Bioinformatic Discovery of Clinically Relevant Splicing Variants in Public Datasets

The Genomics Innovation Unit (GIU) at Guy's and St. Thomas' NHS Trust is a unique, exciting research body within the NHS. The GIU develops novel genomic diagnostic tests for cancer and rare diseases, applying cutting-edge molecular biology and computational techniques in the process. Our research is application-driven, forming the vital link between translational research and clinical implementation, and culminating in tangible improvements to patient care. Our work takes genomic biomarkers through the life cycle of diagnostic test development, from proof-of-concept studies through to validation and accreditation. In the short time since the GIU's creation we have developed a long-read sequencing genomic stratification test for Acute Myeloid Leukaemia (AML), which promises more personalised treatment for thousands of cancer patients and significant savings for the NHS.

The GIU is now looking to expand its capacity strategically by recruiting MRC PhD students to take novel ideas through the proof-of-concept stage. The project for this post involves developing a computational approach to better characterise clinical DNA variants of uncertain significance. In the process of genetic testing, thousands of patients worldwide have their exomes or genomes sequenced. This routine sequencing generates thousands of genetic variants, of which many are predicted to affect protein coding sequence via missense or nonsense mutation. However, the remaining synonymous variants in coding regions and variants in noncoding regions within genes often do not have any obvious consequences. A subset of these variants of uncertain significance (VUSs) likely will affect the transcripts of the genes in which they occur, either by affecting transcript expression levels, or by affecting splicing patterns. Both paradigms leave characteristic signatures that can be searched for in RNA-seq transcript sequencing datasets.

This project involves systematically searching for the effects of VUSs on expression levels and splicing patterns on the genes that contain them. This will be done by massive data mining of the largest global sequencing repository: the Sequence Read Archive from the NCBI (https://www.ncbi.nlm.nih.gov/sra). This data source contains petabytes of data, so specialised computational methods need to be used to efficiently search them for specific DNA sequences.

In this project, you will implement a computational pipeline to put together existing software to do the following:

- Construct a local searchable database of human RNA-seq datasets from the SRA, using tools such as Bifrost
- For each VUS:
    - Search the database for RNA-seq datasets with sequences containing the VUS
    - If any such datasets are found, compare datasets with the VUS to those without the VUS to look for differences in transcript expression and transcript splicing patterns, using existing tools such as kallisto and Trinity.

The VUSs from this pipeline that show the most significant effects will be selected for experimental validation. This study will contribute to clinical knowledge by assigning

function to human variants that are currently labelled as having “uncertain significance”. If you are looking to use your computational skills to help create something tangible that will ultimately improve the patient pathway for patients with rare diseases, this project is for you. You will also have a chance to be the first author on any paper resulting from this work, depending on how much is completed during your project.

Essential Skills / Experience:

- MSc or PhD in computational biology, bioinformatics, or related field (PhD can be in progress)
- Experience analysing genome sequencing data
- Strong programming skills in python
- Ability to think critically and creatively, and constructively challenge colleagues when troubleshooting

Desired Skills / Experience:

- Software development experience (i.e. version control, unit testing, etc.)
- Experience building software with workflow management systems such as Nextflow and/or Snakemake
- Experience building containerised software with Docker