Computational methods for the analysis of single-cell RNAseq data at single-molecule resolution

Lead Groups: Isidro Cortés-Ciriano (EMBL-EBI)
Partner Group: Tommaso Leonardi and Francesco Nicassio (IIT)

Rationale & Hypothesis:

The advent of high-throughput single-cell RNA-seq (scRNAseq) assays now permit to characterize the transcriptomes of single cells at unprecedented scale and resolution. The most widespread scRNAseq technologies used today allow to sequence either the 3' or 5' end of each RNA molecule, together with a barcode and a Unique Molecular Identifier. The resulting short-read data permits to (i) identify which genes are expressed in a given cell by mapping the short reads to the reference transcriptome, and (ii) assign transcripts to individual cells using the sequence of cell-specific barcodes. However, although scRNAseq can now be performed for thousands to millions of cells, the power of scRNAseq technologies to characterize the repertoire of RNA isoforms, aberrant gene fusions, RNA modifications, and poly-A lengths at the single-cell level is limited given that short reads do not provide full-length transcript resolution. As a result, the landscape and functional impact of such biological structures remains poorly understood, even if ample evidence suggests that they play a crucial role in cancer biology.

Nanopore sequencing is an emerging technology that permits continuous reading of individual DNA/RNA molecules with read lengths of over >10kb. Such reads overcome the limitations of short reads outlined above, thus providing unparalleled information to map and reconstruct transcriptomes at unprecedented resolution. However, the higher error rate of Nanopore sequencing data as compared to Illumina sequencing poses a complex challenge when trying to assign individual reads to single cells (demultiplexing), thus limiting the applicability and yield of long-read scRNAseq.

In this project, the fellow will develop computational approaches to demultiplex, analyse, and interpret long-read scRNAseq data for cancer cell lines and primary tumour samples.

Aims:

1. Develop efficient algorithms to demultiplex long-read scRNAseq data in the context of cancer.
2. Develop open-source software for the end-to-end analysis, annotation, visualization and interpretation of long-read scRNAseq data.

Significance & Impact: This project will develop computational tools for the analysis of (aberrant) transcript isoforms and gene fusions at single-cell and single-molecule resolution, which holds great potential to further our understanding of fundamental biological aspects of RNA regulation in healthy tissues and cancer.

Integration of Expertise of Partners: This project represents a multi-disciplinary effort that leverages synergistic expertise at IIT and EMBL-EBI. Specifically, the Leonardi and Nicassio groups will contribute expertise in RNA biology and the analysis of Nanopore sequencing data. The Cortes-Ciriano group will provide expertise in the development of computational tools for the analysis and interpretation of cancer genomics data.