COMPUTATIONAL BIOLOGY

Group Leader

Crispin Miller

Research Scientists
Eva Freckmann
Holly Hall
Alexandrina Pancheva
Andrew Papanastasiou

Graduate Students
Britt van Abeelen
Ekansh Chauhan
Jennifer Muscat
Mayank Sikarwar
Boyu Yu

Rapid advances in technology are leading to a wealth of high-dimensional data describing the behaviour of cells in normal and tumour tissue. We are using computational approaches to interrogate and integrate these high dimensional data in order to develop a more holistic view of the altered regulatory processes that lead to the development and progression of cancer.

A major focus of the Institute is to use multiple 'omics and imaging modalities to generate a more holistic view of the processes that occur in tumour tissue. The goal is to use these data to stratify patient populations to generate a more granular view of the underlying biology of a given tumour. We then use data to position our pre-clinical models of disease against these more tightly defined patient subsets, supporting forward translation from discovery science into the clinic, and back translation from the clinic into our experimental models. Holly Hall has been working, first with the Bird lab to position liver cancer models against patient cohorts, and then to do this at scale using pan-cancer public domain datasets. In parallel, Andrew Papanastasiou is developing novel algorithms and approaches to support the analysis of these more holistic datasets, applying a mixture of techniques from Artificial Intelligence, Large Language Models and Deep Learning. His work is showing that single cell and organoid data are also applicable to these disease positioning approaches as well as in *in vivo* models.

While considerable attention has been directed at the regulation of transcription, many of the downstream processes such as the control of RNA processing, splicing, and mRNA stability are also under tight regulatory control. The translational machinery that governs when, and how these mature mRNAs are translated into correctly folded proteins is similarly constrained. A critical question, therefore, is how is the information that defines these systems encoded within the genome?

Our work exploits the availability of a large and diverse cohort of well annotated genome sequences from different species. This allows comparative genomics to be used to pursue regulatory patterns from an evolutionary perspective. In parallel, the availability of large cohorts of DNA- and RNA-sequenced patient tumour samples makes it possible to explore the evolutionary constraints placed upon different regions of the genome by selection pressure from within the tumour environment. In both cases, the available data are now at sufficient scale to support classical- and neural-network based machine learning algorithms, and we are applying these in combination with mathematical models that draw upon ideas from information theory.

We are collaborating with the Bushell and Le Quesne groups to explore the role of regulatory sequences embedded within coding sequences, how mutations and changes in the regulatory machinery in and around these regions can impact on protein levels. Eva Freckmann is interested in how these regulatory patterns impact on gene expression across human tumours. Britt van Abeelen is exploring how patterns of tRNA usage interact with the translational control machinery and how these are altered in tumour cells. Boyu Yu is investigating the regulatory sequences embedded in the untranslated regions of protein coding genes, and how these sequences are used by cells to regulate mRNA stability and protein translation. In collaboration with Ke Yuan at the University of Glasgow Computer Science, and David Robertson at the University of Glasgow Centre for Viral Research (CVR), we have been supported by the DiRAC high performance computing facility who have enabled us to use their considerable computing power to build state of the art Large Language Models (LLMs) of biological sequences. Alex Pancheva has created an exciting LLM of RNA sequences that are starting to reveal candidate sequence elements with potential relevance in cancer.

We are also part of PREDICT-Meso, a £5m Accelerator project funded through a partnership between CRUK, Fondazione AIRC, and Fundación Científica de la Asociacion Española Contra el Cáncer (FC AECC). Mesothelioma is an incurable cancer that typically develops years after inhalation of asbestos dust and fibres. The factors that underpin the development of mesothelioma are currently poorly understood. We are applying computational approaches to study 'omics data arising from multiple tumour types including mesothelioma, colorectal and liver cancer samples.

Underpinning all these algorithms is a requirement to perform computationally intense calculations across thousands of genome sequences with matched transcriptome and proteomics data. We have worked with Naveed Khan to commission a High-Performance Computing system that underpins our data science efforts across the Institute.

Publications listed on page 124