Compartment repositioning detection algorithm

November 8, 2022

1 Calder genome segmentation

A chromosome is a ordered set of positions $C = \{p_1, p_2, ..., p_N\}$, where N is the total number of positions. Each position is also called *bin* and its size in base pairs (*bin size*) is decided *a-priori*.

Through the Calder algorithm, a chromosome can be partitioned into a set $\mathcal{D}_C$ of *compartment domains*, which are disjoint sets of adjacent bins. If we assume that each bin can be assigned to a compartment domain (which in practice can be false, given lack of data from the Hi-C), then compartment domains are a partition of C.

It is easy to prove that the total number of possible partitions of C in compartment domains is 2^{N-1}

Calder also ranks the compartment domains of each chromosome by imposing a order on $\mathcal{D}_C$. This means that, given two domains d_1 and d_2 belonging to $\mathcal{D}_C$, either $d_1 < d_2$ or $d_1 > d_2$.

We can therefore write:

$$\mathcal{Y}^C = Calder(C) = (\mathcal{D}_C, <)$$

1.1 Properties of $\mathcal{Y}^C$

The order on $\mathcal{D}_C$ can be used to assign a rank to each compartment domain, such that

$$r(d) = \frac{\text{Rank of domain } d}{|\mathcal{D}_C|}$$

where $r(d) \in [0, 1]$.

Going to the bin level, we can simply assign a rank to each bin b based on the compartment domain d they belong to ($b \in d$)

$$r(b) = r(d)$$

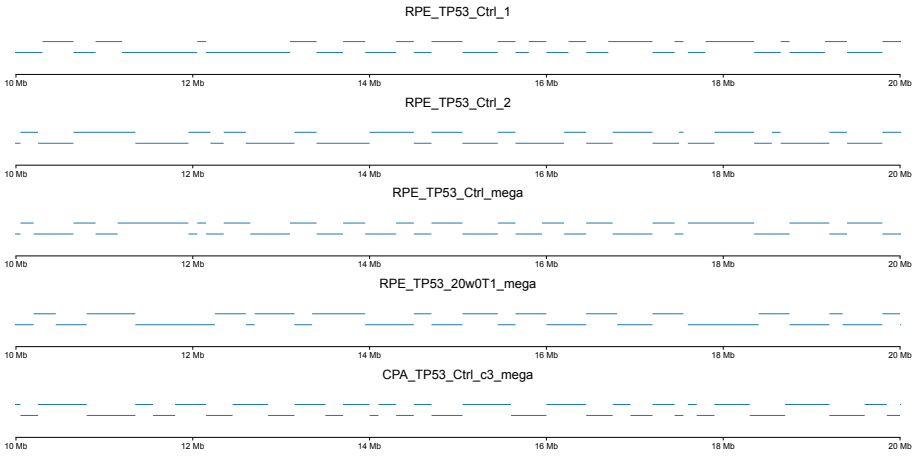

Figure 1: Example of compartment domain partitioning in some Hi-C samples.

2 Compartment repositioning

Let's now consider two experiments E_1, E_2. They will produce two Calder segmentations and rankings $\mathcal{Y}_1^C, \mathcal{Y}_2^C$.