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

by

Dimitrios Zisis

A thesis submitted in fulfilment for the

degree of Doctor of Philosophy in Agricultural Sciences

in the Department of Biometry and Bioinformatics,
Institute of Plant Genetics of the Polish Academy of Sciences

Supervisor:

Prof. Paweł Krajewski

Poznań 2020

Metody analizy danych dla wnioskowania o konfiguracji chromatyny na podstawie eksperymentów 4C-seq

Promotor
prof. dr hab. Paweł Krajewski

Zakład Biometrii i Bioinformatyki
Instytut Genetyki Roślin Polskiej Akademii Nauk
w Poznaniu

Kandydat
mgr inż. Dimitrios Zisis

Poznań 2020

To my family

“ Το ορατό
δεν είναι πάντα βιωμένο,
ούτε το βιωμένο
είναι πάντα ορατό.”

Αντώνης Δ. Σκιαθάς

ACKNOWLEDGMENTS

The last six years have been some of the most challenging, but also most rewarding. This has not been an easy journey for multiple reasons. I wasn't always sure that I had made the right decision to start a PhD but I am glad I persisted and I even ended up learning a few things along the way! This was possible only because of the support of some truly amazing people. In this section I would like to thank all people without whom this work would not have been possible.

First of all, I would like to thank Prof. Paweł Krajewski for giving me the opportunity to enrol as Marie Curie Fellow at the Institute of Plant Genetics of the Polish Academy of Sciences (IPG PAS), for supporting my registration as PhD candidate, and for his continuous help, support and advice. From him I have learnt hard work, strictness, honesty, and persistence as indispensable values a scientist must have to pursue the highest scientific standards. At the end he was the most important person and one of the few who believed in me something that I will never forget and it will be always a pleasure to work with him.

I thank all the members of the Biometry and Bioinformatics team at IPG PAS (Aneta, Pedro, Wojtek, pani Bernardeta, Magdalena) for being the most fun and brilliant people and always supporting and helping. Mainly Hanna for supporting me and helping me with any need I had since my first day in Poznan.

I would like to specially thank all the members of EpiTRAITS project, for the amazing moments. We had the chance to collaborate and learn from experts in their fields and enrich our scientific knowledge not only with practical skills from the training but also with ethical and cultural ones. EpiTRAITS director prof. Maike Stam and her team (Helen Bergman et al.) organized amazing courses and trips giving us the opportunity to create a great scientific network in Europe. My participation in EpiTRAITS gave me the opportunity to make many new friends (Javier, Pawel, Dorota, Massimo, Blaise, Till, Suraj) and spend magnificent moments with them.
I would like to thank prof. Valerie Gaudin (Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, INRA Centre de Versailles) for being great collaborator and helping me as a mentor.

I feel privileged for having been surrounded by greatest scientists in the past, but also in the present. I would like thank the Hellenic Pasteur Institute and all the members of Diana-lab for continuously shaping my passion for science and supporting me to finish a PhD in parallel with other projects and work we had to do in the lab.

I gratefully acknowledge the European Union for supporting scientists at very early stages of the careers through the Marie Curie Fellowship Program, recently renamed as Marie Skłodowska-Curie Fellowship Program in recognition of the Polish origins of the double Nobel Prize winning Polish-French scientist.

I thank the Poznan´ Supercomputing and Networking Center, affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences, for providing high performance computing and file archiving services.
Finally, I would like to thank my parents Kostas and Dora and my sister Vasilina for their infinity support at every decision I have made over the years and their unfailing love. I know it was never close enough but I was lucky to be only few hours and a plane away from the warmth of home and my mom's food. Last but not least I would like to thank my uncle Dimitris Porfiris for his support all these years, with his practical and spiritual advices, helping me to perceive much deeper than the surface in every area.

A special thanks to my fellow travellers, the Veterans, who were always next to me supporting all my decisions.

"You're my river running high, run deep, run wild… "

Contents

Abstract

The circular chromosome conformation capture technique followed by high throughput sequencing (4C-seq) has been used in a number of studies to investigate chromatin structure by identifying interactions between DNA fragments. It is considered as a cost effective and powerful high resolution method which can study all interactions made across the genome by a given site of interest. This dissertation describes a data analysis methodology devoted to finding, on the basis of next generation sequencing data, genomic regions that are characterized by an elevated frequency of interactions. The main goal of this thesis is to present in detail this new analysis schema, which was developed by following two main requirements: adhering to the generally accepted rules of experimentation and striving at a comprehensive description of studied genomic events and signals.

To begin with, we describe the preparation, the characteristics and the design of a 4C-seq experiment. Then, we present the most important 4C-seq data analysis methods. Against this background, we propose the new method of 4C-seq data analysis called *4CseqR*. We describe all its elements, that is, the selected computational and statistical methods. All steps of the proposed analysis are studied and discussed. They proceed from the pre-processing of next-generation sequencing reads, through the mapping and treatment of mapped reads, the normalization of read coverage, until the calling of significant contacts and the comparative statistical analysis of experimental variants. The latter step is based on two statistical approaches: analysis of continuous response variables by linear mixed models and analysis of discrete responses in contingency tables. Finally, a comparison between the results obtained by the proposed *4CseqR* method and one of the existing method, *fourSig*, is presented.

In order to illustrate the workflow and the results of the data analysis, datasets concerning two different species have been used, one from a study devoted to vernalization control in *Arabidopsis thaliana* by the FLOWERING LOCUS C (FLC) gene and another one from a study of long-range chromatin interactions in *Mus musculus* embryonic stem cells.

Keywords: circular chromosome conformation capture (4C); next generation DNA sequencing; chromosomal interactions; NGS data analysis methods; linear mixed models; contingency tables; Fisher exact test.

Streszczenie

Metoda analizy pierścieniowatej konformacji chromosomów (*circular chromosome conformation capture*) połączona z sekwencjonowaniem wysokoprzepustowym (4C-seq) została wykorzystana w wielu badaniach nad strukturą chromatyny poprzez identyfikację interakcji między fragmentami DNA. Uważa się, że jest to opłacalna i wydajna metoda o wysokiej rozdzielczości, która umożliwia badanie ogółu interakcji zachodzących w genomie przy udziale wybranego lokus. W niniejszej rozprawie opisano metodologię analizy danych polegającej na znajdowaniu, na podstawie danych z sekwencjonowania nowej generacji, regionów genomowych, które charakteryzują się podwyższoną częstotliwością interakcji. Głównym celem niniejszej pracy jest szczegółowe przedstawienie nowego schematu analizy, który został opracowany w oparciu o dwa główne wymagania: przestrzeganie ogólnie przyjętych zasad eksperymentowania oraz dążenie do wszechstronnego opisu badanych zdarzeń i sygnałów genomowych.

Na początku opisujemy przygotowanie i charakterystykę eksperymentu 4C-seq. Następnie przedstawiamy najważniejsze istniejące metody analizy danych 4C-seq. Następnie proponujemy nową metodę analizy danych 4C-seq o nazwie *4CseqR*. Opisujemy wszystkie jej elementy, czyli metody obliczeniowe i statystyczne wybrane dla wszystkich etapów. Analiza przebiega od wstępnego przetwarzania odczytów sekwencjonowania nowej generacji, poprzez mapowanie i przetwarzanie mapowanych odczytów, normalizację obserwacji, aż do lokalizacji kontaktów i porównawczej analizy statystycznej wariantów eksperymentalnych. Ostatni krok opiera się na dwóch podejściach statystycznych: analizie zmiennych ciągłych za pomocą liniowych modeli mieszanych oraz analizie zmiennych dyskretnych w tablicach kontyngencji. Na koniec przedstawiamy porównanie wyników uzyskanych proponowaną metodą *4CseqR* z wynikami jednej z istniejących metod, *fourSig*.

W celu zilustrowania przebiegu proponowanej analizy wykorzystano zbiory danych dotyczące dwóch różnych organizmów: zbiór pochodzący z badań poświęconych kontroli wernalizacji u *Arabidopsis thaliana* przez gen FLOWERING LOCUS C (FLC) oraz zbiór z badań nad konfiguracją chromatyny w embrionalnych komórkach macierzystych *Mus musculus*.

Słowa kluczowe: analiza pierścieniowatej konformacji chromosomów (4C); sekwencjonowanie następnej generacji; interakcje chromosomowe; metody analizy danych NGS; liniowe modele mieszane; tablice kontyngencji; dokładny test Fishera.

Declaration of authorship

I, Dimitrios Zisis, declare that this thesis entitled ‘Data analysis methods for inference on chromatin configuration on the basis of 4C-seq experiments’ and the work presented in it are my own. I confirm that:

- This work was done mainly while in candidature for a research degree (PhD level) after enrolling as Marie Skłodowska-Curie Early Stage Research Fellow at the Institute of Plant Genetics of the Polish Academy of Sciences (Poznan´, Poland).
- Where I have consulted the published work of others, this has always been clearly attributed.
- I have acknowledged all main sources of help.
- There are no fragments of publications that have been selected and reproduced without modification from other sources.
- Ethical approval: This work does not contain any studies with human participants or animals performed by any of the authors.

Dimitrios Zisis
Poznań, November 2020

Funding

This work was supported by the EU Marie Curie Initial Training Network EpiTRAITS ("Epigenetic regulation of economically important plant traits", 2013-2016, agreement number 316965).

1. Introduction

The most recent successful systems biology studies depend on data concerning gene action obtained by different, complementary protocols, like RNA-Seq or ChIP-Seq. Many reports indicated that short- and long-range chromosomal interactions can influence gene expression and other gene-related phenomena in physiological processes and different disease states [see, e.g., references in Hövel et al. (2012) and Raviram et al. (2014)]. The first studies in this direction, based mainly on microscopy (Potapova et al., 2019), provided a first view of nuclear organization and revealed the ability of chromosomes to occupy distinct territories, with small mix among them. Then, the rapid development of the chromosome conformation capture (3C) technology transformed the field helping to better identify chromatin interactions at molecular level and study them in detail. At the same time high throughput and genome wide techniques were improved and gave an extra push to further development of genome-scale strategies. 3C technology, referred to as "one versus one" method for analysing interactions of pairs of DNA sites, is the basis of those strategies. These include 4C technology (one-versus-all) (Simonis et al., 2006, Zhao et al., 2006), 5C technology (many-versus-many) (Dostie et al., 2006), ChIA-PET (an approach combining chromatin immunoprecipitation, ChIP and 3C) (Fullwood et al., 2009), and Hi-C (all-versus-all) (Lieberman-Aiden et al., 2009). The main difference between various 3C-based methods is their scope. Each strategy has unique advantages and disadvantages and their choice depends on the specific research question that is to be answered. Hi-C can be used to identify pairwise interactions between fragments whereas 5C can identify interactions in all restriction fragments within a given region, but the interaction analysis will be limited to the primer design and will have relatively low coverage. Circular chromosome conformation capture (4C) combined with next generation sequencing (NGS), can be considered as the best option for high resolution interactions for a specific region of interest. Those different technologies can be collected under one umbrella with other massive DNA sequencing technologies which brought a revolution in genomic research.

1.1. The meaning of 4C

Circular chromosome conformation capture (4C) assay is a method that is derived from 3C, enriched with more innovations than the original 3C protocol. The advantage of 4C is that only the sequence of one of the contacting sites of interest needs to be known, and the locations and annotations of the contacts are inferred from the experimental results and from the reference genome sequence data (assumed to be known). Those genomic sites of interest are called “viewpoints” or “baits”, while contacting regions are called “contacts” or “captures”. Although until now 4C-seq has mainly been used in model organisms, there is an expectation that the further development and applications of this method, for example to sequenced crop plants, can provide useful information about chromosomal interactions that influence important traits of these organisms.

3C protocol is widely used to analyse interactions between pairs of DNA sequences. The first steps of 3C strategy are the formaldehyde cross-linking of interacting parts of DNA, the

digestion of cross-linked DNA by restriction enzymes (REs) and the ligation of cross-linked fragments. The ligated products between two DNA sites are then collected and analysed by PCR using specific primers for the restriction fragment of interest.

The 4C protocol follows the same initial steps as 3C until the crosslinks. After cross-linking and digesting with the first, or primary, restriction enzyme (usually, a 6 bp cutter), a second round of ligation is used to fuse the ends of DNA fragments present. The ultimate outcome of this ligation event are DNA circles containing multiple restriction fragments. A second cutting restriction enzyme, with a different recognition sequence than the first, is used for a second round of digestion and ligation of the DNA in order to create small DNA cycles which include a primary ligation junction. Primers are designed and used to amplify the unknown DNA outward around the circle. The result of this procedure is a library of captures ligated to the viewpoint. With the use of microarray or next generation sequencing, those captured regions can be read and further analysed (van de Werken et al., 2012a). The word "circular" in the protocol name comes from the circular fragments that are created and contain the parts of DNA of interest (van de Werken et al., 2012 a; Stadhouders et al., 2013). As it is mentioned above, to perform the extra round of digestion and ligation it is necessary to use a secondary restriction enzyme (4 bp or 6 bp cutter) in order to increase the robustness of contact profiles, which is important for the reproducibility of the method.

An overview of the 4C protocol is shown in Figure 1A. This overview includes details of all steps of 4C protocol from the digestion and ligation until the NGS product. The NGS data obtained from this process is used for bioinformatics and statistical analysis according to the needs of each experiment.

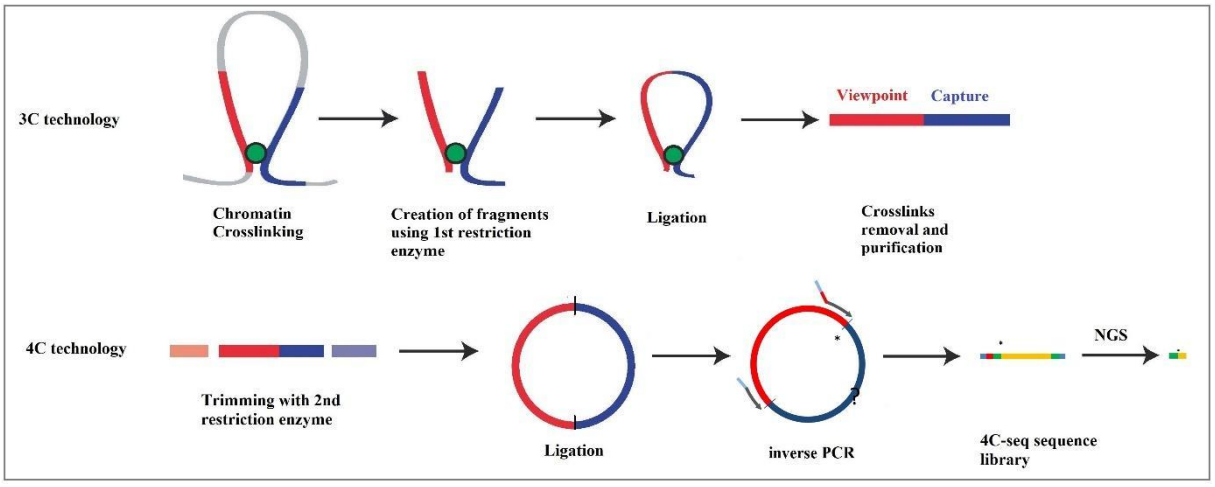

Figure 1: The 4C-seq protocol. Genomic regions that are proximal in the cell nucleus (red and blue) are fixated by formaldehyde induced protein-protein and protein-DNA crosslinks (green). The DNA is cut in fragments using the primary restriction enzyme and the fragments that are close are ligated, after removing crosslinks and molecules purification (3C technology). In 4C-seq the 3C product is digested using a secondary enzyme and circularized in the second ligation step**.** To identify and quantify ligated fragments to the genomic region of interest, an inverse PCR is performed with primers binding outward on the viewpoint and the amplicons are analysed using next generation sequencing. The output of NGS is used as input to bioinformatics methods and tools.

In Fig. 1 a typical 4C ligation product is shown. In blue the bait sequence is shown and in red the captured interacting sequence. The first ligation and restriction site is marked with an asterix. The grey arrows mark the binding site for primers for a PCR reaction. In order to use the PCR product directly for next generation sequencing (NGS) the primers have sequencing tails (free hanging ends in blue, and blue-red). After the use of inverse PCR the corresponding PCR product is presented with different colours, from left to right it contains: blue - a universal anchor to bind sequencing platform, red - a universal sequence that is used as a starting position for the actual sequencing, green - the bait sequence corresponding to where the primer bound, yellow - the captured sequence, green again - the bait sequence corresponding to the second primer, and at the end in blue the NGS anchor. The library of such products is the input to NGS where a platform like HiSeq 2000-2500 is used to provide base-pair resolution of the ligated fragments known as "reads" beginning from anything after the universal starting sequence (red). The first 20-25 bases of the reads will be from the bait sequence and they will end with the restriction enzyme recognition site used for library preparation. The following base pairs will be these of the captured sequence that should be mapped back to the reference genome.

The usual output of a high throughput sequencing process is in FASTQ or FASTA format. A FASTQ file is a text file that contains the sequence data (reads) with quality scores encoded in ASCII. A standard format for a FASTQ file uses four lines per sequence, with the first line to use "@" character followed by header or a name of the sequence, line two with the raw sequence, line three with an optional character like +, and line four with the quality values for each sequence. A FASTA file is a text file that is used to represent nucleotide sequences using single letter codes. A FASTA format begins with a greater-than symbol ">" and the description-title of the sequences and the line two contains the sequence.

In general, the 4C-seq interpretation should take into account the fact that the design of the library of ligation products that represents genomic fragments is relatively complex and the output is derived from a large number of cells that may contain different allelic forms of the viewpoint and contacts. In that way each allele or cell can display different interaction patterns at the time of library preparation, and this can lead to the conclusion that 4C-seq data represent a pattern of interactions being the result of a distribution over all sampled cells. Therefore, the results of the data analysis should be interpreted in terms of frequency of contacts in the tissue under investigation.

1.2. The purpose of 4C

The 4C technique was first introduced as a variation of the 3C technique together with Hi-C (all-to-all) and 5C (many-to-many) techniques. The choice of the technique depends on the biological question that needs to be examined. 4C-seq, in particular, that can uncover all regions interacting with a given "bait" or "viewpoint", is considered as a valuable approach between placed the 3C (one-to-one, based on a hypothesis, gives few data points) and Hi-C technique (all-to-all, no restriction on of interactions, giving big amount of data) (Davies et al., 2017). 4C-seq experiments rely on primer pairs designed specifically for the bait, which reduces a potential bias that can arise from non-specific primer approaches of other similar techniques such as 3C and 5C (Davies et al., 2017).

Since the beginning of 4C technology, scientific community was able to explore its ability in clarifying biological questions through a large number of scientific studies. The first approach that spread the important role of 4C came in two articles in the same journal in 2006 that examined the H19 locus (Zhao et al., 2006) and the organization of chromatin (Simonis et al., 2006). The development of 4C helped to achieve a better understanding of the relation between the 3D chromosome organization and gene expression (Simonis et al., 2006; Zhao et al., 2006; Schoenfelder et al., 2010). For instance, chromosomal contacts made by the β-globin Locus Control Region (LCR) in embryonic liver, in which the gene is active, were shown to be absent in tissue where the gene is inactive (Simonis et al., 2006).

Resolution is important in chromosome conformation capture experiments. The resolution in 4C experiments is determined by the size of the fragments obtained by digestion with the first restriction enzyme. The ability of 4C to provide high resolution in specific interactions, such as enhancers and promoters, made them very useful during the last decade, mainly in understanding phenomena such as chromosome inactivation (Splinter et al., 2011), detection of regulatory enhancers and promoter interactions (Ghavi-Helm et al., 2014 , van de Werken et al., 2012 a), choice of complex translocation partners (Simonis et al., 2009) and the role of three-dimensional architecture in transcriptional regulation (Noordermeer et al., 2011). Today, several studies used 4C-seq to better analyse and understand biological problems like tRNA functions (Raab et al., 2012), breast cancer by analysing chromosome interactions that regulate transcription (Zeitz et al., 2013) and topologically associating domains (TADs) (Denker and de Laat, 2016; Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012) with the use of multiple viewpoints. The identification of regulatory elements plays a significant role in molecular biology, but the volume of 4C experiments may be challenging for some laboratories. A modification of 4C-seq protocol that can reduce the PCR bias and facilitate a greater capture of reads has been proposed (Brettmann et al., 2018)

In addition, 4C gave the ability to reveal the preferential association of imprinted regions with one another (Zhao et al., 2006; Sandhu et al., 2009) and other types of long-range chromosomal interactions such as observed between Polycomb domains in Drosophila (Bantignies et al., 2011, Tolhuis et al., 2011), identification of insulator binding proteins such as the Drosophila Boundary Element-Associated Factor (BEAF) (Shrestha S et al., 2018), or large-scale chromosome organization in Arabidopsis (Grob et al., 2013). Altogether, 4C indicated that large-scale chromatin organization can be conserved in different cell types and to some extent even among different organisms (Woltering al 2014; Lupianez et al., 2015), whereas fine-scale interactions (enhancer-promoter or gene-gene interactions) are rather cell type-specific (Simonis et al., 2006) and take place between regions that share a similar chromatin landscape (Noordermeer et al., 2011a, b; Grob et al., 2013). Nowadays, 4C still shapes our understanding of the relation between chromosomal interactions and gene transcription, also through studies illustrating its importance in diseases (Lupianez et al., 2015; Vicente-García et al., 2017; Loviglio et al., 2017), thus emphasizing the need for user friendly and reliable data 4C analysis tools. Finally, tissue- and cell type-specific 4C-seq analysis may allow the identification of chromosomal interactions that are specific for certain tissues or cell-types (Hövel, 2016)

1.3. Survey of existing data analysis methods

During the last decade, in parallel with discussion around 4C-seq protocol and its applications, various pipelines and tools for the analysis of 4C-seq data have been developed. The first tool which provided a schema for the analysis and visualization of 4C-seq data was *4Cseqpipe* (van de Werken et al., 2012 b). It was followed by *foursig* (Williams et al., 2014), a method using random relocations of data to select significant interactions; by *FourCseq* (Ghavi-Helm et al., 2015) based on modelling of the profile of contacts around the viewpoint; by *4C-ker* (Raviram et al., 2016), a Hidden Markov Model based pipeline to identify genome wide interactions; and by *w4Cseq* (Cai et al., 2016), a computational and statistical approach to analyse 4C-seq data from both enzyme digestion and sonication protocols based on binomial modelling. A thorough comparison of these methods was done by Zisis et al. (2020).

The above methods, through their basic algorithms, can cover all steps of a 4C-seq analysis starting from the pre-processing of next-generation sequencing reads, the creation of in-silico libraries of restriction fragments, alignment of reads (using own algorithms or public domain mappers), and ending with different forms of results that can be used either for further analysis or visualization purposes. They can provide scripts that will return files of count reads or, in some cases, scripts to support comparative analysis of experimental variants.

In addition to the above tools, there are software packages to process the 4C-seq results which are not complete tools. A pipeline described by Stadhouders et al., (2012) provides a protocol with all materials, critical steps and bioinformatics tools required for successful application of 3C-seq technology. Furthermore, useful tools for processing of 4C-seq data are contained in the R package *Basic4Cseq* (Stadhouders et al., 2014), which provides methods for basic operations on 4C reads and sequences, but refers to other methods of bias removal and inference like those of van de Werken al., (2012). The tool *R3Cseq* (Thongjuea et al., 2013) is an R/Bioconductor package designed to perform 3C-seq data analysis in a number of different experimental designs and many operations of this package are the basis of the *foursig* (Williams, et al., 2014) and *FourCseq* (Ghavi-Helm Y et al., 2014; Klein FA et al., 2015) methods.

1.4. Problems of existing methods

The computational tools for analysis of 4C NGS data mentioned above offer different data input options and differ in the treatment of the data. They also offer different summaries and visualizations of the results. The native characteristics of the 4C-seq technology lead to a number of biases related to the properties of the fragments such as the distance of a 4C fragment from the viewpoint or the presence or the absence of the secondary restriction site in the 4C fragment. In 4C-seq experiments fragments containing at least one secondary restriction enzyme recognition site are called “non-blind” and fragments lacking this site are called “blind”. Accordingly, these properties have to be respected and studied during any analysis with different 4C-seq tools.

Another important issue of 4C-seq is connected with the change of the signal coverage according to the location of the contact in the genome. Although the existing methods proposed some treatment of the data obtained for restriction fragments distant from the

viewpoint, the full analysis is usually only performed for interactions in *cis* (on the chromosome arm containing the bait) or even only for interactions very close to the viewpoint (in regions located not more than 5-10 kb away; van de Werken et al., 2012 b) and not in trans (between chromosomes). In 4C-seq analysis also the terms "*near-c*is" or "*far-cis*" are used to characterise the interactions that are close to the viewpoint or long range interactions in the same chromosome with the viewpoint (Raviram et al., 2014). Several authors justify this focus by the larger coverage, hence higher reproducibility of the signal in these regions. The challenge of working with different coverage levels in different regions of the genome consists of two issues (Raviram et al., 2014): low coverage in *trans* and a decreasing coverage in cis with increasing distance from the viewpoint (bait). The first issue suggests that an optimal use of all information obtained from sequencing is important (Raviram et al., 2015). For the second, approaches like the one proposed by *FourCseq* based on modelling of the coverage decay around bait may be useful; however, they do not work if the coverage in this region is low (Zisis et al., 2020).

Another important topic of 4C-seq is connected with the resolution. Until now, mainly 6 bp cutters have been used for *cis* and *trans* interactions (Rocha et al., 2012; de Wit E et al., 2013) and 4 bp cutters for interactions around the viewpoint (Ghavi-Helm et al., 2014; van de Werken et al., 2012; Noordermeer et al., 2011). A six bp cutter enzyme can achieve a resolution of around 3-4 Kbp and a four bp cutter can achieve a resolution around 200 bp. This difference between cutters affects the reproducibility of 4C signal between replicates in far-*cis* and *trans* (Raviram et al., 2015). The selection of restriction enzyme utilized in each experiment can directly play a critical role in discovery of interactions.

In relation to reproducibility and resolution, which are two topics that must be respected in 4C-seq experiments, the concept of "genomic window" has appeared. Genomic windows (or sliding windows) are specified genomic intervals that "slide" across the genome, usually by some constant distance. These windows are mapped to files containing signal or annotations of interest and can overlap or be disjoint. Overlapping windows are often used to "smooth" signal, or to remove or reduce the impact of signal noise. In 4C-seq analysis sliding windows include a desired number of restriction fragments and the total numbers of reads in each window are determined for the observed data (like, e.g., in *foursig* or *r3Cseq* pipelines). It is highly important to properly select the size of the genomic window used to analyse the interactions in *4Cseqpipe*, *foursig*, and *r3Cseq* procedures. In order to be able to find the appropriate size, methods should find the resolution at which interactions are reproducible. Usually it can be assumed that in regions close to the bait the window size can be smaller because the coverage is large, whereas in regions far from the bait the window size should be bigger as the coverage decreases. This relation of coverage and reproducibility with the location of the bait doesn't allow us to apply a uniform approach across the genome (Raviram et al., 2015).

Furthermore, an important issue of 4C-seq data analysis is the data binarization. Some existing methods (like *4Cseqpipe*) use binarization, i.e., a transformation of data to a score of zero or one, of the signal based on the presence or absence of reads in the restriction fragment. To investigate and deal with the PCR artifacts or identical reads that are usually discarded in other techniques, some pipelines are using binarization and some other like Williams et al.

(2014) used identical barcodes included in the reads (between sequencing adapter sequences and 4C genomic primer sequences) in their experimental strategy and found no selective bias in their usage after mapping. This suggests that binary transformation removes experimental information that can be obtained from the number of captured interactions at a given restriction enzyme fragment (Raviram et al., 2014). In case of the *w4Cseq* which uses operations based on binarized coverage and bionomial distribution, the numbers of resulted interactions were limited or zero in experiments in *A. thaliana* or *M. musculus* (Zisis et al., 2020).

Finally, normalization of 4C-seq data, as a special case of NGS data, is a wide topic of discussion. Current methods provide tools to normalize and visualize signal across the genome. However, normalization methods have been designed based on the needs of the experiments and none of the existing 4C data normalization methods have yet been generally accepted as a reference method. The simplest method is the normalization to the total number of obtained reads, a step that can be done in any pipeline, but is not provided as a function. There are methods that suggest normalization mainly for comparison of experimental variants by using tools like *Deseq2* (Williams et al., 2014; Klein et al., 2015). However, only the pipeline of van de Werken et al. (2012) considers particular properties of the restriction fragments that may influence, and, consequently, bias the results in the algorithm. In absence of a universally accepted normalization method, the genomic regions indicated as being in frequent contact with the bait should, in principle, be characterized by higher numbers of reads mapped to fragments within these regions. We found that this requirement holds for some methods like those described by Williams et al. (2014) or Klein et al. (2015) whereas it does not hold for others like the one described by Raviram et al. (2016). The use of different normalization methods may have a large effect on comparisons between experimental variants (differential analysis).

2. The aim of the thesis

The goal of this thesis is to propose a new approach to analysis of 4C-seq data. We analysed and compared existing algorithms that are used to analyse such data (Zisis et al., 2020). These algorithms concentrate – in the first step - on finding, in each sample independently, genomic regions that exhibit large frequency of contacts with the bait called "significant contacts". The comparative analysis of experimental variants, which, according to the standard experimental requirements, must be based on several biological replications, is a second step provided only in some algorithms and is based on (or, is conditioned by) the significant contacts found. Our observation was that the sets of significant contacts obtained with different methods differ considerably, which may cause problems with confidence in the results of experiments. We formulated:

Desideratum 1: Significance of contacts in 4C-seq experiment should be defined in relative terms, based on the purpose and design of the experiment and applied controls.

Our second observation was that the differences between various methods are caused by concentrating on different features of contacts. In general terms, contacts are "genomic signals" (Dougherty et al., 2009), and as such should be characterized by at least three parameters: genomic location, width of the interval, intensity, and possibly others, like stability or frequency within given intervals. We formulated:

Desideratum 2: Characterization of contacts in 4C-seq experiments should be done with respect to various parameters of genomic signals.

To design a methodology that is coherent with these *desiderata* we:

a) Identified transformations of raw sequencing data to variables that reflect well the properties of genomic signals.

b) Identified well-founded statistical methods that can be used for assessing significance of differences between experimental variants based on these variables.

c) Selected data sets with contrasting properties that could be used to assess the quality of functioning of the methods.

To perform the computational tasks, a new pipeline was produced called *4CseqR*. The pipeline is combining required algorithms and statistical methods and is based on existing open source tools and own scripts written in R statistical language.

3. Material and methods

3.1. Data sets

We use two data sets obtained for two different species - *Arabidopsis thaliana* and *Mus musculus*. The *Arabidopsis* data were generated in three biological replications under two different treatments; the mouse data were obtained also in three replications, but for two different cell lines.

Dataset 1

The *Arabidopsis thaliana* (genome size 135 Mb) data used in this study were generated using BglII (AGATCT – 6 bp cutter) as a primary restriction enzyme and NlaIII (CATG – 4 bp cutter) as a secondary restriction enzyme (data obtained by Hövel, 2016, PhD thesis; submitted to Arrayexpress data base, https://www.ebi.ac.uk/arrayexpress/, accession number E-MTAB-7629). The experiment concerned the changes in the genome-wide contacts of the FLOWERING LOCUS C (FLC) gene (At5g10140) responsible for vernalization serving as the bait and included two types of plants (experimental variants): A - non-vernalized plants and B - plants subjected to vernalization for 4 weeks plus 7 days warmth. Three biological replications of each variant were prepared. In the following, we refer to 6 samples by the identifiers At_A and At_B according to variant A and B. By sequencing on the Illumina HiSeq 2000 platform, six data sets of 50 bp NGS reads were obtained, with 10-12 million reads for variant A and 10-16 million reads for variant B per sample, respectively.

Dataset 2

The Mus musculus (genome size 2.7 Gb) data that we use are part of publicly available 4C-seq data obtained for a variety of different lines of embryonic stem cells (ESC) (Denholtz et al. 2013). For our analysis we selected six data sets concerning three biological replications made for the ESCs (experimental variant A; NCBI GEO, https://www.ncbi.nlm.nih.gov/gds, accessions GSM1212827, GSM1212828, GSM1212829 identified as mm_A_1, mm_A_2, mm_A_3) and induced pluripotent stem cells, iPSCs (variant B; accessions GSM1212871, GSM1212872, GSM1212873; identifiers mm_B_1, mm_B_2, mm_B_3). 4C-seq was performed using HindIII (AAGCTT - 6bp cutter) as a primary restriction enzyme and DpnII (GTAC- 4bp cutter) as a secondary restriction enzyme, with the Pcdhb19 (ENSMUSG00000043313) locus as the bait. Sequencing data obtained with Illumina HiSeq 2000 contained 4-14 million reads for ESCs samples and 5–11 million reads for iPSC samples.

Table 1. Characteristics of the data sets used in this study

Characteristics	Data set 1	Data set 2
Reference genome	*Arabidopsis thaliana* (TAIR10)	*Mus musculus* (mm10)
Primary restriction enzyme (recognition site)	BglII (AGATCT)	HindIII (AAGCTT)
Secondary restriction enzyme (recognition site)	NlaIII (CATG)	DpnII (GTAC)
Bait locus	At5g10140	ENSMUSG00000043313
Bait primer sequence	CAAACCCCATAGCAACTCTATAGATCT	TCAAATGGAGGCACATAAAGCTT
Bait chromosome	Chr5	Chr18
Bait start	3173487	37478309
Bait end	3178614	37483219

3.2. General schema

In this section we describe the proposed methodology of analysis of 4C-seq data forming a pipeline called *4CseqR*. The consecutive stages of the pipeline are visualized in Figure 2. The input to *4CseqR* is the NGS data from a 4C-seq experiment in fastq format called "reads". The *4CseqR* pipeline will prepare the library of restriction fragments, pre-process the reads, map reads to the library of fragments, estimate coverage of restriction fragments and normalize the coverage, before submitting data for statistical analysis. Outputs of each step can be used either for visualization with a genome browser or for further statistical analysis. Two statistical models are proposed for comparative analysis of experimental variants with respect to frequency of contacts in the genome. All scripts forming the *4CseqR* pipeline are available at https://github.com/DimitrisZisis/4CseqR_2020.

3.3. Preparation of library of fragments

A proper, or legitimate, 4C NGS read comes from either the 5' or 3' end of a primary restriction fragment, that is, the genomic fragment bound by two consecutive occurrences of the recognition sequence of the primary restriction enzyme. The "fragment ends" of the restriction fragments are defined as the regions between the primary restriction site and the nearest secondary restriction site (van de Werken et al., 2012a). For mapping of reads in the reference genome, we use a genome that is prepared in the form of the library of restriction fragments. In *4CseqR* scripts in R, Python and bash are used to produce the library of fragments.

The script *create_fragments.R* uses the *createVirtualFragmentLibrary* function from Basic4Cseq tool (Walter et al., 2014) and the *BSgenome* Bioconductor (https://www.bioconductor.org) library to produce the list of restriction fragments. The recognition sequences of two restriction enzymes have to be specified. The result is saved in bed format for further operations. Then, *getfasta.sh* script (utilizing *fastaFromBed* function

from *bedtools*) is used to obtain genomic sequences of restriction fragments on the basis of the reference genome.

4CseqR provides also scripts that can extract additional information on restriction fragments, like length. A script *find_blind.py* in Python is provided to characterize each fragment as blind or non-blind; this information is used further in the normalisation step.

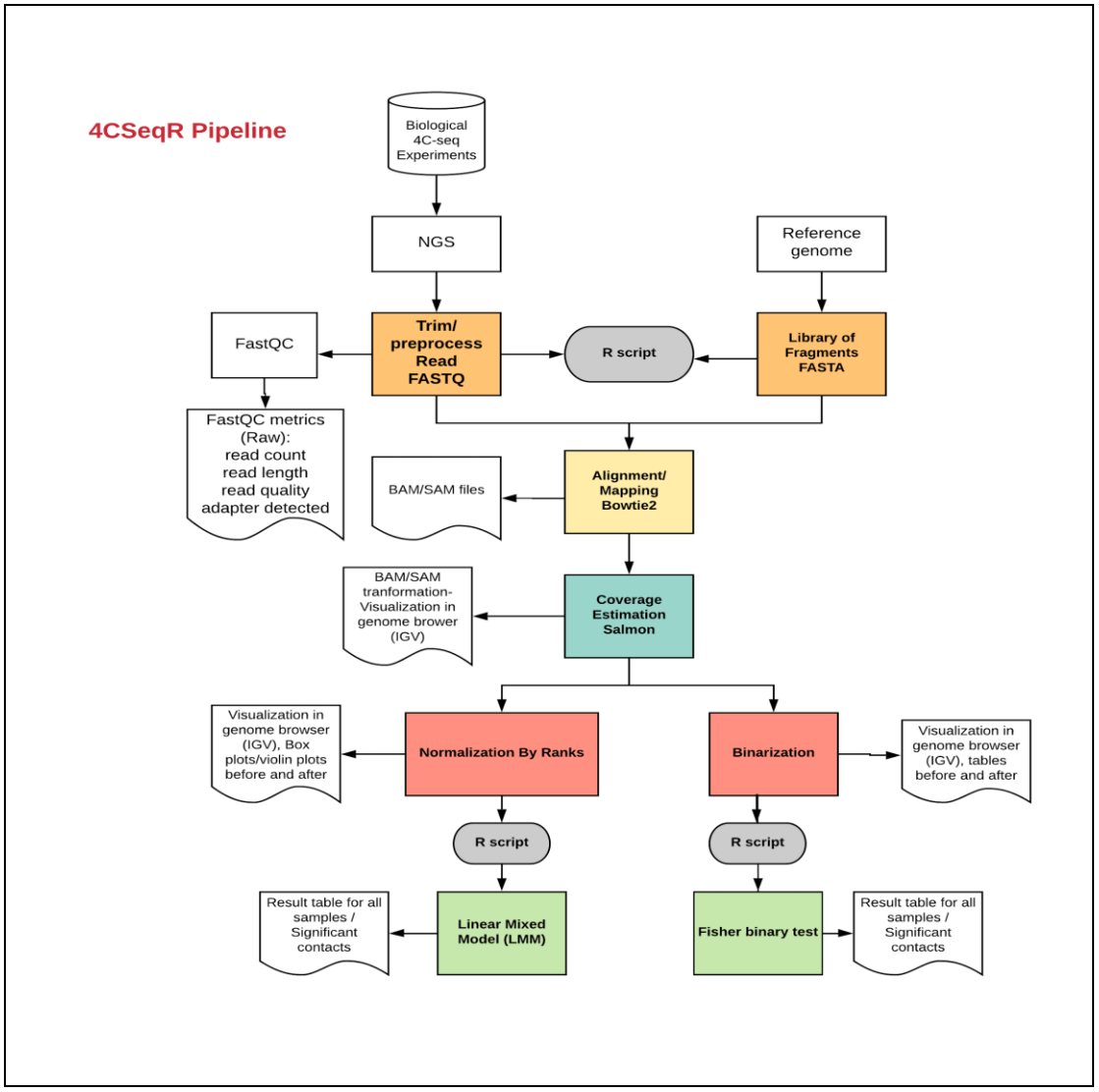

Figure 2: The *4CseqR* analysis workflow starting from raw 4C-seq data in fastq format and going step by step to quality check (*FastQC*), pre-processing by trimming/filtering reads (R scripts), creation of a library of fragments (R script), mapping to a reference genome (*Bowtie2*), coverage estimation (*Salmon*), normalization and binarization (R scripts) until visualizations (*IGV genome browser*) and comparative statistical analysis (R scripts) that declares significant contacts and significant differences in contacts.

3.4. Data preparation and pre-processing

Quality Control

Quality check of 4C-seq data is an important step which should be done before any other processing. Like for other types of NGS data, filtering can be done on the basis of the per base quality scores. To do this *4CseqR* uses the *FastQC* program (Andrews, 2010) which provides a standardized and in-depth report of the quality of the reads. A report file is generated that contains quality metrics of all 4C libraries processed by the pipeline. The quality scores that are presented are defined as per base sequence quality and per sequence quality scores (Phred scores).

Pre-processing

Raw NGS data which are coming from next generation sequencing need special pre-processing that takes into account their origin. In consequence of the applied experimental protocol, each 4C-seq read consists of three parts: the sequence of the bait, the recognition site of the primary restriction enzyme, and the capture sequence - a sequence belonging to the genomic region contacting the bait. Thus, the reads as they are given in the fastq file are not sequences that occur in the reference genome. Before further steps, the contact sequences have to be separated, together with the restriction site, or the viewpoint sequences have to be trimmed from the NGS reads. The reads that do not contain the primary restriction enzyme recognition site must be filtered out.

4CseqR deals with this operation with an R script using a fuzzy algorithm (i.e., allowing for mismatches) to search for the primer sequence in all reads and produce a fastq file with the appropriately trimmed reads. More specifically, *4CseqR* trims and filters the original reads to produce sequences that start directly at a restriction enzyme recognition site. The 4C-seq data obtained for all samples are first processed by filtering out the reads that do not contain the primary restriction site. Then, the trimmed reads are filtered in order to remove reads which are shorter than a limit specified according to the needs of the experiment (very short reads are removed before mapping because they don't provide a proper biological information and create noise in the mapping step afterwards). For filtering and trimming the primer sequences from the raw reads the R function *select.reads* is provided; as input, it takes the raw read files and the enzyme recognition and primer sequences.

3.5 Read alignment

After construction of the library of fragments and selection of legitimate, so potentially mappable, NGS reads, the *4CseqR* pipeline proceeds with mapping the reads to the library of fragments. Mapping is conducted by *Bowtie2* (Langmead et al., 2009) allowing for up to 2 mismatches in the contact region and without mismatches in the restriction enzyme recognition sequence. A script (*mapping.sh*) is provided for this step.

3.6 Fragment coverage estimation

Fragment coverage estimation consists in, first, counting the reads mapped to each restriction fragment on the basis of mapping results. Then, it involves targeting the issue of repetitive genomic sequences and their influence on the estimation results. In general, performing a non-adjusted analysis based on all, also non-uniquely, mapped reads is not valid as the resulting estimated coverage is bigger than the true one. On the other hand, using just uniquely mapped reads means losing some information. The problem can be viewed from two sides: uniqueness of sequences of restriction fragments or uniqueness of read mapping. Estimation of fragment coverage from mapping results is an important step of the data analysis. It should assure the preservation of a maximum of information usable for detection of genomic regions with a frequent contact with the viewpoint and, at the same time, removal of data artefacts that may increase the fraction of false declarations.

We propose estimation of the fragment coverage by the application of the method described by Patro et al. (2017) and implemented in the *Salmon* tool. *Salmon* supports two modes of operation: either with a FASTA file containing a reference genome and a set of reads in a FASTQ file, or a set of pre-computed alignments in a SAM/BAM file. We use the second mode of Salmon by taking the output of mapping with the alignments in SAM format and the library of fragments as reference in FASTA format. *Salmon* is an ultra-fast quantification algorithm, devised for RNA-seq data that uses a probabilistic model to estimate the number of reads coming from genomic sequences contained in the sequenced library. As shown by Zhang et al., (2017), it is preferred to other methods and software like *eXpress* (Roberts and Pachter, 2013) or *Kalisto* (Bray et al., 2016). One of the principles of this method is that reads or alignments are made “random”. That means that alignments are not sorted by target or position.

In case of 4C-seq data, this method can be applied to alignment results (in SAM format) that are obtained by mapping the reads to a reference genome or, like in case of *4CseqR*, to the library of restriction fragments without the requirement of "uniqueness"; such results represent the total numbers of reads that were mapped to restriction fragments. The fragments are specified as "random targets" in the *Salmon* method. The estimated fragment coverage, resulting from distributing the reads over fragments according to the estimated probabilities, is a number that is equal to the total number of mapped reads if all reads were mapped to this fragment uniquely, and a number smaller than that otherwise. A bash script to run *Salmon* (*salmon.sh*) is provided by the *4CseqR* pipeline.

3.7 Finding significant contacts

After the computation of fragment coverage, the analysis continues with the selection of fragments or groups of fragments for which there is an evidence of being in contact with the viewpoint. As indicated in the Introduction, in contrast to other existing algorithms, our method does not find significant contacts for each analysed sample independently, but is looking for significant differences between experimental variants; the relevant methodology is described in Sec. 3.9. However, we use one of the existing methods of 4C data analysis, *foursig* (Williams et al., 2014), for the purpose of comparison of results.

As it is described in Zisis et al., (2019), the existing software packages use different methods to find significant contacts of genomic regions with the bait. *Foursig* is a tool written in R and Perl. It uses a transformation of data into a database of genomic fragments to identify statistically significant interactions of 4C-seq data. The database that is produced is designed for a specific 4C experiment and contains genomic fragments based on the restriction enzymes used to generate a 4C library. When the database is ready the mapped 4C-seq reads are assigned to create a file that is the input in *foursig* program. The program will remove the unmapped fragments of the database and will give the opportunity to the user to define masked areas before any significance analysis. The significance testing is based on the estimated distribution of reads among restriction fragments and on random relocation of reads. The method permits to decide - for a given false discovery rate - the threshold (i.e., the minimum number of reads covering a set of consecutive fragments) for calling a genomic window as being in frequent contact with the bait. Further, the procedure classifies the genomic windows displaying significant contact frequencies as "broad" (Category 1), "intermediate" (Category 2) and "narrow" (Category 3), based on the number of fragments influencing significance within the window.

3.8. Sliding window strategy

In both of the data analysis methods proposed (see Section 3.9), 4CseqR uses a sliding window strategy. Following transformation of fragment coverage data, a set of sliding windows W_{ij} with length i bp and step j bp is considered. All restriction fragments included in the specified window are used in the analysis. Different parameters of sliding windows (length, step) can be applied for different experimental strategies. In what follows, a window (a set of restriction fragments) which is found to be in contact with the bait to a different degree in different experimental variants will be called a "differentially contacting window" (DCW). Each DCW is characterized by estimated parameters of contact intensity or probability, and by a P value corresponding to the test of the null hypothesis saying that there are no differences between the parameters in different the experimental variants.

After application of the sliding window procedure, overlapping differentially contacting windows are merged to produce "differentially contacting regions" (DCR) characterized by differential contacts with the bait. A DCR is characterized by statistical parameters averaged over all DCWs contained in it.

3.9. Comparative analysis of experimental variants

Proposed analysis provides the ability to identify genomic regions with significant differences with respect to contact with the bait which are remote or proximal to the viewpoint. By using two different transformations, which are further described, the fragment coverage is transformed to data characterizing the intensity and width of the contacts. Then, a sliding window strategy is applied to assign statistical significance to genomic regions based on statistical models appropriate for the transformations. In a specific window, a group of fragments is selected and if the (corrected, see Sec. 3.10) P value which is calculated is below

a specific threshold then genomic window represents an area with a statistically significant difference.

3.9.1 Normalization by ranks and linear mixed model

In this section we describe the first of the proposed methods of data analysis aimed at finding genomic regions which contact the bait differently in different experimental variants in a factorial experiment. The method is based on transformation of coverage data to ranks followed by a transformation to normal distribution, and comparison of experimental variants using a linear mixed model (LMM; Searle et al. 2006, Galecki and Burzykowski, 2013).

Van de Werken et al. (2012a) defined a background model for remote contacts by assuming that the coverage may depend on whether the fragment is blind or not blind, or on the total primary restriction fragment length. Such an empirical model is based on, theoretically, different chances of fragments with different characteristics being amplified and sequenced. We use the same principle here. More specific, the coverage estimated by *Salmon* is transformed into ranks within different categories of fragments (blind/non blind, length of fragment ends, total length of fragment).

4CseqR provides a function in R (*normalization_salmon_data.R*) that transforms the coverage estimated by *Salmon* into normalized data. The program uses the basic form of the rank function and it produces a vector that contains the rank of the values in the vector that was evaluated. The counts of reads are normalized according to the procedure based on ranking within classes of fragments. To make the results independent of the numbers of fragments in categories, ranks within each category are expressed in the scale from 0 to 1. In order to achieve a more accurate result in ranking, fragments with zero values are excluded from normalization. By definition, this transformation removes differences in the coverage distribution between the categories.

Furthermore, we transform the non-normally distributed ranks into observations that follow the normal distribution. In order to do this the rank-based inverse normal transformation (ITN) is used (Beasley and Erickson, 2009). It is of the form

$$Y_i^t = \Phi^{-1}\left(\frac{r_{i-c}}{N-2c+1}\right)$$

where r_i is the rank of the *i*-th observation among the N observations and Φ^{-1} denotes the standard normal quantile function. The parameter *c* is recommended to be *c* = 3/8 by Blom (1958). INT transformation is available in statistical packages with the option of setting different values of *c* (e.g., van der Waerden, 1952, suggests *c* = 0). Blom version of this transformation seems to be most commonly used. In case of *4CseqR* the parameter *c* can be specified in the function and *c* = 0 is used as default. After testing different recommended values of the *c* parameters we concluded that changing its value does not make a considerable difference in the results.

The rank and INT transformations provide the values of normalized coverage of restriction fragments. We propose to use the linear mixed model (LMM) as a tool of data analysis in order to find genomic windows which can be inferred to be in contact with the bait to a different degree in different experimental variants. The linear mixed model is an extension of

a simple linear regression model, but with an extra complexity represented by random effects. The LMM may describe data collected in an experiment with one or more factors with fixed effects as well as one or more factors with random effects. The random factors usually are qualitative variables whose studied levels are random samples from a bigger population of levels.

In the general form the LMM for the case of n observations of a quantitative variable is written as (Searle et al. 2006)

$$y = X\beta + Zu + e,$$

where **y** is a $n \times 1$ vector of observations of the outcome variable, X is a $n \times p$ matrix of values of p predictor variables related to factors with fixed effects, β is a $p \times 1$ vector of fixed effects (regression coefficients), Z is the $n \times q$ matrix of values of q predictor variables related to factors with random effects, u is a $q \times 1$ vector of the random effects, and e is a $n \times 1$ vector of the residuals that represent variation of y not explained by the model. In a simple form of LMM, it is assumed that the elements of u are independent and identically distributed random variables with normal distribution N $(0,\sigma_u^2)$ and elements of ε are independent and identically distributed random variables with normal distribution N $(0,\sigma_e^2)$.

In our case, we consider as the outcome variable the vector of normalized coverages of restriction fragments belonging to a genomic window W, where W is any of the windows considered in the moving windows strategy. Denote by y_{vfr} the normalized coverage observed in W for experimental variant v, $v = 1,2...V$, in restriction fragment f, $f = 1,2...F$, in biological replication r, $r = 1,2...R$. The linear model with fixed effects of experimental variants and random effects of restriction fragments is of the form

$$y_{vfr} = \beta_v + u_f + e_{vfr},$$

where β_v denotes the fixed effect of the v-th experimental variant, u_f denotes the random effect of the f-th fragment (with expected value equal to zero and variance σ_u^2), and e_{vfr} denotes the residual related to an individual observation. It can be shown that in this model the fixed effects are equivalent to mean values for experimental variants. The LMM fitting procedure allows to estimate the set of parameters (β_1, β_2,…, β_V, σ_u^2) consisting of mean coverages for experimental variants and variance among restriction fragments within the window W. Standard errors of estimated mean coverages are also obtained. The hypothesis about no influence of experimental variants on the coverage, interpreted as the hypothesis of no differences between variants with respect to contacts with the bait, is written as

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_V ,$$

and in the LMM analysis is tested with the use of the F test and assigned to it P value. Rejection of H_0 means that for at least one experimental variant the coverage in W is different than for other variants, which implies differential contacts. In the special case of $V = 2$, rejection of H_0 means that there is a statistically significant difference between contacts with the bait under two experimental variants.

The *4CseqR* performs a linear mixed model analysis through a function called *LMMPerWin*. For a particular sliding window, in which there is a number of restriction fragments, an LMM

model is created for the normalized coverage as the outcome variable and is fitted using the *lmerTest* package in R. The *lmerTest* package fits the LMM and provides *P* values and summary tables. From these results the mean values for experimental variants, their standard errors, the variance for fragments, and the *P* value for the test of fixed effects are stored for further data analysis.

3.9.2 Binarization and Fisher test

In this section we describe the second of the proposed methods of data analysis aimed at finding genomic regions which contact the bait differently in different experimental variants in terms of the frequency of contacts within a given window. The method is based on binarization of the coverage data and comparison of experimental variants using the Fisher test (Fisher, 1922).

The data binarization process works as follows. For each restriction fragment, its coverage is transformed according to formula

$$C(coverage) = \{1\ if\ coverage > R\ and\ \ 0\ if\ coverage\ \leq\ R\},$$

where *R* is a threshold distinguishing the covered and uncovered fragments. The default value proposed by *4CseqR* is $R = 1$.

After data binarization, a statistical test is performed to compare the coverage observed in different experimental variants in each genomic window according to the sliding windows procedure.

In case of two experimental variants, for a given window, a 2 × 2 contingency table is generated with the counts of covered and not covered fragments in two variants, after summation over biological replications (Table 2).

Table 2. Different populations in each variant

	Covered	Uncovered	Total
Variant 1	*a*	*b*	*a* + *b*
Variant 2	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*n*

In the table, *a*, *b*, *c*, *d* denote the total numbers of fragments with a given property over all biological replications and *n* the grand total.

The homogeneity of distributions of covered/noncovered fragments between variants is tested with the use of the Fisher exact test. The test uses a special formula to obtain the probability of the combination of the frequencies that were actually obtained. It also involves the finding of the probability of every possible combination which indicates more evidence of association. The *P* value of the test can be computed as the sum of all probabilities which are smaller than p. In case of the above table the probability of Fisher exact test is given by:

$$P = (a+b)!\,(c+d)!\,(a+c)!\,(b+d)!/n!\,a!\,b!\,c!\,d!$$

If the *P* value comes out to be smaller than a given threshold (say, 0,05), then the likelihood of similarity of distributions between Variant 1 and Variant 2 is very small and we observe a significant difference.

The *4CseqR* program performs the computations using a function called *FisherPerWin*. This function has been created to calculate the Fisher exact test for the binary results, for a particular sliding window with specific step in which there is a number of restriction fragments, using the *Fisher.test* package in R. The *4CseqR* program returns a table with the coordinates of each window, Fisher *P* values and the probabilities of coverage in each variant. The number of covered and uncovered fragments in each window for each condition is also provided for further exploration.

3.10 *P* value adjustment

In both procedures: the one using normalization and the LMM approach, and the other using binarization and Fisher test, we test a family of hypotheses related with sliding windows, and get a set of *P* values for these tests. We adjust these *P* values to take into account possible errors related to multiple hypothesis testing using the false discovery rate (FDR) correction by the Benjamini-Hochberg (1995) procedure.

The procedure can be described as follows. Let's assume we have $H_1, H_2, \cdots, H_m$ null hypotheses and corresponding P values $P_1, P_2, \cdots, P_m$. These P values are ordered as $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, where $H_{(i)}$ is the null hypothesis with $P_{(i)}$.

We are using the standard Benjamini–Hochberg procedure in which:

- A threshold false discovery rate, α, is specified, which by default is set as 0.05 in our case.
- The largest i (= k) is found for which, for a given α, $P_{(i)} \leq \frac{i}{m}\alpha$.

Then, null hypotheses $H_{(i)}, i = 1, 2, \cdots, k$ are rejected.

Following this procedure, it is true that $FDR \leq \alpha$, thus the false discovery rate is controlled. Genomic windows with FDR-adjusted P values ≤ 0.05 are identified as the ones in which there are significant differences between variants with respect to contacts with the bait.

4. Results

In the following section we present the results of the application of the described statistical methods and of *4CseqR* package to the analysis of data described in Sec. 3.1 for *Arabidopsis thaliana* and *Mus musculus*.

4.1 Results for Dataset 1(*Arabidopsis thaliana*)

4.1.1 Quality Control

Below is the summary of the output of *FastQC* program for two samples that shows boxplots of the data quality across all bases of the NGS reads. The y-axis in the graph represents the quality scores and the x-axis the position in bp in the read. The higher the score the better the base call. The background of the graph is divided into three areas with different colours representing different levels of quality: the green part of y-axis with very good quality, the orange part with reasonable quality and the red part with poor quality. The quality scores that were in the range between 25th and 75th percentile are within the yellow boxes. The red line inside the box represents the median value of quality; it is above 38. The data quality is considered as good if the median is bigger than 20. From the graphs it can be seen that all quality scores, even at the end of a read, are in the green part, which proves that there are no sequences flagged as having poor quality in both samples representing variants A and B.

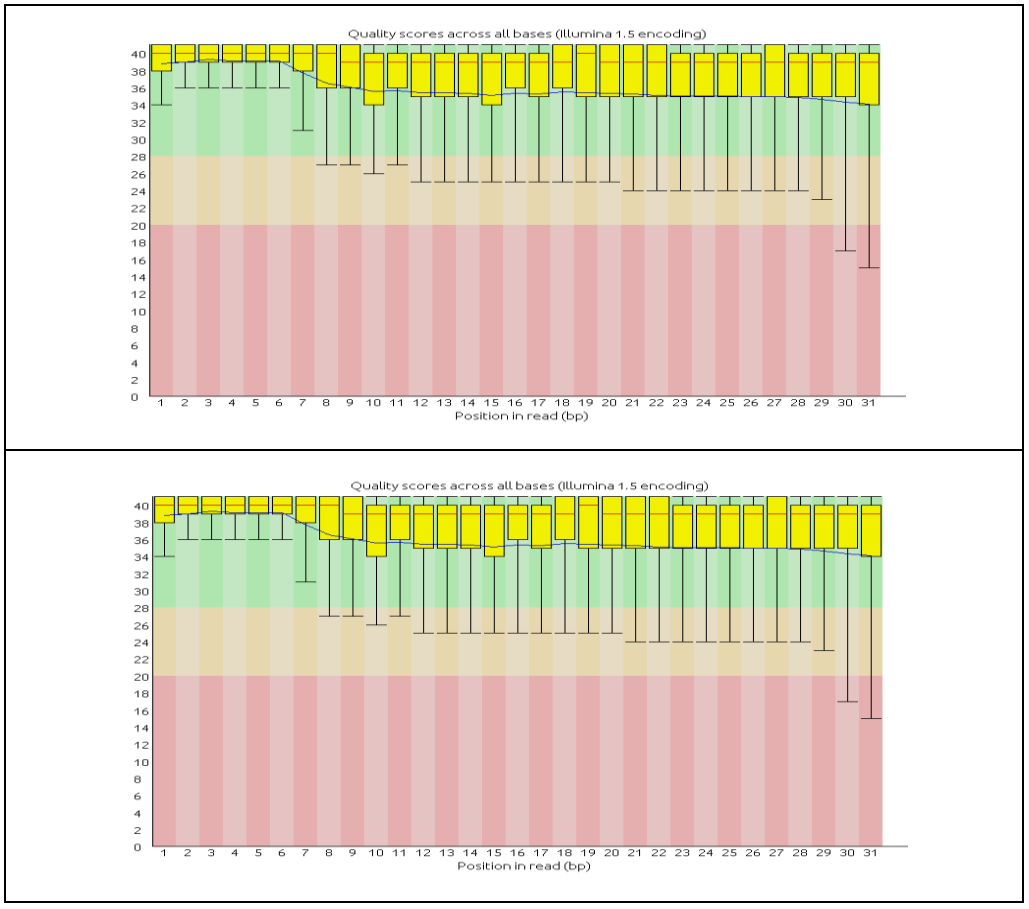

Figure 3. Overview of the quality of NGS reads across bases, at each position, in samples At_A_1 and At_B_1 (first replication of each experimental variant).

4.1.2. Pre-processing and mapping

The NGS data obtained for all samples were first processed by trimming out viewpoint sequences from the reads and removing reads that didn't contain the primary restriction site AGATCT (BgII), as it is described in Methods. The resulting sets of reads were aligned to the library of all AGATCT restriction fragments from the TAIR10 genome used as the reference. For mapping, *Bowtie2* was used with the parameter --score-min L,0,-0.4, in order to achieve mapping with the minimum score = 0.4×31 = 12.4. The "minimum score" is the minimum alignment score needed for an alignment to be considered as "valid". The following Table 3 presents the characteristics of sets of reads related to the pre-processing and mapping process.

Table 3. Results of NGS data processing

Sample	Data file	Number of reads			% reads mapped to genome
		total	after filtering (pre-processing)	mapped to fragment library	
At_A_1	arabidopsis_1_1.fq	16871432	16126982	9928407	61.56%
At_A_2	arabidopsis_2_1.fq	14878396	14448008	10735886	74.31%
At_A_3	arabidopsis_3_1.fq	10365781	10064278	8059346	80.08%
At_B_1	arabidopsis_1_4.fq	12623982	11859951	5735096	48.36%
At_B_2	arabidopsis_2_4.fq	11210240	10990650	7748165	70.50%
At_B_3	arabidopsis_3_4.fq	10702878	10516643	6656010	63.29 %

4.1.3 Coverage estimation

4CseqR contains the *salmon.sh* script which is uses *Salmon* as described in Methods to calculate the estimated coverage, by taking the output of mapping in sam format and the reference genome (library of AGATCT fragments) in FASTA format. A fragment of the output of *Salmon* with the estimated fragment coverage can be seen in the Table 4 and is available in files At_*_salmon.csv (csv format) in Supplementary Files[1] for all samples. The column NumReads is the Salmon's estimate of the number of reads obtained by sequencing protocol from restriction fragments (estimated coverage) as described in Methods and is used for further steps of the analysis in *4CseqR*.

Table 4. *Salmon's* main output is its quantification file (quant.sf). This file is a tab-delimited file with a single header line.

Chromosome	Fragment start	Fragment end	Fragment length	Estimated number of reads from the fragment
chr1	1	7512	7511	344
chr1	7506	10949	3443	0
chr1	10943	13175	2232	0
chr1	13169	20070	6901	1
chr1	20064	28278	8214	0
chr1	28272	29965	1693	0
chr1	29959	36604	6645	0
chr1	36598	37518	920	1
chr1	37512	40365	2853	448
…	…	…	…	…

The distribution of the estimated coverage along all chromosomes and bait chromosome 5 (500 bp around the bait) is shown in Fig. 4. It can be seen that the fragments coverage in the

[1] Supplementary Files are available on CD attached to the manuscript

bait chromosome is greater than in other chromosomes. As to the bait region, it is almost completely covered

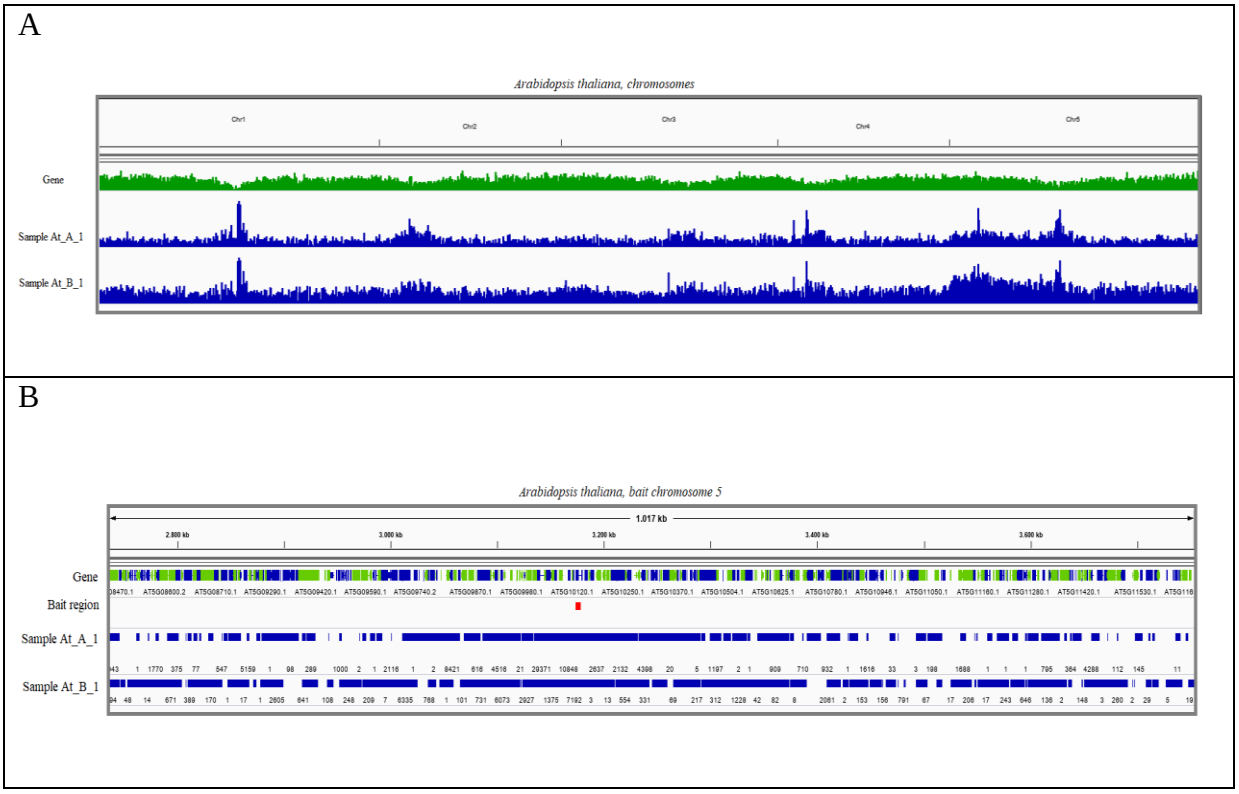

Figure 4. Visualization of the estimated coverage in Integrated Genomic Viewer (Thorvaldsdóttir et al., 2013). A: For all chromosomes, B: for bait chromosome 5.

To further investigate the different levels of estimated coverage of fragments obtained by *Salmon*, *4CseqR* provides an R script that produces violin and bar plots which are presented in Fig. 5. The figure presents data for the first replication of each experimental variant (samples At_A_1 and At_B_1). In both experimental variants the coverage was concentrated mainly in bait chromosome.

The fact that the estimated coverage is higher in bait chromosome than in others can be further seen by calculating metrics like the mean and maximum coverage for each chromosome (Table 5).

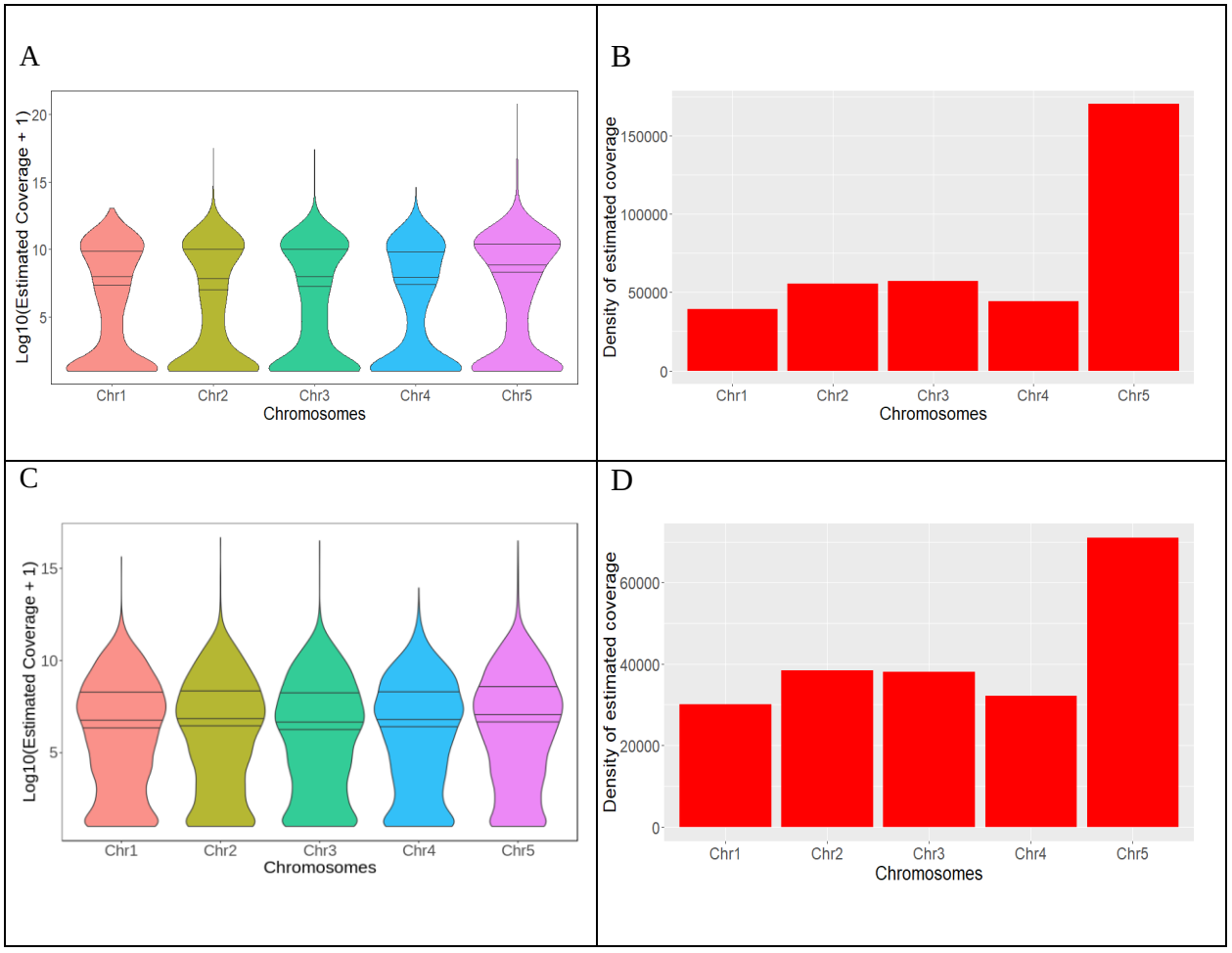

Figure 5: A. Distribution of estimated fragment coverage for sample At_A_1; B. Density of estimated coverage in chromosomes (total estimated genome coverage for a chromosome divided by the length of chromosome in Mb) in A_1. C. Distribution of estimated fragment coverage for sample At_B_1; D. Density of estimated coverage in chromosomes in At_B_1.

Table 5. Characteristics of coverage in sample A_1 in each chromosome

Chromosome	Mean coverage	Maximum coverage
chr1	123.9	8615
chr2	167.5	188493
chr3	167.3	174183
chr4	134.2	24504
chr5	548.7	1793006

4.1.4 Normalization by ranks and linear mixed model

The normalization is performed on the estimated coverage obtained from *Salmon*, with respect to the characteristics of the fragments as it is described in Methods. The distribution of estimated coverage depends on three factors: the length of the fragment, the length of the fragment ends, as well as the existence of the secondary restriction site in the fragment. Those factors create differences in the distribution of estimated coverage and in order to eliminate those differences and improve the analysis we apply a normalization approach described in Methods. Based on this a transformation is proposed of the estimated coverage of fragments into ranks and normally distributed ranks.

The input to the normalization process is coverage of fragments with fragment length, fragment end length (5', 3') and the presence of the secondary restriction site. The following graphs visualize differences in distribution of coverage across different classes and the elimination of those after normalization. The results of normalization with respect to the length of the 5' fragment ends are presented in Fig. 6. This graph is showing violin plots of the distribution of coverage across different categories of fragment end length. It can be seen that fragment ends longer than 150 are relatively less covered. To standardize the data and eliminate as much as possible those differences we normalize the data and the result is visible in Fig. 6B with the normalized ranks and in Fig. 6C with the normally distributed ranks. The resulting distributions are symmetric around 0 with the except of distortion caused by the presence of a large number transformed observations equal to zero (non-covered fragments).

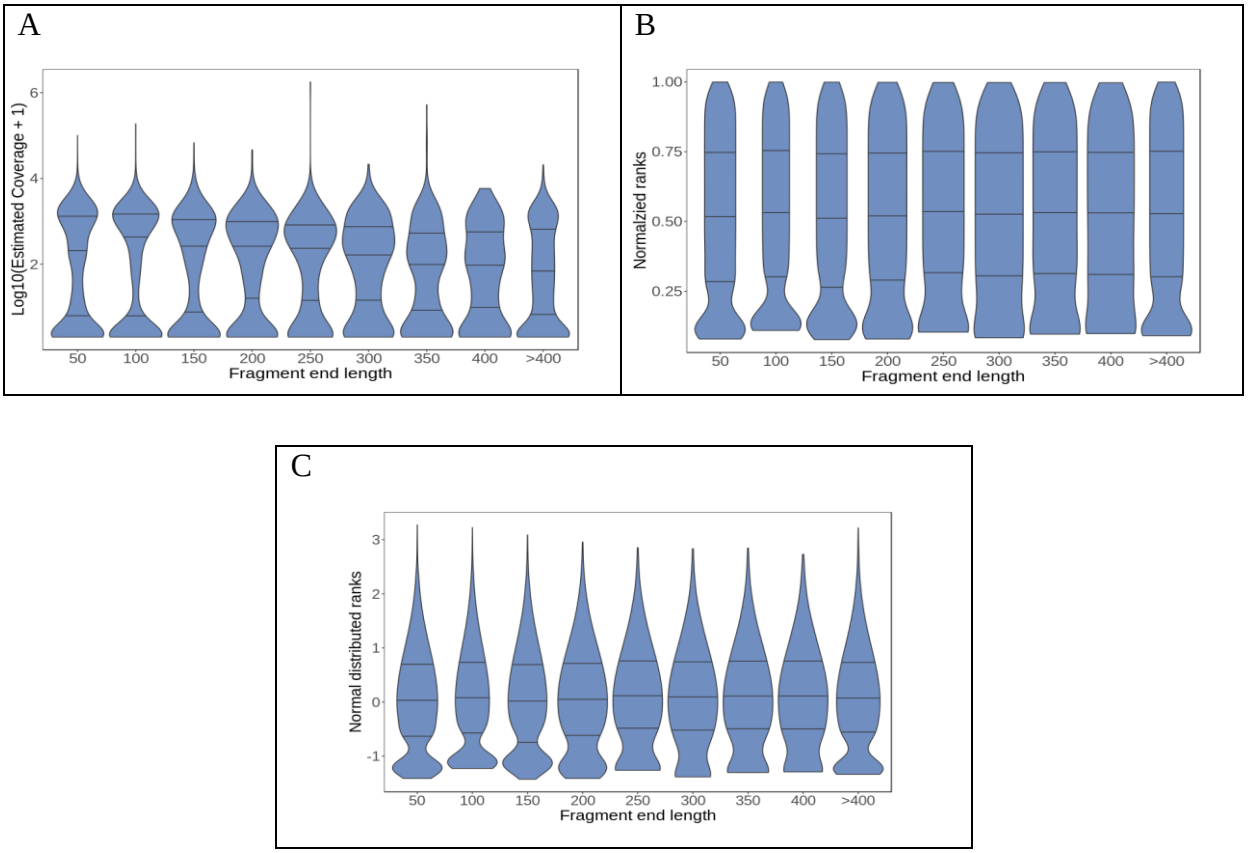

Figure 6. Distributions in categories of 5' fragment end length of the estimated coverage in sample At_A_1. (A) Before normalization, (B) after rank transformation and (C) after normal distributed rank normalization**.**

The transformation is done similarly based on the categories of length of fragments and the results are shown in Fig. 7. In this case in very short fragments, of length about 50 bp, the estimated coverage seems to be smaller than in longer fragments, while the coverage of longer fragments increases with their length (Fig. 7A, Fig. 7B). After transformation into ranks and into normally distributed ranks (Fig. 7C) still there are some differences but the distributions are closer to homogeneity over different classes.

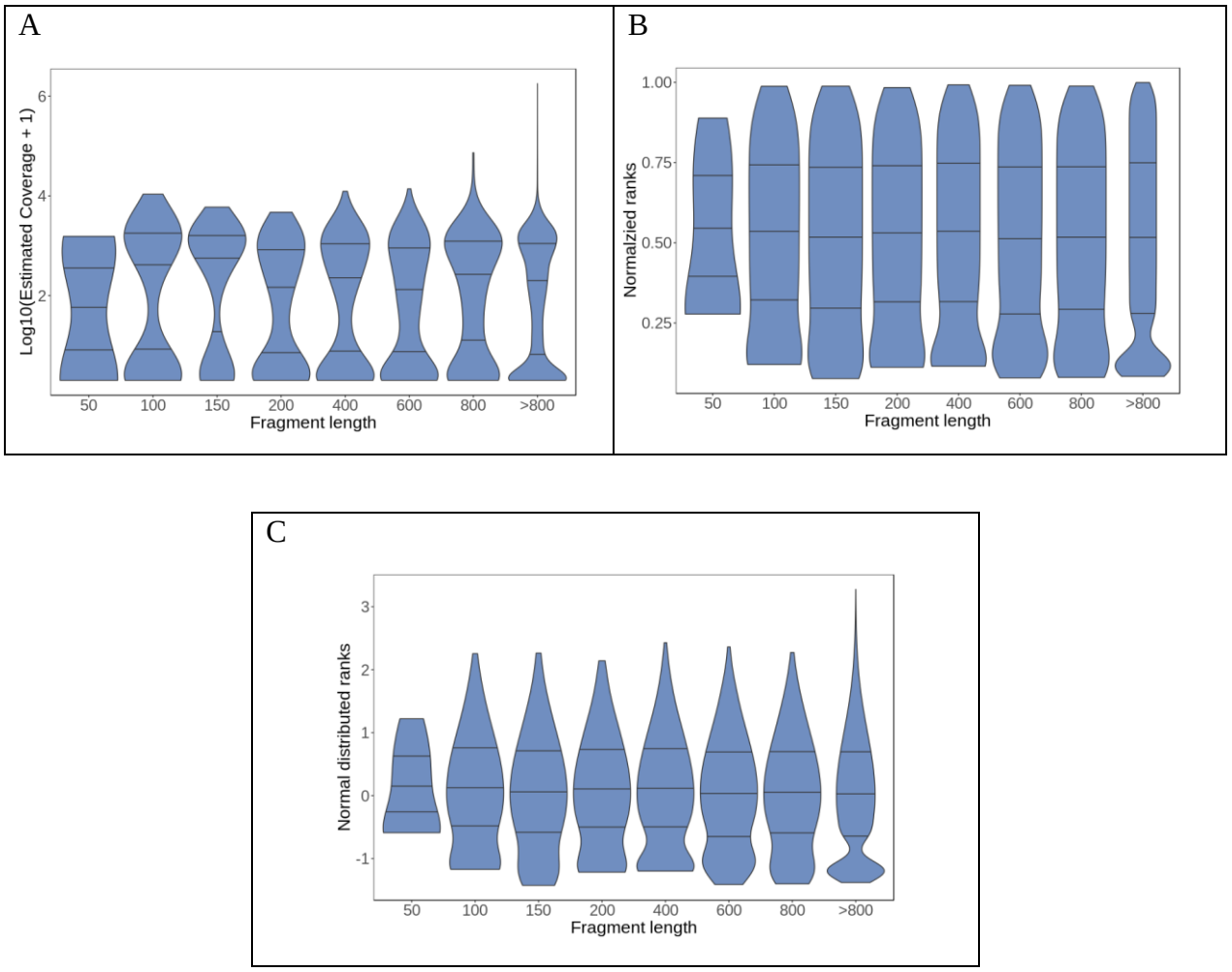

Figure 7: Distributions in categories of fragment length of the estimated coverage in sample A_1. (A) Before normalization, (B) in rank normalization and (C) in normal distributed rank normalization.

The third factor that determines the distribution of estimated coverage is the existence of the secondary restriction site (blind and non-blind fragments). In the following graphs we present the results of the normalization of the estimated coverage of fragments based on this factor. Figure 8 shows that before normalization there are a lot of zero counts in both types of fragments which creates a problem regarding coverage distribution, and blind fragments in general have a lower coverage. By transforming the coverage into ranks, the two distributions are homogenized (Fig. 8B). Finally, by transforming the uniformly distributed ranks into normally distributed ranks we achieve a homogeneous result over the classes of blind and non-blind fragments (Fig. 8C).

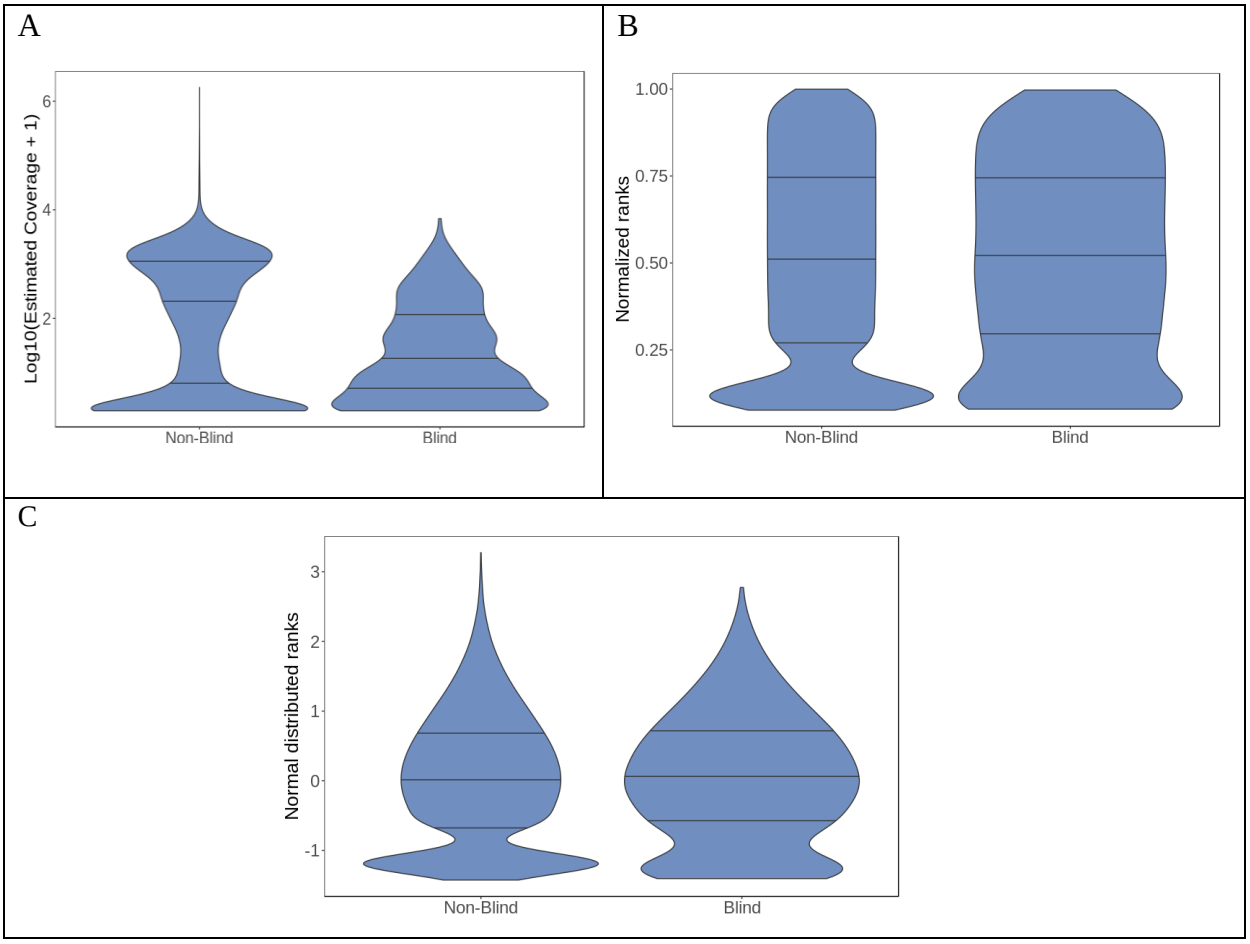

Figure 8. Distribution of estimated coverage in fragments classified by presence/absence of secondary restriction site (non-blind/blind) in sample At_A_1.

From the exemplary results above it is clear that the proposed transformations remove differences in the estimated coverage distribution between the categories.

Following normalization of estimated fragment coverage, a linear mixed model is used as it is described in Methods to identify DCRs. As input to the LMM model we use a table in .csv format with normalized results for all replications under each treatment. A fragment of this table is presented in Table 6. Full table with normalization results of *A. thaliana* is available in Supplementary_file_1.csv in Supplementary Files.

Table 6. Input to the LMM analysis with the normalized results for all samples of each treatment

Chromo-some	Fragment start	Fragment end	At_A_1	At_A_2	At_A_3	At_B_1	At_B_2	At_B_3
chr1	1	7512	0.451	0.219	0.066	0.41	0.102	0.228
chr1	7506	10949	0	0	0	0	0	0
chr1	10943	13175	0	0	0	0	0	0
chr1	13169	20070	0.071	0.776	0.904	0.136	0.898	0.603
chr1	20064	28278	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…

As it is described in methods, a sliding window approach is used to locate DCRs with the LMM model. Different steps and sizes of the sliding windows, in a range from 20000 to 1000000 for the length and 5000 to 50000 for the step, have been tested.

The further analysis uses as input the normalized coverage and the sliding window parameters, and, based on a linear mixed model approach, compares the 2 experimental variants (treatments) and returns a table with the following statistics: estimated means in each variant, standard error of the means, *P* value for testing the difference between mean values, variance component of random effects for fragments, and FDR for each window. Windows characterized by small FDR values (say, < 0.05) can be assumed to be in contact with the bait differently between the two experimental conditions, as it is described in Methods, and these windows can be considered as differentially contacting windows **"**DCW". A fragment of the resulting table can be seen in the Table 7 (Full table with LMM results of *A. thaliana* for sliding window 200000 and step 50000 is available in Supplementary_file_2.csv in Supplementary Files.

Table 7. LMM model output

Chromo-some	Window start	Window end	Estimated mean coverage of window in condition		Std. error of the mean	P-value for difference between conditions	Variance component for fragments	Corrected P-value
			A	B				
chr1	0	199999	-0.028	-0.012	0.0494	0.747	0.00015	0.961
chr1	100000	299999	-0.022	-0.032	0.0358	0.798	0.02352	0.975
chr1	1000000	1199999	0.036	-0.031	0.0417	0.102	0	0.569
chr1	10000000	10199999	0.005	-0.023	0.0349	0.404	0.01272	0.803
...	...	...	...	...	...	...	...	...

The data in this table is used to produce genomic regions (DCR) by merging overlapping windows of restriction fragments. Before merging we are using the estimated mean coverage for each condition to calculate if the significant contact in each window within a merged region is dominating in variant A or variant B. If this is not the case, merging contacts from different variants is not proper. We need the merged sliding windows to include significant regions dominating in the same condition, which is happening without problem in our case after merging. The DCRs are described in bed format for visualisation in genome browser or for further exploration (annotation) and downstream analysis.

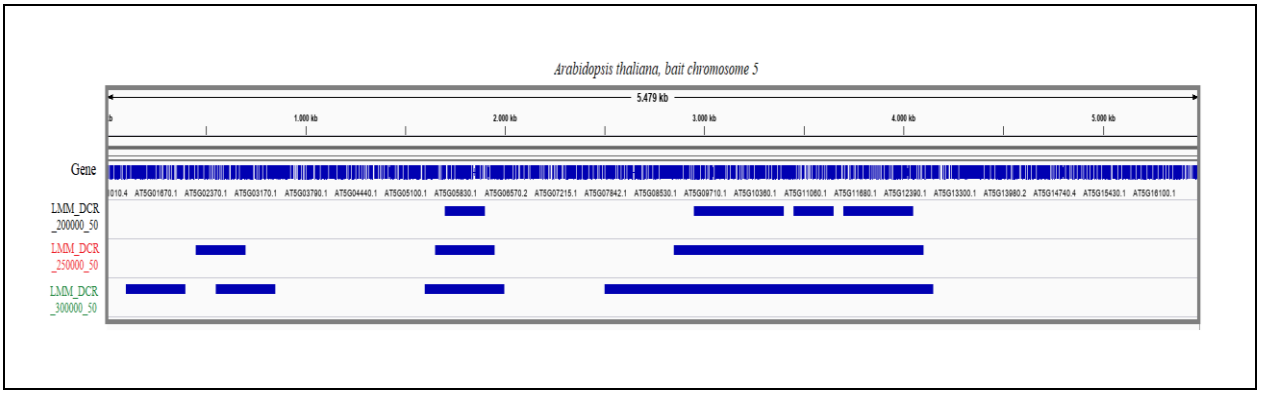

Figure 9. Visual representation of differentially contacting regions (DCR) after LMM analysis for three different sizes of sliding window in bait chromosome 5.

In order to test and validate this approach several sizes of sliding window and step have been analysed by *4CseqR* and the results are shown in Table 8. Also from Table 8 we can extract information about the origin of the dominant coverage in each region which is coming either from condition A or B.

Table 8. Results of LMM analysis for different versions of sliding window and step.

Version	Window size	Step	Number of			
			differentially contacting windows (DCW)	differentially contacting regions (DCR)	DCR – dominant variant A	DCR – dominant variant B
1	10000000	50000	171	1	1	0
2	1000000	50000	100	3	3	0
3	500000	50000	57	6	6	0
4	400000	50000	48	5	5	0
5	350000	100000	17	5	5	0
6	350000	50000	40	5	5	0
7	300000	50000	33	7	7	0
8	250000	50000	27	5	5	0
9	250000	100000	13	3	3	0
10	200000	50000	19	9	7	2
11	200000	100000	10	7	6	1
12	150000	50000	10	4	4	0
13	150000	100000	7	4	4	0
14	100000	50000	4	1	1	0
15	100000	20000	7	1	1	0
16	100000	25000	7	1	1	0
17	80000	20000	5	1	1	0
18	80000	40000	3	1	1	0
19	60000	30000	4	1	1	0
20	50000	25000	3	1	1	0
21	40000	20000	2	1	1	0
22	20000	10000	0	0	0	0
23	20000	5000	0	0	0	0

From Table 8 we can see that the size of the sliding window plays an important role in the number of DCRs that the LMM can locate. Very short (< 40000) and very long (>1000000)

sliding windows do not return a big number of differentially contacting regions. We can conclude that a size of sliding window which is giving the most of DCRs in this case would be between 200000 and 300000 length with step 50000.

The list of DCRs for (200000, 50000) is presented in Table 9. Each DCR is characterized by parameters averaged over all windows contained in it: mean coverage for condition A and B, average standard error of the means, and average corrected *P* value.

Table 9. Differentially contacting regions obtained using normalization by ranks and LMM with sliding window parameters (200000, 50000)

Chromosome	Start	End	Coverage in condition A	Coverage in condition B	Standard error	Corrected *P* value	Variant with dominant contact
chr2	3300000	3499999	0.0905	-0.1499	0.06006	0.01851	A
chr3	250000	499999	0.0650	-0.0714	0.03525	0.02213	A
chr4	5350000	5549999	-0.0826	0.0854	0.04668	0.04783	B
chr4	9400000	9649999	-0.0271	0.1238	0.04153	0.04103	B
chr5	1700000	1899999	0.0968	-0.0923	0.04735	0.01851	A
chr5	2950000	3399999	0.4011	-0.1287	0.08813	0.00101	A
chr5	3450000	3649999	0.0550	-0.1452	0.05139	0.02164	A
chr5	3700000	4049999	0.1001	-0.1017	0.05122	0.02008	A
chr5	23900000	24099999	0.0707	-0.0896	0.04383	0.03969	A

4.1.5 Binarization and Fisher test

The binarization is performed on the estimated coverage obtained from *Salmon*. As it is described in Methods, the binary transformation is done with respect to a threshold (*R*). The proposed transformation takes the estimated coverage of fragments and converts it into 0 or 1 based on the threshold $R = 1$.

Following binarization of estimated fragment coverage, a Fisher exact test is used as it is described in Methods to identify DCWs and DCRs. As input to the Fisher exact test we use a table in .csv format with binary results for all replications under each treatment. A fragment of this table is presented as Table 10. Full table with binary results of *A. thaliana* is available in in Supplementary_file_3.csv in Supplementary Files.

Table 10. Input to the Fisher test for all samples of each treatment

Chromo-some	Fragment start	Fragment end	At_A_1	At_A_2	At_A_3	At_B_1	At_B_2	At_B_3
chr1	1	7512	0	1	0	0	0	0
chr1	7506	10949	0	0	0	0	0	1
chr1	10943	13175	0	0	0	1	0	0
chr1	13169	20070	0	1	1	0	0	0
chr1	20064	28278	0	1	1	0	0	0
…	…	…	…	…	…	…	…	…

Following the Methods, a sliding window approach is used to locate DCRs with the Fisher exact est. Different window sizes and steps have been tested in a range from 20000 to 1000000 for the size of the window and 5000 to 100000 for the step, in different combinations presented in Table 11.

Fisher exact test compares the 2 conditions and returns a table with the following statistics for each window: corrected Fisher *P* value, fragment coverage probability under conditions A and B, and information about the total number of fragments contained in the window and the number of covered fragments under each condition. If Fisher *P* value is smaller than a threshold (< 0.05) it means that the window is in contact with the bait differently (statistical significant difference) under the two experimental conditions (is a DCW) in the sense that the probability of a fragment contacting the bait is different. A fragment of the resulting table can be seen in Table 11. (Full table with Fisher test results of *A. thaliana* for sliding window 20000 and step 10000 is available in Supplementary_file_4.csv in Supplementary Files.

Table 11. Fisher exact test output

Chromo-some	**Window start**	**Window end**	**Fisher corrected p-value**	**Probability of window coverage under condition**		**Number of fragments in the window**	**Number of covered fragments under condition A and B in all replications**	
				At_A	**At_B**		**At_A**	**At_B**
chr1	0	19999	0.316	0.333	0.083	4	4	1
chr1	10000	29999	0.264	0.388	0.166	6	7	3
chr1	100000	119999	0.398	0.233	0.366	10	7	11
chr1	1000000	1019999	1	0.291	0.291	8	7	7
...	...	...	...	...	...	...	...	...

The data in this table are used to produce genomic regions (DCR) by merging overlapping windows of restriction fragments. These are described in bed format for visualisation in genome browser or for further exploration (annotation) and downstream analysis. Exemplary visualisation of results for different parameters used in the sliding window procedure is shown in Fig. 10.

The analysis of this figure shows that there are differences in different sizes of sliding window. There are different situations connected with the length of the sliding window. By increasing the length of the window the result is bigger overlapping regions (Fisher results for size 200000 and step 50000) but less significant regions. By using a different size of sliding window (Fisher results for size 100000 and step 50000) we can find new significant regions that does not exist in bigger windows, but are smaller after overlapping. Whereas in cases of smaller sizes of sliding window (see Fisher results of size 2000 and step 500) it can be seen that there are more fragments captured per sliding window which leads to more smaller significant regions after overlapping. In general, the Fisher exact test seems to work better in cases with increased numbers of covered fragment and returns smaller but more significant regions.

Results of the procedure for all parameters of sliding windows are shown in Table 12. Similarly to LMM, we can extract information about the origin of the dominant coverage in each region which is coming either from condition A or B.

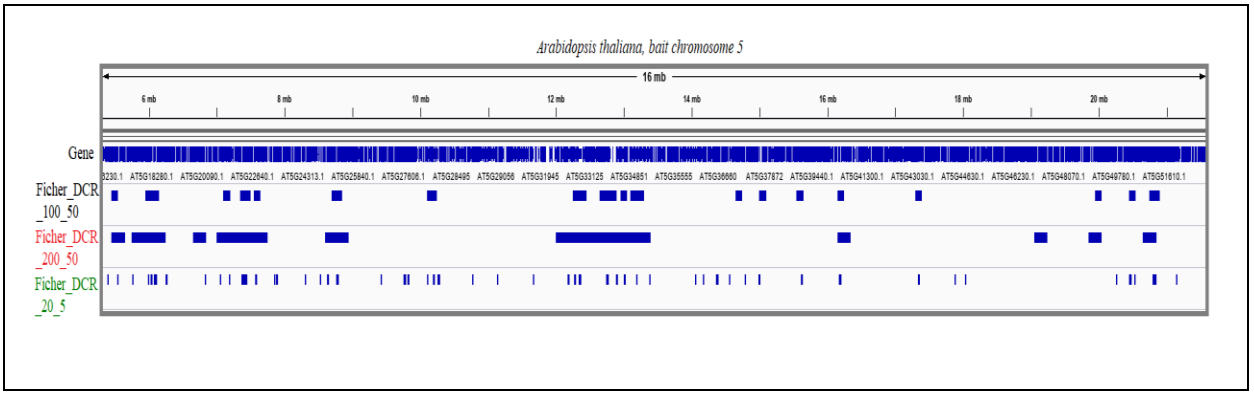

Figure 10. Visual representation of "differentially contacting regions" (DCR) after Fisher exact test for three different sizes and two different steps of sliding windows in bait chromosome 5.

Table 12. Results of Fisher exact test for different versions of sliding window and step.

Version	Window size	Step	Number of			
			differentially contacting windows (DCW)	differentially contacting regions (DCR)	DCR - variant A	DCR - variant B
1	10000000	50000	1250	5	5	0
2	1000000	50000	1045	18	14	4
3	500000	50000	637	31	26	5
4	400000	50000	468	32	25	7
5	350000	100000	204	34	27	7
6	350000	50000	413	36	29	7
7	300000	50000	344	41	36	5
8	250000	50000	261	34	29	5
9	250000	100000	131	31	27	4
10	200000	50000	184	40	35	5
11	200000	100000	97	37	32	5
12	150000	50000	123	41	36	5
13	150000	100000	62	36	32	4
14	100000	50000	61	34	30	4
15	100000	20000	172	38	34	4
16	100000	25000	119	39	35	4
17	80000	20000	82	26	23	3
18	80000	40000	43	22	19	3
19	60000	30000	25	14	13	1
20	50000	25000	14	11	9	2
21	40000	20000	8	5	5	0
22	20000	10000	2	2	2	0
23	20000	5000	4	2	2	0

Results in Table 12 suggest that by using longer windows we obtain more DCWs, however, the number of DCRs is the largest for medium window sizes. In Table 13 we present DCRs obtained for sliding window parameters (200000, 50000). Each DCR is characterized by parameters averaged over all windows that belong to it: average probability of fragment coverage in variant A and B, and average corrected *P* value. The latter parameter takes, for some DCRs, a value bigger than the threshold 0.05. This is due to the presence, within the DCR, of windows which do not express significant difference in contacts between variants; however, the neighbouring overlapping DCWs, with corrected *P* value smaller than the threshold, make a continuous DCR.

Table 13. Differentially contacting regions obtained using binarization and Fisher test with sliding window parameters (200000, 50000)

Chromo-some	Start	End	Probability of coverage in condition A	Probability of coverage in condition B	Corrected *P* value	Variant with dominant contact
chr1	4150000	4349999	0.34	0.20	0.043	A
chr1	8200000	8399999	0.29	0.16	0.012	A
chr1	12200000	12399999	0.31	0.16	0.030	A
chr1	13250000	13949999	0.41	0.24	0.004	A
chr1	14000000	14749999	0.57	0.34	0.010	A
chr1	15050000	15799999	0.63	0.44	0.023	A
chr1	16000000	16199999	0.38	0.25	0.039	A
chr1	16300000	16549999	0.37	0.23	0.035	A
chr1	21500000	21699999	0.40	0.23	0.009	A
chr1	25800000	25999999	0.27	0.15	0.045	A
chr2	1150000	1549999	0.33	0.19	0.048	A
chr2	1950000	2299999	0.33	0.17	0.011	A
chr2	2600000	2999999	0.34	0.22	0.031	A
chr2	3050000	3649999	0.56	0.33	0.003	A
chr2	3700000	4399999	0.44	0.28	0.020	A
chr2	4450000	5649999	0.40	0.26	0.018	A
chr2	6800000	7099999	0.27	0.15	0.066	A
chr2	10500000	10699999	0.28	0.16	0.037	A
chr2	12250000	12449999	0.31	0.16	0.026	A
chr3	9950000	10299999	0.32	0.18	0.021	A
chr3	11300000	11599999	0.35	0.18	0.011	A
chr3	11650000	12899999	0.37	0.20	0.007	A
chr3	13100000	13599999	0.38	0.22	0.012	A
chr3	14300000	14549999	0.44	0.29	0.026	A
chr3	15100000	15349999	0.40	0.26	0.031	A
chr4	2050000	2349999	0.33	0.19	0.017	A
chr4	2800000	2999999	0.48	0.26	0.006	A
chr4	3050000	3499999	0.54	0.38	0.013	A
chr4	4100000	5149999	0.43	0.27	0.007	A
chr4	5250000	5449999	0.33	0.19	0.009	A
chr4	14050000	14249999	0.39	0.22	0.019	A
chr5	1500000	1699999	0.46	0.60	0.048	B
chr5	2500000	2699999	0.47	0.60	0.048	B
chr5	3100000	3299999	0.92	0.80	0.008	A
chr5	5900000	6199999	0.34	0.50	0.007	B
chr5	7250000	7549999	0.33	0.49	0.028	B
chr5	8600000	8899999	0.31	0.45	0.023	B
chr5	12200000	12549999	0.52	0.36	0.018	A
chr5	12550000	12949999	0.42	0.24	0.012	A
chr5	13000000	13299999	0.45	0.33	0.077	A

4.1.6 Comparison of results from LMM and Fisher test

In this section we explore results of LMM and Fisher test analyses of *Arabidopsis* data with the aim of inferring the properties of estimated parameters and test results.

First, by analysing tables of results of LMM model and Fisher test, we can conclude that in both methods the length of the sliding window has an influence on the result. Different sizes of sliding window seems to play a more important role in Fisher test than in LMM model where the numbers of DCRs are limited. For both methods the medium parameters work the best. In both cases very big sizes of sliding windows (> 1000000) return wide regions in all chromosomes which would not lead to any interpretation.

Secondly, we compare the results of LMM analysis and Fisher test analysis obtained for the same sliding windows parameters. Fig. 11 presents a visual representation of a genomic region with DCRs from LMM model and Fisher exact test for the same size (200000) and step (50000) of a sliding window. There are overlapping DCRs obtained by the two methods, but also DCRs that do not overlap.

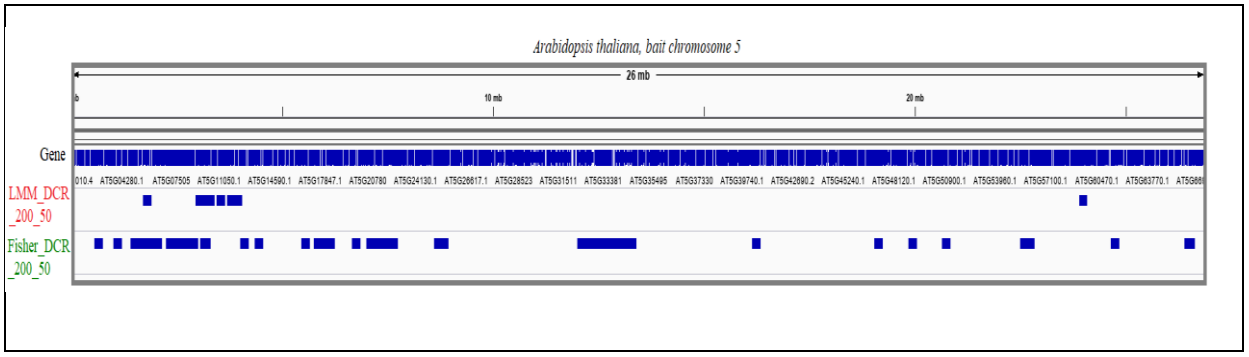

Figure 11. Visual representation of differential contacting regions (DCR) for the same window size (200000) and step (50000), obtained with LMM model and Fisher exact test, in bait chromosome 5.

Thirdly, we explore the characteristics of sliding windows obtained by two methods. The scatterplot of estimated, normalized mean coverage values for two experimental variants indicates that there is no correlation between the coverage in variant A and in variant B (Fig. 12). For windows identified as DCWs, the estimated coverage is positive in one of the variants and negative in the other, with the difference between the two mean values especially big for DCWs of type "A". The absolute differences in mean values between two variants for DCWs are from 0.14 to 0.67. On the other hand, the scatterplot of estimated fragment coverage probabilities indicates that the probabilities for variant A and variant B are correlated. Even in DCWs, the estimated probabilities for two variants are quite close, with the absolute differences being from 0.11 to 0.29. The volcano plots (Fig. 13) visualise the fact that majority of DCWs show dominant contacts under condition A.

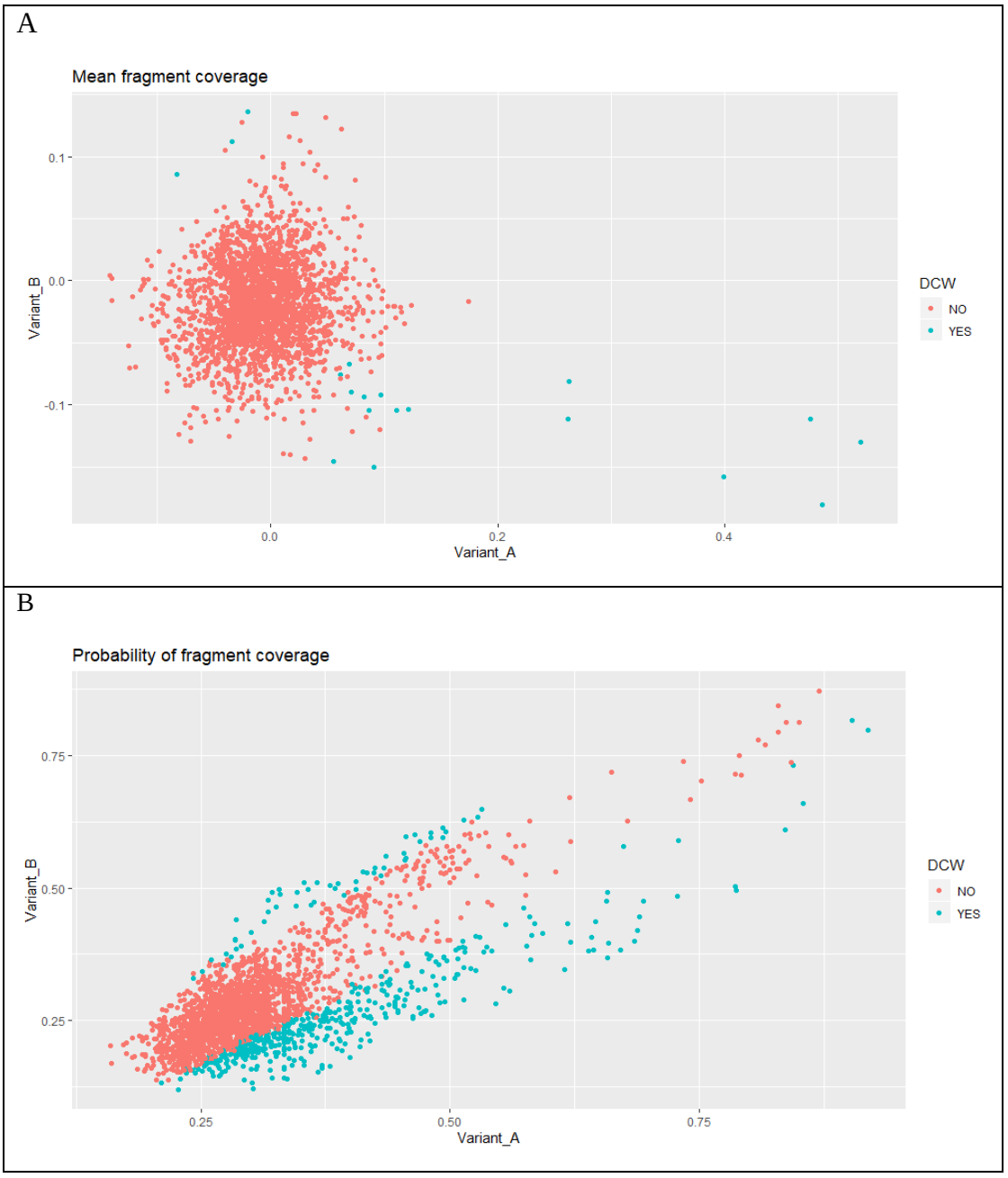

Figure 12. Comparison of parameter estimates in LMM analysis and in analysis by Fisher test. A. Scatterplot of estimated mean fragment coverage in sliding windows for experimental variants A and B. B. Scatterplot of estimated fragment coverage probabilities for variants A and B. Blue dots indicate sliding windows declared as DCW. Results for window length 200000, step 50000 (position 10 in Tables 8 and 12)

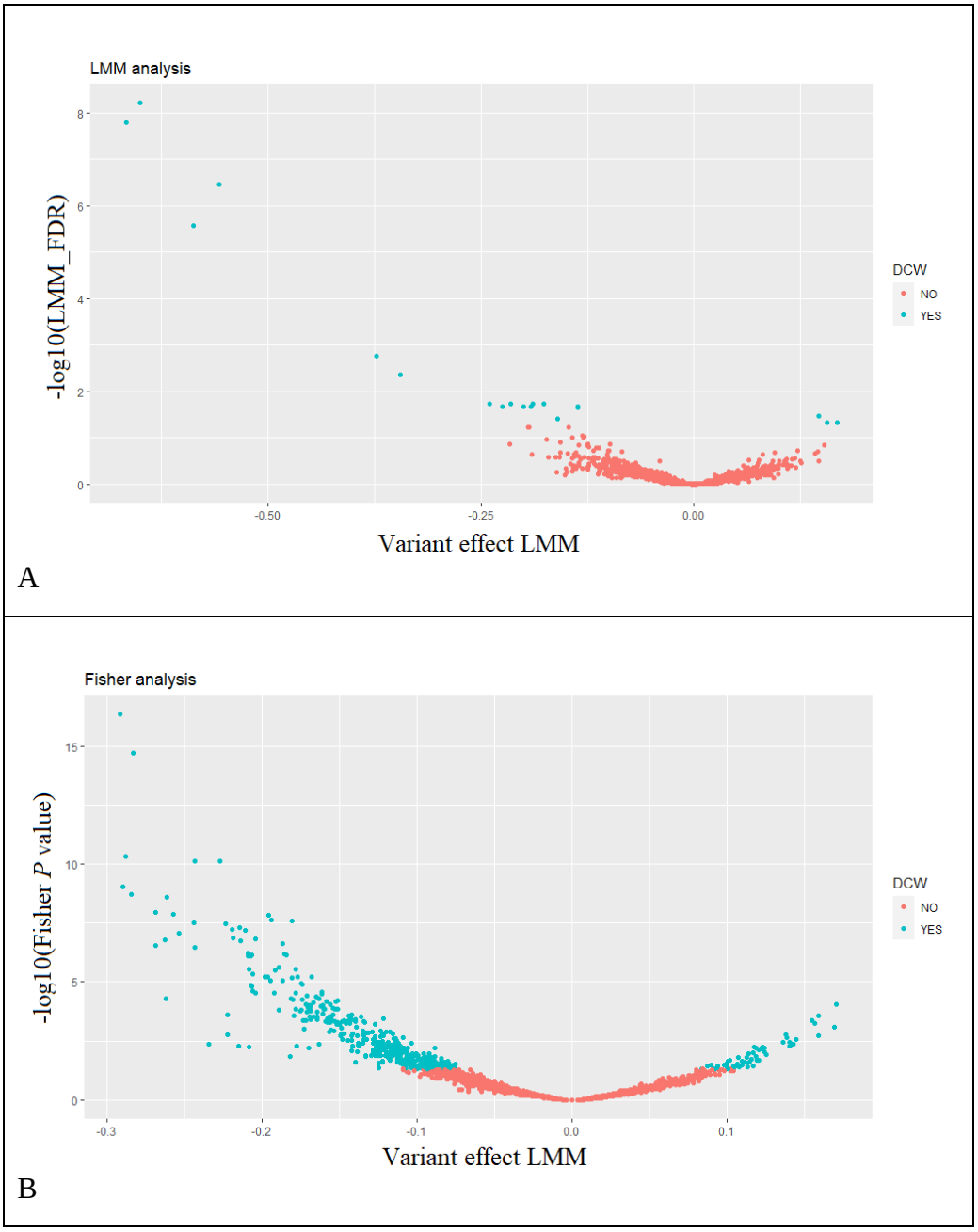

Figure 13. Volcano plots for results of analysis by LMM analysis and by Fisher test. A. Significance scores v. variant effects (mean B - mean A) for LMM analysis. B. Significance scores v. variant effects (probability in B - probability in A) for Fisher test. Blue dots indicate sliding windows declared as DCW. Results for windows length 200000, step 50000 (position 10 in Tables 8 and 12)

Comparison of results obtained by two methods indicates that there is no relationship between the variant effects and significance scores obtained from two methods (Fig. 14). There were two DCWs declared by both methods, both with dominant contact in variant "A".

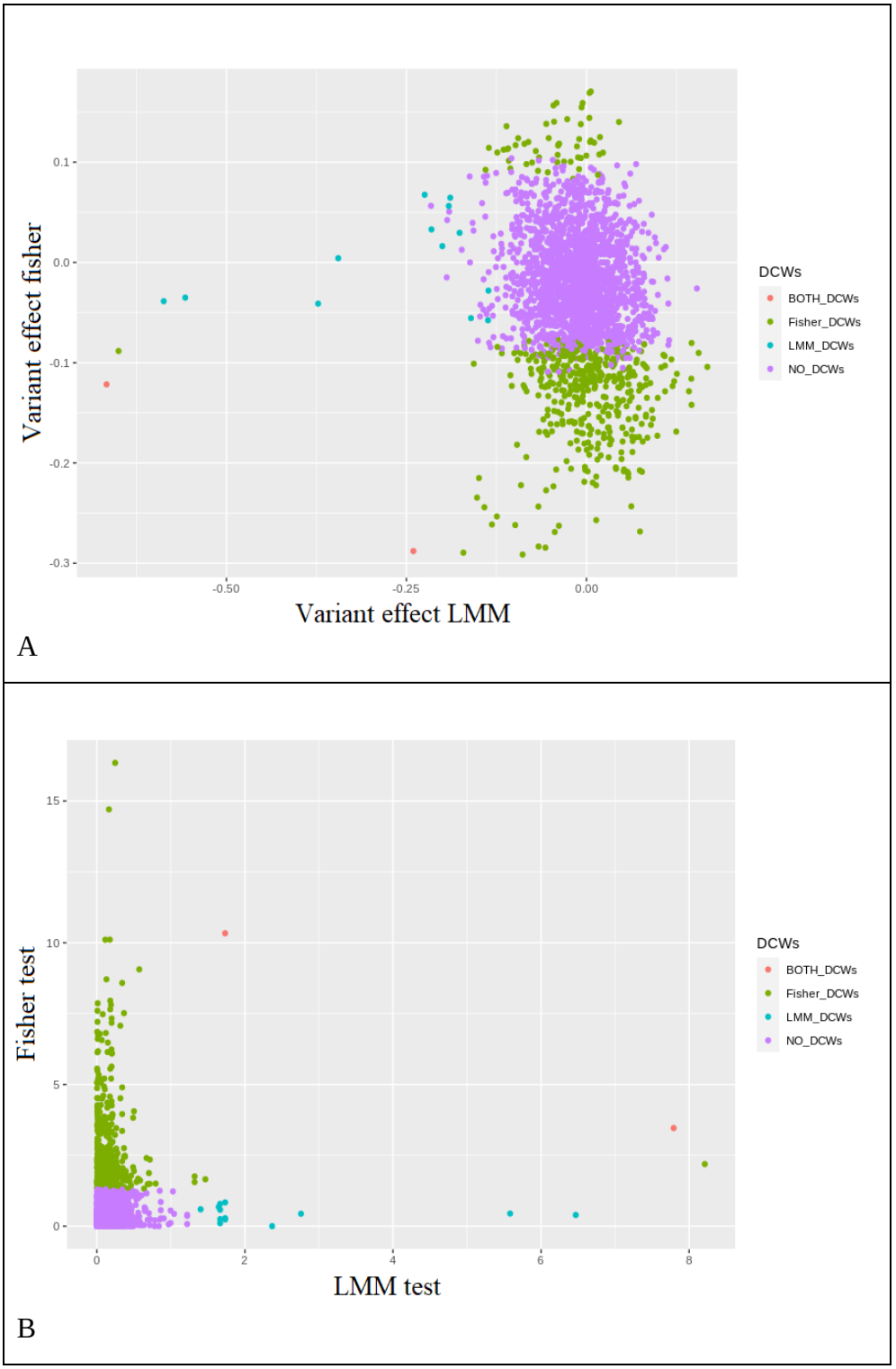

Figure 14. Comparison of results obtained from LMM analysis v. Fisher test analysis. A. Scatterplot of variant effects (B-A; differences in mean values for LMM, differences in probabilities for Fisher test). B. Scatterplot of significance scores -log10 (corrected *P* value). Colored dots indicate DCWs: green - in Fisher test, blue - in LMM analysis, red - in both analyses. Results for window length 200000, step 50000 (position 10 in Tables 8 and 12)

As it is mentioned above, although the two methods use different features to locate the DCRs, there are cases of common DCR's in which there are common DCW's. Table 13 presents a case like that with a common DCW between the two methods. Both the coverage levels and probabilities of coverage are bigger in condition A.

Table 14. DCWs common for two methods

Chromo-some	Start	End	LMM analysis				Fisher test			
			Cove-rage A	Cove-rage B	Diffe-rence	Corr. p-value	Proba-bility A	Proba-bility B	Diffe-rence	Corr. p-value
chr2	3300000	3499999	0.0905	-0.1499	-0.2404	0.0185	0.6705	0.3826	-0.2879	0.003
chr5	3100000	3299999	0.4864	-0.1802	-0.6666	1.615e-08	0.9189	0.7973	-0.1216	0.008

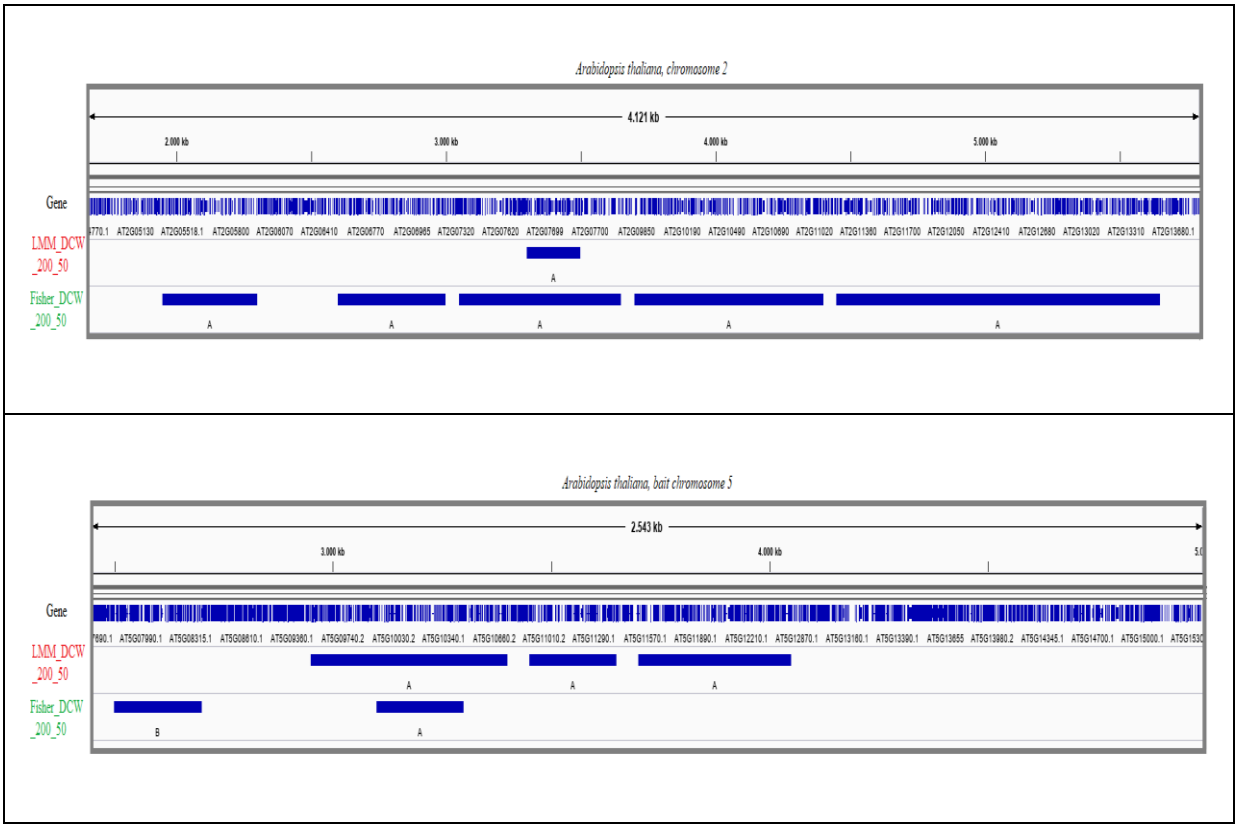

Figure 15. Common differentially contacting windows (DCWs) for the same window size (200000) and step (50000), from LMM model and Fisher exact test, in chromosomes 2 and 5.

Finally, we compare the DCRs found by two methods by finding their intersection. There were 9 DCRs obtained using LMM and 40 obtained by Fisher test, which resulted in 3 common (intersecting) regions shown in Table 15. It can be seen that there is one DCR region in chromosome 4 in which there is a disagreement between the two statistical models, as in LMM the bigger coverage level is under condition B whereas in Fisher test coverage probability is bigger under condition A.

Table 15. Intersecting DCR from LMM and Fisher test for sliding windows (200000, 50000).

LMM						Fisher test				
Chromosome	**Start**	**End**	**Coverage A**	**Coverage B**	**Corrected p value**	**Start**	**End**	**Probability A**	**Probability B**	**Corrected p value**
chr2	3300000	3499999	0.0905	-0.1499	0.01851	3050000	3649999	0.56	0.33	0.003
chr4	5350000	5549999	-0.0826	0.0854	0.04783	5250000	5449999	0.33	0.19	0.009
chr5	2950000	3399999	0.4011	-0.1287	0.00101	3100000	3299999	0.92	0.80	0.008

4.1.7 Comparison to results of *fourSig*

Results from fourSig

The analysis with *fourSig* tool, aimed at finding significant contacts with the bait, is done independently for each replication in each variant. A fourSig_*.bed files produced with the significant contacts for each sample (At_A_1, At_A_2 At_A_3, At_B_1, At_B_2, At_B_3) are available in Supplementary Files. The next Table 16 presents an example of the results in bed format and a visualization in a genome browser (IGV) is presented in Fig. 16.

Table 16. Typical output of *fourSig* algorithm

Chromosome	Start	End	Category of contact
chr5	314938	343218	3
chr5	462475	469301	3
chr5	626919	632859	3
chr5	632859	641001	3
chr5	641001	674612	1
chr5	752136	763833	3
chr5	774061	796535	3
chr5	813088	841505	3

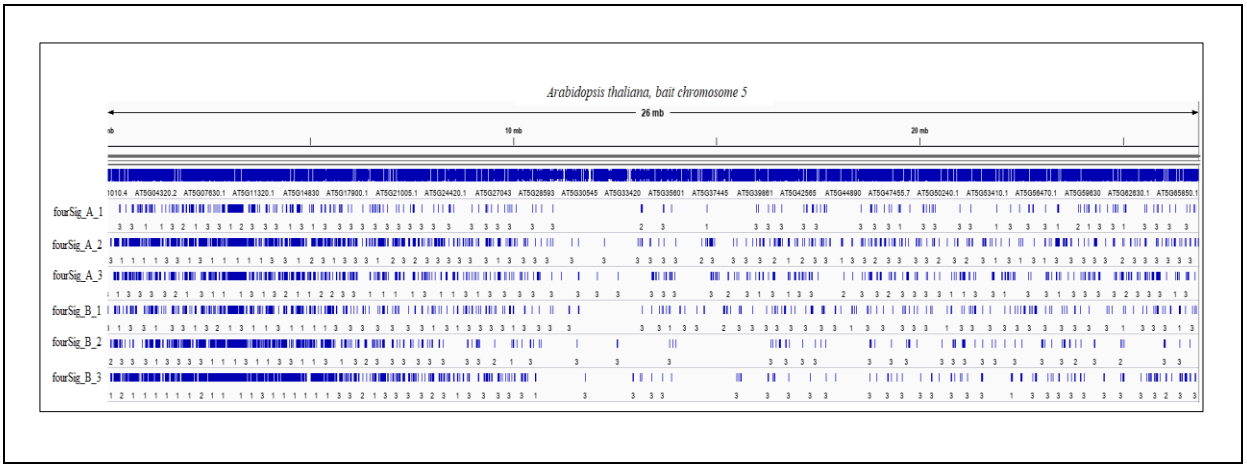

Figure 16. A visual representation of significant contacts in of all samples in bait chromosome 5

Because *fourSig* does not provide any tool for comparative analysis, we performed a comparison of the contacts found in individual replications and experimental variants as follows (Zisis et al. 2020). For each variant separately, we declared as "significant contacts" regions found repeatedly in all three replications. Then, we identified the regions declared as significant for both variants, calling them "common contacts", or for one of the two variants only, calling them "specific contacts" (for variant A or B), in all chromosomes.

In this way we obtained information that in practice could be used for final inference on the outcome of the experiments using *foursig*. Regions significant in variant A and B for all replications are available as *intersect_foursig_varA_Athaliana.bed* and *intersect_foursig_var_ Athaliana B.bed* and significant contacts in both variants A and B are available in bed file with the name *foursig_common_Athaliana.bed* are available in Supplementary Files. A visual representation of common significant contacts is available in Fig. 17 for the bait chromosome 5.

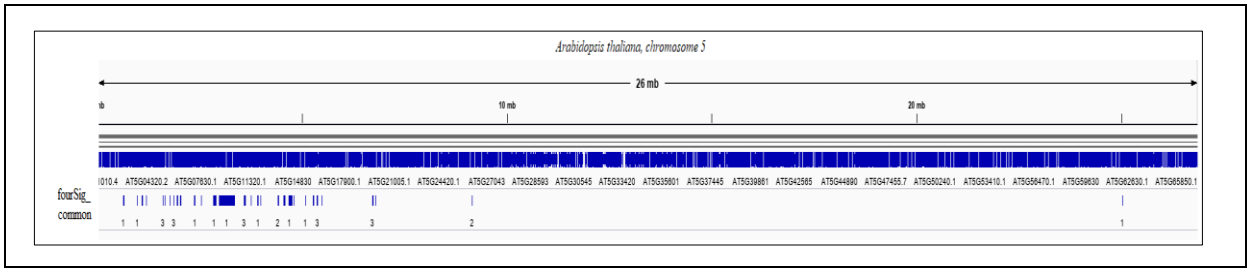

Figure 17. Significant contacts found in the bait chromosome in both experimental variants in the experiment with *A. thaliana*.

Table 17. Results of *foursig* in all replications of each variant and common/specific significant contacts

Groups of significant contacts		**Number of contacts in experimental variant**	
		A	B
Replication	1	1443	1755
	2	1996	1183
	3	1779	1591
Total in variant (common to all replications)		523	496
Specific to variant		416	389
Common		107	

LMM vs foursig

We compare results of LMM analysis and *foursig* by intersecting DCRs of LMM method with regions found as significant contacts in *foursig*. Intersection is done separately for two types of contacts - A and B. From Table 18 we can see that, depending on window parameters in LMM, from 3.4% to 7.1% of *foursig* significant regions in A (out of total 523) are confirmed by LMM analysis (from 5 to 7 DCRs). But very few *foursig* significant regions in B, close to 0% (out of 497), are confirmed by LMM analysis (0-1 DCRs). From Table 19 we can see that from 1.2% to 2.4% of *foursig* varA specific contacts are confirmed by LMM varA-type contacts, and from Table 20 – that from 15,8% to 32.7 *foursig* common contacts are contained

in LMM varA-type contacts. In general, most DCRs identified by LMM intersect with some *foursig* contacts (specific or common).

Table 18. Number of regions with significant contacts from foursig (A 523, B 496) intersecting with DCR from LMM for each condition (varA, varB) and number of DCR from LMM with foursig contacts for each condition (varA, varB)

Window size	Step	Foursig_varA in DCR LMM_varA	Foursig_varB in DCR LMM_varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)
300000	50000	37	0	7 (7)	0 (0)
250000	50000	29	0	5 (5)	0 (0)
200000	50000	23	1	7 (7)	1 (2)
200000	100000	18	1	6 (6)	1 (1)

Table 19. Number of variant-specific contacts from *foursiq* in each condition (A 416, B 389) intersecting with DCR from LMM for each condition (A and B)

Window size	Step	Foursig_specific varA in DCR LMM_varA	Foursig_specific varB in DCR LMM_varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)
300000	50000	10	0	6 (7)	0 (0)
250000	50000	5	0	4 (5)	0 (0)
200000	50000	6	1	3 (7)	1 (2)
200000	100000	6	1	3 (6)	1 (1)

Table 20. Number of common contacts from foursig (107) intersecting with DCR from LMM

Window size	Step	foursig_common in DCR LMM_varA	foursig_common in DCR LMM_varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)
300000	50000	35	0	4 (7)	0 (0)
250000	50000	32	0	4 (5)	0 (0)
200000	50000	24	0	5 (7)	0 (2)
200000	100000	17	0	3 (6)	0 (1)

Fisher test vs foursig

Similar to the previous comparison, we compare results of Fisher approach and *foursig* by intersecting DCRs of Fisher method with regions found as significant contacts in *foursig*. Intersection is done separately for two types of contacts - A and B. From Table 21, we can see that, depending on window parameters, from 0% to 12.8% of *foursig* significant regions in A (523) are confirmed by Fisher analysis (from 0 to 28 DCRs) and from 0% to 9.1% of *foursig* significant regions in B (496) are confirmed by Fisher analysis (0-5 DCRs). From Tables 22 and 23 we can see that more *fourSig* contacts specific to A or B than *fourSig* common contacts are confirmed by DCRs.

Table 21. Number of regions with significant contacts from foursig (A 523, B 496) intersecting with DCR from Fisher test for each condition (varA, varB) and number of DCR after Fisher test with foursig contacts for each condition (varA, varB)

Window size	Step	Foursig_varA in DCR Fisher_varA	Foursig_varB in DCR Fisher _varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)
300000	50000	67	45	28 (36)	5 (5)
200000	50000	22	21	18 (35)	5 (5)
100000	50000	6	5	5 (30)	3 (4)
80000	40000	4	3	3 (19)	3 (3)
40000	20000	0	0	0 (5)	0 (0)

Table 22. Number of significant variant-specific contacts from foursiq in each condition (A 416, B 389) compared to DCR after Fisher test for each condition (A and B)

Window size	Step	Foursig_specific varA in DCR Fisher_varA	Foursig_specific varB in DCR Fisher_varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)
300000	50000	64	39	28 (36)	5 (5)
200000	50000	19	18	16 (35)	5 (5)
100000	50000	5	5	4 (30)	3 (4)
80000	40000	1	3	1 (19)	3 (3)
40000	20000	0	0	0 (5)	0 (0)

Table 23. Number of significant contacts from *foursig* (107) in both conditions (common) in DCR after Fisher test

Window_size	Step	foursig_common in DCR Fisher varA	foursig_common in DCR Fisher varB	Number of DCRs of type A with foursig contacts (all DCRs type A)	Number of DCRs of type B with foursig contacts (all DCRs type B)B
300000	50000	3	6	3 (36)	2 (5)
200000	50000	3	3	2 (35)	2 (5)
100000	50000	1	0	1 (30)	0 (4)
80000	40000	3	0	2 (19)	0 (3)
40000	20000	0	0	0 (5)	0 (0)

4.2 Results for Dataset 2 (*Mus musculus*)

4.2.1 Quality Control

The output of *FastQC* program for two samples shows boxplots of the data quality characteristics across all bases of the NGS reads in the experiment with mouse tissues (Fig. 18). It can be seen that all quality scores are in the green part, which means that there are no sequences having poor quality in both samples for variants A (ESC) and B (iPSC).

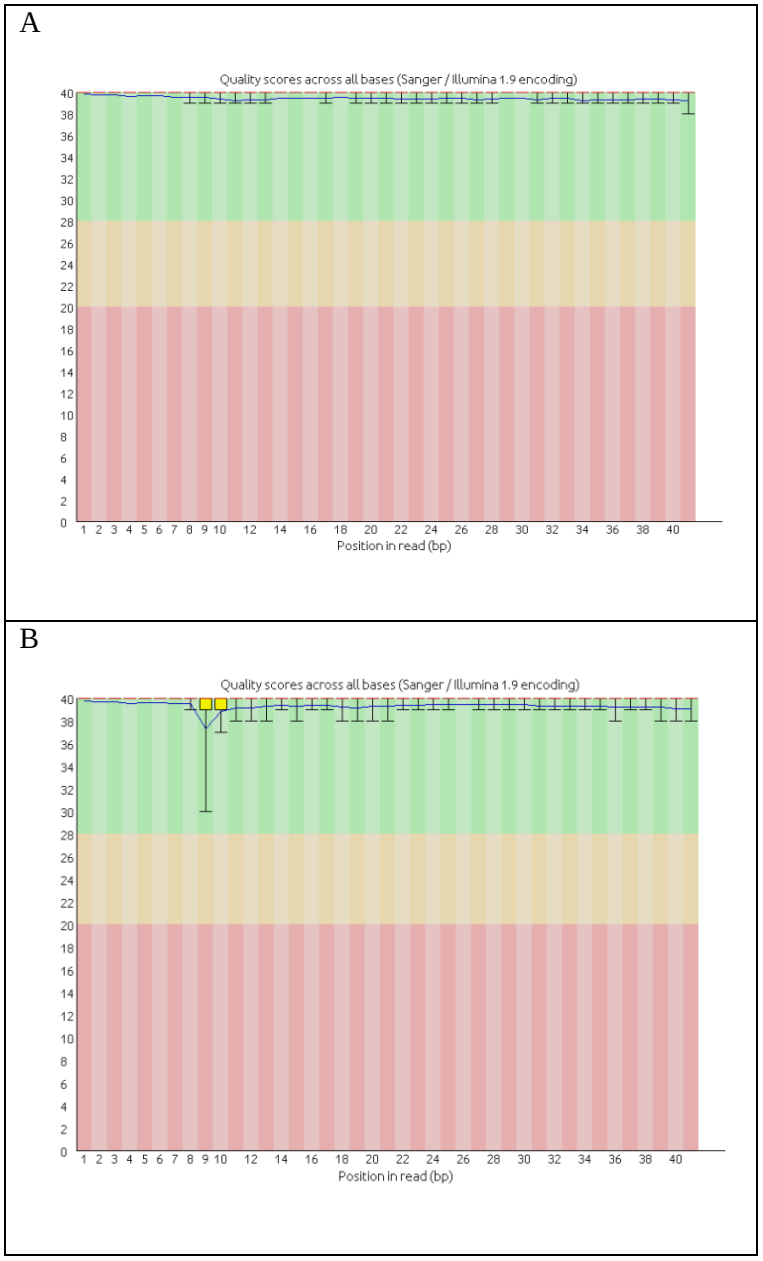

Figure 18. Overview of the quality of NGS reads across bases, at each position, in samples mm_A_1 and mm_B_1 (first replication of each experimental variant).

4.2.2 Preprocessing and mapping

The NGS data obtained for all samples, processed by trimming out viewpoint sequences from the reads and removing reads that didn't contain the primary restriction site AAGCTT (HindIII), were aligned to the library of all AAGCTT restriction fragments from the mm10 genome used as the reference. *Bowtie2* was used with the parameter --score-min L,0,-0.25, which corresponds to the minimum score 0.25×41 = 10.25 for all replications of the first variant (A) and minimum score 0,25×81 = 20.25 for the first 2 replications of the second variant (B) and 0.25×41 =10.25 for the third replication. Table 24 presents the characteristics of data related to the pre-processing and mapping process.

Table 24. Results of NGS data processing

Samples	Data files (ID in the repository)	Number of reads			% reads mapped out of total filtered
		total	after filtering	mapped to library of fragments	
mm_A_1	ESC_Pcdhb19_1.fq	7507857	5144745	4316659	83.90%
mm_A_2	ESC_Pcdhb19_2.fq	2118745	1669878	1339262	80.20%
mm_A_3	ESC_Pcdhb19_3.fq	4653425	1866541	1568671	84.04%
mm_B_1	iPSC_Pcdhb19_1.fq	9609989	8095161	5905514	72.95%
mm_B_2	iPSC_Pcdhb19_2.fq	9130461	8345469	6914012	82.85%
mm_B_3	iPSC_Pcdhb19_3.fq	32457800	18879454	15085709	79.91%

4.2.3 Coverage estimation

The coverage of restriction fragments is computed on the basis of the output of read mapping in sam format and library of AAGCTT restriction fragments in FASTA format. Similar to *Arabidopsis thaliana* data, a fragment of the output of *Salmon* with the estimated fragment coverage is presented in Table 25 and is available in files mm_*_salmon.csv (csv format) and is available in Supplementary Files[2] for all samples. The column NumReads, the estimate of the number of reads, is used for further steps of the analysis in *4CseqR*.

The distribution of the estimated coverage along all chromosomes and bait chromosome 18 (500 bp around the bait) is shown in Fig. 19. The coverage in the bait chromosome is greater than in other chromosomes, especially in the bait region.

[2] Supplementary Files are available on CD attached to the manuscript

Table 25. Salmon's quantification file (quant.sf) for one of the samples

Chromosome	Fragment start	Fragment end	Fragment length	Estimated number of reads from the fragment
chr1	1	3004110	3004109	0
chr1	3004104	3005825	1721	0
chr1	3005819	3008321	2502	0
chr1	3008315	3008899	584	0
chr1	3008893	3009818	925	0
chr1	3009812	3012428	2616	0
chr1	3012422	3015800	3378	1
chr1	3015794	3016189	395	1
chr1	3016183	3024025	7842	0
…	…	…	…	…

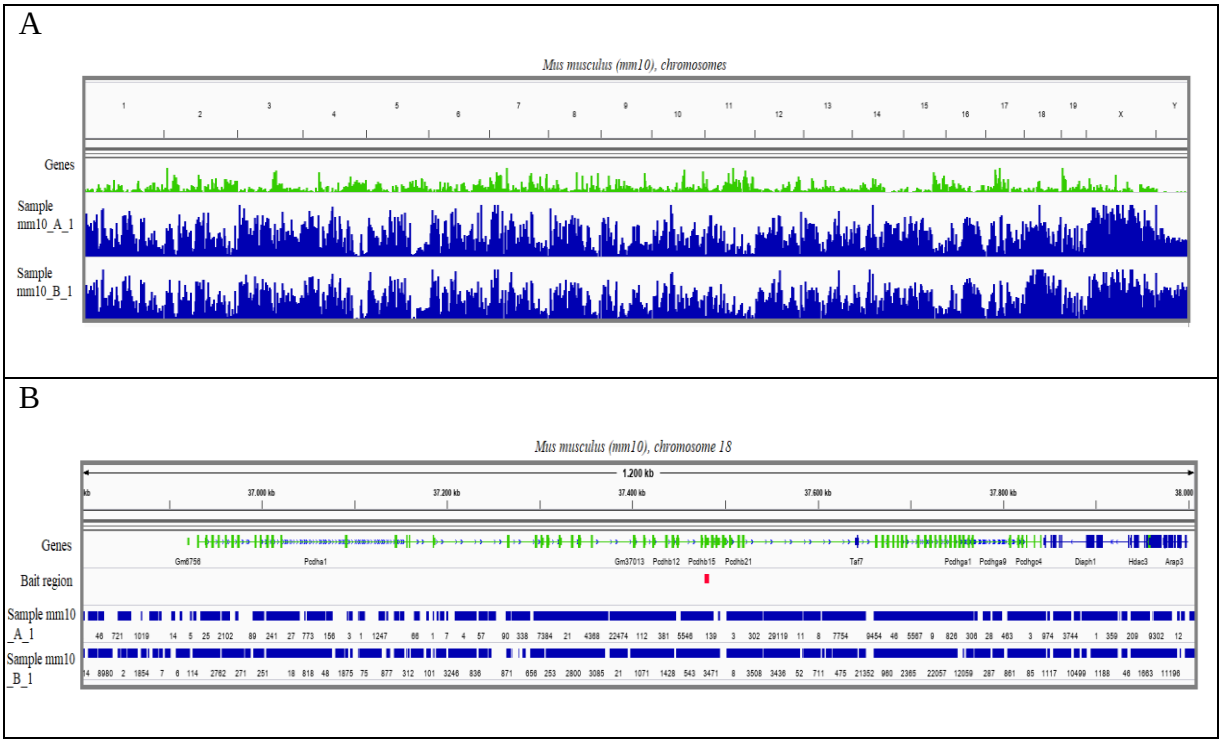

Figure 19: Visualization of the estimated coverage in IGV for mouse samples. A: For all chromosomes, B: for bait chromosome 18.

Distributions of estimated coverage of fragments provided by *Salmon* are presented in Fig. 20. The figure presents data for the first replication of each experimental variant (samples mm_A_1 and mm_B_1). In both experimental variants the coverage was concentrated mainly in the bait chromosome, as can be further seen by calculating metrics for each chromosome (Table 26).

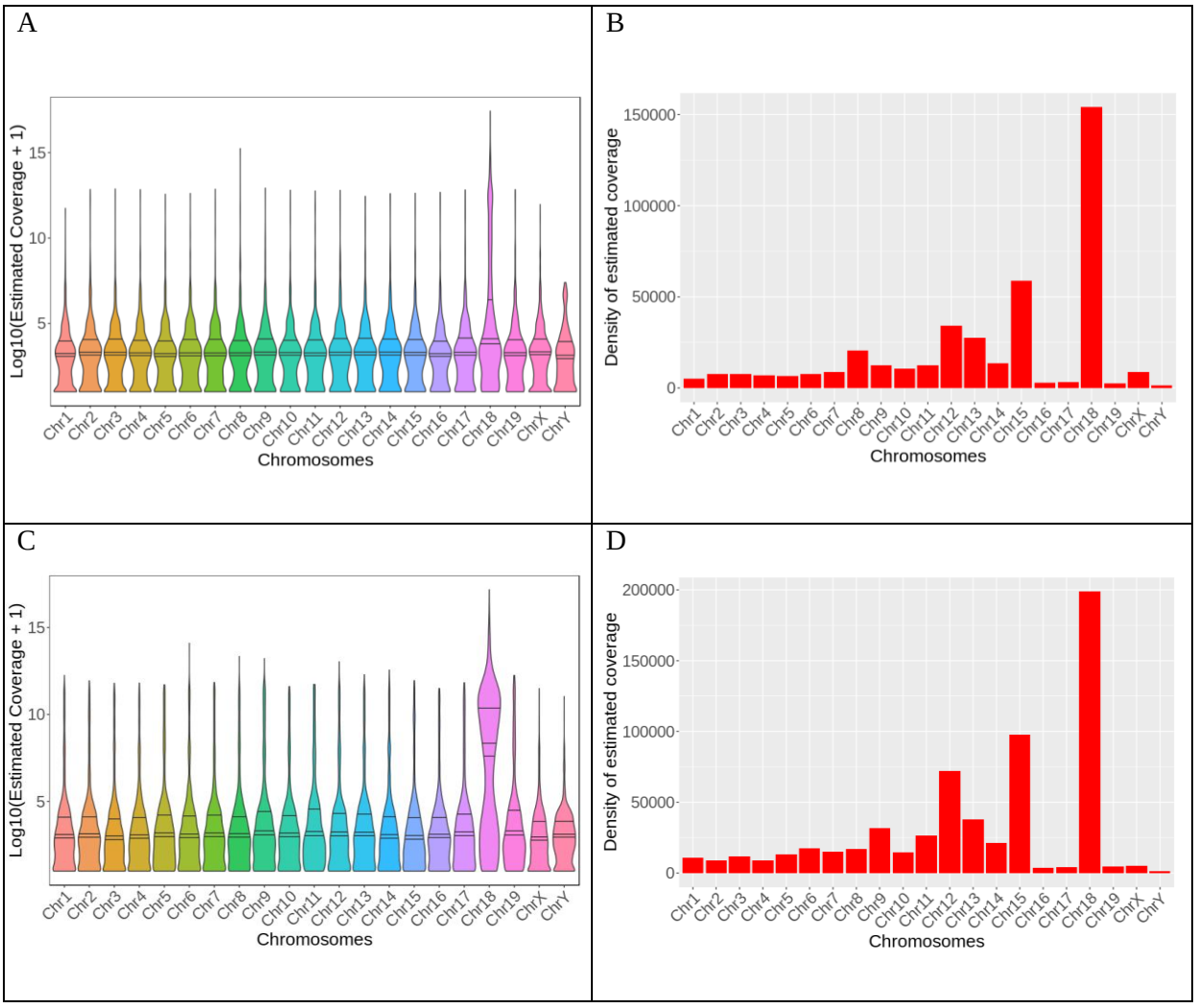

Figure 20. A. Distribution of estimated fragment coverage for sample mm_A_1; B. Density of estimated coverage in chromosomes (total estimated genome coverage for a chromosome divided by the length of chromosome in Mb) in mm_A_1. C. Distribution of estimated fragment coverage for sample mm_B_1; D. Density of estimated coverage in chromosomes in mm_B_1.

Table 26. Characteristics of the restriction fragment coverage by NGS reads in chromosomes of mouse.

Chromosome	Mean coverage	Maximum coverage
chr1	18.20	3427
chr2	33.63	7394
chr3	23.95	7536
chr4	25.22	7334
chr5	24.96	6091
chr6	22.85	6291
chr7	23.22	7471
chr8	55.01	38832
chr9	28.90	7798
chr10	31.76	7160
chr11	31.11	6946
chr12	27.26	7125
chr13	32.52	5588
chr14	29.76	6236
chr15	27.78	6326
chr16	30.93	6565
chr17	37.42	7250
chr18	1105.13	177854
chr19	43.40	7333
chrX	19.82	4014
chrY	14.16	168

4.2.4 Normalization by ranks and linear mixed model

The normalization was performed on the estimated coverage obtained from *Salmon*, in the same way as described in the results of *Arabidopsis thaliana*. The results of normalization with respect to the length of 5' fragment ends are presented in Fig.21.

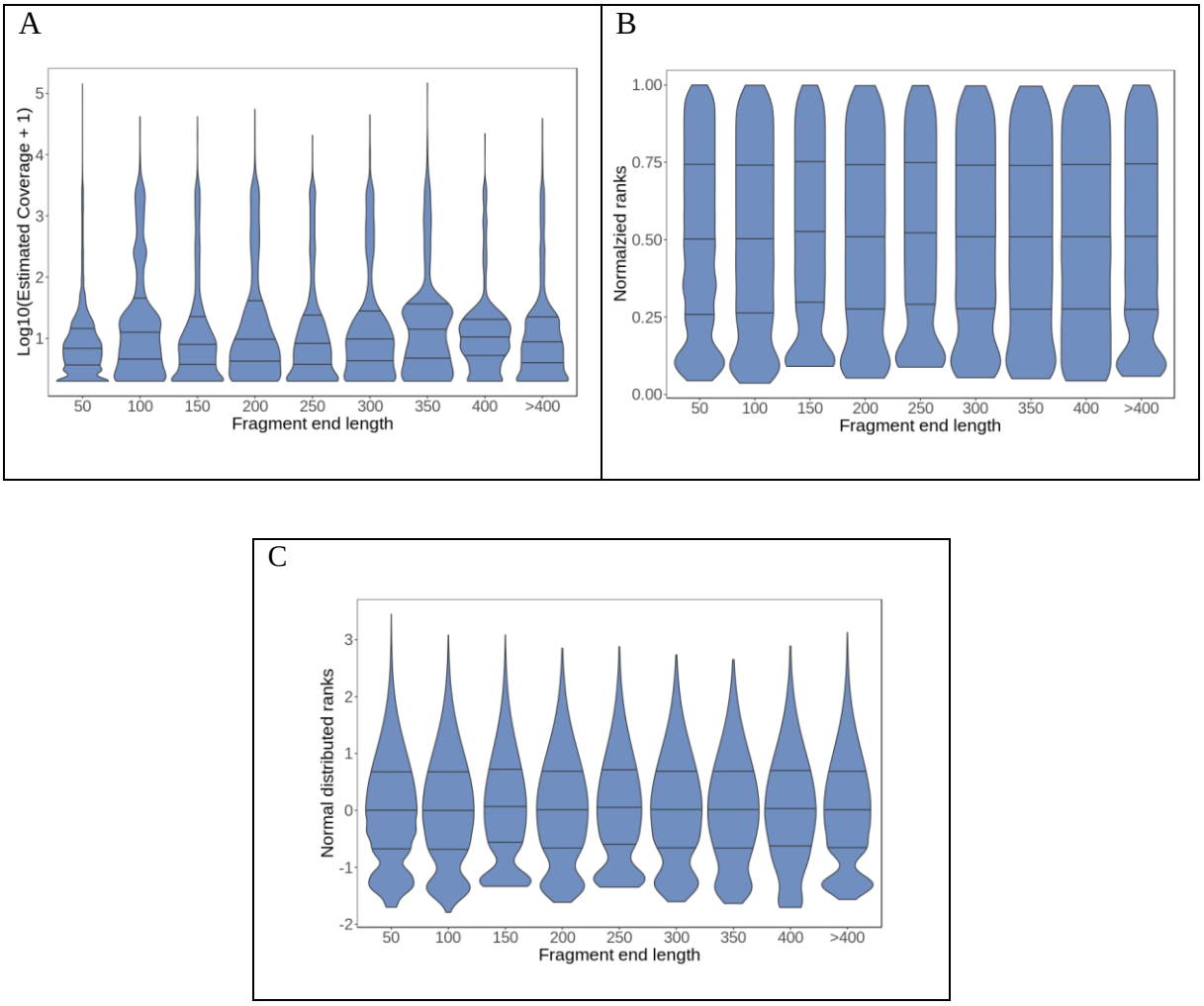

Figure 21. Distributions in categories of 5' fragment end length of the estimated coverage in sample mm_B_1. (A) Before normalization, (B) in rank normalization and (C) in normal distributed rank normalization.

The results of the transformation based on the categories of length of the fragments are shown in Fig. 22. In this case in fragments with very short length of about 500 bp the estimated coverage seems to be smaller than in longer fragments, while the coverage of long fragments (> 4000) decreases with their length (Fig. 22 A, B). After transformation into normally distributed data (Fig. 22 C) the distributions are closer to being homogeneous over different classes.

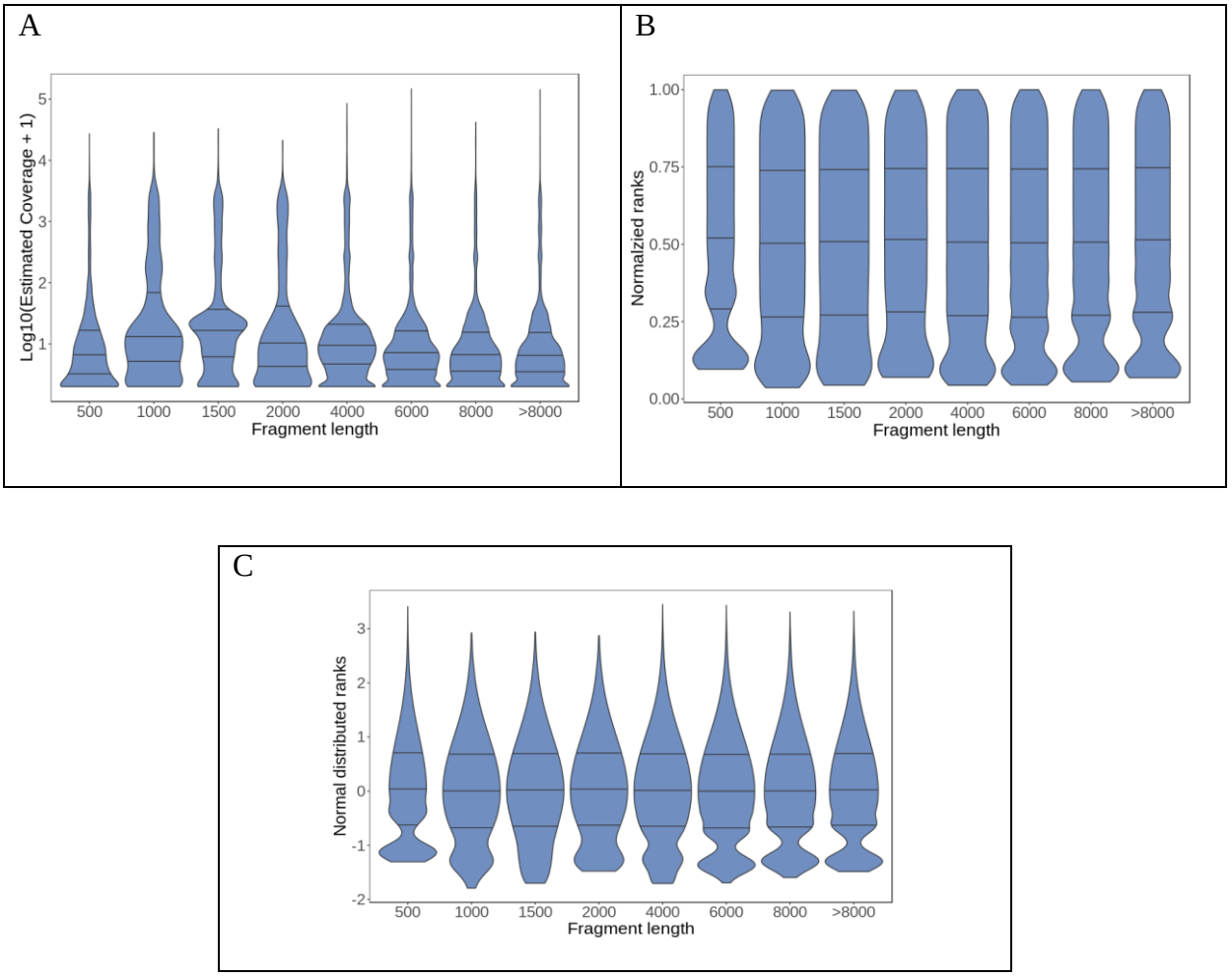

Figure 22. Distributions in categories of fragment length of the estimated coverage in sample mm_B_1. (A) Before normalization, (B) in rank normalization and (C) in normally distributed rank normalization.

The distribution of estimated coverage based on the existence of the secondary restriction site (blind and non-blind fragments) is presented in Fig. 23 A. Blind fragments have a lower coverage. By transforming the coverage into ranks, the two distributions are homogenized (Fig. 23 B) and by transforming into the normally distributed data we achieve a homogeneous result over two classes fragments (Fig. 23 C).

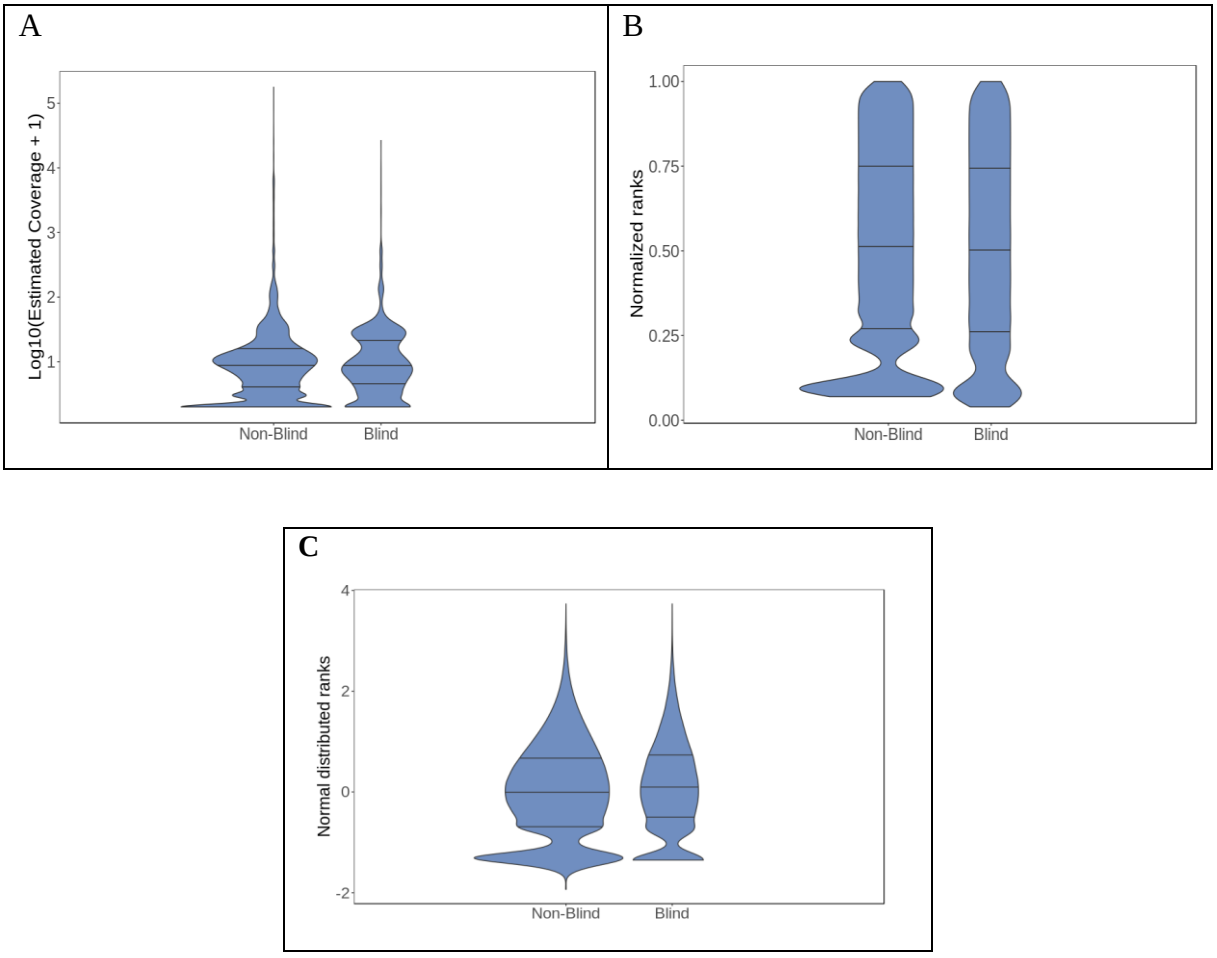

Figure 23. Distribution of estimated coverage in fragments classified by presence/absence of secondary restriction site (non-blind/blind) in sample mm_A_1.

Following normalization of the estimated fragment coverage, a linear mixed model is used as it is described in Methods to identify DCRs. As input to the LMM model we use normalized data for all replications in each variant A and B. A fragment of this table is presented in Table 27. Full table with normalization results of *Mus musculus* 10000 is available in Supplementary_file_5.csv in Supplementary Files. In the sliding window approach, different sizes and steps of the windows in a range from 40000 to 200000 for the length and 20000 to 100000 for the step have been tested.

Table 27. Input to the LMM model with the normalized results for all samples of each variant

Chromosome	Fragment start	Fragment end	mm_A_1	mm_A_1	mm_A_1	mm_A_1	mm_A_1	mm_A_1
chr1	3009812	3012428	0	0	0	0	0	0
chr1	3012422	3015800	0,088	0,158	0,348	0,195	0,067	0,236
chr1	3015794	3016189	0,075	0,175	0,841	0,774	0,389	0,086
chr1	3016183	3024025	0	0	0	0	0	0
chr1	3024019	3025285	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…

Linear mixed model analysis is done in the same way to *Arabidopsis* data and windows characterized by small FDR values (< 0.05) are assumed to be in contact with the bait differently in the two types of cells: ESCs and iPSCs **(**DCW). A fragment of the results can be seen in Table 28 and full table is available in Supplementary_file_6.csv in Supplementary Files. These are used to produce genomic regions by merging overlapping windows of restriction fragments (DCR) (Fig. 24). Table 29 presents results of analysis for different sizes of sliding window and steps.

Table 28. LMM model output for sliding windows (40000, 20000)

Chro-mo-some	Window start	Window end	Estimated mean coverage of window in variant		Std. error of the mean	P-value for difference between variants	Variance component for fragments	Corrected *P* value
			A	B				
chr18	10200000	10239999	-0.024	0.172	0.1221	0.112	0.02739	0.474
chr18	10220000	10259999	-0.028	0.156	0.0959	0.057	0.03192	0.427
chr18	10240000	10279999	0.054	0.109	0.1040	0.597	0.02677	0.822
chr18	10260000	10299999	0.040	-0.013	0.0427	0.207	0	0.544
...	...	...	...	...	...	...	...	...

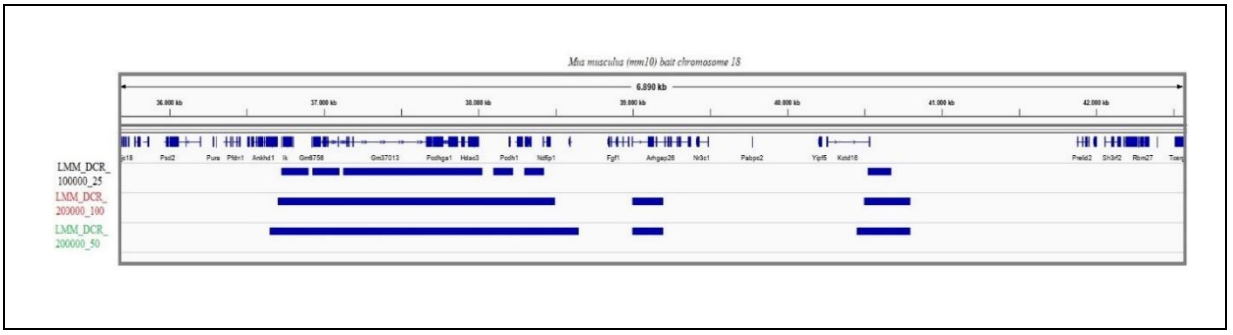

Figure 24: Visual representation of "differentially contacting regions" (DCR) after LMM analysis for three different sizes of sliding window in bait chromosome 18.

Table 29. Results of LMM for different versions of sliding window and step.

Version	Window_size	Step	Number of			
			differentially contacting windows (DCW)	differentially contacting regions (DCR)	DCR – dominant variant A	DCR – dominant variant B
1	200000	100000	61	31	6	25
2	200000	50000	128	37	7	30
3	100000	50000	21	8	6	2
4	100000	25000	49	10	7	3
5	80000	40000	20	5	5	0
6	80000	20000	42	7	7	0
7	40000	20000	6	6	6	0

From Table 29 we can see that the parameters of sliding windows play an important role in the number of DCWs and DCRs that the LMM can find. Very small sliding windows return a small number of DCRs. Sliding windows with size 200000 return more DCRs than windows of other sizes. We can conclude that a size of sliding window which is giving the most of DCRs in this case is 200000, with the step of 50000. The list of 37 DCRs for the sliding

window (200000, 50000) is presented in Table 30. Dominant contacts in chromosomes 2, 15 and 18 are expressed in variant A (ESC cells), while in other chromosomes dominant contacts are in variant B (iPSC cells).

Table 30. Differentially contacting regions obtained using normalization by ranks and LMM with sliding window parameters (200000, 50000) in experiment with mouse cells.

Chromosome	Start	End	Coverage in variant A	Coverage in variant B	Standard error	Corrected *P* value	Variant with dominant contact
chr2	116200000	116449999	0.0349	-0.066	0.023346	0.0130	A
chr4	113900000	114099999	-0.052	0.1482	0.051535	0.0416	B
chr7	5650000	5899999	-0.056	0.0822	0.034630	0.0299	B
chr12	32400000	32599999	-0.069	0.1230	0.047156	0.0265	B
chr14	14550000	14749999	-0.023	0.1641	0.048366	0.0453	B
chr14	35650000	35949999	-0.011	0.1388	0.036768	0.0275	B
chr14	49850000	50199999	-0.001	0.1988	0.045096	0.0128	B
chr14	50300000	50749999	0.0227	0.3159	0.057641	0.0133	B
chr14	51350000	51899999	-0.029	0.2844	0.069257	0.0111	B
chr14	52350000	53249999	0.0064	0.2635	0.056101	0.0107	B
chr14	53300000	53549999	-0.012	0.2029	0.053298	0.0318	B
chr14	55950000	56399999	0.0110	0.2389	0.054441	0.0201	B
chr14	58750000	58949999	-0.005	0.1700	0.042480	0.0208	B
chr14	73800000	73999999	-0.019	0.1497	0.040422	0.0186	B
chr15	24700000	24899999	0.1211	-0.063	0.045232	0.0240	A
chr15	33800000	33999999	0.1611	-0.063	0.057087	0.0368	A
chr18	36650000	38649999	0.6514	0.0257	0.092426	0.0021	A
chr18	39000000	39199999	0.1689	-0.025	0.048081	0.0276	A
chr18	40450000	40799999	0.2428	-0.069	0.069635	0.0127	A
chr18	49450000	49649999	0.1522	-0.008	0.040796	0.0373	A
chrX	20000000	20199999	-0.023	0.1365	0.039340	0.0258	B
chrX	24550000	24899999	-0.088	0.2264	0.070479	0.0106	B
chrX	27550000	27799999	0.0042	0.2855	0.065110	0.0133	B
chrX	27900000	28099999	0.0257	0.2909	0.059155	0.0070	B
chrX	28900000	29099999	0.0079	0.2784	0.068641	0.0373	B
chrX	30000000	30299999	-0.067	0.1420	0.050297	0.0222	B
chrX	35650000	35899999	-0.021	0.1405	0.041462	0.0419	B
chrX	39050000	39399999	-0.051	0.1727	0.052845	0.0203	B
chrX	40150000	40349999	-0.027	0.1195	0.036997	0.0350	B
chrX	45150000	45349999	-0.031	0.1502	0.046631	0.0415	B
chrX	45800000	46049999	-0.044	0.1678	0.048989	0.0124	B
chrX	46950000	47299999	-0.020	0.1378	0.037804	0.0189	B
chrX	48750000	48949999	-0.051	0.2043	0.059251	0.0128	B
chrX	50200000	50699999	0.0336	0.3044	0.059578	0.0159	B
chrX	57400000	57649999	-0.032	0.2406	0.058311	0.0039	B
chrX	58150000	58449999	-0.038	0.1236	0.041182	0.0405	B
chrX	59600000	59799999	-0.044	0.1135	0.040447	0.0402	B

4.2.5 Binarization and Fisher test

The binarization is performed on the estimated coverage obtained from *Salmon* in the same way that was calculated for *Arabidopsis thaliana* data, with respect to a threshold $R = 1$. Then the Fisher exact test is used as it is described in Methods to identify DCRs. The input to the Fisher exact test, a table in .csv format with binary results for all replications under each treatment, is presented partially in Table 31; the full table is available in Supplementary_file_7.csv in Supplementary Files.

Table 31. Input to the Fisher test for all samples of each treatment

Chromo-some	Fragment start	Fragment end	At_A_1	At_A_2	At_A_3	At_B_1	At_B_2	At_B_3
chr1	3082149	3088708	1	1	1	0	0	1
chr1	3088702	3089099	0	0	1	0	0	1
chr1	3089093	3089524	1	0	1	0	0	1
chr1	3089518	3095350	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…

A sliding window approach was used to locate DCRs with the Fisher exact test, with window parameters in the range from 40000 to 200000 for the size of the window and 20000 to 100000 for the step, in different combinations, as presented in Table 29. Significant difference between experimental variants ($P < 0.05$) means that the window was in contact with the bait differently in A and B (is a DCW). A fragment of the resulting table can be seen in the table 32 (Full table with binary results for mm10 with sliding window 200000 and step 50000 is available in Supplementary_file_8.csv in Supplementary Files).

Table 32. Fisher exact test output

Chro moso me	Window start	Window end	Fisher p-value	Probability of window coverage under condition		Number_ of_fragm ents in the window	Number of covered fragments under condition A and B in all replications	
				mm_A	mm_B		mm_A	mm_B
chr1	100150000	100349999	0.840	0.062	0.053	75	14	12
chr1	100200000	100399999	1	0.052	0.052	70	11	11
chr1	100250000	100449999	0.681	0.063	0.049	74	14	11
chr1	100300000	100499999	0.840	0.062	0.053	75	14	12
...	...	...	...	...	...	...	...	...

The data in this table are used to produce genomic regions (DCRs) by merging overlapping windows of restriction fragments. Exemplary visualisation of results for different parameters used in the sliding window procedure is shown in Fig. 25.

We should note here that the Fisher test was less efficient in discovering DCWs and DCRs in mouse experiment than in *Arabidopsis* experiment. As the mouse genome is much bigger, and the number of NGS reads was comparable, there is a number of fragments in many sliding windows which are uncovered, which gives low frequency of coverage. The Fisher test was not able to declare significant differences between those low frequencies. In effect, there were

windows for which the uncorrected *P* values were small, but if a correction (FDR adjustment) of *P* values was done, the number of DCWs was very limited (a few in the whole genome). Therefore, in the inference for mouse data we use uncorrected *P* values; this approach is further commented in section 4.2.6 and in Discussion.

The analysis of Fig. 25 shows that there are differences in results for different sizes of sliding window but they are not as obvious as in the case of *Arabidopsis thaliana*. There are different situations connected with the length of the sliding window. By increasing the length of the window from 100000 to 200000 we can find significant regions that does not exist in smaller windows. Results for windows with parameters (80000, 40000) and (100000, 50000) were similar.

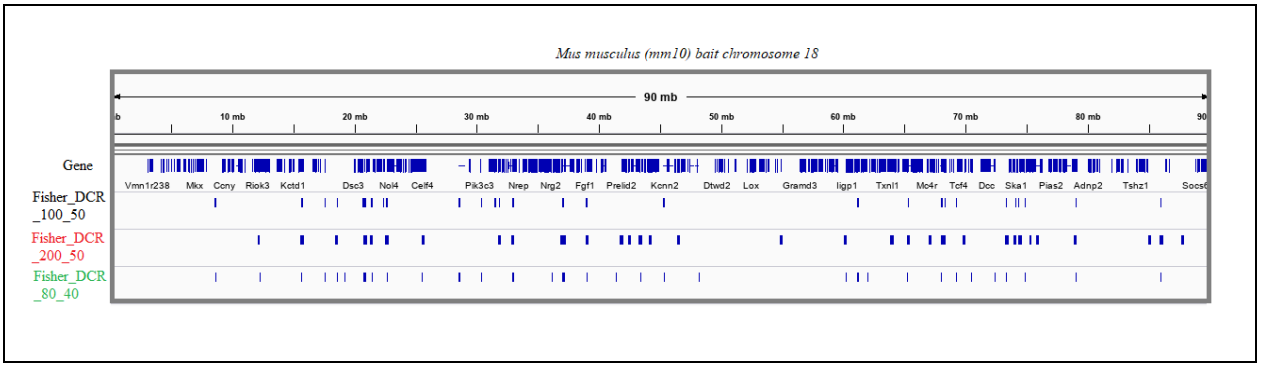

Figure 25: Visual representation of "differentially contacting regions" (DCR) after Fisher exact test for three different sizes and two different steps of sliding windows in bait chromosome 18.

The results for all sets of parameters of sliding windows that have been tested are shown in Table 33. We observe that by increasing the window length we can increase the number of DCWs and DCRs. Increasing the step while keeping the length may produce less significant declarations (see the difference between versions 1 and 2).

Table 33. Results of Fisher exact test for different versions of sliding window and step; DCWs and DCRs are defined using uncorrected *P* values from Fisher test.

Version	Window size	Step	Number of			
			differentially contacting windows (DCW)	differentially contacting regions (DCR)	DCR - variant A	DCR - variant B
1	200000	100000	542	418	29	389
2	200000	50000	1097	558	43	515
3	100000	50000	614	468	33	435
4	100000	25000	1209	619	52	567
5	80000	40000	607	463	37	426
6	80000	20000	1210	618	62	556
7	40000	20000	583	431	30	401

In Supplementary Table 34[3] we present all 558 DCRs obtained for sliding window parameters (200000, 50000) with Fisher test. Each DCR is characterized by parameters averaged over all

[3] Supplementary Table 34 is available on CD attached to the manuscript

windows that belong to it: average probability of fragment coverage in variant A and B, and average *P* value. Similarly as in *Arabidopsis thaliana* genome, there are some DCRs with a *P* value bigger than the threshold 0.05, due to the presence of windows which do not express significant difference in contacts between variants.

4.2.6 Comparison of results from LMM and Fisher test

As it is mentioned above, in the Fisher test we observed an issue with the corrected *P* values due to the big size of the genome and the existence of a big number of uncovered fragments. In Fig. 26 we show the observed distribution of *P* values for both LMM and Fisher tests before and after FDR correction. In Fisher test, before correction, there was a big number of Fisher *P* values close to 1. The FDR correction procedure transformed almost all *P* values to the ones close to 1. Thus, the correction of Fisher *P* values is inefficient.

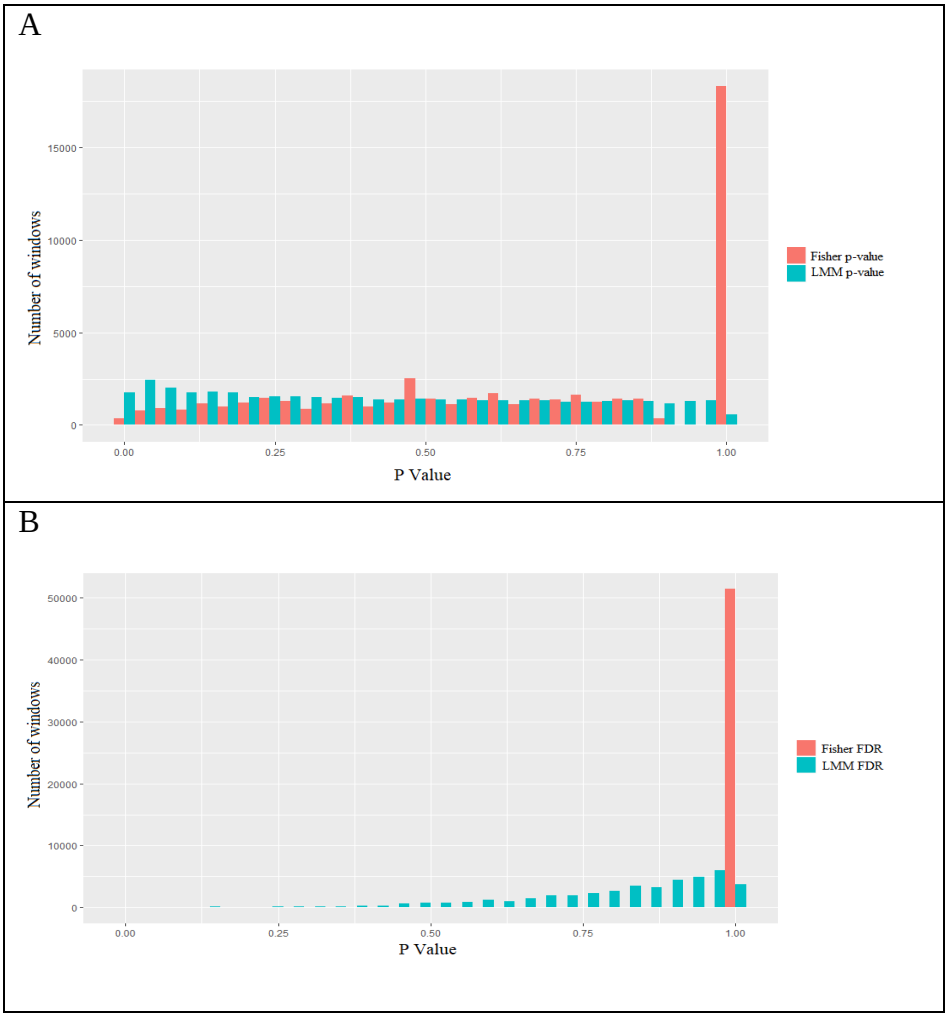

Figure 26. Distribution of: A. *P* values and B. corrected (FDR) *P* values for both LMM (blue bars) and Fisher test (red bars).

By analysing tables of results of LMM model and Fisher test, we can conclude that in both methods the parameters of the sliding windows have an important influence on the results. In both cases very big window sizes (>1000000) return a wide overlapping region in all chromosome which do not lead to any interpretation. Fig. 27 presents a visual representation of DCRs obtained with the LMM model (corrected *P* values) and Fisher exact test (uncorrected *P* values) for the same size and step of a sliding window.

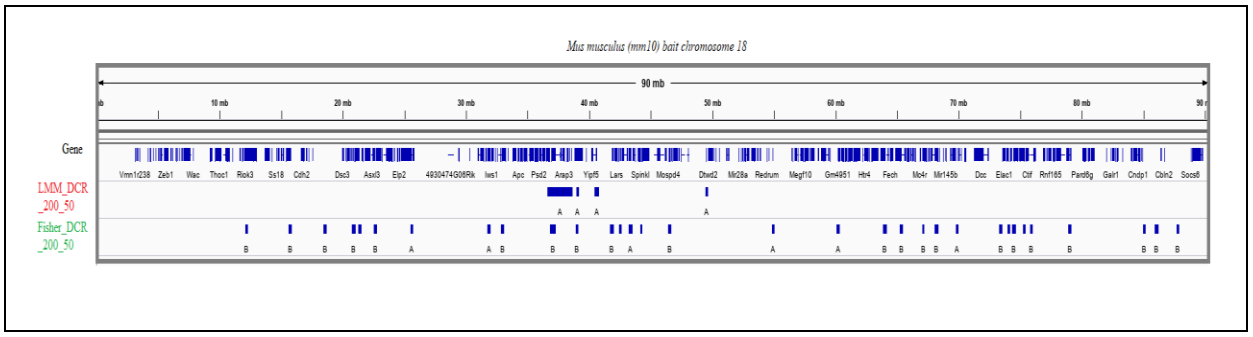

Figure 27. Visual representation of differentially contacting regions (DCR) for the same window (200000) and step (50000), after analysis weith LMM model and Fisher exact test, in bait chromosome 18.

Similarly as in the case of *Arabidopsis* data, we compare the statistical characteristics of genomic regions identified by LMM analysis and Fisher test analysis for the same sliding windows parameters. In Fig. 28 the lack of correlation between the coverage in variant A and in variant B is visible. The difference between the two mean values was especially big for DCWs of type "A". On the other hand, the scatterplot of estimated fragment coverage probabilities indicates that the probabilities for variant A and variant B were correlated. Even in DCWs, the estimated probabilities for two variants were quite close. Comparison of results obtained by two methods indicates that, in general, there is no relationship between the variant effects and significance scores obtained from two methods (Fig. 29).

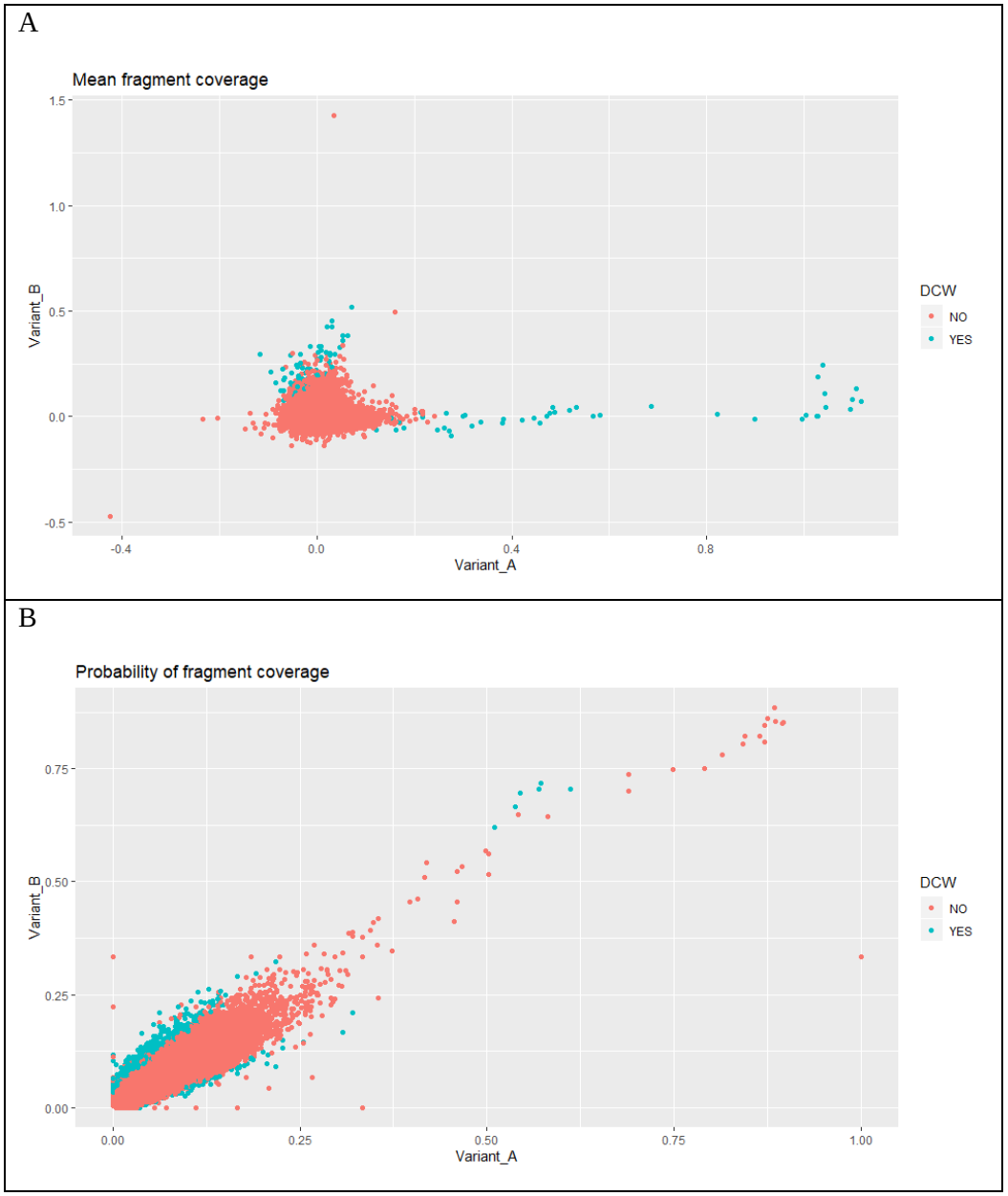

Figure 28. Comparison of parameter estimates in LMM analysis and in analysis by Fisher test. A. Scatterplot of estimated mean fragment coverage in sliding windows for experimental variants A and B. B. Scatterplot of estimated fragment coverage probabilities for variants A and B. Blue dots indicate windows declared as DCW. Results for window length 200000, step 50000 (position 2 in Tables 29 and 33)

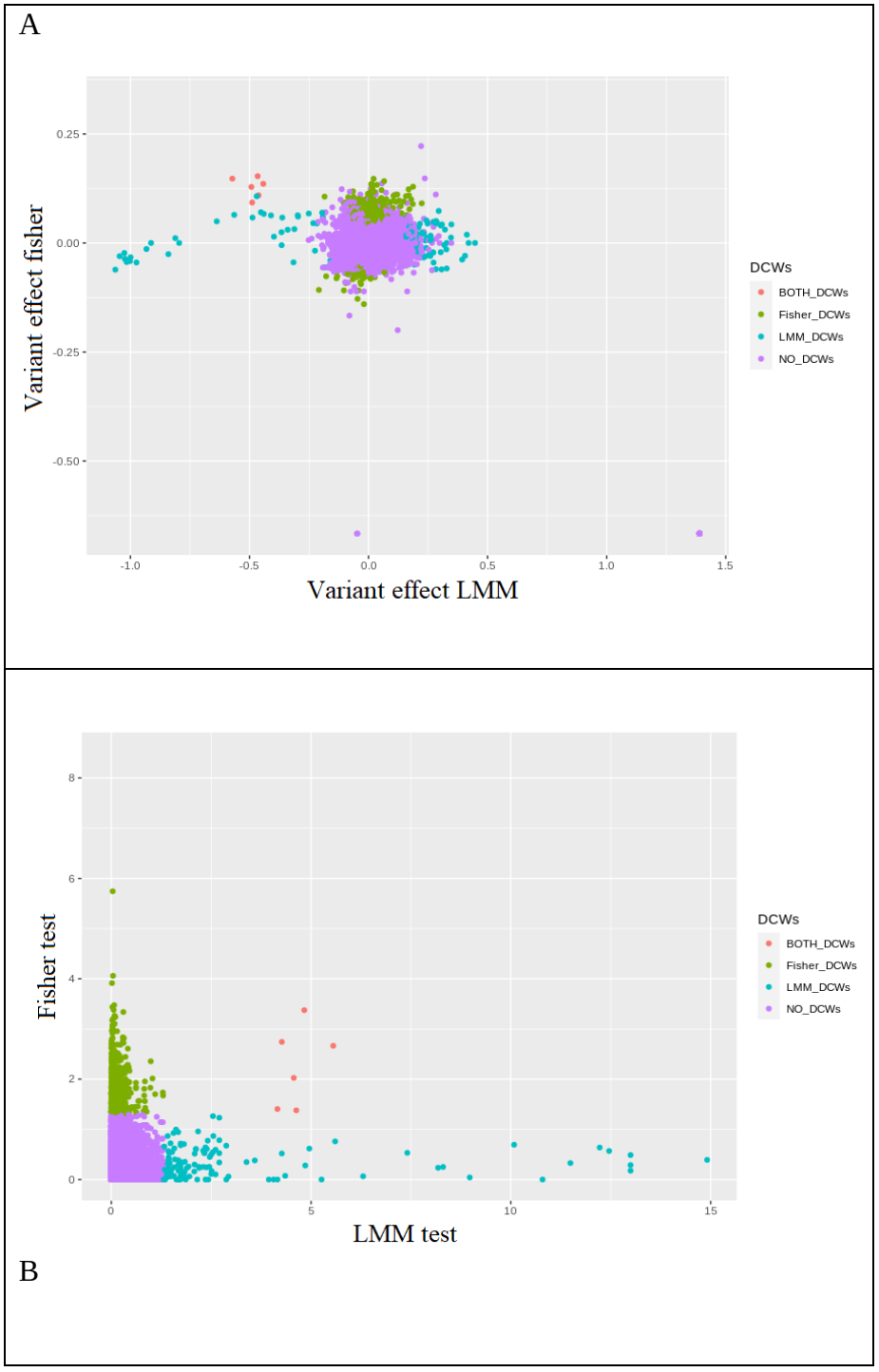

Figure 29. Comparison of results obtained from LMM analysis vs Fisher test analysis. A. Scatterplot of variant effects (B-A) (differences in mean values for LMM, differences in probabilities for Fisher test). B. Scatterplot of significance scores $-\log_{10}$(corrected *P* value for LMM and Fisher *P* value). Colored dots indicate DCWs: green - in Fisher test, blue - in LMM analysis, red - in both analyses. Results for window length 200000, step 50000 (position 2 in Tables 29 and 33).

Table 35 presents exemplary cases of common DCWs between the two methods. In the first DCW it is clear that the coverage level and probability are bigger in variant A, whereas in the second DCW there is a disagreement between the two methods as in LMM we can see bigger coverage in variant A and bigger probability in B. These two cases are also visualized in Fig. 30.

Table 35. DCWs common for two methods

Chromosome	Start	End	LMM analysis				Fisher test			
			Coverage A	Coverage B	Diff.	Corr p-value	Probability A	Probability B	Diff.	Uncorrected p-value
chr14	52900000	53099999	-0.023	0.148	-0.171	0.05	0.047	0.125	-0.078	0.01
chr18	36850000	37049999	0.477	0.015	0.462	6.9e-05	0.039	0.510	-0.471	0.03

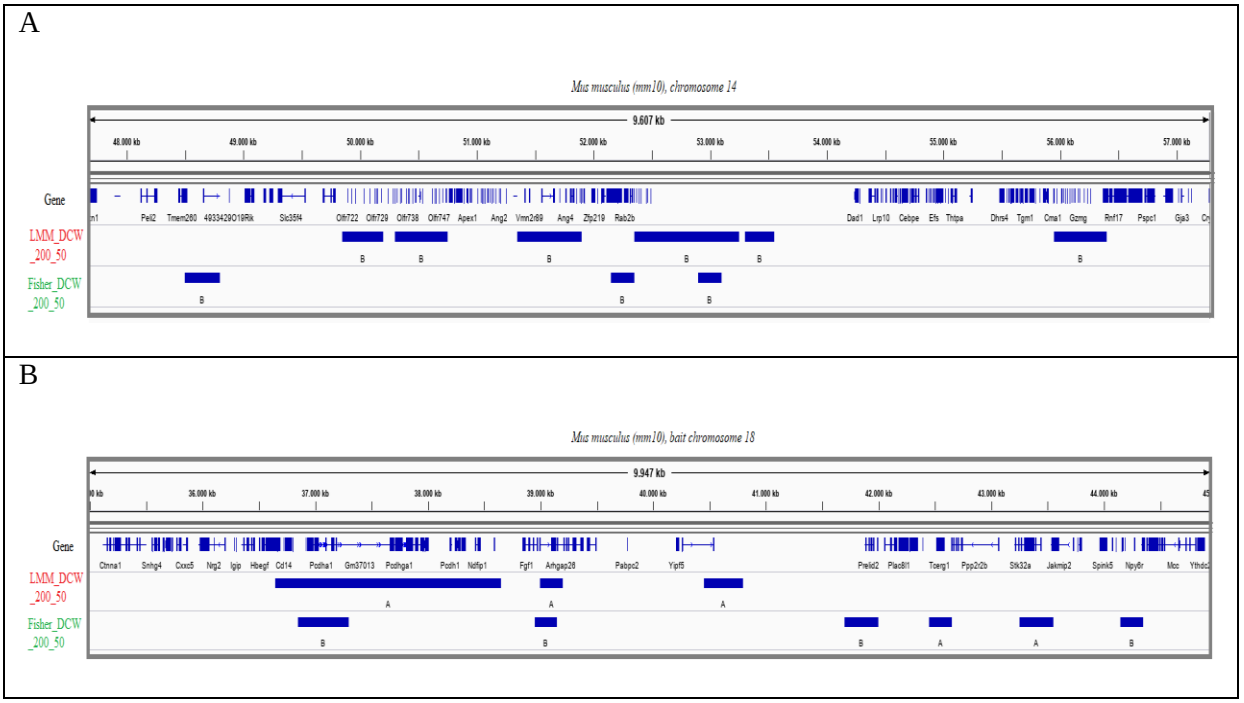

Figure 30: Visual representation of intersecting differentially contacting windows (DCW) obtained by two methods. A) Agreement on the domination variant (B) in chromosome 14 and B) Disagreement on the dominating variant in bait chromosome 18 for the same window (200000) and step (50000), after Fisher exact test and LMM model.

In Table 36 we show the intersection of DCRs found by two methods. There were 37 DCRs obtained using LMM and 558 obtained by Fisher test, which resulted in 7 common (intersecting) regions. In two of them, in chromosome 18 (bait chromosome), the two methods disagree on the variant with dominant contacts.

Table 36. Intersecting DCR from LMM and Fisher test for sliding windows (200000, 50000).

LMM							Fisher test					
Chromosome	**Start**	**End**	**Vari-ant**	**Cove-rage A**	**Cove-rage B**	**Corre-cted *P* value**	**Start**	**End**	**Variant**	**Proba-bility A**	**Proba-bility B**	**Unco-rrected *P* value**
chr14	52350000	53249999	B	0.01	0.26	0.010	52900000	53099999	B	0.047	0.125	0.018
chr18	36650000	38649999	A	0.65	0.03	0.002	36850000	37299999	B	0.557	0.685	0.015
chr18	39000000	39199999	A	0.17	-0.02	0.027	38950000	39149999	B	0.217	0.323	0.019
chr7	5650000	5899999	B	-0.06	0.08	0.029	5500000	5799999	B	0.056	0.120	0.028
chrX	24550000	24899999	B	-0.09	0.23	0.010	24500000	24699999	B	0.084	0.175	0.021
chrX	46950000	47299999	B	-0.02	0.14	0.018	46850000	47049999	B	0.030	0.080	0.046
chrX	50200000	50699999	B	0.03	0.30	0.015	49900000	50349999	B	0.098	0.189	0.016

4.2.7 Comparison to results of *fourSig*

Results from foursig

The analysis with *foursig* tool is done independently for each replication in each sample. A fourSig_*.bed file produced with the significant contacts for each sample (mm_A_1, mm_A _2, mm_A _3, mm_B_1, mm_B_2, mm_B_3) and all bed files are available in Supplementary Files. Similar to *Arabidopsis thaliana* bed files are available and a visualization in a genome browser (IGV) is presented in Fig. 31.

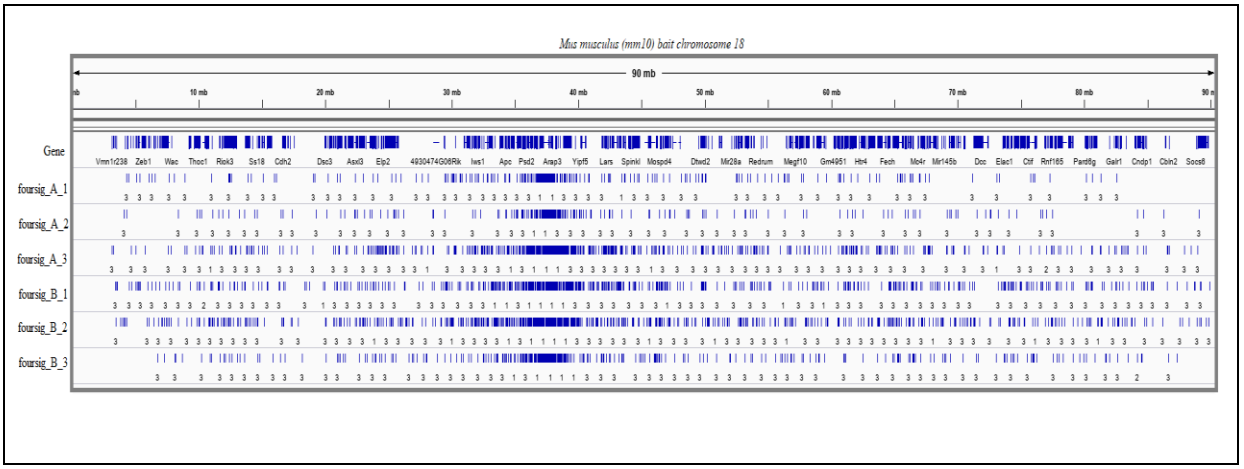

Figure 31. A visual representation of all samples in bait chromosome 18

As it is done with the previous data sets, we performed a comparison of the contacts found in individual replications and experimental variants. For each variant separately, we declared as "significant contacts" regions found repeatedly in all three replications. Then, we counted the regions declared as significant for both variants or for one of the two variants only, in all chromosomes and in the bait chromosome.

Regions significant in variant A and B for all replications are available as *intersect_foursig_varA_mm10.bed* and *intersect_foursig_varB_mm10.bed* and significant contacts in both variants A and B are available in bed file with the name *foursig_common_mm10.bed* in Supplementary Files. A visual representation of common significant contacts is available in the next picture for the bait chromosome 18.

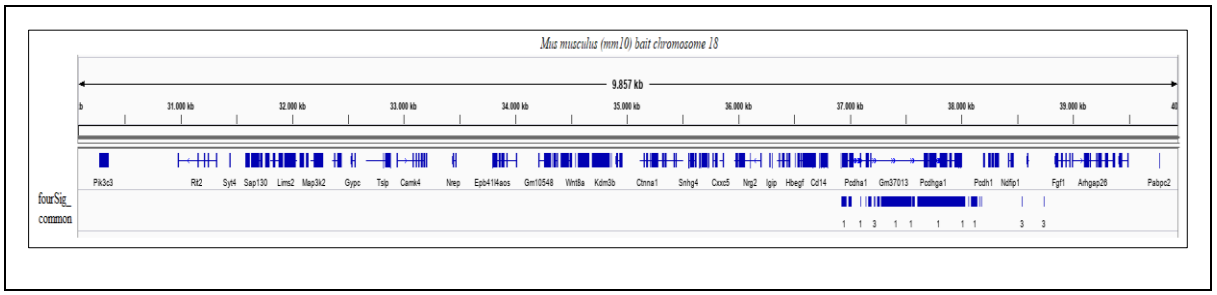

Figure 32. Significant contacts (regions in blue tracks) found in the bait chromosome in both experimental variants in the experiment with mouse cells.

Table 37. Results of *foursig* in all replications of each variant and common/specific significant contacts.

Groups of significant contacts		**Experimental variant**	
		A	B
Replication	1	662	1906
	2	630	2177
	3	2177	1367
Total in variant (common to all replications)		47	64
Specific to variant		25	42
Common		22	

LMM vs foursig

We compare results of LMM and *foursig*, using intersection between DCRs of LMM and regions found as significant contacts in *foursig*. From Table 38, we can see that from 34.0% to 53,1% of *foursig* significant regions in A (47) are confirmed by LMM analysis (1-6 DCRs). But there are no significant *fourSig* regions in B (64) that are confirmed by LMM analysis (0 DCRs). From Tables 39 and 40 it follows that more *fourSig* common contacts are confirmed by LMM than contacts specific to A. More DCRs obtained with windows of smaller length (80000) intersect with *fourSig* contacts than DCRs obtained with longer windows (200000).

Table 38. Number of regions with significant contacts from *foursig* (A 47, B 64) intersecting with DCR from LMM for each variant (varA, varB) and number of DCR from LMM with *foursig* contacts for each variant (varA, varB)

Window size	Step	Foursig_varA in DCR LMM_varA	Foursig_varB in DCR LMM_varB	Number of DCRs of type A with *foursig* contacts (all DCRs type A)	Number of DCRs of type B with *foursig* contacts (all DCRs type B)
200000	100000	21	0	1 (6)	0 (25)
200000	50000	25	0	1 (7)	0 (30)
100000	50000	14	0	5 (6)	0 (2)
100000	25000	21	0	6 (7)	0 (3)
80000	40000	13	0	5 (5)	0 (0)
80000	20000	16	0	6 (7)	0 (0)

Table 39. Number of variant-specific contacts from *foursiq* in each condition (A 25, B 42) intersecting with DCR from LMM for each condition (A and B)

Window size	Step	Foursig_specific varA in DCR LMM_varA	Foursig_specific varB in DCR LMM_varB	Number of DCRs of type A with *foursig* contacts (all DCRs type A)	Number of DCRs of type B with *foursig* contacts (all DCRs type B)
200000	100000	4	0	1 (6)	0 (25)
200000	50000	7	0	1 (7)	0 (30)
100000	50000	3	0	2 (6)	0 (2)
100000	25000	4	0	2 (7)	0 (3)
80000	40000	1	0	1 (5)	0 (0)
80000	20000	4	0	2 (7)	0 (0)

Table 40. Number of common contacts from *foursig* (22) intersecting with DCR from LMM

Window size	Step	foursig_common in DCR LMM_varA	foursig_common in DCR LMM_varB	Number of DCRs of type A with *foursig* contacts (all DCRs type A)	Number of DCRs of type B with *foursig* contacts (all DCRs type B)
200000	100000	17	0	1 (6)	0 (25)
200000	50000	18	0	1 (7)	0 (30)
100000	50000	11	0	3 (6)	0 (2)
100000	25000	17	0	4 (7)	0 (3)
80000	40000	12	0	4 (5)	0 (0)
80000	20000	12	0	4 (7)	0 (0)

Fisher vs foursig

Similarly as for LMM, we compared results of *foursig* analysis with results of Fisher test (Tables 41, 42, 43). No *foursig* contacts found in variant A are confirmed by Fisher test, and from 13.8% to 18.5% of contacts found in variant B intersect with Fisher test DCRs type B. Very small fraction of DCRs of type A or B intersect with *foursig* contacts.

Table 41. Number of regions with significant contacts from *foursig* (A 47, B 64) intersecting with DCR from Fisher test for each variant (varA, varB) and number of DCR after Fisher test with *foursig* contacts for each variant (varA, varB)

Window size	**Step**	**Foursig_varA in DCR Fisher_varA**	**Foursig_varB in DCR Fisher _varB**	**Number of DCRs of type A with *foursig* contacts (all DCRs type A)**	**Number of DCRs of type B with *foursig* contacts (all DCRs type B)**
200000	100000	0	9	0 (29)	2 (389)
200000	50000	0	12	0 (43)	3 (515)
100000	50000	0	9	0 (33)	3 (435)
100000	25000	0	9	0 (52)	3 (567)
80000	40000	0	10	0 (37)	5 (426)
80000	20000	0	11	0 (62)	4 (556)

Table 42. Number of significant variant-specific contacts from *foursiq* in each variant (A 25, B 42) compared to DCR after Fisher test for each variant (A and B)

Window size	**Step**	**Foursig_specific varA in DCR Fisher_varA**	**Foursig_specific varB in DCR Fisher_varB**	**Number of DCRs of type A with *foursig* contacts (all DCRs type A)**	**Number of DCRs of type B with *foursig* contacts (all DCRs type B)**
200000	100000	0	2	0	2
200000	50000	0	5	0	3
100000	50000	0	5	0	3
100000	25000	0	5	0	3
80000	40000	0	6	0	4
80000	20000	0	7	0	4

Table 43. Number of significant contacts from *foursig* (22) in both variants (common) in DCR after Fisher test

Window size	**Step**	**foursig_common in DCR Fisher varA**	**foursig_common in DCR Fisher varB**	**Number of DCRs of type A with *foursig* contacts (all DCRs type A)**	**Number of DCRs of type B with *foursig* contacts (all DCRs type B)B**
200000	100000	0	8	0	1
200000	50000	0	8	0	1
100000	50000	0	4	0	1
100000	25000	0	5	0	1
80000	40000	0	5	0	2
80000	20000	0	6	0	1

4.3 Exemplary downstream analysis of DCRs in *Arabidopsis thaliana* and *Mus musculus*

4.3.1 *Arabidopsis thaliana*

As it was shown in Sec. 4.1.6 above, there were 3 DCRs that overlapped between results of LMM and Fisher approach (Tab. 15). In order to further analyse and explore those differentially contacting regions, an exemplary analysis is presented. A list of genes for *Arabidopsis thaliana* genome was downloaded from Ensemble Plants database by using the *BioMart* tool, and a file in bed format with gene coordinates and IDs was produced. The result of intersection of DCRs and genes is a list with 303 genes, available in Supplementary_Table_44 for further exploration. A fragment of this table is shown below.

Table 44. Results of DCRs intersecting with genes in *Arabidopsis thaliana*

Chromosome	Start	End	Variant	Gene ID
chr5	2950000	3399999	A	AT5G09970
chr5	2950000	3399999	A	AT5G10010
chr5	2950000	3399999	A	AT5G01675
chr5	2950000	3399999	A	AT5G10140
chr5	2950000	3399999	A	AT5G10180
chr5	2950000	3399999	A	AT5G10230
chr5	2950000	3399999	A	AT5G10278
chr5	2950000	3399999	A	AT5G10320
chr5	2950000	3399999	A	AT5G10470
chr5	2950000	3399999	A	AT5G10540
chr5	2950000	3399999	A	AT5G10530
chr5	2950000	3399999	A	AT5G10570
chr5	2950000	3399999	A	AT5G10690
chr5	2950000	3399999	A	AT5G10730
chr5	2950000	3399999	A	AT5G10050
…	…	…	…	…

To further investigate the genes that were found in the 3 DCRs of *Arabidopsis thaliana*, a Gene Ontology overrepresentation analysis can be done. With this approach we can determine whether known biological functions, molecular processes or cellular components are overrepresented in annotations of genes present in genomic regions characterized by differential contacts with the FLC locus. For this purpose, the list of 303 genes was submitted to GO enrichment analysis at Geneontology.org and the results of the analysis for terms describing biological process or cellular components are presented in the Tables 45 and 46.

In those tables the outputs of GO enrichment analysis for *Arabidopsis thaliana* genes are presented. The results are sorted in hierarchical categories and only those with corrected *P* value smaller than 0.05 are shown. First column represents the name of annotation category, second - the number of genes of the reference list that align to the annotation, the third column - the number of genes in our results that align to the annotation, the fourth column - the

expected number of genes for this category, the fifth column shows the fold enrichment of the genes (number in our list divided by the expected number, a value >1 means overrepresentation), sixth column contains either a + or – to characterise over-representation (+) or under-representation (-), the seventh column is the *P* value by Fisher's exact test (can be used also Binomial) and the eighth column is the False Discovery Rate obtained by the Benjamini-Hochberg procedure.

Table 45. Results of GO enrichment analysis for *Arabidopsis thaliana* genes present in DCRs found in the comparison of vernalized and not vernalized plants – GO terms for biological processes.

GO biological process complete	**Arabidopsis thaliana(REF)**	**Genes found in DCRs**					
	#	**#**	**expected**	**Fold Enrichment**	**+/-**	**raw *P* value**	**FDR**
ATP synthesis coupled electron transport	49	5	.25	19.70	+	8.74E-06	1.05E-02
respiratory electron transport chain	57	6	.30	20.32	+	8.98E-07	1.79E-03
electron transport chain	113	7	.59	11.96	+	3.04E-06	4.55E-03
generation of precursor metabolites and energy	377	10	1.95	5.12	+	3.41E-05	2.27E-02
cellular respiration	138	9	.71	12.59	+	7.47E-08	4.47E-04
energy derivation by oxidation of organic compounds	153	9	.79	11.36	+	1.72E-07	5.14E-04
oxidative phosphorylation	57	5	.30	16.94	+	1.73E-05	1.29E-02
ATP metabolic process	145	7	.75	9.32	+	1.44E-05	1.23E-02
aerobic respiration	89	6	.46	13.02	+	1.00E-05	1.00E-02

Table 46. Results of GO enrichment analysis for *Arabidopsis thaliana* genes present in DCRs found in the comparison of vernalized and not vernalized plants – GO terms for cellular components.

GO cellular component complete	**Arabidopsis thaliana(REF)**	**Genes found in DCRs**					
	#	**#**	**expected**	**Fold Enrichment**	**+/-**	**raw P value**	**FDR**
cytochrome complex	30	4	.16	25.74	+	2.78E-05	9.66E-03
respiratory chain complex	109	6	.56	10.63	+	2.99E-05	7.79E-03
respirasome	116	6	.60	9.99	+	4.17E-05	7.24E-03
mitochondrial respirasome	111	6	.57	10.44	+	3.29E-05	6.87E-03
mitochondrial inner membrane	245	7	1.27	5.52	+	3.43E-04	4.47E-02
mitochondrion	4402	61	22.80	2.68	+	3.64E-14	3.79E-11
inner mitochondrial membrane protein complex	177	7	.92	7.64	+	4.91E-05	7.32E-03
mitochondrial protein complex	260	10	1.35	7.43	+	1.47E-06	7.68E-04

4.3.2 *Mus musculus*

In case of experiment with mouse cells, as we have already seen in Sec. 4.2.6, there were 7 DCRs that were overlapping between results of using LMM and Fisher test approach (Tab. 7). We demonstrate further exploration of those differentially contacting regions by intersecting them with genes list downloaded from Ensemble database by using the *BioMart* tool. The result of intersection is a list of 63 genes present in those 7 DCRs available in Supplementary_Table_47 for further exploration.

To further investigate the genes that are found in the 7 DCRs declared in experiment with mouse cells, an overrepresentation analysis is done by using GO enrichment analysis in geneontology.org in the same way as above and the results for the GO terms describing biological processes, molecular functions and cellular components are presented in the Tables 48, 49 and 50.

Table 48. Results of GO enrichment analysis for *Mus musculus* genes present in DCRs found in the comparison of ESCs and iPSCs – GO terms for biological processes.

GO biological process complete	**Mm10(REF)**	**Genes in DCRs**					
	#	**#**	**expected**	**Fold Enrichment**	**+/-**	**raw P value**	**FDR**
cell adhesion	843	17	1.48	11.51	+	1.81E-14	2.86E-10
biological adhesion	853	17	1.49	11.38	+	2.19E-14	1.73E-10
cellular response to stimulus	6554	1	11.48	.09	-	2.91E-05	2.42E-02
multicellular organismal process	7381	1	12.93	.08	-	3.98E-06	5.71E-03
regulation of primary metabolic process	5708	0	10.00	< 0.01	-	2.00E-05	1.97E-02
regulation of metabolic process	6564	0	11.50	< 0.01	-	1.55E-06	3.06E-03
regulation of biological process	11968	1	20.96	.05	-	4.17E-12	1.65E-08
biological regulation	12557	1	22.00	.05	-	4.68E-13	2.46E-09
nitrogen compound metabolic process	5781	0	10.13	< 0.01	-	1.11E-05	1.17E-02
metabolic process	7366	0	12.90	< 0.01	-	3.09E-07	8.13E-04
regulation of macromolecule metabolic process	6082	0	10.65	< 0.01	-	5.66E-06	6.87E-03
organic substance metabolic process	6854	0	12.01	< 0.01	-	7.50E-07	1.69E-03
primary metabolic process	6351	0	11.12	< 0.01	-	2.90E-06	4.58E-03
cellular metabolic process	6425	0	11.25	< 0.01	-	2.78E-06	4.88E-03
positive regulation of cellular process	5709	0	10.00	< 0.01	-	2.00E-05	1.86E-02
positive regulation of biological process	6198	0	10.86	< 0.01	-	5.44E-06	7.16E-03
regulation of cellular process	11461	1	20.08	.05	-	2.99E-11	9.45E-08
negative regulation of biological process	5264	0	9.22	< 0.01	-	3.79E-05	2.99E-02
regulation of nitrogen compound metabolic process	5546	0	9.71	< 0.01	-	2.03E-05	1.78E-02
regulation of cellular metabolic process	5915	0	10.36	< 0.01	-	1.05E-05	1.18E-02

Table 49. Results of GO enrichment analysis for *Mus musculus* genes present in DCRs found in the comparison of ESCs and iPSCs – GO terms for cellular components.

GO cellular component complete	Mm10(REF)	Genes in DCRs					
	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
integral component of plasma membrane	1535	17	2.69	6.32	+	2.26E-10	2.23E-07
cellular anatomical entity	19007	24	33.29	.72	-	2.31E-04	3.51E-02
intrinsic component of plasma membrane	1619	17	2.84	5.99	+	5.11E-10	3.36E-07
Unclassified	1371	14	2.40	5.83	+	4.12E-08	1.63E-05
cytoplasm	11107	4	19.46	.21	-	3.30E-07	7.25E-05
intracellular	14060	5	24.63	.20	-	1.19E-10	2.35E-07
nucleus	7104	2	12.44	.16	-	9.18E-05	1.51E-02
intracellular membrane-bounded organelle	10627	4	18.61	.21	-	9.69E-07	1.91E-04
membrane-bounded organelle	11399	4	19.97	.20	-	1.17E-07	2.89E-05
organelle	12652	5	22.16	.23	-	2.32E-08	1.14E-05
intracellular organelle	12309	5	21.56	.23	-	4.76E-08	1.34E-05
protein-containing complex	5338	0	9.35	< 0.01	-	3.65E-05	6.55E-03

Table 50. Results of GO enrichment analysis for *Mus musculus* genes present in DCRs found in the comparison of ESCs and iPSCs – GO terms for molecular functions.

GO molecular function complete	Mm10(REF)	Genes in DCRs					
	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
Unclassified	1979	26	3.47	7.50	+	1.37E-18	6.38E-15
protein binding	9311	3	16.31	.18	-	4.29E-06	4.98E-03
binding	13713	4	24.02	.17	-	3.77E-11	5.84E-08
catalytic activity	5696	0	9.98	< 0.01	-	1.99E-05	1.84E-02
organic cyclic compound binding	5405	0	9.47	< 0.01	-	3.64E-05	2.81E-02
heterocyclic compound binding	5303	0	9.29	< 0.01	-	3.70E-05	2.45E-02

5. Discussion

The aim of this work was to propose a new alternative method for the analysis of 4C-seq data. After comparing the existing methods at each step of their analysis, as it is described in [Zisis et al. 2020], we were able to understand the mechanisms of existing methods and the focus that they provide. We identified some of their shortages and were able to address the issues by application of methods and algorithms not previously used in the analysis of 4C data. We start discussion by analysing if these methods are applied properly and if they provide advantages in the analysis, and then discuss general recommendations that follow from our study and may concern other applications.

Coverage estimation

Simple counting of the reads mapped to restriction fragments, based on results of a chosen NGS read mapper (like *Bowtie2* used in our pipeline), is not sufficient for proper estimation of relative abundance of reads coming from different genomic loci. Various properties can influence the chances of selection of particular sequences for amplification and sequencing. For RNA-seq experiments, this topic was considered by Patro et al. (2017), who, in particular, described the meaning of the "sequence-specific" bias and of the "GC-content" bias. Their computational pipeline, *Salmon*, deals with these biases using a model in which the probability that a given NGS read was obtained from a given transcript is estimated. This probabilistic model is based on a generative model of RNA-seq experiment described by the authors.

It can be assumed that a similar type of biases are present in the case of 4C-seq experiment. However, in *Salmon*, the methods of dealing with them are specifically designed for transcriptom data. For example, the sequence-specific bias is linked to the properties of sequences surrounding the 5' and 3' ends of transcripts, and is addressed in the computations by training the model on respective genomic windows. This takes place when *Salmon* is used in its first mode of application; in this mode the input data consists of the set of NGS reads in FASTA format and the reference genome. Because the generative model of 4C-seq data is different, we cannot assume that this mode of Salmon would be correct for our data.

On the other hand, in its second mode, *Salmon* uses the set of mapping results, that is, data on the positions and on detailed alignments of the recorded sequences to the reference genome coded in the form of CIGAR strings. An "alignment model", based on first-order Markov formula, is formed and used to estimate the probability that a given read came from a given transcript. We think that this model is applicable to 4C-seq data and, therefore, we use *Salmon* to estimate the restriction fragment coverage on the basis of alignments pre-computed, e.g., by *Bowtie2*.

The four 4C-seq data existing analysis methods mentioned in this thesis deal with this problem in various ways (Zisis et al. 2020). Two of them, *fourSig* and *FourCSeq*, ignore this problem in the sense that they accept, as input, mapping results, which can be adjusted for non-uniqueness if the user wishes to do so (in an extreme approach, only uniquely mapped reads can be used, which is a wasteful approach). The other two, *4C-ker* and *w4Cseq*, use a

library of unique restriction fragments for mapping and filter out reads mapped to different locations with the same quality, respectively. On the basis of the in-depths alignment model described by Patro et al. (2017) and recommendations of Zhang et al. (2017) we think that application of *Salmon* works to the advantage of precision and speed of computations.

Data normalization and transformation

As the second step in our methodology we described the methods of data transformation with the aim of removing the effects of variability of restriction fragments. The considered properties of fragments were: primary restriction fragment length, primary restriction fragment end length, and the existence of the secondary restriction site within the primary restriction fragment. We have demonstrated differences, both in location and shape, between distributions of coverage in classes of fragments differing by the above-mentioned properties.

In this situation, if we want to improve the inference, the solution is to transform the data to homogeneous distribution by either quantile transformation or ranking. In the context of NGS data analysis, these two methods are described by Qiu et al. (2013). Note that, in this context, in most situations, the mentioned transformations are used to homogenize data distributions between samples or experimental variants. In our case, the aim is to remove biases that exist, within each sample, due to properties of genomic sequences varying along the genome.

From the two possible transformations, we have chosen ranking. Ranks are used in statistics in the situations where the information to order of observations is sufficient for inference. Ranking is considered as superior to quantile normalization in terms of robustness to outliers. In our case, this may mean beneficial removing effects of existence of extremely long or short restriction fragments, or restriction fragments with extremely long or short ends.

However, ranking means also loosing information on data features which can be important. The data transformed by ranking have a uniform distribution (over the interval <0,1> if ranks are divided by the number of ranked elements, as it was done here). In terms of statistical inference, ranking leads to application of non-parametric methods. This may lead to an interesting variant of the method that we propose, not investigated by us so far.

However, to stay in the domain of parametric statistics, we use the second function, inverse normal transformation (INT), which transforms the non-normally distributed ranks into observations that follow the normal distribution. Its application in our pipeline is justified by the fact that the initial non-transformed data have a unimodal, although not quite symmetric, probability distribution. So, application of INT recovers this property. As we stated, selection of the value of INT parameter c other than 0 probably has no effect on the analysis, which follows also from the remarks of Beasley and Erickson (2009) who noted that various versions provide almost linear transformations of the result for $c = 0$.

We also use transformation of data to a binary form. Binarization of restriction fragment coverage in 4C-seq data is used in *4Cseqpipe* (van de Werken et al., 2012 b) for the remote (intra- and inter-chromosomal) contacts. A background model based on the characteristics of the unique fragments (their length, length of their ends, and presence of the secondary restriction site – as already mentioned above) is generated and used to obtain the expected

coverage for classes of fragments with similar properties. Then, the ratio of observed and expected coverage is computed for genomic windows of a given width (e.g., 10 kbp), and the data are transformed to the binary form indicating only that a given fragment is covered (read count > 0) or not covered (read count = 0).

Binarization is also used in *w4Cseq* pipeline (Cai et al., 2016) to find contacts both remote and proximal to the viewpoint. After binarization, a sliding window is applied to assign statistical significance to genomic regions. In a specific window, covered sites are seen as "successes" and uncovered sites are seen as "failures". Based on the binomial model, *P* value is calculated as the probability of observing a given number or more of captured sites in the foreground window in comparison to a bigger background window.

Van de Werken et al. (2012 b) claim that binarization of 4C-seq is justified by the analysis of several data sets. They maintain that this approach operates on sufficient data in the view of a low correlation between the number of reads mapped to the 5' and to the 3' ends of the same restriction fragment. Because of this, reducing the data to the binary form does not cause a loss of information. Zisis et al. (2020) showed that this is not always true. In both data sets, analysed also in this thesis, the probabilities of the 3' fragment end being covered and not being covered conditional on value of the 5' fragment end coverage were not constant (Zisis et al., 2020, Figure 6). Therefore, binarization may lead to loss of information.

However, in this work, we use the binarized data in addition to estimated coverage transformed by ranking and INT expressing the quantitative information about contacts with the bait. The information possibly lost in binarization is maintained in the other transformed variable used in the analysis. The dual description of contacts, by continuous signal intensity and binarized signal presence/absence, contributes to meeting requirements set in our work.

Relative approach

As already mentioned, the existing 4C-data analysis tools that provide a differential analysis work in a two-stage mode: firstly, they locate significant contacts in each sample independently, and then use the obtained results to find differences between samples or experimental variants. In the first step, they use various methods and models; for a summary see Zisis et al. (2020). Shortly: *fourSig* tool detects significant interactions using a procedure based on the observed distribution of NGS reads among restriction fragments and on random relocations of the reads to restriction fragments; *FourCSeq* works in the region proximal to the viewpoint and models contacts with function of the genomic distance from the captured fragment to the viewpoint; *4C-ker* uses a three-state Hidden Markov Model, with read counts as input data, to divide the windows into three categories that interact with the viewpoint with high frequency, with low frequency, and those that do not interact; *W4cseq* uses a statistical model approach based on the binarized results to find contacts remote and proximal to the viewpoint. In the second step, *FourCSeq* and *4C-ker* test differences between conditions with *Deseq2*, the tool originally designed for analysis of transcriptomic data.

In our opinion, in the analysis of trials that study some experimental factors using biological replications, this mode of analysis creates a number of issues. Firstly, fitting a model of

contacts using data obtained from a single sample leads to the necessity of distinguishing between analysis of proximal and distant contacts, as it is difficult to find a model appropriate for different levels of coverage characterizing different genomic regions. Secondly, a lack o repeatability between replications, caused by a low coverage in some genomic regions, can influence the fit of the individual models and the statistical power of the analysis, in a way that will be difficult to track and correctly interpret. Thirdly, the differential analysis greatly depends on the first step.

To overcome these issues, we propose to perform a one-stage procedure that reveals the most important results of the experiment, namely, the differences between experimental variants. We use the well-known statistical models to declare as significant such differences between variants that are large in relation to the variability between biological replications. The models are fitted locally, within genomic windows. The first issue mentioned above is mostly overcome, as we use models which, with some limitations, are appropriate for a relatively large spectrum of coverages. The second issue, basically, does not exist: in the extreme case of a very low repeatability no significant differences between experimental variants will be found, but also the user will not be prompted to interpret apparently significant results found in single replications and not repeated in the others. Finally, the analysis is not conditional. We think that this approach, based on general rules of experimentation and of statistical analysis, is useful.

Statistical models

Having well-transformed data and the goal of comparative analysis, we selected statistical models appropriate for the task of comparative analysis between experimental variants.

The choice of doing the analysis of quantitative observations of restriction fragment coverage by a linear mixed model is natural. LMM are used to analyse and summarise results of experiments in many situations arising in biological research. Their theory originates from the problems in analysis of genetic and phenotypic data in human, animal and plant research, but applications have also been made in the area of market analysis or industrial statistics (Galecki and Burzykowski, 2013). LMM allow to analyse both planned and observational studies that involve application of experimental factors to change the behaviour of experimental material. In experiments with plants or animals, the factor levels can consist of various management techniques (like fertilization, feeding) or application of stressors or stimulators. Historically, the traits analysed by LMM were usually measurements of phenotypes like plant yield, plan height or animal weight, to mention the simplest cases. Utilization of LMM on a big scale to molecular-level data started with applications in transcriptomics (Williams et al., 2014) and metabolomics (Wanichthanarak et al., 2019). In these situations, LMM is used to analyse a number of traits in an experiment designed to study a number of factors.

Our application of LMM is similar to the situation of gene expression analysis, in which the data are expression levels quantified by numbers of reads mapped to transcripts, i.e., the coverage of transcripts. Studied properties of transcriptomic data that are important from the point of view of LMM led to development of tools like *Deseq2* (Williams et al., 2014) or

limma (Ritchie et al., 2015). In addition to the estimation of effects and testing their significance, these tools propose also transformations of raw expression data – read counts – to some variables which better fulfil the assumptions required by LMM analysis. For example, the *Deseq2* tool transforms data by the use of the so-called regularized logarithm to stabilize variance of observations in their various ranges. As we already noted, *Deseq2* is used by two of the 4C-seq data analysis tools at the stage of comparative analysis.

Our situation is different. We noted that in case of 4C-seq data a transformation is necessary to remove the bias arising from the inherent presence in the genome of restriction fragments with different properties. We proposed ranking followed by INT to transform the data. After this transformation, the data do not fulfil the assumptions required for, e.g., analysis by *Deseq2,* which is designed to run on raw counts. Therefore, we use the LMM in our own setup.

The exact choice of the form of LMM that we use is dictated by the problem at hand. We use fixed effects of experimental variants to quantify the influence of applied factors. Then, we use random effects of fragments within genomic windows, which means that we allow for some variation of the signal along the window which should be eliminated from the comparison of variants. Then, we have random effects of biological replications which are used to estimate the residual variance. Due to the applied transformation, we can claim that the assumptions made in the LMM concerning normality of random effects hold. We can also note that in case of deviations from normality the methods of estimation of fixed effects n LMM are still valid in terms of unbiasedness.

The choice of the second model that we use is also natural. We propose the binomial model, in which by a success we mean the event of a restriction fragment being covered by at least R NGS reads. Within a genomic window, the number of successes is the number of covered fragments. Due to possibly low number of fragments within window, we use the Fisher exact test (Fisher, 1922) to test the hypothesis about homogeneity of distributions of successes between experimental variants; probably, for longer windows with many fragments, also the usual Pearson chi-squared test could be used. Binarized coverage data are also used for 4C-seq data analysis in *4Cseqpipe* to study distant contacts. The second tool that analyzes binarized data, *w4cseq*, declares significance of contacts on the basis of the binomial model and comparing the probability of fragment coverage in the foreground and background windows. However, these two tools do not provide any comparative data analysis.

Generalizations

As in any problem of designing a methodology of statistical analysis, there is a question if the proposed method can be generalized to a sufficiently large class of experimental situations arising in the particular area of research. We demonstrated the use of the proposed methods on the examples of data sets which are, in some way, minimal in fulfilling the required conditions: both concern factors with two levels, and both were obtained in three biological replications. By searching in the literature and in the data bases of public NGS data, we found the following situations which could be discussed here.

Firstly, there are experiments in which the number of levels of the experimental factor is bigger than two. Actually, the data set that we used here, described by Denholtz et. al (2013), is a fragment of a bigger one in which a range of mouse cell lines was used (ESCs, reprogrammed iPSCs, partially reprogrammed pre-iPSCs, and differentiated MEFs). Another example is the data set obtained by Noordermeer et al. (2011 b) (Arrayexpress, E-GEOD-55342) in which Hox gene clusters were studied in various cell lines (ESCs, forebrain cells, pre-somitic mesoderm cells). A generalization of our methodology in the part that uses LMM is straightforward if we apply general principles of modelling and hypothesis testing in linear models. In the first step, in each instance of sliding window, we would test the general hypothesis saying that the mean coverage is equal under all levels of the studied factor. Then, we could apply the sliding windows procedure to find regions in which a particular comparison (contrast) between factor levels is significant. The same approach could be used for binarized data. There exist a generalization of the Fisher exact test of independence (or homogeneity) which works for the case of $r \times c$ contingency tables (Cochran, 1954). It is computationally intensive for large r or c. We would like to apply this test to an $r \times 2$ table, so only one of the dimension could be (moderately) larger and the computations would be feasible. In the situation, where the number of factor levels is very large, a numerical method proposed by Mehta and Patel (1983) could be applied. After applying this procedure to test the general hypothesis, chosen individual 2×2 tables can be studied to test differences between levels of the factor. We also note that in addition to the exact test also other approaches have been proposed to analysis of contingency tables with small counts, one of them being the method based on permutations implemented, e.g., in R in *chi.perm* function (Beh and Lombardo, 2014).

Secondly, there are experiments in which 4C assay is performed with different genomic loci used as the bait. For example, Matthews and Waxman (2020) (GEO GSE130911) studied 3D genome organization in the context of sex-associated differences in gene expression in mouse by performing 4C-seq using as baits six loci identified previously as enhancers of sex-biased genes. In analysis of data coming from such experiments the challenge would be to work with different levels of coverage for different variants of the factor "bait", because a large proportion of NGS reads is mapped in the bait region. This should not have a big effect on the analysis by LMM model, as the coverage from all samples is transformed to values around zero. Also, the Fisher test based on binarized data should function properly, rejecting the hypothesis of the same frequency of coverage in the regions where the difference in the frequency of restriction fragments covered for two different baits is large.

We note that the generalization of our methods is also possible for multi-factor experiments (possibly with different number of biological replications) which can be found in the literature. For example, Gehrke et al. (2014) (ArrayExpress E-GEOD-61063) performed a two-factor 4C-seq experiment for three Hox genes in three developmental stages of zebra fish. Ghavi-Helm et al. (2014) performed a three-factor 4C-seq experiment using two types of *Drosphila* biological material (nuclei from transgenic embryos, whole embryos) collected at two developmental stages and targeting 92 genomic regions (enhancers) as baits. By using LMM, the effects of all experimental factors and of their interaction can be studied as fixed

effects in the model. To do the Fisher test, combinations of levels of experimental factors could be used to construct $r \times 2$ contingency tables.

One should note that the 4C-seq data analysis pipelines that use *Deseq2* for comparative analysis can also be applied to multi-factor experiments. However, the general practice of using *Deseq2*, suggested in its manual, is to construct a model involving one factor with levels being combinations of all factors used in the experiment, and analyse contrasts between selected combinations. Testing of general hypotheses concerning mean effects of factors and their interactions is not done.

Sliding windows procedure

Application of the sliding windows procedure is common to many methods designed to analyse genomic signals. Of the tools for 4C-seq data analysis, it is used by *w4cseq*. This procedure allows to group restriction fragments into bigger units over which measures of coverage characterizing genomic regions could be estimated. We performed computations using different sets of sliding windows parameters, size and step, for both LMM and Fisher test method. In both cases we found that intermediate window sizes were providing the biggest numbers of significant results. We can explain this as follows. For small window length, coverage accumulated over small number of restriction fragments does not provide many significant differences between variants. For too large window lengths, the number of significant differences decreases because regions are merged into large units containing many potential contacts - too large to provide interpretation. Thus, windows of intermediate size are proper for carrying the analysis. In our analyses, the optimum window size, providing the biggest numbers of DCWs and DCRs, was 200000 bp. Taking into account *Arabidopsis* and mouse genome sizes (135 Mb, 2700 Mb, respectively) and number of primary restriction fragments in those genomes (43663 and 821717, respectively, for the applied primary enzymes), we can compute the average length of a fragment as being approximately 3100-3300 bp in studied genomes. Thus, we can conclude that the optimum window size expressed in terms of the number of restriction fragments was in both cases about 65 fragments/window. On the other hand, if the experiment is performed to locate protein coding genes in the regions contacting the bait, this means a very different resolution for the two species that were studied: about 40 genes/window for *Arabidopsis* and about 1.6 gene/window for mouse. This suggests that maybe suboptimal in terms of the number of DCWs, but providing better resolution in terms of candidate genes in DCWs, window sizes should be used in relatively small genomes.

Signal intensity vs contact frequency

In the analysis of two exemplary data sets we found that the results obtained for ranked and normalized coverage with the use of LMM and the results obtained for binarized data with the use of Fisher test do not always coincide in terms of the location of DCRs and the direction of the differences between variants. In general, there was no correlation between effects of experimental variants estimated by two methods and measures of significance computed for

the same genomic windows. By definition, the two methods measure different properties of the windows. In LMM, the transformed intensity of contact signal is used; this method is more appropriate for finding contacts that involve genomic fragments not necessarily long, but frequent in the pool of cells that were taken for NGS library preparation. In Fisher test, the intensity is transformed to binary values representing the frequency of covered fragments in a window; thus, this method is more appropriate for locating contacts not necessarily frequent in the pool of cells, but relatively wide and continuous (stable over an interval). In our summaries we put more attention to the DCRs that were obtained as intersection of the results of two methods. In practice, the choice of DCRs for interpretation can be different. However, the results obtained for two data sets suggest that the requirement of our analysis, i.e., characterization of genomic contacts in terms of various parameters, has been met.

Comparison with other methods

Out of methods of 4C-seq data analysis that were mentioned in Sec. 1.3, we used *foursig* (Williams et al., 2014) to analyse the two data sets and interpret comparisons with LMM and Fisher test results. *Foursig* uses a quantitative signal measure, operates on the whole genome, so its results can be used for a comparison, and it was found as a method relatively efficient in finding significant contacts (Zisis et al. 2020). We performed the comparison by intersecting the contacts found by *foursig* with DCRs found using LMM and Fisher test. In the analysis of *Arabidopsis* data by LMM, we found that most of identified DCRs intersected with *foursig* specific or common significant contacts; in the analysis by Fisher test, more *foursig* contacts specific to A or B than *fourSig* common contacts were confirmed by DCRs. Thus, we can say that the analysis by methods proposed in this work was able to find a subset of regions that were found as being in contact with the bait by *fourseq.* The situation was slightly different in the analysis of mouse data by LMM, where a majority of DCRs found for shorter sliding windows intersected with, mostly, common *fourSig* contacts, whereas in Fisher test very few DCRs contained *fourSig* contacts. Thus, it is possible that the differences between results obtained by the two methods depend on the genome coverage.

P value adjustment

In the construction of the method by joining several elements of statistical methodology, we encountered some issues. One of them was the inability to use the proposed method of *P* values correction in the analysis of mouse data by Fisher test. As we noted, this was caused by the fact that the sequencing depth was not sufficient for the size of the genome. With a big proportion of uncorrected *P* values being close to one, and even a bigger proportion being so after correction, no DCWs and, consequently, DCRs were declared.

A proposed strategy in this situation would be to find significant regions by using LMM model approach and then apply the Fisher test only within those regions. For such testing, no correction, or a very liberal correction, for simultaneous testing could be done, as within each LMM DCR only a few Fisher tests would be done. In this way our genome-wide procedure

will be based on quantitative signal analyzed by LMM, and the set of identified DCRs would be verified for coverage frequency using binarized data in a local way.

Lessons from two different data sets

We used the constructed method to analyse two data sets obtained for two different species representing different kingdoms – of plants and animals. This was possible because, as it happens now in most of the areas of molecular biology methodology, there are great similarities in protocols that are used, also in data processing and analysis protocols. To illustrate our approach a careful selection of data sets was done, in order to assure a minimum standard of biological replications and an acceptable number of NGS reads in each replication. Also, we did the selection of the data sets for the analysis taking into account genome sizes. The ArrayExpress database contains about 60 data accessions that can be found using the keyword "4C-seq". Among them, the most frequently represented species are: *Mus musculus* (35 accessions), *Homo sapiens* (14), and *Danio rerio*, *Drosphila melangaster* and *Gallus gallus* (each 2). The *Arabidopsis* data set that we used is the only one representing plants. From our experience, the methods that we presented should work for data obtained in species with smaller genome sizes like *D. melanogaster*, if the number of NGS reads is comparable to that used here (10-15 mln reads). However, there exist data sets obtained in *H. sapiens* with 2-3 mln reads per sample; taking into account the size of human genome, this would be definitely not enough if a genome-wide library preparation was done and a corresponding inference is attempted. Such data could be probably only used for the analysis of close contacts. However, with the lowering costs of sequencing, it should be possible to obtain sufficient coverage of NGS reads to properly analyse also distant genomic regions. Our *4CseqR* method does not discriminate between close and distant contacts, so in principle could be used to also study interactions with distant genomic regions – unlike some of the existing methods compared by Zisis D et al. (2020).

We also showed a simple example of a downstream analysis that can follow the statistical part in order to obtain insight into biological phenomena. We do not attempt to make an in-depth interpretation of the reported GO annotation of genes found in DCRs – this is the task reserved to the experiment performers. However, we can note that the specificity of gene functions was more pronounced in the mouse experiment, where enrichment of all three categories of GO terms was found. This could be due to the fact that in this case very specific biological meterial was used – different cell lines, whereas in case of *Arabidosis* a broad selection of material treated with different environmental conditions was used.

6. Conclusions

1. We presented a new method of 4C-seq data analysis and the corresponding computational pipeline *4CseqR* which were constructed with the aim of overcoming shortages of existing methods and extend the usage of well-founded statistical methods in the area of NGS data analysis and, especially, in 4C-seq applications.

2. The presented method is not concentrated on the detection of significant contacts within samples, but provides a full comparative analysis of experimental variants in factorial experiments based on a new approach which is combining two different mathematical models, linear mixed models and analysis of contingency tables by Fisher test.

3. The method describes the significance of contacts in a dual way: by continuous contact signal intensity and by binarized description of presence or absence of contacts.

4. The method does not discriminate between close and distant contacts (*trans* or *cis*), so could be used to also study interactions with distant genomic regions**.**

5. An important feature of the method is that it is not using just raw mapped reads, but data normalized with respect to possible sequence-related biases (by *Salmon*) and restriction fragment-related biases (by rank-based INT), as input to a statistical model.

6. Application of the method to two data sets contrasting in terms of reference genome size allowed to show different features of the proposed algorithms.

References

1. Roberts A, Pachter L (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10(1): 71-73.
2. Anders S, Huber W (2010). Differential expression analysis for sequence count data. Genome Biol 11(10): R106.
3. Andrews S (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Bantignies F, Roure V, Comet I et al. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell 144: 214-26.
5. Beh EJ, Lombardo R (2014). Correspondence Analysis: Theory, Practice and New Strategies. Chichester, Wiley.
6. Benjamini Y and Hochberg Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B 57: 289-300.
7. Blom G (1958). Statistical estimates and transformed beta-variables. Wiley; New York
8. Bray NL, Pimentel H, Melsted P et al. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5): 525-7
9. Cai M, Gao F, Lu W et al. (2016). w4CSeq: software and web application to analyze 4C-seq data. Bioinformatics 32(21): 3333–3335.
10. Zhang C, Zhang B, Lin LL, Zhao S (2017). Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 18: 583.
11. Cochran WG (1954). Some Methods for Strengthening the Common $\chi 2$ Tests. Biometrics 10, 417-451.
12. Cremer T and Cremer M (2010). Chromosome territories. Cold Spring Harb Perspect Biol 2: a003889.
13. Davies JO, Oudelaar AM, Higgs DR et al. (2017). How best to identify chromosomal interactions: a comparison of approaches. Nature Methods 14(2): 125-134.
14. De Wit E, Bouwman BA, Zhu Y et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature 501(7466): 227-31.
15. Denholtz M, Bonora G, Chronis C et al. (2013). Long-range chromatin contacts in embryonic stem cells reveal a role for pluripotency factors and Polycomb proteins in genome organization Cell Stem Cell 13(5): 602-616.
16. Denker A and de Laat W (2016). The second decade of 3C technologies: detailed insights into nuclear organization. Genes Dev 30(12): 1357-1382.
17. Dixon JR, Selvaraj S, Yue F et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485 (7398): 376–380.
18. Dostie J, Richmond TA, Arnaout RA et al. (2006).Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16(10): 1299-309.
19. Dougherty ER, Huang Y, Kim S et al. (2009). Genomic Signal Processing. Curr Genomics 10(6): 364.
20. Brettmann EA, Oh IY, de Guzman Strong C (2018). High-throughput Identification of Gene Regulatory Sequences Using Next-generation Sequencing of Circular Chromosome Conformation Capture (4C-seq). J Vis Exp (140): 58030.
21. Fisher RA (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. Journal of the Royal Statistical Society 85 (1): 87–94.
22. Fullwood MJ, Liu MH, YF Pan et al. (2009).An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462(7269): 58-64.
23. Galecki A and Burzykowski T (2013). Linear Mixed-Effects Models Using R. Springer.

24. Gehrke AR, Schneider I, de la Calle-Mustienes E, et al. (2015). Deep conservation of wrist and digit enhancers in fish. Proceedings of the National Academy of Sciences of the United States of America 112(3): 803-808.
25. Ghavi-Helm Y, Klein FA, Pakozdi T et al. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature 512(7512): 96-100.
26. Grob S, Schmid MW, Luedtke NW, et al. (2013). Characterization of chromosomal architecture in Arabidopsis by chromosome conformation capture. Genome Biol 14(11): R129.
27. Thorvaldsdóttir H, Robinson JT, Mesirov J (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14(2): 178-192.
28. Hövel I, Louwers M, Stam M (2012). 3C technologies in plants. Methods 58: 204-211.
29. Hövel I (2016). Novel insights into gene silencing mechanisms in Zea mays and Arabidopsis thaliana, University of Amsterdam, Swammerdam Institute for Life Sciences (SILS). https://hdl.handle.net/11245/1.535446
30. Klein FA, Pakozdi T, Anders S et al. (2015). FourCSeq: Analysis of 4C sequencing data. Bioinformatics 31(19): 3085-3091.
31. Langmead B, Salzberg S (2012). Fast gapped-read alignment with Bowtie2. Nature Methods 9: 357-359.
32. Li H, Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
33. Lieberman-Aiden E, van Berkum NL, Williams L et al. (2009). Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science 326(5950): 289-293.
34. Loviglio MN, Leleu M, Männik K et al. (2017). Chromosomal contacts connect loci associated with autism, BMI and head circumference phenotypes. Mol Psychiatry 22(6): 836-849.
35. Lupianez DG, Kraft K, Heinrich V et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161: 1012-25.
36. Mehta C and Patel N (1983). A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables. Journal of the American Statistical Association 78(382), 427-434.
37. Noordermeer D, de Wit E, Klous P et al. (2011 a). Variegated gene expression caused by cell-specific long-range DNA interactions. Nat Cell Biol 13(8): 944-51.
38. Noordermeer D, Leleu M, Splinter E et al. (2011 b). The dynamic architecture of Hox gene clusters. Science 334(6053): 222-5.
39. Patro R, Duggal G, Love MI et al. (2017). Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 14(4): 417–419.
40. Qiu X, Wu H and Hu, R (2013). The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis. BMC Bioinformatics 14:124.
41. Raab JR, Chiu J, Zhu J et al. (2012). Human tRNA genes function as chromatin insulators. EMBO J 31(2): 330-50.
42. Raviram R, Rocha PP, Bonneau R et al. (2014). Interpreting 4C-Seq data: how far can we go. Epigenomics 6(5): 455-457.
43. Raviram R, Rocha PP, Müller CL et al. (2016). 4C-ker: A method to reproducibly identify genome-wide interactions captured by 4C-Seq experiments. PLoS Comput Biol 12(3): e1004780.
44. Ritchie ME, Phipson B, Wu D et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.

45. Roberts A, Pachter L (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 10(1): 71-73.
46. Rocha PP, Micsinai M, Kim JR, et al. (2012). Close proximity to Igh is a contributing factor to AID-mediated translocations. Mol Cell 47(6): 873-85.
47. Sandhu KS, Shi C, Sjölinder M et al. (2009). Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. Genes & Dev 23: 2598-2603.
48. Schoenfelder S, Sexton T, Chakalowa L et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nature Genetics 42: 53–61.
49. Searle SR, Casella G, McCulloch CE (2006). Variance Components. John Wiley & Sons.
50. Shrestha S, Oh DH et al. (2018). 4C-seq characterization of Drosophila BEAF binding regions provides evidence for highly variable long-distance interactions between active chromatin. PLoS One 13(9): e0203843
51. Simonis M, Klous P, Splinter E et al. (2006). Nuclear organization of active and inactive chromatin domains uncovered by Chromosome Conformation Capture-on-Chip (4C). Nature Genetics 38: 1348-1354.
52. Stadhouders R, Kolovos P, Brouwer R et al. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc 8: 509-524.
53. Beasley TM and Erickson S (2009). Rank-based inverse normal transformations are increasingly used, but are they merited. Behav Genet 39(5): 580-595.
54. TA Potapova, JR Unruh, Z Yu et al. (2019). Superresolution microscopy reveals linkages between ribosomal DNA on heterologous chromosomes. J Cell Biol 218(8): 2492-2513.
55. Thongjuea S, Stadhouders R, Grosveld FG et al. (2013). r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic Acids Res 41(13): e132.
56. Tolhuis B, Blom M, Kerkhoven RM et al. (2011). Interactions among polycomb domains are guided by chromosome architecture. PLoS Genet 7(3): e1001343.
57. van de Werken HJ, de Vree PJ, Splinter E et al. (2012 a). 4C technology: protocols and data analysis. Methods Enzymol 513: 89-112.
58. van de Werken HJ, Landan G, Holwerda SJ et al. (2012 b). Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods 9(10): 969-972.
59. van der Waerden BL (1952). Order tests for the two-sample problem and their power. Proc Koninklijke Nederlandse Akademie van Wetenschappen. 55:453-458.
60. Vicente-García C, Villarejo-Balcells B, Irastorza-Azcárate I et al. (2017). Regulatory landscape fusion in rhabdomyosarcoma through interactions between the PAX3 promoter and FOXO1 regulatory elements. Genome Biol 18(1): 106.
61. Wanichthanarak K, Jeamsripong S, Pornputtapong N et al. (2019). Accounting for biological variation with linear mixed-effects modelling improves the quality of clinical metabolomics data. Computational and Structural Biotechnology Journal 17: 611-618.
62. Walter C, Schuetzmann D, Rosenbauer F et al. (2014). Basic4Cseq: an R/Bioconductor package for analyzing 4C-seq data. Bioinformatics 30(22): 3268-3269.
63. Williams A, Spilianakis CG, Flavell RA (2010). Interchromosomal association and gene regulation in trans. Trends Genet (4):188-97
64. Williams RL Jr, Starmer J, Mugford JW et al. (2014). foursig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic Acids Res 42(8): e68.

65. Woltering JM, Noordermeer D, Leleu M et al. (2014). Conservation and divergence of regulatory strategies at Hox Loci and the origin of tetrapod digits. PLoS Biol 12(1): e1001773.
66. Zeitz MJ, Ay F, Heidmann JD et al. (2013). Genomic interaction profiles in breast cancer reveal altered chromatin architecture. PLoS One 8(9):e73974.
67. Zhao Z, Tavosidana G, Sjölinder M et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics 38: 1341-1347.
68. Zisis D, Krajewski P, Stam M et al. (2020). Analysis of 4C-seq data: A comparison of methods. Journal of Bioinformatics and Computational Biology 18(1): 2050001.