博士論文

論文題目　Search for susceptibility gene(s) to tuberculosis in the candidate region of Chromosome 20

（20 番染色体上候補領域における
結核の感受性遺伝子の探索）

氏名　中内　彩香

Contents

ABBREVIATIONS……3

ABSTRACT……...6

INTRODUCTION……..7

MATERIALS AND METHODS……...…….....11

RESUTLS…….25

DISCUSSION……..33

REFERENCES…….39

ACKNOWLEDGEMENT……..…….47

TABLES AND FIGURES……....48

Abbreviation

ADAM33	ADAM metallopeptidase domain 33
AIDS	acquired immune deficiency syndrome
bp	base-pairs
CHB	Han Chinese in Beijing, China
Chr.	chromosome
CI	confidence interval
DHS	DNaseI hypersensivity sites
DM	diabetes mellitus
emPCR	emulsion polymerase chain reaction
eQTL	expression quantitative trait loci
gDNA	genomic DNA
GWAS	genome-wide association study
GWLS	genome-wide linkage study
H3K27Ac	acetylation of histone H3 lysine 27
HbA1c	hemoglobin A1c
HIV	human immunodeficiency virus
HLA	human leukocyte antigen
HWE	Hardy-Weinberg equilibrium
IFN	interferon

IMP	inosine monophosphate
indels	insertions and deletions
ISP	Ion Sphere Particle
ITP	inosine triphosphate
ITPA	inosine triphosphatase (nucleoside triphosphate pyrophosphatase)
JPT	Japanese in Tokyo, Japan
LD	linkage disequilibrium
LOD	logarithm of odds
MAF	minor allele frequency
MAVS	mitochondrial antiviral signalling protein
Mbp	mega base-pairs
MDR-TB	multi-drug resistant tuberculosis
MDS	multidimensional scaling
miRNA	microRNA
mRNA	messenger RNA
M. tb	Mycobacterium tuberculosis
NGS	next-generation sequencing
NRAMP1	natural resistance-associated macrophage protein 1
OR	odds ratio
OSA	ordered subset analysis

PANK2	pantothenate kinase 2
PCA	principal component analysis
PCR	polymerase chain reaction
SLC11A1	solute carrier family 11 (proton-coupled divalent metal ion transporter), member 1
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
TB	tuberculosis
TLR	Toll-like receptor
UTR	untranslated region
WHO	World Health Organization
XDR-TB	extensively drug-resistant tuberculosis
95% CI	95% confidence interval

Abstract

Tuberculosis (TB) is a complex disease that both genetic and environmental factors contribute to its development. A number of genetic studies in various populations have been carried out but with little consistency among them.

A previous genome-wide linkage study (GWLS) in Thais identified chromosome 20p13-12.3 as a candidate region for young onset of TB. The present study aims to find any unreported susceptibility genes to young TB within a 1Mbp target region around the top GWLS marker. Next-generation sequencing (NGS) was performed on the region in 13 young patients from Thai multi-case families and functionally interesting SNPs were selected as candidates. Case-control association studies in the Thai population for the SNPs were subsequently carried out.

Among the candidates, rs13830 and rs1127354 in 3'UTR and the second exon on *ITPA* respectively, showed significant associations in young (< 45 years old) cases (P=5.1E-05; OR=0.66; 95% CI=0.54-0.81, and P=1.3E-03; OR=0.72; 95% CI=0.59-0.88 in allelic model, respectively) but not in older cases. These were not identified in the previous Thai genome-wide association study. The finding supports previous studies showing that stratifying by age at onset can be effective in elucidating genetic factors.

To my knowledge, this study is the first attempt at using NGS to gain insight into host genetic factors associated with TB and being the first to report a significant association of *ITPA* with young onset TB. The study also demonstrates the effectiveness of NGS in searching for susceptibility genes in common diseases. To confirm these findings, further genetic and functional studies are needed.

Introduction

Tuberculosis (TB) is one of the three major world-wide infectious diseases along with acquired immune deficiency syndrome (AIDS) and malaria. According to Global Tuberculosis Report 2012 by the World Health Organization (WHO), an estimated 8.7 million incident cases of TB (125 per 100,000 population) was reported globally in 2011, and 1.4 million people were estimated to have died that year [1]. The African and South-east Asian regions are especially burdened, and my interest site, Thailand, is ranked as one of the top 22 high-burden countries, which account for over 80% of the TB cases worldwide **(Figure 1)** [1]. According to the WHO, incidence rates were relatively steady from 1990 to around 2001, before a 2.2% rate of decline was observed between 2010 and 2011 [1]. Thailand experienced a decline in incidence rates within the past 10 years **(Figure 2)**.

Nevertheless, TB still remains a challenging issue for the world's health, especially due to co-infection with human immunodeficiency virus (HIV) and the existence of multidrug-resistant TB (MDR-TB) as well as extensively drug-resistant TB (XDR-TB) [1]. Infection with HIV is known to quicken TB's progression [2]. HIV impairs the immune system and TB-infected patients who are also HIV-positive become sick much more easily than those who are HIV-negative [2]. According to the WHO, 13% of new TB cases are estimated to be co-infected with HIV [1]. The causes of drug-resistant TB include incomplete treatment, where patients stop taking their medicines regularly since they start feeling better, mis-prescription, or unreliable drug supply [2]. MDR-TB is caused by TB pathogen, *Mycobacterium tuberculosis (M. tb)* that is resistant to at least isoniazid and rifampicin, the two main first-line treatment drugs [3]. Furthermore, XDR-TB shows additional resistance to a fluoroquinolone and at least one of kanamycin, amikacin and capreomycin [1]. These

factors make TB control more difficult.

It is estimated that approximately one-third of the world's population is infected with *M. tb*, but only 5-10 percent of infected persons will progress to develop the clinical disease [4]. Early twin study showed that monozygotic twins had a 2.5-times higher concordance rate for TB when compared to dizygotic twins [5]. Referring to several other studies as well, it has been demonstrated that an estimated 36 to 80 percent of heritability can be attributed to genetic factors, which contribute to the outcome of TB [5-10].

In order to identify susceptibility genes, genome-wide linkage studies (GWLS) and case-control association studies of candidate genes were conducted in different populations. Previous GWLS demonstrated several significant loci or suggestive linkage, including chromosome (Chr.) 2q35 in aboriginal Canadians [11], Chr. 8q12-q13 in Moroccans [12], Chr. 10q, 11q and 20q in Brazilians [13] and Chr. 15q and Xq in Africans [14]. However, most of them were not replicated in other populations. Candidate gene studies have also provided us with a variety of clues for susceptibility genes. Until now, quite a few genes have been studied, such as *HLA* [15-19], *SLC11A1* (formerly *NRAMP1*) [20-24], *CD209* [25-28] and *TLR* genes [29-33]. However, not much consistence has been observed among different populations. Candidate genes are usually selected according to prior hypothesis with known biological functions [34], which has been a big limitation when searching for susceptibility genes. Recent advances in genome technology have enabled us to analyse hundreds of thousands up to a million single nucleotide polymorphisms (SNPs) using DNA microarrays. Such studies are known as genome-wide association studies (GWAS). GWAS has an advantage of allowing us to examine the entire genome without a requirement of having prior knowledge of biological functions [35]. One study which combined two GWAS from Ghana and The Gambia followed by a replication analysis reported a locus on Chr. 18q11.2 to be

associated with TB [36]. Apart from the African study, meta-analysis of GWAS in Indonesian cohorts and Russian cohort suggested association for 8 loci, which lie within or near genes that play roles in immune signaling [37]. However, none of the loci reached genome-wide significance level [37], which gives the impression that searching for TB host genetic factors is challenging.

Infection with *M. tb* is caused by inhalation. The disease development is proposed to be classified into three categories; primary TB, reactivation and exogenous re-infection [38]. For primary TB, patients develop the clinical symptoms within one or two years after the first infection [38]. In contrast, secondary TB develops by reactivation of *M. tb* from the first infection after a period of immune protection and/or by exogeneous re-infection [38, 39]. It is speculated that genetic factors may play a major role in primary TB [40]. Young TB patients in developing countries mainly suffer from primary TB, while mainly endogeneous reactivation is seen in the majority of old patients in developed countries with low incidence [41]. These factors allow us to postulate that segregating patients according to age at onset of TB in genetic studies may provide deeper insight into genetic roles in the progression of TB.

In fact, a study in Morocco and Madagascar reported that SNPs in a region were associated with early-onset pulmonary TB before the age of 25 years [42]. Our department has been participating in international genetic studies including Thais, Indonesian, Vietnamese and Japanese populations [16, 28, 33, 39, 41, 43, 44]. We conducted GWAS and replication studies in the Thai and Japanese populations, which identified an at-risk locus in 20q12 in young-onset (< 45 years old) TB, and mentioned that age at onset affects host genetic risks and that stratification may be helpful in elucidating the host genetic factors affecting TB [41].

Our group also conducted a genome-wide SNP-based linkage analysis using Thai affected sib-pair samples [39]. In the study, a region on Chr. 20p was observed to have significant linkage with earlier onset of TB, by an ordered subset analysis (OSA) using minimum age at onset of TB, showing the maximum logarithm of odds (LOD) score of 3.33 (permutation P-value=0.0183) [39].

In this study, the identification of new susceptibility gene(s) and/or SNPs in the young Thai population was attempted in this candidate region on Chr. 20p. There have been limitations that previous GWAS could have missed some possible candidates due to insufficient data about variants and their frequencies, as well as the lack of Thai population information in available public databases. In order to overcome these limitations, next-generation sequencing (NGS), which is a recent novel technology enabling massive and rapid sequencing at comparatively lower cost than conventional methods, was carried out for this region. From the large sequencing data, candidate polymorphisms were selected and case-control association analysis was carried out.

Materials and Methods

1. *Samples*

The current study was approved by the Ethics Review Committees of the Ministry of Public Health in Thailand and Faculty of Medicine, University of Tokyo (Research Project ID: G3356).

1.1 *Samples for variation screening by next-generation sequencing (NGS)*

The current study included case samples from multiplex TB cases per family, which were studied in our previous GWLS [39]. Briefly, families included in this study were largely recruited through a TB surveillance system in Chiang Rai province, located in the northernmost part of Thailand. Diagnosis of TB was by clinical characteristics and microbiological confirmation by sputum culture or at least two out of three positive sputum smears. The patients were tested for HIV using the standard serological test, and HIV patients with TB were excluded from this study. Venous blood samples were collected from the patients after obtaining individual informed consent [39].

Since the previous Thai GWLS, the targeted region, Chr. 20p, was reported to have significant linkage with earlier onset and the range of age at onset was 12-23 years old [39], case samples were selected based on age at onset at 25 years and younger. The youngest available case was selected to represent each family. In total, 13 cases were sequenced **(Table 1)**.

1.2 *Samples for association analysis*

The Thai samples for association analysis included 665 TB patients and 777 healthy

control samples, which were studied in our Thai GWAS [41] (**Table 2**).

The cases were recruited from Chiang Rai, Lampang and Bangkok provinces. Microscopic identification and mycobacterial culture were used for confirmation of TB diagnosis in 98% of TB cases. All cases were HIV-negative when diagnosed with TB. The case samples included 235 samples from patients who are younger than 45 years old (**Table 2**).

The healthy controls were from blood donors in Chiang Rai province. Familial histories of TB were checked for the controls. Individuals with TB associated diseases, such as diabetes mellitus (DM), were excluded. DM status were checked with fasting blood sugar, hemoglobin A1c (HbA1c) or rapid testing of capillary blood and history of DM treatment.

The results of principal component analysis (PCA) performed by our group using multidimensional scaling (MDS) analysis in GenABEL package [45], showed three clusters of population (**Figure 3**). In the GWAS, only the first cluster with largest number of individuals was used for the further analysis and the current study population is also limited to the same cluster.

1.3 *Additional sample set for association analysis*

The additional sample set consisted of 545 TB patients from Chiang Rai province, the Chest Disease Hospital in Bangkok as well as the Payao Hospital in Northern Thailand. Among them, 259 samples were recruited from patients who were younger than 45 years old (**Table 2**).

The 407 control samples were obtained from blood donors in Chiang Rai province and from patients attending hypertension clinics at Chiang Rai regional hospitals. Individual and familial histories of TB were checked for the controls. The control group did not have

any patients with TB associated diseases, such as DM, and they were checked with fasting blood sugar, HbA1c or rapid testing of capillary blood and history of DM treatment.

2. *Definition of the candidate region on chromosome 20*

In our previous Thai GWLS, an ordered subset analysis by minimum age at onset of TB was conducted, which showed that the region Chr. 20p 13-12.3 has significant linkage with earlier onset [39] **(Figure 4)**. Within the region, a SNP marker, rs750702 showed a peak LOD score of 3.33. In order to cover the peak region and to extend the coverage, 1 Mbp around rs750702 was focused on as the candidate region in the current study.

3. *Sequencing of the candidate region on chromosome 20*

3.1 *Designing kit for capturing the candidate region*

The Ion TargetSeq Custom Enrichment Kit for 500Kb-2Mb (Life Technologies) was used to capture the candidate region of approximately 1Mbp around the SNP rs750702, Chr. 20 nucleotides 2,960,407 to 4,050,123. The kit provides a solution-phase DNA probe capture technology which allows selective, specific enrichment of the desired region. I designed and customised by providing the physical position information of the target region to the manufacturer, Life Technologies **(Figure 5)**. The physical position was obtained using the UCSC Genome Browser website, referring to the GRCh37/hg19 assembly [46].

3.2 *Experiments of next-generation sequencing (NGS)*

All experimental procedure was done following manufacturers' protocols. The procedure consisted of the following steps: library preparation, target capturing, pooling the

prepared libraries, template preparation and sequencing.

3.2.1 *Library preparation*

First, genomic DNA (gDNA) samples were sheared into fragments of approximately 200 bases in length using the Ion Xpress Plus Fragment Library Preparation Kit (Life Technologies). Next, the fragmented DNA was ligated to P1 adaptors provided in the library preparation kit and individually barcoded using barcode adaptors provided in the Ion Xpress Barcode Adaptors 1-16 kit (Life Technologies). The barcodes allowed for the detection of which sample each sequenced variant was derived from. The barcoded DNA was then amplified and the libraries for the following experiments were selected according to size by electrophoresis using SOLID Library Size Selection 2% E-gel (Invitrogen) on the E-Gel Safe Image Real-Time Transilluminator (Invitrogen). Fragments with attached adaptors and barcodes which are approximately 330 bases in length were extracted from the gel.

3.2.2 *Target Capturing*

The libraries were purified and amplified according to the protocol provided with the Ion TargetSeq Custom Enrichment Kit. The libraries' molarities were measured using the Agilent 2100 Bioanalyzer High Sensitivity Kit (Agilent Technologies). The purified and amplified libraries were pooled to make sets of libraries of approximately 500ng of gDNA. Each set was then hybridized using the Ion TargetSeq Custom Enrichment Kit protocol, which captured the targeted candidate region.

3.2.3 *Pooling of the prepared libraries*

The hybridized libraries were then checked on the Agilent 2100 Bioanalyzer using High Sensitivity DNA Chips (Agilent Technologies) and pooled for subsequent sequencing.

Three pools were made for 13 libraries in total.

3.2.4 *Template preparation*

Sequencing templates were prepared from the pooled libraries and clonally amplified by emulsion PCR (emPCR) using the Ion OneTouch (Life Technologies) and Ion OneTouch ES systems (Life Technologies) as well as the Ion OneTouch 200 Template Kit v2 DL (Life Technologies). The process of emPCR involves attaching the libraries onto Ion Sphere Particles (ISPs) acrylamide beads provided in the Ion OneTouch 200 Template Kit, by ligation of the previously attached P1 adaptors. The beads are then mixed into micro-droplets made from emulsion oil and PCR reactions take place within these micro-droplets. The libraries attached to the beads are clonally amplified this way [47]. The emulsion oil micro-droplets act as individual compartments to allow simultaneous PCR reactions to take place [48]. The beads with attached template were recovered and the templates enriched by the Ion OneTouch ES system.

3.2.5 *Sequencing by Ion Torrent PGM*

The template beads were annealed with sequencing primers and DNA polymerase and placed in a thermal cycler set to 95ºC for 2 minutes and 37ºC for 2 minutes, according to the procedure in the Ion PGM 200 Sequencing Kit. The beads were next pipetted into an Ion 318 Chip (Life Technologies) according to the protocol. In the loading procedure, the beads were individually deposited into micro-machined wells which are densely packed on the sequencing chip (approximately 11 million wells per Ion 318 Chip) by spinning the chip in a desktop centrifuge [47]. The chips were then loaded into the Ion Torrent PGM for sequencing. The process of sequencing involves having all four nucleotides flowing sequentially over the sequencing chip wells. A nucleotide matching the template base during

the flow will be incorporated in the growing DNA strand by the DNA polymerase and this causes the release of a proton for each nucleotide added in that particular flow [47]. The released proton causes a pH change in the surrounding solution and this pH change is proportional to the total number of nucleotides incorporated in that particular flow. The pH change is detected by a sensor at the bottom of each well and is converted into a voltage signal, which is later automatically changed into a base call by the Ion Torrent PGM signal-processing software [47]. A wash step following each nucleotide flow ensures that the wells are ready to receive the next nucleotide flow [47]. Each sequencing run consists of approximately 500 flows and the data acquired by the PGM is automatically checked to filter out low accuracy reads and quality score values are assigned to the bases. The data is stored in a server for later analysis.

4. *Analysis of Ion Torrent Data*

In order to maintain accuracy, variant detection and subsequent analyses were conducted using two software, Ion Variant Caller plugin and CLC Genomics Workbench v5.5 (CLC Bio).

4.1 *Analysis with Ion Variant Caller plugin*

Sequence reads from the Ion Torrent PGM were automatically mapped to the reference sequence (NCBI GRCH build 37/hg19) and the sequencing primers were trimmed by the Ion Torrent server's Torrent Suite software. The reads were then subjected to the Torrent Suite's Ion Variant Caller plugin variant detection workflow for variant detection. The plugin uses its detection algorithms to detect single nucleotide variants (SNVs) as well as insertions and deletions (indels) variants.

4.2 *Analysis with CLC Genomics Workbench v5.5*

Further analysis was done on the read data extracted from the Ion Torrent server. A separate analysis workflow that was created using CLC Genomics Workbench v5.5 and the reads were put through the analysis workflow. Briefly, the reads were first filtered for quality and trimmed to remove the sequences of attached P1 adaptors. The trimmed and filtered sequence reads were mapped to the reference human genome (NCBI GRCH build 37/hg19) and variant detection was done using CLC Genomic Workbench v5.5's probabilistic variant detection method with the following conditions; 20x minimum depth of coverage for any particular variant and with 95% confidence. The software was set to filter out homopolymer (runs of the same nucleotide bases) indel error calls.

The variant lists of each individually barcoded sample from the Ion Variant Caller plugin were compared against the filtered lists generated from CLC Genomics Workbench v5.5 software. Variants that appeared in both lists were selected to be further analysed.

5. *Selection of candidate SNPs detected by next-generation sequencing (NGS)*

Information of each variant was obtained by using the UCSC Genome Browser, referring to the GRCh37/hg19 assembly.

Among detected variants which were reported to have coverage of 20x or more in both software, variants that were observed in only one sample were excluded from further analysis to increase stringency. Then, variants which were already studied in the Thai GWAS [41] as well as their proxy SNPs were excluded. Proxy SNPs here are defined as SNPs which possess $r^2 \geq 0.8$ in CHB: Han Chinese in Beijing, China +JPT: Japanese in Tokyo population with the studied SNPs. The proxy SNPs were searched using the SNAP website Version 2.2 [49] sourced by 1000 Genomes Project. Variants whose minor allele frequency (MAF) are

less than 0.05 in the HapMap project [50] and/or 1000 Genome Project CHB, if reported, were excluded.

5.1 *Candidate non-synonymous SNPs in exon regions*

All the remaining variants which are reported to be non-synonymous SNPs in exon regions were selected as candidate SNPs **(Table 3)**.

5.2 *Candidate SNPs in 3'untranslated regions (3'UTR)*

Among the remaining variants described in the beginning of this section, those that lay in the 3'UTR of genes were screened to determine if they were located in predicted microRNA (miRNA) binding sites using two databases; microRNA.org - Targets and Expression (http://www.microrna.org/microrna/home.do) [51-53] and miRDB (http://mirdb.org/miRDB/) [54, 55], with prediction score > 80. SNPs which lie in the predicted miRNA binding sites remained as candidates **(Table 4)**.

5.3 *Candidate SNPs in other regions*

In order to prioritise the other polymorphisms which could be functionally interesting, among the remaining variants described in the beginning of this section, the following steps were conducted.

First, each polymorphism was checked to see if it lies in DNaseI hypersensivity sites (DHS) using the HaploReg v2 website (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) [56]. Polymorphisms that are in DHS with motif change were included as candidates. In addition, each candidate's site

was examined with UCSC Genome Browser and polymorphisms which lie in regions with high acetylation of histone H3 lysine 27 (H3K27Ac) markers, indicating that they are located near active regulatory elements, were extracted. Then, polymorphisms which lie in genes were selected. Among them, a SNP shares a proxy SNP, which harbour LD $r^2 \geq 0.8$ in the CHB+JPT population, with another SNP that was studied in Thai GWAS [41] according to the SNAP website Version 2.2 [49] sourced using 1000 Genomes Project and thus excluded. The remaining SNPs became candidates **(Table 4)**.

5.4 *Additional candidate SNP*

The SNP rs1127354 was reported to have strong association with anemia as well as thrombocytopenia induced by pegylated interferon (IFN) and ribavirin therapy for Japanese patients with chronic hepatitis C [57, 58].

Even though the frequency or LD data of rs1127354 on *ITPA* is not available in CHB on HapMap database, a previous report observed high LD in Japanese population between rs1127354 and rs13830, which was detected in the 3'UTR of the *ITPA* gene by NGS [58]. In order to examine the possibility of functional effect of rs13830 on TB susceptibility, the SNP was included as another candidate.

6. *Confirmation of genotypes of detected SNPs*

In order to confirm the genotypes obtained by NGS for the candidate SNPs, conventional Sanger sequencing was performed.

6.1 *Primer design*

First, alignment around each SNP was obtained using the UCSC Genome Browser

and the alignment was inputted to Primer3 v.0.4.0 or its newer version, Primer3web version 4.0.0 [59, 60]. Among the candidates, one or more pairs for each was selected **(Table 5)** and BLAT search using UCSC Genome Browser was conducted to ensure the primers' specificity.

6.2 *Polymerase chain reaction (PCR)*

PCR was performed to amplify the particular region within which each SNP is located to an appropriate amount for the following direct sequencing. Appropriate primer annealing temperatures were determined in advance using reference sample genome and conducting the PCR process with gradient annealing temperatures using TGradient Thermalcycler (Biometra), and checking the amplified DNA bands with electrophoresis **(Table 5)**.

The reaction was performed with FastStart Taq DNA Polymerase (Roche), dNTP (10mM, Roche), 10x PCR Buffer with $MgCl_2$ (Roche), forward and reverse primer sets and 5ng of template DNA on a GeneAmp PCR System 9700 (Applied Biosystems) or a TGradient Thermalcycler (Biometra). The reaction was first incubated at 95℃ for 4 minutes followed by 30 cycles of 95℃ for 30 seconds (denaturation step), 69.1/66.1/64.3/63.3/61.5 ℃, depending on the primer sets, for 30 seconds (annealing step) and 72℃ for 45 seconds (extension step), and by 72℃ for 7 minutes.

6.3 *Sanger sequencing*

The amplicon was labelled using 5x Sequence Buffer (Applied Biosystems), primers, BigDye Terminator v3.1 (Applied Biosystems) and the amplified template DNA. The reaction started with incubation at 96℃ for 3 minutes, followed by 25 cycles at 96℃ for 10 seconds, 50℃ for 5 seconds and 60℃ for 3 minutes, using GeneAmp PCR System 9700 (Applied Biosystems). Following purification of the labelled amplicon with Sephadex G-50

fine (GE Healthcare Bio-Sciences AB), direct sequencing was performed using ABI PRISM 3130*xl* Genetic Analyzer (Applied Biosystems). Results were observed and analysed with Sequence Scanner Software v1.0 (Applied Biosystems).

7. *Selection of tagSNPs of ITPA*

The *ITPA* gene was further studied according to the positive results of the SNPs detected by NGS. For the further investigation, the gene region plus 2kbp upstream and 1.5kbp downstream, Chr. 20p 3,188,006 – 3,206,016, was targeted with the aim to cover the entire gene as well as its promoter region **(Figure 6)**.

First, the positional information was obtained from the UCSC Genome Browser. Then, SNP genotype data of CHB was obtained for close inspection on an Asian population. The data source used here is from HapMap Data Phase III /Rel#2, Feb09, on NCBI B36 assembly, dbSNP b126 [50]. According to the genotype data, its LD block was checked using the software Haploview 4.2 software [61]. The R-squared method was applied for LD score and Confidence Interval method was applied to define the LD block [62]. During this procedure, SNPs which did not pass the following criteria were excluded: Hardy-Weinberg p-value [63] cutoff 0.05, minimum genotyping success rate 95% and minimum MAF 0.05. TagSNPs, which can be representative for other SNPs with r^2 of 0.8 or more, were selected to cover the gene region: rs11087570, rs8362 and rs6139034 **(Figure 6)**.

8. *TaqMan genotyping*

In order to conduct case-control association analysis of the selected candidate polymorphisms, TaqMan assay was conducted for genotyping [64, 65]. A TaqMan probe contains a 5' reporter fluorescent dye, VIC or FAM, and a 3' quencher. First, template DNA

is denatured into a single strand. The forward and reverse primer sets anneal to each strand of the template DNA while either of the probes hybridizes to the target site depending on which allele is present on that strand. The primers extend with the help of DNA polymerase and the polymerase cleaves probes which have bound to the target. The cleavage separates the dye and the quencher, resulting in an increase of fluorescence, which is detected by the machine. According to the difference in intensity of fluorescence from the two dyes, VIC and FAM, the genotype of the target SNP can be determined. When a substantial increase in only VIC-dye fluorescence or FAM-dye fluorescence is observed, the genotype will be determined as a homozygote of either allele. When an increase in both fluorescence is observed, the genotype is called as a heterozygote.

The probes for the candidate polymorphisms were ordered from and manufactured by Applied Biosystems **(Table 6)**. The reagents used in this assay were KAPA PROBE FAST qPCR Master Mix (KAPA Biosystems). 1ng of template DNA was used for each reaction. The assay's reaction started with incubation at 95℃ for 10 minutes, followed by 43-45 cycles at 95℃ for 10 seconds, 60℃ for 1.5 minutes and 72℃ for 1 second, using Light Cycler 480 II (Roche). Genotype clusters were defined and genotypes were determined by the software Light Cycler 480 Software release 1.5.0 SP3 (Version 1.5.0.39). Visual checks were conducted and manual determination used when needed. For genotyping the SNP rs13830 in the additional sample set, Invader assay genotyping [66] was performed by our collaborator, Ms Sukanya Wattanapokayakit.

9. *Case-control association studies*

9.1 *Subgroup analysis in accordance with age in cases*

In order to observe if the SNPs have any association with TB progression, case-

control association studies were conducted. In addition to analysis of cases of all ages versus controls, subgroup analysis was carried out according to age at onset of the cases to examine whether there is any different trend between distinct age groups. In the current study, the cases were divided into two age groups: younger than 45 years old, defined as "young", and 45 years old and above, defined as "old". The 45-year-old age cut-off was applied in this study as epidemiological model and data show TB incidence decreases at around the 40-50 age group in Thailand **(Figure 7)** [67], allowing us to assume there could be different characteristics of TB pathogenesis around 45 years old. In addition, we empirically applied this age cut-off in our previous Asian GWAS and reported a SNP was associated with TB in young cases but not in the old group, showing the distinct pattern according to ages at onset [41].

9.2 *Statistical analysis for case-control association studies*

Chi-square test was applied for the case-control association analysis to evaluate whether there are significant differences in allele and genotype frequencies between cases and controls. For individual SNP analysis, Hardy-Weinberg equilibrium (HWE) test was performed and its p-value was calculated in control samples. The SNPs with HWE p-values under 0.05 were excluded from further analysis. P-values, odds ratios (OR) and 95% confidence intervals (95% CI) were calculated. In the current study, attention was paid not only to the single allele model and genotypic model but also to dominant and recessive models to allow for investigation into any tendency.

In the subgroup analysis by age of the case samples, counts of some columns became small due to small sample numbers. Fisher's exact test was applied to columns with counts less than 6 to improve the statistical accuracy.

10. *Linkage disequilibrium and haplotype analyses*

LD analysis was carried out to examine the degree of linkage disequilibrium between each genotyped SNPs on the *ITPA* gene. Genotype result in this study for rs6115814, rs11087570, rs1127354, rs8362, rs6139034 and rs13830, as well as each sample's status, case or control, were inputted into Haploview 4.2 software [61] and r^2 values were obtained.

Haplotype analysis was carried out to estimate whether a haplotype can contribute to stronger association to TB than each individual SNP. The analysis was performed using Haploview 4.2 software [61] and 10,000 times permuted p-value was calculated for each haplotype.

11. *In silico expression quantitative trait loci (eQTL) analysis for association of rs13830 with ITPA expression*

The following *in silico* eQTL analysis was conducted in order to assess association between rs13830 in 3'UTR of *ITPA* gene and the gene's expression, using Gene Expression Variation (Genevar) database from the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/resources/software/genevar/) [68].

For both eQTL-gene and SNP-gene association analysis, the expression profiling data was obtained from lymphoblastoid cell lines of 80 CHB subjects in HapMap3 project [69] and the NCBI36/Ensembl 50 database was used as reference for SNP and gene data. For correlation and regression, Spearman's rank correlation coefficient (rho) was used as an analysis parameter and 10,000 times permuted p-value was calculated.

Results

1. *Candidate region and its coverage by next-generation sequencing (NGS)*

In the previous Thai GWLS, a region on Chr. 20p13-12.3 was reported to have significant linkage with earlier onset of TB, with a peak marker, rs750702 [39]. In this study, approximately 1Mbp around the marker SNP on Chr. 20 2,960,407 to 4,050,123 was targeted as a candidate region. In order to search for any candidate susceptible polymorphisms, which could not be detected thus far due to the limited data source, the candidate region was sequenced with NGS.

To capture the candidate region, Ion TargetSeq hybridization probes were designed and applied. Probe coverage of the kit for the region was reported to be 87.6 % **(Figure 5)**.

2. *Variants detected by next-generation sequencing (NGS)*

In total, 13 case samples from multiplex TB-affected families, which were studied in the Thai GWLS [39], were used in the NGS analysis. It has been suggested that over 20x sequencing coverage is optimal to reduce inaccuracy in genotype calling [70]. Thus, variants were detected by two software, Ion Variant caller plugin and CLC Genomics Workbench v5.5. Variants with 20x or more coverage on both software were used for further analysis in order to maintain accuracy. The number of detected variants in each sample was 378, 888, 942, 414, 190, 738, 640, 351, 849, 802, 589, 179 and 500, respectively **(Table 7)**. By consolidating duplicated variants among the sequenced samples as one count, 1,878 variants were detected.

3. *Non-synonymous SNPs in exon regions*

3.1 *Selection of candidate non-synonymous SNPs*

Each detected variant was checked with the UCSC Genome Browser (hg19). Among the 1,878 variants, 9 non-synonymous SNPs were already registered; rs6051545, rs1127354, rs2280090, rs2280091, rs709012, rs6037651, rs17857295, rs7262903 and rs2422864. Among them, rs2280091, rs709012, rs6037651 and rs7262903 were observed in only one sample, respectively. In the previous Thai GWAS results [41], rs709012, rs6037651 and rs7262903 were studied and no association with TB was observed (*Mahasirimongkol et al*., personal communication). Furthermore, SNPs rs6051545 and rs22080091 did not pass quality control in the GWAS and their proxy SNPs showed no association. Thus, the 5 SNPs, rs6051545, rs22080091, rs709012, rs6037651 and rs7262903 were excluded. MAF of the SNP, rs2422864, was described to be 0 percent in CHB according to HapMap projects [50], and thus also excluded from further study. After the exclusions, three SNPs, rs1127354, rs2280090 and rs17857295 remained. Further case-control association studies were conducted in order to examine if any of these have association with TB **(Table 3)**.

3.2 *Confirmation of genotypes with Sanger sequencing*

In order to confirm the genotypes obtained by NGS for the candidate SNPs, conventional Sanger sequencing was performed using primers described in **Table 5**. For most of the samples, determined genotypes by NGS and Sanger sequencing were the same and consistent except for only a few samples whose genotypes were difficult to be determined by NGS due to the low coverage. With the high consistency in genotypes, further case-control association studies were conducted.

3.3 *Case-control association analysis*

In the case-control association analysis, 665 unrelated case and 777 unrelated control samples were used **(Table 2)**.

First, these samples were genotyped by TaqMan assay for the SNPs; rs1127354, rs2280090 and rs17857295 **(Table 3)**. Samples which failed to be genotyped were excluded from genotype counts. For each SNP, HWE-test for controls was performed in order to check for the absence of mis-genotyping and/or population stratification. HWE p-values were calculated and observed to be 0.643, 0.711, and 0.947, respectively, showing no deviation from expected genotype counts in controls. All the SNPs were thus included in further analysis.

For the case-control association analysis, cases were divided into three categories depending on age of each case in order to examine if there is any difference in SNP association among distinct age groups; < 25 years old, < 45 years old and ≥ 45 years old. No significant association in the case-control association analysis was observed for two SNPs, rs2280090 on *ADAM33* gene and rs17857295 on *MAVS* gene, in any age group or model **(Tables 8 and 9)**. However, rs1127354 on *ITPA* gene showed significant differences between young cases and controls in the allele and dominant models (P=0.015; OR=0.71; 95% CI=0.53-0.94, P=0.013; OR=0.67; 95% CI=0.49-0.92, respectively, for < 45 years old cases versus controls) while no significant difference was observed in old cases (≥ 45 years old) versus controls **(Table 10)**. The results of these SNPs are summarised in **Table 11**.

4. *Other candidate polymorphisms detected by NGS*

4.1 *Selection of candidate polymorphisms*

Besides non-synonymous polymorphisms, attempts were made to examine other

variants detected by NGS. Since it was difficult to conduct investigations into all the detected polymorphisms, the following criteria were used for selecting candidates for the case-control association analysis.

First, variants observed in only one sample and SNPs that were studied in the previous Thai GWAS and their proxy SNPs, which harbour LD $r^2 \geq 0.8$ with the GWAS SNPs in CHB+JPT were excluded. Then, variants which MAF in CHB were less than 5%, if reported, were excluded.

For polymorphisms detected by NGS and located in 3'UTR of genes, predicted miRNA binding sites were examined. With the database, microRNA.org - Targets and Expression (http://www.microrna.org/microrna/home.do) [51-53], no polymorphism was reported to lie in miRNA binding site. However, with the other database, miRDB (http://mirdb.org/miRDB/) [54, 55], rs1132922 on *MAVS* gene was predicted to lie in miRNA binding site and the SNP was extracted.

Polymorphisms that lie in DHS with motif change and near active regulatory sites as well as on genes were selected. Among them, proxy SNPs, which harbour LD $r^2 \geq 0.8$ with any of them were excluded. With these criteria, 3 SNPs, rs6115814, rs6116080 and rs6084506, were extracted, and all of them locate to promoter regions according to HaploReg v2 database (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) [56].

Finally, 4 SNPs remained in total according to the criteria above **(Table 4)**. Interestingly, one of them, rs6115814, was on *ITPA* gene, whose non-synonymous SNP showed a suggestive association as described. Thus, rs6115814 was determined to be a top candidate for the further study.

In addition, rs13830 in the 3'UTR of the *ITPA* gene, which was detected by NGS, was included to see if the SNP could have functional effect since its proxy SNP, rs1127354, was

reported to be strongly associated with anemia and thrombocytopenia induced by pegylated IFN and ribavirin therapy for Japanese patients with chronic hepatitis C [57, 58]. The SNP rs13830 was the only one detected in *ITPA*'s 3'UTR.

4.2 *Confirmation of genotypes of rs6115814 and rs13830 with Sanger sequencing*

In order to confirm the genotypes of rs6115814 and rs13830 by NGS, Sanger sequencing was performed as mentioned above. For rs6115814, the genotype of one sample was not replicated due to the lack of the available samples, however, for the other samples, genotyping results were consistent between NGS and Sanger sequencing.

For rs13830, one of the sample's genotype was not determined since the site has only 2x coverage in NGS, and another sample's genotype was not consistent between NGS and Sanger sequencing. Thus, the sample was re-sequenced by Sanger sequencing using two other primer sets but the results were unchanged. However, the genotypes for other samples were consistent in both sequencing methods and showed highly consistent results.

4.3 *Case-control association analysis*

In the case-control association analysis, the same sample sets were used for genotyping by TaqMan assay.

None of the SNP showed deviation from expected genotype counts in controls after conducting HWE-test with p-values 0.567 and 0.837, respectively.

In the case-control association analysis, no significant association was observed for rs6115814 in any age group or model **(Table 12)**. However, rs13830 showed significant differences between young cases and controls in allele and dominant models (P=4.4E-03;

OR=0.67; 95% CI=0.50-0.88, P=3.7E-03; OR=0.62; 95% CI=0.45-0.86, respectively, for < 45 years old cases versus controls). No significant difference was observed in old cases, ≥ 45 years old, versus controls **(Table 13)**.

5. *Association study of ITPA with young TB*

5.1 *Case-control association analysis*

According to the results of the SNPs on *ITPA* gene, the gene region was further examined. Considering the LD of the *ITPA* gene region, from 2kbp upstream to 1.5kbp downstream, 3 additional SNPs were selected as tagSNPs to be genotyped; rs11087570, rs8362 and rs6139034 **(Figure 6)**. The SNPs were genotyped by TaqMan assay and analysed.

None of the SNPs showed deviation from expected genotype counts in controls after conducting HWE-test with p-values 0.707, 0.801 and 0.429, respectively.

In the case-control association analysis, marginal p-value was observed for rs6139034 in recessive model and genotypic model for old cases versus controls (P=0.019; OR=1.37; 95% CI=1.05-1.78, P=0.034, respectively). In contrast, no significant result was observed in any model for young cases **(Table 14)**. Marginal p-value was observed for rs8362 in recessive model for young cases (< 45 years old) versus controls (P=0.034; OR=1.52; 95% CI=1.03-2.26) while no significant association was observed in old cases and controls **(Table 15)**. No significant association was observed for rs11087570 in any age group or model **(Table 16)**.

In short, among all the SNPs genotyped in *ITPA* gene region, the initial non-synonymous SNP, rs1127354, and rs13830 in the 3'UTR showed the lowest p-values and no other SNP surpassed their significance level.

5.2 *Case-control association analysis for the two top SNPs with the second sample set*

In order to improve statistical power to see if rs1127354 and rs13830 are possibly significant, genotyping with an additional sample set **(Table 2)** was performed and the results of the first sample set and the second sample set were combined and analysed.

For additional control samples, HWE-test was conducted and none of the SNPs showed deviation from expected genotype counts with p-values 0.758 and 0.868, respectively.

In the allelic and dominant models of case-control association analysis, both SNPs showed lowered p-values, especially in the young, which is younger than 45 years old, versus controls. In this age category, rs1127354 became more significant (P=1.3E-03; OR=0.72; 95% CI=0.59-0.88) in the allelic model. The SNP rs13830 also showed a lower p-value (P=5.1E-05; OR=0.66; 95% CI=0.54-0.81) in the same model, which is more striking than rs1127354. In the dominant model, both rs1127354 and rs13830 became more significant (P=1.1E-03; OR=0.68; 95% CI=0.54-0.86, P=4.5E-05; OR=0.62; 95% CI=0.49-0.78, respectively) **(Table 17)**.

5.3 *Linkage disequilibrium and haplotype analyses of the genotyped SNPs*

Based on the genotyping results, LD of the genotyped SNP pairs were analysed using Haploview 4.2 [61] **(Figure 8)**. LD are represented as r^2 values.

The initial candidate SNP, rs1127354, which lies in the exon region, showed strong LD (r^2=0.88) with the top SNP rs13830 in this studied Thai population. The LD structure for the Thai population did not differ much from the CHB LD structure of SNPs whose frequencies were available on HapMap database.

In addition, haplotype analysis was performed to evaluate whether the SNPs showed an interactive effect as haplotypes. None of the haplotypes reached its permutated significant

p-value, and were less significant than that of the top SNPs, rs13830 and rs1127354 **(Tables 18, 19)**.

6. *In silico eQTL analysis of rs13830*

6.1 *The eQTL-gene association analysis*

The eQTL-gene association analysis for rs13830 was conducted using Genevar database in the CHB population. Observed eQTL is displayed in the line chart in **Figure 9 (a)** and the p-values show association between eQTL and rs13830. Within a 2 Mbp around rs13830, eQTL on *ITPA* gene showed a significant association with the SNP ($P < 0.01$).

6.2 *The eQTL SNP-gene association analysis*

The eQTL SNP-gene association analysis for *ITPA* gene and rs13830 was performed and the functional significance of rs13830 was assessed using Genevar database in the CHB population.

The mRNA expression profile of lymphoblastoid cell lines in each genotype of rs13830 is shown in **Figure 9 (b)**. A significant correlation between mRNA level of *ITPA* and the genotypes of rs13830 was observed (permutated P-value=0.0045), displaying higher expression in accordance with the number of minor A allele **(Figure 9 (b))**.

Discussion

In the previous Thai GWLS, it was revealed that a region on Chr. 20p13-12.3 showed significant linkage with early onset TB [39]. In the current study, identification of new susceptibility gene(s) and/or SNPs in the candidate region in a young Thai population was attempted. The previous Asian GWAS did not report any significant association with TB in the genomic region, even in a young population [41]. However, there have been limitations that the previous GWAS could have missed some possible candidates due to insufficient data about variants and their frequencies, as well as the lack of Thai population information in available public databases. To overcome these limitations, in this study, NGS was carried out for the candidate region. NGS is a recent novel technology which enables rapid and massive sequencing with comparatively lower cost. To my knowledge, there have not been any published reports using NGS to search for human genetic factors for TB susceptibility.

The current study screened for variants within the candidate region using 13 multiplex Thai families. Since there were multiple cases observed in the families, the genetic heritability may be enriched. To maintain accuracy of the variants detected by NGS, stringent criteria were used and at least 20x coverage for the detected variants was required in two software, Ion Variant caller plugin and CLC Genomics Workbench v5.5. In order to ensure the accuracy of the detected variants that were further studied, the samples underwent Sanger sequencing. Overall, the results of the NGS calls and Sanger sequencing genotyping were mostly consistent. In addition to the stringent selection criteria, candidate SNPs were further selected by looking at non-synonymous SNPs as well as capturing potentially functionally interesting ones by looking at two conditions. The first condition was whether a variant was in a miRNA binding site within a 3' UTR, which could result in gene silencing

via translational repression or target degradation [71, 72]. The second condition was whether a variant is located within DHS and near active regulatory elements. Chromatin loses its condensed structure in DHS and variants in the regions are related with transcriptional activity.

The candidate SNPs for the case-control association studies were then genotyped using unrelated samples from the previous Thai GWAS [41]. In the GWAS, PCA was conducted on the samples to see if there was population stratification in the samples used. Three clusters were observed and one cluster was chosen for further analysis to reduce genetic heterogeneity due to population bias **(Figure 3)**. For the current case-control association analysis, the first sample set used for genotyping all the SNPs were derived from the same population cluster and the possibility of population stratification was low. Approximately 5-10% of infected persons will progress to develop the clinical disease in their lifetime [4] and we estimated that to be the maximum percentage of classification bias for TB diagnosis in our controls. We tried to exclude all samples who were at risk of TB in order to minimize the possibility. Thus, the percentage of misclassification should be much lower. In addition, large numbers of control samples were included in the current study to increase statistical power and overcome the misclassification possibility.

Among the non-synonymous SNPs detected by NGS, rs1127354, which is on the *ITPA* gene, showed significant association in young cases **(Tables 10, 11)**. In addition, four SNPs were identified under the functionally interesting conditions **(Table 4)**. Even with several stringent selection criteria and these conditions, another SNP on the *ITPA* gene remained: rs6115814. Furthermore, rs13830 is the only SNP in the 3'UTR of the gene that was detected by NGS. Previous studies showed that a proxy SNP of rs13830, rs1127354, is strongly associated with anemia and thrombocytopenia induced by pegylated IFN and ribavirin therapy for Japanese patients with chronic hepatitis C [57, 58]. Thus the two SNPs

rs6115814 and rs13830 were also genotyped and analysed. Since rs13830 also showed a significant association in addition to rs1127354 **(Table 13)**, I focused on the *ITPA* gene and analysed additional tagSNPs to cover the gene region in order to examine if any of these would show stronger association than the two SNPs.

Among all the studied SNPs, only rs13830 and rs1127354 showed low p-values in young cases and the p-values became lower after increasing statistical power by genotyping additional samples **(Table 17)**. The SNP rs13830 showed a lower p-value (P=5.1E-05; OR=0.66) than rs1127354 (P=1.3E-03; OR=0.72), suggesting a stronger association with young TB and the minor alleles of both SNPs are protective.

It was observed from the case-control association analysis that the SNP rs13830 in the 3'UTR of *ITPA* showed stronger association with TB than the other SNPs tested. The eQTL-gene association analysis of rs13830 showed a significant association with the expression level of *ITPA* ($P < 0.01$) **(Figure 9 (a))**.

The *ITPA* gene encodes the enzyme inosine triphosphate pyrophosphatase (ITPase), which functions to catalyze the hydrolysis of inosine triphosphate (ITP) to inosine monophosphate (IMP) and pyrophosphate [73, 74]. Although the role of ITPase in humans is not well-defined, it is presumed to have some role in maintaining genomic stability by preventing DNA damage and mutagenesis in human cells [75, 76]. There is another study that reported a relationship between low *ITPA* activity and adverse effects by azathioprine, which is an immunosuppressive drug [77]. It has also been speculated that there might be a potential role for *ITPA* in immunity [75].

As mentioned above, rs13830 is located in the 3'UTR of *ITPA* gene. It is known that polymorphisms in 3'UTR of a gene may have a relationship with the regulation of the mRNA transcript and gene expression [78-80]. The *in silico* eQTL analysis assessed that the

expression intensity of *ITPA* differed significantly according to the allele genotype of rs13830, with higher expression levels seen in the minor A allele genotypes (**Figure 9 (b)**). Therefore, it can be speculated that the SNP in this regulatory region be related to mRNA stability [81].

Lymphoblastoid cell lines were used to measure expression levels in the eQTL analysis and these cell lines are constructed using human peripheral B-lymphocytes [82]. Although the traditional understanding of TB disease progression has largely concentrated on the functions and roles of cellular immunity, there has been growing interest recently in the role of B-cells in the immune response to infection by *M.tb* [83]. The expression of *ITPA* may also be affected in T-cells as well based on the genotypes of the alleles of rs13830 and might play a role in the disease progression of TB. Furthermore, a search in the UCSC genome browser (http://genome.ucsc.edu/) revealed that expression of *ITPA* is high in immune cells [46].

Thus, it may be speculated that the minor allele of rs13830 in the 3' UTR of the *ITPA* gene may serve to increase the stability of the mRNA transcript and increase expression of the gene, which in turn may lead to better functioning of the cells involved in the immune response against *M.tb*.

The findings in the current study were observed only in young cases. When all ages were included in the analysis, the association was not as prominent as with only young cases. The current study thus supports the findings of previous studies showing that stratification by age at onset of TB can be effective in elucidating genetic factors [41, 42]. The current study also enables the assumption that different genetic factors contribute to pathogenesis of TB in young and old populations.

To date, NGS has been used for the genetic study of the pathogen, *M. tb* [84]. To the

best of my knowledge, the current study is the first attempt at using NGS to gain insight into host genetic factors associated with TB. The SNP with the strongest association with TB, rs13830, which was not seen in the previous Thai GWAS, was successfully identified by NGS in the current study. In this study, a 1Mbp candidate region was focused on that was previously identified by GWLS [39], however, future studies can include the flanking regions or possibly the entire genome as well. Susceptibility to TB was estimated to have high heritability according to several previous studies, however, the genetic background of TB susceptibility has not been well understood until now. The use of NGS as proposed in the current study may contribute to the elucidation of more genetic factors that play roles in TB susceptibility and progression. Recent studies using NGS on the human exome have successfully uncovered genes that cause disease states in quite a few Mendelian disorders [85, 86]. This study demonstrated that the current method is effective in detecting new potential loci in the study of common diseases. It should be noted that despite the advantages of NGS in identifying new potential loci, there is a need to be aware of the possibility of errors caused by inaccuracies inherent to NGS [87, 88]. Therefore, stringent data filtering and confirmation by Sanger sequencing are required.

In addition, this study successfully detected and provided insights to a new genetic factor for TB susceptibility in the Thai population, whose genome-wide SNP information is lacking from public databases. I believe the findings contribute to the better understanding of TB pathogenesis and can hopefully be useful for studies that attempt to uncover effective drug targets in Thailand, where TB is a large burden on healthcare.

Certain limitations exist in the current study. Even though SNPs in the *ITPA* gene, especially rs13830, were identified, the p-values did not reach genome-wide significance

level. This could likely be due to the limited sample size for young TB patients. It is recommended that the association will be confirmed by increasing sample numbers and/or conducting replication studies in other Asian populations which are genetically close to the Thai population.

The possibility that rs13830 is not the causative SNP cannot be denied either. Further investigation of proxy variants of rs13830 is warranted. Functional studies should also be done to confirm the genetic factor(s) identified.

Another limitation is possible type II error. In the current study, variants with less than 20x coverage were filtered out from the NGS data due to the stringent filtering criteria. However, there may be variants that did not pass the 20x coverage criteria but still play a role in TB susceptibility and progression. Functionally interesting variants such as non-synonymous SNPs and variants in DHS may be selected from this group, and confirmation of the genotypes using Sanger sequencing followed by association studies can be conducted.

In conclusion, this study is the first report of a potential association of *ITPA* gene with young age-at-onset of TB. The study also demonstrates the effectiveness of NGS in searching for susceptibility variants to common diseases.

References

1. World Health Organization, Global tuberculosis report 2012, http://apps.who.int/iris/bitstream/10665/75938/1/9789241564502_eng.pdf
acknackno

2. World Health Organization
http://www.who.int/mediacentre/factsheets/who104/en/print.html

3. Sharma SK, Mohan A: **Multidrug-resistant tuberculosis: a menace that threatens to destabilize tuberculosis control**. *Chest* 2006, **130**(1):261-272.

4. Vynnycky E, Fine PE: **Lifetime risks, incubation period, and serial interval of tuberculosis**. *American journal of epidemiology* 2000, **152**(3):247-263.

5. Comstock GW: **Tuberculosis in twins: a re-analysis of the Prophit survey**. *The American review of respiratory disease* 1978, **117**(4):621-624.

6. Jepson A, Fowler A, Banya W, Singh M, Bennett S, Whittle H, Hill AV: **Genetic regulation of acquired immune responses to antigens of Mycobacterium tuberculosis: a study of twins in West Africa**. *Infection and immunity* 2001, **69**(6):3989-3994.

7. Kallmann F, Reisner D: **Twin Studies on the significance of genetic factors in tuberculosis**. *American review of tuberculosis* 1942, **47**:549-574.

8. Newport MJ, Goetghebuer T, Weiss HA, Whittle H, Siegrist CA, Marchant A: **Genetic regulation of immune responses to vaccines in early life**. *Genes and immunity* 2004, **5**(2):122-129.

9. Kimman T, Janssen R, Hoebee B: **Future prospects in respiratory syncytical virus genetics**. *Future Virology* 2006, **1**(4):483-492.

10. Moller M, de Wit E, Hoal EG: **Past, present and future directions in human genetic susceptibility to tuberculosis**. *FEMS immunology and medical microbiology* 2010, **58**(1):3-26.

11. Greenwood CM, Fujiwara TM, Boothroyd LJ, Miller MA, Frappier D, Fanning EA, Schurr E, Morgan K: **Linkage of tuberculosis to chromosome 2q35 loci, including NRAMP1, in a large aboriginal Canadian family**. *American journal of human genetics* 2000, **67**(2):405-416.

12. Baghdadi JE, Orlova M, Alter A, Ranque B, Chentoufi M, Lazrak F, Archane MI, Casanova JL, Benslimane A, Schurr E *et al*: **An autosomal dominant major gene confers predisposition to pulmonary tuberculosis in adults**. *The Journal of experimental medicine* 2006, **203**(7):1679-1684.

13. Miller EN, Jamieson SE, Joberty C, Fakiola M, Hudson D, Peacock CS, Cordell HJ, Shaw MA, Lins-Lainson Z, Shaw JJ *et al*: **Genome-wide scans for leprosy and tuberculosis susceptibility genes in Brazilians**. *Genes and immunity* 2004, **5**(1):63-67.

14. Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, Bester D, Meyer M, Corrah T, Collin M *et al*: **Genetic susceptibility to tuberculosis in Africans: a genome-wide scan**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(14):8005-8009.

15. Vasilca V, Oana R, Munteanu D, Zugun F, Constantinescu D, Carasevici E: **HLA-A and -B phenotypes associated with tuberculosis in population from north-eastern Romania**. *Roumanian archives of microbiology and immunology* 2004, **63**(3-4):209-221.

16. Yuliwulandari R, Sachrowardi Q, Nakajima H, Kashiwase K, Hirayasu K, Mabuchi A, Sofro AS, Tokunaga K: **Association of HLA-A, -B, and -DRB1 with pulmonary tuberculosis in western Javanese Indonesia**. *Human immunology* 2010, **71**(7):697-701.

17. Saad AH, Ghor MA, Osman A, Sherbiny MA, Zakaria Z: **Association of immune responses to mycobacterium tuberculosis peptide antigens with host genetic factors**. *Journal of the Egyptian Society of Parasitology* 2006, **36**(3):889-910.

18. Lombard Z, Dalton DL, Venter PA, Williams RC, Bornman L: **Association of HLA-DR, -DQ, and vitamin D receptor alleles and haplotypes with tuberculosis in the Venda of South Africa**. *Human immunology* 2006, **67**(8):643-654.

19. Kim HS, Park MH, Song EY, Park H, Kwon SY, Han SK, Shim YS: **Association of HLA-DR and HLA-DQ genes with susceptibility to pulmonary tuberculosis in Koreans: preliminary evidence of associations with drug resistance, disease severity, and disease recurrence**. *Human immunology* 2005, **66**(10):1074-1081.

20. El Baghdadi J, Remus N, Benslimane A, El Annaz H, Chentoufi M, Abel L, Schurr E: **Variants of the human NRAMP1 gene and susceptibility to tuberculosis in Morocco**. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2003, **7**(6):599-602.

21. Fitness J, Floyd S, Warndorff DK, Sichali L, Malema S, Crampin AC, Fine PE, Hill AV: **Large-scale candidate gene study of tuberculosis susceptibility in the Karonga district of northern Malawi**. *The American journal of tropical medicine and hygiene* 2004, **71**(3):341-349.

22. Hoal EG, Lewis LA, Jamieson SE, Tanzer F, Rossouw M, Victor T, Hillerman R, Beyers N, Blackwell JM, Van Helden PD: **SLC11A1 (NRAMP1) but not SLC11A2 (NRAMP2) polymorphisms are associated with susceptibility to tuberculosis in a high-incidence community in South Africa**. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2004, **8**(12):1464-1471.

23. Li HT, Zhang TT, Zhou YQ, Huang QH, Huang J: **SLC11A1 (formerly NRAMP1) gene polymorphisms and tuberculosis susceptibility: a meta-analysis**. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2006, **10**(1):3-12.

24. Leung KH, Yip SP, Wong WS, Yiu LS, Chan KK, Lai WM, Chow EY, Lin CK, Yam WC, Chan KS: **Sex- and age-dependent association of SLC11A1 polymorphisms with tuberculosis in Chinese: a case control study**. *BMC infectious diseases* 2007, **7**:19.

25. Barreiro LB, Neyrolles O, Babb CL, Tailleux L, Quach H, McElreavey K, Helden PD, Hoal EG, Gicquel B, Quintana-Murci L: **Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis**. *PLoS medicine* 2006, **3**(2):e20.

26. Olesen R, Wejse C, Velez DR, Bisseye C, Sodemann M, Aaby P, Rabna P, Worwui A, Chapman H, Diatta M *et al*: **DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans**. *Genes and immunity* 2007, **8**(6):456-467.

27. Vannberg FO, Chapman SJ, Khor CC, Tosh K, Floyd S, Jackson-Sillah D, Crampin A, Sichali L, Bah B, Gustafson P *et al*: **CD209 genetic polymorphism and tuberculosis disease**. *PloS one* 2008, **3**(1):e1388.

28. Kobayashi K, Yuliwulandari R, Yanai H, Lien LT, Hang NT, Hijikata M, Keicho N, Tokunaga K: **Association of CD209 polymorphisms with tuberculosis in an Indonesian population**. *Human immunology* 2011, **72**(9):741-745.

29. Yim JJ, Lee HW, Lee HS, Kim YW, Han SK, Shim YS, Holland SM: **The association between microsatellite polymorphisms in intron II of the human Toll-like receptor 2 gene and tuberculosis among Koreans**. *Genes and immunity* 2006, **7**(2):150-155.

30. Ma X, Liu Y, Gowen BB, Graviss EA, Clark AG, Musser JM: **Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease**. *PloS one* 2007, **2**(12):e1318.

31. Davila S, Hibberd ML, Hari Dass R, Wong HE, Sahiratmadja E, Bonnard C, Alisjahbana B, Szeszko JS, Balabanova Y, Drobniewski F *et al*: **Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis**. *PLoS genetics* 2008, **4**(10):e1000218.

32. Velez DR, Wejse C, Stryjewski ME, Abbate E, Hulme WF, Myers JL, Estevan R, Patillo SG, Olesen R, Tacconelli A *et al*: **Variants in toll-like receptors 2 and 9 influence susceptibility to pulmonary tuberculosis in Caucasians, African-Americans, and West Africans**. *Human genetics* 2010, **127**(1):65-73.

33. Kobayashi K, Yuliwulandari R, Yanai H, Naka I, Lien LT, Hang NT, Hijikata M, Keicho N, Tokunaga K: **Association of TLR polymorphisms with development of tuberculosis in Indonesian females**. *Tissue antigens* 2012, **79**(3):190-197.

34. Zhu M, Zhao S: **Candidate gene identification approach: progress and challenges**. *International journal of biological sciences* 2007, **3**(7):420-427.

35. Stranger BE, Stahl EA, Raj T: **Progress and promise of genome-wide association studies for human complex trait genetics**. *Genetics* 2011, **187**(2):367-383.

36. Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, Gyapong J, Sirugo G, Sisay-Joof F, Enimil A, Chinbuah MA *et al*: **Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2**. *Nature genetics* 2010, **42**(9):739-741.

37. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, Nikolayevskyy V, Drobniewski F, Nejentsev S, Adnan I *et al*: **A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians**. *BMC medical genetics* 2012, **13**:5.

38. Stewart GR, Robertson BD, Young DB: **Tuberculosis: a problem with persistence**. *Nature reviews Microbiology* 2003, **1**(2):97-105.

39. Mahasirimongkol S, Yanai H, Nishida N, Ridruechai C, Matsushita I, Ohashi J, Summanapan S, Yamada N, Moolphate S, Chuchotaworn C *et al*: **Genome-wide SNP-based linkage analysis of tuberculosis in Thais**. *Genes and immunity* 2009, **10**(1):77-83.

40. Alcais A, Fieschi C, Abel L, Casanova JL: **Tuberculosis in children and adults: two distinct genetic diseases**. *The Journal of experimental medicine* 2005, **202**(12):1617-1621.

41. Mahasirimongkol S, Yanai H, Mushiroda T, Promphittayarat W, Wattanapokayakit S, Phromjai J, Yuliwulandari R, Wichukchinda N, Yowang A, Yamada N *et al*: **Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis**. *Journal of human genetics* 2012, **57**(6):363-367.

42. Grant AV, El Baghdadi J, Sabri A, El Azbaoui S, Alaoui-Tahiri K, Abderrahmani Rhorfi I, Gharbaoui Y, Abid A, Benkirane M, Raharimanga V *et al*: **Age-dependent association between pulmonary tuberculosis and common TOX variants in the 8q12-13 linkage region**. *American journal of human genetics* 2013, **92**(3):407-414.

43. Ridruechai C, Mahasirimongkol S, Phromjai J, Yanai H, Nishida N, Matsushita I, Ohashi J, Yamada N, Moolphate S, Summanapan S *et al*: **Association analysis of susceptibility candidate region on chromosome 5q31 for tuberculosis**. *Genes and immunity* 2010, **11**(5):416-422.

44. Vollstedt S, Yuliwulandari R, Okamoto K, Lien LT, Keicho N, Rochani JT, Wikaningrum R, Tokunaga K: **No evidence for association between the interferon regulatory factor 1 (IRF1) gene and clinical tuberculosis**. *Tuberculosis (Edinburgh, Scotland)* 2009, **89**(1):71-76.

45. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis**. *Bioinformatics (Oxford, England)* 2007, **23**(10):1294-1296.

46. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome research* 2002, **12**(6):996-1006.

47. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M *et al*: **An integrated semiconductor device enabling non-optical genome sequencing**. *Nature* 2011, **475**(7356):348-352.

48. Ghadessy FJ, Ong JL, Holliger P: **Directed evolution of polymerase function by compartmentalized self-replication**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(8):4552-4557.

49. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI: **SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap**. *Bioinformatics (Oxford, England)* 2008, **24**(24):2938-2939.

50. **The International HapMap Project**. *Nature* 2003, **426**(6968):789-796.

51. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets**. *PLoS biology* 2004, **2**(11):e363.

52. Betel D, Wilson M, Gabow A, Marks DS, Sander C: **The microRNA.org resource: targets and expression**. *Nucleic acids research* 2008, **36**(Database issue):D149-153.

53. Betel D, Koppal A, Agius P, Sander C, Leslie C: **Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites**. *Genome biology* 2010, **11**(8):R90.

54. Wang X: **miRDB: a microRNA target prediction and functional annotation database with a wiki interface**. *RNA (New York, NY)* 2008, **14**(6):1012-1017.

55. Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals**. *Bioinformatics (Oxford, England)* 2008, **24**(3):325-332.

56. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants**. *Nucleic acids research* 2012, **40**(Database issue):D930-934.

57. Ochi H, Maekawa T, Abe H, Hayashida Y, Nakano R, Kubo M, Tsunoda T, Hayes CN, Kumada H, Nakamura Y *et al*: **ITPA polymorphism affects ribavirin-induced anemia and outcomes of therapy--a genome-wide study of Japanese HCV virus patients**. *Gastroenterology* 2010, **139**(4):1190-1197.

58. Tanaka Y, Kurosaki M, Nishida N, Sugiyama M, Matsuura K, Sakamoto N, Enomoto N, Yatsuhashi H, Nishiguchi S, Hino K *et al*: **Genome-wide association study identified ITPA/DDRGK1 variants reflecting thrombocytopenia in pegylated interferon and ribavirin therapy for chronic hepatitis C**. *Human molecular genetics* 2011, **20**(17):3507-3516.

59. Koressaar T, Remm M: **Enhancements and modifications of primer design program Primer3**. *Bioinformatics (Oxford, England)* 2007, **23**(10):1289-1291.

60. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3--new capabilities and interfaces**. *Nucleic acids research* 2012, **40**(15):e115.

61. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps**. *Bioinformatics (Oxford, England)* 2005, **21**(2):263-265.

62. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M *et al*: **The structure of haplotype blocks in the human genome**. *Science (New York, NY)* 2002, **296**(5576):2225-2229.

63. Wigginton JE, Cutler DJ, Abecasis GR: **A note on exact tests of Hardy-Weinberg equilibrium**. *American journal of human genetics* 2005, **76**(5):887-893.

64. Gilliland G, Perrin S, Blanchard K, Bunn HF: **Analysis of cytokine mRNA and DNA: detection and quantitation by competitive polymerase chain reaction**. *Proceedings of the National Academy of Sciences of the United States of America* 1990, **87**(7):2725-2729.

65. Leutenegger CM, Klein D, Hofmann-Lehmann R, Mislin C, Hummel U, Boni J, Boretti F, Guenzburg WH, Lutz H: **Rapid feline immunodeficiency virus provirus quantitation by polymerase chain reaction using the TaqMan fluorogenic real-time detection system**. *Journal of virological methods* 1999, **78**(1-2):105-116.

66. Olivier M: **The Invader assay for SNP genotyping**. *Mutation research* 2005, **573**(1-2):103-110.

67. World Health Organization, Regional Office for South-East Asia: **Tuberculosis in the South-East Asia Region, The Regional Report: 2008** 2008: 60.

68. Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, Deloukas P, Dermitzakis ET: **Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies**. *Bioinformatics (Oxford, England)* 2010, **26**(19):2474-2476.

69. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, Sekowska M, Smith GD, Evans D, Gutierrez-Arcelus M *et al*: **Patterns of cis regulatory variation in diverse human populations**. *PLoS genetics* 2012, **8**(4):e1002639.

70. Nielsen R, Paul JS, Albrechtsen A, Song YS: **Genotype and SNP calling from next-generation sequencing data**. *Nature reviews Genetics* 2011, **12**(6):443-451.

71. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

72. Kusenda B, Mraz M, Mayer J, Pospisilova S: **MicroRNA biogenesis, functionality and cancer relevance**. *Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia* 2006, **150**(2):205-215.

73. Holmes SL, Turner BM, Hirschhorn K: **Human inosine triphosphatase: catalytic properties and population studies**. *Clinica chimica acta; international journal of clinical chemistry* 1979, **97**(2-3):143-153.

74. Lin S, McLennan AG, Ying K, Wang Z, Gu S, Jin H, Wu C, Liu W, Yuan Y, Tang R *et al*: **Cloning, expression, and characterization of a human inosine triphosphate pyrophosphatase encoded by the itpa gene**. *The Journal of biological chemistry* 2001, **276**(22):18695-18701.

75. Bakker JA, Bierau J, Drent M: **A role for ITPA variants in the clinical course of pulmonary Langerhans' cell histiocytosis?** *The European respiratory journal* 2010, **36**(3):684-686.

76. Menezes MR, Waisertreiger IS, Lopez-Bertoni H, Luo X, Pavlov YI: **Pivotal role of inosine triphosphate pyrophosphatase in maintaining genome stability and the prevention of apoptosis in human cells**. *PloS one* 2012, **7**(2):e32313.

77. Shipkova M, Franz J, Abe M, Klett C, Wieland E, Andus T: **Association between adverse effects under azathioprine therapy and inosine triphosphate pyrophosphatase activity in patients with chronic inflammatory bowel disease**. *Therapeutic drug monitoring* 2011, **33**(3):321-328.

78. Chen JM, Ferec C, Cooper DN: **A systematic analysis of disease-associated variants in the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3' UTR variants**. *Human genetics* 2006, **120**(3):301-333.

79. Misquitta CM, Iyer VR, Werstiuk ES, Grover AK: **The role of 3'-untranslated region (3'-UTR) mediated mRNA stability in cardiovascular pathophysiology**. *Molecular and cellular biochemistry* 2001, **224**(1-2):53-67.

80. Di Paola R, Frittitta L, Miscio G, Bozzali M, Baratta R, Centra M, Spampinato D, Santagati MG, Ercolino T, Cisternino C *et al*: **A variation in 3' UTR of hPTP1B increases specific gene expression and associates with insulin resistance**. *American journal of human genetics* 2002, **70**(3):806-812.

81. Prokunina L, Alarcon-Riquelme ME: **Regulatory SNPs in complex diseases: their identification and functional validation**. *Expert reviews in molecular medicine* 2004, **6**(10):1-15.

82. Ring CJ: **The B cell-immortalizing functions of Epstein-Barr virus**. *The Journal of general virology* 1994, **75 (Pt 1)**:1-13.

83. Maglione PJ, Chan J: **How B cells shape the immune response against Mycobacterium tuberculosis**. *European journal of immunology* 2009, **39**(3):676-686.

84. Loman NJ, Pallen MJ: **XDR-TB genome sequencing: a glimpse of the microbiology of the future**. *Future microbiology* 2008, **3**(2):111-113.

85. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ *et al*: **Genome-wide in situ exon capture for selective resequencing**. *Nature genetics* 2007, **39**(12):1522-1527.

86. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I: **Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders**. *Journal of human genetics* 2012, **57**(10):621-632.

87. Natrajan R, Reis-Filho JS: **Next-generation sequencing applied to molecular diagnostics**. *Expert review of molecular diagnostics* 2011, **11**(4):425-444.

88. Metzker ML: **Sequencing technologies - the next generation**. *Nature reviews Genetics* 2010, **11**(1):31-46.

Acknowledgement

I would firstly like to express my sincere gratitude to Professor Katsushi Tokunaga for his warm supervision and guidance throughout my Master and Ph.D periods. His continuous support enabled me to enjoy this very challenging study. I am also grateful to Dr Hideki Yanai for managing our research collaboration and backup support.

I am very thankful to Dr Surakameth Mahasirimongkol for sharing precious samples and data, as well as providing stimulating discussion. I also thank Ms Sukanya Wattanapokayakit for the experimental support and Dr Taisei Mushiroda for the useful discussions.

Furthermore, I would like to express my cordial gratitude to all the staff members in Department of Human Genetics, Graduate School of Medicine, University of Tokyo, especially Associated Professor Akihiko Mabuchi and Dr Yuki Hitomi, for their helpful advice and instructions. I am also thankful to all my colleagues, who lightened up my days and I am proud of being a member of our department. I would especially like to give special thanks to Mr Wong Jing Hao, who had worked for TB projects together for two years as well as for the active discussions and countless support.

Last but not least, I would like to owe my deepest gratitude to my parents for all the constant and continuous support in all aspects. Without their support, it would have been impossible to pursue my study.