Master projects overview

Introduction to MC-4C

We recently developed Multi-Contact Chromosome Conformation Capture (MC-4C) to uncover 3D interactions that occur simultaneously between multiple genomic loci in single-alleles of DNA (see **Figure.1**). Such a conformation is often formed by elements in the genome to carry out varied function such as activation (or inhibition) of genes. Using MC-4C, we showed for the first time how multiple enhancers collaborate (or compete) to express their target genes in a locus of interest with single-allele resolution (Allahyar & Vermeulen et al. 2018).

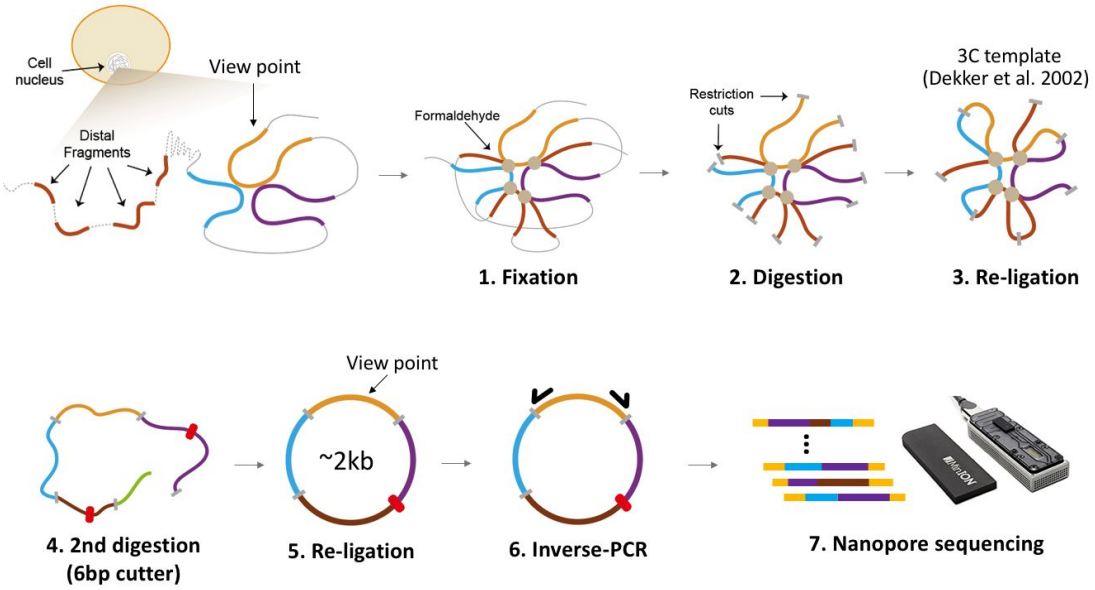

Figure.1 Overview of MC-4C preparation protocol.

[Machine learning] Discovering rare DNA micro-topologies using multi-contact chromatin conformation data with single-allele resolution

The information-rich nature of MC-4C data (~40 runs 3rd generation sequencing; Nanopore Technology) provides excellent opportunities for assessments that was not possible before. For example, MC-4C data can further be explored to detect frequently formed micro-topologies (i.e. enhancer-gene interactions) that are missed when this data is analysed as a homogeneous population of DNA conformations. A promising approach to discover these micro-topologies is unsupervised learning (i.e. clustering). To this end, the multi-way chromatin contacts captured in MC-4C can be seen as matrix of samples where each row represents a single allele and columns represent contacting fragments found in that read (see **Figure.2**). There are several challenges that need to be overcomed in this analysis. An important property of chromatin capture data is that fragments that are linearly close in the genome are often observed together in a read. Therefore, a standard clustering of reads containing linearly proximal fragments do

not capture the true 3D micro-topologies that are formed in the locus of interest. This issue could be alleviated for example by devising a linear-proximity aware distance measure which provides a proxy for determining cluster membership of reads according to captured 3D interactions (represented by enclosed fragments) in each read.

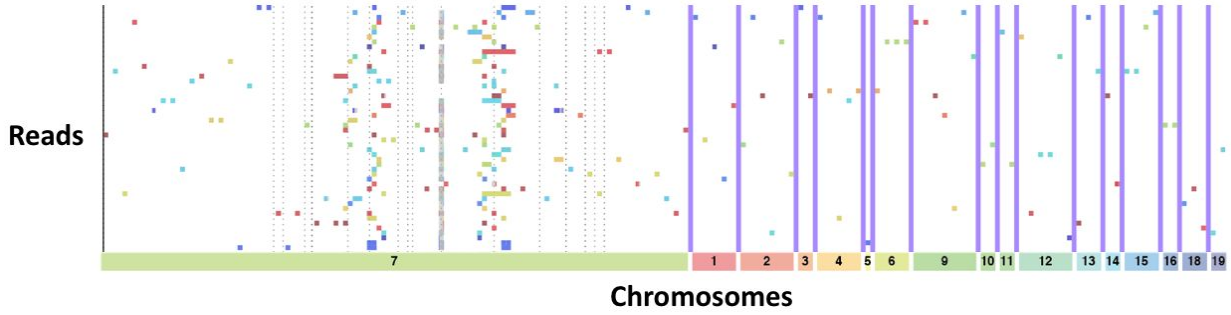

Figure.2 Matrix-based representation of fragments in MC-4C. Each row represents a single read. Each column represent a restriction fragment in the genome.

[Bioinformatics] In-silico Unique-Molecule-Identifier (UMI) to discern single-allele conformations in 3rd generation nanopore sequencing technology

PCR amplification is one of the key ingredients of targeted chromosome conformation capture protocols, including MC-4C. PCR amplification ensures that most (if not all) sequenced reads are coming from a region of interest in the genome (hence the name “targeted”). However, this step also complicates quantitative assessment of sequenced reads. This is because many reads could simply formed by conformation of a single allele (in a cell nucleus) that is amplified multiple times during the PCR amplification.

To resolve this issue and filter out PCR-duplicated reads, we exploited targeted property of MC-4C data. Essentially in MC-4C, we do not expect many fragments to map far away from the region of interest. Therefore, frequently observed trans (or far/cis) fragments in the data can indicate PCR amplification of their corresponding reads. While this approach can reliable identify read duplicates, it can not be used for reads that do not enclose trans (or far/cis) fragments.

An alternative strategy would be to investigate order and orientation of fragments within each read. This way, a pair of read with their fragments arranged in identical order and orientation can be considered as duplicates (see **Figure.3**).

The main challenge in this project is that MC-4C reads are sequenced using 3rd generation sequencing and therefore have many errors in their sequence which complicates the similarity comparison. Additionally, fragments may lose part of their sequence during library preparation or sequencing. As a result, they may map to reference genome in varied length. Taken together, a robust computational method is required to confidently identify fragments and compare order and orientation of fragments for each pair of reads.

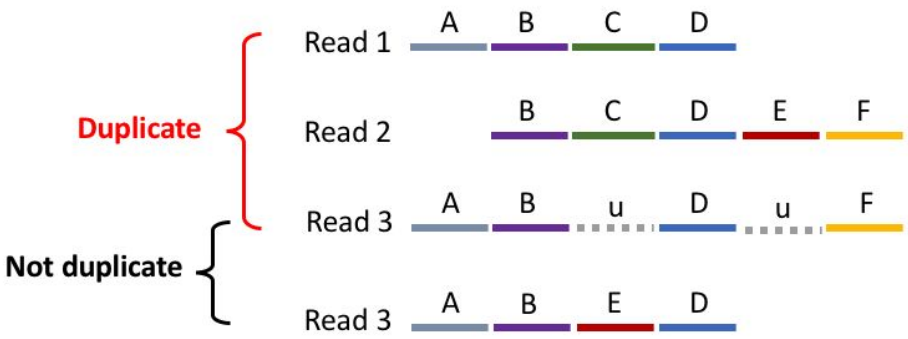

Figure.3 Proposed duplication removal filter. Each row represents a single read. Fragments are represented with colored line. Unmapped fragments are represented with dashed gray line.

[Bioinformatics] Mult-way aware aligner to improve long-read (3rd generation) sequencing mapping quality

In order to assess multi-way contact in MC-4C, long reads must be sequenced using 3rd generation sequencing. However, sequence quality of these reads are often lower than 2nd generation sequencing platform such as Illumina. Consequently, substantial number of fragments in MC-4C reads can not be mapped. This issue calls for an effective mapping strategy to increase number of mapped fragments which in turn increases MC-4C capability in detecting multi-way conformations. A potential avenue for resolving this problem is tendency of observing linearly close fragments in MC-4C reads. In other words, once a fragment is mapped, it is highly likely that the next fragment to be mapped in the vicinity of already mapped fragment (see **Figure.4**). In the standard mapping step of MC-4C, this information is completely hidden from aligner. A specifically-crafted aligner which exploits this information can for example prioritize probable mapped locations of a fragment according to closeness of already mapped fragments in that read.

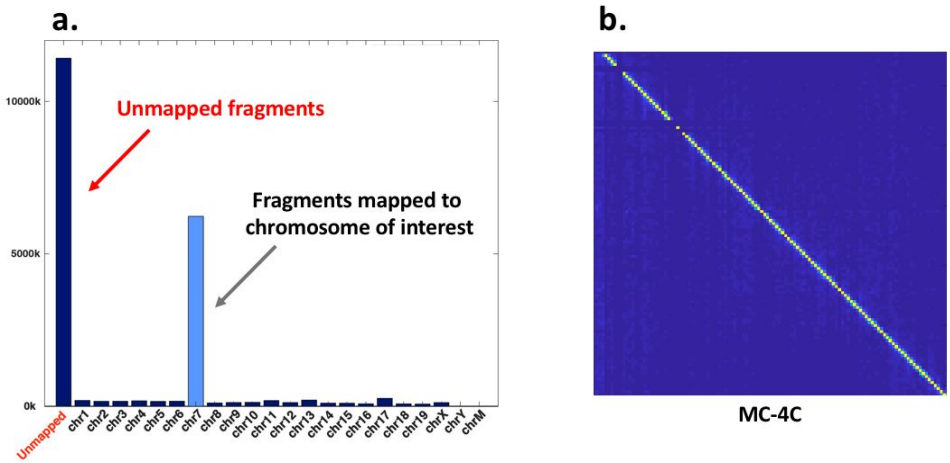

Figure.4 Many fragments can not be mapped due to sequencing errors. **a.** Chromosome coverage of mapped fragments. **b.** Intensity of diagonal line represent the frequency of mapping a fragment (x-axis) when flanking fragment in the read is mapped to a region (y-axis).

[Visualization] Peaking into higher-order chromatin contacts: visualizing multi-way DNA interactions in the nucleus

Visualization of pair-wise contacts in genome conformation analysis is a well-studied area of research. These contacts are often represented in a matrix where each element represents contacts formed by two genomic sites (see **Figure.5**). In this representation, three-way contacts require a cube which is difficult to probe. At this moment, it is still unclear how multi-way interactions should be visualized. Therefore, there is a great need for ideas and methods to visualize these contacts with aim of helping the investigator in hypothesis generation and getting insights from these information rich data.

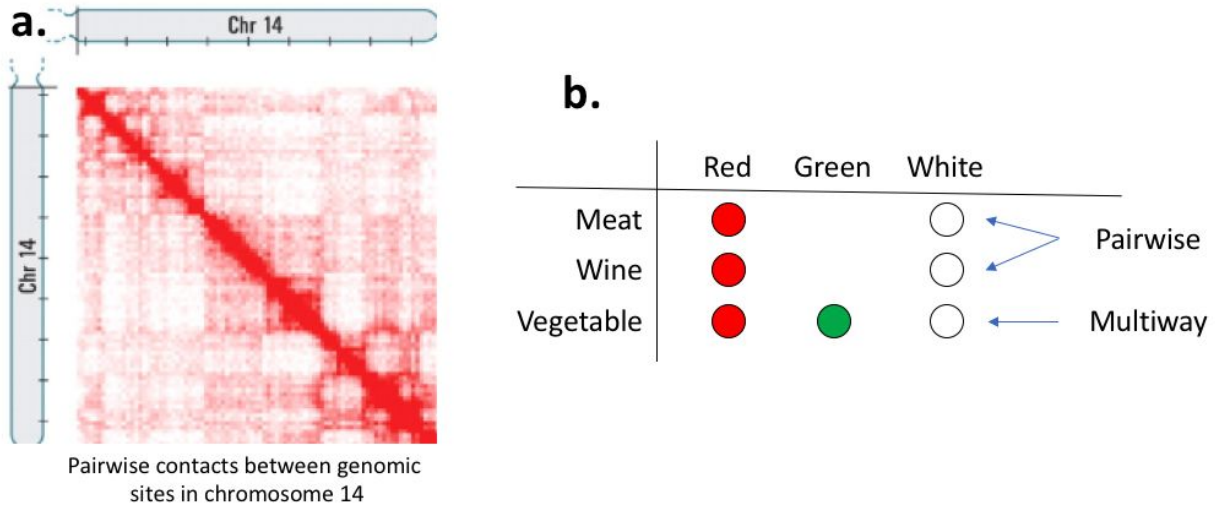

Figure.5 Visualization of 3D DNA interactions. **a.** Most genome conformation visualizations are based on pairwise interactions. **b.** Multi-way interactions between elements can not be represented in a similar manner.

[Statistical modeling] Association test to distinguish linear proximity of sites in the genome from their 3D interactions

Each read in MC-4C data represent multiple fragments that were in close proximity at the moment of fixation. In this construct, linearly close regions have higher chance to be “captured” (i.e. observed). In other words, once a fragment is mapped in a read, other fragments within the read have higher chance to also map in the vicinity of already mapped fragments. This bias introduces a problem in multi-way interaction analysis. This is because, the commonly investigated sites of interest are often very close in terms of linear genome (see **Figure.6**). Therefore, it is difficult to determine whether the observed frequency of co-captured fragments are due to their linear proximity or it is an actual looping mechanism that drives this interaction.

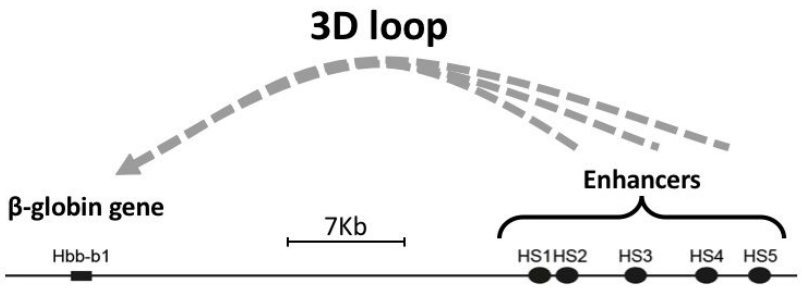

Figure.6 Functional elements of β-globin locus in mouse. Multiple enhancers co-cluster to substantially enhance expression level of β-globin gene.