borders information and the domain level if a hierarchical TAD caller like Sherpa (Krolak and Wilczynski, 2012) was used to determine the domains borders. Other arguments are optional. It is possible to modify the threshold of the q-values that will be considered as significant and returned in the output file (default value is 0.01). By default, the resulting p- and q-values are written to two text files. The user can also choose to make a plot with results as a contact map, with significant contacts between domains highlighted with color (as seen in Figure 1b). One can also change colors of the contact maps and/or interactions on the figure, propose the title or change the distance between ticks on the generated figures.

In summary, the software we have presented is a flexible tool for users interested in identifying interacting loci based on the Hi-C experiments. We provide two different functionalities (bin-to-bin and domain-to-domain contact identification) with several statistical tests that were already shown to be appropriate for each of the scenarios (Weibull distribution for the first and hypergeometric, Poisson and negative-binomial for the second). By providing a tested and easy to use implementation, we hope to make it easier for experimentalists to use these methods without the need to implement them on their own.

Funding

This work has been supported by the Polish National Science Center Grant decision number [DEC 2015/16/W/NZ2/00314].

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Dali, R. and Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic acids research*, **45**(6), 2994–3005.

Gao, T., He, B., Liu, S., Zhu, H., Tan, K., and Qian, J. (2016). EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics (Oxford, England)*, **32**(23), 3543–3551.

Gautier, L. (2012). rpy2: A simple and efficient access to r from python, 2012. *http://rpy. sourceforge. net/rpy2. html*.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, **38**(4), 576–589.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, **9**(3), 90–95.

Krolak, K. and Wilczynski, B. (2012). Sherpa: Simple hierarchical profile aggregation. *https://github.com/regulomics/sherpa*.

Lieberman-Aiden, E., Van Berkum, N. L., 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**(5950), 289–293.

Niskanen, H., Tuszynska, I., Zaborowski, R., Heinäniemi, M., Ylä-Herttuala, S., Wilczynski, B., and Kaikkonen, M. U. (2017). Endothelial cell differentiation is encompassed by changes in long range interactions between inactive chromatin regions. *Nucleic acids research*.

Sazer, S. and Schiessel, H. (2018). The biology and polymer physics underlying large-scale chromosome organization. *Traffic*, **19**(2), 87–104.

Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J., and Holmes, I. H. (2009). JBrowse: a next-generation genome browser. *Genome Research*, **19**(9), 1630–1638.

Spector, D. L. (2003). The dynamics of chromosome organization and gene regulation. *Annual review of biochemistry*, **72**(1), 573–608.

Won, H., de La Torre-Ubieta, L., Stein, J. L., Parikshak, N. N., Huang, J., Opland, C. K., Gandal, M. J., Sutton, G. J., Hormozdiari, F., Lu, D., *et al.* (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27753v1 | CC BY 4.0 Open Access | rec: 23 May 2019, publ: 23 May 2019