Interplay between chromatin contact frequency and expression levels in *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens* genomes

M. D. Samborskaya

Faculty of bioengineering and bioinformatics, Lomonosov Moscow State University, 119991, Moscow, GSP-1, Leninskiye Gory, MSU, 1-73, margarita.samborskaya@gmail.com

E.E. Khrameeva

Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, Building 3, Moscow, 143026, ekhrameeva@gmail.com

A. A. Mironov

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia, mironov@bioinf.fbb.msu.ru

M. S. Gelfand

Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute),

Bolshoy Karetny per. 19, build.1, Moscow 127051 Russia, mikhail.gelfand@gmail.com

The conformation of chromatin in the nucleus plays an important role in many cellular processes, including the regulation of gene transcription and DNA replication. Regulation of gene expression often involves long-range chromatin interactions between regulatory elements. Therefore the spatial organization of chromatin could lend insight into these complex regulatory processes.

The Hi-C method is most commonly used for genome-wide chromosome conformation capture. It enables the interrogation of all loci at once by combining DNA proximity ligation with high-throughput sequencing [1]. Its genome-wide fashion provides insight into the spatial organization of chromatin on a global scale.

However, data obtained by Hi-C have both technical and experiment-induced biases. This results in different regions of the genome having different experimental ”visibility”, which may cause systematic errors. ICE (iterative correction and eigenvector decomposition) is a commonly used method of iterative correction used for elimination of these systematic biases. While ICE is based on the assumption that all loci should have equal ”visibility”, the differences in experimental ”visibility” may be explained not only by technical biases, but also by biological factors such as characteristics of distinct chromatin states. Therefore, elimination of the differences in ”visibility” of chromatin regions could lead to loss of biologically meaningful data. Here, we investigate the interplay between the spatial organization of chromatin and gene expression levels using genome-wide chromatin data sets. We show that cumulative contact frequency is correlated with active transcription rates estimated by enrichment in histone modifications.

To understand the relation between contact frequency and expression levels, we performed computational analysis of 1Mb resolution Hi-C data sets for mouse, drosophila and human genomes before iterative correction [2]. Cumulative contact frequency was obtained by summarizing contact frequencies of each loci. Expression levels were estimated for drosophila [3] and human [4] genomes by chromatin states, combinatorial patterns of chromatin marks. Each state corresponds to a distinct biological enrichment in one type of regulatory element. This approach is more precise than use of individual mark occurrence, as it gives a broader view of regions' expression characteristics, and simplifies data analysis. We used the 15 states described in [4]. Computational analysis was performed on four human cell lines: HMEC, NHEK, HUVEC and K562, mouse cell line CH12-LX and drosophila cell line S2.

As active transcription requires the binding of RNA polymerase and different transcription factors, gene expression demands loose packaging. Tightly packaged regions should, on the contrary, exhibit low gene expression levels. We combined a whole-genome Hi-C map with state percentages' distributions for each region and observed that regions showing high cumulative contact frequencies are more enriched in states corresponding to active transcription than regions showing low contact frequencies, which is at odds with the initial hypothesis. To get a more precise estimate, we visualized the growth of state percentages corresponding to active transcription with increase in cumulative contact frequency. We calculated the correlations between each of the state vectors and the cumulative contact frequency vector to produce a correlation pattern, which enables comparative analysis of different genome regions. The homogeneity of the correlations across the genome was proved with the Stereogene tool [5].

The correlation pattern discussed above displays only the relation between cumulative contact frequency and chromatin states, but not the possible causes for this relation. The dependencies may possibly be explained by some confounding factor interrelated with cumulative contact frequency. If this were the case, then accounting for the influence of this confounding factor would diminish the correlations seen on the pattern. In our search for confounding factors we found that GC content and chromosome length are interrelated with contact frequency. Intrachromosomal contact frequency declines with increase in chromosome length as small chromosomes are more compact while large ones are loosely packaged. Interchromosomal contact frequency increases with chromosome length, which indicates that small chromosomes are spatially isolated and tend to make less interchromosomal contacts than large chromosomes do. However, normalization by chromosome length and subsequent PCA analysis have revealed that, even combined, these effects cannot explain the correlation pattern seen.

Each chromosome has its own unique properties, which cannot be detected when considering the correlation pattern for the full genome. Since each chromosome differs in contact frequency preferences, the correlation patterns must also differ. Indeed, while the first nine chromosomes show a correlation pattern similar to that of the whole genome, smaller chromosomes exhibit individual unique correlation patterns, which have been proved not to be a consequence of small sample size.

We then zoomed in on small chromosomes to investigate the contribution of syntenic regions to the correlations patterns observed. We found that short chromosome syntenic regions indeed exhibit correlations between contact frequency and state enrichments not characteristic of synthetic blocks in long chromosomes. Moreover, syntenic regions have similar preferences in contact frequencies with other syntenic region in human and mouse genomes. These interaction profiles are also similar for syntenic regions mapped at 1Mb resolution using gene-based liftOver [6] in a genome-wide fashion. Then we focused our attention on transitioned syntenic blocks - blocks contained in small human chromosomes and large mouse chromosomes or visa versa. Transitioned blocks were shown to exhibit similar cumulative contact frequencies and TAD structure (as previously shown in [7]) in human and mouse genomes. The interaction preferences observed seem to be intrinsic properties of syntenic blocks as they do not depend on the location of the region in the genome and are inherited in region transitions between chromosomes.

This work was supported by the RSF (grant No14-24-00155).

1. Lieberman-Aiden, Erez, et al. (2009) *Science* 326.5950: 289-293.

2. Rao, Suhas SP, et al. (2014) *Cell* 159.7: 1665-1680.

3. Kharchenko, Peter V., et al. (2011) *Nature* 471.7339: 480-485.

4. Ernst, Jason, et al. (2011) *Nature* 473.7345: 43-49.

5. http://stereogene.bioinf.fbb.msu.ru

6. https://genome.ucsc.edu/cgi-bin/hgLiftOver

7. VietriRudan, M. et al. (2015) Cell reports 10.8: 1297-1309.