OPEN

Journal of Human Genetics (2017) 62, 1015–1022
Official journal of the Japan Society of Human Genetics
www.nature.com/jhg

ORIGINAL ARTICLE

Pathogen lineage-based genome-wide association study identified *CD53* as susceptible locus in tuberculosis

Yosuke Omae[1], Licht Toyo-oka[1], Hideki Yanai[2], Supalert Nedsuwan[3], Sukanya Wattanapokayakit[4], Nusara Satproedprai[4], Nat Smittipat[5], Prasit Palittapongarnpim[6], Pathom Sawanpanyalert[7], Wimala Inunchot[4], Ekawat Pasomsub[6], Nuanjun Wichukchinda[4], Taisei Mushiroda[8], Michiaki Kubo[9], Katsushi Tokunaga[1] and Surakameth Mahasirimongkol[4]

Tuberculosis (TB) is known to be affected by host genetic factors. We reported a specific genetic risk factor through a genome-wide association study (GWAS) that focused on young age onset TB. In this study, we further focused on the heterogeneity of *Mycobacterium tuberculosis* (*M. tb*) lineages and assessed its possible interaction with age at onset on host genetic factors. We identified the pathogen lineage in 686 Thai TB cases and GWAS stratified by both infected pathogen lineage information and age at onset revealed a genome-wide significant association of one single-nucleotide polymorphism (SNP) on chromosome 1p13, which was specifically associated with non-Beijing lineage-infected old age onset cases ($P = 2.54E\text{-}08$, OR = 1.74 (95% CI = 1.43–2.12)), when we compared them to the population-matched healthy controls. This SNP locates near the *CD53* gene, which encodes a leukocyte surface glycoprotein. Interestingly, the expression of *CD53* was also correlated with the patients' active TB status. This is the first report of a pathogen lineage-based genome-wide association study. The results suggested that host genetic risk in TB is depended upon the pathogen genetic background and demonstrate the importance of analyzing the interaction between host and pathogen genomes in TB.
Journal of Human Genetics (2017) **62,** 1015–1022; doi:10.1038/jhg.2017.82; published online 7 September 2017

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*M. tb*), is one of the three most common infectious diseases in the world. Although this pathogen infects one-third of the world population, only 5–15% of infected people develop TB, and the infection in the remaining 90% of infected people stays in a dormant stage throughout their life,[1] suggesting the contribution of host genetic factors to TB onset. The contribution of host genetic factors to TB onset have been demonstrated in twin studies, in which monozygotic twins showed a 2.5 times higher chance of developing TB than dizygotic twins.[2]

To identify host genetic factors in TB, genome-wide association studies (GWAS) have been performed in which differences in genotype frequencies were compared between cases and controls. In infectious diseases, GWAS have successfully identified risk genes with moderate to large effect size (at risk odds ratios >1.5). For example, *HLA-DR -DQ* genes and the *NOD2* gene have been identified as risk genes in leprosy.[3] The sickle hemoglobin (*HbS*) gene in malaria[4] and the Complement factor H (*CFH*) gene in Neisseria meningitides[5] were also identified as risk genes. In contrast, although several GWAS in TB had been reported to date,[6–9] no report has identified a risk gene with a moderate to large effect size.

We previously conducted an age-stratified GWAS by focusing on the heterogeneity of TB onset.[10] After infection by the pathogen, about 5% of infected people develop TB within 2 years, and this is called primary TB.[11] Another 5% of infected people develop TB more than 2 years after their infection, and this is called reactivated TB. It is difficult to distinguish between primary TB patients and reactivated TB patients because surrogate clinical biomarkers or definite clinical definitions are not routinely available. We proposed that the age at TB onset might be one available classifier, and our age-stratified GWAS based on the classification threshold of 45 years of age found an association within chromosome 20q12 in young age onset cases. A

[1]Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; [2]Fukujuji Hospital and Research Institute of Tuberculosis (RIT), Japan Anti-Tuberculosis Association (JATA), Kiyose, Japan; [3]Chiangrai Prachanukroh Hospital, Ministry of Public Health, Chiang Rai, Thailand; [4]Medical Genetics Center, Medical Life Sciences Institute, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand; [5]Tuberculosis Research Laboratory, National Center for Genetic Engineering and Biotechnology, National Science and Technology Development Agency, Pathum Thani, Thailand; [6]Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, Thailand; [7]Food and Drug Administration, Ministry of Public Health, Nonthaburi, Thailand; [8]Laboratory for Pharmacogenomics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan and [9]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
Correspondence: Dr S Mahasirimongkol, Medical Genetics Center, Medical Life Sciences Institute, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Nonthaburi 11000, Thailand.
E-mail: surakameth.m@dmsc.mail.go.th
Received 20 February 2017; revised 11 July 2017; accepted 14 July 2017; published online 7 September 2017

meta-analysis in Thai and Japanese populations reached the genome-wide significance level, and its odds ratio was 1.73 for a moderate effect size.[10] This finding suggests that host genetic risks for TB can be affected by age at onset.

In addition to the heterogeneity of TB onset, TB has heterogeneity in its pathogen genome.[12] Based on the genomic region of difference, six major global lineages have been reported in the world. In Thailand, almost 50% of isolates were accounted for by the Beijing lineage and the remaining 50% were accounted for by East-African Indian (EAI) or other lineages.[13,14] This heterogeneity in one country provides an advantage in comparing the effect of TB lineages in the same patient genetic background.

In the present study, we collected the host genome and the pathogen genome from each patient and considered the heterogeneity of the *M. tb* genome. A pathogen lineage information-based GWAS of TB was conducted to assess a possible interaction between host genetic factors and the pathogen genome. Herein, we report one lineage-dependent TB risk factor.

MATERIALS AND METHODS

Collection of TB patient samples and healthy control samples

TB patients in this study were primarily recruited from Chiang Rai province in Thailand and some patients were recruited from Bangkok or Lampang provinces. Diagnoses of pulmonary TB was confirmed by mycobacterial culture of *M. tuberculosis* from each patient's sputum and several tests as described previously.[15] Any patients infected with the human immunodeficiency virus were excluded from this study. The healthy control samples were recruited from the blood donors in Chiang Rai province and none of the controls had a previous history of TB disease at the time of blood collection. Blood samples and infected *M. tb* samples were collected and used for DNA extraction.

Genome-wide single-nucleotide polymorphism genotyping of host genome samples and its quality control

Host genomic DNA was applied to the Illumina Human610-Quad BeadChip (616 794 single-nucleotide polymorphisms (SNPs)) or Illumina HumanOmniExpressExome-8 v1.2 BeadChip (964 193 Markers for 938 764 independent SNPs) to perform genome-wide SNP genotyping. UCSC hg19 was used as reference genome and overlapping 338 476 SNP genotypes between two platforms were included in this study. Samples with an overall call rate of more than 98% were included and quality controls (QCs) for SNP genotypes from genome-wide genotyping were carried out using the following three thresholds: SNP call rate $\geqslant$95%, minor allele frequency (MAF) $\geqslant$5% and Hardy–Weinberg equilibrium *P*-value $\geqslant$0.001 in healthy controls. In the final analysis, 266 604 autosomal SNPs passed the QC and were used for the association analysis.

All the samples used for GWAS applied to the identity by descent testing to find cryptic relatedness and all the remaining sample pairs after sample filtering showed the PI_HAT values less than 0.1875, which is halfway between the expected identity by descent for second- and third-degree relatives.[16] Principal component analysis (PCA) using the public Hapmap data (GSE17205 (CEU), GSE17206 (CHB+JPT) and GSE17207 (YRI)) as controls revealed that all of the Thai cases belonged to the Asian population (Supplementary Figure S1a). As the Thai population is known to vary in its ethnicity, PCA using only the Thai samples was also performed and 1457 out of 1755 samples were selected based on the PCA result to match the genetic background between cases and controls (Supplementary Figure S1b). The 263 cases who could be tracked by their infected pathogen isolates and 282 healthy controls were included from our previous GWAS report of the Thai population,[10] and named as the first data set. The other 423 cases and 489 healthy controls were newly recruited in this study and were named as the replication data set. In total, 686 cases and 771 controls were included and infected *M. tb* information was available for all of the patients. Clinical characteristics of the study cases and controls are summarized in Supplementary Table S1.

Collection and lineage detection of *M. tb* samples

M. tb samples were cultured on Lowenstein-Jensen medium and their genomic DNA was extracted using an enzymatic lysis method.[13] Smittipat *et al.*[17] performed a PCR-based large sequence polymorphism detection method and spoligotyping and determined the lineage of each *M. tb* isolate. In short, the genomic regions TbD1, RD105, RD239, pks15/1 and RD750 of *M. tb* were analyzed and each isolate was classified into one of five independent lineage groups: EAI (TbD1 present, RD239 absent, also called Lineage 1); Beijing (TbD1 absent, RD105 absent, also called Lineage 2); Euro-American (TBD1 absent, 7 bp deletion at pks15/1, also called Lineage 4); Central Asian strain (CAS; TbD1 absent, RD750 absent, also called Lineage 3); or others (TbD1 absent, the other four markers intact). Classification by spoligotyping was consistent with the PCR-based result. All of the cases in this study had a single lineage infection.

Regional SNP imputation analysis

An imputation method was applied to estimate genotypes in a candidate region by utilizing 1000 Genomes Project (Phase III) data as a reference panel. In this study, IMPUTE2 software was used to predict the genotypes of untyped or missing SNPs.[18] A 1 Mb window size was applied for the candidate region. To control the quality of imputed genotypes, the imputation probability threshold of 0.9 recommended by the developer was applied, and SNPs with more than 1% un-imputed genotype data, an MAF less than 1% and an Hardy–Weinberg equilibrium *P*-value less than 0.0001 were eliminated. Regional association plot was written by LocusZoom.[19] Linkage disequilibrium maps were written by Haploview software using genotype of SNPs shared among our Thai data set, GSE17205 (Hapmap CEU), GSE17206 (Hapmap CHB+JPT) and GSE17207 (Hapmap YRI).[20]

Statistical analysis

In the GWAS and imputation analysis, a χ^2 test was applied to a two-by-two contingency table in an allele frequency model. A quantile–quantile plot of the distribution of test statistics showed that its genomic inflation factor was 1.066. Significance thresholds after Bonferroni correction for multiple testing by the number of QC passed autosomal SNPs were set to 1.88E-07 (0.05/266 604) in this study.

Classification of age at onset was conducted using a 45-year-old threshold based on our previous report.[10] To simplify the association analysis in each subgroup, case genotype frequencies in one subgroup were compared with those in all of the controls. We assumed that control individuals older than 45 years of age do not have the genetic risk factors for young age onset TB, and that control individuals younger than 45 years of age have the potential to progress TB later in their life but that the proportion that progresses to TB is less than 1.7% (one-third of young controls infected by *M. tb* without symptoms, of which almost 5% can develop reactivated TB), which is an acceptable percentage.

All cluster plots for SNPs with *P*-values $<$1E-05 in a χ^2 test of all the subgroups were checked by visual inspection and SNPs with ambiguous genotype calls were excluded.

Analysis of blood expression profiles in public transcriptome data

We analyzed genome-wide transcriptional profiling data from TB patients' blood.[21] The data can be obtained from the GEO database (GSE19435, GSE19439 and GSE19442). GSE19435 is longitudinal blood transcriptional profiles of active TB patients in United Kingdom (UK) before and after drug treatment to identify blood transcriptional signatures for monitoring efficacy of treatment and host response to infection with *M. tb*. GSE19439 is transcriptional profiles in active and latent TB in UK to compare gene expressions between active TB patients, who were symptomatic and confirmed by isolation of *M. tb* on culture of sputum or bronchoalveolar lavage fluid, and latent TB patients, who had no clinical, radiological or microbiological evidence of active infection but positive by tuberculin skin test and Interferon-Gamma Release assay. GSE19442 is transcriptional profiles of active and latent TB in South Africa to analyze the transcriptional profiles in a different population with high TB-burden. Berry *et al.* obtained all the data using the Illumina Human HT-12 V3 BeadChip arrays (Illumina) and performed per-chip normalization by

Illumina's BeadStudio version 2 software to generate signal intensity values from the scans, subtract background and scale each microarray to the median average intensity for all samples.[21] For our gene expression analysis, we applied a QC threshold with detection *P*-value 0.05 to their normalized data and QC passed data were subjected to the analysis.

Blood collection and PBMC isolation in the Thai population

All blood samples were collected at Chiangrai Prachanukroh hospital in Chiangrai province, north of Thailand. Briefly, 15 ml of whole blood samples were collected in a sodium heparin tube and sent to the department of medical sciences, Nonthaburi using overnight courier. On the next day, peripheral blood mononuclear cells (PBMCs) were isolated using Ficoll-Paque Plus (Amersham Biosciences) in Leucosep tubes (Greiner Bio-One, Frickenhausen, Germany). The Leucosep tube was prefilled with 15 ml of Ficoll-Paque Plus, then 15 ml of heparinized blood were diluted with equal volume of phosphate-buffered saline and poured into the prefilled Leucosep tube. The tube was centrifuged at 1000 *g* for 10 min at room temperature to separate the PBMCs. Isolated PBMC was kept in liquid nitrogen until further use.

Multi-color flow cytometry analysis

Frozen PBMCs were slowly thawed and rest at least 6 h in 37 °C, 5% CO_2. After rest, approximately 10^6 viable PBMCs were subjected to staining with fluorescence-labeled antibody against cell surface antigen as follows; anti-human CD3-APC (eBioscience, USA), mouse anti-human CD4-FITC (BD Biosciences, USA), anti-human CD8α-PerCP (R&D Systems, USA) or CD14-PerCP (Molecular Probes, USA) and mouse anti-human CD53 PE (BD Biosciences). Concentrations for each labeled antibody were used according to the manufacturer. Stained PBMCs were then analyzed in a BD FACSCalibur cell analyzer for percentage of CD3+/CD4+ cells, CD3+/CD8+ cells and median fluorescence intensity of CD53 on each cell population.

eQTL analysis

The correlation between the candidate SNP genotype and gene expression on chromosome 1p13 was examined using data available from the GTEx portal database at the BROAD Institute (http://gtex-portal.org/home).[22]

Ethics statement

The protocol of this study was approved by the Human Genome, Gene Analysis Research Ethics Committee of the Graduate School of Medicine, The University of Tokyo, RIKEN Yokohama Campus Ethics Committee and the Institute for the Development of Human Research Protection (IHRP) of the Ministry of Public Health in Thailand. All the experiments were performed in accordance with the relevant guidelines and regulations. All adult subjects provided written informed consent, and a parent or guardian of any child participant provided informed consent on their behalf.

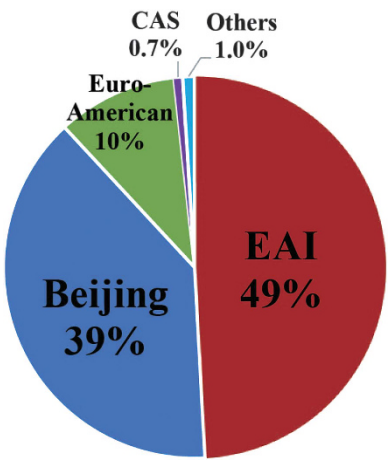

Figure 1 Distribution of each *M. tuberculosis* lineage among 686 TB patients.

RESULTS

We collected both the patient genome-wide SNP genotype data and their infected *M. tb* lineage data for all 978 Thai TB case samples included in this study. The 263 cases that could be tracked by their infected pathogen isolates and 282 controls were included from our previous report.[10] PCA was carried out to reduce population stratification by matching the genetic background between cases and controls (Supplementary Figures S1a and b). After the PCA, 686 cases and 771 controls were selected. The genomic inflation factor between cases and controls after the selection was 1.066, suggesting that population stratification between selected cases and controls is acceptable (Supplementary Figure S2).

Among the 686 selected individual cases, EAI lineage and Beijing lineage were predominant (49% and 39%, respectively) (Figure 1). This result is consistent with previous epidemiological reports in Thailand.[13,14] A small proportion (10%) were Euro-American lineage, and CAS lineage or others were also observed at low frequencies (0.7% and 1.2%, respectively) (Figure 1).

We first conducted the genome-wide association analysis using all of the cases and controls. Unfortunately, after applying Bonferroni correction by the number of individual SNP genotypes tested in the combined data set (significance threshold $\alpha = 1.88E\text{-}07$ from 0.05-/266 604), none of the SNPs showed a statistically significant difference (Supplementary Figure S2). We also considered the age at TB onset in this data set, but we could not find genome-wide significant association from 219 young cases with age at onset less than 45 years of age or 467 old cases with age at onset older than 45 years of age (Supplementary Figures S3a and b).

When we classified cases into Beijing lineage- and non-Beijing lineage-infected group, although the number of cases decreased, one SNP on the chromosome 1p13 locus passed the significance criteria considering the number of analyzed SNPs in non-Beijing lineage-infected group (rs1418425: $P = 1.58E\text{-}07$, OR = 1.62 (95% confidence interval (CI) = 1.35–1.93) (Table 1 and Supplementary Figure S3d). This association was identified in the first data set ($P = 4.30E\text{-}04$, case MAF: 0.356 versus control MAF: 0.246) and the replication data set ($P = 7.05E\text{-}05$, case MAF: 0.380 versus control MAF: 0.278), suggesting the reproducibility of this association in two independent samples (Table 1). Significant association of rs1418425 was not observed in the Beijing lineage infection ($P = 0.0138$, OR = 1.31 (95% CI = 1.06–1.62) (Table 1 and Supplementary Figure S3c). We further considered the age at TB onset in non-Beijing lineage-infected cases, and found that association of rs1418425 was dependent on the old age onset and reached the conservative genome-wide significance level ($\alpha = 5.00E\text{-}$ -08) in the old age onset group ($P = 2.54E\text{-}08$, OR = 1.74 (95% CI = 1.43–2.12) (Figures 2a and b and Table 1). The association of rs1418425 was not observed among non-Beijing lineage-infected and young age onset cases ($P = 0.115$, OR = 1.28 (95% CI = 0.94–1.75) (Supplementary Table S2 and Supplementary Figure S3f). These results suggest the risk of rs1418425 and this locus could be both lineage- and age-dependent. Another SNP (rs1494320), which was genotyped and in moderate linkage disequilibrium ($r^2 = 0.59$) with rs1418425 in the 1000 genome phase I Asian population, also passed Bonferroni corrected significance threshold ($P = 7.84E\text{-}08$, OR = 1.71 (95% CI = 1.40–2.08)) (Table 1). Interestingly, rs1494320 was located in the intronic region of *CD53*, a leukocyte surface glycoprotein and previously reported as a cis-expression quantitative trait locus (eQTL) of *CD53* gene expression level in dendritic cells infected by *M. tb*.[23]

During the genome evolution of *M. tb*, TbD1 deletion occurred and *M. tb* was separated into the EAI lineage (with TbD1 present, also called the ancient strain) and the non-EAI lineage (TbD1 absent, also

Table 1 Significant association of SNPs on chromosome 1p13 in pathogen lineage-based GWAS

						1st data set					*Replication data set*					*Combined*	
rsID	*Chr*	*Position (hg19)*	*Minor/ major*	*Age*	*Lineage*	*Case count*	*Case MAF*	*Control count*	*Control MAF*	P	*case count*	*case MAF*	*control count*	*control MAF*	P	*OR (95%CI)*	P
rs1418425	1	111 468 886	A/G	ALL	Non-Beijing	170	0.356	282	0.246	4.30E-04	249	0.380	489	0.278	7.05E-05	1.62 (1.35–1.93)	1.58E-07
					Beijing	93	0.290	282	0.246	0.235	174	0.339	489	0.278	0.0321	1.31 (1.06–1.62)	0.0138
rs1418425	1	111 468 886	A/G	Old	Non-Beijing	130	0.365	282	0.246	4.34E-04	182	0.404	489	0.278	9.92E-06	1.74 (1.43–2.12)	2.54E-08
					Beijing	58	0.353	282	0.246	0.0174	97	0.294	489	0.278	0.657	1.27 (0.98–1.66)	0.0743
rs1494320	1	111 422 188	G/A	Old	Non-Beijing	130	0.377	282	0.246	1.21E-04	182	0.393	489	0.282	1.01E-04	1.71 (1.40–2.08)	7.84E-08
					Beijing	58	0.371	282	0.246	5.92E-03	97	0.304	489	0.282	0.537	1.33 (1.02–1.73)	0.0319

Abbreviations: CI, confidence interval; Chr, chromosome; MAF, minor allele frequency; OR, odds ratio.

called the modern strain). Later, non-EAI linage was further separated into the Beijing lineage and other lineages including Euro-American, CAS, etc. We therefore checked the pathogen lineage dependency of the significant SNPs by focusing on the risk allele frequencies in each lineage-infected group. This analysis indicated that both SNPs showed higher risk allele frequencies in EAI lineage-infected cases and in Euro-American lineage-infected cases than those in Beijing lineage-infected cases (Supplementary Table S3). This result suggests that the observed risk of these SNPs is similar between EAI and Euro-American lineages.

The most significant SNP in this study, rs1418425, was located in the intergenic region at 26 kbp 3′ of the *CD53* gene and at 21 kbp 3′ of the *LRIF1* gene (Figure 3). CD53 is a leukocyte surface antigen and a member of the transmembrane 4 superfamily, tetraspanin. The CD53 protein is expressed mainly in the lymphoid-myeloid lineage[24] and familial deficiency of CD53 protein expression was associated with recurrent infectious diseases caused by bacteria, fungi and viruses, suggesting its important role in immunity.[25] Ligand-dependent nuclear receptor interacting factor 1 (LRIF1) is a nuclear protein that is known to be involved in the inactivation of the human X chromosome;[26–28] however, its function in immunity is unknown. SNPs around *CD53* gene and *LRIF1* gene constructed mild linkage disequilibrium (LD) in Thai samples (Figure 3). We then analyzed the gene expression of *CD53* and *LRIF1* in published blood transcriptional profiling data sets of human tuberculosis patients in UK or South Africa.[21] The expression of *CD53* was significantly increased in active TB patients compared with healthy controls and its increased expression was suppressed when the treatment of TB patients progressed (Supplementary Figure S4a). On the other hand, the expression of *LRIF1* was not significantly changed in active TB patients (Supplementary Figure S4a). Furthermore, *CD53* gene expression in active TB patients was even higher than that in latent TB patients and the increased expression of *CD53* was consistent among the UK and South African populations (Supplementary Figures S4b and c). Again, these increased expression levels were not observed for *LRIF1* gene (Supplementary Figures S4b and c). These results suggest that the expression of *CD53* gene can be correlated with the patients' active TB status. We further analyzed the cell surface expression of CD53 antigen in Thai TB patients' blood. Multi-color flow cytometry analysis revealed that the surface expression of CD53 on CD4+ and CD8+ T lymphocytes in TB patients were increased compared with healthy controls (Supplementary Figure S4d). The increased surface expression of CD53 in TB patients was not observed on CD14+ monocytes (Supplementary Figure S4d). These results suggest that surface protein expression of CD53 can be increased in TB patients on T lymphocytes.

DISCUSSION

In this study, we first conducted pathogen lineage-based genome-wide association studies and identified two SNPs around *CD53* that are a significant risk for old age TB onset under non-Beijing lineage-infected conditions. These SNPs did not show an association under Beijing lineage-infected conditions, indicating the lineage-dependent risk of host genetic factors. Previous studies using a candidate gene approach have revealed several lineage-dependent host risk factors. To date, variants in *TLR2*,[29] *IRGM*,[30] *SLC11A1*,[31] *LAMP1*,[32] *MTOR*[32] and class I *HLA*[33] have been reported to show a correlation between human genotype and pathogen lineage. However, the sample size in those studies was limited and no study was performed to assess their reproducibility. To the best of our knowledge, this is the first report to analyze the interaction between host genome variation and pathogen genome variation at a genome-wide level and to show an association at a genome-wide significance level. The odds ratio of the most significant SNP observed in this study was 1.74, at moderate effect size. We assume that heterogeneity of the pathogen genome is one factor that is responsible for the current lack of identification of genetic risk factors for TB with moderate to large effect size. Indeed, the genome variation of the *M. tb* is six times higher compared with that of *Mycobacterium leprae*, the pathogen in leprosy.[34,35] We propose that consideration of pathogen genome information is necessary to further understand the pathogenesis of TB.

We showed that the risk of a host genetic factor can differ depending on the lineage of *M. tb*. Phenotypic differences between individual lineages have been suggested from *in vitro* experiments using clinical isolates from each lineage.[35] Beijing lineage isolates were reported to induce lower levels of the cytokines tumor necrosis factor-α, interleukin-6, interleukin-10 and chemokine ligand 1 in monocyte-derived macrophages and dendritic cells than clinical isolates from the EAI lineage, the Euro-American lineage, the CAS lineage and a reference strain H37Rv that originated from the Euro-American lineage.[36–39] As pro-inflammatory cytokines are important mediators of a protective immune response against the pathogen, these phenotypic differences can affect the risk of host genetic factors. In this respect, it is interesting that CD53 was reported to be an important regulator of innate tumor necrosis factor-α levels in genome-wide linkage analysis.[40] In this study, we simply considered the pathogen lineage information. EAI and Euro-American lineages have an intact RD105, which is a 3.5 kbp region that includes four genes (Rv0071–0074), whereas RD105 is deleted from the Beijing lineage. The functions of Rv0071–0074 in *M. tb* are currently unknown. Further genome-wide searches of pathogen genome variant(s) that

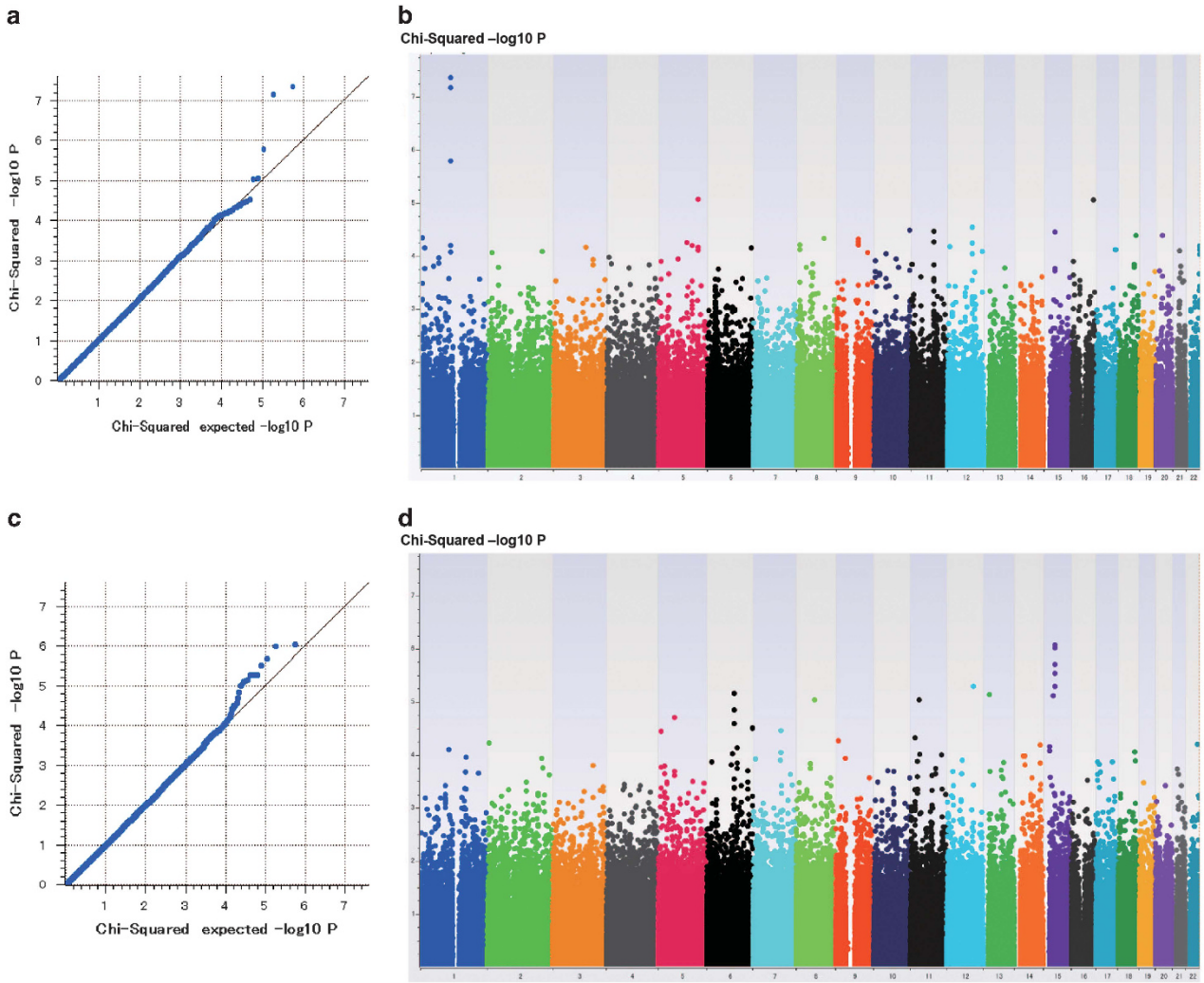

Figure 2 Pathogen lineage-based genome-wide association results. (**a**) Quantile–Quantile (QQ) plot for the comparison of old age onset and non-Beijing lineage-infected cases ($n=312$) and healthy controls ($n=771$). The genomic inflation factor lambda (IF) was 1.044. (**b**) Manhattan plot of old age onset and non-Beijing lineage-infected cases. One SNP on chromosome 1 showed genome-wide significance ($\alpha=5.00E-08$). (**c**, **d**) QQ-plot (IF = 1.028) and Manhattan plot for the comparison of old age onset and Beijing lineage-infected cases ($n=155$) and healthy controls ($n=771$). Each dot represents the −log10 (*P*-value) of each genotyped SNP.

increase the risk of identified host SNP alleles will help to identify the causative variation from the pathogen genome.

We revealed *CD53* gene as a non-Beijing lineage-dependent TB risk factor in old age onset cases, which we assumed are reactivated-TB cases. The *CD53* gene expression was increased in the active TB patients' blood compared with healthy controls and latent TB patients (Supplementary Figures S4a–d). The most significant SNP in this study, rs1418425, located at 3′ region of *CD53*, thus this SNP might affect the *CD53* gene expression through the modulation of enhancer function in this locus. The eQTL analysis in whole blood revealed that risk allele of rs1418425 can increase the endogenous expression level of *CD53*, consistent with the increased expression of *CD53* in active TB patients (Supplementary Figure S5). We investigate the regulatory effect of rs1418425 and other SNPs in LD with rs1418425 using RegulomeDB database, which includes multiple data sources to find noncoding variants that are likely to directly affect binding of transcription factors;[41] however, no functional evidence (Category 1 or 2 in RegulomeDB) was observed (Supplementary Table S4). So far, functional consequence of these SNPs need further validation. Interestingly, another significant SNP, rs1494320, was reported as the cis-eQTL of the *CD53* gene in dendritic cells infected by *M. tb*.[23] The mycobacterium-infected conditions might affect the variabilities in the inducible *CD53* expression level through the transcription factor binding which was not detected under normal condition and contribute to the difference in risk for TB onset. Several possible mechanisms for TB onset can be speculated from the reported functions of CD53, such as, (1) modulation of cytokine responses, (2) protection against oxidative stress and (3) regulation of class II HLA molecule cellular distribution. Regarding the first mechanism, knock down of the *CD53* gene increases inflammatory cytokine production by human monocyte cells,[42] and treatment of neutrophils with tumor necrosis factor-α downregulates the presence of CD53 antigens on the cell surface.[43] These reports indicate a role for CD53 in the modulation and suppression of inflammatory responses. Higher expression of CD53 under infected conditions could result in weaker inflammatory responses in leukocytes and the progression of pathogen survival. Regarding the second mechanism, CD53 has been reported to protect macrophages against oxidative stress through elevated

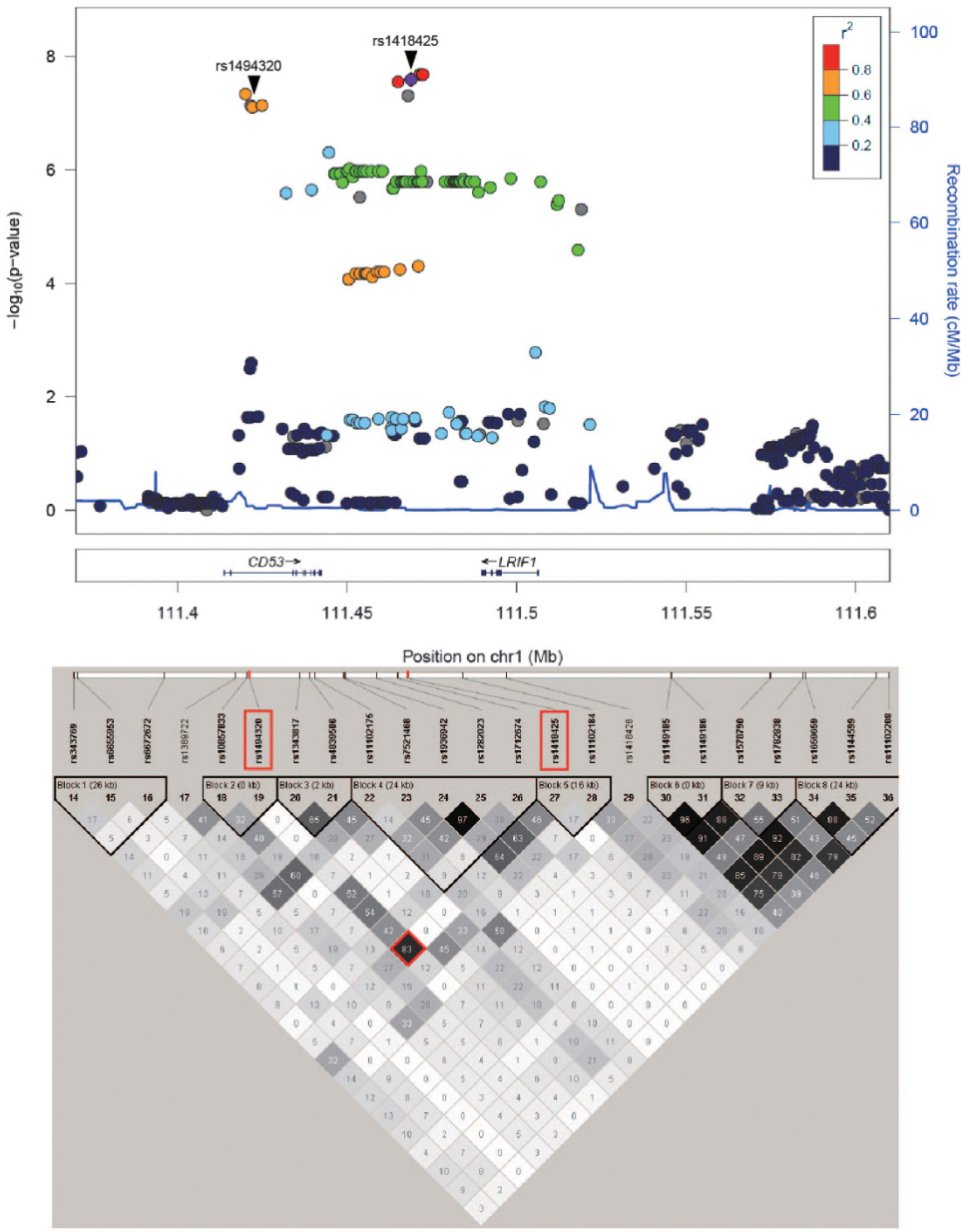

Figure 3 Plot of −log10 (*P*-value) against the physical location and linkage disequilibrium map on chromosome 1p13 locus. Each dot in the upper figure represents the −log10 value (*P*-value) of respective SNP genotyped or imputed in the old age onset and non-Beijing lineage-infected cases. Dots for rs1418425 and rs1494320 were marked by arrow head and the color for each dot represents the pairwise r^2-value against rs1418425 in 1000 genome Asian population. Lower figure represents the LD map around *CD53* and *LRIF1* in a Thai population ($n=1457$) and estimated pairwise r^2 values among 23 SNPs are shown.

intracellular levels of reduced-glutathione (GSH).[44] As reduced-GSH is important for cellular defense against apoptosis from increased oxidative stress, CD53 may contribute to the modulation of cellular apoptosis. Thus, higher expression of CD53 under infected conditions could result in stronger protection against apoptosis and further progression of necrosis, which is another form of cell death that allows *M. tb* to escape from their host cells and infect new cells. Interestingly, GSH redox is also related to human aging.[45] Measurement of GSH redox in the plasma of healthy individuals aged 19–85 showed that reduced-GSH/oxidized-GSH redox was not oxidized prior to 45 years of age and was subsequently oxidized at a nearly linear rate with aging.[46] This decreased capacity of the anti-oxidant system for GSH that occurs after 45 years of age corresponds well with our observation that the risk of *CD53* is specific for patients who are older than 45 years of age. Finally, regarding the third mechanism, CD53 has been reported to co-localize with class I and class II HLA molecules at the surface of B cells and dendritic cells and to change their subcellular localization.[24,47,48] Class II *HLA* was recently reported as a genetic risk factor for TB.[9] Thus, higher CD53 expression could result in a different cellular localization of HLA molecules and a different recognition pattern of the pathogen by HLA molecules. Whether the higher expression of *CD53* in active TB patients depends on the Non-Beijing lineage-infected condition and old age onset cases remains to be validated. Future *in vitro* experiments using monocyte/macrophage

cells with each genotype of associated SNPs and clinical isolates from each pathogen lineage will also help to determine the lineage-dependent mechanism of the *CD53* gene and its risk alleles.

We identified the non-Beijing lineage-dependent risk of rs1418425 and rs1494320, which are in mild LD in a Thai population (Figure 3). From the perspective of the host, LD around *CD53* and *LRIF1* genes found to be weak in European (CEU) or African (YRI) compared with that in Asian (Supplementary Figure S6). In this study, we observed the increased *CD53* expression in active TB patients' blood in the European and African populations (Supplementary Figures S4a–c). Considering the weak LD between rs1418425 and *CD53* region in these populations, further validation in the African and European populations is necessary to conclude whether rs1418425 is the causative variant for TB onset among different populations. From the perspective of the pathogen, the EAI lineage is distributed around East Africa and the Oceanic region, including the TB high burden countries of India, the Philippines, Vietnam and Myanmar. The Euro-American lineage is distributed around Europe, North and South America, and the North African region. In contrast, the Beijing lineage is predominant in East Asia, Central Asia, Russia and South Africa.[12] Previous genome-wide association studies in TB were conducted mainly in the African and Russian populations.[6–9] Based on the lineage-dependent risk of the *CD53* gene, we expect that future replication analysis in EAI and Euro-American lineage distributed regions will confirm the association of *CD53* with TB onset.

Previous genome-wide association studies of TB reported several risk loci at chromosome 18q11.2 and 11p13, and risk genes of *ASAP1*, class II *HLA* and *MAFB*.[6–10] We recently showed that the risk of class II *HLA* alleles is dependent on the specific strain of *M. tb* in a Thai population.[49] This finding suggests that class II *HLA* is another example of pathogen genome-dependent host genetic risk factors. Although our sample size was limited to detect a significant association, we observed that rs6071980 which we previously reported as a young age onset TB associated SNP in a Thai population showed non-Beijing lineage-dependent association (Supplementary Tables S5 and S6). Additionally, risk of rs2057178 on chromosome 11p13 and rs4331426 on chromosome 18q11.2 showed non-Beijing lineage dependency and EAI lineage dependency, respectively (Supplementary Table S5). These observations suggest that consideration of the heterogeneity of the pathogen genome is vital for identification of consistent genetic risk factors for TB among different populations.

In this study, significant and lineage-dependent association of *CD53* locus with TB onset was identified from the pathogen lineage-based GWAS. As CD53 is a modulator of inflammatory responses, and the ability of non-Beijing lineages to induce inflammatory responses differs from that of the Beijing lineage as discussed above, specific interaction between CD53 function and non-Beijing lineages seems a promising possibility. In addition to the GWAS through the division by Beijing lineage and non-Beijing lineage, we have conducted the GWAS through the division by EAI lineage- and non-EAI lineage with each age stratification (Supplementary Figures S3g–l). Although additional significantly associated loci have not been identified, we listed SNPs whose association was suggested to be lineage-dependent ($P < 1.00E-05$) and replicated in our independent data set (Supplementary Table S6). This list includes an intronic variant of *EBF1* gene, which was previously reported from a GWAS in an Indonesian population,[50] and a missense variant of the *AGER* gene, whose association with pulmonary function was reported in a previous GWAS.[51,52] Further meta-analysis using our SNP list shall identify other lineage-dependent genetic risk factors and contribute to determination of the mechanism of TB onset. More detailed subgroup analysis based on the pathogen genome variations will also help facilitate the identification of host genetic factors that are significantly associated with TB from the loci that are suggested to be associated in this study. We expect these future analyses considering the heterogeneity of the pathogen genome can provide a clue to identify the consistent genetic risk factors for TB among different populations and contribute to the effective control of TB.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all the participants in this study. This work was supported by Japan International Cooperation Agency/Japan Agency for Medical Research and Development under Science and Technology Research Partnership for Sustainable Development (SATREPS) project, Grant-in-Aid for Young Scientists (B) (JSPS KAKENHI grant number 15K19039), Grant-in-Aid for JSPS Fellows (grant number 25·10599) and Grant-in-Aid for Scientific Research (B) (grant numbers 15H05271, 24406010). Sample collection was also done by International Collaboration Research funding to the Research Institute of Tuberculosis—Japan Anti-Tuberculosis Association and Japan Science and Technology Agency-National Science and Technology Development Agency.

Author contributions: YO designed the study, coordinated the analyses and drafted the manuscript. PS, KT, NW and SM participated in design of the study and advised on entire analyses. YO, SM, EP and LT performed data QC and genotype imputation. HY, SN and SM coordinated the sample collection in Thailand. WI, SW and NW managed the human genome sample collection. YO, NSa and SM conducted the gene expression analysis. NSm, SM and PP coordinated the pathogen genome analyses. TM and MK performed SNP genotyping and QC. All authors approved the final manuscript.

1 World Health Organization. Global Tuberculosis Report 2015, WHO Press (Geneva, Switzerland) (2015).
2 Comstock, G. W. Tuberculosis in twins: a re-analysis of the Prophit survey. *Am. Rev. Respir. Dis.* **117**, 621–624 (1978).
3 Zhang, F. R., Huang, W., Chen, S. M., Sun, L. D., Liu, H., Li, Y. *et al.* Genomewide association study of leprosy. *N. Engl. J. Med.* **361**, 2609–2618 (2009).
4 Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat. Genet.* **41**, 657–665 (2009).
5 Davila, S., Wright, V. J., Khor, C. C., Sim, K. S., Binder, A., Breunis, W. B. *et al.* Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat. Genet.* **42**, 772–776 (2010).
6 Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42**, 739–741 (2010).
7 Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* **44**, 257–259 (2012).
8 Curtis, J., Luo, Y., Zenner, H. L., Cuchet-Lourenco, D., Wu, C., Lo, K. *et al.* Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet.* **47**, 523–527 (2015).
9 Sveinbjornsson, G., Gudbjartsson, D. F., Halldorsson, B. V., Kristinsson, K. G., Gottfredsson, M., Barrett, J. C. *et al.* HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet.* **48**, 318–322 (2016).
10 Mahasirimongkol, S., Yanai, H., Mushiroda, T., Promphittayarat, W., Wattanapokayakit, S., Phromjai, J. *et al.* Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *J. Hum. Genet.* **57**, 363–367 (2012).
11 American Thoracic Society/Centers for Disease Control and Prevention and the Infectious Diseases Society of America. Controlling tuberculosis in the United States. *Am. J. Respir. Criti. Care Med.* **172**, 1169–1227 (2005)
12 Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S. *et al.* Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc. Natl Acad. Sci. USA* **103**, 2869–2873 (2006).
13 Palittapongarnpim, P., Luangsook, P., Tansuphaswadikul, S., Chuchottaworn, C., Prachaktam, R. & Sathapatayavongs, B. Restriction fragment length polymorphism study of Mycobacterium tuberculosis in Thailand using IS6110 as probe. *Int. J. Tuberc. Lung Dis.* **1**, 370–376 (1997).

14 Thong-On, A., Smittipat, N., Juthayothin, T., Yanai, H., Yamada, N., Yorsangsukkamol, J. *et al.* Variable-number tandem repeats typing of Mycobacterium tuberculosis isolates with low copy numbers of IS6110 in Thailand. *Tuberculosis (Edinburgh, Scotland)* **90**, 9–15 (2010).
15 Kent, P. T. & Kubica, G. P. *Public Health Mycobacteriology: A Guide for the Level III Laboratory*, (U. S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, 1985)
16 Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
17 Smittipat, N., Juthayothin, T., Billamas, P., Jaitrong, S., Rukseree, K., Dokladda, K. *et al.* Mutations in rrs, rpsL and gidB in streptomycin-resistant Mycobacterium tuberculosis isolates from Thailand. *J. Glob. Antimicrob. Resist.* **4**, 5–10 (2016).
18 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
19 Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics (Oxford, England)* **26**, 2336–2337 (2010).
20 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)* **21**, 263–265 (2005).
21 Berry, M. P., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A., Oni, T. *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973–977 (2010).
22 Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
23 Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C. & Gilad, Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl Acad. Sci. USA* **109**, 1204–1209 (2012).
24 Escola, J. M., Kleijmeer, M. J., Stoorvogel, W., Griffith, J. M., Yoshie, O. & Geuze, H. J. Selective enrichment of tetraspan proteins on the internal vesicles of multivesicular endosomes and on exosomes secreted by human B-lymphocytes. *J. Biol. Chem.* **273**, 20121–20127 (1998).
25 Mollinedo, F., Fontan, G., Barasoain, I. & Lazo, P. A. Recurrent infectious diseases in human CD53 deficiency. *Clin. Diagn. Lab. Immunol.* **4**, 229–231 (1997).
26 Nozawa, R. S., Nagao, K., Igami, K. T., Shibata, S., Shirai, N., Nozaki, N. *et al.* Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* **20**, 566–573 (2013).
27 Grolimund, L., Aeby, E., Hamelin, R., Armand, F., Chiappe, D., Moniatte, M. *et al.* A quantitative telomeric chromatin isolation protocol identifies different telomeric states. *Nat. Commun.* **4**, 2848 (2013).
28 Brideau, N. J., Coker, H., Gendrel, A. V., Siebert, C. A., Bezstarosti, K., Demmers, J. *et al.* Independent mechanisms target SMCHD1 to trimethylated histone H3 lysine 9-modified chromatin and the inactive X chromosome. *Mol. Cell Biol.* **35**, 4053–4068 (2015).
29 Caws, M., Thwaites, G., Dunstan, S., Hawn, T. R., Lan, N. T., Thuong, N. T. *et al.* The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. *PLoS Pathog.* **4**, e1000034 (2008).
30 Intemann, C. D., Thye, T., Niemann, S., Browne, E. N., Amanua Chinbuah, M., Enimil, A. *et al.* Autophagy gene variant IRGM -261T contributes to protection from tuberculosis caused by Mycobacterium tuberculosis but not by M. africanum strains. *PLoS Pathog.* **5**, e1000577 (2009).
31 van Crevel, R., Parwati, I., Sahiratmadja, E., Marzuki, S., Ottenhoff, T. H., Netea, M. G. *et al.* Infection with Mycobacterium tuberculosis Beijing genotype strains is associated with polymorphisms in SLC11A1/NRAMP1 in Indonesian patients with tuberculosis. *J. Infect. Dis.* **200**, 1671–1674 (2009).
32 Songane, M., Kleinnijenhuis, J., Alisjahbana, B., Sahiratmadja, E., Parwati, I., Oosting, M. *et al.* Polymorphisms in autophagy genes and susceptibility to tuberculosis. *PLoS ONE* **7**, e41618 (2012).
33 Salie, M., van der Merwe, L., Moller, M., Daya, M., van der Spuy, G. D., van Helden, P. D. *et al.* Associations between human leukocyte antigen class I variants and the Mycobacterium tuberculosis subtypes causing disease. *J. Infect. Dis.* **209**, 216–223 (2014).
34 Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A. *et al.* Comparative genomic and phylogeographic analysis of Mycobacterium leprae. *Nat. Genet.* **41**, 1282–1289 (2009).
35 Coscolla, M. & Gagneux, S. Consequences of genomic diversity in Mycobacterium tuberculosis. *Semin. Immunol.* **26**, 431–444 (2014).
36 Wang, C., Peyron, P., Mestre, O., Kaplan, G., van Soolingen, D., Gao, Q. *et al.* Innate immune response to Mycobacterium tuberculosis Beijing and other genotypes. *PLoS ONE* **5**, e13594 (2010).
37 Sarkar, R., Lenders, L., Wilkinson, K. A., Wilkinson, R. J. & Nicol, M. P. Modern lineages of Mycobacterium tuberculosis exhibit lineage-specific patterns of growth and cytokine induction in human monocyte-derived macrophages. *PLoS ONE* **7**, e43170 (2012).
38 Portevin, D., Gagneux, S., Comas, I. & Young, D. Human macrophage responses to clinical isolates from the Mycobacterium tuberculosis complex discriminate between ancient and modern lineages. *PLoS Pathog.* **7**, e1001307 (2011).
39 Chen, Y. Y., Chang, J. R., Huang, W. F., Hsu, S. C., Kuo, S. C., Sun, J. R. *et al.* The pattern of cytokine production in vitro induced by ancient and modern Beijing Mycobacterium tuberculosis strains. *PLoS ONE* **9**, e94296 (2014).
40 Bos, S. D., Lakenberg, N., van der Breggen, R., Houwing-Duistermaat, J. J., Kloppenburg, M., de Craen, A. J. *et al.* A genome-wide linkage scan reveals CD53 as an important regulator of innate TNF-alpha levels. *Eur. J. Hum. Genet.* **18**, 953–959 (2010).
41 Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
42 Lee, H., Bae, S., Jang, J., Choi, B. W., Park, C. S., Park, J. S. *et al.* CD53, a suppressor of inflammatory cytokine production, is associated with population asthma risk via the functional promoter polymorphism-1560 C>T. *Biochim. Biophys. Acta* **1830**, 3011–3018 (2013).
43 Mollinedo, F., Martin-Martin, B., Gajate, C. & Lazo, P. A. Physiological activation of human neutrophils down-regulates CD53 cell surface antigen. *J. Leuk. Biol.* **63**, 699–706 (1998).
44 Kim, T. R., Yoon, J. H., Kim, Y. C., Yook, Y. H., Kim, I. G., Kim, Y. S. *et al.* LPS-induced CD53 expression: a protection mechanism against oxidative and radiation stress. *Mol. Cells* **17**, 125–131 (2004).
45 Townsend, D. M., Tew, K. D. & Tapiero, H. The importance of glutathione in human disease. *Biomed. Pharmacother.* **57**, 145–155 (2003).
46 Jones, D. P., Mody, V. C. Jr., Carlson, J. L., Lynn, M. J. & Sternberg, P. Jr. Redox analysis of human plasma allows separation of pro-oxidant events of aging from decline in antioxidant defenses. *Free Radic. Biol. Med.* **33**, 1290–1300 (2002).
47 Szollosi, J., Horejsi, V., Bene, L., Angelisova, P. & Damjanovich, S. Supramolecular complexes of MHC class I, MHC class II, CD20, and tetraspan molecules (CD53, CD81, and CD82) at the surface of a B cell line JY. *J. Immunol.* **157**, 2939–2946 (1996).
48 Engering, A. & Pieters, J. Association of distinct tetraspanins with MHC class II molecules at different subcellular locations in human immature dendritic cells. *Int. Immunol.* **13**, 127–134 (2001).
49 Toyo-Oka, L., Mahasirimongkol, S., Yanai, H., Mushiroda, T., Wattanapokayakit, S., Wichukchinda, N. *et al.* Strain-based HLA association analysis identified HLA-DRB1*09:01 associated with modern strain tuberculosis. *HLA* **90**, 149–156 (2017).
50 Png, E., Alisjahbana, B., Sahiratmadja, E., Marzuki, S., Nelwan, R., Balabanova, Y. *et al.* A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. *BMC Med. Genet.* **13**, 5 (2012).
51 Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. A., Loehr, L. R., Marciante, K. D. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat. Genet.* **42**, 45–52 (2010).
52 Repapi, E., Sayers, I., Wain, L. V., Burton, P. R., Johnson, T., Obeidat, M. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat. Genet.* **42**, 36–44 (2010).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)