# COMPUTATIONAL BIOLOGY

Group Leader
## Crispin Miller

Research Scientist
Tamara Luck

Graduate Student
Boyu Yu

The computational biology group is interested in how the processes that control gene expression are altered in tumour cells, how these changes occur, and how they drive oncogenic transformation and tumour progression. We are studying these systems by using classical- and deep-machine learning approaches to study multiomics datasets arising from clinical and *in vitro* studies.

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.

The group was established in 2019, and over the last year we have continued to develop our research programme. Tamara Luck, a postdoc in the group, is interested in regulatory sequences embedded within coding sequences, and how mutations in and around these regulatory sites can impact on protein levels. Boyu Yu, a new graduate student, co-supervised with the RNA and Translational Control in Cancer Group, led by Martin Bushell, 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.

Rapid advances in technology have made it possible to generate simultaneous measurements across the same cell and tissue samples. These can describe a diversity of changes in genome structure and organisation, mRNA expression and protein levels. These present a computational challenge not only in terms of the mathematical models required to properly integrate and analyse these complex multi-omics datasets, but also in the mapping of these data into clinical datasets arising from, for example, tumour RNA-seq. We are particularly interested in strategies that support the joint analysis of single-cell and bulk sequencing datasets.

Underpinning all these algorithms is a requirement to perform computationally intense calculations across thousands of genome sequences with matched transcriptome and proteomics data. Over the last year we have been working with the Information Services team to expand the High-Performance Computing infrastructure that will underpin our data science efforts across the Institute.

Publications listed on page 98