Determination of DNA Active and Inactive Compartments Using Intra- and Inter-Chromosomal Hi-C Contact Maps

SUMMARY

INTRODUCTION

Team presentation

Our team SB-3 is composed of five French students from the Sorbonne University. Our two technical experts, Adrien and Valentin, led the coding process and were real drivers of the project throughout the whole semester. Corentin and Julien, the scientific experts, have been essential in understanding the biological implications of our subject and the significance of our results. Jeanne, manager and writer of the team, was the anchor of the group who made the teamwork possible with her scheduling talents, and the writer of this report.

We were paired with the SB-1 team, composed of Oktawia Scibior, Mikal Daou, and Maxime Gueudré.

Acknowledgements

We would like to thank our teachers for their thorough help during this project, especially Leopold Carron without whom we could not have written a single line of functioning code, Elodie Laine for her clarity on what was expected of us and her guidance, and Juliana Bernardes for keeping a close and benevolent eye on our progress. We would also like to thank Carl Herrmann, Alessandra Carbone, Andres Quintero and Magdalena Machnicka for the organization of this course.

Objectives and context of the subject

For this year's issue of Meet-Eu, we worked on DNA compartments and more precisely on how to determine them thanks to Hi-C contact maps. This subject is indubitably one of the rising topics of today's Biology: Ten years ago in 2011, 6 articles were published on PubMed that contained "Hi-C" in their title. In 2021, there were 506 such articles. Moreover, in 2009, studies constructed spatial proximity maps of the human genome with Hi-C at a resolution of 1 megabase [1], whereas today it is possible to work at a 25 kilobase resolution with our personal computers. Those advances in cell imaging are making it possible to better understand the importance of DNA compartments in cellular processes and gene expression.

Our work is largely based on an article by Rao *et al.* [2], in which the authors interest themselves in the spatial distribution of DNA in the cell nucleus. Indeed, the spatial distance of two DNA regions in the cell nucleus does not always reflect how far apart those regions are along the unfolded DNA sequence. Thus, regulatory elements such as gene enhancers or promoters can exert their influence on regions of the DNA genetically distant from them. Therefore, it seems relevant to assume that DNA regions that are geographically closer together, *i.e.*, that interact preferentially together, are probably functional and regulatory clusters.

To assess this hypothesis, Rao *et al.* generated Hi-C contact maps, that is contact matrices that indicate how often two regions of DNA have been observed "touching" each other. They noticed that the compartments determined solely based on these interaction patterns exhibited distinct genomic and epigenomic content, and that the distribution of histone marks was more correlated to those compartments than with the position within the DNA sequence. These compartments seem to correlate strongly with gene expression, and for instance in *Arabidopsis thaliana*, less packaged domains are enriched in activating epigenetic marks such as H3K4me3, while more densely packaged domains are enriched with inactivating epigenetic marks such as H3K27me3 [3]. It has also been shown that in drosophila and several other eukaryotes, the transcriptional state is a major predictor of Hi-C contact maps [4].

Based on this literature, we can propose our definition of a compartment: a cluster of interacting (geographically close) DNA regions that display a same tendency for gene expression; the compartment with the highest gene density (A) is the active one, whereas the other (B) is the inactive one. We can interpret those as euchromatin and heterochromatin.

We will determine the active and inactive compartments for Hi-C data from five human cell types, firstly by considering intra-chromosomal interactions, then by enlarging our code to account for inter-chromosomal interactions. Our work aims at comparing the intra- and inter-chromosomal methods, and notably at assessing if they lead to the same results.

Workflow figure

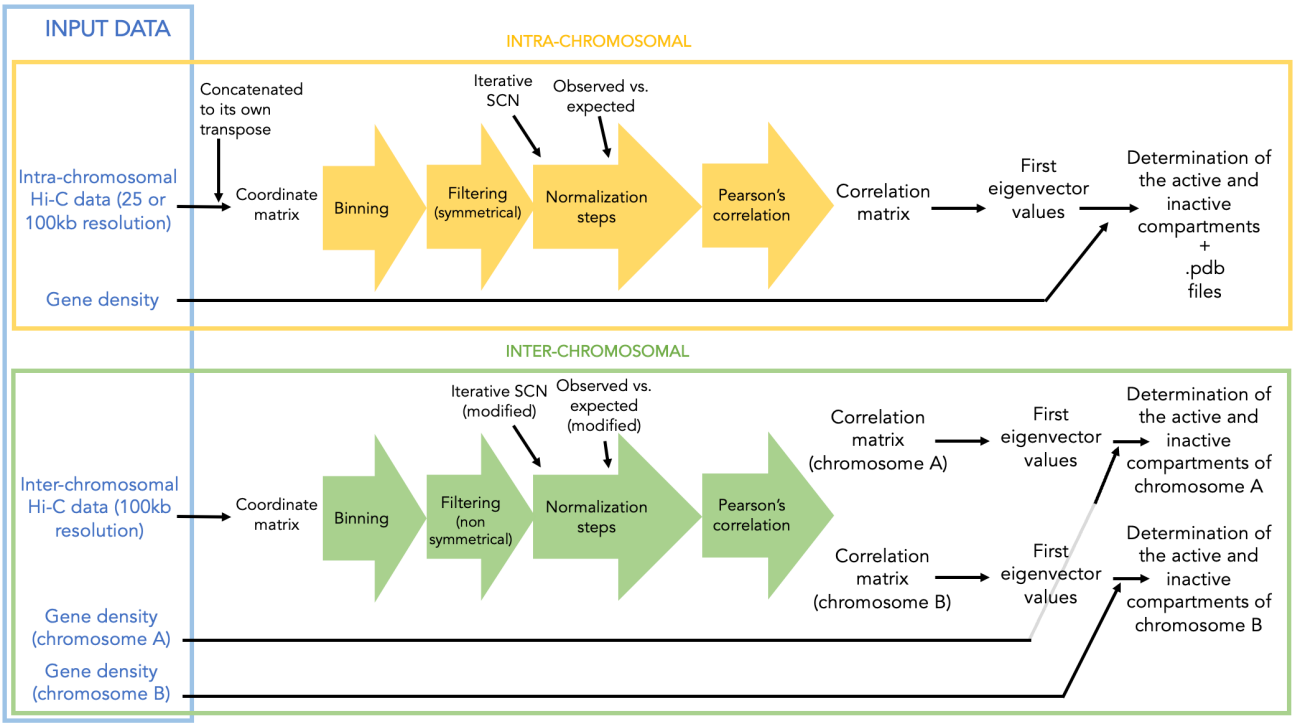

Figure 1: *Workflow figure. The blue box corresponds to the input data, the yellow one to the intra-chromosomal pipeline, and the green one to the inter-chromosomal pipeline.*

MATERIAL AND METHODS

Available Data

We used our pipeline on 5 human cell types. For each cell type, intra-chromosomal Hi-C maps were available at both 25kb and 100kb resolution, for 23 chromosomes (chromosomes 1 to 22 and chromosome X). As for inter-chromosomal Hi-C maps, they were available for all these chromosomes at 100kb resolution. Our cell types are the following:

- GM12878 is a lymphoblastoid cell line (mesoderm cell lineage)
- HMEC corresponds to the human mammary epithelial cell line (ectoderm)
- HUVEC stands for Human Umbilical Vein Endothelial Cells (mesoderm)
- IMR90 is lung fibroblast cell line
- NHEK represents epidermal keratinocytes (ectoderm)

IFB core cluster

The IFB (Institut Français de Bio-informatique) core cluster is a computing cluster composed of 4300 hyper-threaded cores. It allows registered users to use its computing power for projects related to bioinformatics. User accounts are created on demand and are free. The use of this cluster allowed us to generate our results on all available data quickly. For instance, for all chromosomes of all cell types, at both 25kb and 100kb resolution, it took 1 day and 10 hours to generate all intra-chromosomal compartments and the associated .pdb files. For the inter-chromosomal compartments, it took 7h50, because there was only one available resolution (100kb) in the input data, and because we didn't generate .pdb files.

Pipeline

Our pipeline, especially the intra-chromosomal part, is very similar to Leopold Carron's. We will therefore not dwell on this part of the code, even though we will describe the major steps. We will mainly focus on the adaptations we have made to study inter-chromosomal contacts.

One of the first steps of our work was to code an algorithm enabling us to retrieve all the input data file names, in order to automatize our process, once launched on the IFB computing cluster.

Firstly, the Hi-C matrix is loaded as an integer matrix, and it is reformatted into a coordinate matrix. In the intra-chromosomal process, it is concatenated to its own transpose to turn it into a symmetrical square matrix (in the intra-chromosomal case, [i;j] and [j;i] have the same value). Whereas, in the inter-chromosomal case, the Hi-C data already contains information about [i;j] and [j;i], which may not have the same value, so the concatenation is not necessary.

The matrix is then transformed into a binned map and the bins undergo a filtering step (the bins are either 25kb or 100kb long, according to the resolution). In the intra-chromosomal case, a bin is filtered (*i.e.*, removed) if it has no contact at all (this corresponds to empty lines or columns) or if, on the contrary, it has too many contacts and is a clear outlier. In the inter-chromosomal case however, this filtering step has to be non-symmetrical, as one coordinate x can correspond to an empty line for the chromosome placed "vertically" in the matrix, and also to a non-empty column for the chromosome placed "horizontally". The saved and removed bins for each chromosome are therefore not the same, and the modified filtering function designed for inter-chromosomal analysis must therefore generate not one but two lists of saved bins.

Then the matrix is normalized through two steps: the Sequential Component Normalization (SCN) and the Observed vs Expected correlation. The SCN consists in applying an iterative correction on the rows and columns of the matrix so that the sum of the contacts of each row and each column is the same. In the inter-chromosomal case, we modified the function so it doesn't multiply the matrix with its own transpose, because the matrix is most likely not squared and we want to keep it that way, so as to not lose data. Then, since regions of DNA that are closer along the sequence have a greater chance of being geographically close together, we divide each point by the average value of the diagonal on which it is located. This Observed over Expected matrix is also generated by a slightly modified function in the inter-chromosomal case, so as to consider that the matrix is not square.

The correlation matrix (or matrices, in the inter-chromosomal case) is computed with Pearson's correlation. From now on, we will only describe the intra-chromosomal pipeline, but the inter-chromosomal process is the same, except that it is applied to two correlation matrices. Indeed, the data is correlated for each chromosome from its "point of view".

The eigenvalues and eigenvectors are computed for the correlation matrix, and an ACP coupled with the gene density analysis allows us to identify the two compartments, A for active and B for inactive. In the intra-chromosomal case, we also generate .pdb files that allow for a visualization of the compartments in 3D.

Our script is designed to be launched on the IFB core cluster. Details on how to run our code on the cluster are available in the README of our GitHub page : https://github.com/meet-eu-21/Team-SB3

Evaluation of our results

Our pipeline generates, among others, the following files:

- In the case of the intra-chromosomal pipeline, a text file as long as the binned chromosome, with one number per line: -1.0 if the bin has been filtered (*i.e.*, there was no contact), 0.0 if it belongs to compartment A (active), and 1.0 if it belongs to compartment B (inactive).
- In the case of the inter-chromosomal pipeline, two such files, one for each of the studied chromosomes.

We used these files to compute the similarity of our results with those of Leopold Carron, which we consider as gold standard, and also the similarity of our results with those of the SB1 team. The similarity value is computed, for each chromosome, as the percentage of bins for which our conclusion (either filtered, active or inactive) is the same as either the gold standard or the SB1 team's results.

RESULTS

Intra-chromosomal compartments

We obtained the results for the intra-chromosomal compartments in 1 day and 10h on the IFB core cluster. An example of compartment determination is shown in figure 2. The correlation matrix illustrates two clusters, one with positive first eigenvector values in red, and the other with negative values in blue. The gene density represented above the matrix allows us to determine that the red compartment is the active one (A), and the blue compartment is the inactive one (B). A visual validation of our results is given in figure 3, where 3 examples of .pdb files generated through our pipeline are represented, and the spatial segregation of the two compartments is clear.

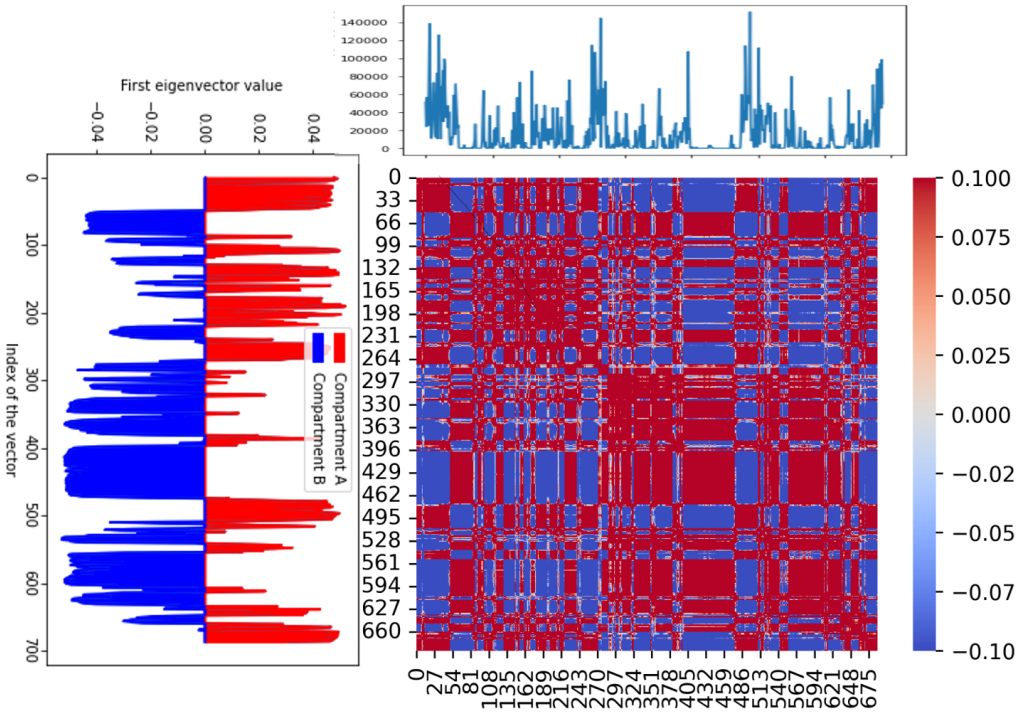

Figure 2: *Intra-chromosomal compartments of chromosome 16. The curve on top is the gene density of the chromosome. On the left, the first eigenvector values of the matrix. The compartment A (active) is in red, compartment B (inactive) is in blue. This is for cell type GM12878. Resolution : 100kb.*

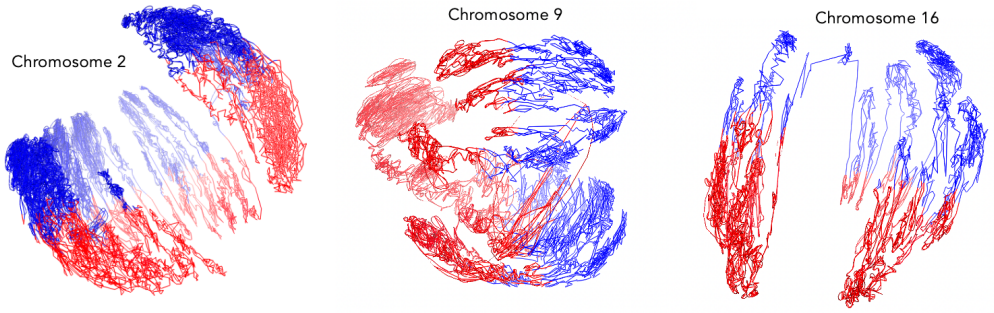

Figure 3: *Examples of .pdb files generated by the intra-chromosomal pipeline (for cell type GM12878). The A (active) compartment is colored in red, the B (inactive) compartment is colored in blue.*

To further validate our results, we compared them to Leopold Carron's gold standard, as described in "Evaluation of our results", page 5. The results of this comparison are shown on figure 4. The global performance of our pipeline to find similar results as the gold standard is slightly above 80% for all cell types, at both 25kb and 100kb resolution. However, this performance varies widely depending on the chromosome and chromosomes 1 and 2 for instance show poor similarity to the gold standard compared to the others. One hypothesis could be that this is due to them being longer than the other chromosomes, but we do not observe a correlation between chromosome size and similarity to gold standard otherwise.

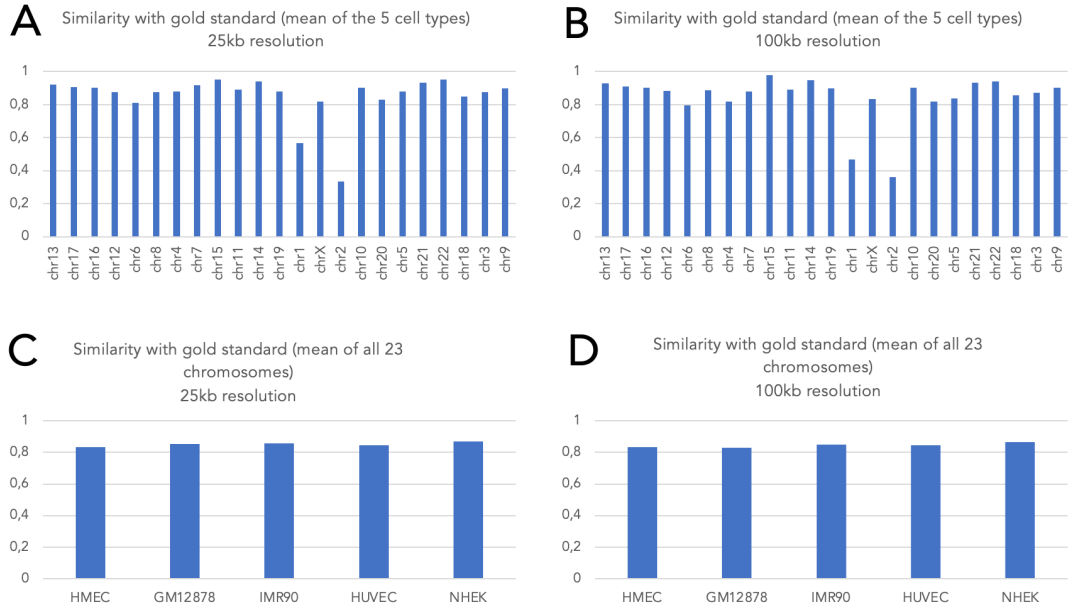

Figure 4: *Similarity or our intra-chromosomal compartments with Leopold Carron's gold standard compartments. A and B: similarity of the compartments per chromosomes (at respectively 25kb and 100kb resolution). C and D: similarity of the compartments per cell type (at respectively 25kb and 100kb resolution)*

Inter-chromosomal compartments

We obtained our results for the inter-chromosomal compartments in 7h50min on the IFB core cluster. An example of intra-chromosomal vs. inter-chromosomal compartment determination in shown in figure 5. The results are different of course, and in this case the right side of the figure has been generated via the interactions of chromosome 16 with chromosome 15. Had it been with another chromosome, we would have found yet another result. The first eigenvector values are not identical in both cases but there seems to be common trends.

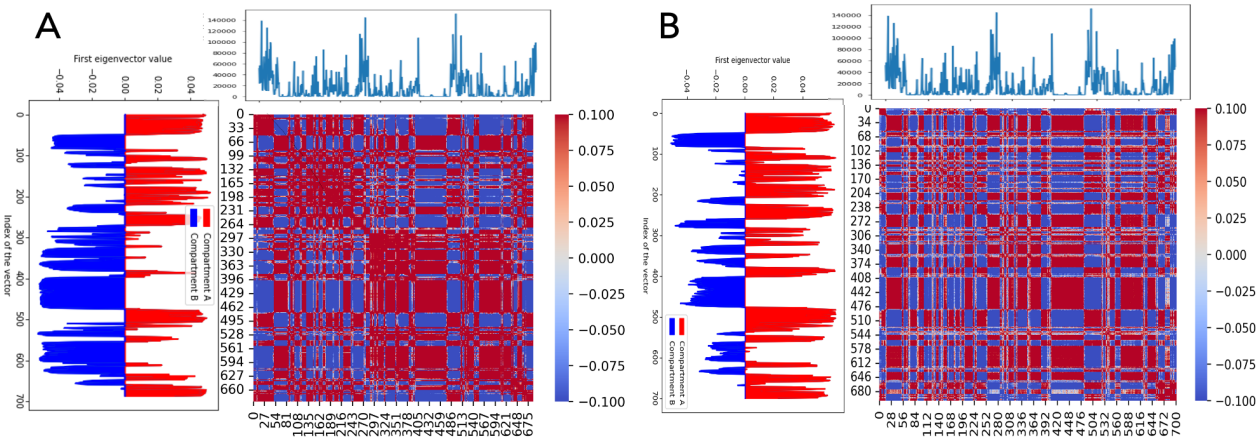

Figure 5: *A: Intra-chromosomal compartments of chromosome 16 (similar to figure 2). B: Inter-chromosomal compartments of chromosome 16, generated with its Hi-C data relative to chromosome 15. On top is the gene density of chromosome 16, and on the left the first eigenvector values of the matrix. The compartment A (active) is in red, compartment B (inactive) is in blue. Resolution : 100kb.*

The similarity of our inter-chromosomal compartments with the gold standard is given on figure 6. For each cell type, the value on line i and column j gives the percentage of similarity between the intra-chromosomal gold standard for chromosome i, and our inter-chromosomal results for chromosome i, in interaction with chromosome j. The performance of our inter-chromosomal pipeline is globally as good as the intra-chromosomal one, and we observe that chromosomes 1 and 2 are once again the ones with the poorest similarity. In this case, we can see that the percentage is systematically lower for chromosome 1 when it is paired with chromosomes 11, 12, 14, 16, 17 and 18, and for chromosome 2 when it is paired with chromosomes 19 and 21. One possible explanation would be that those chromosomes are far apart in the cell nucleus, and therefore had very little contacts, not enough to properly assign the regions of those chromosomes to a compartment or another.

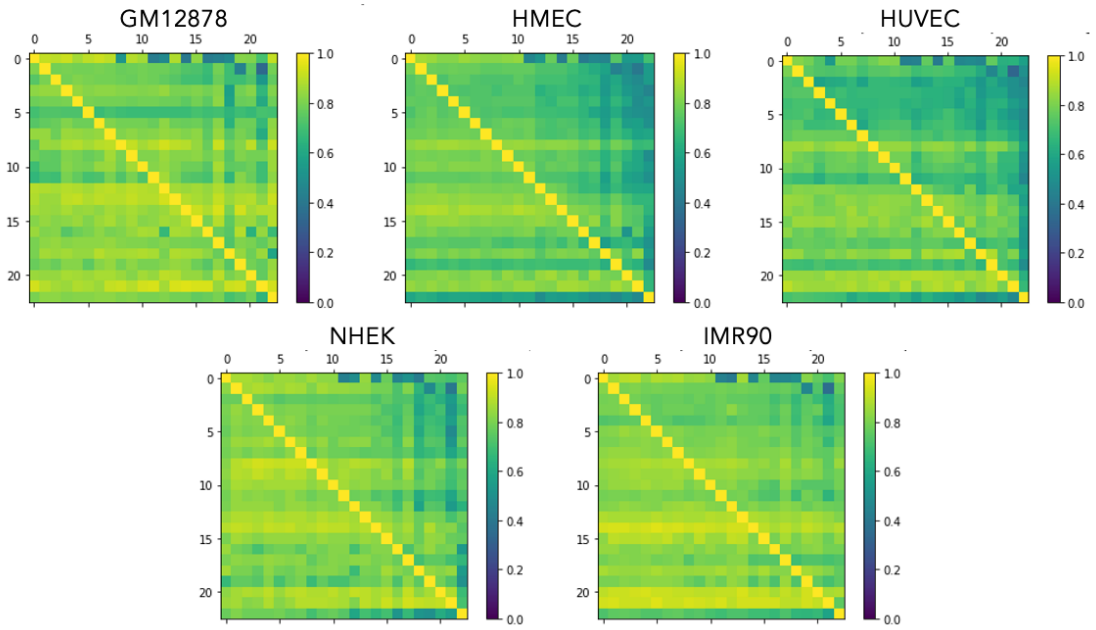

Figure 6: *Similarity between our inter-chromosomal compartments and Leopold Carron's gold standard. For each cell type, the value on line i and column j corresponds to the percentage of similarity between our compartments found by the inter-chromosomal pipeline, for chromosome i in interaction with chromosome j, and the gold standard for chromosome i. The values on the diagonal have been arbitrarily set to 1. Chromosome X is labeled as 23.*

DISCUSSION AND COMPARISON

Comparison with the other team

To assess the work and results of the SB1 team, we firstly compared the compartments they had found to the gold standard and to our own compartments, as described in "Evaluation of our results", page 5. We only compared intra-chromosomal compartments, as the SB1 team did not work on inter-chromosomal ones. These comparisons for two cell types are given in figure 7.

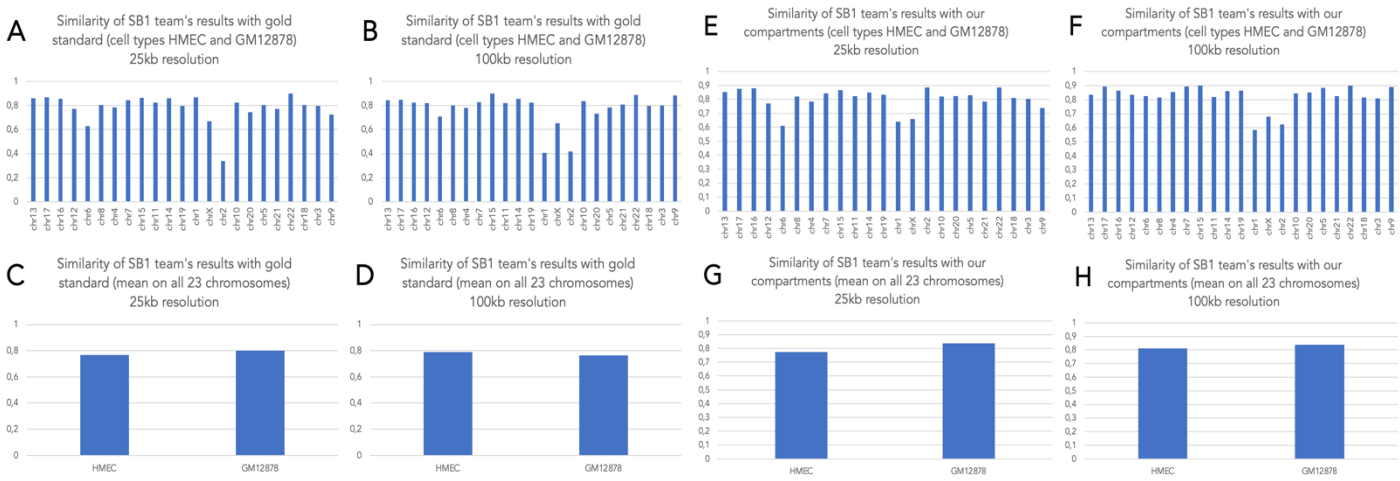

Figure 7: A, B, C and D: *Similarity of intra-chromosomal compartments of the SB1 team with Leopold Carron's gold standard compartments, for cell types HMEC and GM12878. A and B: similarity of the compartments per chromosomes (at respectively 25kb and 100kb resolution). C and D: similarity of the compartments per cell type (at respectively 25kb and 100kb resolution)*
E, F, G and H: Similarity of intra-chromosomal compartments of the SB1 team with our results, for cell types HMEC and GM12878. E and F: similarity of the compartments per chromosomes (at respectively 25kb and 100kb resolution). G and H: similarity of the compartments per cell type (at respectively 25kb and 100kb resolution)

Their results are approximately as good as ours in term of similarity to gold standard, and as a matter of fact our results are closer together than they are to Leopold Carron's. We even seem to find the same "mistakes" concerning chromosomes 1 and 2, as they are also far from the gold standard like we are for these two, but the comparison between us gives better results. This is obviously due to our methods being quite similar, as both teams' work is based on Leopold Carron's code. Notably, the preprocessing and normalization steps at the beginning of our pipelines are mostly the same.

Our questions were not the same; while we wanted to focus on the inter-chromosomal approach, they oriented their work more towards the determination of the optimal number of compartments with Hidden Markov Models and clustering methods such as K-means, silhouette clustering, hierarchical clustering and spectral clustering. Their work is comprehensive, detailed, and their results conclusive.

Biological significance

Our results show that it is possible to distribute regions of DNA as small as 25kb long into two compartments based solely on contact maps, and that this distribution leads to coherent clusters, both spatially and regulatorily. Notably, the 3D visualizations shown on figure 3 allow to clearly see the two compartments as well-defined geographical clusters, whereas the correlation matrices with the gene density demonstrate that the compartments define two groups of more or less active genes. Those results are consistent with the literature, and we added to it that it is also possible to efficiently determine the compartments based on inter-chromosomal contact maps. This reinforces the idea that spatial proximity is often linked to functional clusters, and in particular to remote gene activation and inhibition circuits.

Challenges and further development opportunities

There are a few possibilities of enhancement of our project that we noticed as we were working on it :

- The way we calculated the similarity of our results to the gold standard (or to the SB1 team's results), by comparing line by line, is very sensitive to the slightest insertion or deletion. By chance, our results seem coherent so it seems like there was no such problem, but a prior alignment, with the Needleman-Wunsch algorithm for instance, could have been a great enhancement.
- In the comparison matrices of figure 6, we arbitrarily set to 1 the value of the diagonals. In fact, the value on the diagonal should be, for each cell type and each chromosome, the similarity score of this chromosome's intra-chromosomal compartments compared to the gold standard.

And then there are many other possible ameliorations that are less relevant and too numerous to write down here, and of course all those that we didn't think about.

Such projects on the organization of DNA and its distribution in the cell nucleus contribute to a finer understanding of genetics, down to the level of the promoter and its target, and open perspectives towards a better comprehension of epigenetics. We know today that several diseases are linked to epigenetics and knowing how DNA can be either expressed or not based on its spatial location and its epigenetic marks seems like one of the most promising questions for the future of genetics.

Meet-Eu has been a pretty unique experience in our student years, for it is rare to conduct a project on such a long period of time. Moreover, working in a team of five is quite a challenge and has been an opportunity to mature our sense of responsibility, organization, and comprehension of the others. We were fortunate to work as a close-knit team, and we got along well throughout this adventure. We regret, of course, that the final presentation could not take place in Heidelberg because of sanitary precautions, but we understand this decision. We would also have preferred to work with a team from another university, but even with its flaws Meet-Eu remains a great experience.

REFERENCES

[1] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct 9;326(5950):289-93. doi: 10.1126/science.1181369. PMID: 19815776; PMCID: PMC2858594.

[2] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. doi: 10.1016/j.cell.2014.11.021. Epub 2014 Dec 11. Erratum in: Cell. 2015 Jul 30;162(3):687-8. PMID: 25497547; PMCID: PMC5635824.

[3] Grob S, Schmid MW, Grossniklaus U. Hi-C analysis in Arabidopsis identifies the KNOT, a structure with similarities to the flamenco locus of Drosophila. Mol Cell. 2014 Sep 4;55(5):678-93. doi: 10.1016/j.molcel.2014.07.009. Epub 2014 Aug 14. PMID: 25132176.

[4] Rowley, Michael & Nichols, Michael & Lyu, Xiaowen & Ando Kuri, Masami & Rivera, Sarahi & Hermetz, Karen & Wang, Ping & Ruan, Yijun & Corces, Victor. (2017). Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Molecular Cell. 67. 10.1016/j.molcel.2017.07.022.

[5] Hu, B., Wang, N., Bi, X. *et al.* Plant lamin-like proteins mediate chromatin tethering at the nuclear periphery. *Genome Biol* **20,** 87 (2019). https://doi.org/10.1186/s13059-019-1694-3

[6] Nicolas Servant, Bryan R. Lajoie, Elphège P. Nora, Luca Giorgetti, Chong-Jian Chen, Edith Heard, Job Dekker, Emmanuel Barillot, HiTC: exploration of high-throughput 'C' experiments, *Bioinformatics*, Volume 28, Issue 21, 1 November 2012, Pages 2843–2844, https://doi.org/10.1093/bioinformatics/bts521

[7] Liu, Y., Nanni, L., Sungalee, S. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat Commun* **12,** 2439 (2021). https://doi.org/10.1038/s41467-021-22666-3

The IFB core cluster:
https://www.france-bioinformatique.fr/cluster-ifb-core/

Leopold Carron's GitHub and GitLab pages:
https://gitlab.com/LeopoldC
https://github.com/LeopoldC

The Bioinfo-fr.net website:
https://bioinfo-fr.net

Our GitHub page :
https://github.com/meet-eu-21/Team-SB3