Engineering AI to map broken and repaired DNA during genome replication.

Principal supervisor's name: Dr **Michael Boemo**
Principal supervisor's email address: mb915@cam.ac.uk

Principal supervisor's CRUK CC theme: Cell and Molecular Biology Programme

Department for student registration: Department of Pathology
Department or institute where research will take place: Department of Pathology

Postgraduate scheme: **MRes + PhD (1 + 3-year non-clinical applicants only)**

MRes project outline:

Over the next hour, your body will replicate enough DNA to span about 28 round trips from the Earth to the Moon. It is critical that DNA replication is timely and accurate: The frequent stalling or slowing of replication forks, often called “replication stress”, is a hallmark of cancer and can lead to genomic instability. However, replication stress can also targeted by chemotherapeutic agents both in the clinic and in clinical trials. There is a pressing need to understand how replication stress leads to genomic instability, both to better understand oncogenesis and to learn how to exacerbate replication stress to selectively kill cancer cells. The Boemo Lab in the Cambridge Department of Pathology is a computational biology group that approaches this problem through a combination of mathematical modelling, data science, and AI. It includes members from the biological, physical, and mathematical sciences and involves interdisciplinary and collaborative work with labs in Cambridge, around the UK, and around the world. The purpose of this project is to develop new algorithms to detect sites in cancer genomes that were broken and subsequently repaired during DNA replication, leading to the first high-resolution, genome-wide maps of these sites across the genomes of different cancer types. The project lies at the intersection of AI and genomics at a time when both of these fields are undergoing a revolution and it presents the opportunity to learn, or build upon, computational skills in a supportive environment.

MRes experimental plan:

The Boemo Lab has developed software that can detect thymidine analogues (such as BrdU) incorporated into nascent DNA sequenced on the Oxford Nanopore platform (Müller & Boemo, et al. Nature Methods 2019; Totañes, et al. Nucleic Acids Research 2023). Oxford Nanopore have recently released duplex sequencing which, in the context of DNA replication, enables the sequencing of both the nascent and parental strands. Stalled or collapsed replication forks can be converted into double-strand breaks that can subsequently be restarted by homologous recombination. To generate a negative control, human RPE1 cells will be cultured in the thymidine analogue BrdU for one cell cycle. Positive control cells will be treated with BrdU for one cell cycle together with DNA alkylating agent methyl methanesulphonate (MMS). We will further test an additional condition of cells treated with MMS together with an ATR inhibitor (VE-821) to block the intra-S-phase checkpoint. In the negative control, we anticipate that duplex sequencing with Oxford Nanopore followed by analogue detection will reveal one unlabelled strand (parental) and one strand labelled with high levels of BrdU (nascent). In cells treated with DNA damaging agents, we anticipate duplex sequencing to reveal patterns of BrdU incorporation indicative of HR-mediated fork restart, as labelled and unlabelled patches on single molecules

reveal the strand exchange between parental and nascent DNA. The student will develop algorithms to efficiently identify molecules containing these patterns and map them to the genome.

PhD project outline:

Having established this method in human RPE1 cells with induced DNA damage, the student will transition to cancer genomes. In particular, the Boemo Lab has access to a panel of breast cancer cell lines where some lines are resistant to DNA replication inhibitors while others are susceptible. The student embark on a longer-term trajectory of determining where forks stall in these lines, whether these forks collapse or are restarted, the pattern of stalling and restart across the genome, how this correlates to genomic features, and how and why these patterns change in response to treatment with chemotherapeutic agents.

PhD experimental plan:

Preparation of wet lab samples will be similar to that of the MRes project, but the PhD phase will bring a significant increase in the scope of software development. The development of this technology is expected to be reliant on the accurate alignment of Oxford Nanopore signals to the reference genome. Therefore, in addition to the bioinformatics and data analysis objectives outlined above, the PhD phase will involve the development of new algorithms and strategies to perform these alignments using deep learning methods (such as connectionist temporal classification) with a particular focus on making these algorithms robust to structural variations. Taken together, this will result in a novel computational method which we anticipate will create a step change in the field. This method will be developed into portable, easy-to-use software and published together with the biological insights outlined above.

Main techniques:

- Bioinformatics skills (e.g., sequence alignment, BAM/SAM/fasta/fastq files)
- Opportunity to learn, or improve upon, software engineering in Python/C/C++/shell as well as version control with Git
- Use of both CPU and GPU high-performance compute nodes
- Deep learning with Google TensorFlow
- Management and efficient parsing of large datasets

Key references:

- Totanes, F.I.G., Gockel, J., Chapman, S.E., Bartfai, R., Boemo, M.A.†, Merrick, C.J.† (2023) A genome-wide map of DNA replication at single-molecule resolution in the malaria parasite Plasmodium falciparum. Nucleic Acids Research.
- Jones, M.J.K.†, Rai, S.K., Pfuderer, P.L., Bonfim-Melo, A., Pagan, J.K., Clarke, P.R., McClelland, S.E., Boemo, M.A.† (2022) A high-resolution, nanopore-based artificial intelligence assay for DNA replication stress in human cancer cells. bioRxiv.
- Shaikh, N.*, Mazzagatti, A.*, Bakker, B., Spierings, D.C.J.E., Wardenaar, R., Maniati, E., Larsson, P., Wang, J., Boemo, M.A., Foijer, F., McClelland, S.E.† (2022) DNA replication stress generates distinctive landscapes of DNA copy number alterations and chromosome scale losses. Genome Biology 23:223.
- Mueller, C.A.*, Boemo, M.A.*, Spingardi, P., Kessler, B. Kriaucionis, S. Simpson, J.T., Nieduszynski, C.A.† (2019) Capturing the dynamics of genome replication on individual ultra-

long nanopore sequencing reads. Nature Methods 16:429-436.
- Gaillard, H., García-Muse, T., Aguilera, A.† (2015) Replication stress and cancer. Nature Reviews Cancer 15:276-289.