Caleb Lareau
BST 290 Advanced Computational Biology
Prepared 17 October 2016

Primary Summary–HiC rao et al.

As I've spent the better part of the past year analyzing data from genome topology experiments, I set up my presentation to frame this paper in the context of past and present work, including alternate methods of genome topology investigation. The paper selected is quickly becoming a mainstay in the literature for genome topology as it 1) provides a detailed experimental protocol for in situ Hi-C, 2) defines many key terms and features concerning insights from Hi-C data, and 3) provides the highest resolution mapping at nearly 5 billion reads. Some of the key discussion points that I highlighted included the definition of topologically associated domains, boundaries, subcompartments, loops, and resolution.

Since I was pretty familiar with the work, I was probably more critical than most other presenters. Namely, with my experience in developing the diffloop software, I've thought quite a bit about how loops are defined from these assays. Not only does in situ Hi-C fail to generate the data to maximize loop calls (compared to other assays like ChIA-PET, HiChIP, and PLAQ-Seq) but the algorithm proposed in this paper (HiCCUPS) is lacking. These will be points of innovation over the next few years.

One component of the analysis that I find useful is the description of the variable maternal/paternal haplotypes and how these data can determine genomic imprinting through variation in the topological domains. Using the GM12878 phased heterozygote alleles, Rao et al. was able to split the reads based on the maternal/paternal origin. While Hi-C fails to provide the resolution needed to assess the potential causal effect of a variant in altering topology, a similar mechanism for splitting reads based on phased alleles has been implemented in ChIA-PET data (Tang et al. 2015–Cell) where they were able to discern common variants that alter topology. This section of the paper provides some unique links to other statistical genetics concepts and may enhance the interdisciplinary links between these fields.

Ultimately, I referenced the paper by Brad Bernstein's group concerning insulator (boundary) disfunction leading to proto-oncogene activation. This localized model provides the best conceptional understanding for why genome topology experiments will be critical to our understanding of complex diseases like cancer. I will discuss this model system in the GBM secondary, but it of course begs the question whether this phenomenon is specific to the locus identified by Bernstein's group or whether this notion of weakened boundaries permeates complex phenotypes.

The discussion involving the potentially poor definition of resolution by this paper is one of the clear weaknesses and provides some opportunity for enhanced statistical metrics to determine the quality of Hi-C data. However, other technological advances such as Liu proposed in capture Hi-C present a different set of biases and require a unique set of statistical methodology to determine significant loops. Ultimately, I expect significant development and innovation on both the computational and experimental fronts that will allow topology data to be discerned in more phenotypes and at higher resolutions.