

Exploiter l'intégralité des données de séquençage de *Mycobacterium tuberculosis* : une plateforme pour l'analyse génomique in-silico à grande échelle

La tuberculose demeure un problème de santé publique majeur à l'échelle mondiale. Selon l'Organisation mondiale de la santé, en 2022, la tuberculose était la deuxième cause de décès dans le monde due à un seul agent infectieux. Depuis plus d'une décennie, l'étude de cette maladie peut également se faire par l'analyse des génomes obtenus de nombreuses souches de tuberculose. Avec le développement des méthodes de séquençage et la réduction de leur coût, on compte actuellement plus de 160 000 séquences de génomes disponibles publiquement. Il s'agit d'une source de données d'une grande richesse mais qui reste une source de données brute, qu'il reste à analyser. Cette thèse introduit la plateforme TB-Annotator qui a permis l'analyse de l'intégralité des données de séquençage de *Mycobacterium tuberculosis* publiquement disponibles. Cette plateforme met en oeuvre de nombreuses méthodes in-silico spécialement développées et adaptées à l'extraction de toutes les caractéristiques génomiques propres au MTBC, notamment dans la détection des variations structurales et dans l'optimisation des pipelines bioinformatiques. Les capacités d'analyse en temps réel de cette masse d'information par TB-Annotator ont permis plusieurs découvertes pour la phylogénie et la phylogéographie du MTBC. Cette thèse décrit notamment la découverte d'une nouvelle lignée rare, la lignée 10, mise en évidence par 2 séquences publiques non classifiées après analyse des 160 000 séquences disponibles. Plusieurs améliorations de la définition de lignées et sous-lignées sont aussi décrites, notamment L2, L4, L6 et L7. L'ouverture de cette plateforme devrait permettre de simplifier l'accessibilité des analyses à grande échelle pour l'épidémiologie, la phylogénétique, la phylogéographie et la résistance aux médicaments.

Mots-clés : : Tuberculose, Santé publique, Génomique, Séquençage, *Mycobacterium tuberculosis*, Variations structurales, Phylogénie, Phylogéographie, Résistance aux médicaments, Lignées

Leveraging the entire sequencing data of *Mycobacterium tuberculosis* : a platform for large-scale in-silico genomic analysis

Tuberculosis remains a major public health issue worldwide. According to the World Health Organization, in 2022, tuberculosis was the second leading cause of death globally due to a single infectious agent. For over a decade, the study of this disease has also been possible through the analysis of genomes obtained from numerous tuberculosis strains. With the development of sequencing methods and the reduction in their cost, there are currently more than 160,000 publicly available genome sequences. This is a rich data source, but it remains raw data that needs to be analyzed. This thesis introduces the TB-Annotator platform, which enabled the analysis of all publicly available *Mycobacterium tuberculosis* sequencing data. This platform implements numerous specially developed in-silico methods adapted to extract all genomic features specific to the MTBC, particularly in detecting structural variations and optimizing bioinformatics pipelines. The real-time analysis capabilities of this vast amount of information by TB-Annotator have led to several discoveries regarding the phylogeny and phylogeography of the MTBC. This thesis notably describes the discovery of a new rare lineage, lineage 10, identified through 2 unclassified public sequences after analyzing the 160,000 available sequences. Several improvements in the definition of lineages and sub-lineages are also described, notably L2, L4, L6, and L7. The opening of this platform should simplify the accessibility of large-scale analyses for epidemiology, phylogenetics, phylogeography, and drug resistance.

Keywords: Tuberculosis, Public health, Genomics, Sequencing, *Mycobacterium tuberculosis*, Structural variations, Phylogeny, Phylogeography, Drug resistance, Lineages