BIOINFORMATICS AND DATA SCIENCE

Head
Crispin Miller

Scientific Computing Specialist
Naveed Khan

Bioinformaticians
Beto Bermudez Barrientos
Robin Shaw

Software Engineers
Mayank Sikarwar
Ifedayo Ojo

The variety of data generating platforms within the Institute make it possible to generate 'deep tissue phenotyping' data in which different modalities combine to provide a more holistic view of tumour tissue. The Unit provides support across the Institute for the analysis of this diversity of data. A major aspect of our work is to develop the data management strategies to deal with the high volumes of multimodal and imaging data.

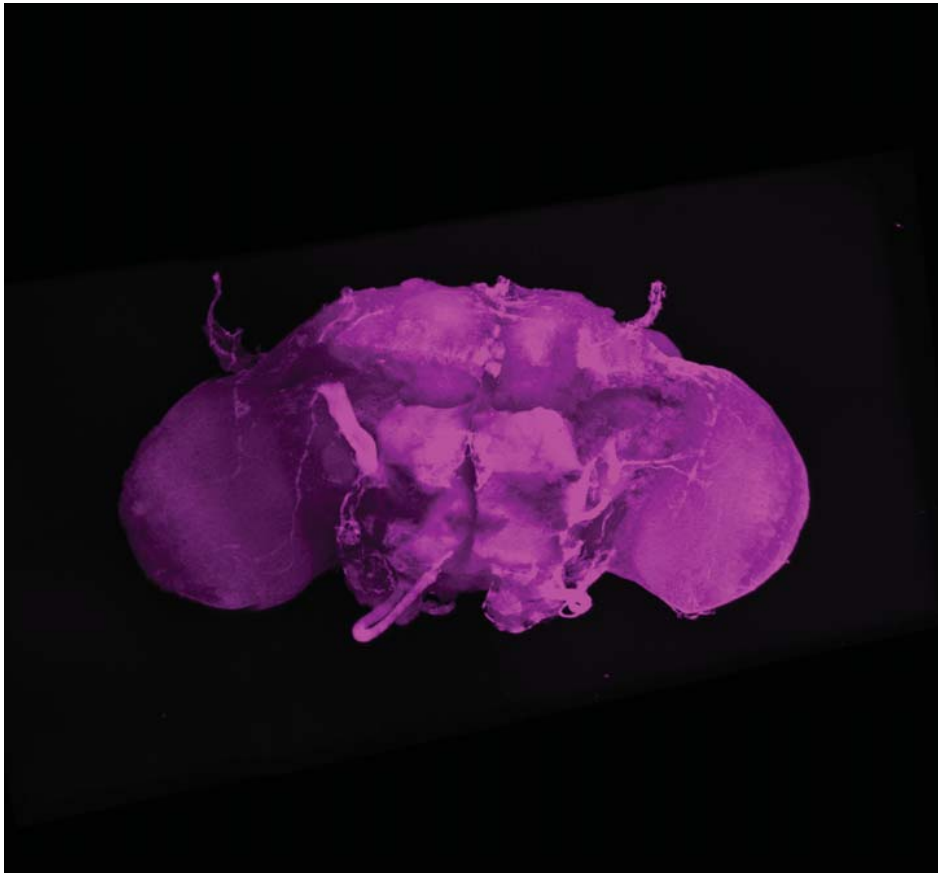
Our ultimate goal is to provide insights that enhance our understanding of cancer biology. The need for DNA and RNA sequencing analyses has continued to grow, and this has been accompanied by continued interest in using computational and machine learning approaches to interpret imaging and proteomics data. The Institute has access to a range of spatial transcriptomics platforms including Xenium and CosMX, and the data management and analysis demands of these platforms is becoming increasingly challenging, not least because of the volumes of data they produce. In collaboration with the Computational Biology Group, we are developing Next Flow workflows to support the systematic processing and management of spatial transcriptomics data.

Advances in technology are leading to a rapid increase in the size of the data we are analysing, leading to significant increase in our computing requirements. Naveed has commissioned a High Performance Computing (HPC) system that combines conventional processing with GPUs and a fast filesystem in order to support our data science, AI and deep learning needs. Naveed is also working closely with IT Services on the provision of Virtual Machines (VMs) to support non-HPC tasks. Robin and Naveed are also working together with IT services to standardise data processing workflows through our computing platforms and the different filesystems within the Institute.

Mayank and Ifedayo are working together to generate additional software to tools to support data sharing according to FAIR principles (that data should be **F**indable **A**ccessible **I**nteroperable and **R**eusable) and are developing NextFlow workflows to provide standardised analysis workflows for a range of modalities in addition to the spatial transcriptomics platforms described above. This is also exploiting synergies between the needs of the CRUK PREDICT-Meso accelerator and parallel needs across the Institute and the CRUK Scotland Centre. To this end Ifedayo is working with a team from the West of Scotland Safe Haven and NHS Greater Glasgow and Clyde.

Data analysis and modelling is performed using a variety of open-source software environments, programming languages and scripting tools, including Python, R, Bioconductor, Bash, PHP and Perl. We frequently make use of analytical routines that have been developed in-house, and/or in collaboration with our colleagues from the areas of mathematics, statistics, computer science and biology. We use a mixture of academic software tools for functional annotation, clustering, enrichment, ontology and pathway analysis, as well as commercial tools.

The unit also provides support and guidance to graduate students and postdocs in other research groups who are using computational approaches to analyse their data. This includes advice on R scripting (by appointment), experimental design, and data presentation. Our team also participates in delivering part of the Cancer Research & Precision Oncology MSc programme at the University of Glasgow.



Expansion Microscopy of Drosophila Brain - *Nikki R. Paul and Jack Holcombe*