DiADeM: differential analysis via dependency modelling of chromatin interactions with generalized linear models

Rafał Zaborowski[1], Bartek Wilczyński[1*]

1 Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland

* bartek@mimuw.edu.pl

Abstract

High throughput Chromosome Conformation Capture experiments have become the standard technique to assess the structure and dynamics of chromosomes in living cells. As any other sufficiently advanced biochemical technique, Hi-C datasets are complex and contain multiple documented biases, with the main ones being the non-uniform read coverage and the decay of contact coverage with distance. Both of these effects have been studied and there are published methods that are able to normalize different Hi-C data to mitigate these biases to some extent. It is crucial that this is done properly, or otherwise the results of any comparative analysis of two or more Hi-C experiments is bound to be biased. In this paper we study both mentioned biases present in the Hi-C data and show that normalization techniques aimed at alleviating the coverage bias are at the same time exacerbating the problems with contact decay bias. We also postulate that it is possible to use generalized linear models to directly compare non-normalized data an that it is giving better results in identification of differential contacts between Hi-C matrices than using the normalized data.

Introduction

Eukaryotic chromosomes are very tightly packed biopolymers in the space of the nucleus. Despite the ultra-dense packing, their biological functions such as the ability to replicate and to express genes, depends on reproducible structural dynamics [Rada-Iglesias et al., 2018]. Importantly, their structure is not necessarily completely reproduced between the cells in the same epigenetic state, however, certain contacts between parts of the chromosomes, such as the enhancer-promoter contacts or insulator-mediated loops are essential to ensure proper gene regulation [Eagen, 2018]. Chromosome Conformation Capture and related techniques have been developed in the last years in order to give the researchers the ability to measure the frequency of such contacts and make inferences about the regulatory relationships based on the relative contact frequencies [Dekker et al., 2002]. In particular, the development of the hi-throughput chromosome conformation capture technique (Hi-C) has allowed us to measure genome-wide statistical chromosome contact patterns from large populations of cells [Lieberman-Aiden et al., 2009]. The introduction of this technique a decade ago, and its subsequent improvements, gave us now access to multiple datasets containing measured average frequencies of contacts across wide variety of cell types in multiple species [Schmitt et al., 2016, Rudan et al., 2015].

While the importance of the Hi-C technique for studying chromosome structure and dynamics should not be underestimated, as with any sufficiently complex biochemical protocol, we need to take into account the natural biases of the method when analysing the data [Yaffe and Tanay, 2011]. Especially when one is attempting to make statements comparing two or more Hi-C datasets originating from different cell populations, one needs to be very cautious whether the inferred conclusions are not due to any intrinsic biases of the measurements itself. This is especially important in the case of Hi-C data, where because of the very high cost of obtaining the data, researchers are frequently facing the necessity to make comparative analyses without sufficient number of replicate experiments.

There are two very well documented biases in the Hi-C data: the non-uniform coverage, that needs to be removed if relative statements are to be made comparing contact frequencies between different regions in the same data; and the contact decay bias, that gives rise to the observed decrease in contact frequency of regions that are distant in the linear space of the chromosome [Imakaev et al., 2012, Nagano et al., 2015]. The first bias is usually taken care of by the standard normalization techniques, however the second one is usually considered to be a biological phenomenon and not accounted for in the normalization. However, when comparing different Hi-C datasets one can clearly see that even in datasets coming from seemingly the same type of cells, significant differences in the decay rate can be observed between experiments, that cannot be explained by the biological differences between the studied cell populations. In such cases, when one is interested in identifying differential contacts between Hi-C maps, the results of the comparative analyses can be severely and systematically biased by

the difference in decay rates.

In this paper we take on the problem of identifying and measuring the effects of both biases and whether the currently available normalization methods for Hi-C data are adequate for countering the observed biases in the data. We also discuss possible solutions to the problems identified with current methods.

Results

In this section we report the main results of our study. We start by describing the coverage and contact decay effects and explain how they affect differential analysis of Hi-C data. Then we show that existing Hi-C normalization methods fail to account for all issues resulting from differences in coverage and contact decay. Finally we present that raw Hi-C interaction patterns at given genomic distance are highly correlated even between different cell line datatsets. This observation enables to construct a linear background model allowing for discovery of local changes in contact intensity by testing for deviations from the expected pattern. Finally we introduce, a simple algorithm for detection of long range differentially interacting regions.

Dominating signals in Hi-C data

There are 2 factors, which exhibit significant variation across Hi-C datasets: the coverage and contact decay. The variation is particularly large when samples were derived from different studies. However, significant shift in either of these two signals may be observed even during comparison of putative technical replicates of single sample (see for instance IMR90 replicates from GSE63525 study - supplementary fig. 1, 2) [Rao et al., 2014]. In perfect scenario we would like both datasets to have both coverage and contact decay constant (in terms of distribution). This would allow to detect local differences in the number of interactions between pairs of chromatin regions, which is usually the purpose of Hi-C differential analysis. In absence of the above situation there is no straightforward way for finding local contact differences. Below we demonstrate the influence of coverage and contact decay on Hi-C differential analysis.

Hi-C coverage bias

Like any sequencing based protocol an outcome of Hi-C experiment depends on the sequencing depth. Higher the number of reads, larger signal to noise ratio may be achieved. However often reasearcher is forced to compare Hi-C datasets obtained from different studies, which usually implicates diverging coverages of corresponding contact maps (fig. 1a). Because of this direct comparisons would be severely biased leading to enrichment of most cells in higher coverage matrix (supplementary fig. 3). To avoid this problem contact matrices are usually normalized before conducting differential analysis. Among many normalization methods the most commonly used seems to be the Iterative Correction approach

based on assumption of uniform visibillity of every region [Imakaev et al., 2012]. This method is largely effective in removing coverage bias (fig. 1b). Unfortunately, as we point out in the next section such normalization may even amplify issues stemming from other biases during Hi-C differential analysis. Most other normalization methods are focusing on this effect ignoring the contact decay.

Figure 1. Coverages of human IMR90 and MSC cell Hi-C data, chromosome 19. a) Raw data. **b)** ICE normalized.

Hi-C contact decay bias

The contact decay refers to the rate of change in number of interactions between a pair of chromatin regions as a function of their linear separation on the chromosome. Usually it is measured as the mean number of interactions between all pairs of loci at given distance (see fig. 3 and Materials and methods for details). Intuitively this relationship reflects chromatin compaction and is very pronounced in Hi-C data (fig. 2 top). Such effect masks virtually any local differences allowing to only compare global compaction of chromatin. Iterative correction method mentioned in previous section is not particularly helpful this time as it does not remove contact decays difference (fig. 2 center and supplementary fig. 4). What is worse different normalization methods may yield opposing patterns of relative contact decays between compared conditions (fig. 2 bottom and supplementary fig. 4, 5). As can be seen in fig. 2 initial relationship between human IMR90 and MSC contact decays exhibit much stronger signal

for IMR90 data set at almost entire range. Upon normalization this pattern changes significantly. Depending on which method is used MSC contact intensity starts to exceed IMR90 at around 80th (HiCnorm) or 6th (ICE) diagonal [Hu et al., 2012]. Both these methods show consistent behaviour for higher ranges of contact decays (fig. 2, third column, rows 2 and 3) however they diverge significantly at close genomic distances (fig. 2, second column, rows 2 and 3), where the most reliable interaction data is collected.

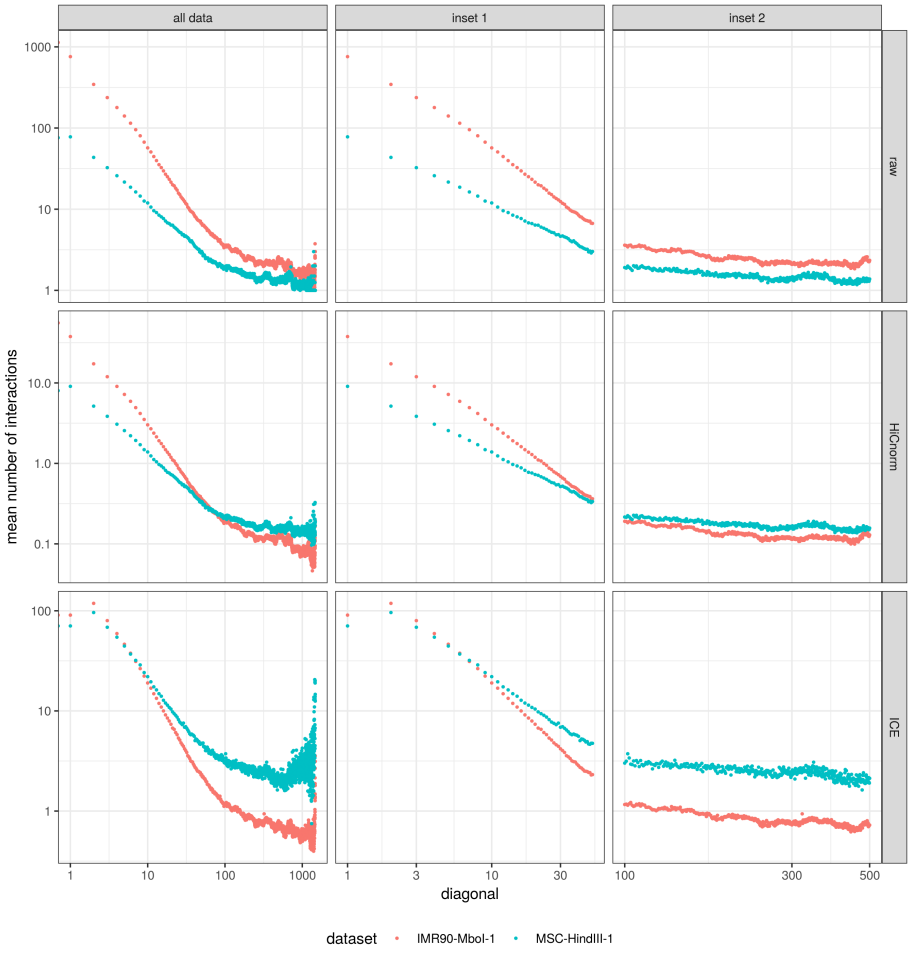

Figure 2. Contact decays of human IMR90 and MSC cell Hi-C data, chromosome 19. Top row - raw data, middle row - HiCnorm normalized, bottom row - ICE normalized.

Correlations between raw contact intensities

When analyzing different Hi-C matrices we discovered that relationship between raw interaction intensities of two Hi-C datasets at given loci separation exhibit approximately linear behaviour with non constant (funnel like) variance (supplementary fig. 6, 7, 8). Despite different coverages Pearson correlations between contact intensities are evident and statistically significant for pairs of regions separated by up to few megabases (fig. 3a). Surprisingly this trend is not limited to replicates (supplementary fig. 9, 10), but it is also pronounced between different study datasets, being much weaker, but still highly significant (fig. 3b). The maximum separation of interacting regions, which preserves statistical significance between interaction intensities varies, typically ranging from 4 to 8 or even 16 Mbp. The precise value of separation threshold depends on the dataset quality (notably sequencing depth) and the chromosome size (supplementary table 1). We realize that maximum separation distance suggested by us comprise minor fraction of Hi-C decay range however we argue that more distant contacts are severly affected by noise and will be difficult to quantify for differential interactions anyway. This is also reflected by the fact that Hi-C data is known to follow power law distribution and the cumulative number of contacts at as little as 10 percent of diagonals comprise from 60 to even 90 percent of all contacts within the entire chromosome (supplementary table 2) [Lieberman-Aiden et al., 2009].

Modelling interaction dependency using GLM

As described in previous section we observe statistically significant linear relationship between interaction intensity of two contact maps. However as can be seen in fig. 4a and 4b, small fraction of observations disobey linear pattern. We hypothesise such pairs of regions to be potential differential interactions. Due to apparent heteroscedasticity we can't use ordinary least squares regression to model the dependency. Instead we decided to use Negative Binomial regression (see Materials and methods), which is known to perform well in analysis of overdispersed count data in general and in NGS data in particular [Hilbe, 2011, Robinson et al., 2010, McCarthy et al., 2012, Love et al., 2014]. Our model aims to capture the relationship between the number of contacts in contact map A (predictor) and number of contacts in contact map B (response) for every pair of regions given their genomic separation. Additionally, as we expect a fraction of observations to be outliers (differential interactions), we first apply a simple procedure to filter them out. Filtering process is based on performing robust regression and removing points with zero weight (see Materials and methods). When the model is estimated the significance of observation enrichment can be calculated based on conditional distribution of the response given predictor (supplementary fig. 11-16). Lastly, p-values are adjusted for multiple hypothesis testing using Benjamini-Hochberg method. The summary of estimation procedure is illustrated in fig. 4. The whole process is repeated for every diagonal from prespecified range. Results can be visualised using significance matrix

Figure 3. Correlations between corresponding decays of human IMR90 and MSC cell Hi-C raw data, chromosome 19. a) Pearson, Spearman and Kendall correlations as a function of diagonal. **b)** Significances (negative log p-value) of correlations.

where each cell express how likely given pair of regions is enriched in interactions in response Hi-C map. As there are 2 ways of choosing predictor and response the upper and lower triangles of the significance matrix correponds to each of them respectively (fig. 5c).

As there is no golden standard in detection of Hi-C differential interactions and thus there is no dataset with "true" differential chromatin contacts we decided to check the performance of our method on a pair of IMR90 bootstrap and IMR90 biological replicate maps. Both of this datasets are expected to exhibit much less interaction difference as compared with different cell lines data. The results are presented in figure 5a-c and supplementary figures 17-20. The comparison of p-value distribution as well as visual inspection of significance maps provides evidence of valid model.

Detection of long range differentially interacting regions

As our testing is done on semi-diagonal separately we further reasoned that to increase the confidence of differential analysis it is straight forward to look for aggregates of significant differentially interacting cells rather than individual isolated ones. Such analysis seems to be more robust against noisy data as large

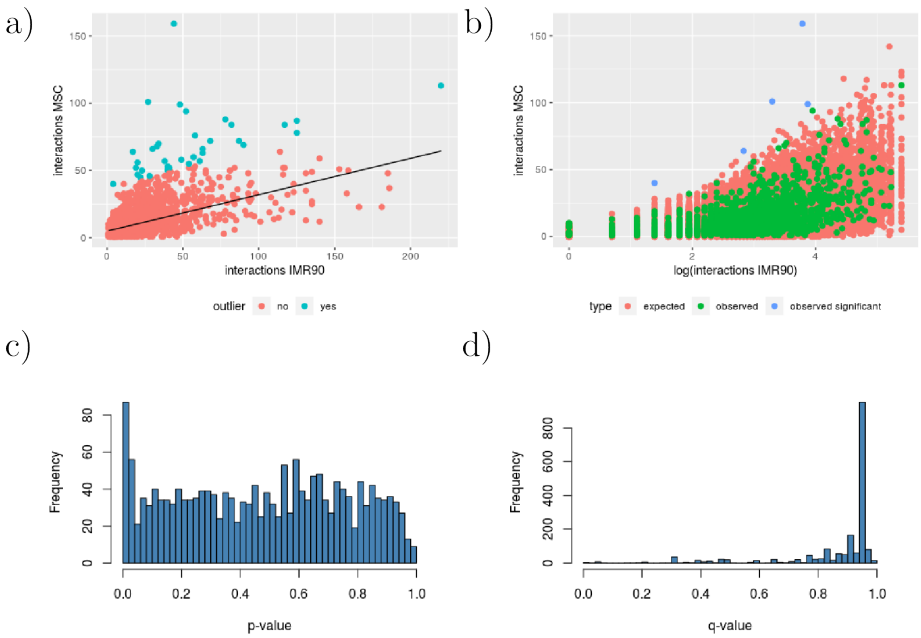

Figure 4. Model construction procedure. Steps a-d are performed for every diagonal from prespecified range. **a)** Robust regression on paired interactions set allows to determine outliers (potential differential interactions) and exclude them. **b)** The Negative Binomial regression is fitted to remaining data and used as expected model to calculate enrichment p-value for each paired interaction. **c)** Enrichment p-value distribution indicate deviation from null model. **d)** Finally p-values are adjusted using Benjamini-Hochberg method. Resulting q-values are used to draw significance maps.

group of significant neighboring cells is more likely to reflect biologically relevant differential interaction as opposed to individual cell, especially for small bin sizes (high resolution). For that purpose we propose a 2 step procedure. During first step, the q-value threshold is determined, so interactions can be labelled as either significant or not. A simple method is to let the researcher specify significance threshold. Alternative approach would be an adaptive method, which uses bilinear model fit. Fitting bilinear model is based on the idea that most interactions are insignificant and therefore their negative log q-values should be small and almost constant. On the other hand we might expect a fraction of significantly enriched interactions, with their negative log q-values following a steep linear trend (see supplementary fig. 21). In the second step significant cells are used to construct incidence matrix, which is then scanned for connected components, i.e. groups of neighboring significantly enriched cells (see Materials and methods). When algorithm completes the connected components search, a list containing smallest rectangle-like regions including significantly enriched interactions (LRDI) is returned. Resulting list may be filtered to keep only the

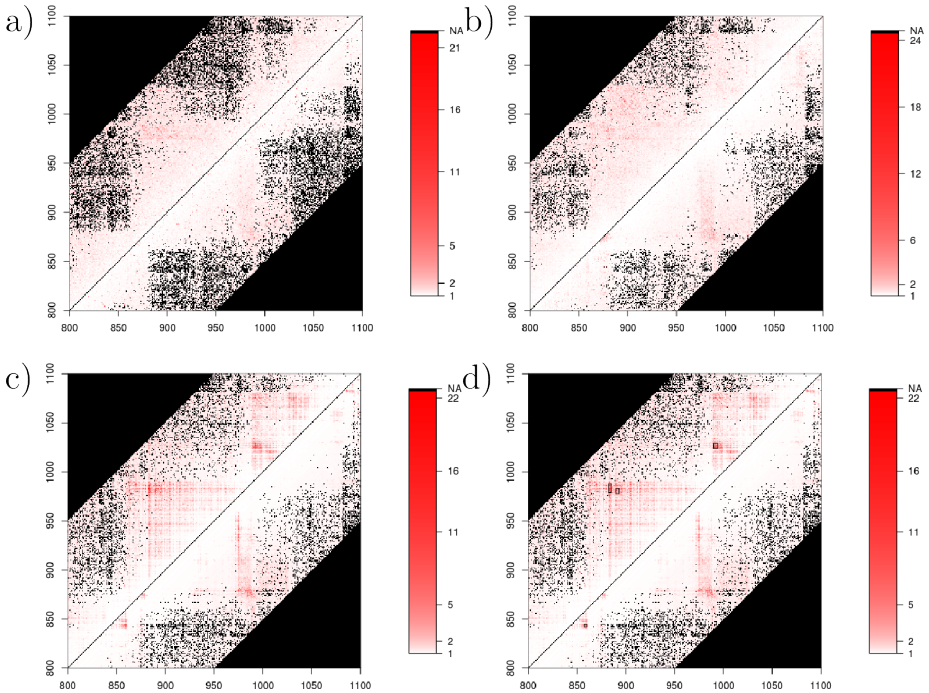

Figure 5. Significance maps. Every cell contain negative log of adjusted p-value (q-value). Enrichment of second condition with respect to first is shown on upper triangle and the opposite on bottom triangle. **a)** IMR90 bootstrapped data 1 vs 2. **b)** IMR90 biological replicates 1 vs 2. **c)** IMR90 vs MSC. **d)** IMR90 vs MSC with LRDI regions of size at least 10.

largest LRDIs (see fig. 5d).

To asses the performance of LRDI detection method we compared the LRDI size distribution of different cell lines dataset to that from pairs of biological replicates and pairs of bootstrapped datasets. The results (fig. 6) indicate that not only the number of LRDI is higher for the dataset with different cell lines, but also detected LRDIs correspond much larger regions.

Discussion and Conclusions

In this paper we have shown several analyses indicating that currently employed normalization techniques can affect the results of comparative analyses of multiple Hi-C datasets. What is even more worrying, the results of such analyses can be biased systematically and the direction of the bias may depend on the particular normalization technique used giving researchers opposite results with respect to the number of enriched interactions. While this may sometimes be a minor issue (e.g. when researchers do not analyze data for differential contacts), there are situations in which there is no clear answer to the question of which normalization

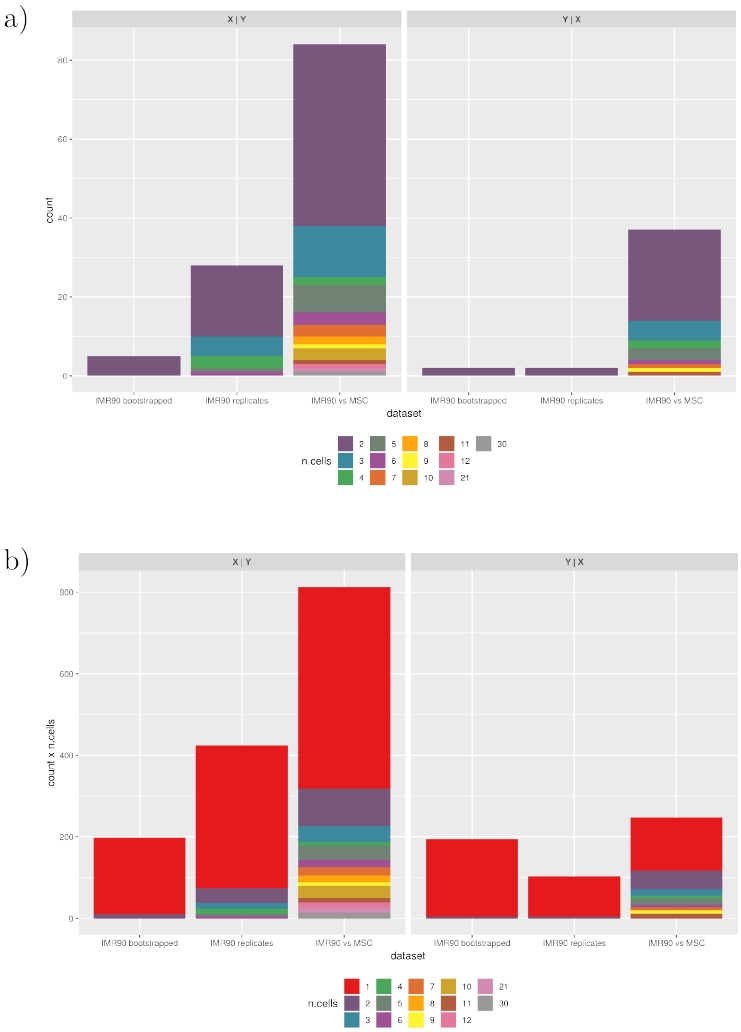

Figure 6. Comparison of LRDI regions size distribution. a) Each bar represents the number of LRDI regions of given size (i.e. including the number of significantly enriched cells) in 3 pairs of datasets: IMR90 bootstrap, IMR90 biological replicates and IMR90 vs MSC. LRDI regions of size 1 were removed for better visibility. **b)** Similar as in a), but now y axis indicates the number of LRDI regions of given size times the size of that LRDI region.

method is giving better results.

The natural choice in this situation is to go back to the raw, non-normalized data and see if we can get the answer from the state of the data before normalization. As we have shown, in typical cases we see a very good correlation between measurements of semi-diagonals in datasets before normalization that deteriorates after applying standard coverage uniformization. Even though this

signal is only visible in the range of relatively close interactions, it happens to be the range where we are most interested in accurate prediction of differential contacts. This suggests, that the approach that might be productive would be to attempt to compare data without the uniform coverage normalization.

Since we still need to have some sort of mapping between the raw counts that would take into account the global effects such as the total read count and the observable variance differences, we chose to use the generalized linear models approach to represent the relationship between counts found in the compared Hi-C experiments for each semi-diagonal of the compared maps. This allowed us to have a relatively flexible model, that enables for representation of the heteroscedacticity of the data and the expected contact decay with increased distance while avoiding the deterioration of the signals stemming from over-normalization of the coverages.

Using this model, we are able to not only identify regions that are significantly different in their contact frequency, but also aggregate them into larger regions that are likely to coincide with biologically relevant events. We show that the proposed approach gives much more differential contacts in comparison of biologically different samples that betwen replicates and that the detected differential regions are also significantly larger. We have implemented this approach in the software tool called DIADEM and made it available to the public.

In summary, we consider the results very promising. We hope that the fact that we documented the significant side-effects of the widely used normalization techniques will lead to better understanding of the results of Hi-C analyses. Even more importantly, we hope that the solution to the identified problem that we provide will lead to wide adoption of the methods we propose and in turn to more unbiased results in future Hi-C analyses.

Materials and methods

Software availability

The software is freely available as R package DIADEM under MIT license at: https://github.com/rz6/DIADEM.

Data

The Hi-C data used in this study comes from publicly available GEO repositories, studies: GSE63525, GSE52457, GSE87112 and GSE96107.

Definitions

Throughout the article we use following notation:

- Hi-C contact maps A, B:

$$A \in \mathbb{Z}_{\geq 0}^{n,n},\ A^T = A,\ a_{ij} \coloneqq (A)_{ij}$$

$$B \in \mathbb{Z}_{\geq 0}^{n,n},\ B^T = B,\ b_{ij} \coloneqq (B)_{ij}$$

- k diagonal elements of matrices A, B:

$$A_k = \{a_{ij} \mid k = |i-j|,\ a_{ij} \neq 0\}$$

$$B_k = \{b_{ij} \mid k = |i-j|,\ b_{ij} \neq 0\}$$

- paired interactions set $\mathcal{D}_k$ of matrices A, B at distance (diagonal) k:

$$\mathcal{D}_k = \{(a_{ij}, b_{ij}) \mid k = |i-j|,\ a_{ij} \neq 0,\ b_{ij} \neq 0\}$$

Coverage calculation

The coverage of contact map X is defined as a sum of contacts across bins:

$$s(i) = \sum_{j=1}^{N} x_{ij}$$

Contact decay calculation

The contact decay of contact map X is defined as a mean of contacts across diagonals:

$$d(k) = \frac{1}{|X_k|} \cdot \sum_{x_{ij} \in X_k} x_{ij}$$

Diagonal correlations calculation

Correlations are calculated based on paired interactions set:

$$\forall_{1\ \leq i \leq |\mathcal{D}_k|}\ (x_i, y_i) \in \mathcal{D}_k$$

Then $r(\mathbf{x}, \mathbf{y})$, $\rho(\mathbf{x}, \mathbf{y})$ and $\tau(\mathbf{x}, \mathbf{y})$ denote Pearson r, Spearman rho and Kendall tau correlation coefficients respectively.

Model estimation

Given paired interactions set $\mathcal{D}_k$, we define our model $\mathcal{M}_k$ as:

$$Y_i | x_i \sim NB(\beta x_i, \theta) \text{ s.t. } (x_i, y_i) \in \mathcal{D}_k \tag{1}$$

and estimate model parameters β and θ using glm function of MASS package [Venables and Ripley, 2002]. Prior to fitting Negative Binomial regression we fit robust linear model using iteratively reweighted least squares (IRLS), with bisquare psi function and remove paired interactions with zero weight (putative outliers). The process is performed with package robustreg [Johnson, 2017].

Significance calculation

Given paired interaction $(x_i, y_i) \in \mathcal{D}_k$ and model $\mathcal{M}_k$ we calculate probability of seeing a paired interaction at least as extreme as the observed one according to formula:

$$p_i = 1 - P(Y < y_i), \text{ where: } Y \sim NB(\beta_i x_i, \theta) \tag{2}$$

The p-values for model $\mathcal{M}_k$ are then adjusted using Benjamini-Hochberg procedure.

Detection of LRDI

Detection of LRDI regions starts with determination of significance threshold. For that purpose cells of significance matrix (negative log q-values) are arranged into vector in ascending order. Then every index of vector is used as split point and both resulting vectors are used to fit 2 linear models. The threshold is determined based on split which maximizes the overall fit (in terms of minimizing MSE).

In the next step significance map is converted to incidence matrix by substituting cells above threshold with one and all other with zero. The resulting matrix is scanned for connected components using raster package [Hijmans, 2018].

References

Dekker et al., 2002. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *science*, 295(5558):1306–1311.

Eagen, 2018. Eagen, K. P. (2018). Principles of chromosome architecture revealed by hi-c. *Trends in biochemical sciences*, 43(6):469–478.

Hijmans, 2018. Hijmans, R. J. (2018). *raster: Geographic Data Analysis and Modeling*. R package version 2.8-4.

Hilbe, 2011. Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hu et al., 2012. Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). Hicnorm: removing biases in hi-c data via poisson regression. *Bioinformatics*, 28(23):3131–3133.

Imakaev et al., 2012. Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999.

Johnson, 2017. Johnson, I. M. (2017). *robustreg: Robust Regression Functions*. R package version 0.1-10.

Lieberman-Aiden et al., 2009. 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.

Love et al., 2014. Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550.

McCarthy et al., 2012. McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297.

Nagano et al., 2015. Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B.-M., Wingett, S. W., and Fraser, P. (2015). Comparison of hi-c results using in-solution versus in-nucleus ligation. *Genome biology*, 16(1):175.

Rada-Iglesias et al., 2018. Rada-Iglesias, A., Grosveld, F. G., and Papantonis, A. (2018). Forces driving the three-dimensional folding of eukaryotic genomes. *Molecular systems biology*, 14(6):e8214.

Rao et al., 2014. 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.

Robinson et al., 2010. Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rudan et al., 2015. Rudan, M. V., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., and Hadjur, S. (2015). Comparative hi-c reveals that ctcf underlies evolution of chromosomal domain architecture. *Cell reports*, 10(8):1297–1309.

Schmitt et al., 2016. Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., et al. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell reports*, 17(8):2042–2059.

Venables and Ripley, 2002. Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Yaffe and Tanay, 2011. Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11):1059.