Skoltech

Skolkovo Institute of Science and Technology

Jury Member Report – Doctor of Philosophy thesis.

Name of Candidate: Aleksandra Galitsyna

PhD Program: Life Sciences

Title of Thesis: Chromatin folding in individual cells

Supervisor: Professor Mikhail Gelfand

Name of the Reviewer: Anton Goloborodko

I confirm the absence of any conflict of interest (Alternatively, Reviewer can formulate a possible conflict)	 **Date: 21-09-2021**

The purpose of this report is to obtain an independent review from the members of PhD defense Jury before the thesis defense. The members of PhD defense Jury are asked to submit signed copy of the report at least 30 days prior the thesis defense. The Reviewers are asked to bring a copy of the completed report to the thesis defense and to discuss the contents of each report with each other before the thesis defense.

If the reviewers have any queries about the thesis which they wish to raise in advance, please contact the Chair of the Jury.

Reviewer's Report

Reviewers report should contain the following items:

- Brief evaluation of the thesis quality and overall structure of the dissertation.
- The relevance of the topic of dissertation work to its actual content
- The relevance of the methods used in the dissertation
- The scientific significance of the results obtained and their compliance with the international level and current state of the art
- The relevance of the obtained results to applications (if applicable)
- The quality of publications

The summary of issues to be addressed before/during the thesis defense

The Ph.D. thesis of Aleksandra Galitsyna advances our understanding of the 3-dimensional structure of chromosomes. This field of molecular biology has experienced explosive growth within the last decade, fueled by the invention of novel experimental techniques as well as steady improvements in methods of genetic engineering, microscopy, and DNA sequencing. Despite these rapid advances, the field still has many fundamental open questions, such as characterizing the chromosome structures in various organisms and cellular contexts, understanding the molecular mechanisms that shape these structures, as well further improving experimental methods. Aleksandra had prolific Ph.D. studies and managed to address a number of open questions in all of these categories. Overall, I find this thesis to be of the highest quality, both in terms of scientific contribution, as well as in its structure.

Most of the work presented in this thesis centers on the novel experimental method called Hi-C. Hi-C characterizes the structure of chromosomes inside living cells by measuring the frequencies of physical contacts between different locations in the genome. Briefly, in Hi-C experiments, researchers use a series of biochemical reactions to modify the DNA molecules of the studied cells and then analyze these molecules using high-throughput DNA sequencing machines. The resulting data are difficult both to process - a typical experiment produces tens of gigabytes of sequences, which then get converted into contact data via a complex multi-step pipeline - and to analyze, as the resulting data contain information on many levels of genome organization that typically varies a lot across the studied population of cells. As a result, working with Hi-C data requires special education in computer science in bioinformatics, attracting bright young minds such as Alexandra.

In her thesis, Aleksandra demonstrated the knowledge and skill with the existing methods of Hi-C data analysis, that are available to very few researchers in the world, as well significantly contributed to the arsenal of computational tools of this field.

The results obtained by Aleksandra in the course of her thesis work, represent a major contribution to the field of 3D genomics. On the side of biology, her work significantly expanded our understanding of the 3D genome structure of Drosophila fly, one of the most popular and convenient model organisms in biology, that however remained surprisingly understudied in this field. The work by Aleksandra and her collaborators demonstrated that the fly genomes fold according to principles similar to those in mammals, but also show a number of unique structural features. Importantly, these discoveries were enabled by Aleksandra's significant methodological developments in data analysis.

Specifically, in chapter 4, Aleksandra characterized the spatial segregation of inactive and active parts of the fly genome, known as "genomic compartmentalization". Compartmentalization is an evolutionarily conserved feature, whose functional role and mechanism are not fully understood. In mammals, the formation of inactive compartments correlates with two independent molecular mechanisms: phase separation of heterochromatin and its attachment to the nuclear wall, or, lamina. In this chapter, Aleksandra and her colleagues studied the role of lamina attachment in the fly genome folding by depleting the lamina proteins and observing the changes in chromatin architecture with microscopy and Hi-C. They showed that, while chromosomes show some degree of decompaction upon lamina depletion,

the genome architecture remains largely similar, thus highlighting the multiple mechanisms behind genome compartmentalization.

In Chapter 5, Aleksandra characterized another important layer of fly genome organization, topologically associating domains, or, TADs. TADs are another evolutionarily conserved structure of the genome, whose function remains largely mysterious. In mammals, TADs are formed by proteins complexes called cohesins via the mechanism of loop extrusion and are demarcated by DNA-bound CTCF proteins and sites of active expression. In Drosophila, the mechanisms of TAD formation seem even more complicated, as cohesins do seem to play a role in TAD formation, the existence of loop extrusion is debated, and TADs are demarcated by a broad collection of proteins. Together with her colleagues, Aleksandra asked what DNA-binding proteins or histone modifications can be responsible for TAD demarcation in the fly genome. They used a wide collection of available datasets and cutting-edge techniques of machine learning. They found that the strongest TAD signal came from a protein called Chromator/Chriz, as well as from histone modification H3K4me3, typically occurring at transcriptionally active regions. These findings narrow down the set of potential candidates for TAD forming factors in flies and will aid the design of future experiments.

In Chapter 6, Aleksandra addressed an important question of data decomposition in Hi-C. One of the key issues with Hi-C data analysis is that the resulting datasets are extremely complex and represent a convolution of different structures, that vary across cells, as well are contaminated with several experimental biases. As a result, the current methods of Hi-C data analysis are based on a collection of agreed-on heuristics and offer great potential for further improvement. Here, Aleksandra and her colleagues, dissect one of such heuristics, called the assumption of equal visibility, and show that it may result in a loss of important biological information.

In Chapter 7, Aleksandra addresses the question of the structural variability of chromosomes in individual cells of Drosophila. To characterize this question, she and her collaborators used a cutting-edge experimental modification of the Hi-C experiments called single-cell Hi-C. The standard Hi-C experiment measures the average contact frequency within a large population of cells (typically, 105-106 cells); the single-cell experiments can resolve contacts between different cells. These experiments truly represent the cutting-edge technology - at the time when Aleksandra and her colleagues started working on this project, there were only four such experiments ever published. Importantly, all four datasets were obtained in mammalian cells, leaving open the question of the variability of chromosome structures in Drosophila. This study demonstrated that fly cells have much less variability in TAD structure from cell to cell, comparing to mammalian cells. It is also important to note that the quality of data produced in single-cell experiments varies greatly between different labs, and, on average, remains far from optimal. Aleksandra and her collaborators have convincingly shown that their new datasets are on par and better than the existing ones by several metrics, such as coverage and genomic resolution, thus representing a great resource for the community. Finally, in this chapter, Aleksandra makes a significant contribution to the methods of single-cell Hi-C analysis via a novel software package called ORBITA.

Finally, in Chapter 8, Aleksandra provides a thorough review of the state of the art of the young field of single-cell Hi-C experiments and data analysis. I found the review to be refreshingly substantial, striking a nice balance between addressing high-level conceptual questions and focusing on important technical questions in the field.

Aleksandra had an extremely productive Ph.D., as all of these chapters were published in respectable peer-reviewed journals. I found all five of these papers to be very well designed, technically sound, and well written. As all of the chapters of the thesis have already undergone strict peer review by my colleagues, I found it very difficult to spot major issues.

For the thesis defense and our future scientific discussions, I would like to propose a couple of intentionally open-ended questions about the future of the field (the answers to which I may not know myself):

What are the fundamental limits for bias extraction in Hi-C? In other words, can we expect one day to be able to cleanly extract all experimental biases from Hi-C, w/o removing any useful information? What kind of additional experiments could one design to help with this question?

What would single-cell Hi-C maps look like if we had an "ideal" experiment, where 100% of DNA fragments would get ligated and recovered? Would we be able to tell protein-supported loops from sporadic contacts? What would TADs and compartments look like? What kind of additional information would we be able and not able to extract from such maps? How badly would the presence of sisters and homologs confuse us?

Provisional Recommendation

☒ *I recommend that the candidate should defend the thesis by means of a formal thesis defense*

☐ *I recommend that the candidate should defend the thesis by means of a formal thesis defense only after appropriate changes would be introduced in candidate's thesis according to the recommendations of the present report*

☐ *The thesis is not acceptable and I recommend that the candidate be exempt from the formal thesis defense*