Mise en place d’un pipeline de déconvolution de données de transcriptomique spatiale et déploiement d’un outil de benchmarking avec application au cancer (glioblastome)

Level: Stage troisième année d'école d'ingénieur

Development environment: Python, Snakemake, R, Docker, Git?

Scientific background

In organs and tissues, functions are performed by individual cells working together in a specific spatial arrangement. In biology these functions are often assessed by measuring gene expression. Spatial expression data acquired for tissue samples is not only important to understand normal organ development and function, but also to investigate how cells are perturbed in diseased conditions like cancer.

However, the resolution of spatial expression data (where different cells are located within a tissue), can sometimes be low, causing data from multiple cells of different types to be mixed together in one spot. This “mixing” can obscure the true values of gene expression in each cell type, leading to errors in how we interpret the tissue's biology and reconstruct its cellular makeup. To address this, an essential part of analyzing spatial transcriptomics data (which combines gene expression data with spatial information) is "**cellular deconvolution**" (see Figure 1). This process uses computational techniques to separate out the mixed data into estimates of how much each cell type contributes to the sample, helping to clarify the true composition and behavior of cells in the tissue.

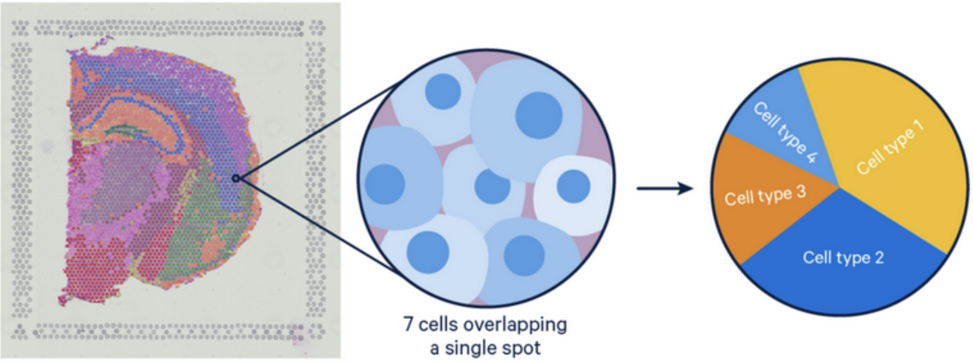

Figure 1. Deconvolution is required to determine the cell type composition of each spot.

Spatial transcriptomics being a rapidly evolving research field, there is a large number of deconvolution methods whose performance is difficult to assess. Benchmarking pipelines that enable the user to apply different/several deconvolution algorithms and quantify their performance do exist. One such pipeline is **Spotless** (https://github.com/saeyslab/spotless-benchmark). Spotless comes under the form of a Nextflow pipeline that uses either *Singularity* or *Docker* container platforms to run and compare the performance of both R- and Python-based deconvolution methods that are most widely used by the spatial transcriptomics research community. Spotless developers also provide a platform to generate synthetic spatial transcriptomics data (**synthspot,** https://github.com/saeyslab/synthspot).

Current and future spatial *omics* analysis carried out in the lab could potentially benefit from the in-house availability of both the benchmarking and the dataset simulation tools.

Internship objectives

Some tools for spatial deconvolution analysis have already been tested in our team. However, a reproducible, easy-to-use pipeline is still to be developed.

The successful intern will showcase and sharpen her or his coding skills deploying the **Spotless** benchmarking framework on our servers and testing it on the provided data set:

1) The **Spotless** code base is written in **R** and runs using **Nextflow.** The intern will provide a prototype of the benchmarking framework in **Python** for selected deconvolution algorithms in **Python** using the **Snakemake**

workflow available in house. She/he will ensure that **Python** codes produce the same results as original **R** code on a gold standard dataset.

2) The intern will extend the workflow to include selected containerized methods based on the docker images provided in **Spotless.** This step will serve as a template to extend the benchmarking to other deconvolution methods that are so far absent from the benchmarks.

3) The intern will generate synthetic spatial transcriptomics data via **synthspot**. Time permitting, the deployed benchmarking framework and prototype will also be applied to such simulated data.

The intern will work at a project that sits at the crossroads between computer science and biology, with the internship outcome potentially helping the researchers in the team in their current and future spatial *omics* analysis.

Candidate profile

We are looking for a highly motivated intern with an interest in applying their computer science knowledge to health-related scientific questions. Good communication skills and team spirit are expected.

Hard skills

Required skills

□ Good command of Python programming language
□ Good command of singularity or docker
□ Good command of SnakeMake
□ Good command of Unix/Linux environment

Preferred skills

□ Git revision control system

Desired skills

Command of R programming would be a plus

Environment

The internship will take place at the IBGC (https://www.ibgc.cnrs.fr/en/welcome/) under the supervision of Maialen Arrieta-Lobo, Slim Karkar (MCU) and Macha Nikolski (IBGC team leader and CBiB director).

Contact

Interested candidates should send their resume/CV and cover letter to Macha Nikolski (macha.nikolski@u-bordeaux.fr), Slim Karkar (slim.karkar@u-bordeaux.fr) and Maialen Arrieta-Lobo (maialen.arrieta@u-bordeaux.fr).