Integrative analysis of single-cell genomics data

Abstract for the general public

Genome carries the complete genetic information of an organism. This information manifests itself in the process of gene expression, by which DNA sequences of individual genes are decoded into functional products, such as proteins. Proper gene expression is a fine-tuned process controlled by DNA regulatory elements found in non-protein-coding parts of the genome. This process has been extensively studied by experimental techniques using DNA and RNA sequencing. However, traditional sequencing methods could only obtain the average signal over multiple cells, losing the information on cell-to-cell differences (cellular heterogeneity). This has changed with the development of single-cell technologies, which are capable of sequencing the genome or transcriptome of single cells. The use of single-cell technologies reveals cell population differences and allows for distinguishing rare cell types.

The objective of the project is to integrate genomics data from both single-cell and population-level methods to get a more complete understanding of gene expression. The mechanisms that regulate gene expression occur within a complex three-dimensional architecture that helps bring functional fragments of DNA into spatial proximity. The 3D organization of the genome allows for regulatory elements to be far away on DNA strand from their target gene, possibly with multiple other genes in between. While the 3D organization of the genome is essential for appropriate interpretation of genetic information, experimental methods quantifying this organization are usually performed on population level to provide sufficient resolution. For this purpose, we will complement them with data from single-cell sequencing experiments measuring gene transcription (single-cell RNA sequencing) and genome accessibility (single-cell ATAC-seq).

The 3D organization of the genome spans multiple scales, including well-demarcated physically interacting domains, known as Topologically Associated Domains (TADs) or chromatin domains. Chromatin domain structure correlates with gene expression, allowing to define active and inactive chromatin domains at the population level. However, the trend observed at population level can be reversed in individual cell types. We will use single-cell RNA sequencing data to test how often it is the case. We will also look at the variation of gene expression, observed in individual cell types within chromatin domains, and test to which extent the genes with high variation in their expression are co-localized in the same domain.

A more fine-grained level of genome organization ensures that the enhancers come into proximity to the promoters of their target genes. These enhancer-promoter contacts can be observed in population-level Hi-C data, but these data alone do not reveal whether the contact is occurring in all cell types of the population. To reveal cell population differences of these contacts, and to predict in which cell types they are taking place, we will incorporate single-cell ATAC-seq data and test the chromatin accessibility at the interacting regions. Finally, we will integrate chromatin accessibility data with matching single-cell RNA sequencing datasets. Using additional information on transcription factor binding and methods of DNA sequence analysis, we expect to identify the factors responsible for transcriptional activation or repression in individual cell types.