

Big Data Institute
Li Ka Shing Centre for Health Information and Discovery



Current address
Wellcome Trust Centre for Human Genetics
Roosevelt Drive, Oxford, OX3 7BN
Tel: +44(0)1865 287534
mcvean@well.ox.ac.uk
www.well.ox.ac.uk/gil-mcvean

Gil McVean FRS FMedSci
Professor of Statistical Genetics
Director of the Oxford Big Data Institute
Li Ka Shing Centre for Health Information and Discovery

The Editor
Bioinformatics

03 January 2017

Please find attached the manuscript *Deconvoluting multiple infections in Plasmodium falciparum from high throughput sequencing data*, which we wish to be considered for publication as an original paper in Bioinformatics.

Here, we introduce statistical and computational methodology to solve an important problem in pathogen genomics, namely the widespread occurrence of individuals infected with more than one strain of pathogen. With the growing use of whole genome sequencing as a tool in routine infectious disease epidemiology, the lack of suitable methods for coping with such multiply-infected individuals will limit our ability to infer key details of epidemiology, such as transmission routes and the spread of drug resistance. There are existing algorithms for estimating the number of different strains present, however we believe this is the first method capable of inferring both strain number and their haplotypes. We have validated the method through application to experimentally mixed samples and have considered how factors such as reference panel size and composition influence both scaling and accuracy. We make available source code and an R implementation to enable others to use the algorithms.

We believe that the method will have widespread interest to readers of Bioinformatics, both because of the statistical novelty and the applications it makes possible. The software is already being used in an international project to provide an open-access resource for malaria genomics, the Pf3k Project, and, with minimal adjustments, can be used in many other contexts where multiple strains may be present in sequence data.

If you would like any further information, please do not hesitate to ask.

Yours sincerely,

A handwritten signature in black ink, appearing to read "G. McVean".

Gilean McVean FRS FMedSci
Professor of Statistical Genetics and Director of the Oxford Big Data Institute
University of Oxford