PhD project

Title: Exploring the genetic traces linked to the epidemic success of the Beijing *Mycobacterium tuberculosis* lineage

Research group: Infection, Genetics and Epidemiology of Emerging Pathogens (IGEPE) I2BC, CNRS-CEA-Université Paris-Saclay, Gif-sur-Yvette.
Supervisor : Guislaine Refrégier
Co-supervisor: Christophe Sola

Key words: Whole genome sequencing, *Mycobacterium tuberculosis*, evolutionary genetics

Overview:

Tuberculosis is still the world's leading infectious disease. In this project, we propose an extension of the molecular study of our in-house DNA extracts from clinical strains, the analysis of the corresponding sequences, and a massive exploitation of public databases to understand the most currently transmitted tuberculosis lineage in the world, the L2/Beijing lineage. We will characterize a wide range of DNA samples in candidate loci using high throughput genotyping methods and will use public genome and SNP databases to identify further evidence of the link between specific mutations and virulence. Our goal is to improve the characterization of virulence genetic determinants to suggest new paths for disease surveillance.

Thematic:

Tuberculosis (TB) is still the world's first fatal infectious disease (WHO, 2017). Further understanding of *Mycobacterium tuberculosis* pathogenesis is needed to identify new targets for future drugs and/or features to design new vaccines and other therapies providing adequate and sustainable immunity.

Domain: Bioinformatics, Evolutionary genetics, Health, Diagnostics.

Goal: We aim at performing an evolutionary genetic analysis of L2/Beijing strains from various sources (local DNA extracts, public genomes and SNP databases) to identify genes involved in virulence.

Context:

Lineage L2/Beijing is one of the seven currently identified main lineages of human tuberculosis (Wan et al, 2011; Shitikov et al, 2017). Certain L2/Beijing strains are more virulent than the average in animal models (Hanekom et al, 2011), and have been responsible for several epidemics in humans, regardless of their antibiotic-susceptibility status.

Public genome and SNP databases are now massive information sources to explore traces of selection in tuberculosis evolution, with 65,000 sequence read archives linked to biological samples referred to as "*Mycobacterium tuberculosis*" in the databases as of November 2018, plus an additional 100,000 genomes from the “100,000 genomes project”. This information makes it possible to test whether specific sequences found in clonal virulent strains have been selected for or against (Pourcel et al, 2017, Aguileta et al, 2009).

New techniques emerge that can help recover deeper and/or more extensive information on pathogen diversity. The Oxford Nanopore Next-Generation Sequencing Technology MinIon,

producing long reads, allows the exploration of not only Single Nucleotide Polymorphisms (SNPs) in non-repeated regions and the approximate position of any deletions, but also SNPs in repeated sequence regions and the exact position of deletions. New genotyping platforms can be used to target specific variants at a very high throughput and at a low cost (up to 10 targets for less than 2 euros per sample) (Abadia et al, 2011).

Methods:

We will proceed with Whole Genome Sequencing of a set of L2/Beijing strains of interest. Using Illumina sequencing, we will retrieve high quality SNPs that serve as reference, and Nanopore long-read sequencing to retrieve full genomes including repeated regions, to be able to explore all sources of diversity, including rearrangements inside highly similar Proline and Glutamine-rich so-called PE-PPE genes known to be involved in host-pathogen interactions (Ates et al, 2018). This method has only recently become reliable (Bainomugisa et al, 2018) and corresponding data for a wide range of tuberculosis strains are not yet available.

Raw Illumina sequencing data will be analyzed in parallel on BioNumerics software, and using standard Unix tools to check for reproducibility. The partners of this project have extensive experience in whole genome SNP investigations, including the detection of admixtures of strains in read archives (Sobkowiak, 2018). Sequencing data will be used to correlate virulence to genome sequence.

We will use a SNP-candidate approach to check for recurrent SNP selection within tuberculosis diversity based on the analysis of 1) public database genome sequences (*in silico* data) and 2) retrievable genotypes from a 10 year local world-wide representative DNA collection using high-throughput genotyping methods (*in vitro* data). This approach relies on the rationale that an advantageous SNP will occur independently at a frequency related to the selective advantage it provides. For instance, drug resistance markers have occurred many millions of times independently. Several SNPs have already been identified as outliers by population genetics methods, likely because they confer increased virulence; corresponding genes are implicated in host-pathogen interactions (Merker et al, 2015; Osorio et al, 2013). In addition, specific SNPs within L2/Beijing evolutionary branch have also been described as potentially advantageous. For *in vitro* data, we will select a subgroup of these candidates SNPs: eight SNPs of functional relevance (Rv0176_290E; Rv1872c_3V; Rv1872c_109T; Rv2524c_ 2844573H; Rv1270c_ 1419658T; Rv1527c_ 1722228R; mutT2_58; mce3F_410) and 5 SNPs for phylogenetic assignment to lineages L1, L2, L3, L4, or L0-5-6.

Expected Results:

The combination of whole genome sequencing and targeted SNP genotyping analyses will identify the frequency at which SNPs of interest were selected for during tuberculosis evolution. This frequency will allow us to classify SNPs according to their potential role in virulence. Additional correlation studies using geographical data and available clinical information may help to document whether some SNPs may be prone to selection in specific contexts such as for drug-resistance. Additional correlations may be drawn, depending on the results obtained *in vitro* by collaborators studying a macrophage infection model.

The most significant SNPs will in the end be used to build a diagnostic test to detect them in *M. tuberculosis* isolates. This diagnostic test could contribute to accurate surveillance of tuberculosis in Europe where L2/Beijing prevalence is increasing, but could also be proposed to Asian countries that suffer from very high L2/Beijing burden.

References

Abadia, Refregier, Sola et al. Infect Genet Evol. 2010 Oct;10(7):1066-74. doi: 10.1016/j.meegid.2010.07.006. Epub 2010 Jul 17. PMID:20624486.
Aguileta G, Refrégier G, et al. Infect Genet Evol. 2009; 9(4):656-70. Review. EPub 6 Apr 2009. PMID: 19442589.
Ates LS, et al. Nat Microbiol 2018; 3(2):181-188. doi: 10.1038/s41564-017-0090-6. Epub 2018 Jan 15.
Bainomugisa et al, Microb Genom. 2018 Jul;4(7). doi: 10.1099/mgen.0.000188. PMID:29906261.
Hanekom, Tuberculosis(Edinb). 2011 Nov;91(6):510-23.doi:10.1016/j.tube.2011.07.005. PMID:21835699
Merker et al. Nat Genet. 2015 Mar;47(3):242-9. doi: 10.1038/ng.3195. PMID:25599400
Osorio et al, Mol Biol Evol. 2013 Jun;30(6):1326-36. doi: 10.1093/molbev/mst038.
Pourcel, Vergnaud et al, PLoS One. 2017 Jan 6;12(1):e0169684. doi: 10.1371/journal.pone.0169684.
Shitikov et al, Sci Rep. 2017 Aug 23;7(1):9227. doi: 10.1038/s41598-017-10018-5. PMID:28835627.
Sobkowiak et al. BMC Genomics. 2018 Aug 14;19(1):613. doi: 10.1186/s12864-018-4988-z.
Wan, Vergnaud, Pourcel et al, PLoS One. 2011;6(12):e29190. doi: 10.1371/journal.pone.0029190.
WHO, Tuberculosis annual report 2017. www.who.int/tb/publications/global_report/fr/

Material and financial conditions :

We have a running lab of 35 m². Large bioinformatic resources in the team and the institute are noteworthy.

International openness:

All TB projects are international. The team has worldwide connections that are very active (part of a South-American European group exploring TB diversity worldwide), exchanges with international collaborators (Russia, Netherlands, Switzerland, Australia, etc.).

Collaborations:

A specific collaboration has been set-up with a neighbouring research group and a group involved in clinics from Lyon. Additional collaborations are underway with Spain, the Netherlands (RIVM), etc.

Valorization:

Scientific publications, contributions to international conferences (ESM, MEEGID, IMMEM, etc.)

Searched profile and skills :

Bioinformatics, Evolutionary genetics, Molecular Biology. The candidate should be acquainted with the basics of all the three domains of the project. It is acceptable if one or two of the domains has only been studied from a theoretical point of view. Entering a team of professors in an institute with large human resources and with numerous collaborations, the candidate should prove autonomous and interact with a variety of collaborators (both on site and distant).

Supervision methods: daily interactions (out of teaching duties), mensual meetings, email exchanges.

French level: none

English level: B2