Common variants at 22q12.2 are associated with susceptibility to Tuberculosis

Xuling Chang[1,2,3*], Zheng Li[4], Phan Vuong Khac Thai[5], Dang Thi Minh Ha[5], Nguyen Thuy Thuong Thuong[6,7], Matthew Silcocks[1], Cynthia Bin Eng Chee[8], Nguyen Thi Quynh Nhu[6], Chew-Kiat Heng[2,3], Yik Ying Teo[9,10], Jian-Min Yuan[11,12], Woon-Puay Koh[13,14], Maxine Caws[15,16], Chiea Chuen Khor[4,17], Rajkumar Dorajoo[2,4]and Sarah J Dunstan[1*]

[1]Department of Infectious Diseases, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Parkville, Victoria, Australia.

[2]Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore.

[3]Khoo Teck Puat – National University Children's Medical Institute, National University Health System, Singapore 119074, Singapore

[4]Genome Institute of Singapore, Agency for Science, Technology and Research, 138672, Singapore

[5]Pham Ngoc Thach Hospital, Ho Chi Minh City, District 5, Viet Nam

[6]Oxford University Clinical Research Unit, Ho Chi Minh City, District 5, Viet Nam

[7]Centre for Tropical Medicine, Nuffield Department of Clinical Medicine, Oxford University, Oxford, UK.

[8]Tuberculosis Control Unit, Tan Tock Seng Hospital, Singapore, Singapore

[9]Department of Statistics and Applied Probability, National University of Singapore, Singapore.

[10]Saw Swee Hock School of Public Health, National University of Singapore, Singapore

[11]Cancer Epidemiology and Prevention Program, UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232, USA.

[12]Department of Epidemiology, School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA.

[13]Healthy Longevity Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117545, Singapore

[14]Singapore Institute for Clinical Sciences, Agency for Science Technology and Research (A*STAR), Singapore 117609, Singapore

[15]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom

[16]Birat Nepal Medical Trust, 257 Lazimpat, Kathmandu, Nepal

[17]Singapore Eye Research Institute, Singapore

* Corresponding author: Correspondence should be addressed to Assoc Prof Sarah Dunstan, University of Melbourne (sarah.dunstan@unimelb.edu.au) or Dr Xuling Chang, University of Melbourne (shirley.chang@unimelb.edu.au)

Abstract

Tuberculosis (TB) continues to be a leading cause of morbidity and mortality worldwide. Past genome-wide association studies (GWAS) have explored TB susceptibility across various ethnic groups, yet a significant portion of TB heritability remains unexplained. In this study, we conducted GWAS in the Singapore Chinese and Vietnamese, followed by a comprehensive meta-analysis incorporating independent East Asian data, and identified a novel pulmonary TB (PTB) susceptibility locus at 22q12.2 [rs6006426, OR(95%Cl)=1.097(1.066, 1.130), $P_{meta}=3.31\times10^{-10}$]. Our lead SNP was found to affect the expression of *SF3A1* in various immune-related cells (*P* ranging from 1.48×10^{-9} to 6.17×10^{-18}). Furthermore, a significant association was observed between rs6006426 and cigarette smoking ($P<0.044$). When exploring the interplay between genetic marker, smoking and TB, our findings indicated that smoking status significantly mediated the effect of rs6006426 on PTB ($\beta_{indirect\text{-}effect}=-0.004$, $P_{indirect\text{-}effect}=0.020$). Our findings offer novel insights into the genetic factors underlying TB and reveals new avenues for understanding its etiology.

Introduction

Tuberculosis (TB), an infectious disease caused by bacteria *Mycobacterium tuberculosis* (*Mtb*), remains a leading cause of morbidity and mortality worldwide. It poses a significant challenge to global health [1] and a major contributor to the rising global burden of antimicrobial resistance [2]. Approximately one quarter of the global population has been infected with *Mtb*, yet only about 5-15% of these individuals progress to active disease in their lifetime [3-5]. The incidence of TB is not uniform across the globe. It predominantly affects individuals in low- and middle-income countries, with the highest burden observed in Southeast Asia, the Western Pacific, and African regions [1]. Individuals with compromised immune systems, including those living with HIV, suffering from malnutrition or diabetes, or tobacco users, are at a heightened risk of developing active TB [1, 6]. The most predominant form of the disease is pulmonary TB (PTB), however in some cases, *Mtb* can spread from the lungs to other organs of the body. The World Health Organisation (WHO) 2023 global tuberculosis report indicates a record high of 7.5 million new TB cases in 2022, exceeding the pre-COVID baseline of 7.1 million in 2019. In 2022, TB accounted for approximately 1.30 million deaths globally, with the COVID-19 pandemic estimated to have contributed to nearly half a million excess TB deaths from 2020 to 2022 compared to pre-pandemic projection [1]. Despite recent advances in treatment and diagnosis, the challenge to optimally disrupt transmission and ensure positive outcomes for all of those infected remains. This enduring challenge is primarily attributed to several factors, including but not limited to the ineffectiveness of the current vaccine in fully preventing disease, the lengthy and complex multi-drug treatment regimens required, and the rising incidence of multidrug-resistant (MDR) and extensively drug-resistant (XDR) TB infections, which are considerably more difficult to treat [7]. Hence a deeper understanding towards the etiology of TB is crucial for

identifying people at high risk and developing targeted treatments and effective vaccines, to control this significant global health threat.

TB susceptibility is influenced by a complex interplay of genetic and environmental factors [8], such as socio-economic conditions [9], smoking [10] and diabetes [11]. Twin studies and mouse models demonstrate a strong host genetic influence on TB susceptibility [8, 12, 13] with heritability reported to be more than 50% [14]. Research investigating the relationship between host genetics and TB via genome-wide association studies (GWAS) have been conducted across multiple ethnic groups, including those from Africa [15, 16], Russia [17], Iceland [18] and Asia [19-21]. However, these studies have not fully accounted for the genetic risk and have shown considerable location or ethnicity specific genetic associations, with minimal replication across populations [17, 18, 22].

In the present study, we have performed GWAS and meta-analysis to identify genetic variants shared among East-Asian population groups associated with PTB. We employed PTB case control datasets from both Singapore and Vietnam, and to increase statistical power and robustness of our findings, we incorporated public data from additional East Asian populations, including Chinese Han [20] and Japanese [21], into a meta-analysis.

Results

Genome-Wide Association Study in the Singaporean Chinese and Vietnamese

To evaluate genetic susceptibility to PTB we performed a GWAS among Chinese in Singapore [data from Singapore Chinese Health Study (SCHS)] [23], which consisted of 1,610 PTB cases and 24,015 population controls. Concurrently, a second PTB GWAS was

performed on 1,598 individuals with microbiologically confirmed PTB and 1,267 populations controls from Ho Chi Minh City, Vietnam.

Analyses were performed separately in Singaporean Chinese and Vietnamese populations. After PCA-adjustment, no significant inflation was observed; inflation factors λ of 1.007 for PTB GWAS in the Singaporean Chinese (Supplemental Figure 1) and 0.985 for PTB GWAS in the Vietnamese (Supplemental Figure 2), suggesting that effects from potential population stratification were well-controlled. No SNP associations surpassing genome wide significance ($P<5\times10^{-8}$) were observed in these cohorts alone (Supplemental Figure 1-2).

A meta-analysis was performed incorporating all available PTB GWAS datasets from East Asia. Published summary statistics from a Han Chinese dataset (833 PTB cases and 1,220 controls) and a Japanese dataset (data from Biobank Japan; 7,800 PTB and 170,871 controls) [20, 21] were analysed along with the Singaporean Chinese and Vietnamese datasets (Supplemental Table 1). This integration resulted in a combined dataset comprising 11,841 PTB cases and 197,373 controls. The meta-analysis of the 4 East-Asian datasets indicated minimal inflation (λ = 1.028) and two loci were identified that surpassed genome-wide significance ($P<5\times10^{-8}$) (Figure 1). This includes the previously reported *HLA-DQB1* locus in the Japanese [21]. The lead SNP rs140780894 in the Japanese was not present in the Singaporean Chinese and it showed a significant association with PTB in the Han Chinese population (P = 0.034) but not in the Vietnamese (Supplemental Table 2). The sentinel SNP in this meta-analysis was rs9274669 [OR (95%Cl) = 1.160 (1.121, 1.200), $P_{meta} = 1.92\times10^{-17}$] (Supplemental Table 3), which is in weak linkage disequilibrium (LD) with rs140780894 (r^2=0.114, JPT in 1000 Genome Project). The association of rs9274669 was primarily driven by data from Japan ($P = 7.84\times10^{-16}$) and was also significant in the Singapore Chinese dataset

(P = 0.002) while showing directional consistency in the Vietnamese and Han Chinese datasets (Figure 1, Supplemental Table 3). A novel locus was identified at 22q12.2 [rs6006426, OR (95%Cl) =1.097 (1.066, 1.130), P_{meta} = 3.31×10^{-10}, Table 1, Figure 1], demonstrating significance associations across all datasets included in this study (Supplemental Table 4). The lead SNP rs6006426, an intergenic SNPs near gene Oncostatin M (*OSM*) (Figure 2), exhibits a minor allele frequency (MAF) ranging from 37%-47% in East Asian (EAS) populations according to the 1000 Genome Project [24].

A gene-based test was conducted by MAGMA (v1.08), which uses a multiple regression approach to properly incorporate LD between markers and to detect multi-marker effects, [25] implemented in FUMA GWAS (Functional Mapping and Annotation of Genome Wide Association Studies) [26] using the result from the meta-analysis. In addition to the Major Histocompatibility Complex (MHC) region on chromosome 6, the *OSM* gene located on chromosome 22 was associated with PTB (ZSTAT=5.013; P = $2.68\text{x}10^{-7}$; P_{adj} = 0.005) surpassing the genome wide significance threshold ($P<2.65\times10^{-6}$, Supplemental Figure 3, Supplemental Table 5).

Association with smoking and mediation analysis

The impact of cigarette smoking on PTB is well-documented [6, 10], with evidence showing that smoking not only elevates the risk of contracting TB but also increases the likelihood of recurrence [27]. Additionally, smoking has been found to adversely affect treatment outcomes [28].

The established link between smoking and active PTB has been reaffirmed by previous findings within SCHS [29]. In light of this, we explored the relationship between rs6006426

and PTB, stratifying the analysis by smoking status (Supplemental Table 6). Our findings revealed a significant association between rs6006426 and PTB in individuals who have ever smoked [OR (95%Cl) = 1.146 (1.031, 1.273), P = 0.012], but this association was not observed in non-smokers (P = 0.229). Nonetheless, the interaction between smoking status and the rs6006426 variant in relation to PTB did not reach statistical significance ($P_{interaction}$ = 0.327).

We further investigated the relationship between the novel locus, smoking behaviour and PTB susceptibility using data from SCHS and Biobank Japan. rs6006426 showed significant association with cigarette smoking in both SCHS [OR (95%Cl) = 0.944 (0.898, 0.992), P = 0.022] and Biobank Japan [OR (95%Cl) =0.983 (0.966, 1.000), P = 0.044] [30]. Carriers of the A allele, which was associated with increased PTB risk, were found to be less likely to smoke (Table 2). To elucidate the influence of rs6006426 on PTB risk via smoking behaviour, we applied mediation analysis using Structural Equation Modeling (SEM), accounting for age and sex as covariates. The analysis indicated that overall while the A allele of rs6006426 reduces the likelihood of smoking, it concurrently elevates the risk for PTB ($\beta_{total-effect}$ = 0.093, $P_{total-effect}$ = 0.012). Interestingly, the mediation effect of rs6006426 on PTB through smoking was demonstrated to be protective ($\beta_{indirect-effect}$ = -0.004, $P_{indirect-effect}$ = 0.020, Table 3). Moreover, a significant direct effect was also observed ($\beta_{direct-effect}$ = 0.097, $P_{direct-effect}$ = 0.009). This finding suggests that rs6006426 exhibited a pleiotropic effect, as evidenced by its associations with various traits in the GWAS catalog. It implies that the relationship between rs6006426 and PTB is multifaceted, involving pathways both related and unrelated to smoking behaviour.

Functional mapping and annotations

The lead SNP rs6006426 and those in LD ($R^2 > 0.2$) with it, are noted in the GWAS catalog [31] for its associations with a range of traits, predominantly cardiovascular disease [21, 32, 33], heel bone mineral density [34, 35] and certain brain-related phenotypes [36, 37] (Supplemental Table 7), yet there have been no associations reported with TB. Through analysis of expression quantitative trait loci (eQTL) from publicly available databases, rs6006426 has been observed to influence the expression of *SF3A1* in various immune-related cell, such as monocytes, B cells and T cells. Notably, the allele associated with increased risk for PTB (allele A) was associated with a reduced expression of *SF3A1* in these cells (Supplemental Table 8).

Discussion

TB remains a significant global health burden, responsible for millions of infections and deaths annually, predominantly in low- and middle-income countries [1]. Genetic, environmental and socio-economic factors significantly contribute to TB susceptibility [8-11], and more recently the focus of GWAS has been to identify the genetic loci involved [15-21]. In this study, we performed a comprehensive PTB GWAS among East Asia population groups [20, 21] using new as well as previously published data. This approach led to the identification of a novel susceptibility locus for PTB at 22q12.2. Intriguingly, the lead SNP rs6006426 demonstrated a significant association with cigarette smoking, a well-established risk factor for TB [6, 10, 27, 28], in both the SCHS and Biobank Japan. When exploring the interplay between genetic marker, smoking and TB, our findings indicate that smoking significantly mediates the effect of rs6006426 on PTB. Notably, allele A of rs6006426 is associated with an increased risk of PTB, yet it paradoxically correlates with a lower likelihood of smoking. Our mediation analysis revealed that while the overall impact of the A

allele of rs6006426 increased the risk for PTB, its effect when mediated through smoking actually diminishes the risk (A allele, decreased the likelihood of smoking).

Previous GWAS have explored the genetic predisposition to TB in diverse populations, including those from Africa [15, 16], Russia [17], Iceland [18] and Asia [19-21]. However, these studies have yielded inconsistent results, which have been rarely replicable between studies. For instance, GWAS conducted in Ghanaian and Gambian identified loci at 18q11.2 and 11p13 as being associated with TB [15, 16], but only the 11p13 locus was replicated in studies in Russia [17] and the South African coloured population [22]. Additionally, the *ASAP1* gene was reported as a susceptibility locus for TB in a Russian cohort [17], but this finding was not replicated in an Icelandic population [18]. These inconsistencies across GWAS could stem from various factors, such as differing LD structures in distinct populations or phenotypic heterogeneity in case or control ascertainment. As *Mtb* lineages display substantial geographic variation, pathogen genomic variation might also underlie some of the difficulties in replicating susceptibility loci across populations [38]. In our study, we examined the association of SNPs identified in previous TB GWAS (Supplemental Table 2), and we did not observe consistent replication for any previously associated SNPs across all four datasets included in our study. The most promising replicated signal was rs557011located in the HLA region of chromosome 6, initially reported in the Icelandic study [18], which was also recently replicated in a multi-ancestry meta-analysis [39]. This association was significant in the Vietnamese, Han Chinese, and Japanese datasets and directionally consistent in the Singapore Chinese cohort. These findings underscore the importance of exercising caution when extrapolating GWAS results to different populations, highlighting the challenges of translating genetic associations across diverse ethnic groups.

In our study, we identified 22q12.2 to be a novel susceptibility locus for PTB through meta-analysis and independent replication across four East Asian populations. This finding was substantiated through both single-variant association tests and gene-based analysis. eQTL analysis indicated that rs6006426 modulates the expression of *SF3A1* in various immune-related cells, with the A allele correlating with decreased gene expression levels. The *SF3A1* gene encodes a component of the splicing factor 3a (SF3A) protein complex, which forms part of the mature U2 small nuclear ribonucleoprotein particle (snRNP). A previous study has demonstrated that inhibition of SF3A1 or SF3B1 leads to increased production of a short form of MyD88 mRNA ($MyD88_S$), which is a negative regulator of innate immunity through Toll-like receptor (TLR) signalling [40]. The levels of $MyD88_S$ are critical in determining the production of inflammatory cytokines in murine macrophages [41]. SF3A and SF3B mRNA splicing complexes function together in TLR signalling to regulate the production of $MyD88_S$, and thereby control the extent of innate immune activation [40, 41]. Considering that innate immune cells are the initial line of defence against *Mtb* [42] and that *Mtb* can potentially be eradicated by the innate immune system before the onset of an adaptive immune response [43], the modulation of *SF3A1* expression is particularly significant. Our findings suggest that individuals carrying the A allele of rs6006426 may exhibit reduced *SF3A1* expression, potentially leading to increased levels of $MyD88_S$ mRNA level. This increase could limit the activation of the innate immune system, rendering carriers more susceptible to active TB disease due to inadequate innate immune responses. A candidate gene study in the Chinese Han population has previously established a connection between SF3A1 and TB [44]. The SNPs identified in their study, rs2074733 and rs10376, exhibit weak LD with our sentinel SNP, with r^2 values ranging from 0.233 to 0.336 and 0.056 to 0.099, respectively, in the 1000 Genomes Project East Asian (EAS) population. While rs2074733 showed a significant association with PTB in the Singapore Chinese, Han Chinese, and Japanese, the direction of

this association was contrary to that in the candidate gene study in the Chinese Han population. This inconsistency could potentially be attributed to variations in the composition of the case groups across different studies. Our study found no association between rs10376 and PTB across all datasets analysed. *OSM* was significantly associated with PTB in gene-based test. A previous study reported OSM to be a novel mediator in the pathogenesis of TB and may be a potential therapeutic target to minimise the tissue damage associated with TB [45]. In addition, *OSM* was reported to be one of the key characteristic genes related to inflammation during the progression of latent tuberculosis infection (LTBI) to active TB [46].

The primary strength of our study lies in the meta-analysis of 4 East Asian populations by the integration of our data with publicly available summary statistics from East Asia, significantly augmenting the sample size and enhancing statistical power. However, it is important to acknowledge a limitation inherent in our methodology: the use of the general population as control subjects. This approach combines individuals with a range of TB exposure statuses, including those never exposed to *Mtb*, those who have been *Mtb* infected but remain asymptomatic (ie individuals with a positive IGRA, including latent individuals or subclinical TB), and those exposed but not infected. This heterogeneity within the control group potentially leads to an underestimation of the true effect size regarding the relationship between genetic markers and the risk of PTB but is unlikely to result in false positive findings.

In conclusion, our meta-analysis has led to the identification of 22q12.2 as a novel susceptibility locus for PTB. The PTB risk allele of the lead SNP reduces the expression level of *SF3A1*. SF3A1 function together with TLR signalling to regulate the production of $MyD88_S$, influencing the extent of innate immune activation. This finding reveals new

avenues for research, and further studies are necessary to elucidate the underlying mechanisms of this association.

Material and Methods

Study samples

The Singapore Chinese participants included in this study were enrolled in the Singapore Chinese Health Study (SCHS), a long-term prospective population-based cohort study focused on genetic and environmental determinants of cancer and other chronic diseases in Singapore [23]. A total of 63,257 participants (27,959 men and 35,298 women) who were 45 to 74 years old were enrolled between April 1993 and December 1998. At recruitment, subjects were interviewed in-person using a structured questionnaire, capturing data on demographics, anthropometrics and lifestyle factors including tobacco use. Since April 1994, a random sample of 3% of the subjects were re-contacted for donation of blood/buccal cells and spot urine specimens. In January 2001, the collection of biospecimens was expanded to include all enrolled participants who provided consent, with around half of the cohort eventually contributing blood or buccal samples. PTB cases were ascertained via linkage with the National Tuberculosis Notification Registry [47], leveraging Singapore's legal mandate for reporting all suspected and confirmed TB cases within 72 hours of initiating treatment and/or receiving laboratory confirmation. Diagnosis is predominantly driven by passive case detection, when patients present with symptoms, such as persistent cough, blood-stained sputum, fever, chills, and night sweats. Cases were diagnosed by positive sputum smear and confirmed by microbiological culture of *Mycobacterium tuberculosis*. Notifications primarily originate from public hospitals and the Tuberculosis Control Unit, supplemented by electronic records from the two Mycobacterial laboratories in Singapore, ensuring comprehensive data capture in the National Tuberculosis Notification Registry [47]. This

study was approved by the Institutional Review Board at the National University of Singapore, and written informed consent was obtained from all study participants.

Blood samples for DNA were collected from pulmonary TB patients from Ho Chi Minh City (HCMC) Vietnam who were recruited as part of a larger clinical study [48]. Briefly 2,091 newly diagnosed pulmonary TB patients attending Pham Ngoc Thach Hospital outpatients department or one of 8 District TB Units (DTUs) in HCMC were recruited to our genetics study between December 2008 and July 2011. Patients sampled for genetics were 18 years or older, HIV negative, had provided written informed consent, had no previous history of TB treatment and were sputum smear positive. DNA was extracted using Qiagen Blood Midi kits (Qiagen). DNA from 1,650 of these patients underwent whole genome genotyping at the Genome Institute of Singapore. The study was approved by the institutional Review Boards of the Hospital for Tropical Diseases HCMC Vietnam, Pham Ngoc Thach Hospital for Tuberculosis and Lung Disease HCMC Vietnam, Health Services of HCMC Vietnam, the Oxford Tropical Research Ethics Committee, Oxford University UK and the University of Melbourne Human Research Ethics Committee, Melbourne Australia (ID 21973). Vietnamese Kinh population controls are otherwise healthy adults with primary angle closure glaucoma who have been previously described [49].

Genotyping and Imputation

In SCHS, 27,308 DNA samples were whole genome genotyped using the Illumina Global Screening Array (GSA), including 18,114 samples genotyped on v1.0 and 9,194 samples genotyped on v2.0. An additional 2,161 independent subjects from the SCHS CAD-nested case-control study were genotyped on Illumina HumanOmniZhonghua Bead Chip. Comprehensive information regarding genotyping and quality control (QC) protocols has

been previously published [50-52]. After QC, single nucleotide polymorphism (SNP) alleles were standardized to the forward strand and mapped to the hg38 reference genome. Minimac4 (version 1.0.0) [53] was employed to impute additional autosomal SNPs using a local population-specific reference panels compromised of 9,770 whole-genome sequences of local Singaporean population samples obtained from the SG10K initiative (SG10K Health) [54] on the Research Assets Provisioning and Tracking Online Repository (RAPTOR) [55].

The Vietnamese samples were genotyped using either the Infinium OmniExpress-24 or OmniExpress-12 array. Detailed QC procedures are provided in Supplemental Table 9. 1,650 PTB cases were combined with 1,357 Vietnamese kinh population controls for analysis. Briefly, samples with call-rate below 95% and those exhibiting extremes heterozygosity (beyond ±3 standard deviations, N = 47) were excluded. Identity-by-state (IBS) analyses were conducted to identify first and second-degree related samples, with the lower call-rate sample from each detected pair removed from subsequent analyses (N = 83). PCA together with 1000 Genomes Projects reference populations and within the Vietnamese samples were performed to identify possible outliers from reported ethnicity and 12 samples were excluded. For SNP QC (Supplemental Table 10), SNPs that were monomorphic or rare (MAF < 1.0%) and SNPs with low call-rates (<95.0%) were excluded (N = 112,360). Sex chromosome and SNPs shown different missingness were removed (N = 7,088), together with SNPs displaying gross Hardy–Weinberg equilibrium (HWE) deviation ($P < 10^{-6}$, N = 13,601). Alleles for all SNPs were coded to the forward strand and mapped to hg19. IMPUTE v2 [56] was used to mutually impute variants with cosmopolitan 1000 Genomes haplotypes as reference panel (Phase 3) [24]. SNPs with low impute information score (< 0.6), MAF < 1.0% as well as non-biallelic SNPs were excluded from subsequent analyses.

Statistical analysis

Single-variant association tests were performed using an R package SAIGE (v1.3.0) [57], adjusted for the first three principal components in the association analysis to correct for the population stratification. SAIGE accounts for sample relatedness based on the generalized mixed models and manages case-control imbalance of binary traits. Genome-wide significance threshold was set at 5×10^{-8}. Effect size is estimated through Firth's Bias-Reduced Logistic Regression. To assess potential inflation in the study results, the genomic inflation factor (λ) was calculated, revealing only marginal inflation (λ ranging from 0.985 to 1.007, Supplementary Figure 1-3). To maximise sample size and increase statistical power, publicly available association summary statistics from Biobank Japan [21] (https://pheweb.jp/pheno/PTB) and a Han Chinese dataset [20] (https://doi.org/10.6084/m9.figshare.7006310) were integrated with PTB GWAS data from SCHS and Vietnam. An inverse variance-weighted meta-analysis was employed, assuming a fixed effects model to derive overall association values using META (v1.7) [58]. Heterogeneity among the combined data was evaluated using Cochran's Q [59] and a Cochran's Q *P*-value ($P_{\text{heterogeneity}}$) < 0.05 was determined to be significantly heterogeneous. To investigate the interplay between genetic markers, smoking habits, and PTB risk, Structural Equation Modeling (SEM) was utilized. Using the "gsem" command in Stata, we determined total effect of SNP on PTB (sum up of direct and indirect effect), direct effect of SNP on PTB (the effect of SNP on PTB absent smoking) and indirect effect of SNP on PTB through smoking (the effect of SNP on PTB that works through smoking) [60]. Age at interview and gender were included as covariates. Analyses were conducted using Stata/SE15.1. Two-sided $P < 0.05$ was considered statistically significant.

Functional mapping and annotation

Lead SNPs discovered in this study were functionally annotated using the SNP2GENE function in Functional Mapping and Annotation (FUMA, v1.6.1) [26]. SNPs in LD (r^2 > 0.6 in 1000G ASN panel) with sentinel SNPs were identified. Genes located within a 10 kb region flanking each lead SNP were mapped as regional genes at the locus of interest. FUMA performs MAGMA gene analysis (v1.08) [25] using the default SNP-wide mean model and 1000G EAS population as a reference panel. Expression quantitative trait loci (eQTL) was conducted on all regional genes using cell-type specific data for circulating immune cells sourced from van der Wijst et al. [61] (scRNA eQTLs), eQTL Catalogue [62] and the DICE (Database of Immune Cell Expression, Expression quantitative trait loci (eQTLs) and Epigenomics) project [63]. The significance of eQTLs was determined at the genome-wide level.

Data Availability

Summary statistics for PTB GWAS in Singaporean Chinese and Vietnamese will be available in figshare on publication.

Acknowledgements

We wish to thank the individuals who participated in our studies. We acknowledge the clinical staff who recruited patients in the Vietnam study from the following District TB Units (DTUs) in Ho Chi Minh City: Districts 1, 4, 5, 6 and 8, Tan Binh, Binh Thanh and Phu Nhuan; and also our colleagues from Pham Ngoc Thach Hospital for Tuberculosis and Lung Disease, particularly Nguyen Ngoc Lan and Nguyen Huu Lan and from Oxford University Clinical Research Unit, Hoang Thanh Hai and Vu Thi Ngoc Ha. We thank Siew-Hong Low of the National University of Singapore for supervising the fieldwork in the Singapore Chinese Health Study.

This work was supported by the National Health and Medical Research Council, Australia (Investigator grant APP1172853 to SJD); NHMRC (APP1056689) /A*STAR (12/1/21/24/6689) joint call to SJD/YYT; the USA National Institute of Health (U19AI162583 to SJD); This research was funded by Wellcome in whole, or in part (research training fellowship 081814/Z/06/Z to MC, 206724/Z/17/Z to NTTT and 089276/Z/09/Z to Major Overseas Program in Vietnam). The Singapore Chinese Health Study was supported by the National Institutes of Health (NIH) of the United States (Grants R01CA080205, R01CA144034, and UM1CA182876 to J-MY) and the National Medical Research Council, Singapore (NMRC/CIRG/1456/2016). W-P Koh was supported by the National Medical Research Council, Singapore [CSA-SI (MOH-000434)].

Author contributions

X.C., R.D, CCK, S.J.D were responsible for conceptualization and X.C., R.D, S.J.D wrote the manuscript. PVKT, DTMH, NTTT, NTQN, MC and SJD were responsible for the clinical study in Vietnam (including patient diagnosis, recruitment and sampling). CBEC, C-KH, J-MY, W-PK, RD, CCK were responsible for data collection and data generation that contributed to this manuscript from the Singapore Chinese Health Study. ZL, CCK were responsible for all data generation and QC for the Vietnamese datasets. All analysis was performed by XC and RD. SJD, CCK, MC, YYK was responsible for funding acquisition for the Vietnamese study. All authors reviewed and edited the manuscript.

References

1. Organization, W.H. *Global tuberculosis report 2023*. (World Health Organization. 7 November 2023).
2. Singh, R., et al., *Recent updates on drug resistance in Mycobacterium tuberculosis.* J Appl Microbiol, 2020. **128**(6): p. 1547-1567.
3. Houben, R.M. and P.J. Dodd, *The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling.* PLoS Med, 2016. **13**(10): p. e1002152.
4. Gupta, R.K., et al., *Discovery and validation of a personalized risk predictor for incident tuberculosis in low transmission settings.* Nat Med, 2020. **26**(12): p. 1941-1949.
5. Getahun, H., et al., *Latent Mycobacterium tuberculosis infection.* N Engl J Med, 2015. **372**(22): p. 2127-35.
6. Silva, D.R., et al., *Risk factors for tuberculosis: diabetes, smoking, alcohol use, and the use of other drugs.* J Bras Pneumol, 2018. **44**(2): p. 145-152.
7. García, J.I., et al., *New Developments and Insights in the Improvement of Mycobacterium tuberculosis Vaccines and Diagnostics Within the End TB Strategy.* Current Epidemiology Reports, 2021. **8**(2): p. 33-45.
8. Comstock, G.W., *Tuberculosis in twins: a re-analysis of the Prophit survey.* Am Rev Respir Dis, 1978. **117**(4): p. 621-4.
9. Nidoi, J., et al., *Impact of socio-economic factors on Tuberculosis treatment outcomes in north-eastern Uganda: a mixed methods study.* BMC Public Health, 2021. **21**(1): p. 2167.
10. Alavi-Naini, R., B. Sharifi-Mood, and M. Metanat, *Association between tuberculosis and smoking.* Int J High Risk Behav Addict, 2012. **1**(2): p. 71-4.
11. Restrepo, B.I., *Diabetes and Tuberculosis.* Microbiol Spectr, 2016. **4**(6).
12. Möller, M. and E.G. Hoal, *Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis.* Tuberculosis (Edinb), 2010. **90**(2): p. 71-83.
13. Apt, A. and I. Kramnik, *Man and mouse TB: contradictions and solutions.* Tuberculosis (Edinb), 2009. **89**(3): p. 195-8.
14. Abel, L., et al., *Human genetics of tuberculosis: a long and winding road.* Philos Trans R Soc Lond B Biol Sci, 2014. **369**(1645): p. 20130428.
15. Thye, T., et al., *Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2.* Nat Genet, 2010. **42**(9): p. 739-741.
16. Thye, T., et al., *Common variants at 11p13 are associated with susceptibility to tuberculosis.* Nat Genet, 2012. **44**(3): p. 257-9.
17. Curtis, J., et al., *Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration.* Nature Genetics, 2015. **47**(5): p. 523-527.
18. Sveinbjornsson, G., et al., *HLA class II sequence variants influence tuberculosis risk in populations of European ancestry.* Nat Genet, 2016. **48**(3): p. 318-22.
19. Qi, H., et al., *Discovery of susceptibility loci associated with tuberculosis in Han Chinese.* Human Molecular Genetics, 2017. **26**(23): p. 4752-4763.
20. Zheng, R., et al., *Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese.* Nature Communications, 2018. **9**(1): p. 4072.

21. Sakaue, S., et al., *A cross-population atlas of genetic associations for 220 human phenotypes.* Nature Genetics, 2021. **53**(10): p. 1415-1424.
22. Chimusa, E.R., et al., *Genome-wide association study of ancestry-specific TB risk in the South African Coloured population.* Hum Mol Genet, 2014. **23**(3): p. 796-809.
23. Hankin, J.H., et al., *Singapore Chinese Health Study: development, validation, and calibration of the quantitative food frequency questionnaire.* Nutrition and cancer, 2001. **39**(2): p. 187-195.
24. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
25. de Leeuw, C.A., et al., *MAGMA: Generalized Gene-Set Analysis of GWAS Data.* PLOS Computational Biology, 2015. **11**(4): p. e1004219.
26. Watanabe, K., et al., *Functional mapping and annotation of genetic associations with FUMA.* Nature communications, 2017. **8**(1): p. 1826.
27. Pourali, F., et al., *Relationship between smoking and tuberculosis recurrence: A systematic review and meta-analysis.* Indian J Tuberc, 2023. **70**(4): p. 475-482.
28. Chi, C.L., et al., *Smoking adversely affects treatment response, outcome and relapse in tuberculosis.* European Respiratory Journal, 2015. **45**(3): p. 738.
29. Soh, A.Z., et al., *Alcohol drinking and cigarette smoking in relation to risk of active tuberculosis: prospective cohort study.* BMJ Open Respir Res, 2017. **4**(1): p. e000247.
30. Masahiro, K., et al., *Insights from complex trait fine-mapping across diverse populations.* medRxiv, 2021: p. 2021.09.03.21262975.
31. Sollis, E., et al., *The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource.* Nucleic Acids Res, 2023. **51**(D1): p. D977-d985.
32. Koyama, S., et al., *Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease.* Nature Genetics, 2020. **52**(11): p. 1169-1177.
33. Aragam, K.G., et al., *Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants.* Nature Genetics, 2022. **54**(12): p. 1803-1815.
34. Morris, J.A., et al., *An atlas of genetic influences on osteoporosis in humans and mice.* Nature Genetics, 2019. **51**(2): p. 258-266.
35. Kim, S.K., *Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture.* PLOS ONE, 2018. **13**(7): p. e0200785.
36. Zhao, B., et al., *Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (n = 17,706).* Mol Psychiatry, 2021. **26**(8): p. 3943-3955.
37. Smith, S.M., et al., *An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank.* Nature Neuroscience, 2021. **24**(5): p. 737-745.
38. Gagneux, S., et al., *Variable host-pathogen compatibility in Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2006. **103**(8): p. 2869-73.
39. Schurz, H., et al., *Multi-ancestry meta-analysis of host genetic susceptibility to tuberculosis identifies shared genetic architecture.* eLife, 2024. **13**: p. e84394.
40. De Arras, L. and S. Alper, *Limiting of the innate immune response by SF3A-dependent control of MyD88 alternative mRNA splicing.* PLoS Genet, 2013. **9**(10): p. e1003855.
41. O'Connor, B.P., et al., *Regulation of Toll-like Receptor Signaling by the SF3a mRNA Splicing Complex.* PLOS Genetics, 2015. **11**(2): p. e1004932.

42. Ravesloot-Chávez, M.M., E.V. Dis, and S.A. Stanley, *The Innate Immune Response to Mycobacterium tuberculosis Infection.* Annual Review of Immunology, 2021. **39**(1): p. 611-637.
43. Lerner, T.R., S. Borel, and M.G. Gutierrez, *The innate immune response in human tuberculosis.* Cell Microbiol, 2015. **17**(9): p. 1277-85.
44. Zhang, J., et al., *Association between a single nucleotide polymorphism of the SF3A1 gene and tuberculosis in a Chinese Han population: a case□control study*. 21 November 2022, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2252919/v1].
45. O'Kane, C.M., P.T. Elkington, and J.S. Friedland, *Monocyte-dependent oncostatin M and TNF-alpha synergize to stimulate unopposed matrix metalloproteinase-1/3 secretion from human lung fibroblasts in tuberculosis.* Eur J Immunol, 2008. **38**(5): p. 1321-30.
46. Ma, S., et al., *Predicting the Progress of Tuberculosis by Inflammatory Response-Related Genes Based on Multiple Machine Learning Comprehensive Analysis.* J Immunol Res, 2023. **2023**: p. 7829286.
47. Chee, C.B. and L. James, *The Singapore Tuberculosis Elimination Programme: the first five years.* Bull World Health Organ, 2003. **81**(3): p. 217-21.
48. Thai, P.V.K., et al., *Bacterial risk factors for treatment failure and relapse among patients with isoniazid resistant tuberculosis.* BMC Infectious Diseases, 2018. **18**(1): p. 112.
49. Khor, C.C., et al., *Genome-wide association study identifies five new susceptibility loci for primary angle closure glaucoma.* Nat Genet, 2016. **48**(5): p. 556-62.
50. Dorajoo, R., et al., *Loci for human leukocyte telomere length in the Singaporean Chinese population and trans-ethnic genetic studies.* Nature Communications, 2019. **10**(1): p. 2491.
51. Chang, X., et al., *Low frequency variants associated with leukocyte telomere length in the Singapore Chinese population.* Communications Biology, 2021. **4**(1): p. 519.
52. Chang, X., et al., *Utility of genetic and non-genetic risk factors in predicting coronary heart disease in Singaporean Chinese.* Eur J Prev Cardiol, 2017. **24**(2): p. 153-160.
53. Das, S., et al., *Next-generation genotype imputation service and methods.* Nature Genetics, 2016. **48**(10): p. 1284-1287.
54. Wong, E., et al., *The Singapore National Precision Medicine Strategy.* Nat Genet, 2023. **55**(2): p. 178-186.
55. Shih, C.C., et al., *A five-safes approach to a secure and scalable genomics data repository.* iScience, 2023. **26**(4): p. 106546.
56. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
57. Zhou, W., et al., *Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.* Nature Genetics, 2018. **50**(9): p. 1335-1341.
58. Liu, J.Z., et al., *Meta-analysis and imputation refines the association of 15q25 with smoking quantity.* Nat Genet, 2010. **42**(5): p. 436-40.
59. Conover, W.J., *Practical nonparametric statistics*. Vol. 350. 1999: john wiley & sons.
60. Burgess, S., et al., *Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways.* International journal of epidemiology, 2015. **44**(2): p. 484-495.

562 61. van der Wijst, M.G.P., et al., *Single-cell RNA sequencing identifies celltype-specific cis-*
563 *eQTLs and co-expression QTLs.* Nat Genet, 2018. **50**(4): p. 493-497.
564 62. Kerimov, N., et al., *eQTL Catalogue 2023: New datasets, X chromosome QTLs, and*
565 *improved detection and visualisation of transcript-level QTLs.* PLOS Genetics, 2023.
566 **19**(9): p. e1010932.
567 63. Schmiedel, B.J., et al., *Impact of Genetic Polymorphisms on Human Immune Cell Gene*
568 *Expression.* Cell, 2018. **175**(6): p. 1701-1715.e16.
569

Table 1. Summary statistics of genome-wide associations identified after meta-analysis of 4 East Asian populations for pulmonary tuberculosis.

					11,841 cases/197,373 controls		
SNP	chromosome	position	Other allele	Effect allele	OR (95%Cl)	*P*	$P_{heterogeneity}$
rs6006426	22	30669883	G	A	1.097 (1.066, 1.130)	3.31×10^{-10}	0.079

OR: Odds ratio; Cl: confidence interval.

Table 2. The association between smoking (never or ever smoking) and rs6006426 in SCHS and Biobank Japan.

					SCHS		Biobank Japan	
SNP	chromosome	position	Other allele	Effect allele	OR (95%Cl)	*P*	OR (95%Cl)	*P*
rs6006426	22	30669883	G	A	0.944 (0.898, 0.992)	0.022	0.983 (0.966, 1.000)	0.044

OR: Odds ratio; Cl: confidence interval.

Table 3. Effect of rs6006426 on pulmonary tuberculosis mediated through smoking in the SCHS.

	Direct effect			Indirect effect			Total effect		
	β	se	*P*	β	se	*P*	β	se	*P*
rs6006426	0.097	0.037	0.009	-0.004	0.002	0.020	0.093	0.037	0.012

se: standard error

Figure 1. Meta-analysis for PTB in East Asia. **a** Manhattan plot of the meta-analysis of 4 Asian populations. Two hits at chromosome 6 and 22 were identified beyond the genome-wide significance threshold ($P < 5 \times 10^{-8}$, red line) **b** QQ-plot of observed compared to expected *P*-values indicated minimal inflation ($\lambda = 1.028$).

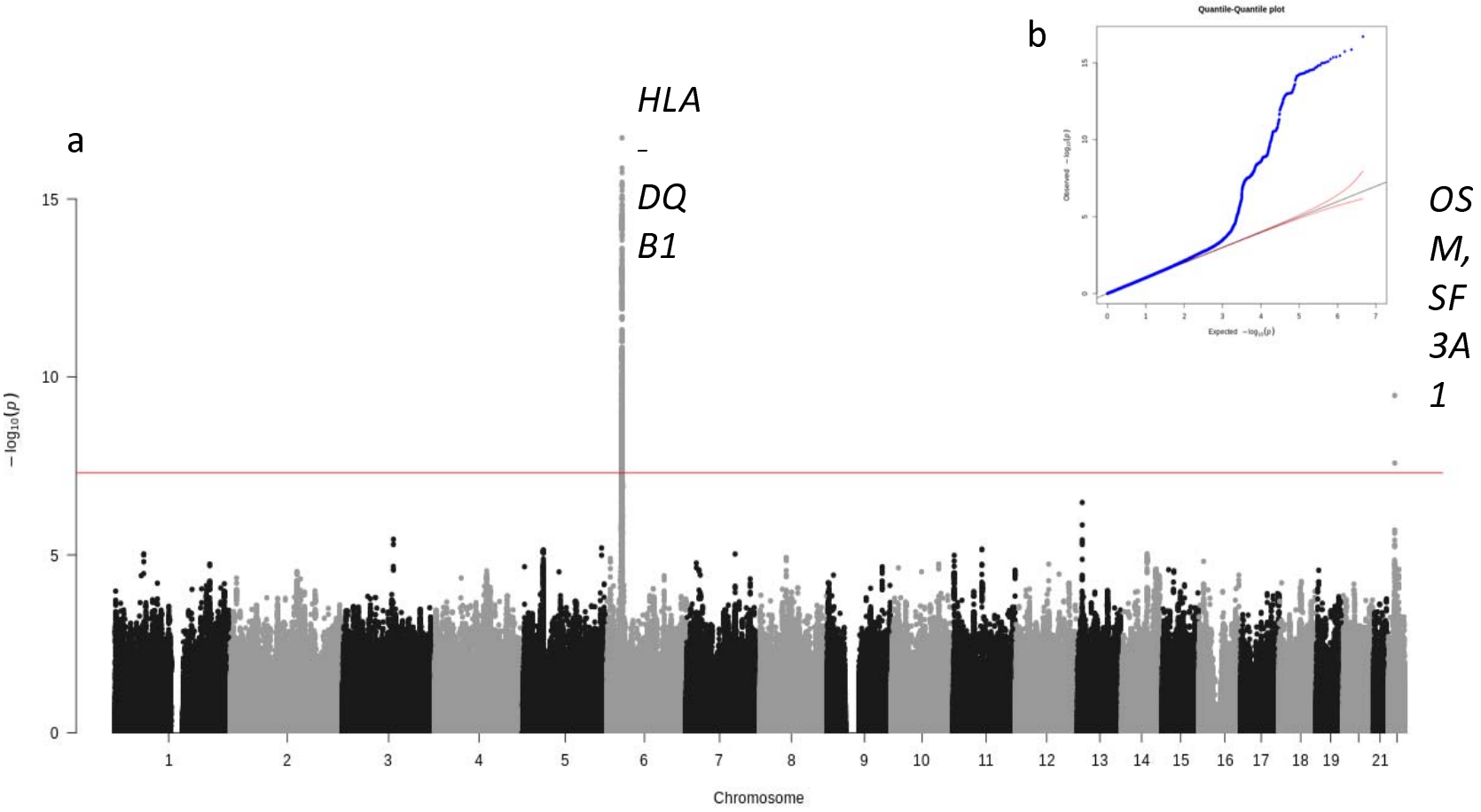

Figure 2. Regional SNP associations at 22q12.2 in the meta-analysis of 4 Asian populations. Lead SNP indicated as purple diamonds. LD (r^2) data of SNPs based on ASN panels of 1000Genome database. Plots plotted using LocusZoom (http://locuszoom.org/).