PI profile

## **Jerome Kelleher**

|  |  |
| --- | --- |
|   | **Dr. Jerome Kelleher****Titles**: Robertson Fellow in Biomedical Data Science, Group Leader**Location**: Big Data Institute**Department**: Nuffield Department of Medicine **Group**: Kelleher **Webpage**: <https://www.bdi.ox.ac.uk/Team/jerome-kelleher>**Email**: jerome.kelleher@bdi.ox.ac.uk |

GMS themes:

* Genomic analysis (bioinformatics and statistical genetics)

### Research Overview

Computational genomics is increasingly being driven by very large datasets consisting of hundreds of thousands of whole genomes. Existing computational methods and infrastructure are struggling to cope with this scale, and there is an urgent need for efficient algorithms and high-quality software. Our research focuses on two fundamental problems in population and statistical genetics: the ability to analyse genetic variation datasets at the population scale and inference of the genealogical history of these vast datasets. The methods revolve around the “succinct tree sequence” data structure, which is a highly efficient encoding of genetic ancestry that has the potential to solve the most pressing problems of scale, as well as provide deep insights into evolutionary processes. We develop efficient algorithms, and implement them in production quality, open-source software using state-of-the-art software development practices, in collaboration with a globally distributed network of scientists and software engineers.

Project areas: Large scale genomics, genealogical analysis, population genetics, software development

### Specific project proposals:

* ‘Improved inference of genetic ancestry’

Please contact directly for further information.

***These pages were reviewed/updated:2021-07-23***

Project proposal

# **Title**: Improved inference of genetic ancestry

Supervisors: Jerome Kelleher

Wet/dry lab mix (approx): 0% wet lab, 100% dry lab

### Description:

Recent breakthroughs in computational genomics have made it possible to infer genetic ancestry in recombining organisms at scale for the first time, making numerous downstream applications possible. A number of different methods have recently been proposed, providing deep insights into human evolution. However, these methods are in their infancy, and much work remains to be done before they are ready for mainstream genomics. Our group developed the "tsinfer" method, which is capable of accurately inferring genetic ancestry for millions of whole genomes, based on the "succinct tree sequence" data structure. This method of encoding genetic ancestry has also lead to performance improvements of multiple orders of magnitude in genome simulation and statistical computation, and has the potential to solve many of the major computational problems facing large scale genomics. For example, the data compression levels achieved by the tree sequence data structure are so high that it is in principle possible to store the ancestral history of 10 billion humans in around 1TB of storage.

In this rotation project and potential extension into a DPhil, you will use simulations and human data to investigate areas in which tsinfer's accuracy and computational performance can be improved. You will implement updates to the core algorithms in tsinfer's Python and C codebase as part of an open-source development process. Specific areas for development include:

* Better heuristics for ancestral haplotype generation
* Improved recombination breakpoint detection
* Detailed analysis of patterns of recurrent mutations and their relation to sequencing error
* Incorporation of uncertainty via probabilistic ancestor generation and stochastic HMM traceback
* Better performance via more fine-grained parallelisation strategies
* Better scalability by distribution across multiple machines

### Training Opportunities:

This project will suit a student interested in a DPhil focused on computational statistical genomics, and will include a large software development component.

### Background reading / references:

* Kelleher et al. **(2019)**. Inferring whole-genome histories in large population datasets. Nature Genetics. <https://doi.org/10.1038/s41588-019-0483-y>
* Wohns et al. **(2021)**. A unified genealogy of modern and ancient genomes. Preprint. <https://doi.org/10.1101/2021.02.16.431497>
* Project website: <https://tskit.dev>