Supplementary material

A genome-wide association study of nontuberculous mycobacterial pulmonary disease

Jaeyoung Cho, Kyungtaek Park, Sun Mi Choi, Jinwoo Lee, Chang-Hoon Lee, Jung-Kyu Lee, Eun Young Heo, Deog Kyeom Kim, Yeon Joo Lee, Jong Sun Park, Young-Jae Cho, Ho Il Yoon, Jae Ho Lee, Choon-Taek Lee, Nayoung Kim, Kyu Yeong Choi, Kun Ho Lee, Joohon Sung, Sungho Won, Jae-Joon Yim

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Supplementary text

Diagnosis of nontuberculous mycobacterial pulmonary disease (NTM-PD)

Diagnosis of NTM-PD was made based on the criteria of the American Thoracic Society/Infectious Diseases Society of America guidelines, which include the following: 1) pulmonary symptoms with radiological abnormalities suggestive of NTM-PD, 2) appropriate exclusion of other diagnoses, 3) positive culture results from at least two separate sputum samples or at least one bronchial wash or lavage, or lung biopsy with mycobacterial histopathologic features and positive culture for NTM.

The first quality control (QC) stage before merging the nontuberculous mycobacterial pulmonary disease (NTM-PD) and Healthy Twin Study, Korea (HTK) cohorts in the discovery data

The NTM-PD and HTK cohorts were analyzed separately before merging them to remove unknown bias. First, participants were removed where 1) genders were differently predicted from their clinical value, 2) the call rates were less than 97%, 3) the differences in the heterozygosity rates from the means were larger than three times the standard deviation, and 4) the estimates of identity-by-descent with other participants were larger than 0.185, with the highest missing rates to remove participants with first- or second-degree relationships. Next, variants were excluded whose minor allele frequencies were smaller than 1% and call rates were less than 95%. The HTK cohort went through the same process. For pooling and the steps that followed, only unambiguous and overlapped autosomal variants between the NTM-PD and HTK cohorts were used.

The first quality control (QC) stage before merging the NTM-PD and healthy cohorts in

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

the replication data

Replication data were collected in the form of CEL files. We called their genotypes and converted them to PLINK file format using Affymetrix Power Tools (v1.18) and Axiom Analysis Suite (v3.1.51). Genotypes for patients with NTM-PD and two healthy cohorts were separately called through the programs. In the calling step for the NTM-PD cohort, we removed participants whose 1) dish QC values were less than 82%, 2) call rates were less than 97%, and 3) plate's average call rates and pass rates were less than 98% and 95%, respectively. Subsequently, we removed variants whose 1) call rates were less than 95%, 2) minor allele frequencies were less than 5%, 3) Fisher's linear discriminant was less than 3.6, and 4) heterozygous strength offset was less than −0.1 and so on as default options in the program. For the healthy cohorts, we followed almost the same steps as above, however, samples were removed whose plate's average call rates were less than 99% and variants were removed whose call rates were less than 97%, and additionally, missing rates between plates were significantly different ($p < 0.00001$). Then, the two healthy cohorts were merged with only unambiguous and overlapped variants. The NTM-PD and the merged healthy cohorts went through the same QC steps as the discovery cohort. For pooling and the steps that followed, only unambiguous and overlapped autosomal variants between the NTM-PD and healthy cohorts were used.

Multidimensional scaling (MDS) plot

The MDS plot was based on the raw Hamming distances between the genotype-imputed samples using PLINK. The top two dimensions were used to represent the genetic distances among the discovery and replication cohort participants and the 1000 Genomes Phase 3 samples.

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Topologically associating domain (TAD) analysis

To identify chromatin interactions in the region that the genome-wide study identified as putatively significant, we used a publicly available web-based query tool, the 3D Genome Browser (http://promoter.bx.psu.edu/hi-c/). We checked the TAD around the putatively significant variants, and protein-coding genes belonging to the TAD were investigated. We chose human species with hg19 assembly as the reference and lung tissue with Donor-LG1-raw type as the Hi-C data source.

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Table E1 Medical comorbidities of patients with nontuberculous mycobacterial pulmonary disease

	Discovery cohort (n = 403)	Replication cohort (n = 184)
Past medical history		
Pulmonary tuberculosis	148/390 (38.0%)	77/183 (42.1%)
Measles	79/272 (29.0%)	38/150 (25.3%)
Pertussis	16/238 (6.7%)	10/134 (7.5%)
Comorbidities		
Diabetes mellitus	27/403 (6.7%)	10/184 (5.4%)
Malignancy	37/403 (9.2%)	15/184 (8.2%)
Asthma	6/403 (1.5%)	6/184 (3.3%)
COPD*	59/333 (17.7%)	27/115 (23.5%)
Bronchiectasis	383/403 (95.0%)	171/184 (92.9%)
Liver cirrhosis	3/403 (0.7%)	2/184 (1.1%)
Chronic kidney disease	2/403 (0.5%)	2/184 (1.1%)
Solid organ transplantation	1/403 (0.3%)	0/184 (0.0%)

COPD: chronic obstructive pulmonary disease.

*COPD is defined on the basis of airflow limitation (post-bronchodilator forced expiratory volume in 1 s/forced vital capacity < 0.7).

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Table E2 Genomic regions identified by combining the results from the discovery and replication cohorts using the Fisher's method

Chr	Region* (Mb)	Variant	Location (BP)	Combined p-value	Risk / Protective alleles	Risk Allele Frequency	
						Discovery (NTM-PD /Control)	Replication (NTM-PD /Control)
1	239.1-239.1	rs9428778	239093672	7.05×10^{-6}	G/A	0.10/0.06	0.12/0.07
3	–	rs12487736	47459679	9.62×10^{-6}	C/T	0.56/0.46	0.56/0.49
3	167.1-167.1	rs2699943	108875914	3.44×10^{-6}	C/T	0.07/0.03	0.11/0.06
7	**43.6-43.9**	**rs849177**	**43783788**	$\mathbf{4.92 \times 10^{-8}}$	**C/T**	**0.70/0.57**	**0.68/0.63**
8	–	rs28709344	143047788	5.95×10^{-6}	G/A	0.38/0.30	0.46/0.38
12	–	rs7301800	106481036	8.23×10^{-6}	G/A	0.60/0.57	0.77/0.67
13	–	rs9511572	25564326	7.62×10^{-6}	C/T	0.52/0.42	0.49/0.48
13	29.0-29.0	rs9508036	29015877	3.50×10^{-6}	G/A	0.42/0.39	0.46/0.35
15	25.1-25.1	rs72693656	25106554	7.77×10^{-6}	T/G	0.38/0.34	0.38/0.48
16	–	rs77953657	23825724	9.85×10^{-6}	C/T	0.83/0.73	0.84/0.76
21	–	rs45593631	45854740	2.11×10^{-6}	T/C	0.25/0.19	0.15/0.25

Chr: chromosome; BP: base pair; NTM-PD: nontuberculous mycobacterial pulmonary disease.

*Region is defined by the locations of variants with p-value $< 1.00 \times 10^{-5}$ and linkage disequilibrium $r^2 > 0.1$ with the most strongly associated variant, whose location is shown in BPs. A dash represents where only one variant exists.

Table E3 Candidate genes and tissues where gene expression levels were significantly affected by rs849177 based on the Genotype–Tissue Expression database

Gene	p-value	Tissue
STK17A	3.1×10^{-19}	Heart - Left Ventricle
	7.1×10^{-19}	Lung
	2.4×10^{-16}	Thyroid
	1.3×10^{-9}	Heart - Atrial Appendage
	1.2×10^{-5}	Skin - Not Sun Exposed
	1.4×10^{-5}	Skin - Sun Exposed
	2.0×10^{-5}	Artery - Aorta
COA1	6.8×10^{-20}	Skin - Sun Exposed
	5.8×10^{-19}	Artery - Tibial
	4.8×10^{-18}	Skin - Not Sun Exposed
	6.1×10^{-15}	Oesophagus - Muscularis
	1.6×10^{-14}	Adipose - Subcutaneous
	7.2×10^{-14}	Nerve - Tibial
	2.5×10^{-13}	Artery - Aorta
	4.0×10^{-12}	Oesophagus - Mucosa
	4.8×10^{-12}	Thyroid
	2.5×10^{-10}	Pancreas
	4.4×10^{-10}	Colon - Sigmoid
	6.4×10^{-10}	Colon - Transverse
	5.0×10^{-6}	Oesophagus - Gastroesophageal Junction
	5.9×10^{-6}	Artery - Coronary
	7.4×10^{-6}	Adipose - Visceral
	3.2×10^{-5}	Muscle - Skeletal
	4.5×10^{-5}	Cells - Transformed fibroblasts
BLVRA	7.7×10^{-10}	Whole Blood
	1.7×10^{-8}	Oesophagus - Mucosa
	1.1×10^{-6}	Artery - Tibial
POLR2J4	3.3×10^{-6}	Artery - Aorta
AC004951.5	1.1×10^{-8}	Oesophagus - Mucosa
AC004985.12	1.4×10^{-6}	Oesophagus - Mucosa

Table E4 Gene Ontology (GO) terms related to apoptotic processes that were significant in both datasets of GSE72821 and E-MTAB-1101

GO biological process	GSE72821		E-MTAB-1101	
	p-value	FDR	p-value	FDR
Regulation of apoptotic process (GO:0042981)	1.25×10^{-17}	2.81×10^{-15}	1.28×10^{-9}	3.93×10^{-7}
Positive regulation of apoptotic process (GO:0043065)	8.80×10^{-14}	1.20×10^{-11}	3.87×10^{-6}	2.88×10^{-4}
Regulation of apoptotic signaling pathway (GO:2001233)	2.20×10^{-11}	2.24×10^{-9}	7.65×10^{-8}	9.90×10^{-6}
Negative regulation of apoptotic process (GO:0043066)	1.55×10^{-10}	1.44×10^{-8}	1.02×10^{-9}	3.31×10^{-7}
Apoptotic process (GO:0006915)	1.27×10^{-9}	1.05×10^{-7}	2.98×10^{-8}	4.52×10^{-6}
Regulation of extrinsic apoptotic signaling pathway (GO:2001236)	1.31×10^{-7}	7.82×10^{-6}	3.55×10^{-7}	3.65×10^{-5}
Regulation of intrinsic apoptotic signaling pathway (GO:2001242)	1.64×10^{-6}	7.89×10^{-5}	3.42×10^{-4}	1.23×10^{-2}
Apoptotic signaling pathway (GO:0097190)	2.26×10^{-6}	1.06×10^{-4}	1.17×10^{-3}	3.24×10^{-2}
Regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0043281)	2.56×10^{-6}	1.18×10^{-4}	3.89×10^{-5}	2.15×10^{-3}
Negative regulation of apoptotic signaling pathway (GO:2001234)	4.26×10^{-6}	1.86×10^{-4}	9.56×10^{-8}	1.19×10^{-5}
Extrinsic apoptotic signaling pathway (GO:0097191)	4.32×10^{-5}	1.44×10^{-3}	2.60×10^{-4}	9.86×10^{-3}
Negative regulation of extrinsic apoptotic signaling pathway (GO:2001237)	1.29×10^{-4}	3.79×10^{-3}	3.39×10^{-5}	1.92×10^{-3}
Negative regulation of intrinsic apoptotic signaling pathway (GO:2001243)	2.11×10^{-4}	5.87×10^{-3}	2.34×10^{-5}	1.41×10^{-3}
Negative regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO:0043154)	2.65×10^{-4}	7.14×10^{-3}	1.45×10^{-4}	6.32×10^{-3}

FDR: false discovery rate.

Table E5 Common variant and gene set analyses of 43 genes previously reported to be associated with nontuberculous mycobacterial pulmonary disease

Gene	Common variant analysis			Gene set analysis	
	Locus of the most significant variant (hg19)	p-value	FWER	p-value	FDR
RCOR3	**chr1:211363628**	**5.66×10^{-5}**	**0.030**	0.684	0.865
NFATC2	chr20:51549716	1.53×10^{-4}	0.093	**7.47×10^{-4}**	**0.032**
SLC29A1	chr6:44212688	2.88×10^{-4}	0.110	0.072	0.621
ISG15	chr1:920503	4.93×10^{-4}	0.164	0.022	0.481
IL12RB1	chr19:17967816	1.03×10^{-3}	0.445	0.250	0.718
ANKRD6	chr6:89624219	1.07×10^{-3}	0.628	0.052	0.564
TARP	chr7:38332854	1.43×10^{-3}	0.896	0.161	0.709
IFNGR2	chr21:33437648	1.83×10^{-3}	0.938	0.115	0.709
MUC12	chr7:101046298	2.65×10^{-3}	1	0.388	0.782
SLC11A1	chr2:218438843	4.88×10^{-3}	1	0.439	0.782
SAMD3	chr6:130310686	5.11×10^{-3}	1	1	1
CRTAM	chr11:122811024	7.94×10^{-3}	1	0.477	0.782
MAP2K4	chr17:12195571	8.22×10^{-3}	1	0.477	0.782
TTK	chr6:80117150	9.52×10^{-3}	1	0.731	0.899
CFTR	chr7:117550950	0.0101	1	0.127	0.709
TPBG	chr6:82408766	0.0120	1	0.315	0.753
NELL2	chr12:44966063	0.0120	1	1	1
PSPH	chr7:56096567	0.0134	1	0.450	0.782
KRT83	chr12:52257245	0.0138	1	0.052	0.564
IRF8	chr16:85967604	0.0158	1	0.545	0.782
IL2RB	chr22:37143875	0.0206	1	0.569	0.789
GATA2	chr3:128419054	0.0212	1	0.232	0.713
LDHB	chr12:21536230	0.0214	1	0.168	0.709
GZMK	chr5:55127417	0.0219	1	0.546	0.782
AK5	chr1:77312414	0.0221	1	0.657	0.857
IFNLR1	chr1:24054800	0.0222	1	0.861	1
IFNG	chr12:68249467	0.0230	1	0.185	0.709
STAT1	chr2:190968863	0.0239	1	0.601	0.808
TIGIT	chr3:114233728	0.0285	1	0.502	0.782
PMS2P1	chr7:100347079	0.0296	1	0.153	0.709
ORC3	chr6:87621339	0.0298	1	0.220	0.713
XCL1	chr1:168647094	0.0331	1	0.198	0.709
FLJ45825	chr6:37576163	0.0337	1	1	1
XCL2	chr1:168624717	0.0430	1	0.285	0.753
IFNGR1	chr6:137253162	0.0478	1	1	1
PZP	chr12:9172646	0.0631	1	0.442	0.782
A2M	chr12:9172646	0.0631	1	0.514	0.782
FAHD2A	chr2:95485567	0.0653	1	0.305	0.753

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

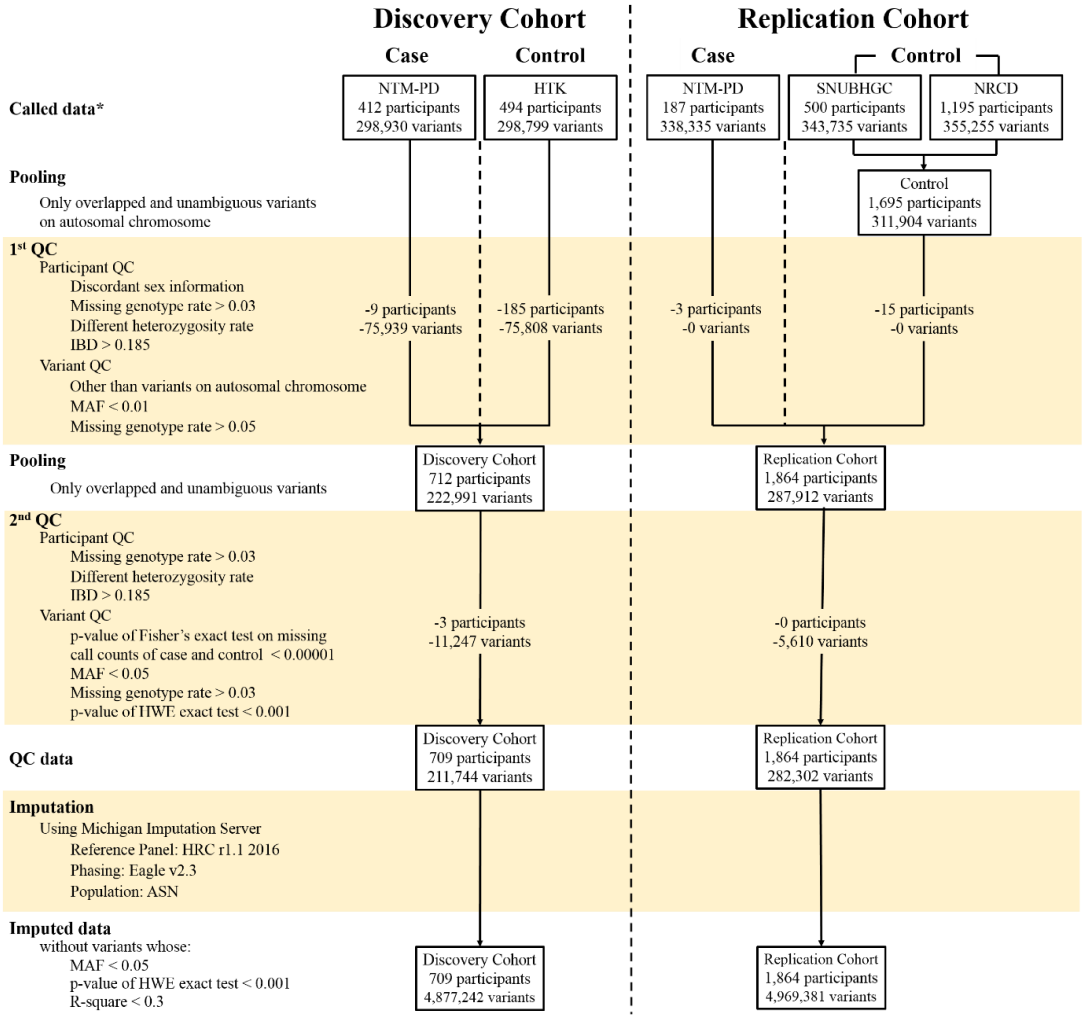
PPIH	chr1:42576537	0.0814	1	0.397	0.782
FCRL3	chr1:157694755	0.1410	1	1	1
MPEG1	chr11:59249666	0.1443	1	0.390	0.782
MST1R	chr3:49979727	0.1536	1	0.763	0.912
GUSBP14	chr5:70309998	0.3003	1	1	1

AK5: adenylate kinase 5; *A2M*: alpha-2-macroglobulin; *ANKRD6*: ankyrin repeat domain 6; *CFTR*: cystic fibrosis transmembrane conductance regulator; *CRTAM*: cytotoxic and regulatory T cell molecule; *FAHD2A*: fumarylacetoacetate hydrolase domain containing 2A; *FCRL3*: Fc receptor like 3; FDR: false discovery rate; *FLJ45825*: uncharacterized LOC100505530; FWER: family-wise error rate; *GATA2*: GATA binding protein 2; *GUSBP14*: GUSB pseudogene 14; *GZMK*: granzyme K; *IFNG*: interferon gamma; *IFNGR1*: interferon gamma receptor 1; *IFNGR2*: interferon gamma receptor 2; *IFNLR1*: interferon lambda receptor 1; *IL2RB*: interleukin 2 receptor subunit beta; *IL12RB1*: interleukin 12 receptor subunit beta 1; *IRF8*: interferon regulatory factor 8; *ISG15*: ISG15 ubiquitin like modifier; *KRT83*: keratin 83; *LDHB*: lactate dehydrogenase B; *MAP2K4*: mitogen-activated protein kinase kinase 4; *MPEG1*: macrophage expressed 1; *MST1R*: macrophage stimulating 1 receptor; *MUC12*: mucin 12, cell surface associated; *NELL2*: neural EGFL like 2; *NFATC2*: nuclear factor of activated T cells 2; *ORC3*: origin recognition complex subunit 3; *PMS2P1*: postmeiotic segregation increased 2 pseudogene 1; *PPIH*: peptidylprolyl isomerase H; *PSPH*: phosphoserine phosphatase; *PZP*: PZP alpha-2-macroglobulin like; *TARP*: TCR gamma alternate reading frame protein; *TIGIT*: T cell immunoreceptor with Ig and ITIM domains; *TPBG*: trophoblast glycoprotein; *TTK*: TTK protein kinase; *RCOR3*: REST corepressor 3; *SAMD3*: sterile alpha motif domain containing 3; *SLC11A1*: solute carrier family 11 member 1; *SLC29A1*: solute carrier family 29 member 1; *STAT1*: signal transducer and activator of transcription 1; *XCL1*: X-C motif chemokine ligand 1; *XCL2*: X-C motif chemokine ligand 2.

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Figure E1 Workflow of two stages of quality control (QC). The same QC criteria were applied to both discovery and replication cohorts. The different heterozygosity rates in the participant QC steps represent the removal of participants whose differences in the heterozygosity rates from the means were larger than three (or four, in the last QC steps for the replication cohort) times the standard deviation. *Calling steps in the replication cohort included a step to remove rare variants (MAF < 5%). In this step, 459,607, 450,525 and 425,869 variants were removed from the NTM-PD, SNUBHGC and NRCD cohorts, respectively. Called data only included variants passing the QC steps for genotype calling. HTK: Healthy Twin Study, Korea; HWE: Hardy–Weinberg equilibrium; IBD: identity-by-descent; MAF: minor allele frequency; NRCD: National Research Center for Dementia; NTM-PD: nontuberculous mycobacterial pulmonary disease; SNUBHGC: Seoul National University Bundang Hospital Gastric Cancer; QC: quality control.

Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430



Cho J, *et al. Thorax* 2021; 76:169–177. doi: 10.1136/thoraxjnl-2019-214430

Figure E2 Illustrative diagram showing Mendelian randomization assumptions. Assumption ①: the genetic variants are independent of unknown confounders, assumption ②: the genetic variants are associated with the expression level of a candidate gene, assumption ③: the genetic variants are independent of the risk of nontuberculous mycobacterial pulmonary disease (NTM-PD) conditional on unknown confounders and the expression level of a candidate gene. The coefficient β represents the causal effect of the expression level of a candidate gene on the risk of NTM-PD. Potential violations of the assumptions are shown by dotted lines.

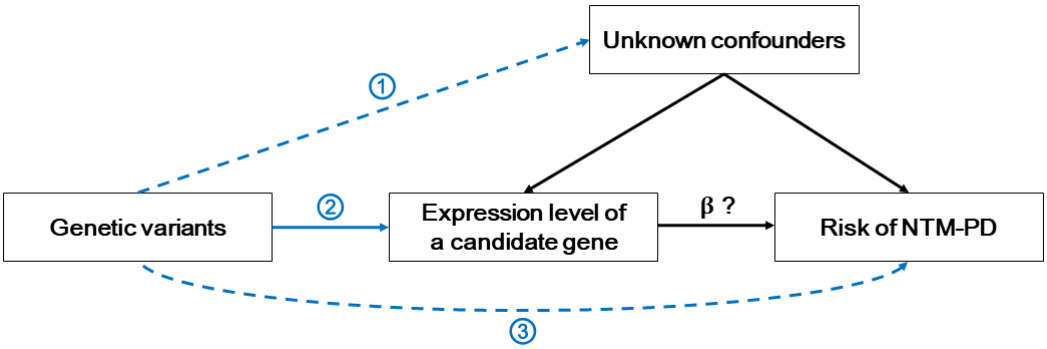


Figure E3 Multidimensional scaling plots with top two dimensions in (A) the discovery, replication cohorts and 1000 Genomes Phase 3 samples and (B) the discovery and replication cohorts. AFR: African; AMR: ad mixed American; EAS: East Asian; EUR: European; NTM-PD: nontuberculous mycobacterial pulmonary disease; SAS: South Asian.

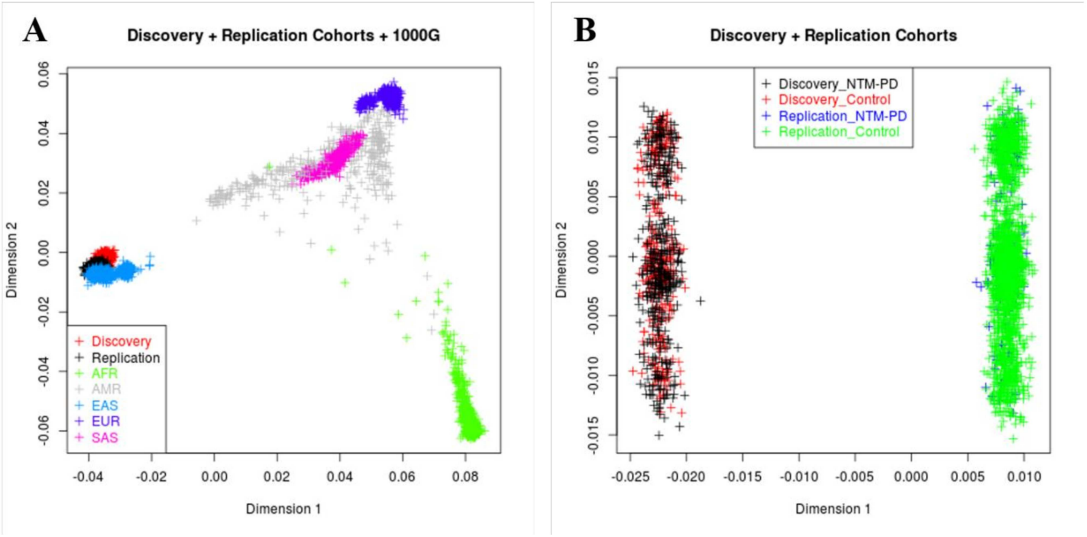


Figure E4 Logistic analysis without adjusting for body mass index in the discovery cohort. (A) Manhattan plot. Red and blue lines indicate significance levels of 1.00×10^{-7} and 1.00×10^{-5}, respectively. (B) Quantile-quantile plot. Confidence intervals of the estimated p-values are 0.95 and are colored in grey. BMI: body mass index; GIF: genomic inflation factor.

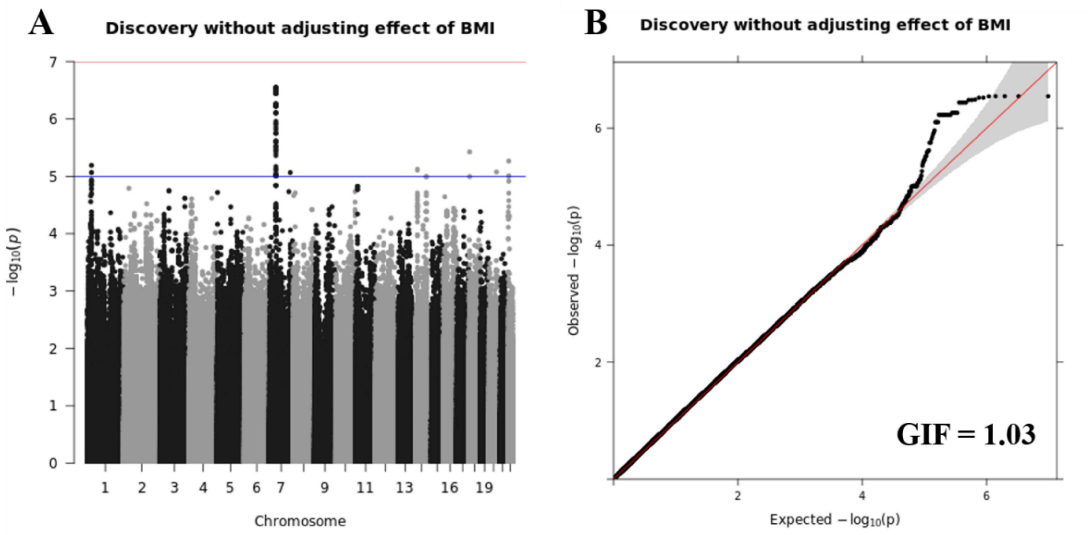


Figure E5 Conditional analysis on rs849177 in the discovery cohort. (A) Manhattan plot. The blue line indicates the significance level of 1.00×10^{-5}. (B) Quantile-quantile plot. Confidence intervals of the estimated p-values are 0.95 and are colored in grey. GIF: genomic inflation factor.

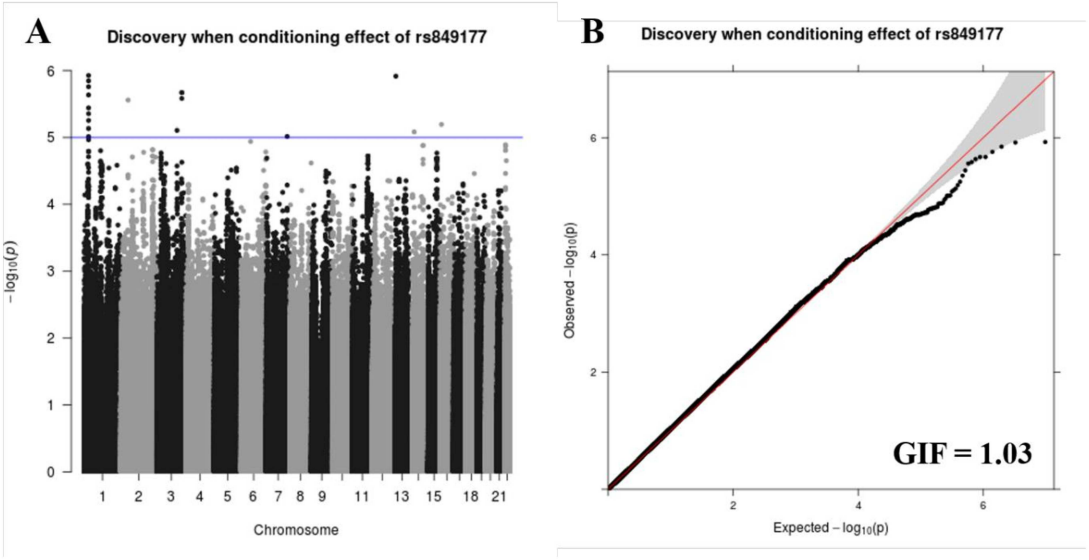


Figure E6 Topologically associating domain (TAD) near the locus of rs849177. Several genes, including *STK17A, COA1 and BLVRA*, are in the same TAD (blue block) with rs849177 (in the black dotted circle).