Efficient processing of Hi-C data and application to cancer

N. Servant (Institut Curie, INSERM U900, Mines Paris Tech)

Over the past decade, major advances in high-throughput sequencing have allowed the development of new epigenetics approaches. Among them, the Hi-C technique was proposed as a genome-wide method to explore the chromatin organization in three-dimension (3D). Since then, the spatial organization of the genome and the physical interactions occurring within and between chromosomes has been described as a key factor of gene regulation and genome functions in general.

However, as any genome-wide sequencing data, Hi-C usually requires several millions to billions of paired-end sequencing reads, depending on genome size and on the desired resolution. Managing these data thus requires optimized bioinformatics workflows able to extract the contact frequencies in reasonable computational time and with reasonable resource and storage requirements. In this context, we developed a couple of years ago, HiC-Pro (https://github.com/nservant/HiC-Pro), an optimized and flexible pipeline for processing Hi-C data from raw reads to normalized contact maps. Today, I would like to focus on a new collaborative project called nf-core-hic (https://github.com/nf-core/hic) which is a Nextflow-based pipeline for Hi-C data analysis. The current version of nf-core-hic is dedicated to data processing, but in the coming months, we could like to encourage the community to further develop this pipeline, including additional analytical steps.

In addition, I will discuss the current computational challenges that emerge when Hi-C is applied on cancer cells. Given the important recent insights that chromosome conformation techniques have provided into 3D genome organization in a normal context, the application of such approach to a disease context offers the possibility to further explore the genome organization of cancer cells, and its impact on cell regulation. I will demonstrate why the Hi-C cancer data require dedicated normalization method, and how we can solve these issues through two recent normalization methods that we have developed in this purpose.