School of Mathematics and Statistics
Faculty of Science, Technology, Engineering and Mathematics

2022 PhD Projects

Project title	Statistics for Biologists: Leveraging user design for transcriptomics analysis
Principal supervisor	Wendi Bacon (LHCS)
Second supervisor	Stefanie Biedermann (M&S)
Further supervisors	Paul Mulholland (KMI); Andrew Stubbs (Erasmus Medical Center in the Netherlands); Berry Kriesels, Omnigen (Netherlands)
Discipline	Statistics/Biological informatics/Mathematical biology
Research area/keywords	Human computer interaction, statistics, bioinformatics, computer science, next generation sequencing
Suitable for	Full time applicants, Part time applicants
Industry Partner Details / ££	Enhanced studentship (approx. £1500/month) for 4 years; 3-6 month placement with industry partner: Omnigen in the Netherlands, housing/travel fully covered while on placement, Student visa fees included (if necessary)

Project background and description

Statistics and mathematics underpin the algorithms that biologists use to analyse the world - often with minimal statistical training. Single-cell RNA-seq analysis (scRNA-seq) is a cutting-edge bioinformatics field aiming to identify all the cell types and subtypes within an organism, as demonstrated in the global, Chan-Zuckerberg Initiative funded Human Cell Atlas project[1]. ScRNA-seq is increasingly becoming a necessity for biological research, uniting computational biologists, mathematicians, statisticians, computer scientists, bioinformaticians, and biologists. Because of this breadth of expertise, the distance between biologist or bioinformatician using the algorithms and the mathematics underpinning them is far. Fancy algorithms are great – but not if people can't use them. How do we bridge the gap?

The Galaxy Platform allows users to analyse data without programming skills, which helps bridge the gap. The Galaxy Training Network[2] and the annual GTN Smörgåsbord (led by Dr Hiltemann, Stubbs lab, Erasmus MC) provides a platform for high quality bioinformatics training using Galaxy. However, users can still struggle to apply the analyses to their own messy data, and are ever limited by the tools that currently exist in Galaxy. Given the popularity of scRNA-seq, and the growing demand for advanced techniques (particularly for spatial data analyses), the student will focus on this field for their project. They will interrogate private and public datasets with our

industry partner Omnigen as proof of principle, identifying biomarkers in disease. They will develop much-needed tools and training materials for advanced single-cell and spatial analyses within the Galaxy platform. They will use human computer-interaction design methods to assess and improve the tools, training, and most importantly, the decision-making by users, with an emphasis on engaging non-mathematicians in the vital statistics they (often incorrectly) use. They will use the psychology of algorithm comprehension to embed statistics into these materials, both within the tools and the training materials themselves. This student will evaluate this unique pipeline in how we develop accurate analyses – and analysts – of the future.

Background reading/references

1. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. and Teichmann, S. A. The Human Cell Atlas: from vision to reality. Nature 550, 451–453 (2017).
2. Batut, B. et al. Community-driven data analysis training for biology. Cell Syst. 6, 752-758.e1 (2018).