博士論文（要約）

Fine-mapping for Tuberculosis (TB) susceptibility and Genome-wide association study for drug-resistant TB

(結核感受性領域における高密度マッピング及び薬剤耐性結核のゲノムワイド関連解析)

ウオン　ジン　ハオ

Wong Jing Hao

博士論文の要約

論文題目　Fine-mapping for Tuberculosis (TB) susceptibility and Genome-wide association study for drug-resistant TB

（結核感受性領域における高密度マッピング及び薬剤耐性結核のゲノムワイド関連解析）

氏名　Wong Jing Hao

Tuberculosis (TB) is a serious infectious disease caused by the inhalation of the pathogen *Mycobacterium tuberculosis* (*M.tb*), and affects patients globally and is one of the biggest causes of death, along with infection by the Human Immunodeficiency Virus (HIV). The majority of cases of TB in 2014 were seen to occur within the South East Asian and Western Pacific regions, with approximately 58% of incident cases reported in these regions. Recently, the rise of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB has made the situation even more serious. MDR-TB refer to *M.tb* pathogen that are resistant to at least both isoniazid and rifampicin, the two most powerful first-line anti-TB drugs. XDR-TB refers to *M.tb* pathogen that are resistant to one fluoroquinolone and a second-line injectable anti-TB drug, in addition to both isoniazid and rifampicin. These factors make TB extremely difficult to control and treat. The immunopathogenesis of TB involves the actions of a number of immune cells and cytokines in the host that result in the formation of a granuloma structure, with the core of the structure made up of macrophages that have up-taken the pathogen. The granuloma is thought to act as a barrier to contain the bacteria and stop the infection from spreading. A patient would enter a period of dormancy in containment of the *M.tb* pathogen is successful, but will suffer a re-activation of the disease should containment by the granuloma fail.

An estimated one-third of the world's population is infected with the *M.tb* pathogen, however, only about 5-15% of those infected ever develop the clinical symptoms of TB disease within two years, leading to the speculation that there may be genetic factors involved in the host susceptibility to TB. Previous studies done using twins also showed that there was a 2.5-fold higher concordance rate for TB in monozygotic twins compared to dizygotic twins, further reinforcing the idea that genetic factors are involved in TB susceptibility. Numerous studies have been conducted in a many different populations to try to identify these genetic factors, including

genome-wide linkage studies (GWLS), candidate gene association studies and more recently, genome-wide association studies (GWAS). A large number of genes, genetic regions and loci have been seen to be associated with TB susceptibility, however, the lack of replicability among the regions and genes identified among the different populations point to the complex nature of host genetic susceptibility factors for TB. Additionally, many studies have also been conducted in the genome of the *M.tb* pathogen, and many mutations in the pathogen genome have been identified that confers resistance to anti-TB drugs. However, pathogen genome mutations cannot completely explain susceptibility of the host to drug-resistant TB, and studies to identify host genetic susceptibility factors are also needed. A number of studies have been conducted in the Human Leukocyte Antigen (*HLA*) region and a number of *HLA* alleles have been identified in the Indian and Korean populations, as well as the *SLC11A1* gene in Japanese, to be associated with drug-resistant TB susceptibility of the host. The present study consisted of two parts. In the first part, I attempted to identify new, as-yet un-identified genetic factors affecting TB susceptibility in Thais, using next-generation sequencing (NGS) and a variant detection, filtering and prioritization workflow. In the second part of the study, I conducted a GWAS in an Indonesian population to search for genetic susceptibility factors affecting host susceptibility to drug-resistant TB.

A previous GWLS conducted in Thai family samples identified a susceptibility region on the chromosome (chr.) 5q23.2-31.2 region that showed suggestive evidence of linkage for susceptibility for TB. Subsequently, another single nucleotide polymorphism (SNP) association study was conducted in Thai family samples on the chr.5q31 region, which was part of the earlier identified susceptibility region. A number of SNP haplotypes were identified to be significantly associated with TB susceptibility. For the first part of this current study, a candidate region was defined by centering on the top SNP in the most significantly associated SNP haplotype in the earlier Thai family association study and extended in both directions up to approximately 1Mb. This candidate region was captured and extracted from 18 Thai samples using custom-made in-solution hybridization probes and subsequently sequenced using the Ion Torrent Personal Genome Machine (PGM) NGS platform. Two software were used to detect variants from the sequencing data. After applying stringent filtering criteria, such as requiring at least 20x read depth coverage of detected variants, and variants needing to be detected in both software as well as seen in at least two or more sequenced samples, a total of 904 variants were identified. Variants that were located in genes with minor allele frequencies (MAF) of at least 5% in the 1000 Genomes Project Southern Chinese (CHS) population were prioritized, and SNPs that were previously studied in another GWAS done in Thais were excluded. A total of 188 SNPs remained after this filtering step and from this list, 35 tagSNPs were identified before being put through the Ensembl Variant Effect Predictor (VEP) software to further

prioritize candidate SNPs according to their predicted functional consequences. Five SNPs were identified as candidates for further analysis, and the NGS calls for these SNPs were confirmed using Sanger direct sequencing, before a further case-control association analysis was conducted using the Taqman genotyping assay in another set of 663 Thai TB cases and 774 controls. It was observed from the association analysis that a SNP, located in the 3'-Untranslated Region (UTR) of a gene, showed moderate evidence of association with TB susceptibility in the genotypic model (p = 0.012) and the dominant model (p = 0.029, OR = 1.43, 95% CI = 1.04 – 1.96).

In the second part of the study, a GWAS was conducted using the Illumina HumanOmniExpressExome-8 v1.2 BeadChip array, that interrogates approximately 960,000 SNPs across the human genome. The case samples used in the study were 160 drug-resistant TB cases and control samples consisted of 192 drug-sensitive TB patients. The samples were collected from the Indonesian islands of Java and Madura. Quality control of the data was conducted, which included filtering out SNPs that had less than 99% genotype call rates and excluding SNPs that had MAF > 5%, as well as SNPs that failed the Hardy-Weinberg Equilibrium test. Samples which had call rates of less than 99% were also excluded, as well as those that showed cryptic relatedness. Principal component Analysis (PCA) also identified a number of individuals that showed some genetic differences from the main Indonesian sample group and these samples were excluded. Association analysis was conducted on the genotyped SNPs after quality control of the data was performed. The association analysis identified ten SNPs that showed suggestive evidence of association ($p \leq 1.0 \times 10^{-05}$) with drug-resistant TB, with the most significant association seen for a SNP located in a gene on chr. 4 (p-value = 2.83×10^{-07}, OR = 2.33). A number of genes were also observed to contain a number of SNPs showing suggestive and moderate ($p \leq 1.0 \times 10^{-04}$) evidence of association.

Regional genotype imputation was conducted for these gene regions and for regions that contained SNPs with suggestive evidence of association to try to identify potential un-typed SNPs that may show more significance than the ones in the GWAS. An imputed SNP located in chr. 6 was seen to have a lower p-value ($p_{imputation} = 2.04 \times 10^{-06}$) compared to the most significant SNP seen in the GWAS for that region. The association results SNPs showing suggestive evidence of association, as well as the imputed SNP, were validated using Taqman genotyping, and genotype concordance between the two genotyping platform was high (~98%). Given the previous reports of association between *HLA* alleles and drug-resistant TB, *HLA* allele imputation was conducted on the GWAS results. *HLA* imputation identified a number of class I and class II *HLA* alleles showing moderate association with drug-resistant TB in Indonesians, with the most significant associations seen in the class I *HLA-B* and *HLA-C* alleles (p = 0.003). Furthermore, due to previous reports of association between the *BTNL2* gene and TB, as well

as signals in some class II *HLA* alleles being in high linkage disequilibrium with signals in *BTNL2*, a conditional analysis for the *HLA* region on chr. 6 on SNPs showing moderate association in *BTNL2* was conducted. A SNP, located within a cluster of class II *HLA* genes, was seen to have a more significant p-value ($p = 1.29 \times 10^{-04}$) compared to the original GWAS results ($p = 1.53 \times 10^{-01}$). However, the main signals in the region were still seen to be SNPs in the *BTNL2* gene.

A gene-based association analysis was also conducted using the results of the GWAS using the VEGAS2 software, and significant results of the gene-based association analysis (genes with p-values ≤ 0.01) were used to conduct gene pathway, process networks, and gene ontology enrichment analysis. The enrichment analysis was conducted using the Metacore program. A number of gene pathways, process networks, and gene ontologies related to immune system function and immune response to bacterial infection were seen to be significantly enriched ($p \leq 0.05$). The genes identified by the GWAS may have functions related to kidney function and subsequent drug clearance from blood plasma, granuloma formation or maintenance, as well as also potentially being involved in apoptosis and necrotization of the granuloma core, possibly leading to lowered drug penetration and exposure and increasing the risk and chance of developing drug-resistant TB. Further studies are required to elucidate the functional effects of the SNPs on the genes. The results of the enrichment analyses also identified a number of pathways that may be interesting avenues for further investigation into their effects on the development of drug-resistant TB.

In conclusion, the present study manages to identify a new 3'-UTR SNP, located in a candidate susceptibility region on chr.5q31.1, that showed moderate evidence of association with TB in a Thai population. The variant prioritization and filtering workflow was successful in identifying novel SNPs as candidates from NGS data and it would be worthwhile to further expand and improve the workflow. This study is also the first report of a GWAS for drug-resistant TB in an Indonesian population, and identified a number of potential genes and gene pathways, biological process networks, and gene ontologies that are associated with the development of drug-resistant TB.