Dr hab. Mirosław Kwaśniewski
Centrum Bioinformatyki i Analizy danych
Uniwersytet Medyczny w Białymstoku

Katowice, 10.05.2021

The review of the Ph.D. dissertation by Dimitrios Zisis entitled:
"Data analysis methods for inferring chromatin configuration on the basis of 4C-seq experiments"

The doctoral dissertation submitted for the review was conducted at the Department of Biometry and Bioinformatics, Institute of Plant Genetics of the Polish Academy of Sciences. The thesis supervisor is prof. dr hab. Paweł Krajewski, a recognized expert in the field of development of computational methods used in genomics research.

1. The scope of the work

   The PhD dissertation of Dimitrios Zisis concerns the creation of a new scheme for the analysis of data obtained in one of the methods of studying the structure of chromatin: 4C-seq. Based on the previously conducted and published by the author comparative analyzes of the existing programs used so far in the analysis of data generated in the 4C-seq method, the author attempts to develop a new, better method that takes into account the weaknesses of the programs used so far. The results of the work carried out and the operation of the newly developed analytical pipeline called *4CseqR* were illustrated by analyzing the existing 4C-seq data from experiments carried out on *Arabidopsis thaliana* and *Mus musculus*.

2. General description of the thesis

   The dissertation is written in a classic way. The work consists of 6 chapters in which the author reviews the literature on the subject of the work, formulates the aim of the work, describes the methods and materials used, and presents the results of the work, which then discusses and draws final conclusions.

   The introductory chapter outlines the methods of studying chromatin organization and chromosome conformation in terms of the interaction of chromatin fragments in cell nuclei. Particular emphasis was placed on the description and explanation of the 4S-seq method and its application in research in the field of functional genomics. Later in the chapter, the author briefly describes the bioinformatics methods used so far in the analysis of data generated with the 4C-seq technology. The chapter ends with an important, critical analysis of the advantages and disadvantages of existing software packages, especially with regard to the specific features of the 4C-seq method. This part

of the introductory chapter logically leads the reader to the set goal of the doctorate thesis. The introduction is comprehensive, supported by a large number of literature references, written in a concise and clear manner.

The aim of the work, which was to develop a new approach to the analysis of data obtained by the 4C-seq method, was presented as a consequence of the author's observations and conclusions regarding the non-optimal performance of other programs and methods used so far for the analysis of 4C-seq data. Describing the aims of the work, the author refers to his published experiments comparing the existing algorithms used in the 4C-seq data analysis and precisely defines what elements of the analytical pipeline he intends to address by developing a new method.

The Material and Methods chapter begins with a description of two data sets that the author uses later in the work to evaluate the performance of the software developed. In the work, the author uses the data sets obtained by the 4C-seq experiments for two organisms with significantly different genome sizes: *Arabidopsis thaliana* (genome size 135 Mb) and *Mus musculus* (genome size 2.7 Gb). Importantly, the author does not conduct biological 4C-seq experiments on his own but uses publicly available data sets generated in other studies.
In the further part of the Materials and Methods chapter, the author presents the general scheme of the analytical procedure used in the 4CSeqR software package he created and describes in detail the key elements of the created analytical pipeline. Descriptions of each of the pipeline elements, the strategies and statistical methods used, and the rationale for their use are clearly explained. Additionally, the descriptions refer to the created and published analytical pipeline scripts deposited on the author's GitHub platform account.

The main part of the work is the chapter devoted to the description of the obtained results. The author describes the results of using the created 4CSeqR software package and the statistical methods used in experiments carried out on two independent data sets, Dataset 1 for *Arabidopsis thaliana* and Dataset 2 for *Mus musculus*, respectively. For each of the data sets, the results of the software operation at each stage of the 4C-seq data analysis were presented separately, and additionally, the obtained results were compared with the results generated by the popular *foursig* software package. The results obtained at each stage of the analysis were described in detail and clearly, and were supported by charts and tables, which significantly facilitated their understanding. In the final part of the Results chapter, the author conducts a brief analysis of the biological significance of the obtained results for the experiments carried out on *Arabidopsis thaliana* and *Mus musculus*.

In the final part of the work, the author thoroughly discusses the obtained results and summarizes the work carried out.

The doctoral dissertation ends with a list of references (68 items). A CD with a copy of the doctoral dissertation and a set of 41 supplementary files containing detailed results and * .bed files generated during the analyzes were also attached to the dissertation.

3. Detailed evaluation of the dissertation

The development of Next-Generation Sequencing (NGS) methods has revolutionized biological sciences, particularly in the field of genomics research. One of the latest trends in this area is research on the influence of the structure and spatial conformation of chromatin in the cell nucleus on gene expression. Several laboratory methods have been developed in recent years to enable such analyzes. All of them use NGS technology to precisely identify chromatin fiber contact sites in the nucleus and can be used to evaluate the potential effect of chromatin rearrangement and contacts on the regulation of gene expression in cell differentiation or in cellular response to external factors.
The raw results of NGS sequencing obtained in chromatin conformation studies, including the 4C-seq method, require advanced bioinformatic analyzes to enable the selection of the actual contact sites. Therefore, the development of laboratory methods entails a strong development of bioinformatics methods that enable the interpretation of the obtained data. To date, a number of software packages have been developed to facilitate the analysis of results from 4C-seq experiments. The author of the dissertation has previously assessed the existing software (Zisis et al., 2020). Taking into account his experience and conclusions from the research published so far related to the use of the 4C-seq method in genomic research, Ph.D. candidate has decided to propose, create and test a new analytical approach and a new software package for analyzing data from 4C-seq experiments - in principle, better and more reliable than the existing solutions.

The author proposed a new analytical approach, linking proven and useful methods used successfully to solve problems of a similar nature (*Salmon*, sliding window approach) with a general approach characteristic of experimental research (biological replicas, well-founded statistical methods). As a consequence, the author created a new software package called *4CseqR*, which enables convenient data processing at all stages of NGS data analysis in 4C-seq experiments. The author tested the created solution using two publicly available data sets from experiments conducted on *Arabidopsis thaliana* and *Mus musculus* and compared the obtained results with the results generated by the *foursig* software.

The created method is not concentrated only on the detection of significant contacts within samples but enables comparative analysis of experimental variants using new approach which is combining two different mathematical models: linear mixed models (LMM) and analysis of contingency tables by Fisher's test. Consequently, the method allows to describe the significance of contacts by continuous contact signal intensity and by binarized description of presence or absence of contacts. Also, the method enables analysis of contacts in cis and trans at the same time. Therefore, the created software package facilitates the analysis of 4C-seq data from the beginning to the end and gives the user the opportunity to select the optimal parameters

of the analysis (for example by changing the size of the sliding window, which significantly affects the DCR identification).

Unfortunately, the author did not attempt to verify in more detail whether the obtained results can be realistically related to biological processes of experimental datasets used in his study. The author only briefly performed an analysis of Gene Onthology overrepresentation using the genes present in the identified DCRs. In my opinion, the author could carry out a more detailed analysis of the obtained results and refer to the knowledge about the processes related to vernalization in Arabidopsis or stem cell differentiation based on the results of previously published work. Such analyzes would allow a better assessment of the effectiveness of the created solution in solving real biological problems.

4. Comments and questions

After reading the dissertation and published comparison of the usefulness of existing programs that enable the analysis of 4C-seq data, I have the impression that the results obtained are very different for different programs. Therefore, the basic questions arise:

a. Why were the data from the foursig program only selected for the comparative analysis of the results performed on the Arabidopsis thaliana and Mus musculus datasets, especially as the 4CSeqR and foursig results only slightly overlapped? Perhaps a comparison of the 4CSeqR results with other programs, published in Zisis et al., 2020, would produce better (at least different) results and show more common elements?
b. Considering the different results obtained from the analyzes with different programs, how can you check and prove which programs are producing the correct (reliable) results? What positive and negative controls can be used in data analyses of the 4C-seq experiments, especially those which were already published?
c. Does it make sense to use a large sliding window (200000; 50000) for a very compact genome? Are we not devoting too much to the biological question of the role of contact in regulating the expression of specific genes? It would be interesting to see what genes carry a single significant DCR identified in the small sliding window in Arabidopsis (40000; 20000) and whether they have anything to do with vernalization.
d. What is the minimum data set that can be used to reliably analyze 4C-seq contacts in cis and in trans? Was the mouse dataset sufficient to perform the analyzes? How to explain the very large differences in the coverage of the chromosome containing the bait in mice and in Arabidopsis, despite such large differences in average coverage on other chromosomes?

5. Conclusions

The presented comments should not obscure the fact that the reviewed work presents a high scientific level in the field of bioinformatics and statistical analyzes used in genomics of eukaryotic organisms. Undoubtedly, the Ph.D candidate showed great skills in creating new

and adapting existing bioinformatics solutions and statistical methods to support the experimental work methodically. The conducted research is part of the mainstream research on the influence of chromatin conformation in the cell nucleus on gene expression, and thus on differentiation or cell response to external factors. Unquestionably, the obtained results bring us closer to getting to know these mechanisms better. On this basis, I am asking the Scientific Council of the Institute of Plant Genetics of the Polish Academy of Sciences to admit mgr Dimitrios Zisis to further stages of the doctoral dissertation.