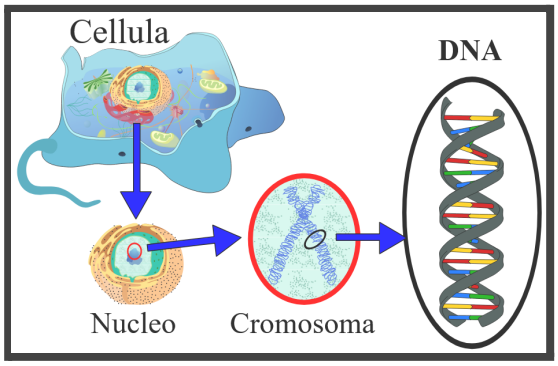
# ScGenomics

Chromosomes topological domains role

in gene expression variability

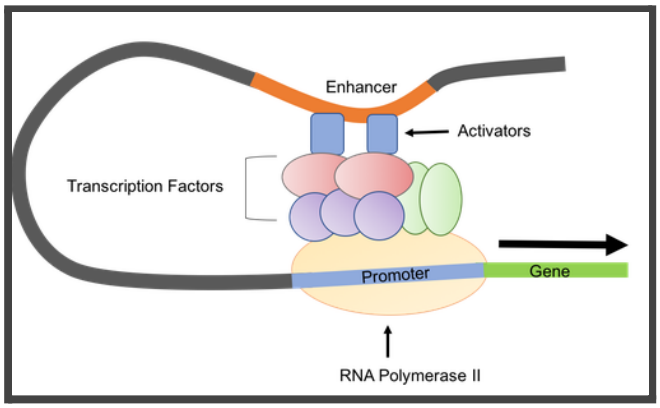
# I - Context

## Eukaryotic genomes



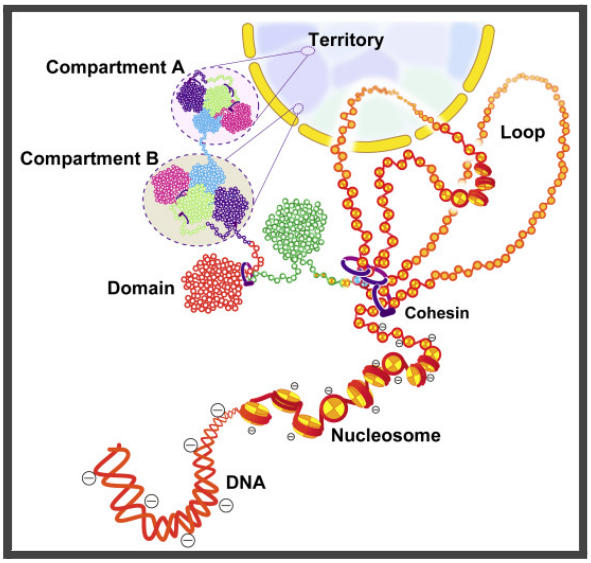

All cells have the same DNA but differentiated cells do not express the same genes. Gene expression is directed by many factors:

- ★ Enhancer sequences
  - Activators
- ★ Transcription factors
  - Proteins that regulate the transcription of genes
- ★ 3D structure of DNA
  - Including TADs (whether active or inactive)
- ★ Compartments
  - Open euchromatin
  - Compact heterochromatin
  - Facultative heterochromatin
- ★ Epigenetic marks
  - H3K27me3 = heterochromatin




*(A RNA polymerase (RNAP), or ribonucleic acid polymerase, is a multi subunit enzyme that catalyzes the process of transcription where an RNA polymer is synthesized from a DNA template.)*

Additionally, these different elements are or can potentially be more or less linked and can diverge according to cell types. Today, it's unknown how closely these elements driving gene expression are linked, nor how much they influence gene expression. By the way epigenetics correspond to mechanisms modifying in a reversible, transmissible (during cell divisions) and adaptive manner the expression of genes without changing their nucleotide sequence.

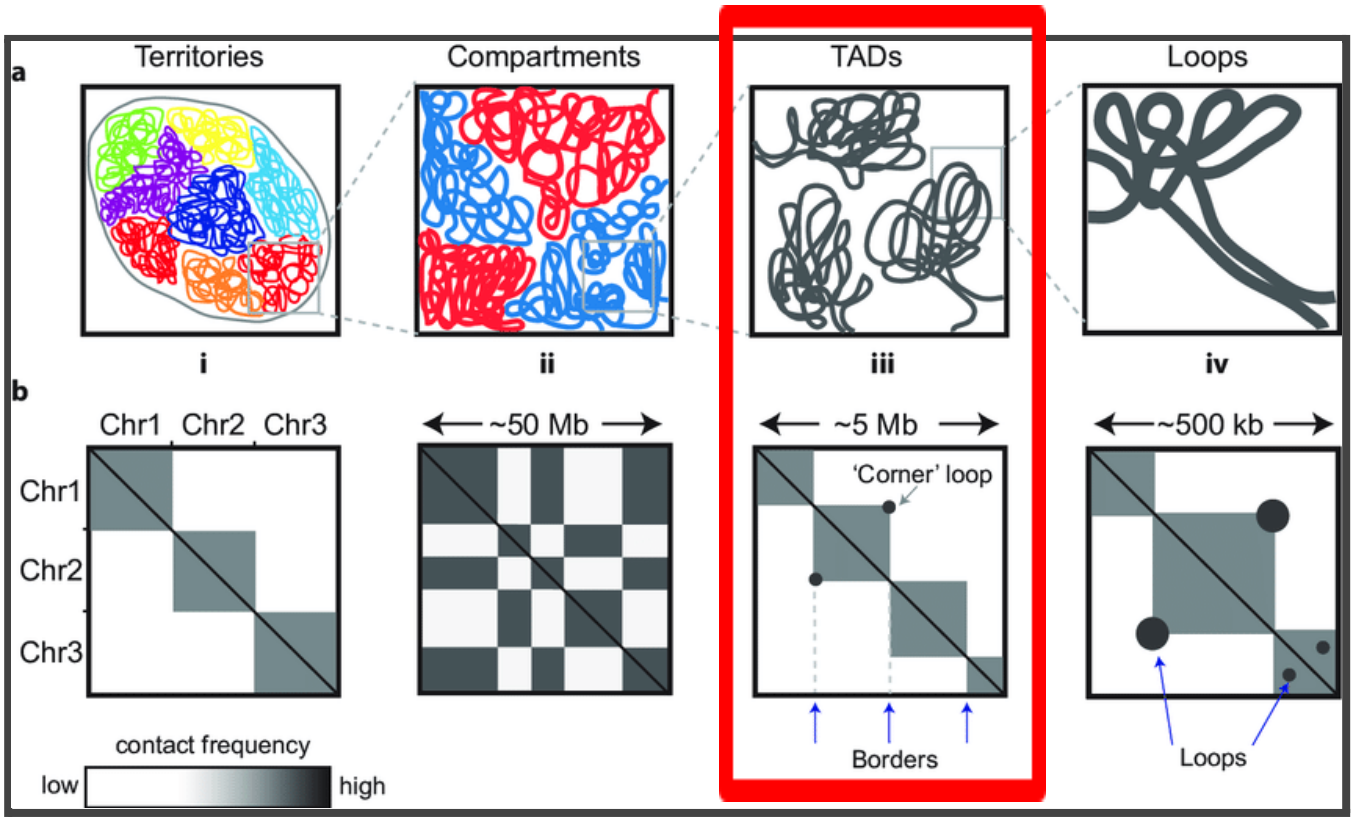



Also, eukaryotic genomes encode genetic information in their linear sequence. And as a matter of fact, mammalian genomes are composed of roughly three billion base pairs that code for the instructions needed for cellular function.

Stretched end to end, the genome is equivalent to almost two meters in length; yet, it is housed in the cell nucleus whose diameter is at the micrometer scale. So to achieve such a level of compaction, a multi-layered genome organization structure is essential.

## What's a TAD (Topologically Associating Domain)?

What's more, three-dimensional genome organization can be characterized at different scales. Globally, chromosomes exist in discrete territories in the cell nucleus. On a sub-chromosomal scale, chromatin physically compartmentalizes into topologically associating domains. Those TADs are megabase-long genomic regions that self-interact, but rarely contact regions outside the domain. Then there are DNA loops which won't be discussed today. Thus TADs are fundamental units of chromosome folding and can be termed as high-frequency chromatin interaction domains.




## Why do TADs matter?

So, the human genome is organized into those TADs, which represent contiguous regions with a higher frequency of intra-interactions. The reason why TADs matter is because TADs and TAD-boundary disruptions have been implicated in rare-disease pathogenesis. They could be associated with human limb malformations or some cancers for instance.

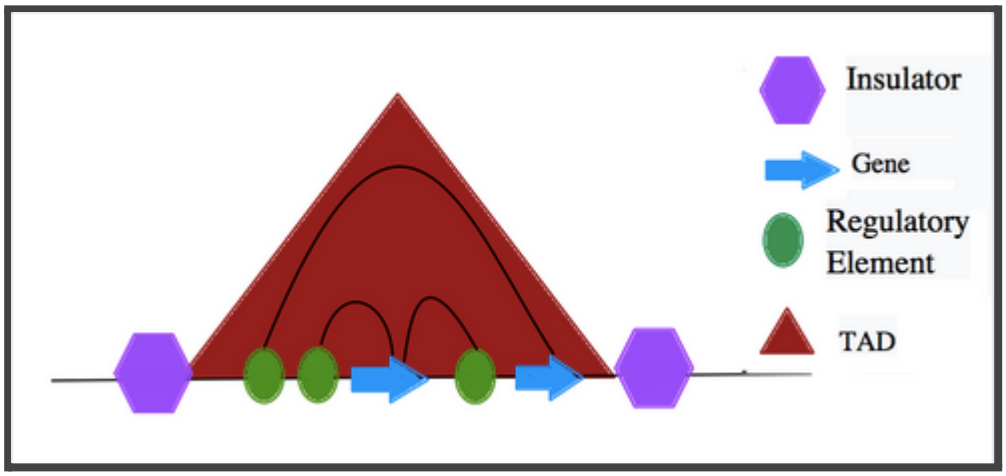
In the concrete, understanding how different attributes of three-dimensional genome architecture influence disease risk is crucial. It lets us interpret human variation and, ultimately, moving from disease associations to an understanding of disease mechanisms.

## Mechanisms underlying TAD formation?

Though the mechanisms underlying TAD formation are complex and not yet fully elucidated. Indeed, the extent to which chromatin three-dimensional topology affects gene expression is still debated. In short the theory consists in the following statement : "TADs contribute to gene expression regulation by restricting the interactions between their regulatory elements.".

## What's within a TAD?

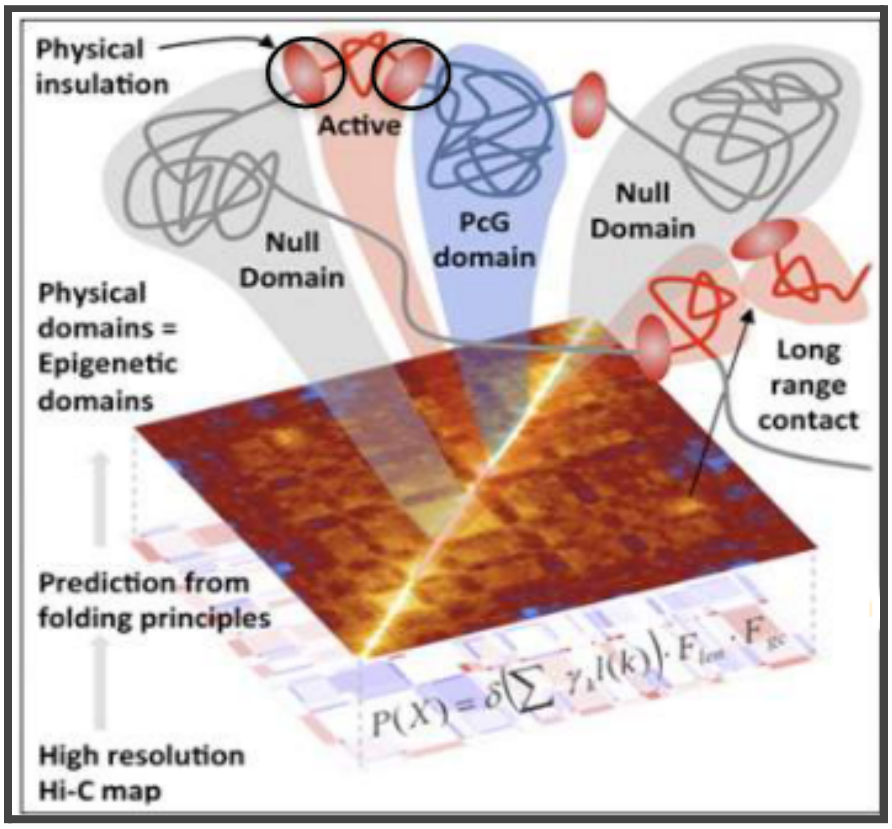
TADs are self-associating, loop-like domains which contain interacting regulatory elements and target genes. Then, there are TAD boundaries which are insulatory elements that restrict interactions of regulatory sequences, like enhancers, to target genes.




Besides, TADs often contain clusters of co-regulated genes. Plus, TAD boundaries are enriched for housekeeping genes and transcription start sites. But for the rest of the presentation we'll consider the three basic elements : genes, regulatory elements and insulators.

## Another TAD illustration

Below lies a picture which can be divided into two parts. The upper part shows a three-dimensional visual which stands for self-folding elements. There's a clear distinction between the domains and we look closely at the active ones here. The red dots circled in black are insulators. Thanks to the two-dimensional visual below, acquired with the HiC technique, we can interpret further TAD structure.




*Gene activation range*

The luminous diagonal line pictures the chromatin. The square bright spot shows interaction frequency within the TAD whereas the square size exhibits its range. Indeed we can observe the tiny bright spot where starts the arrow labelled "Long range contact". And if we draw two segments between this spot and the diagonal, we can imagine a large triangle.

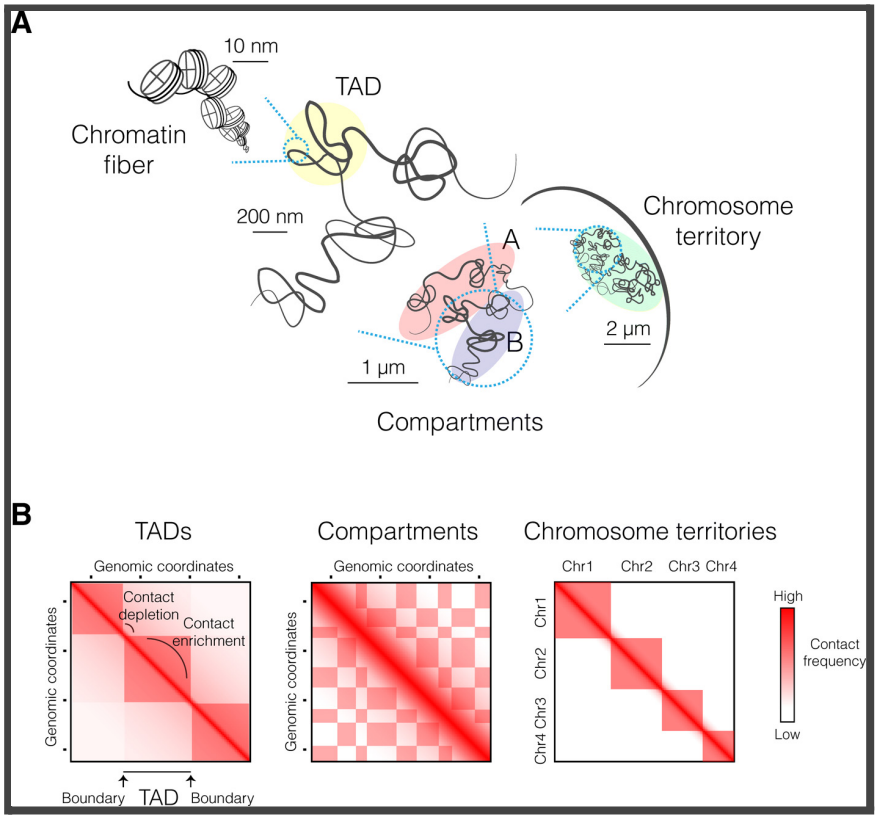
Now we're able to make the link between the position of insulators in both visuals. So this picture with two active domains highlights the fact that gene activation range can be small as well as considerable.

So, a TAD is a genomic region interacting with itself. It means that the DNA sequences within a TAD physically interact more frequently than the sequences outside the TAD. Therefore, a TAD is a genomic regulatory unit.

The chromatin and Hi-C single cell tracing experiments coupled with super-resolution microscopy have shown that there are some variations in TAD structures from cell to cell, that's why the value of single cell work is evident. It involves isolating a single cell and this method provides a better understanding of the function of an individual cell in the context of its microenvironment.
Besides, Hi-C is a genome-wide sequencing technique used to study the conformation of 3D chromatin inside the nucleus.

The organization of chromosomes in TAD is explained by the loop extrusion delimited by the boundary elements, creating domains of interaction relevant for gene regulation.

TADs delimit the gene interaction space in the sense that the activation of genes is dictated by the accessibility of enhancers. Removal of insulators removes activator-promoter interactions. The abrogation (dismantling) of these highly organized structures by removing insulators appears to have a relatively small effect on gene expression. In addition, these structures are of low functional importance in themselves but rather provide information on the activity and organization of regulatory elements.



# II - Main interests

1 - Quantify leaks and gene expression variability

2 - Test if there are correlations between the two variables in wild-type cells?

3 - Test if mutant-type cells have an increasing gene expression variability

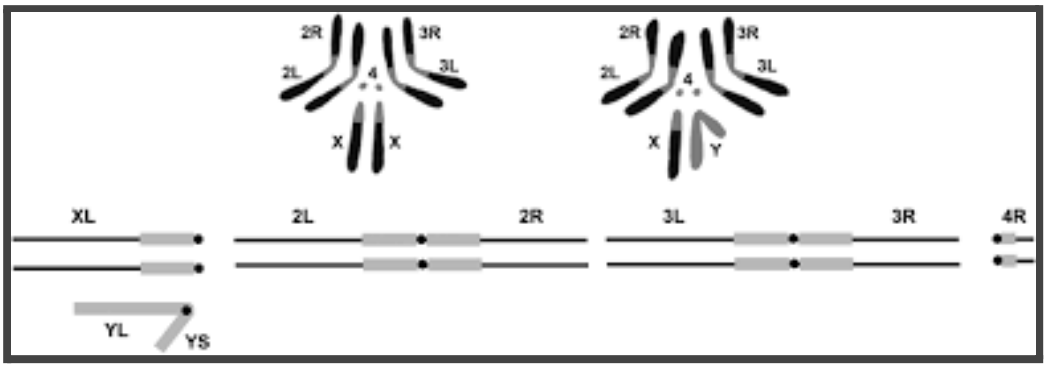
We therefore have three objectives that we set for ourselves, the first is to quantify the leaks and determine the influence on gene expression. Our second objective is to find out if there is a correlation between different wild-type cells.

Eventually, we want to know if in a mutant, therefore a depletion of the insulator, we have an increase or a decrease in the variability of gene expression, we specify here it's about the variability of gene expression and not only the gene expression.

# III - Material and methods

## Model organism: Drosophila Melanogaster

For our single cell Genomics project, the study model organism is Drosophila Melanogaster. It has multiple strengths and almost non-existent drawbacks. The drosophila karyotype is simple and its genome, small in size, has been completely sequenced since the year 2000. In addition, its complexity remains very close to vertebrate organisms.



Besides, in contrast to mammals, the genetic manipulation tools available in flies have allowed the characterization of several proteins that, like the transcriptional repressor CTCF, are capable of inhibiting enhancer-promoter interactions.

Aside, in mammals, TAD-boundaries are occupied by CTCF and the Cohesin complex. Diversely, plants and bacteria lack CTCF homologs but also show TAD-like structures. Hence, it is possible that additional factors are involved in the formation of TADs.

Now in Drosophila, insulators are detected at chromatin accessible regions enriched for active histone marks. What's more, they are occupied by multiple architectural proteins including CP190, Beaf32 and CTCF among others. Also, there are Mes4 and NSD as well as HypB and Set2 which regulate the deposition of repressive histone marks. The mutant-type cells in our project are either mes4-depleted or HypB-depleted.

That being said, for the rest of the report we'll remain to basic terms: insulators, wild-type cells and mutant-type cells.

First of all, we had four large datafiles. The first one being "Genes with NSD" where we could find the variability of gene expression. Then, we had the "Spreading score" which provided us with information on regulatory elements. Next, we had the “Leaks score", where a score is associated with leaks and we mainly found TAD leaks here. The last datafile is the “Fold-Change”, containing the difference in expression between two groups.

Moreover some tools such as *scatter plots* were used to search correlations between leaks and gene expression variability. Plus, we produced *box plots* to study the influence of leaks. What’s more, *Wilcoxon* tests were performed to ascertain if our hypotheses are accepted or rejected.

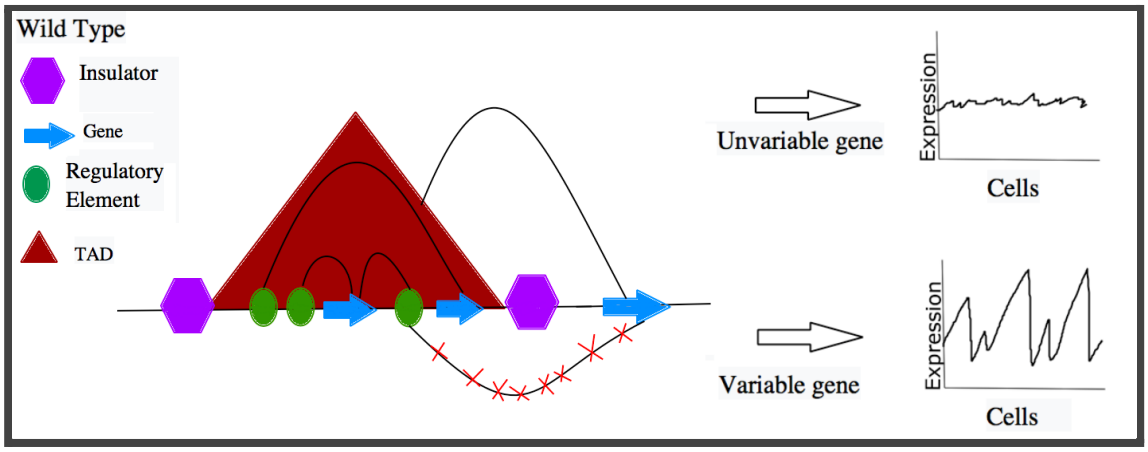
Most of our studies are done in R Studio. We therefore first of all had to associate a leak score to each of the corresponding genes. Nethertheless we encountered an issue in the middle of our work. We have associated several vectors, particularly the NSD vector and the leak vector, however there were times where a gene could be present in the data of the leak vector but not in the NSD vector, and vice versa . Consequently we obtained the so-called NA values. This term means an absence of value, so we decided to remove the genes related to NA from our further tasks.

By the way, single-cell sequencing is a set of molecular biology techniques that allows the analysis of genetic information at the scale of a single cell using next-generation sequencing technologies. In comparison

with traditional sequencing techniques, this technology makes it possible to study cellular differences with optimal resolution and permits to understand the particularity of a cell within its micro-environment. Here it revealed to us the heterogeneity in the cell population. Therefore we noticed the cells' differences.

Below are the two different cell conditions we worked with throughout the project. We decided to use two conditions to answer our questions. Here the first condition:

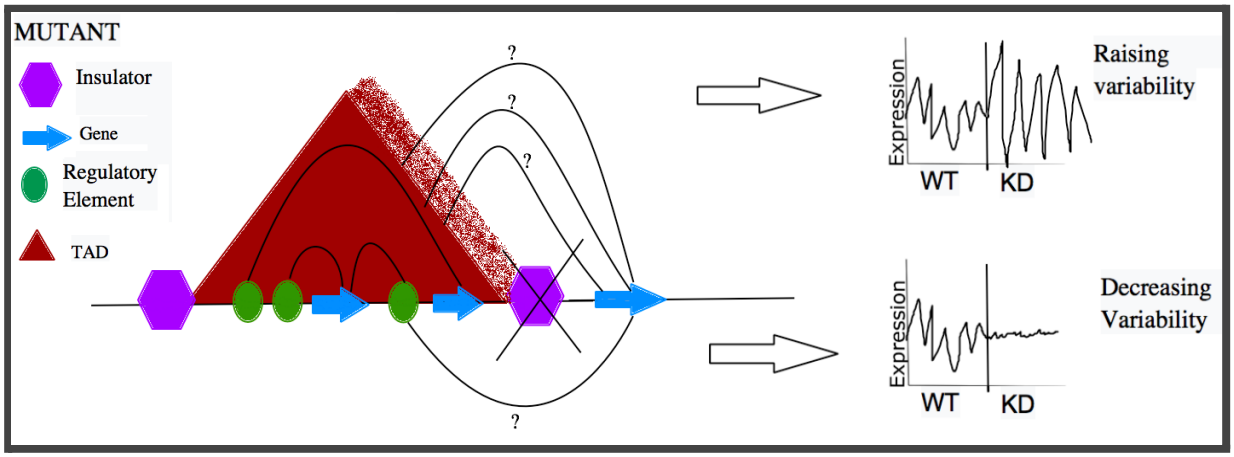
*1st condition*




Remark: we have chosen to represent one single TAD but experimentally the study comported several. In this first condition, we consider wild-type cells. So the TAD is pictured by the red triangle. Within it we have the regulatory elements represented in green which cannot interact outside the TAD. Indeed, the presence of an insulator borders the TAD. We know that there is however a slight leak, and we tried to find out if there is a variability in the expression of genes, as we can see on the right of the diagram.

In mutant-type cells, the insulator is depleted:

*2nd condition*

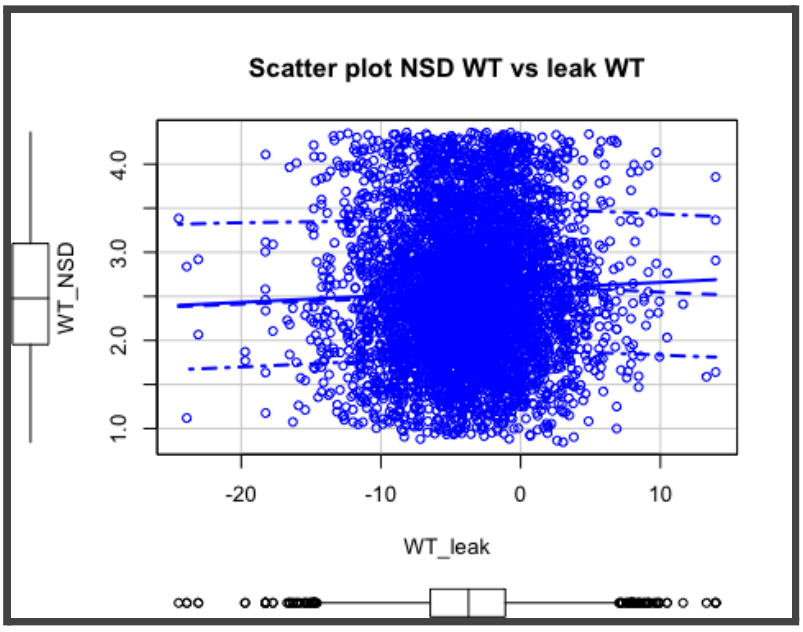


Although, what is a leak ? A leak appears when the regulatory elements can interact with the genes outside the TAD. The above figure depicts a leak represented by the red dot cloud. Here we sought to know, if there are more leaks when we deplete the insulator, and if this increase in leakage will have an increase or a decrease in the expression of the genes as we could observe it on the graphic on the right.

## IV - Results

Our scenario of interest considered long-distance contact. The vector Leak/Gene were constructed by:
- recollecting of all active genes present in the Leak window associated with the loops;
- combining leak scores and genes (each gene possesses its corresponding leak score);
- sorting of genes by leak scores;
- evaluating some correlations between leak scores and NSD / Fold-change.

In statistical terms, we used correlation to denote association between two quantitative variables. The degree of association is measured by a Pearson's correlation coefficient.

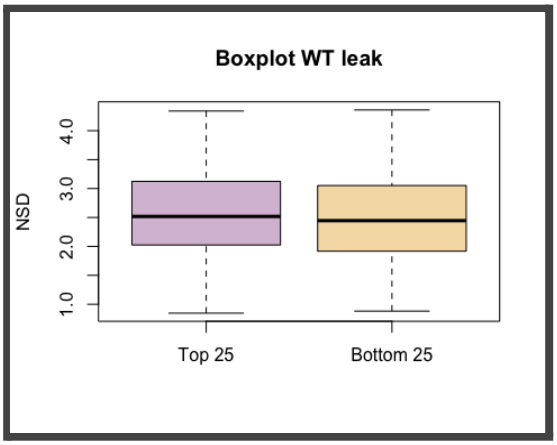


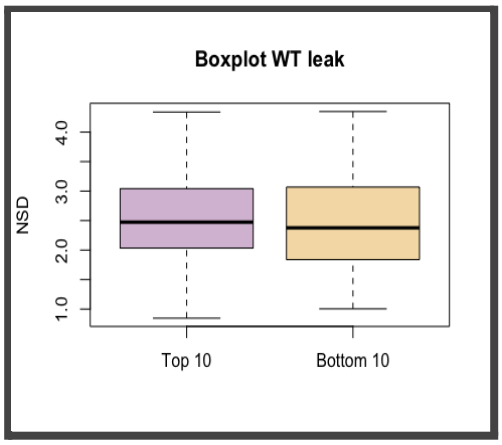

```
cor(WT_NSD, WT_leak, method = c("pearson"))
[1] 0.04051684
```

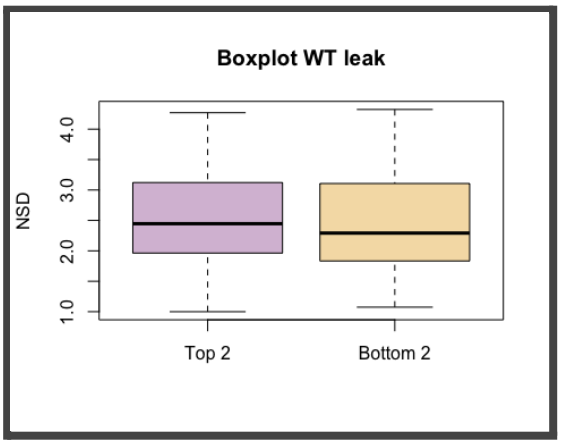
Complete correlation between the two variables should be expressed by either +1 or -1. This value, close to 0, indicates a complete absence of correlation between long distance leaks and gene expression variability for the full set of genes.

1) **Gene expression variability in Wild type cells seems to be influenced by long-distance contact, relating to our data.**

Using all the data, we've been able to find that the two groups containing WT cells with the most (Top 10%) and fewest (Bottom 10%) long-distance contact leaks are statistically significantly different. However, this difference doesn't exist for the Top 2% et Bottom 2% group genes probably due to the low statistical power.

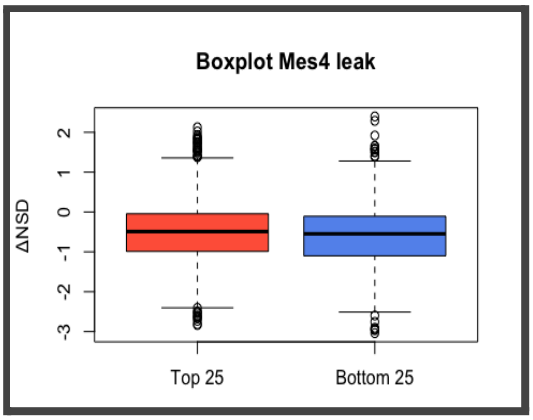


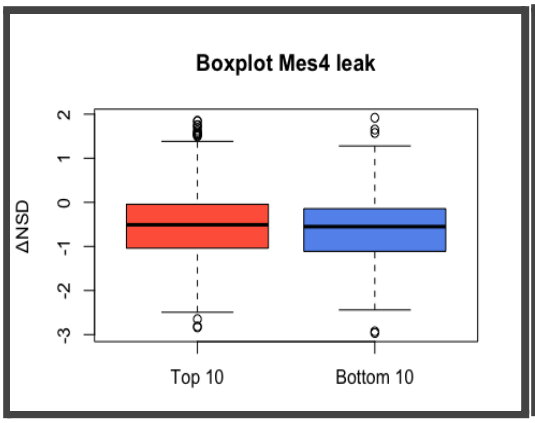



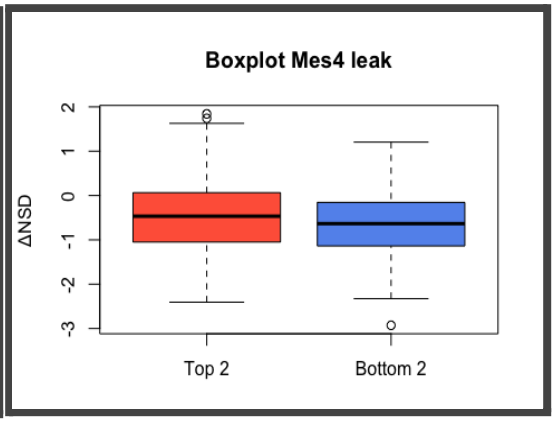




2) **Gene expression variability raising in Mes4-depleted cells does not appear to be influenced by long-distance contact, with regards to our groups of genes.**

No significant differences in ΔNSD have been observed between the two groups containing Mes4-depleted cells respectively with the most (Top xx%) and fewest (Bottom xx%) leaks. Furthermore, various levels of analysis are necessary. Indeed, we have found some differences in groups containing a high quantity of data. But those have not been confirmed by other tests on a lower quantity of genes.

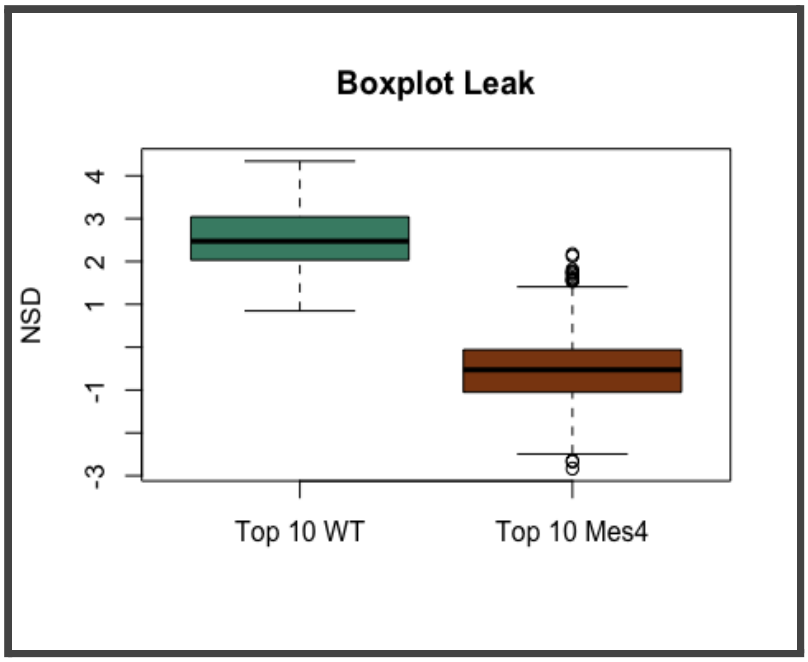


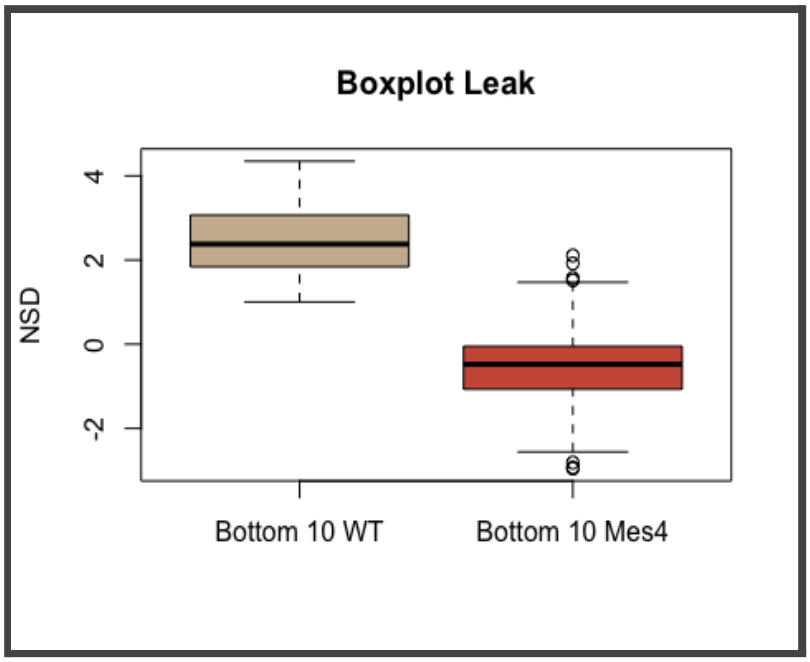




3) **The depletion of a boundary element seems to have an influence on gene expression variability, according to our data.**

After testing the two cell conditions independently, we chose to compare WT to Mes4-depleted groups containing 1000 genes. What we found is that WT and Mes4-depleted groups are significantly different which is statistically confirmed. However, gene expression variability is similar for high and low leaks scores.



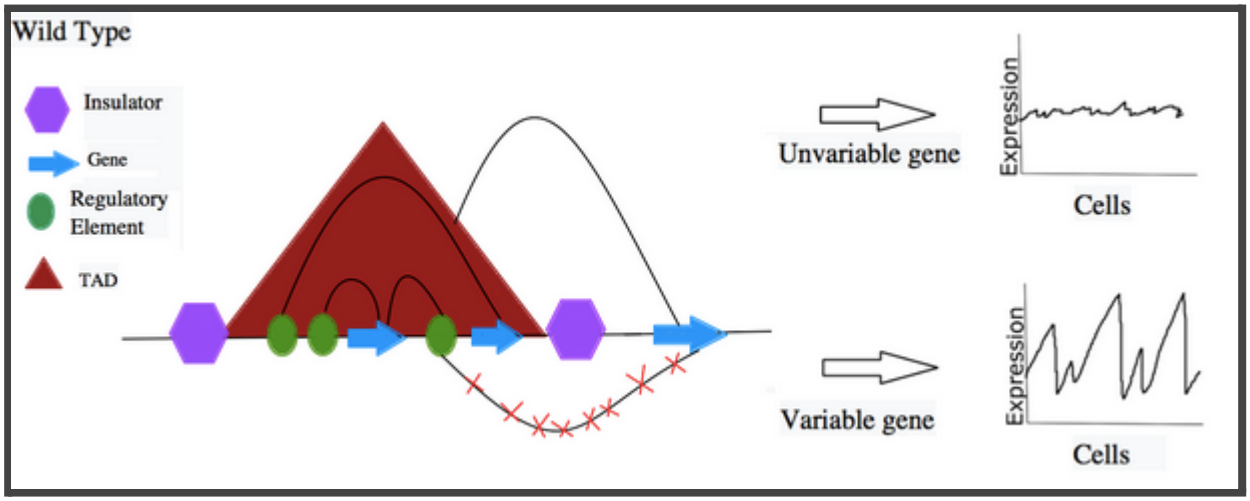




```
data:  WT_top10 and Mes4_top10
W = 265742, p-value < 2.2e-16
```

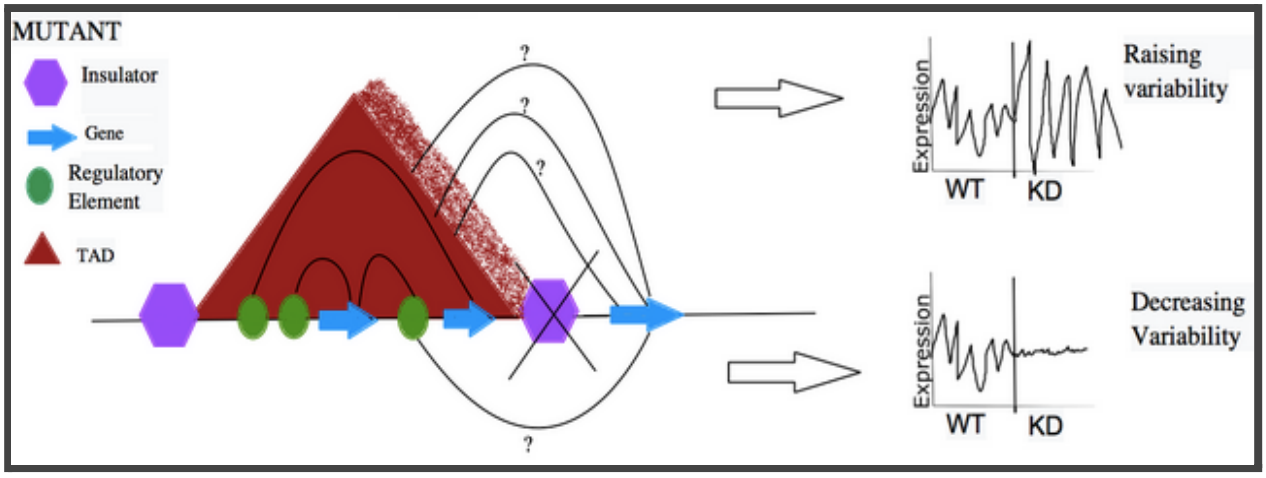
```
data:  WT_bottom10 and Mes4_bottom10
W = 318629, p-value < 2.2e-16
```

# V - Conclusion

To conclude our single cell Genomics project, TADs are definitely high-frequency chromatin interaction domains. According to the results, insulators help to reduce leaks. But we couldn't observe a change of gene expression variability in leaks.




Firstly, we noticed that gene expression variability in wild-type cells seems to be influenced by leaks. What's more, in wild-type cells, genes with more leaks have a greater gene expression variability than genes with fewer leaks.




Then we perceived that raising gene expression variability in mutant-type cells doesn't appear to be influenced by leaks.

To pursue, we've been able to tell that leaks influence gene expression variability while comparing directly wild-type data to mutant-type data. Eventually, we can say that insulator depletion influences gene expression variability.

Lastly, the purposes of this project were exploring raw data using R language, trying to discover correlations between our variables and biologically interpreting our results. So far, the mechanisms underlying TADs formation are not fully discovered.

# VI - References


Single cell analysis pushes the boundaries of TAD formation and function - Jennifer M Luppino and Eric F Joyce

The relationship between genome structure and function - A. Marieke Oudelaar and Douglas R. Higgs

The epigenetic basis of cellular heterogeneity - Benjamin Carter and Keji Zhao

Insulator-mediated 3D chromatin looping : New co-factors, new roles, new methods - Dépierre*, Perrois*, Schaak, Heurteau et al. Cuvier

CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression - Gang Ren, Wenfei Jin, Kairong Cui, ..., Zhiying Zhang, Daniel R. Larson, Keji Zhao