Radzików 2.02.2021 r.

Sz. Pan

Prof. dr hab. Paweł Krajewski

Dyrektor Instytutu Genetyki Roślin PAN

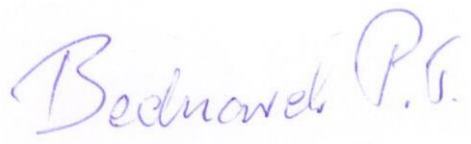
Strzeszyńska 34,

60-479 Poznań

Szanowny Panie Profesorze,

W nawiązaniu do Pana Pisma (L.dz. 3311/2020) dotyczącego wykonania recenzji rozprawy doktorskiej pana mgr D. Zisisa pt: ”Data analysis methods for inference on chromatin configuration on the basis of 4C-seq experiments” (Uchwała Rady Naukowej IGR PAN nr 28/VI/2020 RN IGR PAN) serdecznie dziękuję za okazane mi zaufanie i możliwość zapoznania się z bardzo ciekawymi rozwiązaniami dotyczącymi sposobów analizy danych eksperymentalnych uzyskiwanych metodą 4C-seq.

Z poważaniem,



P.S. W załączeniu przesyłam kopię elektroniczną oraz pocztą wydrukowane egzemplarze recenzji.

INSTYTUT HODOWLI I AKLIMATYZACJI ROŚLIN
PAŃSTWOWY INSTYTUT BADAWCZY
PLANT BREEDING AND ACCLIMATIZATION INSTITUTE
NATIONAL RESEARCH INSTITUTE
tel. centrala:+(4822)7334500, fax:+(4822)7254714, e-mail: postbox@ihar.edu.pl
http://www.ihar.edu.pl, REGON 000079480, NIP 529-000-70-29, KRS 0000074008
Nr konta:PEKAO I/O Błonie, 54 1240 2164 1111 0000 3561 7204

Zakład Biochemii i Fizjologii Roślin
PL 05-870 Błonie, Radzików
e-mail: p.bednarek@ihar.edu.pl
tel. +(48 22) 725 45 33

Radzików, dnia 2.02.2021 r.

Prof. dr hab. Piotr T. Bednarek
Plant Breeding & Acclimatization Institute – NRI
Plant Biochemistry and Physiology Dept..

Review on Thesis for degree of Doctor of Philosophy in Agricultural Science entitled:

"Data analysis methods for inference on chromatin configuration on the basis of 4C-seq experiment"

Thesis supervisor: prof. dr hab. Paweł Krajewski

By MSc Dimitrios Zisis

The Review was done base on the resolution Nr 28/VI/2020 RN IGR PAN of the Scientific Council of the Institute of Plant Genetics – PAS, Poland, Poznań
and on
Art 13 ust. 1 (14.03.2003) concerning titles and degrees in arts (Dz U. 2014 poz 1852 z późn. zm.) in connection with article 179 act 1 (3.07.2018). According to regulations introducing an act – Prawo o szkolnictwie wyższym i nauce (Dz. U. 30.08.2018 poz. 1669) and in the case when Doctoral Thesis is a standalone and separated part of collective work, also the evaluation of individual input of the Candidate for the work (par 6 ust. 5 Rozporządzenie Ministra Nauki I Szkolnictwa Wyższego concerning detailed rules and conditions of doctoral Thesis, habilitation and procedure conferring professor title).

Review

The presented Doctoral Thesis entitled: **"Data analysis methods for inference on chromatin configuration on the basis of 4C-seq experiment"** is written in English. It is organized according to common rules and consists of the following sections: Introduction, Aims, Materials & Methods, Results, Exemplary downstream analysis, Discussion, Conclusions, and References. It is proceeded by Acknowledgments and short Abstract written both in English and Polish.

A six-page long Introduction encompasses information on chromosome conformation capture (3C) and its derivatives, with particular emphasis on the circular chromosome conformation capture approach (4C). The Author illustrates a general scheme of the 4C method, its purposes, and existing alternative approaches, as well as their limitations concerning data analysis. Assuming the limitations, the Author puts the aims of his Thesis concerning the significance of contacts in the 4C-seq experiment that need to be defined in relative terms based on the experimental and applied controls' purpose and design. Furthermore, it is being suggested that the so-called contacts in 4C-seq experiments should be done for various parameters of genomic signals. To solve the aims, Mr. D. Zisis suggests a methodology that, in his opinion, should unravel the addressed issues (the identification of transformation of raw sequencing data; well-founded statistical methods to assess the significance of differences between experimental variants; selection of data sets with contrasting properties) if a new pipeline is evaluated.

Materials & Method section is devoted to two experimental sets, the one for *Arabidopsis thaliana* and the other for *Mus musculus* consisting of repeated and contrasting data. In the first case, changes in genome-wide contacts of the flowering locus C (FLC) gene responsible for vernalization in non-vernalized and four weeks (7 days warmth) vernalized plants with three repetitions are considered. The second set comprises data related to different lines of embryonic stem cells. Six data sets concerning three biological repetitions were selected. The general scheme of the 4C-seq data analysis forming pipeline is made up of a library of fragments, data preparation, read alignment, estimation of fragment coverage, finding contacts, sliding window strategy, comparative analysis of experimental variants, and adjustment of P-value.

The Results section describes the outcomes obtained using a pipeline developed by the Author based on two data sets. First of all, Mr. D. Zisis takes care of data quality. For that reason, the FastQC program delivering a boxplot of the data quality is being used. It is demonstrated that the sequencing data used are of high quality as the median is bigger than twenty.

The Authors' reasoning then is to trim viewpoint sequences from reads, remove reads without the primary restriction site, and align the resulting reads to a library of restriction fragments. The best possible alignment is achieved by applying a minimum score parameter allowing proper alignment.

A critical step implemented in the 4CseqR pipeline assumes the evaluation of coverage estimation and distribution of estimated coverage along all chromosomes and a bait chromosome. The estimated coverage is subjected to normalization (concerning the length of fragments, length of fragment ends, secondary restriction site in the fragment) by ranks to eliminate putative differences in the estimated coverage distribution between categories.

Two alternative or complementary approaches to further data analysis were implemented in the pipeline. The first one uses the linear mixed model (LMM) and the other one data binarization combined with a sliding window approach needed to locate differently contrasting regions. The false discovery rate (FDR) and the Fisher exact test are implemented to distinguish between sequences in contact with the bait. Only the windows with small p values of FDR or the Fisher test (differently contacting windows -DCW) are assumed to contain the so-called differently contacting regions (DCR). The DCRs are assigned to contrasting experimental data (e.g., vernalized vs. non-vernalized *Arabidopsis thaliana* plants or contrasting *Mus musculus* stem lines). If possible overlapping windows of restriction fragments are merged. The two

approaches' implementation demonstrated that the sliding window size does matter and results in varying the DCR located by the LMM. It was demonstrated that the sliding window size is essential in the number and location of DCRs, mainly when the Fisher test is used.

Furthermore, the two alternative approaches' utilization resulted in few common DCR regions on *Arabidopsis thaliana* affecting chromosomes 2, 4, and 5. In *Mus musculus*, the identification of such regions was problematic. According to the Author, the explanation relies on the genome size with better results for smaller genomes.

The presented pipeline implements the third approach, allowing for identifying significant contacts with the bait using the fourSig tool. The fourSig does not provide any tool for comparative analysis. The analysis is performed independently for each replication, and each variant and only contacts present in three replicates are considered significant. Comparing results of fourSig, LMM, and Fisher test showed that few shared contacts were detected. Furthermore, the percentage of contacts varied from as few as 1.2 to 32.7% and depended on the experimental variant. Comparison of the Fisher test and fourSig outcomes resulted in up to 12.8% of contacts and dependent on the experimental variant in *Arabidopsis thaliana*. Analysis of *Mus musculus* data showed that the number of significant contacts shared for the LMM and fourSig rose to 53.1%, whereas Fisher and fourSig results to 18.5%.

Finally, the results section encompasses the downstream analysis of DCRs in *Arabidopsis thaliana* and *Mus musculus*, demonstrating the application of the data evaluated using the pipeline.

In the Discussion section, Mr. D. Zisis focuses on statistical issues implemented in the pipeline giving the rationale for their implementation. Such aspects as coverage estimation, data normalization and transformation, relative approach, and statistical models are being discussed. Among others, the Author explains differences between implemented models and shows that identifying significant contacts could be generalized for multi-factor experiments. Mr. D. Zisis discusses the problem of identifying proper settings in the case of sliding window depending on genome size being analyzed and explains why the approach based on the LMM and Fisher test may result in non-coinciding outputs (due to window properties). Some attention is also paid to the fourSig approach that the Author assumes to be efficient as it operates on quantitative signal measures and identifies common and specific contacts shared between the fourSeq and LMM or Fisher approach in *Arabidopsis thaliana*.

Conclusions presented in the Thesis are well defined and summarize the most important aspects of the presented variant approach evaluated to identify important contacts using the 4C method.

While the Thesis is well written and all issues are properly addressed, below are some minor remarks or comments that might be of interest to the Author. They arose mainly because of my interest in the presented analysis. They did not influence positive opinion of the work. Furthermore, I have some general questions for Mr. D. Zisis to consider overall aspects not necessarily discussed in the Thesis but linked to the study.

Remark/comment on data quality

Although not necessary, implementation of additional options that allow, i.e., the evaluation of data quality using cumulative plots (resulting in the average quality of each read giving a value of average quality) or base content at each position (demonstrating whether some bases are not overrepresented in the read) might improve quality control of the data. Although not necessary when the alignment option is used in the pipeline, it would be interesting to have the opportunity to remove sequencing errors. Possibly, this would improve further analysis.

Remark on sequence error elimination

Is it not essential to implement some algorithms that would allow for the elimination of sequencing errors? What kind of algorithms could be useful here and would not significantly slow down the pipeline? If

applied, would they significantly improve the analysis in combination with restrictive alignment settings?

Remark on exploitation non-parametric test for the evaluation of contacts

The proposed approaches implemented for the identification of the DCRs are based on parametric tests. Do you think that non-parametric tests (which of them) would be of value in distinguishing contacts? Do you think that i.e., machine learning variants (i.e., artificial neural networks) could be used here?

Remark concerning biological aspects of results

Chromosome 5 of A. thaliana encodes FLC present close to the centromeric region. Does the contact identified on chromosome 5 coincide with the location of the FLC? How do you explain the presence of intersecting DCR from the LMM and Fisher test on chromosomes 2, 4, and 5?

Is there any biological reason for contacting chromosome 2 (receptor kinase gene ERECTA) and 4 (FRIGIDA gene involved in FLC regulation is present)? Can you explain the presence of common DCRs by biological phenomenon?

If the two approaches identify different, differently contacting regions, which of the methods is preferential? Are there any statistical (or other) criteria that allow distinguishing which approach is the right one? Or the two approaches should be run, and only shared results should be considered?

Remark on genome complexity

As you have mentioned, genome size might have affected the approaches used to identify DCRs. Do you think that problem is the genome size or its complexity (presence of TEs, microsatellites, ploidy)? Is it possible to overcome the problem?

Remark concerning Discussion

I would expect that an extended discussion to compare existing pipelines dedicated to the 4C analysis would be presented in the Discussion. Furthermore, having experimental data from two distinct kingdoms and exciting results from their analysis, I feel that the results' biological background is not presented. This could be done, i.e., of A. thaliana, where FLC locus functioning is well studied, and interaction between different genes was suggested in the literature. I do think that even a short discussion (about a page) would make the Thesis even more fascinating for biologists demonstrating that real biological phenomena form the background for 4C analysis.

Additional issues

1. Is it possible to combine your method of 4C-seq data analysis with the QTL approach? If so, should it improve the discriminative ability of the approach?

2. Assuming 4C analysis could be performed not on contrasting experimental data but on a large sample size that fulfills minimum sample size requirements supporting sufficient effect size, is it possible to evaluate relationships (mathematical models) between contacting regions?

3. Mathematical modeling could be applied for demonstrating processes reflecting biological phenomena. Is it possible to use the pipeline results to model chromatin conformational changes in response to environmental stresses?

Final Conclusions

The Thesis under review is very engaging and reading it was a great pleasure to me. The text is written in a condensed, precise manner without unnecessary divagations, making it easy for reading despite presenting complicated tasks. The author demonstrated that the presented method could be successfully applied to identify significant contacting regions in contrasting experimental data. Indeed, the presented Thesis reflects the knowledge and Mr. D. Zisis’s fluency in developing prospective statistic-based tools that

could be useful for analyzing the biologically important phenomenon. Thus, Mr. Dimitrios Zisis fulfills the requirements necessary for applying for a Doctoral degree. Furthermore, assuming all mentioned above, the Thesis is worth distinction.

Profesor dr hab. Piotr T Bednarek