HGEN 47900 - Lab 2: Hi-C

Sébastien Bastide

April 6th, 2023

Molecular methods

Hi-C is a method of the Chromatin Conformation Capture (3C) family. In contrast to its one-to-one (3C) and one-to-many (4C) relatives, it generates a genome-wide (all-to-all) map of chromatin contacts.

Although conceptually simple, Hi-C is multi-step protocol. The main ones are:

1. **Cell fixation.** Cells are typically fixed using 1% formaldehyde for 10 min. This is a light fixation, often used in ChIP-seq, etc. This freezes, through protein bridges, the spatial organization of the genome.

2. **Nuclei extraction.** For *in situ* Hi-C (Rao, et al., Cell 2014), nuclei are extracted. The following steps are performed in crosslinked nuclei in order to reduce the noise arising from free-floating chromosomes.

3. **DNA fragmentation.** DNA is digested, typically with a 4-cutter restriction enzyme (DpnII or MboI). 4-cutters are used for their high cutting frequency. There are $4 \times 4 \times 4 \times 4 = 256$ possible random DNA 4-mers. Thus, a fully digested genome is expected to generate fragments with sizes around 256 bp. Using a 5-cutter would generate fragments that are on average 4 times larger (1024 bp).

   Note: Micro-C uses Micrococcal Nuclease (MNase) in order to produce even smaller fragments ($\tilde{1}50$ bp), thereby increasing resolution.

4. **Fill-in.** In contrast to 3C and 4C, which use the cohesive ends generated by the restriction digest, Hi-C uses blunt ligation. The difference is a Hi-C-specific fill-in step. In this step, reverse complement strands of the cohesive ends left by restriction enzymes are polymerized using the Klenow fragment. Among the nucleotides used for the fill-in, biotin-dATP or biotin-dCTP will enable the pull down on chimeric fragments.

   Note: The Klenow fragment possesses $3' \rightarrow 5'$ exonuclease activity but not $5' \rightarrow 3'$.

5. **Proximity ligation.** Because DNA fragments are maintained in place by crosslinked protein structures, re-ligation of those fragments depends on the physical distance between their ends.

   Note: This is the fundamental principle of the 3C-based methods. Importantly, when we talk about "contact frequency", what we really mean is *ligation frequency*.

6. **Reverse crosslinking and DNA extraction.** Ligation products are incubated at 55°C with proteinase K in order to degrade proteins and then heated to 68 °C to reverse the crosslinking. DNA is then extracted using isopropanol or PCIA extraction.

7. **DNA fragmentation.** The ligation step typically generates fragments that are several kilobases long and cannot be sequenced as is using short read sequencing. Those large fragments are thus sheared using sonication to an average size of $\tilde{3}00$ bp.

8. **Biotin pull down.** Because biotin was incorporated at the ends of the initial digested fragments during the fill-in step, biotin marks the junctions between re-ligated fragments. Therefore, by pulling down on biotin, we enrich the pool of fragmented DNA for chimeric molecules.

9. **Library preparation.** The library preparation is standard and involves: end-repair and A tailing, adapter ligation and PCR amplification.

   Note: The end-repair step uses T4 polymerase which possesses $5' \rightarrow 3'$ exonuclease activity. This means that the polymerase comes and chews on the 5' ends of the double stranded fragments for a bit before starting to polymerize. This activity is crucial to remove biotinylated adenosine and cytosine from the ends of the fragments and prevent the enrichment of non-informative molecules.

10. **Sequencing.** Paired-end sequencing is performed in order to produce read pairs corresponding to chimeric DNA molecules.

Computational methods

Data preprocessing and QC

Hi-C (and micro-C) sequencing data processing involves many steps. Those steps are typically wrapped into end-to-end pipelines. The most popular ones are `Juicer` (produces `.hic` files) and `distiller` (produces `cooler` or `.cool` files).

The general workflow consists in:

1. Align of the paired reads obtained from chimeric DNA molecules.
2. Parse the paired `.sam` file to generate Hi-C pairs (in the `BEDPE` format).
3. Filter out PCR duplicates based on their mapping coordinates and orientation.
4. Aggregate pairs into binned matrices of Hi-C interactions.

The main end result is a heatmap where the x-axis and y-axis represent genomic positions, and each tile represents the frequency of interactions between the two corresponding genomic regions.
It also generates a file with some important statistics about the Hi-C data. The most important ones are:

1. The fraction of mapped reads ($\frac{\text{Mapped reads}}{\text{Total reads}}$). This largely depends on the quality of the genome sequence and the read length. Better genome sequence and longer reads result in more unambiguous mapping.
2. The fraction of duplicated reads ($\frac{\text{Non-duplicated reads}}{\text{Mapped reads}}$). The amount of duplicated reads depends on the number of PCR cycles used for the final library amplification. This can be optimized using qPCR to catch the middle of the exponential phase.
3. The *cis*/*trans* ratio ($\frac{\textit{Cis} \text{ read pairs}}{\textit{Trans} \text{ read pairs}}$) or fraction of *cis* pairs ($\frac{\textit{Cis} \text{ read pairs}}{\text{Total read pairs}}$). Experiments that aim at uncovering local structures (such as TADs and loops) ideally have high fractions of *cis* pairs ($\geq 75\%$). This depends on several factors:

   Biological: the size and number of chromosomes (larger, more numerous chromosomes will produce more *trans* interactions).

   Technical: the integrity of the nuclei (*in situ* Hi-C), the crosslinking time and efficiency, the ligation efficiency.
4. The *cis* short/long-range ratio ($\frac{\text{Short-range read pairs}}{\text{Long-range read pairs}}$) or fraction of long-range read pairs ($\frac{\text{Long-range read pairs}}{\text{Total read pairs}}$). This largely depends on the crosslinking time and efficiency.

One of the most studied characteristics of 3D genome folding is the relationship between frequency of interaction and distance. Under a simple polymer model (no loop extrusion), this relationship is predicted to decrease exponentially (linear in log-log space). The addition of the self-blocking Loop Extrusion Factor (LEF, cohesin) creates a "shoulder" of increased short-range interaction frequencies.

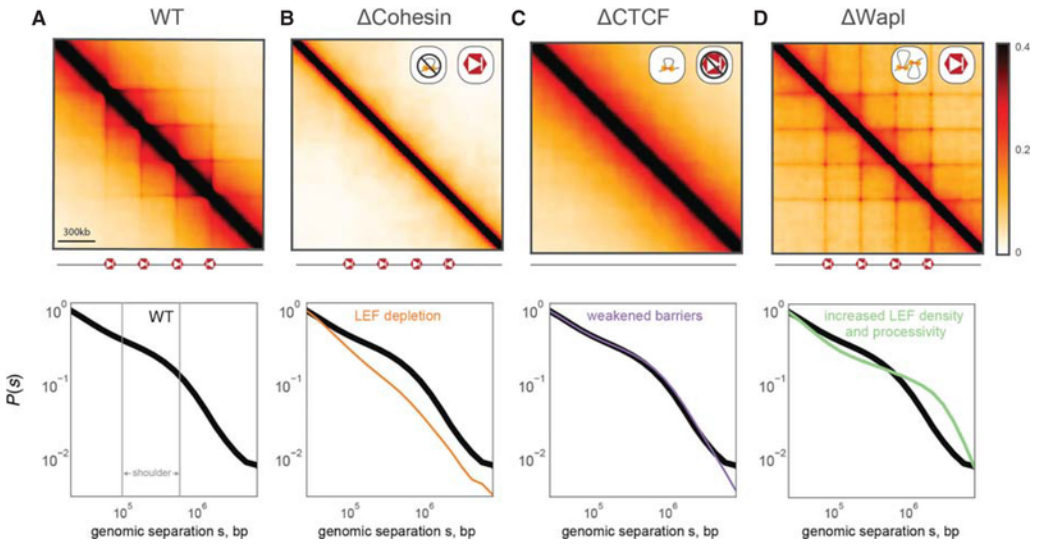

Figure 1: Loop extrusion polymer simulations predict the consequences of cohesin and CTCF perturbations. (Top row) Simulated Hi-C maps for indicated perturbations. (Bottom row) P(s) for indicated perturbation compared to WT P(s). From Fudenberg, *et al.* (2017).

A/B compartments

Although A and B compartments are quite visually obvious on the heatmaps, they are also computationally defined. Because compartments are the largest source a variation in Hi-C datasets, they can be easily captured through Principal Component Analysis (PCA). It was first used in Lieberman-Aiden, *et al.* (2009). PC1 typically represents compartments as you can see on the figure below.

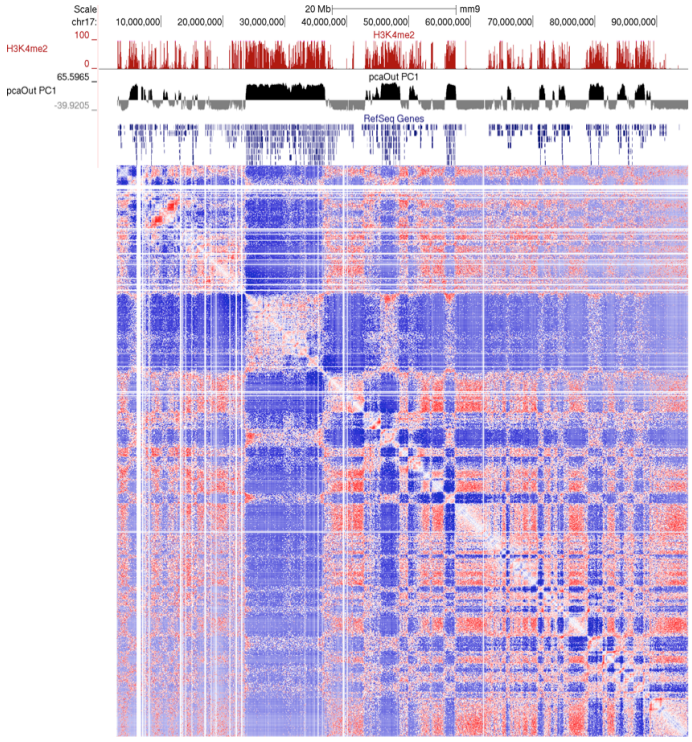

Figure 2: (Top panel) H3K4me2 ChIP-seq data (Note that we would now use H3K27ac as the mark of active chromatin, due to its better predictive power). (Middle panel) PC1 obtained from PCA of the Hi-C map. (Bottom panel) Hi-C map. From the HOMER website.

A good way to validate that those are compartments is to produce a ”saddle plot”. This essentially is a heatmap that shows the average interaction frequencies as a function of the PC1 scores.

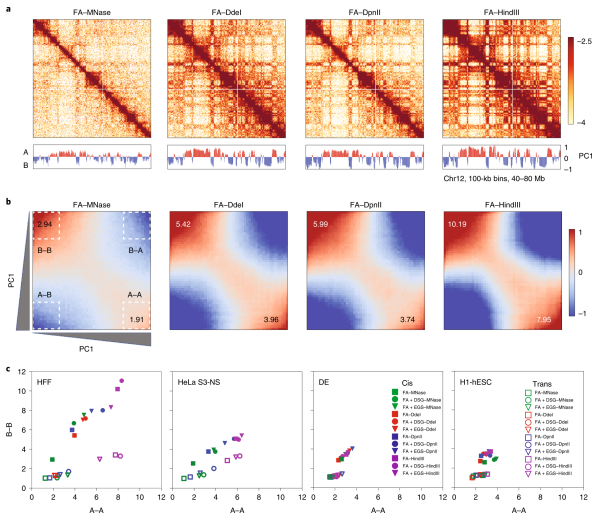

Figure 3: (Top row) Maps from different Hi-C protocols and the corresponding PC1. (Middle row) Saddle plots. From Akgol Oksuz, Yang, Abraham, *et al.* (2021).

TAD calling

There are many TAD callers. They are use different algorithms and generate different (although thankfully similar) results.

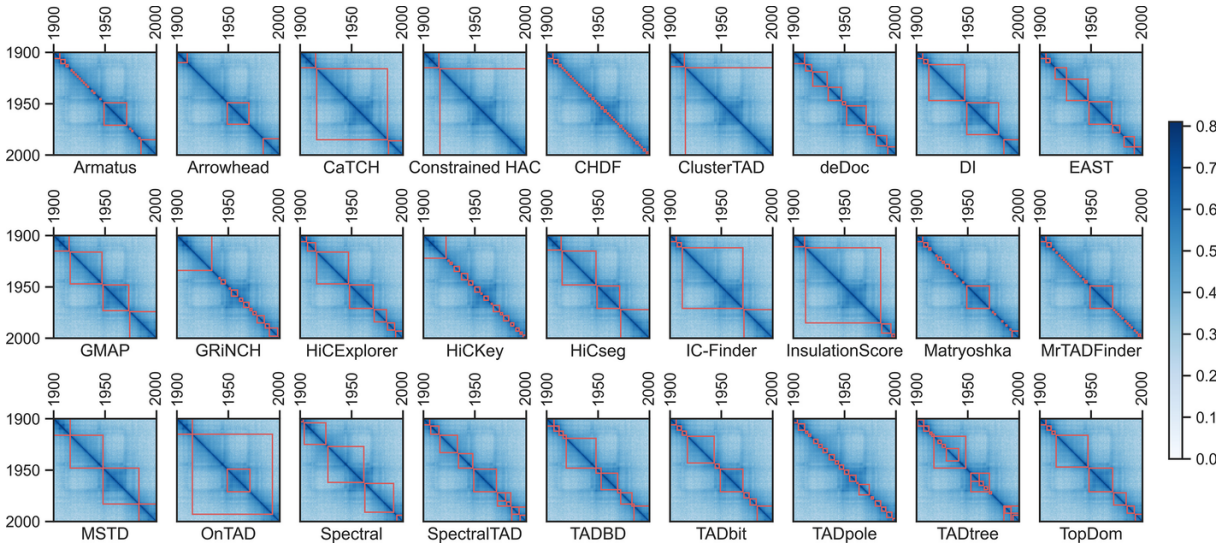

Figure 4: From Sefer (2022)

.

Some of the most widely used methods are:

1. **Arrowhead** from Rao, *et al.* (2014). Some transformation is applied to the normalized count matrix. This transformation produces deformed arrowhead-shaped domains. From this matrix, it is possible to assign pixels to domains.

2. **Insulation Score** from Crane, *et al.* (2015), implemented in `cooltools`. The insulation score is defined as the average interaction between regions upstream and downstream of a bin (diamond). Domain boundaries correspond to local minima of this insulation score. A derivative of the insulation score is computed along the chromosomes and thresholded to identify bins. This has two arbitrary parameters: the window size (for the upstream and downstream regions) and the threshold.

3. **CaTCH** from Zhan, *et al.* (2017). A Reciprocal Insulation (RI) score can be computed for pairs or arbitrarily size adjacent bins. A thresholding method is then used to merge those bins into domains. Because of the flexibility of the threshold, this methods results in hierarchical domains. The threshold can be picked in such a way that CTCF sites are enriched at domain boundaries (thus corresponding to TADs).

4. **HiCExplorer** from Ramírez, *et al.* (2018), implemented in `HiCExplorer`. The data is normalized with respect to the genomic distance that separates the bins (ie., the distribution of contact frequencies for a given distance is used to compute a z-score). The TAD separation score is then computed as the average z-score of the diamond above a given genomic bin (interactions between bins on the left and bins on the right, similar to the Insulation Score method). This is done at multiple resolutions, combined and converted into a p-value.

5. **TADbit** from Serra, *et al.* (2017). TADbit leverages asseumptions about distribution of contact as a function of distance (exponential decay) and the number of reads sequences (Poisson distribution) to statistically infer (most likely) deviations from the expected values using Poisson regression.

Data exploration

1. Open HiGlass.
   The inferface is pretty simple. Each "window" has 5 regions (top, right, bottom, left and center) where you can add tracks. Tracks are added by clicking on the + at the upper right corver of the window.
   After a track is added, the name of the genome to which the data was mapped appears at the bottom right corner, as well as the name of the dataset and the current resolution. The color scale is at the upper right corner (you can change the lower and upper limits).

2. The ability to discern 3D structure in Hi-C maps depends on the resolution of the maps (and thus directly on the sequencing depth).

   Question 1. In order to simulate the effect of a difference in sequencing depth, we can look at the same dataset at different resolutions. To do this, load the WT mESC data from Hsieh, *et al.*, (2019).
   Go to the coordinates `chr10:79,456,726-81,654,520 & chr10:79,595,224-81,428,597`. Duplicate the window and lock the zoom and location across the two views.
   To do this, click on the small wheel on the upper right corner of the heatmap on the right and click on "Take and lock zoom and location with" and then on the heatmap on the left.
   Now, put your mouse on one of the heatmaps and click the small wheel, then go to *Configure Series > Zoom Limit > 10 kb* to lock the zoom to a 10 kb resolution, even when zooming in. Zoom in such that the resolution goes to 1 kb on the unlock view and count the number of TADs (squares) at both resolutions.

3. Change the dataset to the ESC data from Bonev, *et al.* (2017) and the genome and annotation mm10.

   Question 2. Look at the heatmap from the perspective of the whole genome (zoom out). You can see clear blocks of increased self-interaction. What do those blocks correspond to and what type of nuclear organization do they represent?

   Question 3. Zoom in to see the heatmap at the scale of a single chromosome. There are white bands. Compare their location in human and mouse. Where are those bands positioned, why is this different between mouse and human and why are they white?

   Question 4. Look at this dataset. How can you explain this strange contact pattern?

4. Using the Bonev, *et al.* (2017) ESC dataset, zoom enough to reach the 128 kb resolution. The plaid/checkerboard pattern appears. Those are the megabase-scale A (active)/B (inactive) compartments.

5. You can also add Hi-C heatmaps to the top, right, bottom and left panels to obtain the heatmaps with triangles, which typically with visualization . Add the Bonev, *et al.* (2017) ESC track to the top. Add the ESC TADs track from Alavattam, *et al.* (2019) ESC TADs and the CTCF ChIP-seq from Bonev, *et al.* (2017) to the top panel.

   Question 5. How are the CTCF ChIP-seq peaks positioned relative to the TAD boundaries?

6. Duplicate the window and replace the heatmaps with the NPC dataset from Bonev, *et al.* (2017).

   Question 6. Zoom in to reach the 8 kb resolution and scroll a bit. Do **TADs** in the NPC data correspond to TADs in the ESC data?

   Question 7. Zoom out again to the 128 kb resolution and scroll a bit. Do **compartments** in the NPC data correspond to compartments in the ESC data?

7. The modern theory for the formation of self-interacting domains in vertebrates is the loop extrusion model.

   Question 8. In your own words, briefly explain the loop extrusion model.

   Question 9. (a) Explain what "flames" (or "stripes") are and their mechanistic basis. Are they typically shared between ESCs and NPCs? Do you find CTCF ChIP-seq peaks at their root?

   (b) Load the TAM dataset from Schwarzer, *et al.* (2017) and the untreated ESC dataset from Nora, *et al.* (2017). Find two examples of flames (provide the genomic coordinates where the flames are clearly visible).

   (c) Now, load the NIBPL KO dataset from Schwarzer, *et al.* (2017), and the Auxin-treated data from Nora, *et al.* (2017). How do those stripes behave? Why?

8. Hi-C data readily allows to identify large scale inter-chromosomal structural variants such as translocations, inversions, deletions and duplications. For example, a deletion would appear as a reduced frequency of interactions between the deleted region and other genomic regions, resulting in a diagonal gap in the heatmap. Conversely, a duplication would result in a higher frequency of interactions between the duplicated region and other genomic regions, resulting in a diagonal stripe of increased interaction frequencies. Similarly, an inversion could cause regions that were previously not in close proximity to interact more frequently, resulting in off-diagonal squares with higher interaction frequencies. Finally, a translocation between two chromosomes would result in off-diagonal squares between the two chromosomes with aberrant reinforced interactions frequencies.

Question 10. (a) Load the data from GB176 (Harewood, *et al.* (2017)) and change the genome to hg19. Identify and describe at least two inter-chromosomal rearrangements (ie., what type of rearragenement? Between which chromosomes?).

(b) Same question with the HeLa control dataset from Wutz, *et al.* (2017).

(c) Now zoom to chromosome 11 on the HeLa dataset. What phenomenon can explain such a strange interaction pattern?

9. Additionally, Hi-C allows to detect smaller scale structural variants.

Question 11. (a) Go here. What kind of structural variant is this?

(b) Same question with this one.