Title : Accounting for population structures in bacterial genome-wide association studies

Research project within an international company ; no PhD is forseen.

Advisor: Magali Jaillard Dancette, magali.dancette@biomerieux.com , +33 4 78 87 2000

Laboratory : bioMérieux company, trainee located at Marcy l'Etoile (West suburbs of Lyon)

Context: Genome-wide association studies (GWAS) aims at studying correlations between genetic variants and a phenotypic trait of interest, observed in a population. The main statistical issue is the identification of significant associations from millions of explanatory variables, measured by high-throughput technologies such as next-generation sequencing.

The acceleration of the acquisition of new resistances to antibiotics by bacteria has become a major worldwide public health concern. In this context, GWAS seem an appropriate tool, to better understand the genetic bases of resistance mechanisms and to identify new markers.

However, the clonal reproduction of bacterial strains makes their genomes highly correlated in clade structures which is a source of confusion increasing the risk to identify false associations. Several methods, including linear mixed models, allows for the integration of this population structure in the association model, in order to correct its effect.

Objective : The main objective of the training is to evaluate several strategies to assess the population structure to be integrated to the linear mixed model. Most strategies indeed use a matrix of single nuclear polymorphism (SNP) observed among the genes which are in common to all strains (core genes). Here we propose to also evaluate strategies allowing to account not only for the core genes, but also the accessory genome, mostly resulting from horizontal gene transfers.

This work may be directly applied to complex association studies such as inter-species GWAS, and may be valorized as a participation to the redaction of a scientific publication.

Competences required : We are looking for a master student in biostatistics, or in statistics (MSIAM master track Statistics (STAT)) with a great interest for biological applications, or in bioinformatics with good skills in statistics. The trainee will read an produce documents in English and will use R statistical language and scripting tools (shell, perl, python, …) for data processing.

Bibliography :

Hoffman, G. E. (2013). **Correcting for population structure and kinship using the linear mixed model: theory and extensions**. PLoS One, 8(10), e75707.

Earle, S. G., *et al.* (2016). **Identifying lineage effects when controlling for population structure improves power in bacterial association studies**. Nature Microbiology, page 16041.

Jaillard, M., et al (2017). **Representing Genetic Determinants in Bacterial GWAS with Compacted De Bruijn Graphs**. Cold Spring Harbor Labs Journals, doi:10.1101/113563.

Contacts : magali.dancette@biomerieux.com. Application on bioMerieux recruitment plateforme at http://biomerieux-recruitment.com, using job reference : 48485.