**Data science approaches for integrative modelling in spatial proteomics**

Aberrant protein localisation has been implicated in numerous diseases, including cancers (Kau et al., 2004), obesity (Siljee et al., 2018), and multiple others (Laurila and Vihinen, 2009). Characterising the sub-cellular localisation of proteins in health and disease is therefore of critical importance in order to understand the pathobiological mechanisms of many diseases. We have recently established the value of adopting a Bayesian probabilistic modelling approach for inferring the sub-cellular localisation of proteins on the basis of mass spectrometry (MS) based spatial proteomic data (Crook et al., 2018). Other recent advances have demonstrated the benefits of adopting integrative approaches, in which additional data sources such as gene ontology (GO) annotations and protein-protein interaction (PPI) datasets are exploited in order to improve the inference of protein sub-cellular localisation (Breckels et al., 2016). In this project, we will extend the model introduced in Crook et al. (2018) to allow additional data sources to be incorporated, according to the MDI ("Multiple Dataset Integration") model for Bayesian integrative genomics (Kirk et al., 2012). We will fit the resulting model to a number of proteomics datasets, and will assess how well it is able to predict sub-cellular protein localisation compared to our existing (single data source) model. The method will then be applied to proteomic data from the human osteosarcoma cell line (U2-OS) to obtain novel predictions of protein sub-cellular localisations in cancer cells. We can then validate any predictions using fluorescence microscopy data made available by the Human Protein Atlas (Thul et al. 2017).

The project will be supervised by Paul Kirk and Oliver Crook at the MRC Biostatistics Unit, as well as by Laurent Gatto from the Computational Proteomics Unit. This project offers an exciting opportunity to learn about data science, Bayesian inference, proteomics, and coding in the R statistical programming language.

**References:**


