**EPSRC industrial CASE studentship in “Statistical methods for assessing and using chromatin structure and accessibility to understand gene regulation and human phenotypes”**

**Supervisor:** Chris Wallace (MRC Biostatistics Unit, University of Cambridge) in collaboration with David Willé (GSK)

**Location:** MRC Biostatistics Unit, University of Cambridge

**Sponsors:** EPSRC and GSK ([www.gsk.com](http://www.gsk.com))

**Project overview**

Assays such as DNase1 and ATAC-seq and Chi-C are becoming increasing widespread to investigate gene regulation but continue to pose many analytical challenges. This proposal would consider statistical approaches to assays for open chromatin and/or chromatin contact with special focus on differentiation between different cell types and their relation to other 'omics technologies, e.g. single cell RNA-seq. Methods development in any of these is possible as is consideration of the distribution and organisation of accessible regions or hypersensitivity sites, the effects of time and stimuli and similarities between related cell types. Applications to the use of GWAS SNPs and causal inference could also be considered. The work could involve additional collaboration with GSK scientists, EMBL and the OpenTargets consortium.

**Detailed description**

Following recent advances in molecular biology, in particular ENCODE and other similar consortia, there is an increasing awareness that chromatin availability and structure is important for gene regulation and the differential expression of genes between different cell and tissue types and states. Many molecular techniques – DNase1 and ATAC-seq being two of the best known – have been proposed to investigate the chromatin state across the genome but have to date received less statistical attention than for example the analysis of gene expression data. Techniques such as Hi-C are used to probe chromatin structure, but resolution can be low. Current methods are limited, and new ones are urgently required given the importance of these data. This PhD proposal is an attempt to address that. We aim that the student will shape the project according to their interests, within the broad topic outlined.

Particular topics for investigation could include, but are not limited to:

- the distribution and organisation of such assessable regions or hypersensitivity sites across the genome and how best to incorporate their relationships into any related analysis
- the variation of such signals between cell types, or between the same cell type in different environments (eg stimulated vs unstimulated), taking into account their hierarchy and relationships borrowing information between cell types or lines as required, with the aim of identifying drivers of cell type differentiation
- the relationship between genotype, chromatin state and/or structure, cell type and gene expression (either bulk or single cell). We are particularly interested in the use of such assays to understand the mechanisms underlying genetic associations with human phenotypes. Statistically key challenges could include the construction of appropriate schemes of inference to model the relationships between different signals and platforms, the use of Bayesian methods and where possible the use of data from different platforms to elucidate causal relationships between genes and cell types and states.

The project will be based on existing public domain data sources (eg ENCODE, BluePrint). We will also utilise data generated through the OpenTargets collaboration within which the industrial partner GSK has an active role and future projects such as the Human Cell Line Atlas which all look to epigenetic and gene expression signals and their relationship to differences between different cell types and states. Particular interest within the OpenTargets consortium includes efforts to characterise the disease relevance of particular cell lines to specific cancers or more
general models as well as potentially the regulation of epigenetic controls in response to changes in cell state. Any software developed will be made publicly available, e.g. as an R package.

**Details of the research setting**

The student will be primarily based at the MRC BSU, an internationally recognised research unit specialising in statistical modelling with application to medical, biological or public health sciences. Details of the work carried out in the Unit appear on our Research page [https://www.mrc-bsu.cam.ac.uk/research-and-development](https://www.mrc-bsu.cam.ac.uk/research-and-development). The supervisor, Chris Wallace, is a statistician and scientist with an interest in using genomics to understand human disease, for more information see [http://chr1swallace.github.io](http://chr1swallace.github.io).

The iCASE studentship is co-sponsored by GSK – one of the world’s leading research-based pharmaceutical and healthcare companies – is committed to improving the quality of human life by enabling people to do more, feel better and live longer (for further information please visit [www.gsk.com](http://www.gsk.com)). The student will gain valuable research experience in both an academic and an industrial setting with appropriate research supervision provided by both the Academic (Dr. Chris Wallace) and Industrial (Dr. David Willé) supervisors throughout the course of this collaborative research project. The collaboration will involve the student spending a minimum of three-months at a GSK research facility.

**Start date:** 1 October 2018 or earlier.

Informal enquiries addressed to cew54@cam.ac.uk are welcomed.

All application queries regarding eligibility should be directed to phdstudy@mrc-bsu.cam.ac.uk

**How to Apply**

Applications should be made on-line via [http://www.graduate.study.cam.ac.uk/applicant-portal](http://www.graduate.study.cam.ac.uk/applicant-portal) selecting course details MDBI22 PhD in Biostatistics