Title: Strategies for integrating spatial data across individuals

EBI PI: John Marioni
Sanger PI: Sarah Teichmann

Fully understanding cell type identity in a complex multicellular system fundamentally requires the integration of a cell's expression profile with its spatial location within the tissue under study. To date, transcriptomics studies have focused primarily on profiling dissociated populations of cells where precise spatial information is lost. From the perspective of a cell's location with the 3D architecture of a tissue, single-molecule RNA FISH and other related approaches allow gene expression levels to be measured whilst preserving spatial location. Typically, this has and is possible only for a subset of genes rather than the whole transcriptome. Recent technological advances promise to overcome these limitations – using approaches such as highly-multiplexed RNA FISH or spatial transcriptomics, they promise to simultaneously profile the expression of hundreds or thousands of genes within single cells, while reporting on spatial location within the tissue of interest.

In parallel to these technological advances, it is critical to develop computational methods for analysing such data. Our groups have been pioneers in this domain, developing approaches for uncovering patterns of spatial heterogeneity (Svenson *et al*., Nat Methods, 2018; Ghazanfar et al., Nat Methods (in press)) and for integrating dissociated datasets with spatially resolved atlases (Achim et al., Nat Biotechnol., 2015; Petit et al., PLoS Comp Biol, 2014). However, numerous questions remain unanswered. In this postdoc, we propose that the candidate will tackle a fundamental challenge: how can spatially resolved expression profiles from multiple individuals be combined to generate a coherent atlas?

By taking an entire tissue and dissociating it, scRNA-sequencing can, in principle, provide an unbiased representation of all cell types present. However, approaches for spatial transcriptomics are limited to sampling only a small region of a tissue. Additionally, given different physical characteristics between individuals, it is extremely difficult to know whether the same spatial location is being sampled across individuals.

Overall, we propose to address the question of how can such data be combined to create a meaningful spatially-resolved reference atlas?

We expect that this project will explore warping and integration algorithms, drawing from machine learning and deep learning. During the course of this project, we expect that we will be able to answer specific questions relating to experimental design such as:

i) what is the number of distinct spatially resolved samples from an individual that are needed to construct a reference?
ii) what is the resolution and plexity of spatial data that is required?
iii) is it better to sample more individuals or to sample comprehensively from a single individual?
iv) how can uncertainty about 3D location of a sample be accounted for?

v) what data from different types of spatial technologies can be integrated in an informative way?

To this ends, we will take advantage of the newly launched High-throughput Spatial Genomics Initiative (HTSG) at the Wellcome Sanger Institute and EMBL-EBI. We will consider published data, as well as unpublished heart,thymus and other tissue and organ sample datasets generated within HTSG to address the overarching computational challenges described above.