Computational Challenges for the Analysis of High-throughput Sequencing Data

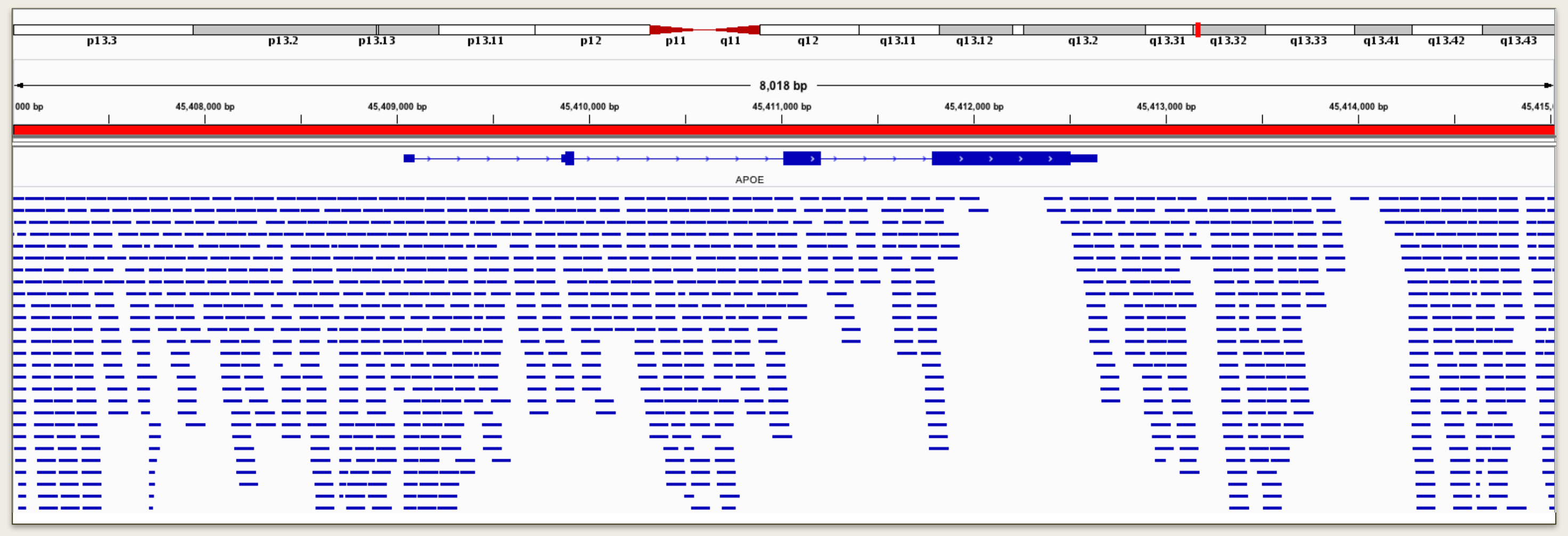
THE UNIVERSITY OF CHICAGO

Ittai Eres & Joyce Hsiao
Department of Human Genetics, University of Chicago

We gratefully acknowledge the support of:
Genetics and Regulation Training Grant (T32 GM07197)
University of Chicago Research Computing Center

INTRODUCTION

High-throughput sequencing (HTS) technology can be used to study gene regulation associated with many different molecular phenotypes, such as gene expression, transcription factor binding and chromatin accessibility. We at the Gilad lab routinely incorporate the latest-generation HTS assays to understand the impact of regulatory and genetic variation within humans and between species (focused on human and closely related primates). HTS assays (e.g., RNA-seq, ChIP-seq, Hi-C) generate high-resolution data on how molecular traits vary along genomic locations for each sample (e.g., a single cell or a combined sample of cells collected from the same individual).

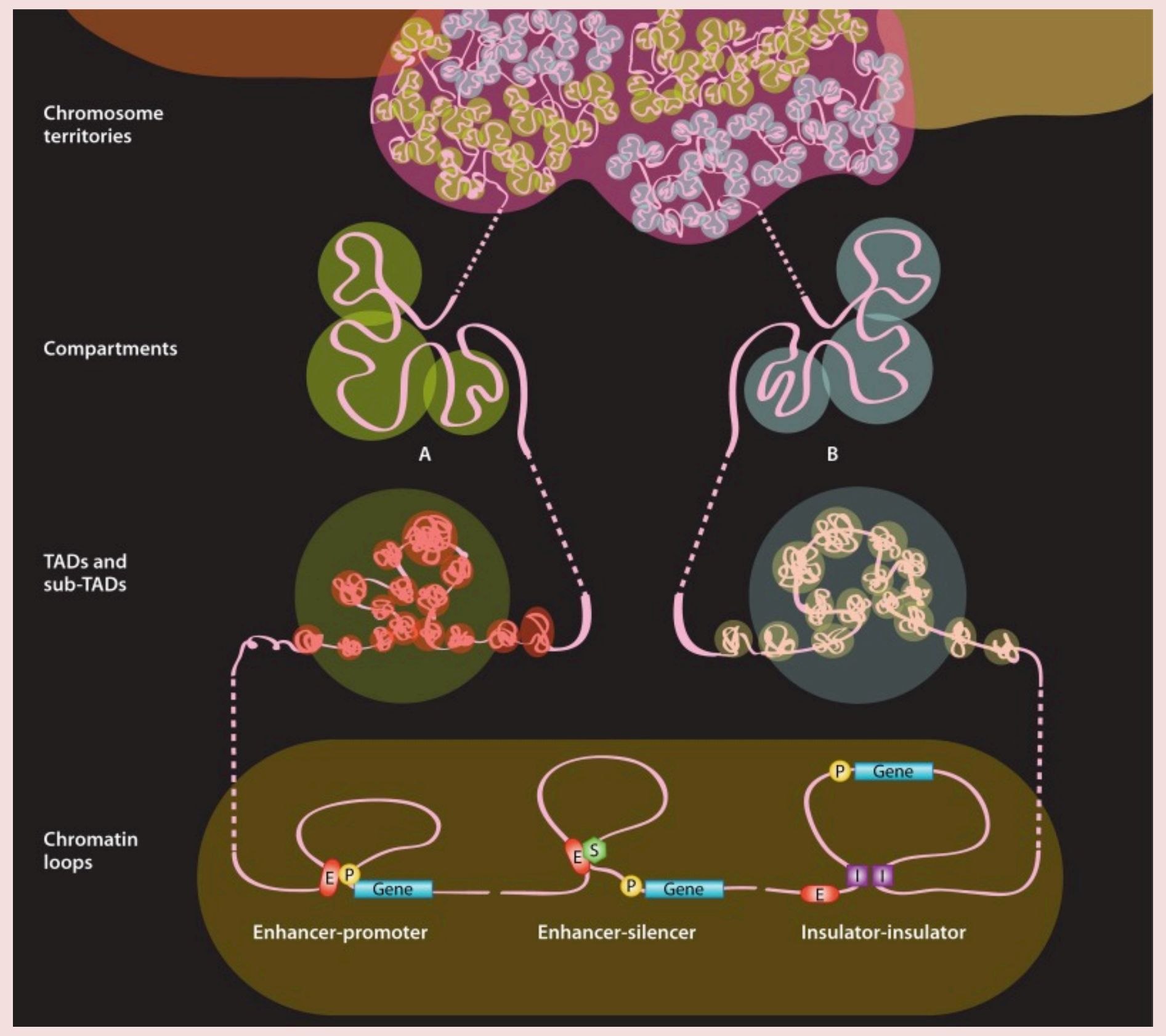


Snapshot of Hi-C sequencing reads mapped back to the human reference genome on chromosome 19.

DETECTING CHROMATIN INTERACTIONS

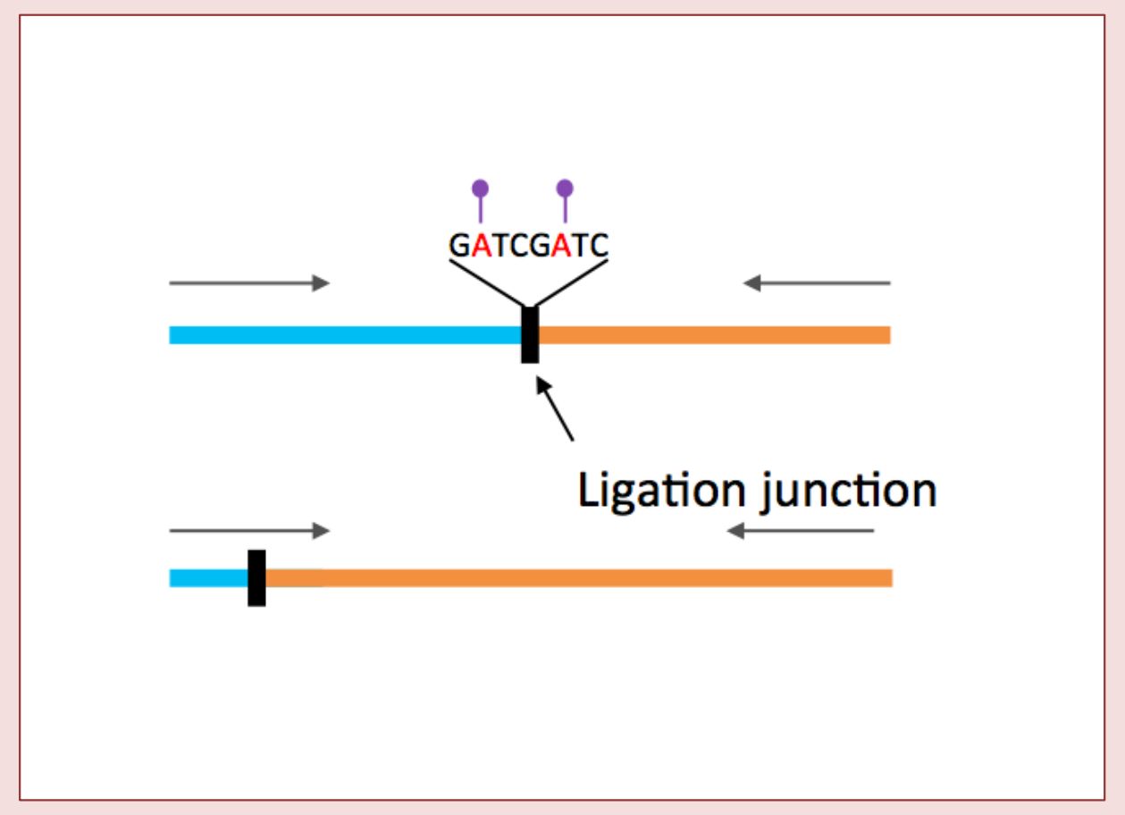
Our DNA is compartmentalized at many different levels inside a cell. ~2 meters of DNA is packed into each micrometer-scale nucleus. Folding and looping of the chromosomes brings linearly distant loci on the genome into proximity. Figure credit: Fraser et al. 2015.

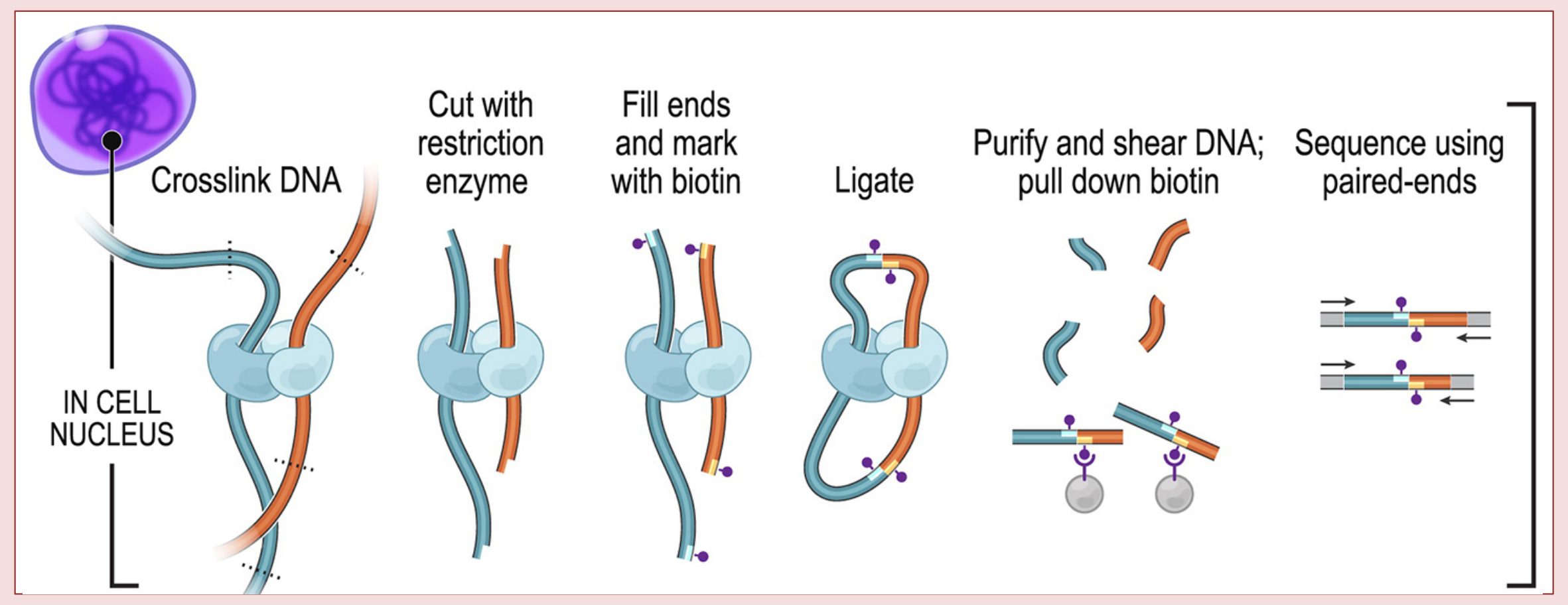
Hi-C allows for detection of this spatial structure, and has led to the discovery of many topologically associated domains (TAD) on the genome. These TADS represent large regions of DNA in which sequences are more likely to contact each other than they are to make contact with sequences outside the TAD.



High-throughput Chromosome Conformation Capture (Hi-C)

Hi-C is the latest development in the chromosome conformation capture technology that allows us to interrogate DNA-DNA contacts on the genome-wide scale. Figure credit: Rao et al., 2014.




Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends

PIPELINE

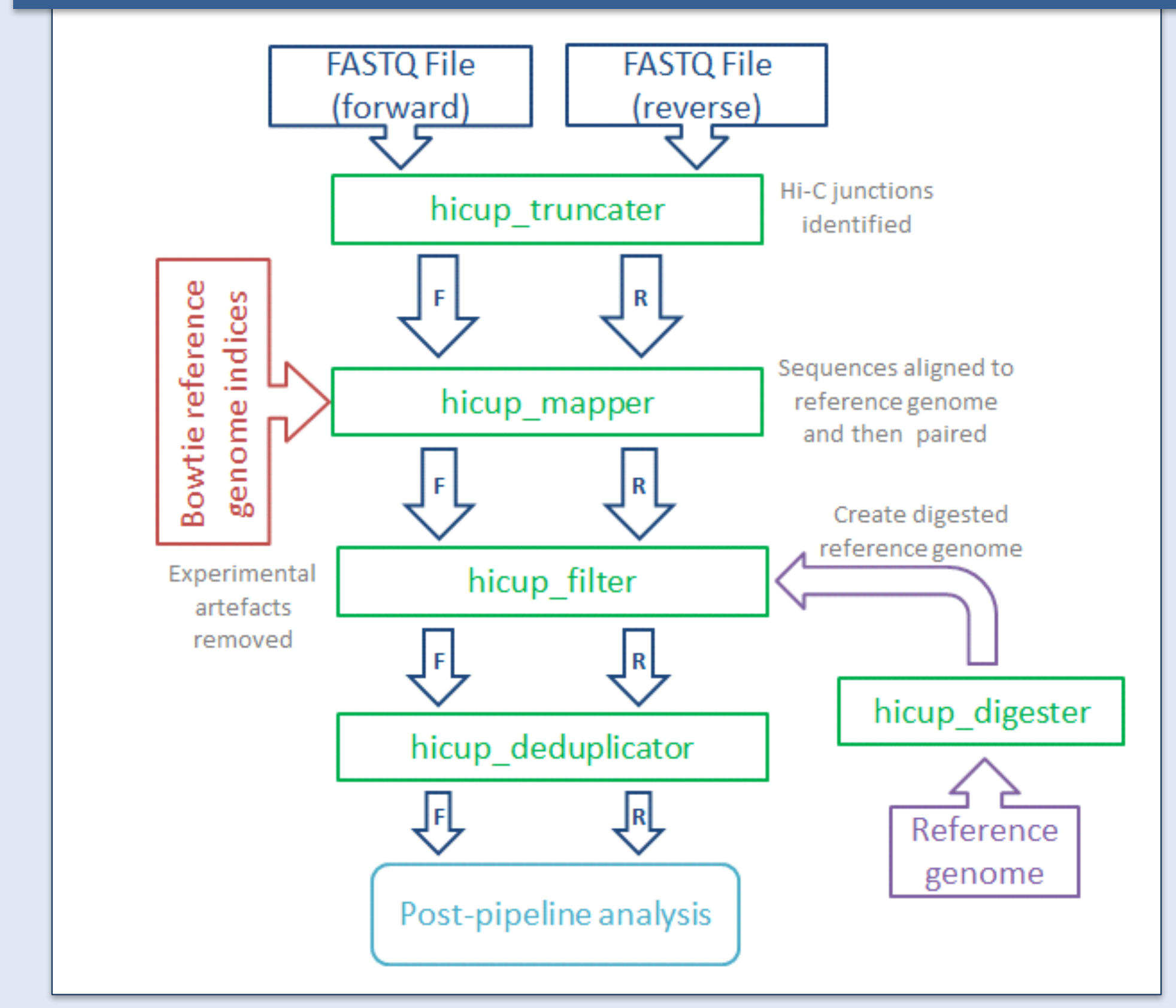
Broadly speaking, the HTS analysis pipeline can be broken down into three components:

(1) Preprocess each sample individually: The workflow involved varies widely between the type of sequence data. For RNA-seq – a popular technology for quantifying gene expression on the genome-wide scale, the workflow is well-established and typically includes applying bioinformatic tools for quality control and mapping sequence reads to reference genome. However, the workflow for preprocessing Hi-C samples is still under development, because Hi-C is still a relatively new HTS technology.

(2) Combine and summarize data from multiple samples: This step is usually similar between the different HTS data. Data from individual samples are pooled together and summarized at the level of genes, exons, exon-exon splice junctions or at the base-pair resolution.

(3) Analyze large-scale patterns along the genome: The input data at this stage are usually stored in a table format. R, especially the Bioconductor packages, are often used to visualize and analyze data.

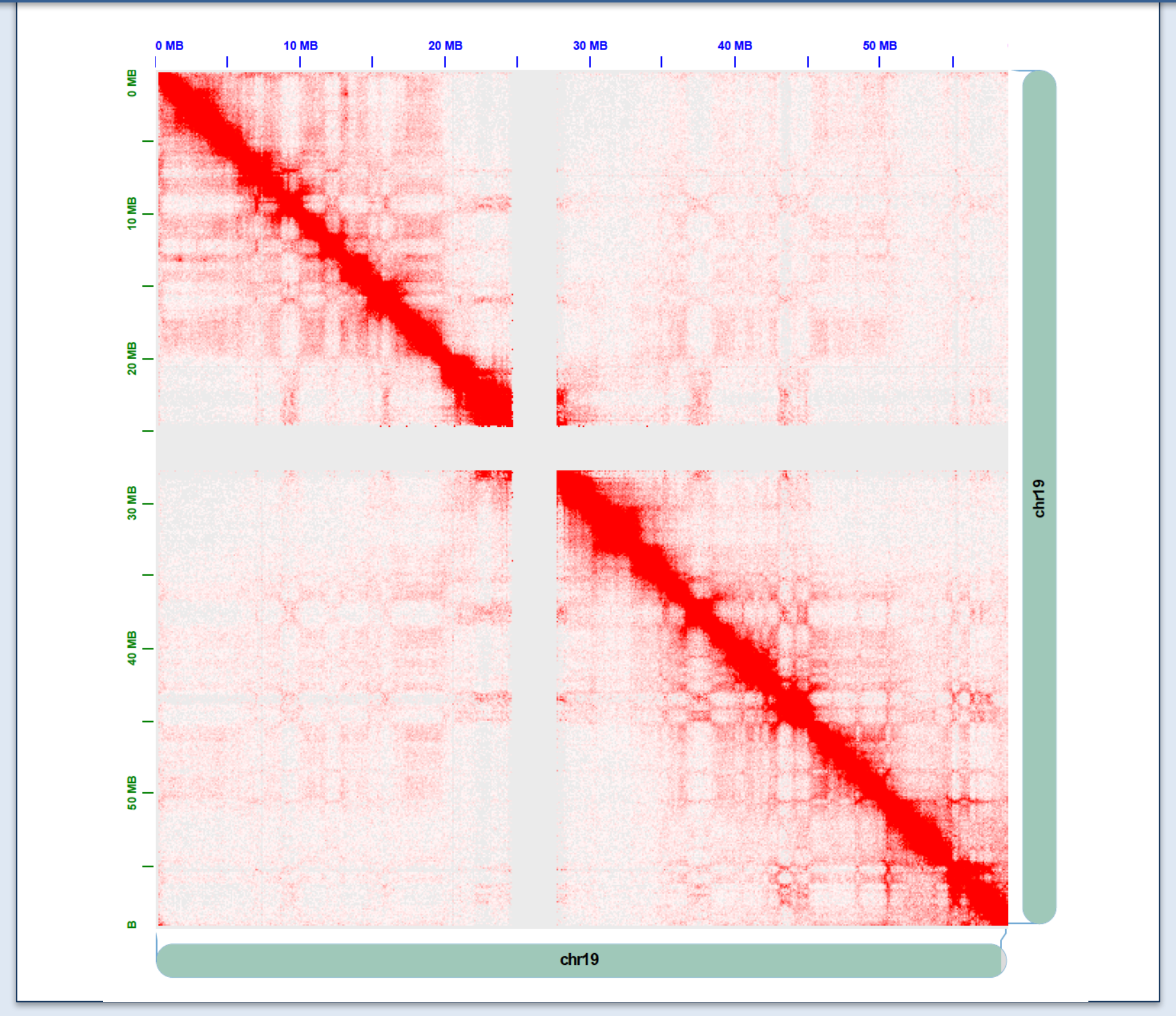
Analyzing Hi-C data


FASTQ File (forward)
FASTQ File (reverse)
hicup_truncater
Hi-C junctions identified
F
R
Bowtie reference genome indices
hicup_mapper
Sequences aligned to reference genome and then paired
Create digested reference genome
hicup_filter
Experimental artefacts removed
hicup_digester
hicup_deduplicator
Reference genome
Post-pipeline analysis

First, we go through sequence reads in parallel and truncate the reads that contain Hi-C ligation junctions. The truncated reads are mapped independently to the reference genome. After mapping the truncated reads, we use an in-silico genome digest to filter out reads which do not represent valid Hi-C ligation products. In the last step we remove PCR duplicates. The figure on the left outlines our current workflow for processing Hi-C samples. Figure credit: Wingett et al. 2015

After processing sequence reads, the major challenge lies in visualizing and analyzing locus-locus interaction across chromosomes. These large-scale patterns are difficult to visualize and require high-performance computing environments. Finally, there is currently no consensus on how to separate the true signal in the Hi-C data from the noisy measurements introduced by technical artifacts.

Hi-C Contact Heatmap



This heatmap represents a Hi-C contact matrix, with binned genomic loci of 100kb plotted symmetrically to show pairwise relationships. Each cell corresponds to a possible locus-locus interaction along chromosome 19. Darker-colored cells suggest the two loci are likely making frequent contact with each other inside the cell, while lighter-colored cells indicate that contact between the two loci is unlikely . Of note is the strong proximity-based signal seen along the diagonal, as well as larger squares emanating from it representing TADs and other sub-domains.

REFERENCES

Fraser, James et al. An Overview of Genome Organization and How We Got There: From FISH to Hi-C. Microbiology and Molecular Biology Reviews, 2015. 79(3): p. 347-72.
Rao, Suhas S.p., et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell, 2014. 159(7): p. 1665-680.
Durand, Neva C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Systems 2016. 3(1): p. 94-98.
Wingett, Steven et al. HiCUP: Pipeline for Mapping and Processing Hi-C Data. F1000Research, 2015. 4: 1310