Bayesian graphical models for strain resolution from hybrid metagenomics sequencing

Dr Christopher Quince (Warwick Medical School) and Professor Xavier Didelot (SLS – Statistics)

The diversity of microbial communities, or microbiomes, and their importance to both human health, for example the gut, and the environment, has only become apparent through the development of improved DNA sequencing technologies. These are now cost effective enough to enable sequencing of an entire community of microbes, i.e. the metagenome. However, the challenge of entirely resolving the genomes of the individual species and strains present in a community from metagenomics sequence has yet to be solved. Recently, long read technologies, such as Oxford Nanopore enable far longer reads than before but these sequences are noisy. This project proposes to develop methods that draw on Bayesian graphical models to integrate these long noisy reads with the short accurate reads that are available from standard next generation sequencing, this is termed a 'hybrid' assembly approach.

We have already developed methods for representing the key structure in genome assembly, the assembly graph, as a Bayesian graphical model and demonstrated that it can be solved for short reads with methods such as variational inference. Currently, though, our results are limited to local regions of the assembly graph enabling strain resolution on conserved genes but not the reconstruction of entire genomes. This project would build on this to develop novel graphical models that can represent both long and short reads.

We will begin by generating synthetic hybrid data sets from known metagenomes using existing software. This will provide us with a ground truth for methods development. We will construct assembly graphs from the short reads and then map the long reads onto them again using existing software. We will represent this combined data set as a single graphical model and explore strategies for its solution ranging from exact MCMC methods such as Gibbs samplers to approximate solution from variational Bayes or expectation propagation. Successful methods will then be implemented in this or follow up projects as a bioinformatics pipeline for hybrid metagenome resolution. We will apply these methods to an existing data set from a clinical study of changes in the human gut microbiome during a course of dietary treatment for Crohn's disease.

This project has the potential to transform the study of the human and environmental microbiomes with great relevance both for bacterial pathogens and the symbiotic microbiota. This Masters project could lead to a PhD through either of the two NIHR funded Health Protection Units based at the University of Warwick, Gastrointestinal Infections or Genomics and Enabling Data. This would allow the PhD to join a dedicated training programme in an area of high medical relevance. Please e-mail Dr Christopher Quince with any enquiries regarding this project (c.quince@warwick.ac.uk).