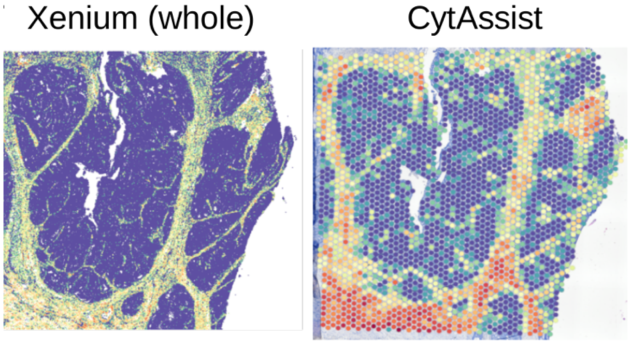
Master Project: Spatial Processes for Single-Cell Transcriptomics

- **Keywords:** stochastic processes, computational statistics, multiple testing, spatial transcriptomics
- **Location:** Laboratoire Jean Kuntzmann (Univ. Grenoble Alpes) or Laboratoire de Biologie et Modélisation de la Cellule (ENS Lyon)
- **When?** 4-6 months from March/April 2025
- **Supervisors:** JF Coeurjolly (Univ. Grenoble Alpes) and Franck Picard (ENS Lyon)
- **Prerequisites:** strong background on statistics and probability, software implementation



Expression of gene ACTA2 in serial sections of colorectal samples with Xenium (left) and CytAssist (a spot-based technology like Visium. The Visium technology is based on spots that are regularly spaced (lattice data).

Single-cell transcriptomics now allows for the quantification of gene expression at the scale of individual cells, encoded in count matrices containing thousands of observations (cells) and tens of thousands of features (gene expression values). The analysis of such data requires new methodological frameworks dedicated to their complexity and size. Recent technological breakthroughs have further enabled the profiling of the spatial heterogeneity of gene expression, with high-throughput sequencing directly performed on tissue sections, revealing tremendous potential. Recent publications have started to explore and benchmark the analysis of spatial transcriptomics, but methodological challenges remain before the scientific community can fully exploit the potential of these technologies.

Despite its tremendous potential, the spatial process framework has been little explored for modeling the spatial component of gene expression, particularly in capturing heterogeneities in this variability. Spatial point processes, which are stochastic processes dedicated to spatial random variations, could serve as a foundation for developing a more robust statistical test. In this project, the candidate will investigate the spatial heterogeneity and dependence of the expression of a single gene (univariate) to start with. Since the technology captures the sub-cellular location of target transcripts, the standard exploratory analysis will involve examining the first- and second-order structures of the point pattern. The first-order intensity function serves as a starting point [1], as it roughly measures the probability that the expression of the target gene is located in the vicinity of a point. The second one helps to capture clustering or repulsive characteristics of the point pattern. Thus, when identifying spatially variable genes, the two following questions are of interest: are the locations homogeneous or inhomogeneous in space and do these genes locations depend on each other or not. We classically inspect the intensity function say $\rho(x)$ which measures the probability to observe a point in the vicinity of x and the Ripley's-K function, $K(r)$ which measures the normalized mean number of points in a ball centered in each point with radius r.

The tested hypothesis ($\rho(x) = \rho_0$ or $K(r) = K0(r)$ where $K0(r)$ is the K-function for a Poisson point process which models independence) can be global, thus testing the overall homogeneity of the spatial pattern or the dependence at each scale. However, given the spatial complexity of tissue (especially cancerous tissue), this global hypothesis is likely to be rejected in most cases. Therefore, we propose adopting a more localized approach by testing. This strategy presents a particular challenge known as continuous testing [2], where an infinite number of hypotheses must be considered, as there are as many hypotheses as locations. Thus, the multiplicity of tests needs to be properly accounted for.

The candidate will develop a statistical test procedure based on point patterns and will implement a false discovery control for this continuum of hypothesis testing, at least under the assumption that the point pattern originates from a Poisson point process (the reference model for generating points without interaction) to treat the two problems stated before (inhomogeneity and independence). For the inhomogeneity problem, using 2D continuously sliding scans, the output will be a map highlighting areas with significant deviations from the homogeneity assumption. Based on standard benchmarks, we expect to find numerous inhomogeneous areas in tissue sections, making stringent false discovery control crucial. For the independence problem, the output will be a union of intervals of r values such that we observe a departure to the Poisson model. For the two problems, mathematical developments (to obtain statistical guarantees in terms of false discovery rate for instance) and software implementation (within R for example) are expected. The works will have to be compared to existing literature. Finally, it is expected that procedures are applied to real spatial transcriptomics datasets.

The candidate will be co-supervised by Franck Picard (CNRS, ENS Lyon) and Jean-Francois Coeurjolly (UGA, Grenoble), experts in computational statistics, statistical learning and point processes. The candidate will work at the ENS de Lyon or at the University of Grenoble, in an interdisciplinary environment, between mathematics, computer science and biology. Moreover, the candidate will benefit from the Artificial Intelligence for single-cell based medicine (AI4scMed) PEPR project that gathers an interdisciplinary consortium in machine learning / IA dedicated to single cell genomics, with experts in machine learning, optimal transport and statistics.

Contact: franck.picard@ens-lyon.fr , jean-francois.coeurjolly@univ-grenoble-alpes.fr

[1] J. Møller and R. P. Waagepetersen. Statistical inference and simulation for spatial point processes. CRC press, 2003
[2] F. Picard, P. Reynaud-Bouret, and E. Roquain. Continuous testing for poisson process intensities: a new perspective on scanning statistics.Biometrika, 105(4):931–944, 2018
[3] Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. Genome Biol 21, 31 (2020).