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

Our team is composed of two technical experts, Adrien Pauron and Valentin Gherdol, two scientific experts, Julien Kot and Corentin Delhay, and a manager and writer, Jeanne Bauduin. Our work aims at efficiently determining DNA compartments based on inter-chromosomal contact maps.

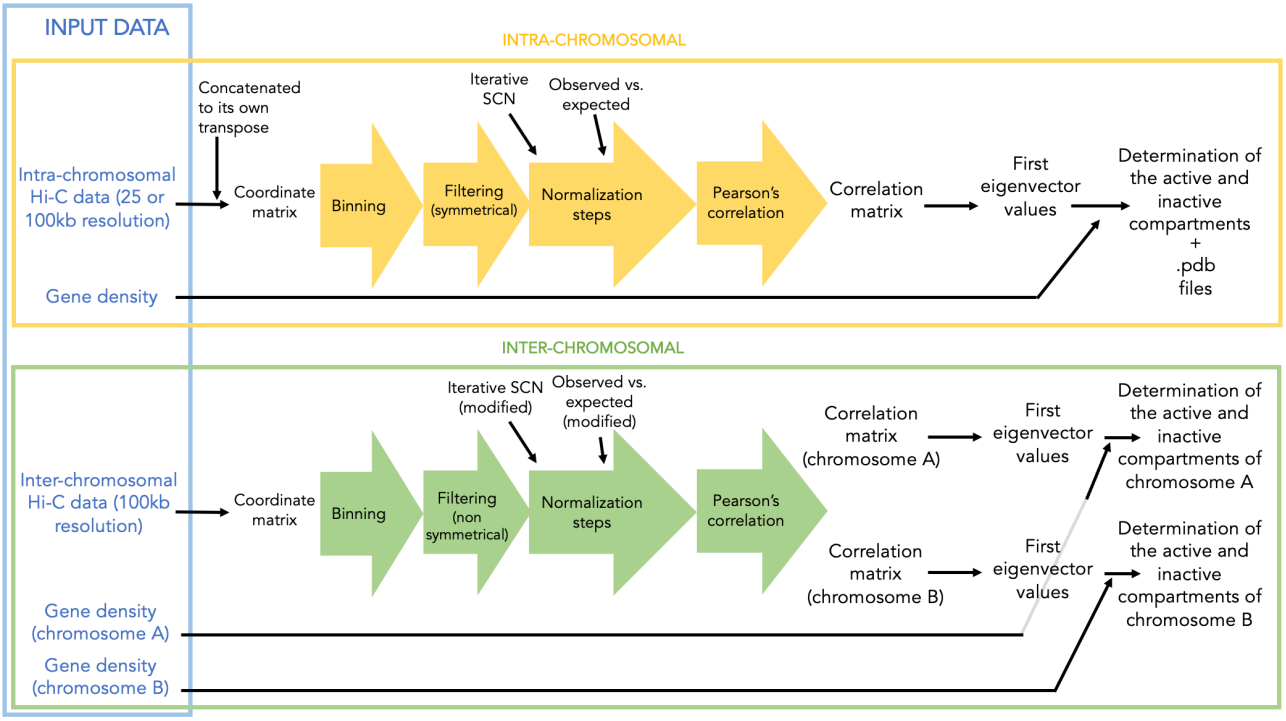

Figure 1: *Workflow figure. The blue box corresponds to the input data, the yellow one to the intra-chromosomal pipeline, and the green one to the inter-chromosomal pipeline.*

Method presentation

We worked on intra-chromosomal and inter-chromosomal Hi-C contact maps for 23 human chromosomes in 5 human cell types, at both 25kb and 100kb resolution. As shown in our workflow figure (figure 1), we built two parallel and resembling pipelines, one for the intra-chromosomal approach and one for the inter-chromosomal approach. The first one is largely inspired by Leopold Carron's work, and the second one is an adaptation of the first to tolerate inter-chromosomal data, notably non-square matrices. In both processes, our Hi-C data is first reformatted into a coordinate matrix, binned, and those bins are filtered, which means the bins that had no contact at all or too many contacts (outliers) are removed. Then the matrix undergoes two normalization steps, a Sequential Component Normalization and an Observed over Expected correlation (with slight differences between intra- and inter-chromosomal pipelines). The computation of the Pearson's correlation gives, according to the case, one or two correlation matrices. The eigenvector values are computed for these matrices, and a gene density analysis allows us to discriminate the active (A) compartment from the inactive (B) one. In the intra-chromosomal approach, we also generate .pdb files to visualize our compartments in 3D.

This whole pipeline is designed to be ran on the IFB core cluster, and the instructions to run it are available on our GitHub page : https://github.com/meet-eu-21/Team-SB3

Results

- We obtained the compartments, the matrices and the .pdb files for all the available data in 1 day and 10h for the intrachromosomal approach. We obtained the compartments and the matrices in 7h50min for the inter-chromosomal approach.
- The intrachromosomal approach resulted in .pdb files that show a clear distinction between the two compartments. Our compartments show, on average, 80% similarity with Leopold carron's gold standards, for all 5 cell types and both resolutions.
- The inter-chromosomal compartments show the same similarity with the gold standard as the intra-chromosomal compartments for all cell types.
- Not all chromosomes display the same similarity; notably, the compartments of chromosomes 1 and 2 are consistently the furthest from the gold standard.

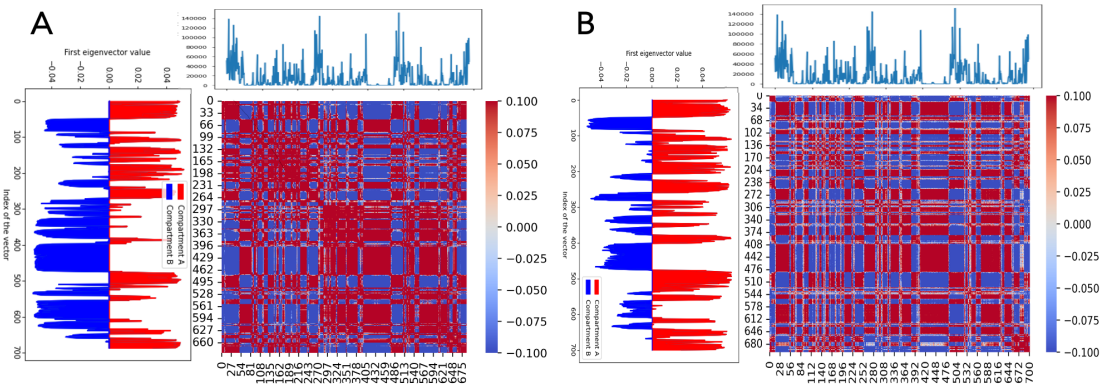

Figure 2: *A: Intra-chromosomal compartments of chromosome 16. B: Inter-chromosomal compartments of chromosome 16, generated with its Hi-C data relative to chromosome 15. On top is the gene density of chromosome 16, and on the left the first eigenvector values of the matrices. The compartment A (active)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

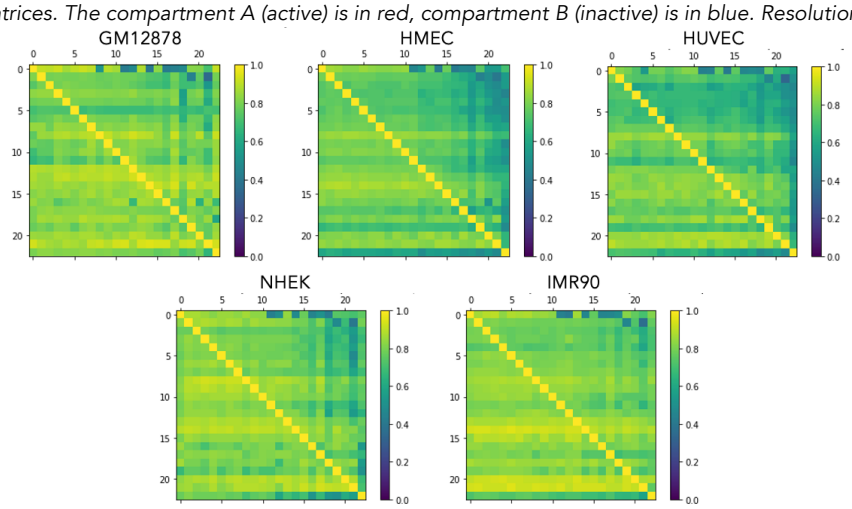

Figure 3: *Similarity between our inter-chromosomal compartments and Leopold Carron's gold standard. For each cell type, the value on line i and column j corresponds to the percentage of similarity between our compartments found by the inter-chromosomal pipeline, for chromosome i in interaction with 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.*

Comparison with team SB1

- The intra-chromosomal compartments found by team SB1 display the same similarity to gold standard as our own compartments.
- The results of team SB1 and ours are closer together than they are to the gold standard, due to a very similar code, especially during the preprocessing of the matrices.
- The team SB1 interested themselves not on an inter-chromosomal approach as we did, but rather on the estimation of an optimum number of sub compartments, via Hidden Markov Models and silhouette metrics (K-means, hierarchical and spectral clustering).