MRes Proposal

Project title: Evolution of Long QT syndrome in human populations

Student: Shiyun Liu
Contact email: s.liu18@imperial.ac.uk

Supervisor: Dr. Matteo Fumagalli
Contact email: m.fumagalli@imperial.ac.uk

Keywords:
Population genetics; Evolution genetics; Positive selection; Machine learning; Convolutional neutral network; Long QT syndrome.

Background:
Long QT syndrome is featured by an extended QT interval on electrocardiogram. This prolonged interval reflects the abnormality of heart repolarization after a heartbeat. Patient with Long QT syndrome have increased chance of palpitation, drowning, fainting and sudden death [1]. Candidate gene mutations associated with high risk of developing Long QT syndrome have been identified and reported [2]. Moreover, it has been proposed that natural selection on some alleles in key genes (e.g. NOS1AP), which are linked to the disease phenotype, is conferring to the high frequency of this disease [3].
In recent years, an alternative approach to classic association studies on revealing genetic basis of complex phenotypes has emerged. The evolutionary approach is established based on the principle that natural selection is prone to carry important functions to the carrier, and the selection process can leave certain patterns on the genome. For example, positive selection can be recognized as the rapid increase of a selected allele can result in a long haplotype/linkage disequilibrium around it [4]. Therefore, part of the evolution of genetic diseases can be revealed by detecting the positive selection signatures. One of the most recent and powerful approaches to locate positive selection events is by machine learning techniques, where convolutional neural network based machine learning pipeline is becoming a promising tool to estimate selection events and strength. [5]

Problem and Aim:
In this project, we intend to explore the evolution of long QT syndrome in human populations by analysing population genomics data via machine learning.
In general, we want to:

1. Retrieve genetic variants associated to Long QT syndrome
2. Implement a convolutional neural network(CNN) to detect positive selection patterns that may contribute to the disease phenotype and also quantify the strength of positive selection.
3. Compare this new method with previous machine learning methods on positive selection detection. Test the capacity of CNN on a clinical-related genetic problem and find out potential problems and possible optimizations of the method.
4. Infer functional impacts of the relative selections occurred in human populations and its linkage to the disease phenotype.

Proposed methods:

The genetic basis of the disease can be collected from previous association studies by literature review.

The population genomics data could source from publicly available data sets, such as 1000 Genomes project data base.

The main approach that will be using and tested in this project is the convolutional neural network implemented in the pipeline. The procedures include:

1. Generation of training and testing sets for machine learning using simulations of population genomic data.
2. Conversion of genomic data to images that can be recognized, compared and calculated by CNN pipeline.
3. Training and testing of the network with the image sets and calculate the discrete probability distribution for selection coefficient.
4. Examine the Long QT syndrome genomics data and suggest positive selection events.

The functional impact might be evaluated via combination of our study result and human genomic database.

Anticipated outputs:

We currently expect this new method to perform a more efficient and accurate detection of positive selection pattern along with selection coefficient. This can help us better understanding the evolution of Long QT syndrome and its genetic basis as well as functional impacts.

Project feasibility:

1st month: Literature review on Long QT syndrome and algorithm study of CNN.

2nd-4rd month: Implement the pipeline on targeted genetic problem.

5-6th month: Evaluate the result and carry out improvement and comparison.

6th month: Functional impact mining and project report wrapping up.

7th month: Re-evaluate of the whole project and get feedback from the supervisor. Modify the report and finishing up the project by tiding up the data, coding scripting and result.

Project budget:

Possible transportation and accommodation fee for attending useful conferences (e.g. held by Data Science Institute): 250£

Reference:

[1] Morita et al. The QT syndromes: long and short. Lancet (2008)

[2] Pfeufer et al. Common variants at ten loci modulate the qt interval duration in the qtscd study. Nature genetics (2009)

[3] Newton-Cheh et al. Genetic determinants of qt interval variation and sudden cardiac death. Current opinion in genetics & development (2007)

[4] Pavlidis et al. A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research (2017)

[5] Schrider, D.R., Kern, A.D.: Supervised machine learning for population genetics: A new paradigm. Trends in Genetics (2018)

I have seen and approved the proposal and the budget.

Supervisor: Dr. Matteo Fumagalli

Date: 10 Dec 2018

Signature: