Université Paul Sabatier

Report MUT3D project

M1 Bio-informatique et biologie de systèmes

Claire DELAMARE-DEBOUTTEVILLE, Youcef BEN MOHAMMED,
Julie CAMPOS Y SANSANO
12/05/2023

Table des matières

1) Context and objectives 2
2) Materials and Methods 5
3) Results 7
4) Conclusion and perspectives 12
5) ANNEXES 13

For more details and results, visit the GitLab page of the Mut3D team: https://gitlab.com/bbs_m1/mut3d

1) Context and objectives

The spatial organisation of the genome is intimately related to numerous key biological functions (including gene expression and DNA replication regulations). As a result, some cancers and developmental disorders might be associated with alterations in the three-dimensional (3D) genome architecture in space and time (Li et al, 2019, ref 24 in annex 0).

This project was realised in the team of Dr. Olivier CUVIER (CBI, Toulouse, https://cbi-toulouse.fr/eng/equipe-cuvier), where they study how chromatin dynamics govern cell fate and cell proliferation. They investigate how specific chromatin modifiers (methylases) and regulators (insulator and co-factors) impact transcription or cell replication through chromatin dynamics. The mechanisms by which molecular drivers functionally organise the 3-Dimensional (3D) conformation of the genome, such as topologically associating domains (TADs), remain to be explored (Raphaël Mourad, Olivier Cuvier 2018, ref 1).

AIM

In this project, we will explore the relation between 3D organisation and transcriptional regulation. To do so, we will compare the genome-wide physical contact maps of 3D-disturbed-mutants (double mutant:hpl2/lin-61 and a triple mutant: set-25/set-32/Met-2) versus the wild type (N2) in C.*elegans* (nemathod, a frequently used laboratory organism in epigenetic and 3D organisation issues). We want to investigate whether the disruption of TADs, a 3D structure, in the mutants have an impact on:

1/ the regulation of transcription of genes,
2/ the contact between enhancers and promoters (de novo enhancer-promoter couples)
3/ the creation of de novo couples that affect the regulation of genes (up or down regulation).

METHODS

By studying the epigenetic modifications and chromatin organisation in these strains, we can gain insights into the mechanisms of gene regulation in C.*elegans* and how they are affected by the mutations. Specifically, we can investigate the distribution of euchromatin and heterochromatin, and the organisation of TADs in the genome of each strain. This information can help us identify the specific genes and pathways that are affected by the mutations and provide a more comprehensive understanding of the epigenetic regulation of gene expression in C.*elegans*.

As a first step, we will analyse DNA sequences coming from all the relaxed parts of the genome (open chromatin, with ATACseq data). Then we will compare those DNA relaxed regions with the cartography of several epigenetic marks (histone modifications, specifically found in transcriptionally active regions, like H3k27ac or in transcriptionally inactive regions like H3k9me3) provided by CHIP-Seq data. We will combine the results from these two experiments and get a list of transcriptionally active/inactive regions found in the relaxed part of the genome. We will create two lists: one for probable enhancer regions, and one for probable promoter regions (1 kbp around the beginning of gene sequences). As soon as we have those two lists, we will combine them and create a list of active enhancer-promoter couples.

In the next step, we will compare those couples with a physical contact map of DNA sequences (HIC data), informing us about DNA region interactions. We will look at different scales (inside the whole chromosome, inside TADs region: covering 40 to 300kbp distance) and find out if the apparition of these de-novo couples are linked with the disruption of chromatin 3D structures (TADs), that play a role of insulating genomic regions favouring contacts inside these structures.

In a final step, we will compare these results with gene expression levels (RNAseq data). This will indicate if this de novo contacts (bringing some enhancers aberrantly closer to genes) in the mutants lead to an up or down regulation of the involved gene.

CONTEXT

In higher eukaryotic cells, chromosomes are folded inside the nucleus. In humans for example, the genome would measure 2 metres linear. To fit this 2 metres genome length inside a 10 µm nucleus, the genome (DNA) is compacted with proteins through several layers of compaction. This compact form is called the chromatin (DNA and proteins). Based on microscopy observations, we can distinguish two forms of chromatin; euchromatin (for the lower compact forms) and heterochromatin (the most compact form). It appears as dark areas inside the nucleus). For example, we can see this in human blood cells:

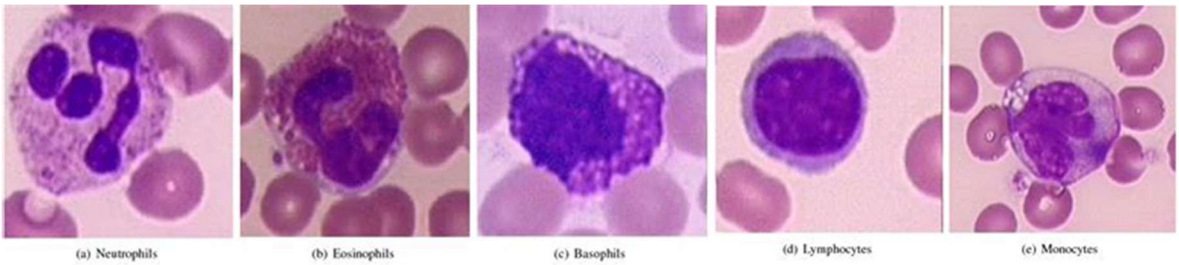


Figure 1. Optical microscopy of Human white blood cells. DNA is coloured in dark purple (nucleus).

Higher-order chromatin organisations vary among cells, tissues, and species depending on the developmental stage and/or environmental conditions (figure 1). Chromosomal interactions and topological changes in response to the developmental and/or environmental stimuli affect gene expression.

3D ORGANISATION OF THE GENOME

The fundamental units of genome organisation are the nucleosomes. These are grouped into fibres, loops, domains, and compartments, as well as chromosomes. These features organise genomes at multiple levels and length scales:

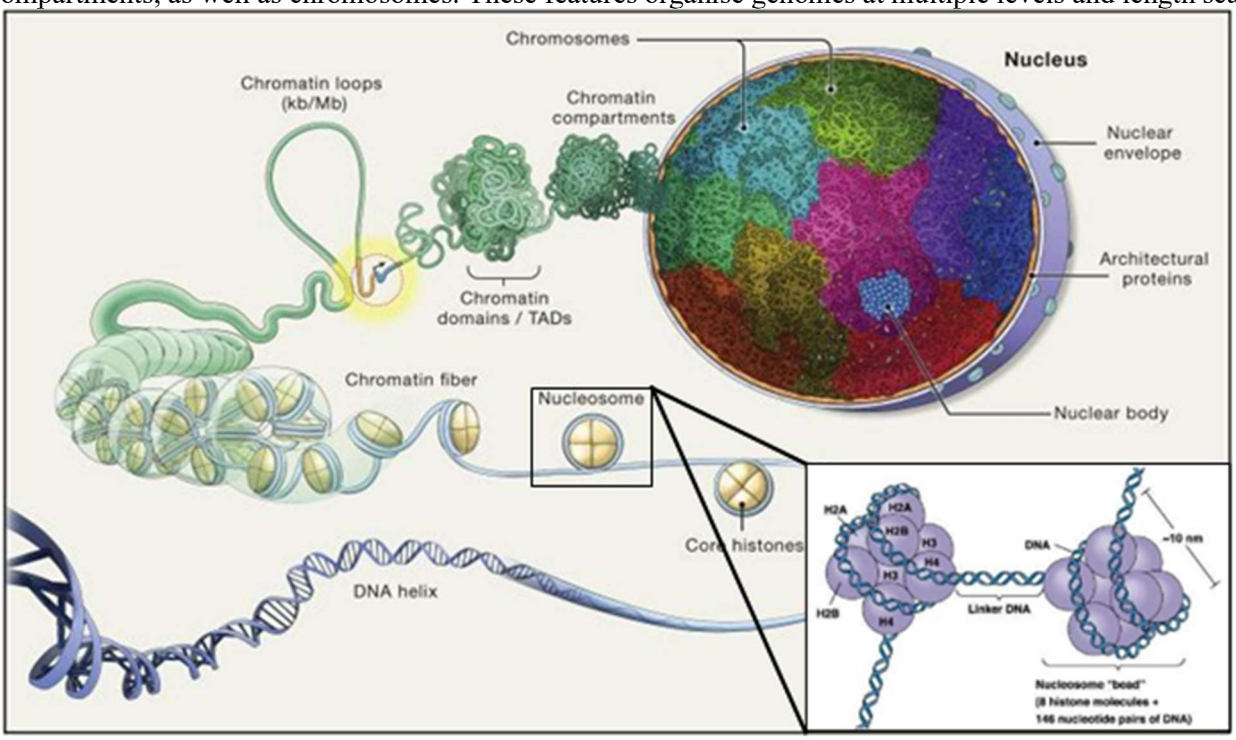


Figure 2. **Schematic representation of 3D ORGANISATION OF GENOME in eukaryotes** (10. Review Cell, Misteli, 2020). **Detailed structure of two nucleosomes** (Addison Wesley Longman, 1999) .

Genomes are as such organised at multiple levels. At the lowest layer DNA is wrapped around the nucleosome, which is composed of an octamer of core-histones. Nucleosomes are the structural basis of chromatin (figure 2). In the nucleus chromatin fibres can form loops, often bringing upstream gene regulatory elements (yellow), such as enhancers, near to the promoters of genes (gold/blue) to control their transcription (black arrow). These chromatin loops, that represent physical proximity between distantly located genomic regions, are highly frequent inside larger genomic domains called TADs. Activation of gene transcription and enhancers seem to be performed at the level of TADs (Dixon et al 2015, ref 12, Delaneau et al, Science 2019, ref 25), and TADs can be grouped into active and inactive compartments. These active TADs are generally regrouped at the centre of the nucleus forming the active

compartment, while inactive TADs are found principally at the periphery of the nuclear membrane and near the nucleolus (11). The DNA of each chromosome occupies a distinct volume, or chromosome territory (multiple colours), within the cell nucleus, generating non-random patterns of chromosomes and genes. In the DNA-free space, the nucleus also contains RNA and proteinaceous protein aggregates which form nuclear bodies (blue) (10. Review Cell, Misteli, 2020).

TADs

With the development of new techniques, such as HI-C (high-throughput chromatin conformation capture), a more detailed map of the genome organisation was obtained. HI-C was first described in 2009 by Lieberman-Aiden et al.(11) and has made it possible to highlight new hierarchical domains, such as the TADs (figure 2, 3). In 2012, several publications (Dixon et al,(12) Sexton et al (13) and Nora et al(14)) described chromatin domains (TADs) and demonstrated higher frequencies of contact with each other than contact with genomic regions outside of the domain.

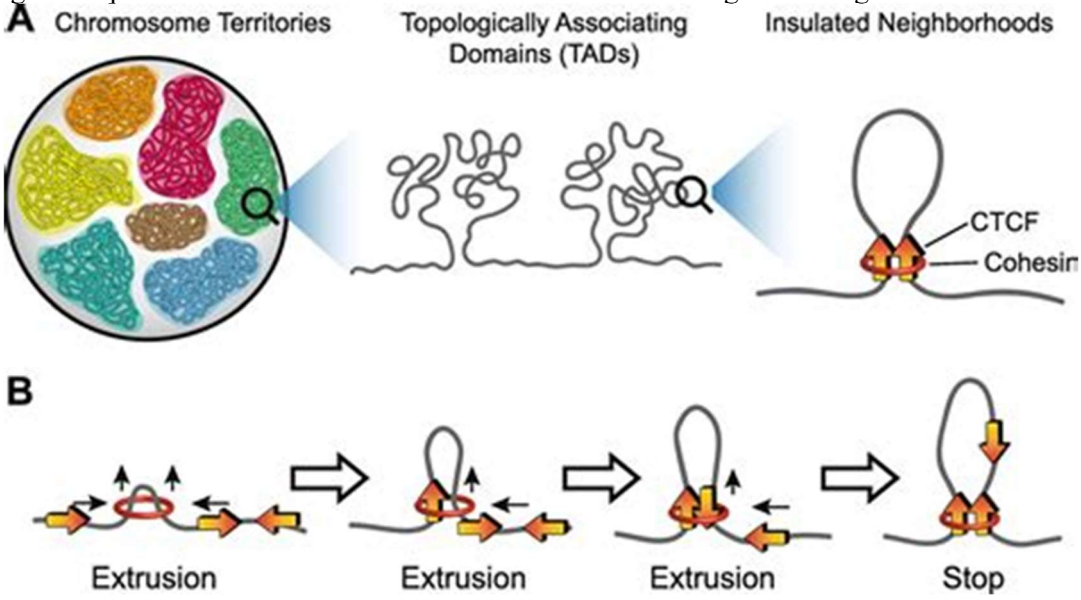


Figure 3. Schematic representation of TADs and their dynamics with insulators (Fu et al, 2018, review, (19))

These domains have been shown to play a critical role in gene regulation, as they can act as boundaries that prevent the spread of epigenetic marks and regulate the accessibility of genes to the transcriptional machinery. TADs are highly dynamic structures(19). TADs have been reported to be relatively constant between different cell types (in stem cells and blood cells, for example), and even across genomes in specific cases (ref 12,16,17,18).

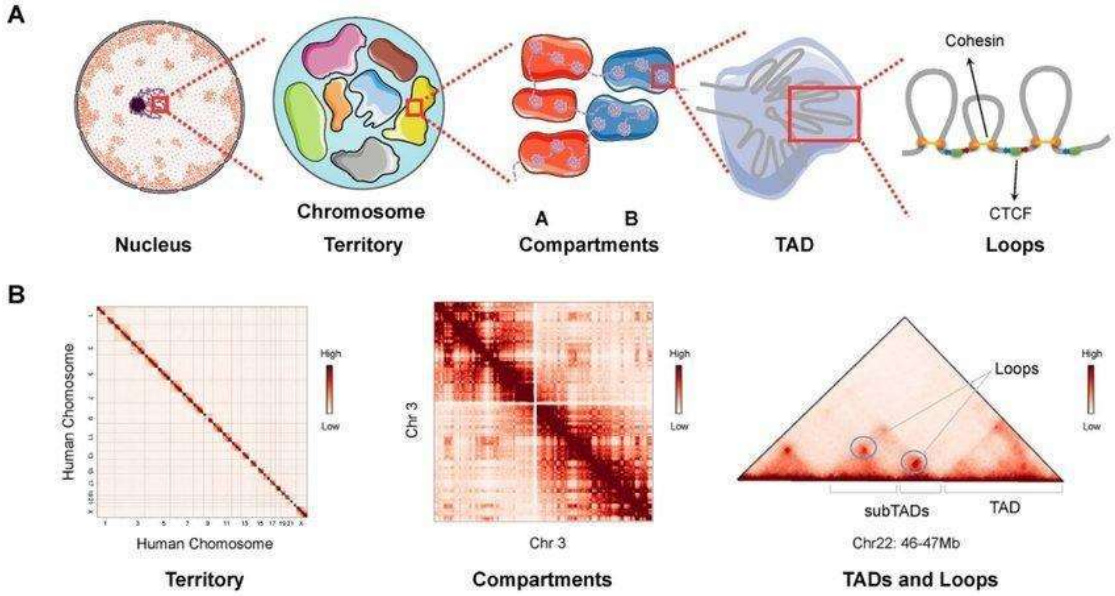


Figure 4. Hierarchical genome organisation. (A) Multilevel 3D genome organisations. Chromosome territory, compartments, TADs and loops can be observed from left to right. Each chromosome territory is denoted by different colours. Compartment A and B are indicated in red and blue background, respectively. (Hao et al 2021, BB).

HISTONES MODIFICATIONS

Epigenetics is the study of mechanisms that regulate transcription without altering DNA sequences. This includes the post-transcriptional modifications occurring on the proteins linked to DNA, particularly histone proteins. As mentioned earlier, nucleosomes correspond to a 147 long DNA wrapped around an octamer of histone proteins. The modifications of histones composing the nucleosomes can lead to a change of the chromatin compaction and the accessibility of the underlying DNA sequence to the transcriptional machinery. Specific patterns of histone modifications are associated to transcriptionally silenced or active parts of the genome. Frequently, methylation is related with compaction (inactivation) and acetylation with relaxation of the chromatin (ref21: Kim et al 2017,EMM). For example, linked with transcriptionally silent genomic regions, repressive histone marks **H3K9me or H3k27me3** can be found abundantly enriched within the heterochromatin **(**Mourad et al 2018, Padeken et al 2019, Woodhouse et al 2018, Nicetto et al 2020 (ref 1,2,3, 23**)).** H3K27me3 is a facultative heterochromatin mark, genes marked in their gene body are transcriptionally inactive generally speaking. And some others are related with transcriptionally active areas, like **H3K27ac** or **H3K4me.** These histone modifications were found to be associated with active promoters: **H3K9ac or H3K4me3 (ref 22.** Igolkina et al 2019, Cells**).**

PROTEINS OF INTERESTS:

Triple mutant set32 set25 MET2:

MET-2, SET-25 and SET-32 are 3 histone methyltransferases. They add a methyl group (METH) on histones. In Particular, they catalyse the methylation on lysine 9 of histone H3(H3k9me)(ref 2: Padeken et al 2019). Mutations of these 3 proteins are associated with alteration of heterochromatin compaction. It was shown that the triple mutant presents a low level H3K9me (compared to the WT) and a down-regulation of proteins involved in ovary development (ref 2,3,4,5,6). This showed a link between, H3K9me, heterochromatin formation and transcription regulation.

Double mutant hpl2 lin61

LIN-61 plays a role in the regulation of cell division and differentiation during development(ref 7: Padeken et al 2021). Specifically, LIN-61 is a member of the LIN-15A/B and LIN-35 protein complex, which is involved in the regulation of the G1 phase of the cell cycle. Predicted to enable chromatin binding activity and histone binding activity(Wormbase, NCBI, Uniprot).

Hpl2 enables H3K27me3 modified histone binding activity and chromatin binding activity. Involved in several processes, including developmental processes involved in reproduction; nematode larval development; and regulation of gene expression. Located in the nuclear periphery (Wormbase, NCBI, Uniprot).

These double and triple mutants seem to disrupt TAD structures (data not published). In this study, we wanted to confirm that the removal of TADs induce de novo interactions.

2) Materials and Methods

To evaluate the effects of the double and triple mutations on genome organisation and gene expression, data from ATAC-seq, HiC, ChiP-seq and RNA-seq were provided for our analysis (cf Annex 1 to see the description of each method we used).

Analysis of ATAC-seq data

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) is a genome analysis method that allows the identification of regions of DNA accessible for transcription. In order to explore the data from ATAC-seq, we use the following pipeline (cf Annex 2: linux/bash commands):

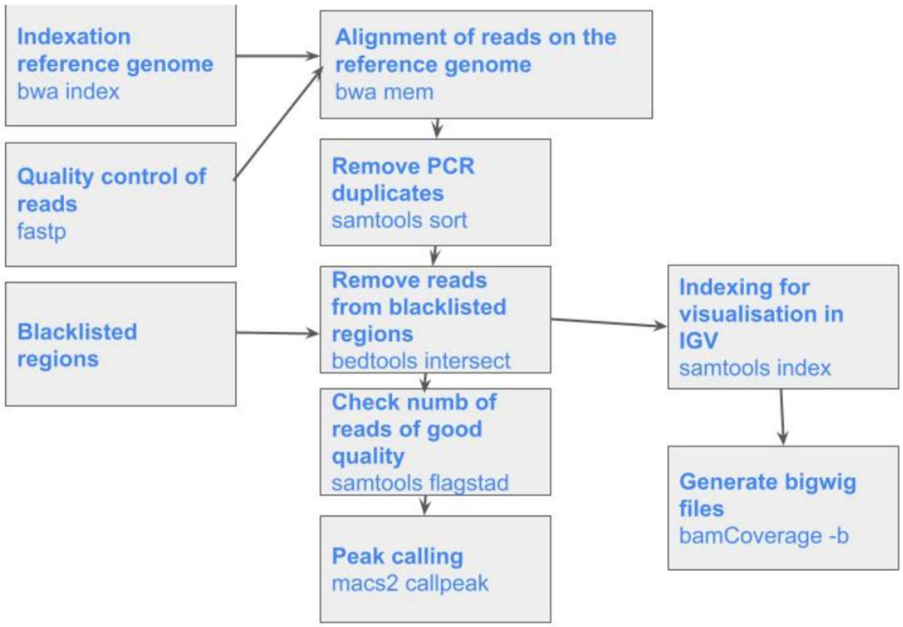


Figure 5. Pipeline ATAC seq for pre-processing raw data.
The peaks obtained from a peak calling analysis represent regions of the genome that are accessible to the transposase enzyme used in the ATAC-seq protocol, and thus indicating regions of open chromatin and potentially regulatory regions. The workflow we used for our samples: N2, hlp2/lin-61 and set- 25-32/met-2 mutants, is detailed in annex 3.

We used raw data files (3 samples: N2, hlp2/lin-61 and set- 25-32/met-2 mutants, format .fastq) obtained from ATAC sequencing from **Gene2i**. We used the reference genome C.elegans: Caenorhabditis_elegans.WBcel235.99 (or ce11).

Hi-C data and map creation with HicAggR package

HiC is a sequencing-based method that allows for the study of interactions between different regions of the genome (cf the introduction section and annex 1). This technique enables the mapping of physical contacts between DNA fragments that are in close proximity in the three-dimensional space of the cell nucleus. We will use a R library upon development in Prof. Cuvier's Laboratory, the HicAggR package.

The HicAggR package in R provides functions to aggregate the Hi-C data into larger genomic bins of a user-defined size, which can help reduce the noise in the data and increase the resolution of the analysis. The package also includes several visualisation tools to help users explore the aggregated data, including heatmaps and distance plots.

The main purpose of the HicAggR package is to preprocess the Hi-C data and perform genome-wide interaction studies between specific features of the genome. It allows to aggregate signals of chromatin contacts between for example enhancers and genes found within their surroundings. The aggregation process would assess if chromatin contacts between enhancers and genes are globally present across the genome or if the contacts are sporadic and restricted to particular sites of enhancer-gene couples. Overall, the package can facilitate the analysis of large-scale Hi-C datasets and help researchers better understand the organisation and mechanism of action of certain features like enhancers. In order to explore the data from HiC, we will use the following HicAggR pipeline (cf Annex 4).

The Hi-C data we used for the study were already cleaned and normalised (format .bed, .cool, by Gen2i).

Other datas

We used other pre-cleaned data (from Gen2i) : CHIP-seq (format .peaks, narrow and broad) and RNAseq (format .csv).

3) Results

ATACseq

Raw data pre-processing

Before we start the treatment of the data, we need to check the quality of the reads. From the raw data, (format .fastq) we generate HTML files which summarise the quality results (cf gitlab: https://gitlab.com/bbs_m1/mut3d).

	Before filtering			After filtering			Filtering result	CONCLUSION	
	total reads	Q20	Q30	total reads	Q20	Q30	reads passed filtered	Observations	Quality
N2	67.54M	92%	86%	67.17M	97%	92%	99%	good amout of reads, with high % of high qualitty	GOOD
htl2 lin-61	67.56	94%	87%	67.11%	97%	91%	99%	good amout of reads, with high % of high qualitty	GOOD
set 25 set 32 met-2	49,9M	92%	86%	49,7M	97%	92%	99%	good amout of reads, with high % of high qualitty	GOOD

Figure 6. CQ Summarise with fastp. version: fastp 0.23.2, (cf annex 5)

The HTMLs generated by fastp, about 97% of our data present less than an error every 100 bp(q20) , and more than 90% of q30 (1 error per 1000 bp). These results indicate **the good quality of our data**.
After that, we followed the pipeline of ATAC_seq (cf annex 4), and obtained a curated list of peaks (potential enhancer regions, in the relax chromatin part) with macs2:

STEP 2: FILTRATION							
		Q20	DUPLICATES		PEACK CALLING		
			total duplicated	total reads	p value = 0,05	p value = 0,02	p value = 0,01
WT	N2	1,56 M	253,2 k	1 562 877	80	6001	9094
DOUBLE MUTANT	htl2 lin-61	45,15 M	11.2M	33.93M	807	-	12277
TRIPLE MUTANT	set 25 set 32 met-2	43,3 M	10,56 M	36,9 M	8926	-	26770

Figure 7. Result summary after filtering steps (output: bed files).
We test different adjusted p values (from 0.05 to 0.01) for filtering significant peaks. The best results were obtained with an adjusted p-value of 0.01 (less stringent). With the number and the good quality of the reads, we expected more than 20k peaks after filtering.

Visualising ATAC seq peaks

Now we obtain a list of peaks for each sample, corresponding to DNA sequences in the open part of the chromatin (transcriptionally active part). Let's compare the samples together and check whether there are specific peaks in double or triple mutant. With the IGV tool, we check several regions among the C.*elegans* genome (ce11, annex 6):

Figure 8. IGV window, focus at a region of 5161 bp inside the chromosome X (7.7 M bp position)**. Annex 6.**

In green, we highlight regions (in green, figure 8) where ATACseq peaks were exclusively found in mutants (h2pl2/lin-61 and set-25/set-32/Met-2). This means that new relaxed chromatin areas were created in the double and triple mutants. In addition, we found some regions with more peaks in the double mutant (in orange, figure9) and less peaks than the wild type sample (N2). Those regions might correspond to new relaxed chromatin areas within the lin-61/hpl2 mutant and a slight compaction inside the triple mutant (set-25/set-32/Met-2).

These results look promising. The next step would be to look at the new contacts created from these new areas in the mutants (creation of contact maps with the R package HicAGGr (APA maps). We want to see if the disruption of TADs region induces new contacts between enhancer and promoter, and induces their transcription.
We will first create lists of (de novo) couple enhancers (intergenic region)/promotor (around 500 bp from the beginning of the genes) for each condition.

SELECTION OF TRANSCRIPTIONALLY ACTIVE REGIONS (R, cf RMarkdown)

From the narrow peaks files, we first transform them into GenomicRanges objects (with R package rtracklayer). We want to extract regions of interest, such as active peaks, active TSS, and couples of active enhancers-promoters. We will first combine the ATAc seq peak lists, with CHIP-seq data (for definition see annex 1). CHIP-seq capture DNA regions presenting defined histone marks, such as:

- **Histone modification specific to transcriptionally inactive regions:** H3K27me3
- **Histone modification specific to transcriptionally active regions:**

⇒ **For intergenic regions:** we choose the double histone marks **(H3K27ac/H3K4me).** This should include active enhancer regions.
⇒ **For promoter regions:** we choose to select DNA sequences presenting the histone modification **H3K27ac, combined with RNApol II sites**. The combination of the two should define active promoter (active TSS in the figure 9) regions. Additionally, we restrict DNA sequences to 1 kbp around the beginning of the genes (500bp before and 500bp after TSS, for Transcriptionally Starting Site). The results are summarised in the table below:

	Wild-Type	Double-Mutant	Triple-Mutant
ATAC seq peaks alone	9094	12277	26770
Active Peaks	3263	2491	5535
Active TSS	1060	775	1593

Now we select all ATAC seq peaks presenting either intergenic (potentially including enhancers) or promoters transcriptionally active histone specific modifications. We will use the **library HicAggR** (development on going) for the next steps.

CREATION OF LISTS OF ENHANCER/PROMOTER COUPLES (R, cf RMarkdown)

We want to create lists of enhancer/promoter couples for each condition by combining the two selected regions (result of search pairs with HicAggr library). First, to extract and plot the matrices of couples enhancers-promoters with the library HicAggr, we have set the minimum and the maximum distances between an enhancer and its promoter, to 40 kbp as minimum, and 300 kbp as maximum (neighbourhood TADs association), and the sizes of matrices in output to default (21x21)(Annex 4).

CONTACT MAPS - APA plot (R, cf RMarkdown, Hic AGGR)

We want to compare the interaction of enhancers/promoters inside TADs (genomic constraints, with 10-25 and 50 kbp resolution) and inside the whole chromosome (no constraints). We use pre-cleaned, normalised and balanced Hi-C data (format .bed, .cool) to define the TADs regions. At this step, we need to be careful to choose the same genomic constraints, same resolution for the indexing and extraction of the matrix steps. To settle down the optimum parameters, we first look at the 50 kbp resolution (faster and less memory consuming) and without TADs constraints. Since we detect ATAC peak specifics to the double and triple mutants, we expect to see “loops” in APA-plot. This is how to understand a APA (aggregated peak analysis) plot:

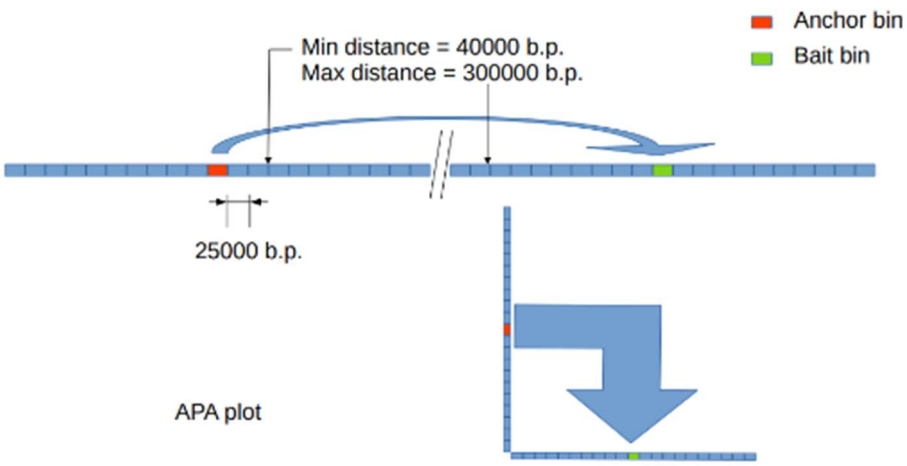


Figure 9. How to understand an APA plot (Dr.Robel Tesfaye), with a 25 kbp resolution.

Why look first without genomic constraints? Firstly because we want to get the maximum number of enh/prom contacts. Secondly, it was shown that it was easier to see loops inside the whole chromosome than within TADs constraints (Dr.Robel Tesfaye). As an example, in 2014 from 2.5Ga reads from Human cell lines, the researchers managed to get only 2.5 k loops (contact between enh/promoters) from roughly 30k enhancers!

We met some technical issues:

- the 10 kbp resolution Hi-C data demands a lot of informatics resources (more than 16GO RAM)
- the lin61-hpl2 mutants do not have TADs domains. (we use the N2 (WT) TADs domains)

The figures below show Hic-plot obtained from couples enhancers-promoters, close to all the genes. The intensity of interaction is represented by colours, ranging from blue (very low interaction) to red (higher interaction).

First, we test for the three conditions: WT(wild-type), Lm(Double-mutant) (figure 10) and Tri(Triple-Mutant) (Annex 8), at a resolution of 50kbp with no TADs constraints (inside the whole chromosome), between a minimum distance of 40 kbp and a maximum distance of 300kbp:

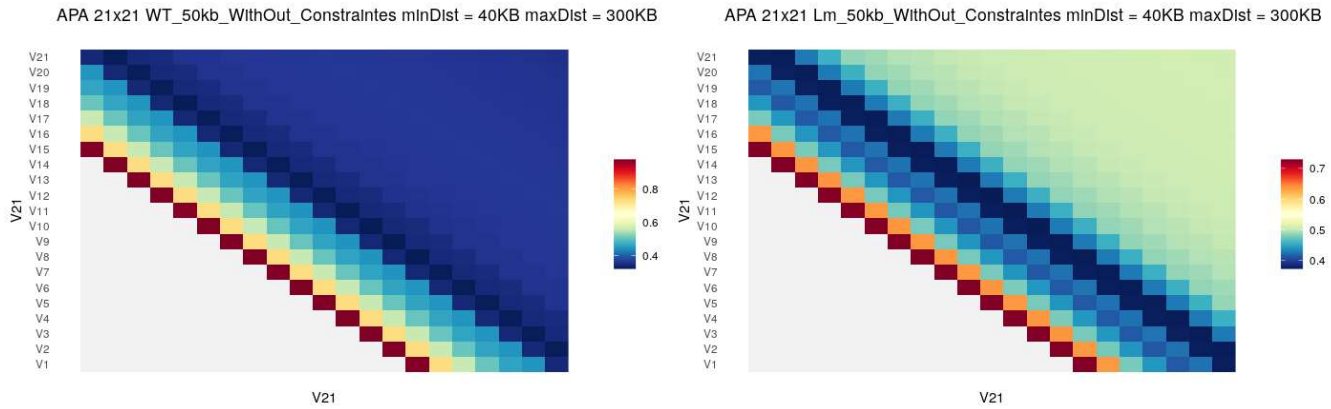


Figure 10. APA (aggregated peak analysis) plot - Contact Map -WT and double mutant (lm) between 40 to 300 kbp. (Annex 8) 50 kbp resolution

This plot represents a matrix of 21 bins in the x axis vs 21 bins in the y axis. Each bin represents 50 kbp (resolution). The data are symmetric. The white area represents regions where it was not possible to detect contact information between the given ATAC peaks and TSS. Because the distances we chose (40 to 300 kbp) were too short, there is a part where we can not measure any contacts (cf figure 9).

We want to detect interaction between a distance ranging from 40 to 300 kbp. And we have data with a 50 kbp resolution. This means that we want to look between less than 1 bin to 6 bins in the APA plot. As first suggested, the initial parameters like the high dimension of the APA matrix (21x21) and/or the distance parameters were not optimum. To solve this problem we tried to adjust the parameters by:

- either increasing the initial distances between the enhancers-promoters (250 kbp instead of 40 kbp as minimum), and 500 kbp instead of 300 kbp as maximum (we even try with 1M and more),

- and/or reduce the dimension of the APA from 21x21 to 11x11 (as a focus).

After readjusting the new parameters, increasing the range of distances between enhancers-promoters, and reduce the dimension of the APA, we obtained the plots in figure 11, for the three conditions, in the presence of genomic constraints (the enhancer and its TSS are located in same chromosome), or absence (the enhancer and its TSS can be located in different chromosomes) (Annex 8).

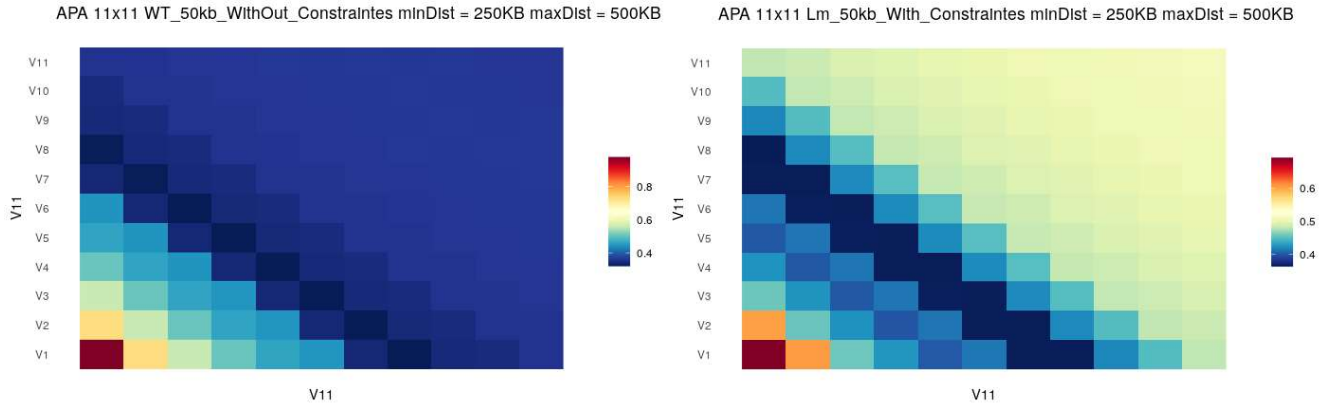


Figure 11. APA plot - Contact Map -WT and double mutant (lm).between 250 to 500 kbp. 50 kbp resolution.

We tried several parameters (different min and max distances, such as 600 kbp and 1Mbp) at different resolutions. None of them allow us to detect the expected loop with any or the condition (Annex 8). One hypothesis would be that the number of active enhancer/promoter couples are too low to be able to detect any contact.

To have more enhancer/promoter couple, we will try not to restrict the promoter regions to the active ones (with the histone mark H3K27ac, combined with RNApol II sites). As the promoters lists, we will only select ATAC peak (DNA sequence in the open chromatin area) close to the genes (1 kbp of length):

	Wild-Type	**Double-Mutant**	**Triple-Mutant**
Active TSS	1060	775	1593
TSS	7804	-	13041

Figure 12. Number of ATAC peaks inside a 1 kbp region around genes (promoter lists).

We will use (for each condition/sample):

- **As anchor:** the transcriptionally active ATAC peaks (ATAC seq results combined with CHIP-sep data: the double histone marks **H3K27ac/H3K4me**), as enhancers list
- **As bait:** all the ATAC peaks (ATAC seq results) restricted to a region close to 1 kbp around all genes, as promoter list.

After indexing, we used the search pairs function of HicAggR to create enhancer/promoter couples (matrices). Another hypothesis might explain why we were not able to see a loop. Until now we worked with all the genes. However, we found specific ATAC peaks (cf Figure 8) in the double and the triple mutant. Those new ATAC peaks correspond to new DNA sequences detected inside the relax part of the chromatin. Indeed, the disruption of TADs (double and triple mutant) induces new relaxed chromatin regions. We are interested only in these new regions. What we can do now, is to select only the specific ATAC peaks, forming new enhancer/promoter couples. To do so, we filtered the sub matrices we wanted: the one corresponding to the anchor (ATAC peaks combined with **H3K27ac/H3K4me**) of the triple mutant. This is performed with the FilterInteractions function of HicAggR: selection with a boolean column.

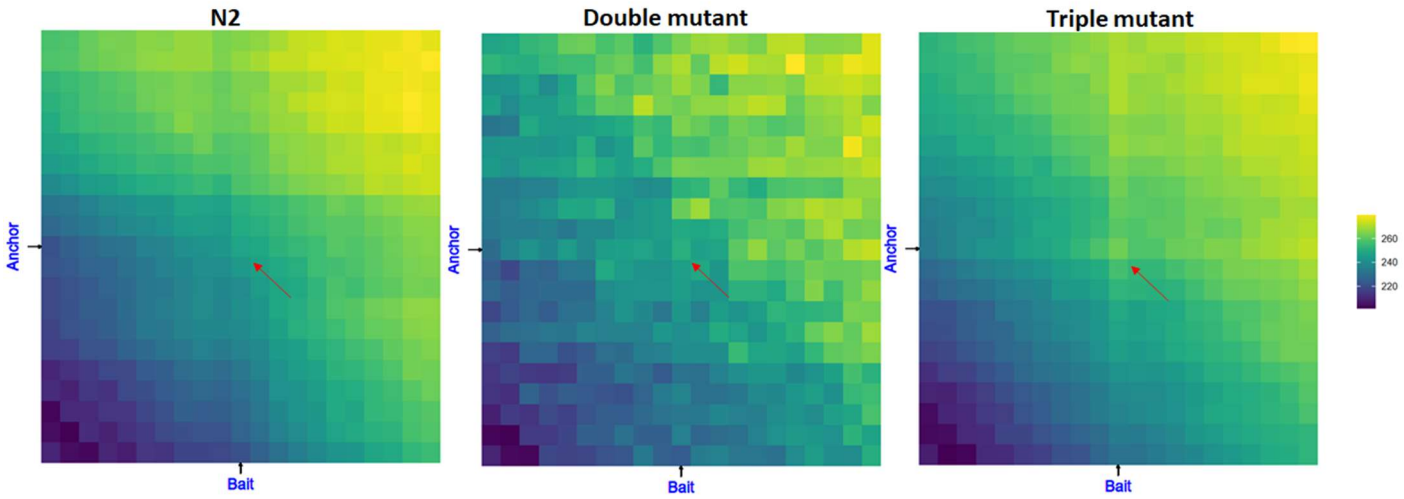


Figure 13. APA-CONTACT MAPs, with anchor (active ATAC peaks intergenic) specific to the triple mutant. 10 kbp resolution (1bin =10 kbp), distance 200-600 kbp, no TADs. Triple mutant matrix vs matrix HiC (WT)
Anchor -> enhancer, bait -> TSS (cf annex 8 for more details)
At the intersection of the anchor and Bait line, in the triple mutant only, we start to detect a yellow square, as "loop". This plot shows that the number of interactions between promoters and enhancers increase after the disruption of TADs (within the triple mutants set25/set32/met2).
The next step would be to extract the central values (matrices), with "getQuantif" function, and identifying particular enhancer/promoter couples (to have more contrast).

4) Conclusion and perspectives

TECHNICAL CONCLUSION:

In this study, we first learn how to pre-process ATACseq data:

- We checked their quality (fastp, Generate an interactive summary HTML report using MultiQC.11),
- We trimmed sequencing adapters,
- We mapped the reads to the reference genome (Bwa) and transformed them into bam, bai and bed files. We manipulated alignment files using SAMtools16 or sambamba17 to get coordinate-sorted BAM files.
- We filtered aligned reads by:
    - selecting only high quality data (minimum Q20, one error every 100 nucleotides)
    - removing duplicated (from PCR step), mapping artefacts
- We identified peaks, with peak calling (macs2)

We played with parameters of ATACseq preprocessing steps, and were careful to follow good practical rules. (https://haibol2016.github.io/ATACseqQCWorkshop/articles/ATACseqQC_workshop.html)

In a second time, we used the cleaned ATACseq peaks and compared them with CHIP-sep data. We discovered the GenomicRanges with the R package rtracklayer and learned how to manipulate GenomicRanges objects. This package is a particularly important package for treatment of genomics data. We selected DNA sequences presenting histone modifications related to transcriptionally active regions. We created lists of promoters, enhancers and enhancer-promoter couples close to the promoters of genes from WT and mutants (hpl2/lin 61 and set 25/32/met-2). We then used these lists and created map contacts. We looked at different conditions (inside the whole chromosome and inside TADs), distance ranges and resolution (HI-C data). We tried several parameters to find the best combination of resolution, interaction region to be able to see a loop like structure.

We accomplished 3 of the 4 expected missions.

This study gave us a great opportunity to play with various types of data: ATACseq, CHIP-seq, RNA-seq and HI-C data. We first discovered these techniques and were made aware of the difficulties of producing interpretable results and apprehend some of their limits.

BIOLOGICAL CONCLUSION

With the cleaned ATAC seq data, we were able to detect new open chromatin regions (specific ATAC peaks) with lin-61/hlp2 and set/met mutants. In this study, we observed that:
- the lin-61/hlp2 mutant lost their TADs domains (HiC data)
- up-regulated genes were found in the lin-61/hlp2 and the set-25/set-32/Met-2 mutants (RNAseq data)
- the disruption of TADs induces new contacts between promoters and enhancers in the set-25/set-32/Met-2 mutant.

The next steps would be to:
-to extract the central values (matrices), with "getQuantif" function, and identifying particular enhancer/promoter couples (to have more contrast).
-compare with RNAseq data and see if this de novo enhancer/promoters couples induced by the loss of TADs regions, lead to a up regulation for the concerned genes
-in a second time, we might focus on the loss of contact and the down regulation genes
-what is missing is the gene annotation part and the functional analysis of the up-regulated genes (with gene ontology) from the list we made, to see which genes and functions were affected.

SUGGESTION FOR OPTIMISATIONS

- Having replicates for the ATAC-Seq data could allow them to identify reproducible ATAC-peaks to perform differential accessibility analysis between mutants and wt N2 cells.
- We also suggest to perform either more deeply sequenced HiC data or to perform HiC pre-processing steps with higher resolution parameters such as using a bin size of 5000 or even 1000. We believe that at these resolutions, the minimal and maximal distances requested by the clients to assess enhancer-promoter contacts could be achieved.

We would like to give a special thanks to Dr. Robel Tesfaye!

5) ANNEXES

ANNEXE 0: REFERENCES

WEBSITES

⇒ Equipe Dr. Olivier CUVIER:

https://cbi-toulouse.fr/eng/equipe-cuvier

⇒ Publications, articles:

pubmed, NCBI,

⇒ GENERAL INFORMATIONS:

WORMBASE,
WIKIPEDIA,
UNIPROT

⇒ 3D organisation of the genome:

https://www.4dnucleome.org/
https://www.researchgate.net/

ARTICLES

1. TAD-free analysis of architectural proteins and insulators
Raphaël Mourad, Olivier Cuvier
Nucleic Acids Res. 2018 Mar 16;
46(5):e27. doi: 10.1093/nar/gkx1246.

2. Synergistic lethality between BRCA1and H3K9me2 loss reflects satellite derepression
Jan Padeken, Benjamin Towbin and SUSAN GASSER
February 2019 Genes & Development 33(7-8)
DOI:10.1101/gad.322495.118

3. Chromatin Modifiers SET-25 and SET-32 Are Required for Establishment but Not Long-Term Maintenance of Transgenerational Epigenetic Inheritance
Rachel M Woodhouse,Alyson Ashe
Cell Rep. 2018 Nov 20;PMID: 30463020

4. Transgenerational Epigenetic Inheritance Is Revealed as a Multi-step Process by Studies of the SET-Domain Proteins SET-25 and SET-32
Rachel M Woodhouse 1, Alyson Ashe
Epigenetic Insights. 2019
PMCID: PMC6466464,DOI: 10.1177/2516865719844214

5. C. elegans Heterochromatin Factor SET-32 Plays an Essential Role in Transgenerational Establishment of Nuclear RNAi-Mediated Epigenetic Silencing
Natallia Kalinava 1, Julie Zhouli Ni 2, Zoran Gajic 3, Matthew Kim 4, Helen Ushakov 5, Sam Guoping Gu
Cell Rep. 2018 Nov 20;
PMCID: PMC6317888, DOI: 10.1016/j.celrep.2018.10.086

6. Histone H3K9 methylation promotes formation of genome compartments in Caenorhabditis elegans via chromosome compaction and perinuclear anchoring
Qian Bian, Barbara J Meyer
Proc Natl Acad Sci USA, 2020 May
PMCID: PMC7261013, DOI: 10.1073/pnas.2002068117

7. Argonaute NRDE-3 and MBT domain protein LIN-61 redundantly recruit an H3K9me3 HMT to prevent embryonic lethality and transposon expression
Jan Padeken 1, Susan M Gasser

Genes Dev. 2021.
PMCID: PMC7778263,DOI: 10.1101/gad.344234.120

8. Understanding 3D Genome Organization and Its Effect on Transcriptional Gene Regulation Under Environmental Stress in Plant: A Chromatin Perspective
Suresh Kumar , Simardeep Kaur , Karishma Seem , Santosh Kumar , Trilochan Mohapatra
Cell Dev. Biol., 08 December 2021
Volume 9 - 2021 | https://doi.org/10.3389/fcell.2021.774719

9. Machine Learning Methods for Exploring Sequence Determinants of 3D Genome Organization
Muyu Yang and Jian Ma
JMB, Volume 434, Issue 15, August 2022,
167666

10. The Self-Organizing Genome: Principles of Genome Architecture and Function
Tom Misteli
Cell,Volume 183, Issue 1, 1 October 2020, Pages 28-45
https://doi.org/10.1016/j.cell.2020.09.014

11. Comprehensive mapping of long range interactions reveals folding principles of the human genome
Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender,Mark Groudine,Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander and Job Dekker
Science. 2009 Oct 9;
326(5950): 289–293.

12.Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions
Jesse R. Dixon,Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen,Ming Hu, Jun S. Liu and Bing Ren
Nature, 2012 May 17; 485(7398): 376–380.
doi: 10.1038/nature11082

13. Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome
Tom Sexton ,Eitan Yaffe ,Ephraim Kenigsberg,Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello,Amos Tanay,Giacomo Cavalli
Cell, Public ArchivePublished:January 19, 2012
DOI:https://doi.org/10.1016/j.cell.2012.01.010

14. Spatial partitioning of the regulatory landscape of the X-inactivation centre
Elphège P Nora [1] , Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, Edith Heard
Nature, 2012 Apr 11;485(7398):381-5.
doi: 10.1038/nature11049.

15. The Role of Chromosome Domains in Shaping the Functional Genome
Tom Sexton 1 and Giacomo Cavalli
Cell,Review,2015

16. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture
Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, Hadjur S
Cell Reports. (March 2015).
doi:10.1016/j.celrep.2015.02.004. PMC 4542312. PMID 25732821

17. Coupling 1D modifications and 3D nuclear organisation: data, models and function
Jost D, Vaillant C, Meister P (February 2017)
Current Opinion in Cell Biology. 44: 20–27
doi:10.1016/j.ceb.2016.12.001. PMID 28040646.

18. Comparing 3D Genome Organization in Multiple Species Using Phylo-HMRF
Yang Y, Zhang Y, Ren B, Dixon JR, Ma J .
Cell Systems, June 2019,**8** (6): 494–505.e14.
doi:10.1016/j.cels.2019.05.011. PMC 6706282. PMID 31229558.

19. From association to mechanism in complex disease genetics: the role of the 3D genome.
Fu, Y., Tessneer, K.L., Li, C. *et al.*
Arthritis Res Ther **20**, 216 (2018).
https://doi.org/10.1186/s13075-018-1721-x

20. Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles
Huan Tao Hao Li Kang Xu[...] Xiao-Chen Bo
January 2021,· Briefings in Bioinformatics 22(5)
DOI:10.1093/bib/bbaa405

21. Epigenetic regulation and chromatin remodeling in learning and memory
Kim, S., Kaang, BK.
Exp Mol Med 49, e281 (2017).
https://doi.org/10.1038/emm.2016.140

22. H3K4me3, H3K9ac, H3K27ac, H3K27me3 and H3K9me3 Histone Tags Suggest Distinct Regulatory Evolution of Open and Condensed Chromatin Landmarks
Anna A. Igolkina,[1,2,*] Arsenii Zinkevich,[3] Kristina O. Karandasheva,[4] Aleksey A. Popov,[3] Maria V. Selifanova,[3] Daria Nikolaeva, Victor Tkachev, Dmitry Penzar,[3,6] Daniil M. Nikitin and Anton Buzdin
Cells. 2019 Sep; 8(9): 1034
doi: 10.3390/cells8091034

23. Role of H3K9me3 Heterochromatin in Cell Identity Establishment and Maintenance
2020 May 16
Dario Nicetto and Kenneth Zaret
doi: 10.1016/j.gde.2019.04.013

24.**Regulation of 3D Organization and Its Role in Cancer Biology**
Peng A, Peng W, Wang R, Zhao H, Yu X, Sun Y.. Front Cell Dev Biol.
2022 Jun 8;10:879465.
doi: 10.3389/fcell.2022.879465. PMID: 35757006; PMCID: PMC9213882.

ANNEXE 1: METHOD PRINCIPLES

ATAC-seq

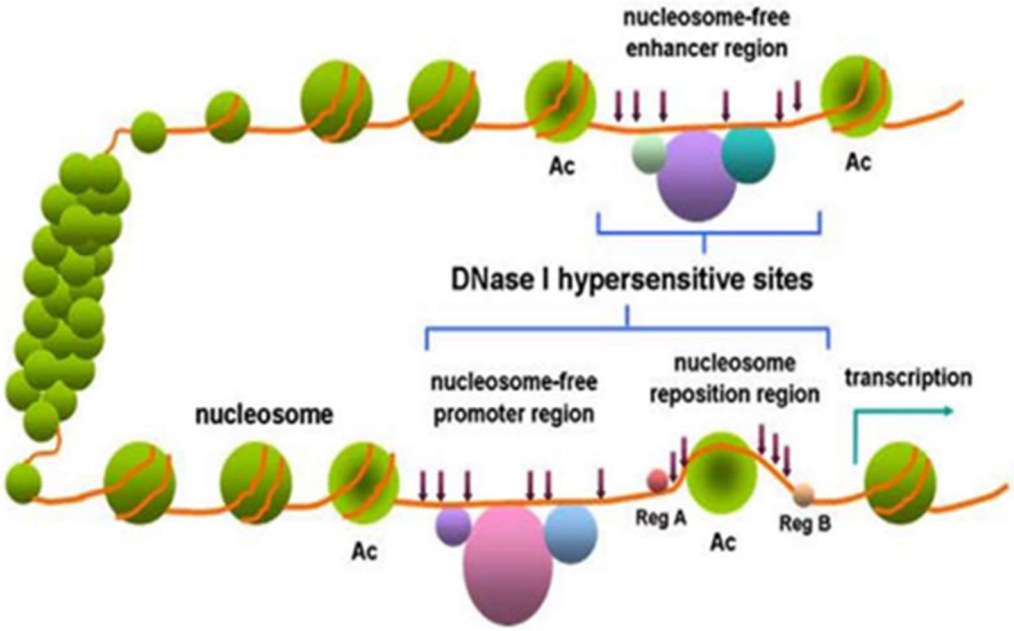


Figure 1. Principle of ATAC sequencing.

It relies on the use of an enzyme called transposase, which marks accessible DNA regions by inserting specific PCR adapter sequences. The principle of ATAC-seq data analysis might involve several steps, including:

1. Quality control and filtering of sequencing reads to remove low-quality and adapter-contaminated reads.
2. Alignment of reads to a reference genome to determine their genomic locations.
3. Identification of regions of open chromatin.
4. Quantification of chromatin accessibility by counting the number of reads that map to open chromatin regions.
5. Identification of differences in chromatin accessibility between the WT, the double and triple mutants.

HiC

HiC is a sequencing-based method that allows for the study of interactions between different regions of the genome. This technique enables the mapping of physical contacts between DNA fragments that are in close proximity in the three-dimensional space of the cell nucleus. Hi-C is a powerful method for studying the conformation of chromatin, which is the form in which DNA is organised inside the nuclei of cells.

The principle of the HiC method is to cross-link interactions between DNA regions in the cell, fragment the DNA, and ligate the ends of fragments that are in physical contact. The ligated fragments are then sequenced in paired-end mode, and the sequence data is analysed to map interactions between DNA regions.

The analysis of HiC data allows for the generation of maps of physical contacts between different regions of the genome. These maps enable the visualisation of chromatin structure at different scales, from short-range interactions between DNA regions to long-range interactions between regions on different chromosomes.

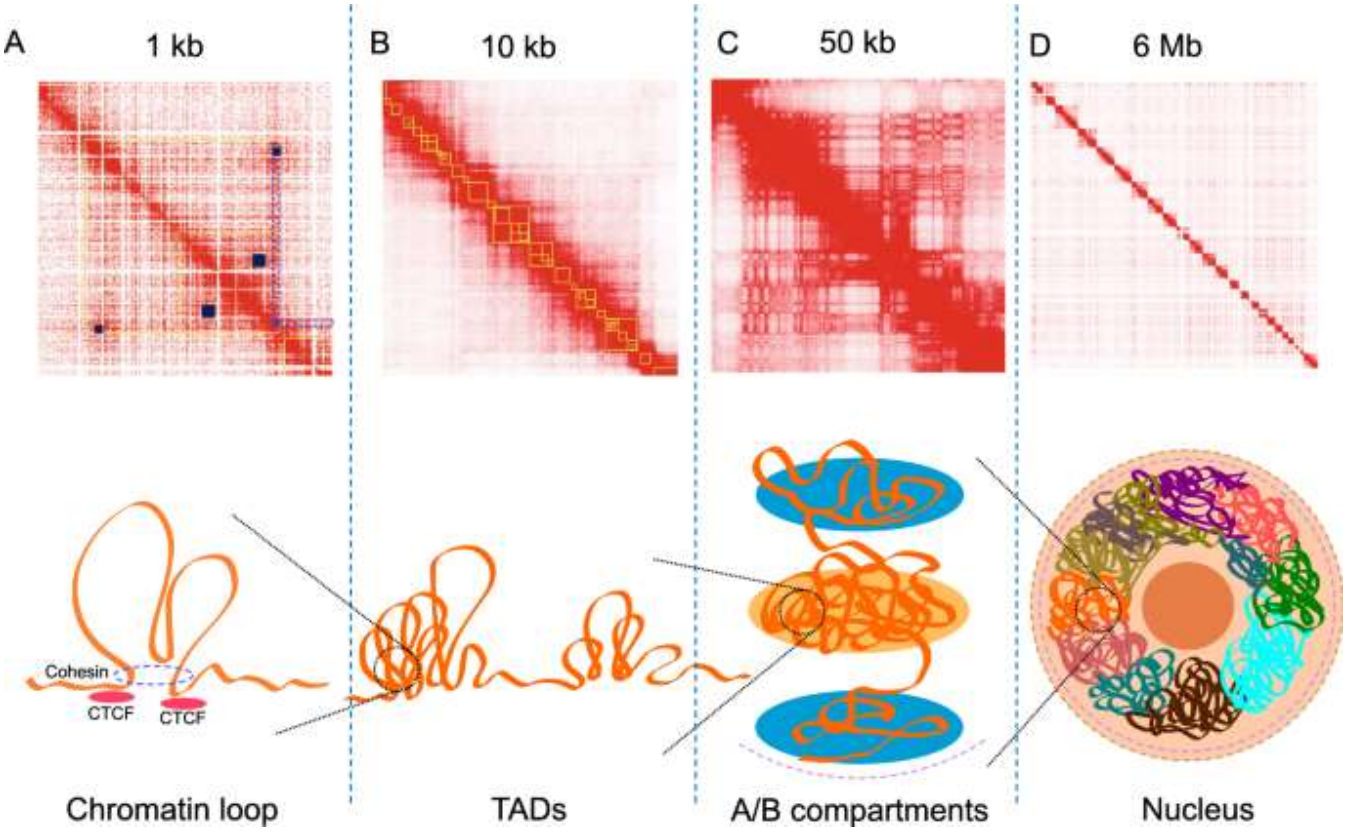


Illustration of genome architecture and the corresponding Hi-C interaction maps. Top panel: interaction heatmaps A, B, C, D are in different scales (kb or Mb per pixel) to correlate with the diagrams of 3D structures in the bottom panel, yellow boxes in A and B are identified TADs and small blue boxes in A indicate chromatin loops. The purple box in A is a frequently interacting region, with its classical "V" shape pattern coloured in purple dotted lines. Heatmaps were generated using Juicebox [29] with published Hi-C data of GM12878 [3]. Bottom panel: diagrams of 3D structures in the genome.

(Liu, N., Low, W.Y., Alinejad-Rokny, H. *et al.* Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C. *Epigenetics & Chromatin* **14**, 41 (2021). https://doi.org/10.1186/s13072-021-00417-4)

ChiP-seq

ChIP-seq stands for Chromatin Immunoprecipitation followed by sequencing. It is a high-throughput technique that is used to study protein-DNA interactions and the epigenetic modifications that regulate gene expression.

In ChIP-seq, a specific protein of interest is cross-linked to DNA in living cells, and the protein-DNA complexes are then isolated and fragmented. Antibodies are used to selectively pull down the protein of interest, along with any DNA fragments that are bound to it. The resulting protein-DNA complexes are then sequenced using next-generation sequencing technologies to identify the genomic regions where the protein of interest is bound.

RNA-seq

RNA-seq (RNA sequencing) is a high-throughput sequencing technique that allows researchers to profile the transcriptome of cells or tissues. In RNA-seq, RNA molecules are extracted from cells or tissues and converted into cDNA fragments. These fragments are then sequenced using next-generation sequencing technologies to generate millions of short reads that can be mapped back to the reference genome or assembled de novo to generate a transcriptome. By analysing the number and distribution of reads that map to each gene or transcript, RNA-seq can be used to quantify gene expression levels.

ANNEXE 2: Pipeline Atac-seq

```
# Modifier les noms des chromosomes (UCSC to Ensembl).
Par ex: dans R rtracklayer::import.bed("/home/robel/mount/cbi-filer/grpcuvier/Robel/PTUT/RefGenome/ce11-blacklist.v2.bed.gz") %>% `seqLevelsStyle<-`("Ensembl") %>%
export.bed("/home/robel/mount/cbi-filer/grpcuvier/Robel/PTUT/RefGenome/ce11-blacklist.v2_ens.bed")

# dans R:
rtracklayer::import.bed("/home/cdelamare/Bureau/Ptut/data/ce11-blacklist.v2.bed.gz") %>% `seqLevelsStyle<-`("Ensembl") %>% export.bed("/home/cdelamare/Bureau/Ptut/data/
ce11-blacklist.v2_ens.bed.gz")

BLACKLIST_DCC = "/home/cdelamare/Bureau/Ptut/data/ce11-blacklist.v2_ens.bed"


# Indexer le génome de référence, a faire 1 seule fois.
bwa index Caenorhabditis_elegans.WBcel235.dna_sm.toplevel.fa.gz



# Vérifier la qualité des reads, filtrer les reads de mauvaise qualité et trimer (enlever) les adapteurs de séquençage

fastp -i /home/cdelamare/Bureau/Ptut/data/Atac-seq/X201SC21033194-Z01-F001/raw_data/N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_1.fq.gz -o /home/cdelamare/Bureau/Ptut/
data/Atac-seq/X201SC21033194-Z01-F001/raw_data/N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_1_FILTERED.fq.gz -I /home/cdelamare/Bureau/Ptut/data/Atac-seq/
X201SC21033194-Z01-F001/raw_data/N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_2.fq.gz -O /home/cdelamare/Bureau/Ptut/data/Atac-seq/X201SC21033194-Z01-F001/raw_data/
N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_2_FILTERED.fq.gz --detect_adapter_for_pe -q 20


# Aligner les reads au génome de référence
bwa mem -t 8 /home/cdelamare/Bureau/Ptut/data/GenomeReference/Caenorhabditis_elegans.WBcel235.dna_sm.toplevel.fa /home/cdelamare/Bureau/Ptut/data/Atac-seq/
X201SC21033194-Z01-F001/raw_data/N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_1_filtered.fq.gz /home/cdelamare/Bureau/Ptut/data/Atac-seq/X201SC21033194-Z01-F001/
raw_data/N2_I_R1/N2_I_R1_DKDL210002613-1a_HYFHHDSXY_L2_2_filtered.fq.gz | samtools view - -@8 -S -b -q 20 > N2_I_R1_q20.bam
```

```
# Enlever les duplicats PCR

samtools sort -@ 4 -T N2_I_R1_q20.tmp.bam -n N2_I_R1_q20.bam | samtools fixmate --threads 4 -m - N2_I_R1_q20_fixedMate.bam

samtools sort --threads 3 N2_I_R1_q20_fixedMate.bam | samtools markdup -s -r --threads 3 - N2_I_R1_q20_noDup.bam 2> N2_I_R1_q20.dupStats


# Enlever les reads dans des régions black listé et avec flagstat de 1804 (Voir https://broadinstitute.github.io/picard/explain-flags.html). Ce flag repère les reads non alignés, les mates (R1 ou R2) non alignés, alignement sur multiples sites du génome etc.

# source pour les régions sur black list (https://github.com/Boyle-Lab/Blacklist/blob/master/lists/ce11-blacklist.v2.bed.gz). Nécessite de changer les noms des chromosomes de chrI (style UCSC) à I (style Ensembl).

bedtools intersect -v -a N2_I_R1_q20_noDup.bam -b /home/cdelamare/Bureau/Ptut/data/ce11-blacklist.v2_ens.bed | samtools view -bu -F 1804 - | samtools sort -@ 4 -T
N2_I_R1_q20_noDup.tmp.bam -o N2_I_R1_filtered.bam -


# Indexer pour loader sur IGV si nécessite visualisation
samtools index -@ 6 N2_I_R1_filtered.bam


# Pour voir le nombre de reads de bonne qualité qu'il reste
samtools flagstat N2_I_R1_filtered.bam > N2_I_R1_filtered.bam.flagStat


# le peak calling

macs2 callpeak -t N2_I_R1_filtered.bam --keep-dup all -f BAMPE -g ce --outdir peak_macs2 -n N2_I_R1 -p 1e-1 --broad --broad-cutoff 1e-5 &> N2_I_R1.macs2.log


# Générer un fichier bigwig pour faire les figures et les visualisations (plus léger que le bam et normalisé). Nécessite de faire samtools index avant. Pour effective genome size (https://deeptools.readthedocs.io/en/develop/content/tools/bamCoverage.html). La taille de nos reads est de 150 nt (zcat L_tm_I_R1_DKDL210002614-1a_HYFHHDSXY_L2_1.fq.gz | head).
bamCoverage -b N2_I_R1_filtered.bam --ignoreDuplicates --effectiveGenomeSize 98721253 --skipNonCoveredRegions --normalizeUsing RPKM -p 5 --extendReads -o N2_I_R1_rpkm.bw 2>
N2_I_R1_deeptools.log
```

ANNEXE 3: Workflow Atac-seq

from Dr.Robel Tesfaye

ANNEXE 4: HicAggR pipeline

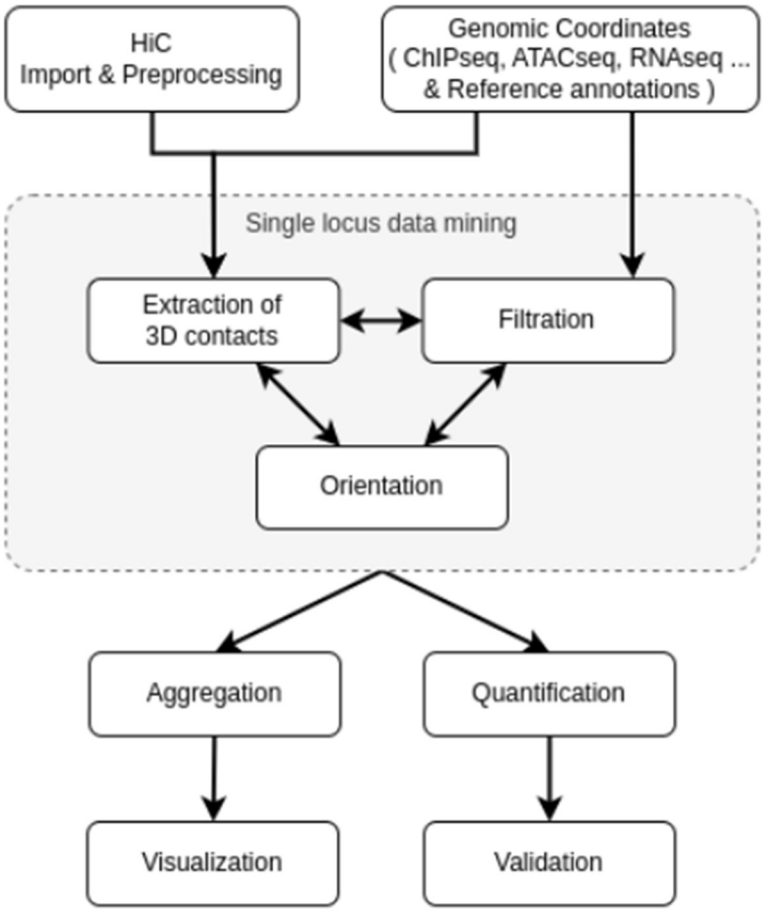


Figure 1. Workflow of HI-C data analysis.

More informations can be found in Prof. Cuvier's git (library not published yet) :
https://cuvierlab.github.io/HicAggR/

1) Additional genome information;

Before the Hi-C pipeline, we need to provide additional information about the genome, like chromosome names and sizes, and the binSize used for the HiC matrices, this information is stored as a data frame.

2) Import HiC data;

As the first step, we need to provide the Hic data (supported files; .hic, .cool, .mcool, .h5, .hdf5), cool is used in this pipeline.

3) Balancing and distance-based normalization:

for our case is not required because the Hic data obtained are already normalized.

4) Import genomic coordinates and indexing features:

In this step, we import the genomic coordinates data that include both the anchor data (such as peaks from ChIP-seq or other assays) and the bait data (TSS). This step ensures that our genomic coordinates align with the same reference genome used for the Hi-C data. (such as binSize and chromosome sizes).

5) Search Pairs:

By using the indexed genomic features from the previous step, this step identifies pairs of anchor data and bait data that are within the minimum and maximum distances provided. The purpose of this step is

to focus on interactions between the anchor and the bait that occur within a biologically relevant distance range. In our study, we initially set this range between 40KB and 300KB to identify significant genomic interactions (as requested by the clients).

6) SubMatrices extractions :

The putative couples are formed and the HiC data are ready for downstream analysis. contact signals for the regions of interest (ATAC peaks as anchors and gene promoters as baits for example) and their surroundings are extracted.

7) Plot and visualisation :

Submatrices are aggregated as sum, average or median. Then, the aggregated matrix is plotted as a heatmap of contact frequencies, the intensity of the interaction represented by a color code, from low interaction (blue) to high interaction (red).

ANNEXE 6: RESULTS FASTP

RESULTS CLEANING RAW DATA ATACseq

STEP 1: CONTROL QUAL	FASTQC								
		Before filtering			After filtering			Filtering result	CONCLUSION	
		total reads	Q20	Q30	total reads	Q20	Q30	reads passed filtered	Observations	Quality conclusion
WT	N2	67.54M	92%	86%	67.17M	97%	92%	99%	good amout of reads, with high % of high qualitty	GOOD
DOUBLE MUTANT	htl2 lin-61	67.56	94%	87%	67.11%	97%	91%	99%	good amout of reads, with high % of high qualitty	GOOD
TRIPLE MUTANT	set 25 set 32 met-2	49,9M	92%	86%	49,7M	97%	92%	99%	good amout of reads, with high % of high qualitty	GOOD

STEP 2: FILTRATION							
		Q20	DUPLICATES		VIABILITY		
			total du	total read	p value = 0,0	p value =	p value = 0,01
WT	N2	1,3 M	253,2 k	1562877	80	6001	9094
DOUBLE MUTANT	htl2 lin-61	45,15M	11.2M	33.93M	807	-	12277
TRIPLE MUTANT	set 25 set 32 met-2	43,3 M	10,56 M	36,9 M	8926	-	26770

Example of html result with Lm (double mutant)

fastp report

Summary

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (150 cycles + 150 cycles)
mean length before filtering:	150bp, 150bp
mean length after filtering:	112bp, 112bp
duplication rate:	8.276294%
Insert size peak:	65
Detected read1 adapter:	CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
Detected read2 adapter:	CTGTCTCTTATACACATCTGACGCTGCCGACGA

Before filtering

total reads:	67.562970 M
total bases:	10.134445 G
Q20 bases:	9.541136 G (94.145611%)
Q30 bases:	8.881377 G (87.635550%)
GC content:	45.391725%

After filtering

total reads:	67.215960 M

total bases:	7.560550 G
Q20 bases:	7.346867 G (97.173706%)
Q30 bases:	6.941395 G (91.810718%)
GC content:	41.154586%

Filtering result

reads passed filters:	67.215960 M (99.486390%)
reads with low quality:	319.976000 K (0.473597%)
reads with too many N:	4.864000 K (0.007199%)
reads too short:	22.170000 K (0.032814%)

Adapters

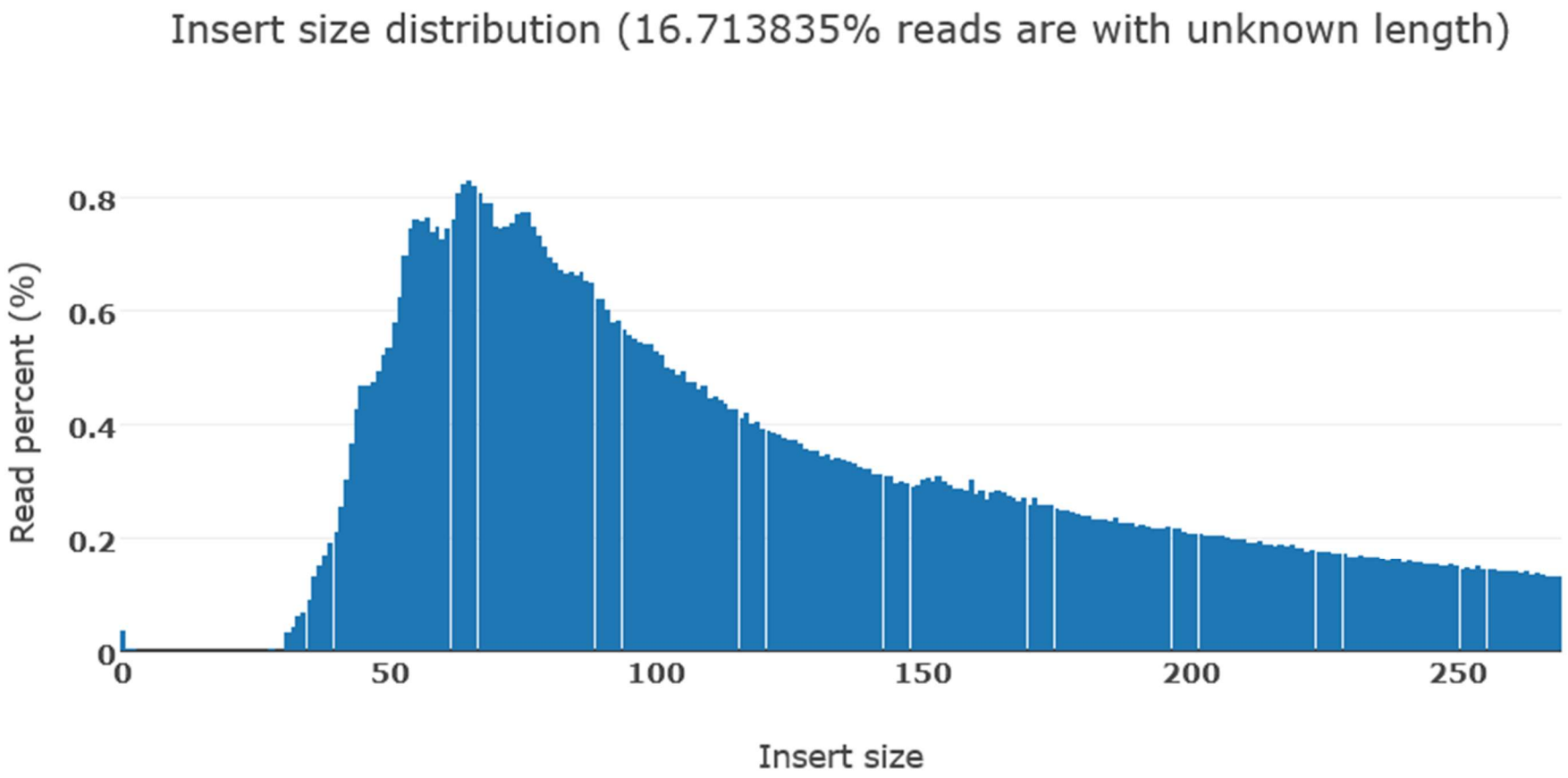
Adapter or bad ligation of read1

Sequence	Occurrences
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCCTGAGCATCGCGTATGCCGTCTTCTGCTT	387969
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTT	792436
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCCTGAGCATCGCGTATGCCGTCTTCTGCTTGAAAA	465782
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTGAAAA	1119590
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTCCTGAGCATCTCGTATGCCGTCTTCTGCTTGAAAAT	306240
other adapter sequences	16940075

Adapter or bad ligation of read2

Sequence	Occurrences
CTGTCTCTTATACACATCTGACGCTGCCGACGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAA	374538
CTGTCTCTTATACACATCTGACGCTGCCGACGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAA	231404
other adapter sequences	19390197

Insert size estimation

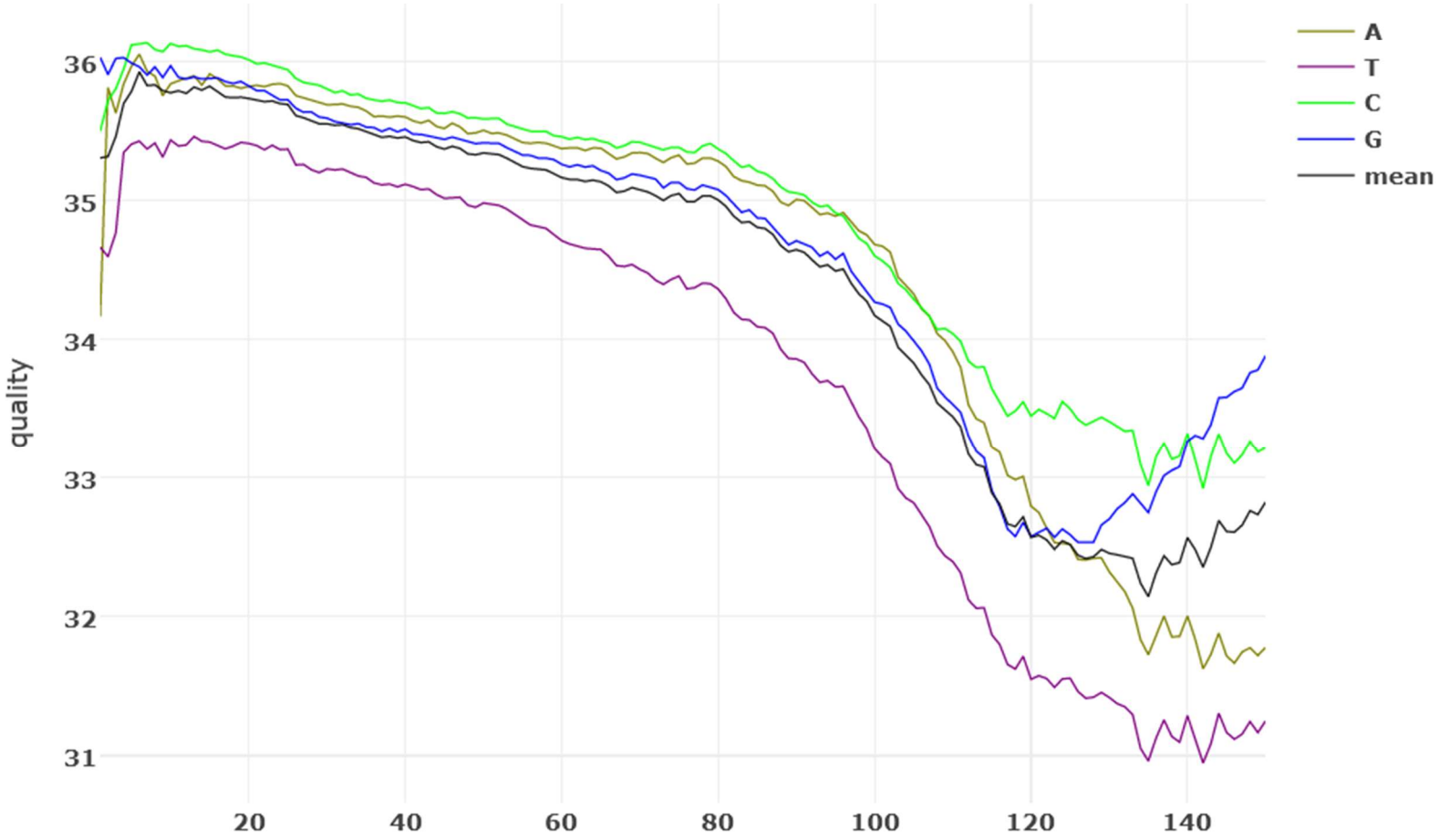


This estimation is based on paired-end overlap analysis, and there are 16.713835% reads found not overlapped.
The nonoverlapped read pairs may have insert size <30 or >270, or contain too much sequencing errors to be detected as overlapped.

Before filtering

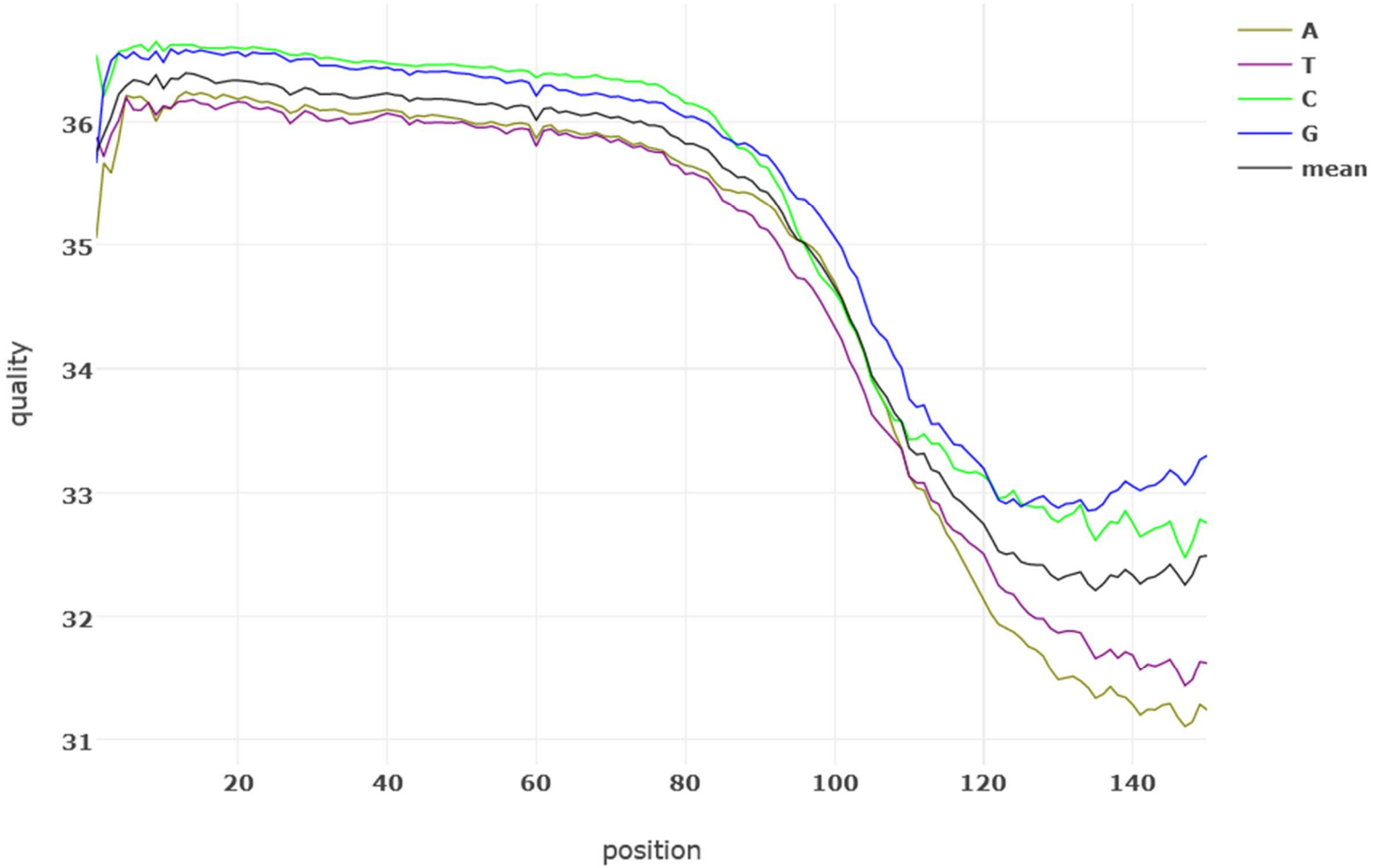
Before filtering: read1: quality

Value of each position will be shown on mouse over.



Before filtering: read2: quality

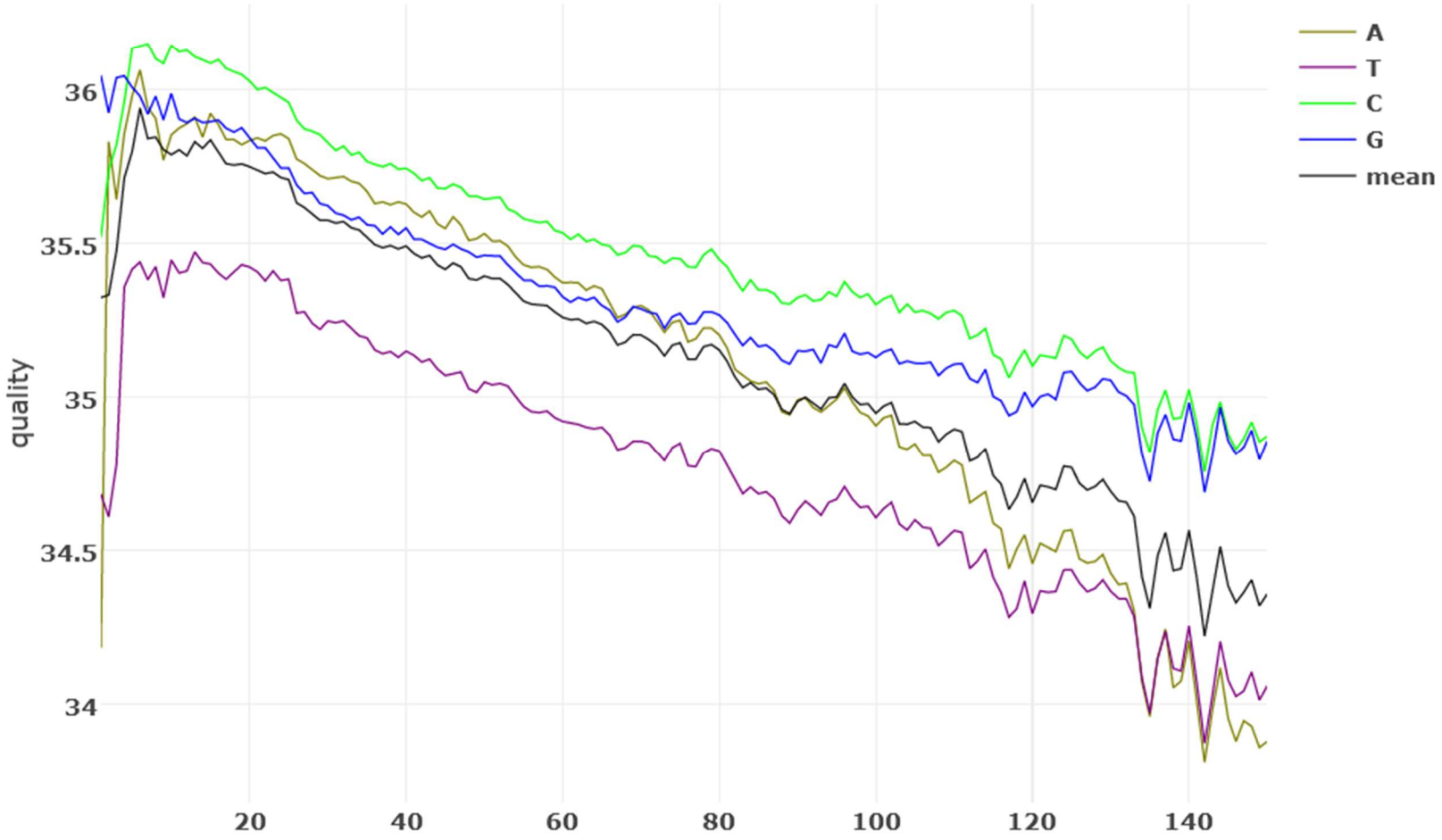
Value of each position will be shown on mouse over.



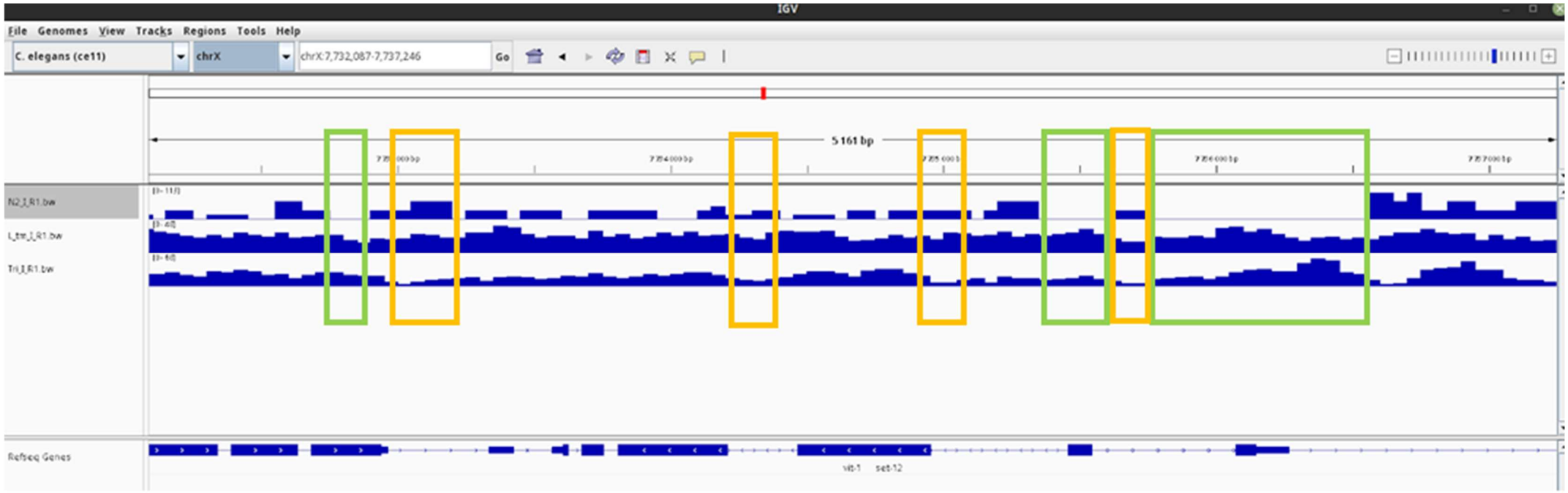
After filtering

After filtering: read1: quality

Value of each position will be shown on mouse over.

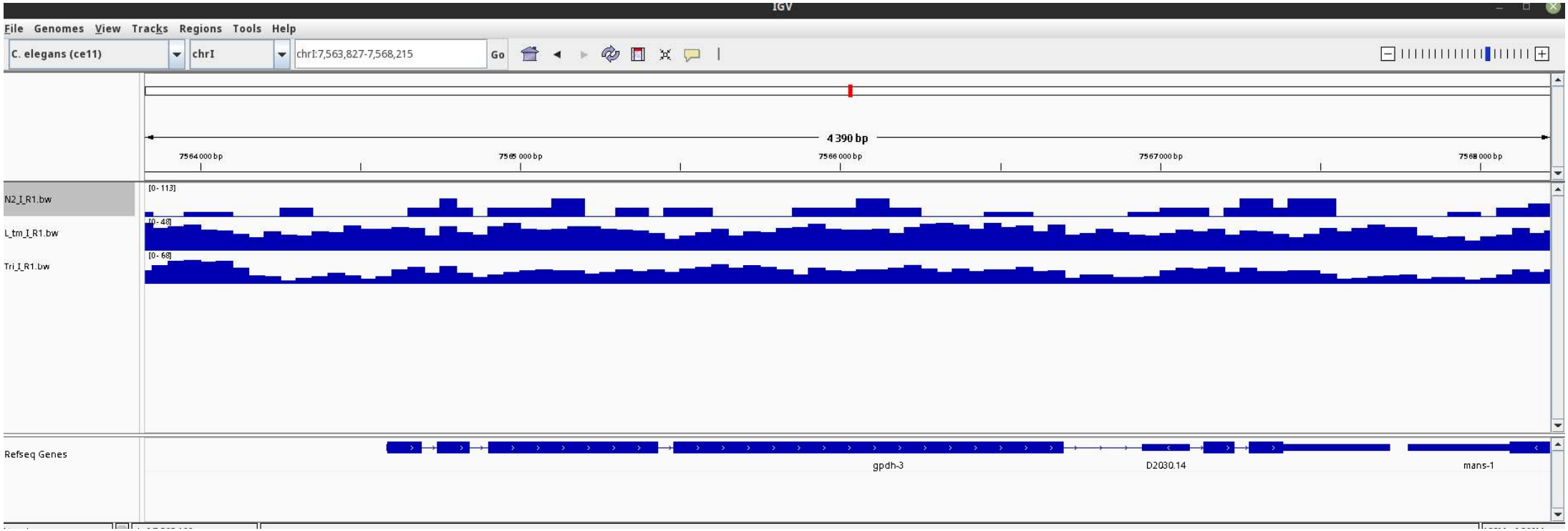


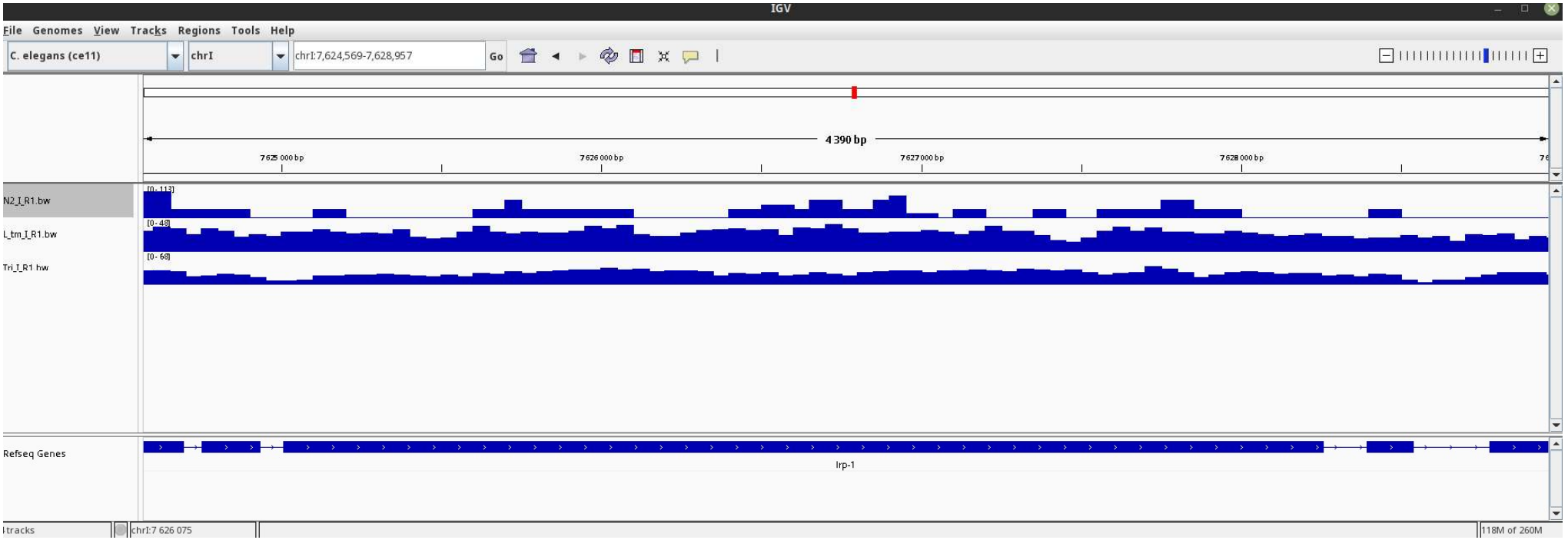
ANNEXE 6: RESULT IGV



Peak specifics to mutants

More peaks in the double mutant
Reduction of peaks in the triple mutant


IGV
File Genomes View Tracks Regions Tools Help
C. elegans (ce11)
chrI
chrI:7,563,827-7,568,215
Go
4 390 bp
7564 000 bp
7565 000 bp
7566 000 bp
7567 000 bp
7568 000 bp
N2_I_R1.bw
[0 - 113]
L_tm_I_R1.bw
[0 - 48]
Tri_I_R1.bw
[0 - 68]
Refseq Genes
gpdh-3
D2030.14
mans-1


IGV
File Genomes View Tracks Regions Tools Help
C. elegans (ce11)
chrI
chrI:7,624,569-7,628,957
Go
4 390 bp
7625 000 bp
7626 000 bp
7627 000 bp
7628 000 bp
N2_I_R1.bw
[0 - 113]
L_tm_I_R1.bw
[0 - 48]
Tri_I_R1.bw
[0 - 68]
Refseq Genes
lrp-1
tracks
chrI:7 626 075
118M of 260M

ANNEXE 7: RESULT COUPLE CREATION

ENHANCER PROMOTOR COUPLE

we summurize the number of objects we found after selection (genomic ranges)

DATA CHIP seq (narrow peak)

	FULL GENOME	ACTIVE REGION spe ENHANCER		ACTIVE REGION spe PROM		INACTIVE REGION
HISTONE MARKS	None	H3K27ac	H3K27ac / H3K4me	RNApol II	H3K27ac/RNApol II	H3K27me3
GENES	49 904	15 436	8 778	4 812	4 808	10 206
TSS	49 904	7 727	-	3 240	2 402	6 140

STEP 1: SELECTION OF ATACseq PEAKS IN ACTIVE AND INACTIVE REGION: OVERLAPPING DATA ATACseq WITH DATA CHIPSEQ

		RAW ATACseq	ENHANCER ACTIVES	PROMOTOR ACTIVES	INACTIVE REGION
	HISTONE MARKS	None	H3K27ac / H3K4me	H3K27ac/RNApol II	H3K27me3
WT	N2	9 094	2 739	1060	1633
DOUBLE MUTANT	htl2 lin-61	12 277	2 018	775	663
TRIPLE MUTANT	set 25 set 32 met-2	26 770	4 576	1 593	2 785

number of peak (g.ranges)

		ENHANCER	PROMOTOR
	HISTONE MARKS	H3K27ac / H3K4me	H3K27ac/RNApol II
WT	N2	2 739	1633
DOUBLE MUTANT	htl2 lin-61	2 018	663
TRIPLE MUTANT	set 25 set 32 met-2	4 576	2 785

ANNEXE 8: RESULT HIC-AGGR

1/ Test 1 with: Initial parameters:

21*21 matrix,

min distance = 40 kbp

max distance = 300 kbp

no genomic constraints (whole chromosomes)

remove 0 = true (default parameter, good do not change it)

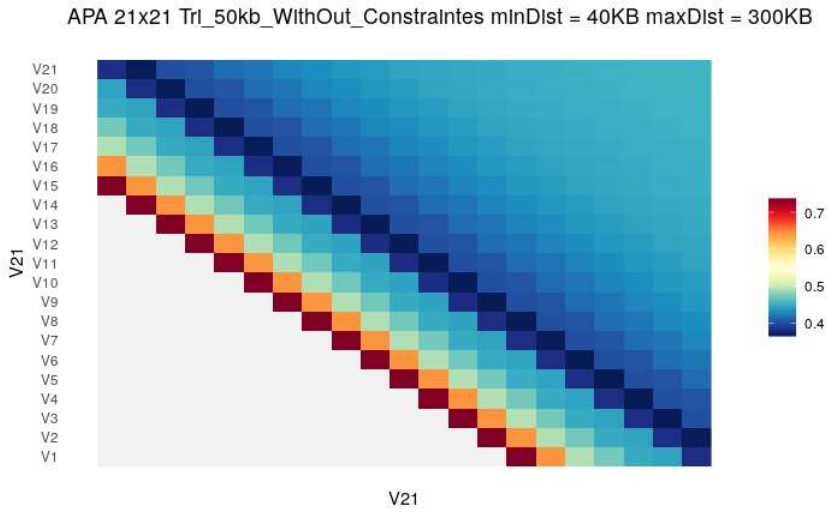
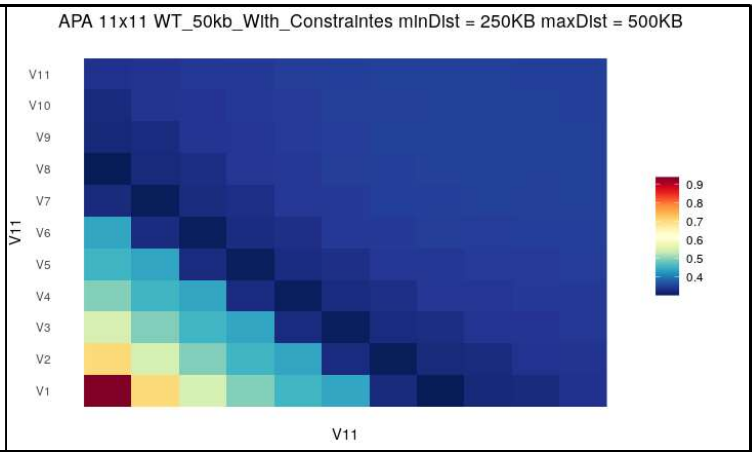
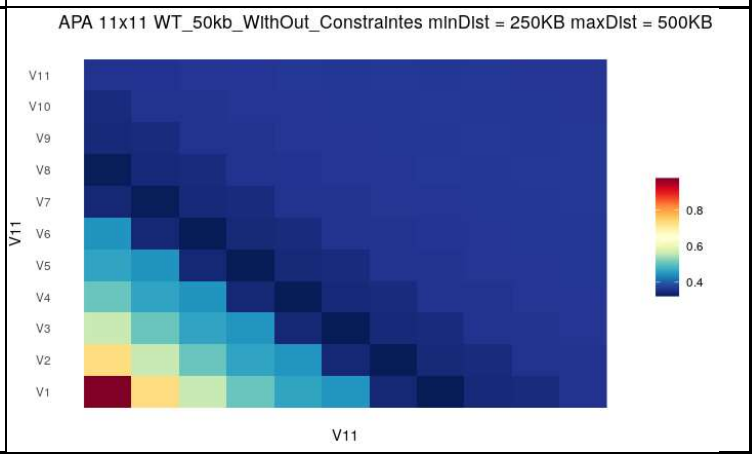
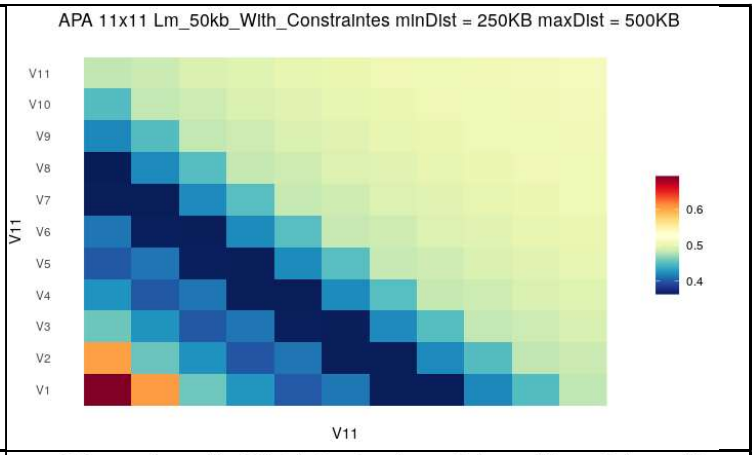
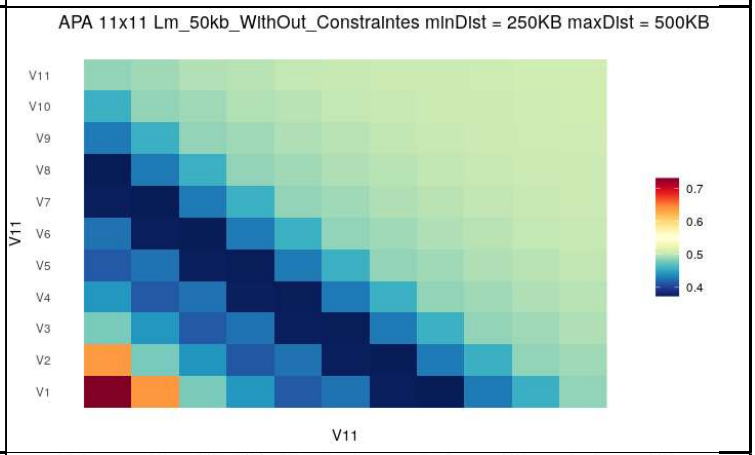
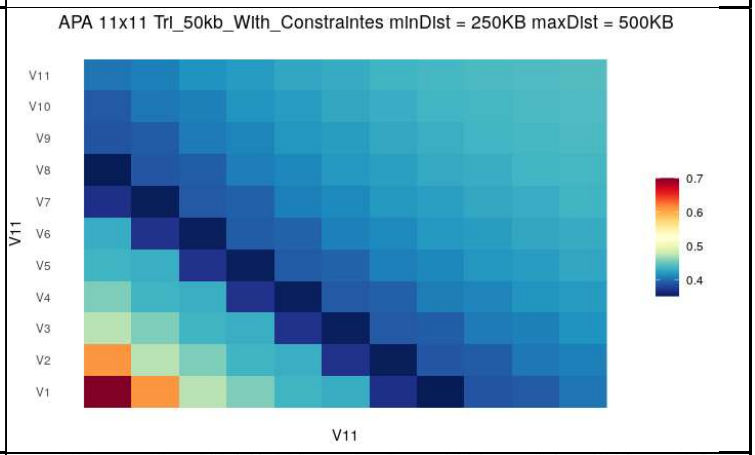
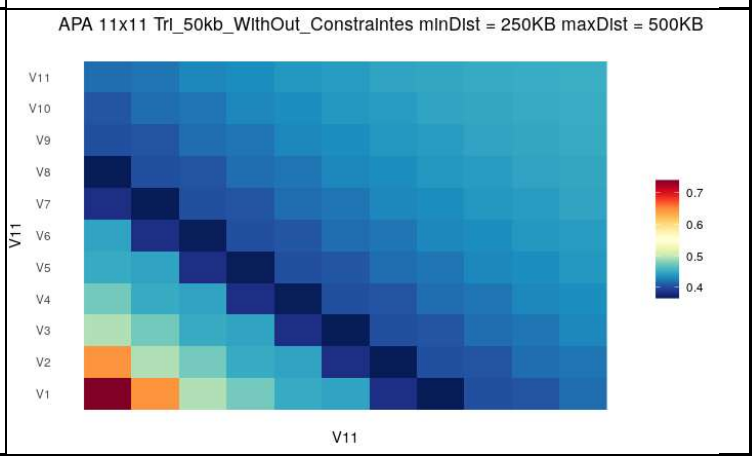


Figure 1. APA plot - Hic Contact map of triple mutant (Tri).

2/ Test 2: compare several parameters

Wild-Type	With Constraints	
	Without Constraints	

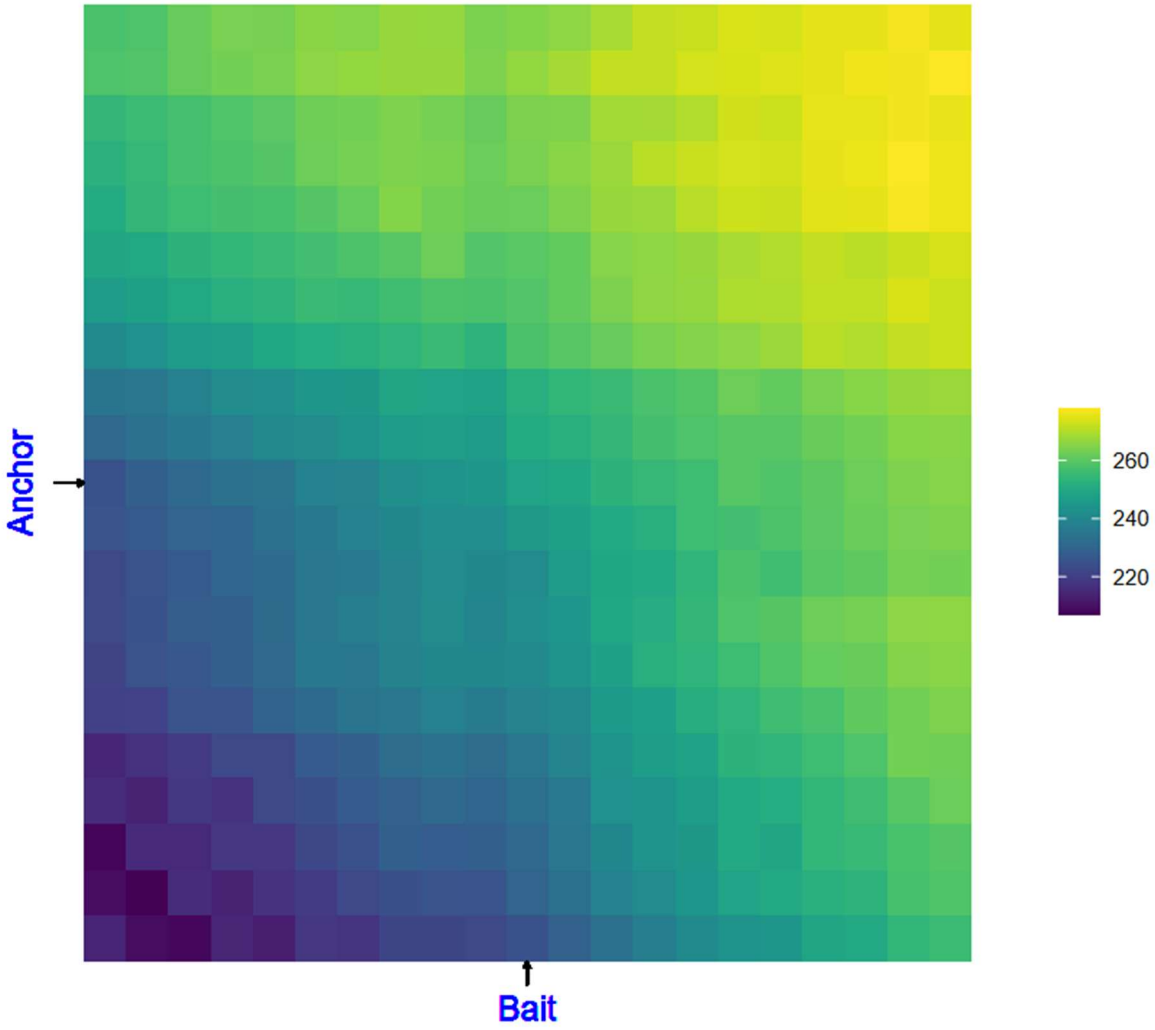
Double-Mutant	With Constraints	
	Without Constraints	
Triple-Mutant	With Constraints	
	Without Constraints	

Conditions used (Dr. Robel Tesfaye):
- 10 kbp resolution (files .cool)
- No TADs (whole chromosome, no genomic constraints)
- distance: min 200 kbp - 600 kbp

⇒ N2

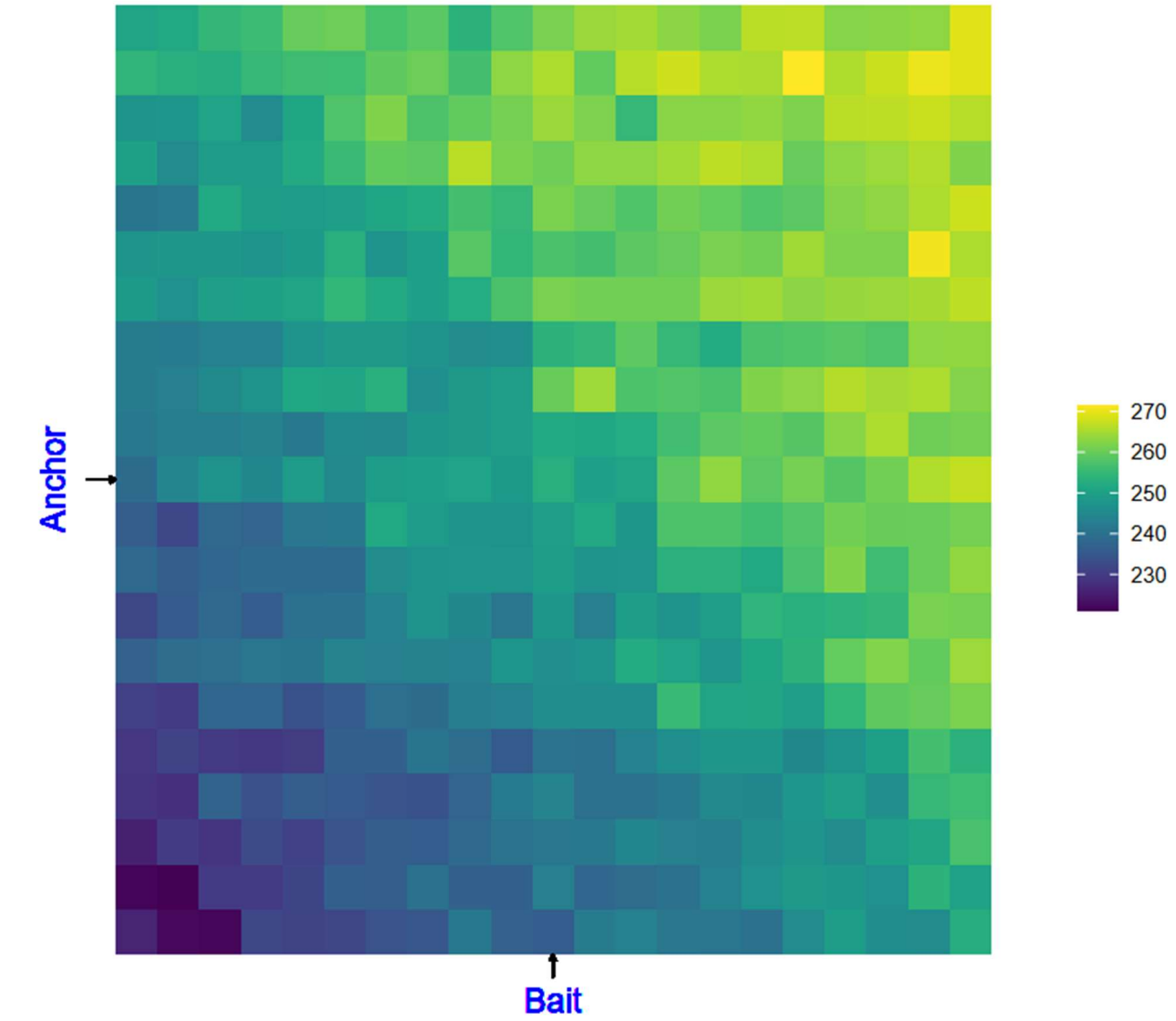
Agregation

scale (auto), center()



	value
totalMatrixNumber	17877
filteredMatrixNumber	17877
minimalDistance	NA
maximalDistance	NA
zeroRemoved	TRUE
resolution	10000
referencePoint	pf
matriceDim	21
transformationMethod	matrix(dplyr::percent_rank(c(x))*500,21,21)
aggregationMethod	mean(x,na.rm=T,trim=0.01)

⇒ **double mutant: lin-61/hpl2**

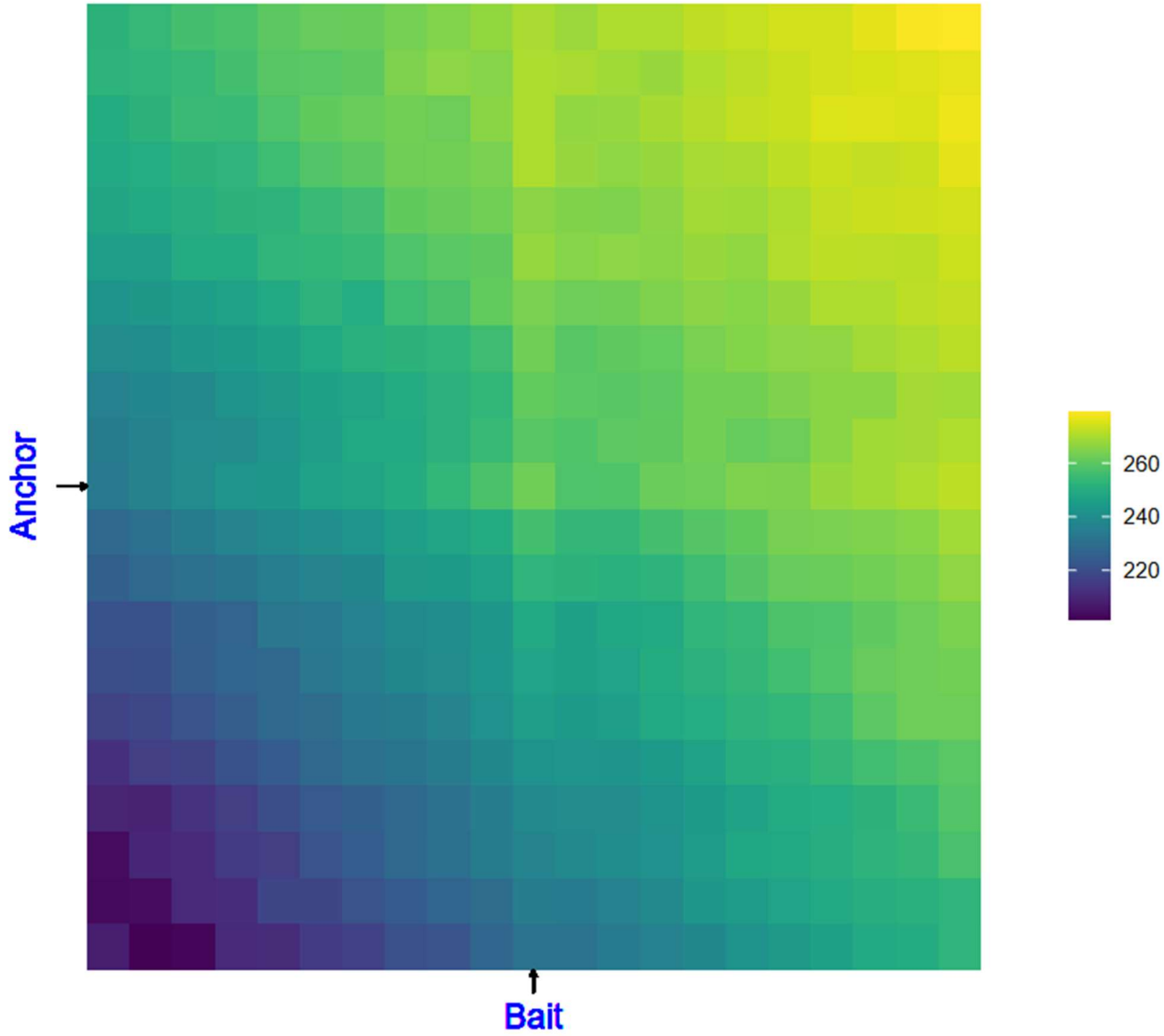


	value
totalMatrixNumber	3230
filteredMatrixNumber	3230
minimalDistance	NA
maximalDistance	NA
zeroRemoved	TRUE
resolution	10000
referencePoint	pf
matriceDim	21
transformationMethod	matrix(dplyr::percent_rank(c(x))*500,21,21)
aggregationMethod	mean(x,na.rm=T,trim=0.01)
selection	Reduce(intersect,list(anchor.N2_ATAC.bln,anchor.dble_ATAC.bln,; bait.TSS_Lm.bln))

⇒ **Triple mutant: set25/set32/met2**

Agregation

scale (auto), center()



	value
totalMatrixNumber	17877
filteredMatrixNumber	17877
minimalDistance	NA
maximalDistance	NA
zeroRemoved	TRUE
resolution	10000
referencePoint	pf
matriceDim	21
transformationMethod	matrix(dplyr::percent_rank(c(x))*500,21,21)
aggregationMethod	mean(x,na.rm=T,trim=0.01)
selection	Reduce(intersect,list(anchor.triple_ATAC.bln,anchor.N2_ATAC.bln,; anchor.dble_ATAC.bln,bait.TSS_tri.bln))

ANNEXE 9: PROJECT MANAGEMENT

Sharing tasks:

- **“Cahier de spécification”:** We share all the writing parts
- **literature:** Each one of us is doing bibliography to update with the issues of our project.
- **Pipeline ATACseq:** Each person of the group has a sample and is doing the full pipeline of ATACseq.
- **Pipeline HiCAggR:** All members of the group explore the library with example data.. We analyse our samples with the library . We take Julie and mainly Youcef's results because they manage to go further into the analysis.
- **Report writing:** We share all the redaction
- **presentation realisation** (ppt): Mainly Claire

REPARTITION DES TACHES			Rstudio			
RESULT COLLECTING			rtracklayer	HicAGGR		
		PRE PROCESSING/CLEANING ATACseq DATA	CREATION COUPLE	discovery	First try	Testing various parameters
	N2	Claire	Claire	Claire	Youcef /Julie	Youcef
	Lm	Youcef	Youcef	Youcef	Youcef /Julie	Youcef
	Trim	Julie	Julie	Julie	Youcef / Julie	Youcef
	N2 vs Lm	-	-	-	Youcef	-
	N2 vs Tri	-	-	-	Julie	-

LIVRABLES					Redaction			Gitlab
		pipeline linux	Rmardown HicAGGr	Organisation	Cachier charge	rapport	ppt	
	Claire	x			xxx	xx	xxx	xx
	Youceff		x		x	xx	x	xx
	Julie			x	xx	xx	x	x

Organisation of group work:

- With the team, we had frequent meetings, every 2 weeks and at least once a week toward the end of the project.
- Meetings with the tutor were once every two weeks and once a week toward the end of the project.
- We reported the meeting to CR, and sent it to all of us. The tasks were split based on people preferences, and maintain frequent contact (mail, phone, zoom). We help each other at every step.
- We created TO-DO lists, with deadlines (SMART method).
- Gitlab
- Google Drive

CONCLUSION

Our project was realised with an equal implication, and participation. We shared the work equitably, based on each person's strengths. This job was done in a real team based work, in a good atmosphere.