이학석사 학위논문

# 크론병과 나병의 유전적 구조 비교 연구

## Comparison of genetic architecture between Crohn's disease and leprosy using meta-analysis data

울 산 대 학 교 대 학 원

의 과 학 과

박 도 훈

# 크론병과 나병의 유전적 구조 비교 연구

지 도 교 수　　송　규　영

이 논문을 이학석사 학위 논문으로 제출함

2022 년　2 월

울 산 대 학 교 대 학 원

의　　과　　학　　과

박　도　훈

박도훈의 이학석사학위 논문을 인준함

심사위원 예 병 덕 ( 인)

심사위원 송 규 영 ( 인)

심사위원 이 호 수 ( 인)

울 산 대 학 교 대 학 원

2022 년 2 월

# ABSTRACT

**Background**

Genome-wide association studies (GWAS) of Crohn's disease (CD) in European and leprosy in Chinese population suggested that CD and leprosy might share genetic risk loci related to nonspecific innate immunity. Pleiotropic variants within these loci showed opposite allelic effects between CD and leprosy.

**Methods**

Using CD meta-analysis of 2,354 CD patients and 4,907 healthy controls in Korean and leprosy meta-analysis of 2,960 leprosy patients and 3,747 healthy controls in Chinese, we compared the genetic architecture of CD and leprosy using linkage disequilibrium score regression analysis (LDSC) and polygenic risk score (PRS) analysis.

**Results**

Most of the shared loci between CD and leprosy showed an opposite allelic effect except the *RIPK2* and *LACC1* loci. Investigation of the genetic correlation using cross-trait LDSC showed a significant negative genetic correlation between CD and leprosy ($r_g$[SE] = -0.30 [0.12], $p$ = 1.5 x $10^{-2}$). Phenotype variance explained by the polygenic risk score derived from Chinese leprosy data explained up to 5.27 % of variance of Korean CD. When the directions of the effects of the *LACC1* and *RIPK2* loci were flipped to account for the original directions between CD and leprosy, the explained variance increased up to 8.21 %. After removing both the *MHC* and *TNFSF15* regions, the most significant signals in CD, genetic correlation and overlap between the two diseases were decreased.

**Conclusions**

Our study represented the first systemic effort to compare the genetic basis of CD and leprosy in samples of East Asian origin. Our findings showed that CD and leprosy shared a substantial number of genetic susceptibility loci in East Asians with risk allele effects in the opposite directions. In addition, out data suggest that the most significant CD susceptibility loci, *MHC* and *TNFSF15*, may be the main driver of higher overlap between CD and leprosy.



**Key words:** Crohn's disease; leprosy; polygenic risk scores; genetic correlation

# CONTENTS

# LIST OF TABLES



# LIST OF FUGURES

# ABBREVIATIONS

| | |
|---|---|
| CD | Crohn's disease |
| IBD | Inflammatory bowel disease |
| UC | Ulcerative colitis |
| GWAS | Genome-wide association study |
| LD | Linkage disequilibrium |
| MHC | Major histocompatibility complex |
| OR | Odds ratio |
| SNP | Single nucleotide polymorphism |
| PRS | Polygenic risk score |

# 1. INTRODUCTION

Genome-wide association studies (GWAS) have suggested that Crohn's disease (CD) and leprosy might share a common underlying genetic susceptibility.[1-6] By simple cross-comparison of genome-wide significant loci between leprosy GWAS of Chinese population and CD GWAS of European population[7-8] followed by replication of European CD or ulcerative colitis (UC) susceptibility loci in leprosy samples of Chinese origin, 10 susceptibility loci were identified to be shared between CD and leprosy (*IL23R*, *IL18RAP*, *IL12B*, *RIPK2*, *TNFSF15*, *ZNF365-EGR2*, *CCDC88B*, *LACC1*, *NOD2* and *IL27*). These genes were related to immune response, NOD signaling and immune response regulation, and suggested that common immunologic features could be involved in both CD and leprosy. Furthermore, most of shared risk loci showed the opposite allelic effects. However, there have not been genome-wide scale comparisons between CD and leprosy susceptibilities in Asian population.

Estimating genetic correlation is a key step toward understanding the shared genetic architecture between complex traits and disease. The genetic correlation parameter describes how genome-wide genetic effects align between two complex phenotypes. To estimate genetic correlation using GWAS data, there are two widely used approaches- when only GWAS summary statistic data are available, linkage disequilibrium score regression (LDSC)[9]; when individual data available, genetic correlation is commonly estimated by restricted maximum likelihood (REML).[10]

Recently, Polygenic risk scores (PRS) have attracted increasing interest from the clinical community for their predictive value for multiple common diseases.[11-13] A PRS estimates an individuals' genetic liability to disease based on genotype profile and relevant GWAS data. PRSs are calculated by summing risk alleles, which are weighted by effect sizes derived from GWAS results.[14] PRS can be used to compare the genetic architecture of related traits. To investigate genetic relationship between CD and leprosy, genetic overlap between CD and leprosy meta-analysis data was estimated using two different approaches: LDSC and PRS analyses.

# 2. MATERIALS AND METHODS

## 2.1. Korean IBD datasets

The three IBD GWAS datasets used in this study include previously reported Korean GWAS (cohort I : 1,469 IBD cases (896 CD and 573 UC) and 4,041 controls),[15] Asian screening array (cohort II : 1,726 IBD cases (725 CD and 1,001 UC) and 378 controls),[16] and Immunochip

v2.0 (cohort III : 1,334 IBD cases (738 CD and 601 UC) and 488 controls).[17] In total, 4,529 IBD patients (2,359 CD and 2,175 UC) and 4,907 healthy controls were included in IBD datasets (Table 1). Patients' clinical characteristics of CD are summarized in Table 2. All IBD patients were recruited from IBD Clinic of Asan Medical center.

### 2.2. Chinese Leprosy datasets

We used summary statistics of 3 previously published GWAS including cohort IV (706 cases and 1,225 controls),[1] cohort V (842 cases and 925 controls)[4] and cohort VI (1,412 cases and 1,597 controls),[5] genotyped using Human610-Quad BeadChip, Omni Zhonghua chips, and Human 660K-Quad BeadChip, respectively. In total, 2,960 leprosy patients and 3,747 healthy controls were included in leprosy datasets (Table 1).

**Table 1. Study cohorts.**

| Disease | Ethnicity | Cohort | Study | Platform(Illumina) | No. of SNPs | No. of samples | |
|---|---|---|---|---|---|---|---|
| | | | | | | Case (CD/ UC) | Control |
| CD / UC | Korean | Cohort I | GWAS | Omni1-Quad arrays | 6,610,963 | 896/573 | 4,041 |
| | | Cohort II | ASA | Asain Screening arrays | 6,597,252 | 725/1,001 | 378 |
| | | Cohort III | Immunochip v 2.0 | Infinium ImmunoArray-24 v2.0 BeadChip | 2,869,283 | 738/601 | 488 |
| Leprosy | Chinese | Cohort Ⅳ | GWAS | Human610-Quad BeadChip | 7,164,077 | 706 | 1,225 |
| | | Cohort Ⅴ | | Human 660K-Quad BeadChips | 7,808,757 | 842 | 925 |
| | | Cohort Ⅵ | | Omni Zhonghua chips | 6,413,152 | 1,412 | 1,597 |

GWAS, genome-wide association study; SNP, single nucleotide polymorphism.

**Table 2. Clinical characteristics of Crohn's disease subjects in Koreans.**

| | Cohort I | | Cohort II | | Cohort III | | Cohort I, II and III | |
|---|---|---|---|---|---|---|---|---|
| | CD | Control | CD | Control | CD | Control | CD | Control |
| No. of samples | 896 | 4,041 | 725 | 378 | 738 | 488 | 2,359 | 4,907 |
| Male (%) | 633 (70.6) | 1,602 (39.6) | 561 (77.4) | 190 (50.3) | 551 (74.7) | 259 (53.1) | 1,745 (74.0) | 2,051 (41.8) |
| Mean age at sampling (yr) | 25.5 ± 9.1 | NA | 27.6 ± 9.2 | NA | 28.4 ± 9.6 | NA | 27.1 ± 9.4 | NA |
| Mean age at diagnosis (yr) | 22.3 ± 8.2 | | 24.2 ± 8.8 | | 24.8 ± 8.8 | | 23.7 ± 8.6 | |
| Age group at diagnosis (%) | | | | | | | | |
| ≤ 16 | 237 (26.5) | NA | 104 (14.6) | | 73 (9.9) | NA | 414 (17.6) | NA |
| 17~40 | 621 (69.3) | NA | 566 (79.5) | | 612 (82.9) | NA | 1,799 (76.7) | NA |
| ≥ 40 | 38 (4.2) | NA | 42 (5.9) | | 53 (7.2) | NA | 133 (5.7) | NA |
| NA | | | 13 | | | | 13 | |
| Location, no. (%) | | | | | | | | |
| Ileum | 158 (18.0) | | 106 (20.4) | | 190 (25.9) | | 454 (21.3) | |
| Colon | 48 (5.5) | | 28 (5.4) | | 20 (2.7) | | 96 (4.5) | |
| Ileocolon | 674 (76.6) | | 385 (74.2) | | 525 (71.4) | | 1,584 (74.2) | |
| NA | 16 | | 206 | | 3 | | 225 | |
| Behavior, no. (%) | | | | | | | | |
| Inflammatory | 343 (39.1) | | 267 (49.1) | | 345 (47.0) | | 955 (44.3) | |
| Stricturing | 173 (19.7) | | 98 (18.0) | | 122 (16.6) | | 393 (18.2) | |
| Penetrating | 362 (41.2) | | 179 (32.9) | | 267 (36.4) | | 808 (37.5) | |
| NA | 18 | | 181 | | 4 | | 203 | |
| Perianal fistula, no. (%) | | | | | | | | |
| No | 325 (38.5) | | 264 (38.1) | | 392 (53.2) | | 981 (43.1) | |
| Yes | 519 (61.5) | | 429 (61.9) | | 345 (46.8) | | 1,293 (56.9) | |
| NA | 52 | | 32 | | 1 | | 85 | |

CD, Crohn's disease

### 2.3 Fixed-effects meta-analysis

We performed disease-specific meta-analyses of CD and leprosy, respectively, using summary statistics of GWAS datasets calculated using SNPTEST[18] based on the additive model of frequentist association test. A meta-analysis was carried out using META v1.7[19] software based on the fixed effects model. Summary statistics of 3 datasets were utilized for the CD meta-analysis, and 3 independent GWAS datasets for leprosy meta-analysis. As a part of quality control, we excluded SNPs with significant heterogeneity *p*-value ($p < 0.05$). Finally, the number of available SNPs was 2,592,453 SNPs for CD meta-analysis and 5,539,402 SNPs for Chinese leprosy meta-analysis. A total of 2,289,680 SNPs overlapped between CD and leprosy datasets.

### 2.4 Shared genetic background analysis

#### 2.4.1 Genetic correlation

To examine the shared genetic architecture of CD and leprosy, we estimated the genetic correlation using LD score regression (LDSC) v1.0.0.[9] Summary statistics of CD (cohort I, II and III) and leprosy (cohort IV, V and VI) meta-analysis with 2,289,680 overlapping SNPs were used as input data. To further examine effect of two largest effect size of CD GWAS, we removed *MHC* (chromosome 6: 25 ~ 34 Mb, hg19) or *TNFSF15* (chromosome 9: 117.4 ~ 118.7 Mb, hg19) region. For LD reference panel, the East Asian data (JPT + CHB) from the 1000 Genomes Project was used.[20]

#### 2.4.2 Polygenic risk score analysis

We performed polygenic risk score (PRS) analysis using PRSice-2.[14] To estimate how much variance of CD are explained by PRS derived from the leprosy GWAS ($PRS_{leprosy}$), $PRS_{leprosy}$ was calculated by summing the risk alleles associated with CD weighted by the effect size estimated by a meta-analysis of leprosy GWAS. To achieve the largest sample size possible, we used the leprosy meta-analysis of cohort IV, V, and VI as the base data for estimating risk allele effect size, and the CD meta-analysis of cohort I, II, and III as the target data for estimating PRS. To minimize overfitting due to tight LD in the *MHC* region (chromosome 6: 25 ~ 34 Mb), we selected only the most significant variant (rs9271011) in the *MHC* region. First, Eight lead SNPs reaching the genome-wide significance threshold ($p < 5 \times 10^{-8}$) in the meta-analysis of leprosy were included in the calculation of PRS. Then, to maximize the variance explained by $PRS_{leprosy}$ for CD, we manually flipped the direction of effect of two susceptibility loci (*LACC1* and *RIPK2* loci), which

showed directionally consistent associations between CD and leprosy in East Asians. The clumping was set to create clumps of SNPs spanning 250 kb in LD with an $r^2$ threshold greater than 0.1 using the East Asian (CHB + JPT) 1000 Genomes data as a reference panel to calculate LD. We also calculated PRS after removing the *TNFSF15* (chromosome 9: 117.4 ~ 118.7 Mb, hg19) or *MHC* (chromosome 6: 25 ~ 34 Mb, hg19) region which showed the largest effect size in CD GWAS. To compare the variance explained by $PRS_{leprosy}$ among the clinical phenotypes of CD with respect to disease location, we calculated PRS after stratifying CD samples by disease location, colonic, ileocolonic, and ileal CD (Table 2). We then compared the full model (including the PRS) with null model (with the PRS variance excluded) and estimated the variance explained using Nagelkerke's pseudo-$R^2$.

### 2.5 Protein-protein interaction (PPI) network and pathway enrichment analyses

To construct the PPI network of target genes in the shared susceptibility loci between CD and leprosy, we used the Search Tool for the Retrieval of Interacting Genes/Proteins database (STRING v11.5).[21] Using a list of proteins as input data, the STRING predicts functional interactions of the proteins based on co-expression, text-mining, biochemical data from experiments, previously curated pathway, and protein-complex knowledge from datasets. We applied default setting of full STRING network type, medium confidence of interaction score (0.4), and false discovery rate (FDR) < 0.05. The STRING also performed a pathway enrichment analysis based on the background gene list of Gene Ontology resource[22] to identify enriched biological pathway associated with the proteins in the PPI network (FDR < 0.05).

# 3. RESULTS

### 3.1 Meta-analysis of CD and leprosy

A disease-specific fixed-effects meta-analysis of CD and leprosy datasets showed 13 and eight loci at genome-wide significance (Table 3 and 4), respectively. Of those known loci, three loci (*MHC*, *TNFSF15,* and *IL12B*) showed genome-wide significant association in both CD and leprosy datasets. Of lead SNPs within the three loci in CD and leprosy, the SNPs within the *MHC* region are in moderate LD whereas the SNPs within the *TNFSF15* and *IL12B* loci are in high LD (*MHC*: $r^2 = 0.60$; *TNFSF15*: $r^2 = 0.94$; *IL12B*: $r^2 = 0.74$, the East Asian LD reference (JPT + CHB) of the 1000 genomes). Most of the susceptibility loci shared between CD and leprosy showed the opposite genetic effect except *RIPK2* and *LACC1* loci.

### 3.2 Genetic correlation between CD and leprosy

To explore the shared genetic architecture of CD and leprosy, we estimated the genetic correlation and *p* values using LDSC[9] with shared 2,289,680 SNPs between the meta-analysis summary statistics of CD and leprosy datasets. It showed a negative correlation ($r_g$[SE] = -0.30 [0.12]) with statistically significant *p* value of $1.5 \times 10^{-2}$. To further examine this negative correlation between CD and leprosy, we re-estimated genetic correlation between CD and leprosy after removing the largest-effect loci in CD, *TNFSF15* (chromosome 9: 117.4~118.7 Mb, hg19) or *MHC* (chromosome 6: 25 ~ 34 Mb, hg19). The correlation and significance of *p*-value was decreased (excluding both *MHC* and *TNFSF15*: $r_g$[SE] = -0.24 [0.15], *p* -value : $1.03 \times 10^{-1}$) (Table 5).

### 3.3 Estimation of variance explained by polygenic risk scores

We examined the extent of the overlap of the genetic architecture between CD and leprosy by estimating the variance of CD explained by the polygenic risk scores (PRS) derived from leprosy PRS ($PRS_{leprosy}$). First, we calculated variance explained in CD by the $PRS_{leprosy}$ with eight genome-wide significant variants using leprosy GWAS effect sizes (Table 4). $PRS_{leprosy}$ explained up to 4.27 % of phenotype variance of CD (Table 6).

As most of the susceptibility loci shared between CD and leprosy have the opposite genetic effect except *RIPK2* and *LACC1* loci, we manually flipped those two loci and calculated variance explained. $PRS_{leprosy}$ explained up to 8.21% of CD phenotype. Using PRSice-2, bar plot of PRSice-2 for leprosy explaining CD showed that the best *p*-value threshold was $1.71 \times 10^{-5}$ (Figure 1). As

PRS value of a sample should approximate the normal (Gaussian) distribution, we checked whether the normal distribution is satisfied. At the best *p*-value threshold, the skewed distribution of PRS was observed, probably due to mixture of normal distributions. Indeed, PRS showed normal distributions when the samples were stratified based on genotypes of the most significant signal (rs9271011) at the *MHC* locus (Anderson-Darling normality *p*-value: AA: 0.510, AG: 0.365, GG: 0.063) (Figure 2). Mean value of $PRS_{leprosy}$ based on stratified genotype of rs9271011 was decreased when the number of the CD risk allele G was increased (Figure 2). The high resolution best-fit $PRS_{leporsy}$ explained 5.28 % variance of CD and was based on 39 SNPs (Table 6).

To further examine this substantial overlap between CD and leprosy, we re-calculated the variance explained by $PRS_{leprosy}$ after removing the largest-effect loci in CD, *TNFSF15* (chromosome 9: 117.4 ~ 118.7 Mb, hg19) or *MHC* (chromosome 6: 25 ~ 34 Mb, hg19). The variance explained by $PRS_{leprosy}$ was decreased sharply (excluding *TNFSF15*: 2.76%, excluding *MHC*: 2.32%, excluding both *MHC* and *TNFSF15*: 0.71%) (Table 7). In addition, we stratified CD samples by disease location, and compared difference of overlap of the genetic architecture among clinical phenotypes. The variance explained by $PRS_{leprosy}$ was highest in ileocolonic CD and lowest in colonic CD (colonic CD: 1.27 %, ileocolonic CD: 4.40 %, ileal CD: 1.27 %) (Table 8). When the *RIPK2* and *LACC1* loci were manually flipped, $PRS_{leprosy}$ explained up to 8.35% of ileocolonic CD phenotype. To minimize the difference of overlap of genetic architecture among CD clinical phenotype due to sample size, we re-calculated the variance explained by $PRS_{leprosy}$ after random sampling. The variance explained by $PRS_{leprosy}$ was decreased (ileocolonic CD: 2.72 %, ileal CD: 2.07 %) (Table 9).

We also examined the overlap of the genetic architecture between UC and leprosy, and compared the explained variance of $PRS_{leprosy}$ for CD and UC. At genome-wide significant *p*-value threshold, variance explained by $PRS_{leprosy}$ for CD is far better than that for UC (4.27 % vs 0.03 %) (Table 6 and 10).

### 3.4 Protein-protein interaction and pathway enrichment analyses

Previously, 10 susceptibility loci were identified to be shared between CD and leprosy. To construct a functional association network for the 11 genes in 10 loci, we performed a protein-protein interaction (PPI) network analysis using the STRING database (https://string-db.org/). One PPI network involving 9 of 11 proteins were found with a significant network connectivity compared with a random set of proteins of the same size ($p < 1.0 \times 10^{-16}$) (Figure 3). To further evaluate this connection, we performed a gene ontology enrichment analysis in biological processes. Applying the threshold of false discovery rate (FDR) < 0.05, we identified a total of 91 biological pathways including regulation of cytokine production involved in immune response,

regulation of T helper 1 type immune response, regulation of T-helper cell differentiation (Table 11).

**Table 3. Lead SNPs with $p < 5 \times 10^{-8}$ in the meta-analysis of CD.**

| Locus | SNP | Candidate gene | Position (hg19) | Risk allele* | CD | | Leprosy | |
|---|---|---|---|---|---|---|---|---|
| | | | | | OR | *p* | OR | *p* |
| 9q32 | rs6478109 | *TNFSF15* | 117,568,766 | G | 2.09 | 1.06E-76 | 0.69 | 1.30E-22 |
| 6p21 | rs9270965 | *HLA* | 32,573,471 | G | 2.12 | 1.20E-58 | 0.56 | 1.06E-42 |
| 4p14 | rs73243351 | *TBC1D1* | 38,335,067 | A | 1.56 | 6.09E-20 | 0.89 | 1.67E-02 |
| 2q37 | rs3749172 | *GPR35* | 241,570,249 | A | 0.72 | 1.37E-14 | 1.02 | 6.28E-01 |
| 10q21 | rs224135 | *ZNF365* | 64,466,802 | A | 0.77 | 8.49E-11 | 0.95 | 2.29E-01 |
| 5q33 | rs755374 | *IL12B* | 158,829,294 | T | 1.32 | 1.42E-10 | 0.79 | 4.28E-08 |
| 20q13 | rs2315647 | *ZBTB46* | 62,379,853 | G | 0.59 | 1.90E-10 | 1.00 | 9.64E-01 |
| 17q21 | rs9895473 | *STAT3* | 40,515,722 | G | 0.77 | 5.57E-10 | 0.98 | 5.28E-01 |
| 22q12 | rs5756393 | *CSF2RB* | 37,300,290 | A | 0.77 | 7.07E-10 | 1.02 | 5.27E-01 |
| 1p31 | rs12033764 | *IL23R* | 67,734,482 | T | 1.28 | 9.29E-10 | 0.98 | 6.58E-01 |
| 2q37 | rs56049444 | *ATG16L1* | 234,157,688 | T | 1.28 | 4.80E-09 | 0.94 | 1.30E-01 |
| 1p36 | rs11249215 | *RUNX3* | 25,297,184 | G | 1.26 | 7.01E-09 | 1.03 | 3.98E-01 |
| 16p11 | rs12446008 | *IL27* | 28,656,150 | C | 1.42 | 1.58E-08 | 0.74 | 2.60E-03 |

CD, Crohn's disease; hg19, human genome version 19; Position, chromosome position; SNP, single nucleotide polymorphism; OR, odds ratio; *p* , *p* value.

*Risk allele in CD GWAS.

**Table 4. Lead SNPs with $p < 5 \times 10^{-8}$ in the meta-analysis of leprosy.**

| Locus | SNP | Candidate gene | Position (hg19) | Risk allele* | CD | | Leprosy | |
|---|---|---|---|---|---|---|---|---|
| | | | | | OR | *p* | OR | *p* |
| 6p21 | rs9271011 | *HLA* | 32,574,676 | G | 1.81 | 1.48E-29 | 0.52 | 4.25E-50 |
| 13q14 | rs9567307 | *LACC1* | 44,471,877 | G | 1.17 | 1.47E-04 | 1.67 | 2.46E-37 |
| 16q12 | rs9302752 | *NOD2* | 50,719,103 | T | 1.14 | 2.68E-03 | 0.65 | 3.04E-26 |
| 9q32 | rs10817678 | *TNFSF15* | 117,579,457 | A | 2.09 | 1.40E-76 | 0.69 | 1.04E-22 |
| 8q21 | rs39761 | *RIPK2* | 90,772,920 | T | 1.13 | 3.09E-03 | 1.31 | 1.01E-12 |
| 6q24 | rs13215778 | *RAB32* | 146,900,563 | T | 1.06 | 2.57E-01 | 1.33 | 3.44E-10 |
| 5q33 | rs4921493 | *IL12B* | 158,836,107 | C | 1.28 | 2.23E-09 | 0.79 | 4.15E-09 |
| 1p36 | rs1801133 | *MTHFR* | 11,856,378 | A | 1.11 | 8.32E-03 | 0.81 | 3.32E-08 |

CD, Crohn's disease; hg19, human genome version 19; Position, chromosome position; SNP, single nucleotide polymorphism; OR, odds ratio; *p* , *p* value.

*Risk allele in CD GWAS.

**Table 5. Genetic correlation estimates by LDSC.**

| Data | $r_g$ | SE | *p* | Number of SNPs used |
|---|---|---|---|---|
| All | -0.30 | 0.12 | 1.45E-02 | 2,289,680 |
| Excluding *MHC* region | -0.33 | 0.17 | 4.78E-02 | 2,254,154 |
| Excluding *TNFSF15* region | -0.25 | 0.12 | 4.21E-02 | 2,287,523 |
| Excluding *MHC* and *TNFSF15* region | -0.24 | 0.15 | 1.03E-01 | 2,251,997 |

$r_g$, genetic correlation estimate; SE, standard error; *p* , *p* value.

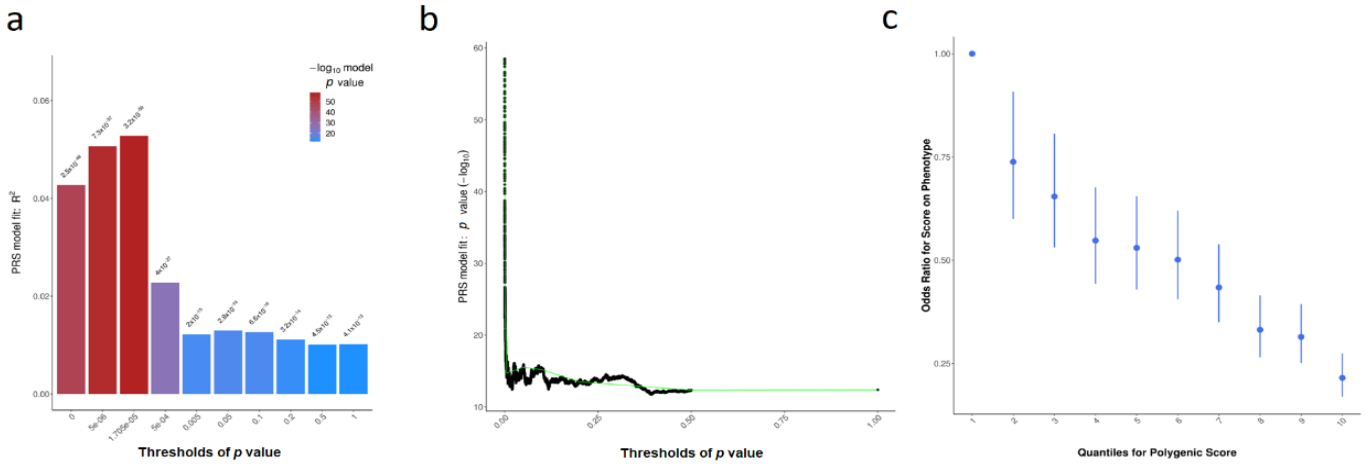
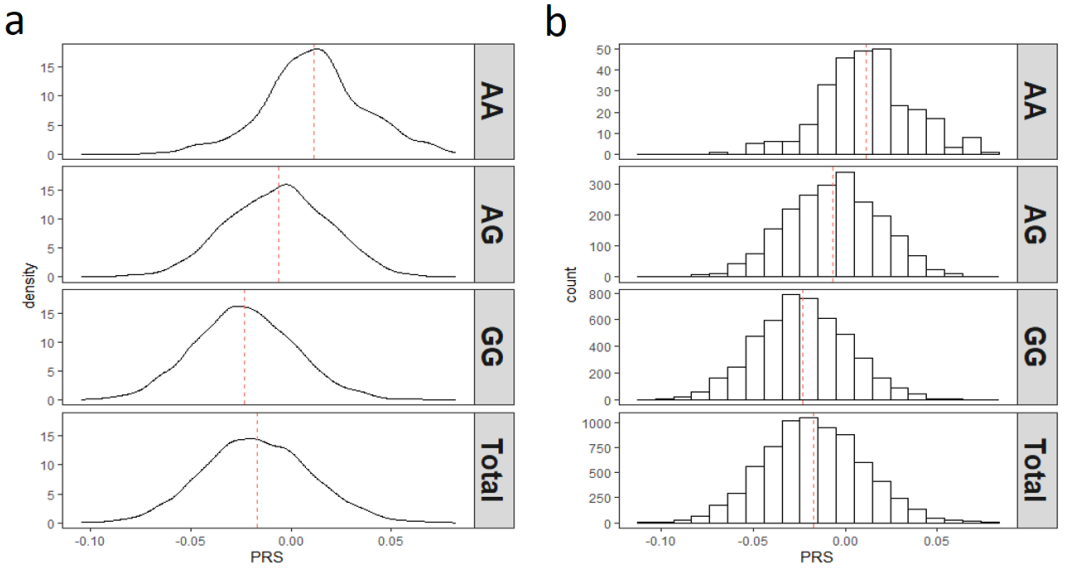



**Figure 1**. **Three different plots representing the result of polygenic risk score (PRS) analysis to estimate the variance of CD explained by leprosy GWAS.** (a) The bar plot shows variance of CD explained by the $PRS_{leprosy}$ at multiple *p* value thresholds. The x-axis represents the *p* value threshold and the y-axis represents the Nagelkerke's pseudo-R-squared fit. (b) The high-resolution plot shows the significance of PRS model fit at all *p* value thresholds. The x-axis represents the *p* value threshold and the y-axis represents the significance of Nagelkerke's pseudo-R-squared fit. A green line connects points of the *p* value thresholds used in the bar plot. (c) The quantile plot provides an illustration of the effect of increasing PRS on predicted risk of phenotype. The x-axis shows the range of different quantiles and the y-axis shows the odds ratio when comparing PRS from different quantiles with the reference quantile. The bars represented 95% confidence intervals of the odds ratio.




c

| | Kolmogorov-Smirnov | | | | Anderson-Darling | | | |
|---|---|---|---|---|---|---|---|---|
| | Total | rs9271011 | | | Total | rs9271011 | | |
| | | AA | AG | GG | | AA | AG | GG |
| SNPs at the best *P*-value threhold | 1.11E-02 | 0.033 | 0.261 | 0.090 | 1.77E-03 | 0.510 | 0.365 | 0.063 |

**Figure 2**. **Mixture of normal distribution of PRS for CD stratified by the rs9271011 genotypes**. (a) density plot, (b) histogram, and (c) statistical tests whether variables have normal distributions based on PRSs stratified by the most significant SNP (rs9271011) in the *MHC* region (chromosome 6: 25~34 Mb). Red line indicates the mean value of PRS.

**Table 6. Variance of CD explained by polygenic risk scores ($PRS_{leprosy}$) with five different thresholds for including SNPs.**

| Threshold | Base file | Variance explained* | *p* | Number of SNPs used |
|---|---|---|---|---|
| **5.00E-08** | Leprosy | **4.27%** | **2.46E-48** | **8** |
| 5.00E-07 | | 5.14% | 2.27E-57 | 13 |
| 5.00E-06 | | 4.94% | 1.42E-55 | 27 |
| 1.71E-05 | | 5.28% | 3.17E-59 | 39 |
| 5.00E-05 | | 4.41% | 3.51E-50 | 63 |
| 5.00E-04 | | 2.27% | 3.95E-27 | 322 |
| **5.00E-08** | Flipped the direction of effect of the *LACC1* and *RIPK2* | **8.21%** | **7.09E-89** | **8** |
| 5.00E-07 | | 9.10% | 1.39E-97 | 13 |
| 5.50E-07 | | 9.10% | 1.39E-97 | 13 |
| 5.00E-06 | | 8.14% | 2.05E-88 | 27 |
| 5.00E-05 | | 6.83% | 1.83E-75 | 63 |
| 5.00E-04 | | 3.27% | 5.48E-38 | 322 |

*p* , *p* value for variance explained; PRS, polygenic risk scores; SNP, single nucleotide polymorphism.

*Variance explained was calculated by SNPs captured by $PRS_{leprosy}$.

**Bold : variance explained by PRS calcculated by SNPs with $P < 5 \times 10^{-8}$ .**

**Table 7. Variance of CD explained by polygenic risk scores ($PRS_{leprosy}$) excluding *TNFSF15* or *MHC* region at best *p*-value threshold.**

| Threshold | Base file | Variance explained* | *p* | Number of SNPs used |
|---|---|---|---|---|
| 1.71E-05 | SNPs at the best *p* threshold | 5.28% | 3.17E-59 | 39 |
| 4.55E-06 | Excluding *TNFSF15* region | 2.76% | 1.78E-32 | 70 |
| 1.71E-05 | Excluding *MHC* region | 2.32% | 1.29E-27 | 38 |
| 2.71E-01 | Excluding both *TNFSF15* and *MHC* region | 0.71% | 1.18E-09 | 26,729 |

*p* , *p* value for variance explained; PRS, polygenic risk scores; SNP, single nucleotide polymorphism.

*Variance explained was calculated by SNPs captured by $PRS_{leprosy}$.

**Table 8. Variance explained by polygenic risk scores ($PRS_{leprosy}$) according to CD location in best *P* threshold.**

| Target file | *P* threshold | Base file | Variance explained* | *p* | Number of SNPs used |
|---|---|---|---|---|---|
| Total (2,354 vs 4,907) | 1.71E-05 | Leprosy | **5.28%** | 3.17E-59 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **9.10%** | 1.39E-97 | 13 |
| Colon (96 vs 4,907) | 1.71E-05 | Leprosy | **1.27%** | 9.93E-04 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **1.51%** | 3.67E-04 | 13 |
| Ileocolon (1,584 vs 4,907) | 1.71E-05 | Leprosy | **4.40%** | 4.95E-42 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **8.35%** | 1.87E-75 | 13 |
| Ileum (454 vs 4,907) | 1.71E-05 | Leprosy | **3.47%** | 4.88E-19 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **6.29%** | 4.07E-32 | 13 |
| Colon + Ileocolon (1,684 vs 4,907) | 1.71E-05 | Leprosy | **4.38%** | 5.89E-43 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **8.17%** | 5.78E-76 | 13 |
| Ileum + Ileocolon (2,038 vs 4,907) | 1.71E-05 | Leprosy | **4.96%** | 2.29E-52 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **9.21%** | 2.29E-92 | 13 |

*p* , *p* value for variance explained; PRS, polygenic risk scores; SNP, single nucleotide polymorphism.

*Variance explained was calculated by SNPs captured by $PRS_{leprosy}$.

**Table 9. Variance explained by polygenic risk scores ($PRS_{leprosy}$) according to CD location in best *P* threshold (random selection).**

| Target file | *P* threshold | Base file | Variance explained* | *p* | Number of SNPs used |
|---|---|---|---|---|---|
| Colon (96 vs 4,907) | 1.71E-05 | Leprosy | **1.27%** | 9.93E-04 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **1.51%** | 3.67E-04 | 13 |
| Ileocolon (100 vs 4,907) | 1.71E-05 | Leprosy | **2.72%** | 1.16E-06 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **4.16%** | 2.88E-09 | 13 |
| Ileum (100 vs 4,907) | 1.71E-05 | Leprosy | **2.07%** | 2.09E-05 | 39 |
| | 5.50E-07 | Flipped the direction of effect of the *LACC1* and *RIPK2* | **2.72%** | 1.38E-06 | 13 |

*p* , *p* value for variance explained; PRS, polygenic risk scores; SNP, single nucleotide polymorphism.

*Variance explained was calculated by SNPs captured by $PRS_{leprosy}$.

**Table 10. Variance of UC explained by polygenic risk scores ($PRS_{leprosy}$) with five different thresholds for including SNPs.**

| Threshold | Base file | Variance explained* | *p* | Number of SNPs used |
|---|---|---|---|---|
| **5.00E-08** | Leprosy | **0.03%** | **2.23E-01** | **8** |
| 5.00E-07 | | 0.01% | 5.01E-01 | 13 |
| 5.00E-06 | | 0.01% | 4.08E-01 | 27 |
| 5.00E-05 | | 0.13% | 9.66E-03 | 63 |
| 5.00E-04 | | 0.01% | 5.33E-01 | 322 |
| 1.00E+00 | | 0.27% | 2.59E-04 | 48,418 |
| **5.00E-08** | Flipped the direction of effect of the *LACC1* and *RIPK2* | **0.20%** | **1.36E-03** | **8** |
| 5.00E-07 | | 0.13% | 1.10E-02 | 13 |
| 5.00E-06 | | 0.01% | 4.44E-01 | 27 |
| 5.00E-05 | | 0.04% | 1.71E-01 | 63 |
| 5.00E-04 | | 0.00% | 9.23E-01 | 322 |
| 1.00E+00 | | 0.25% | 4.31E-04 | 48,418 |

*p* , *p* value for variance explained; PRS, polygenic risk scores; SNP, single nucleotide polymorphism.

*Variance explained was calculated by SNPs captured by $PRS_{leprosy}$.

**Bold : variance explained by PRS calcculated by SNPs with $P < 5 \times 10^{-8}$.**

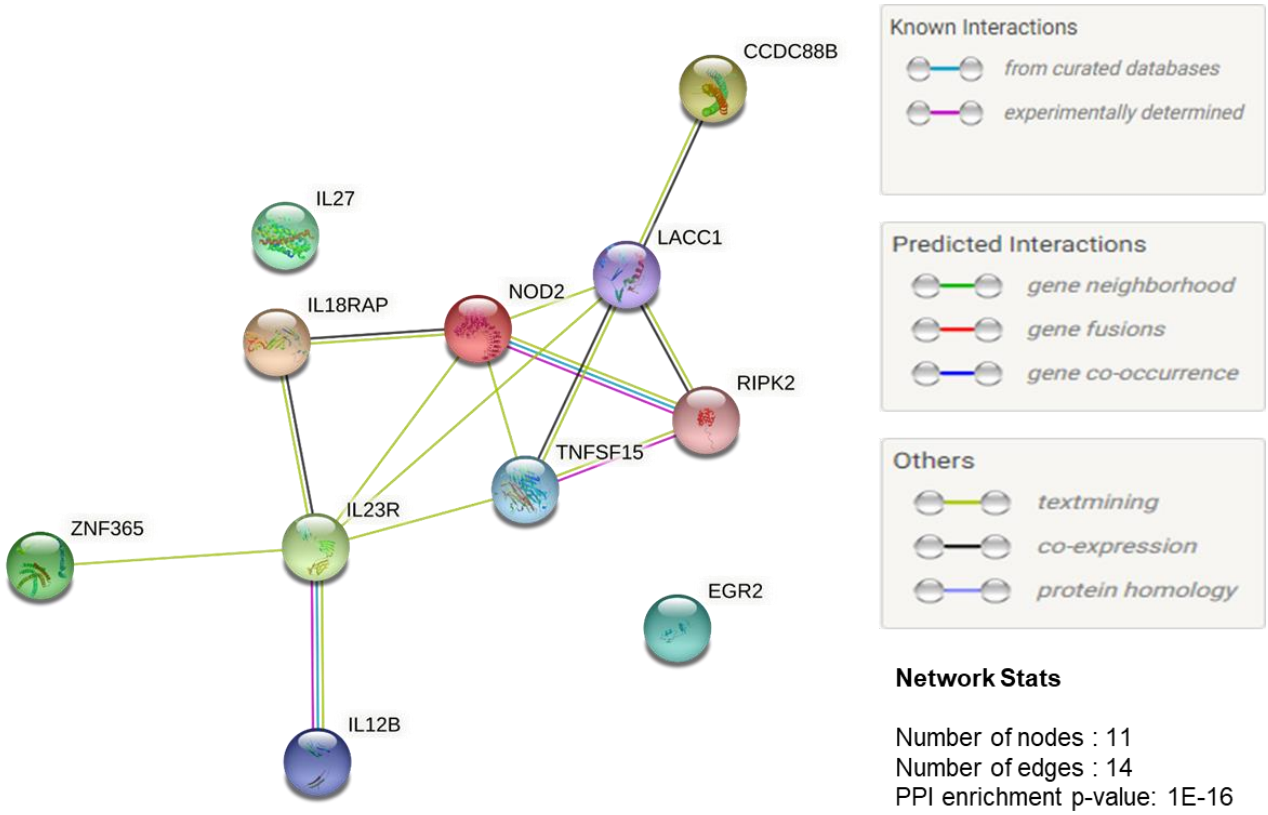



**Figure 3**. **Protein-protein interaction (PPI) network visualized by STRING**. Eleven genes derived from the 10 shared loci were used for the PPI network analysis using the STRING database. The color of lines between proteins indicates types of protein-protein interactions. The significant PPI enrichment *p* value showed that proteins in the PPI network have more interactions among themselves than what would be expected for a random set of proteins of the same size.

**Table 11. Top 20 biological processes significantly over-represented among the 10 loci shared between CD and leprosy.**

| GO term | Desciption | Strength* | FDR | Involved genes |
|---|---|---|---|---|
| GO:0002367 | Cytokine production involved in immune response | 2.45 | 2.23E-05 | *IL12B,IL18RAP,NOD2,LACC1* |
| GO:0002825 | Regulation of T-helper 1 type immune response | 2.42 | 2.23E-05 | *RIPK2,IL12B,IL23R,IL27* |
| GO:0045622 | Regulation of T-helper cell differentiation | 2.31 | 2.56E-05 | *RIPK2,IL12B,IL23R,IL27* |
| GO:0050863 | Regulation of T cell activation | 1.51 | 4.52E-05 | *RIPK2,IL12B,NOD2,IL23R,CCDC88B,IL27* |
| GO:0019221 | Cytokine-mediated signaling pathway | 1.26 | 5.10E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,IL27,TNFSF15* |
| GO:0031347 | Regulation of defense response | 1.27 | 5.10E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,IL27,LACC1* |
| GO:0042129 | Regulation of T cell proliferation | 1.73 | 5.10E-05 | *RIPK2,IL12B,IL23R,CCDC88B,IL27* |
| GO:0070431 | Nucleotide-binding oligomerization domain containing 2 (NOD2) signaling pathway | 2.82 | 5.10E-05 | *RIPK2,NOD2,LACC1* |
| GO:0002697 | Regulation of immune effector process | 1.41 | 6.46E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,IL27* |
| GO:0006952 | Defense response | 1.04 | 7.04E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,CCDC88B,IL27,LACC1* |
| GO:0032729 | Positive regulation of interferon-gamma production | 2.01 | 7.58E-05 | *RIPK2,IL12B,IL23R,IL27* |
| GO:0050870 | Positive regulation of T cell activation | 1.63 | 7.58E-05 | *RIPK2,IL12B,NOD2,IL23R,CCDC88B* |
| GO:0001819 | Positive regulation of cytokine production | 1.36 | 7.67E-05 | *RIPK2,IL12B,NOD2,IL23R,CCDC88B,IL27* |
| GO:0002699 | Positive regulation of immune effector process | 1.60 | 7.83E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R* |
| GO:0080134 | Regulation of response to stress | 1.00 | 9.28E-05 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,IL27,ZNF365,LACC1* |
| GO:0098542 | Defense response to other organism | 1.14 | 9.28E-05 | *RIPK2,IL12B,NOD2,IL23R,CCDC88B,IL27,LACC1* |
| GO:0032740 | Positive regulation of interleukin-17 production | 2.50 | 1.00E-04 | *IL12B,NOD2,IL23R* |
| GO:0042102 | Positive regulation of T cell proliferation | 1.87 | 1.30E-04 | *RIPK2,IL12B,IL23R,CCDC88B* |
| GO:0002827 | Positive regulation of T-helper 1 type immune response | 2.43 | 1.40E-04 | *RIPK2,IL12B,IL23R* |
| GO:0006955 | Immune response | 0.95 | 1.40E-04 | *RIPK2,IL12B,IL18RAP,NOD2,IL23R,IL27,TNFSF15,LACC1* |

FDR, false discovery rate; GO, Gene Ontology

$Log_{10}$(observed/ expected). This measure describes how large enrichment effect is.

## Discussion

In this study, we performed genetic pleiotropy analysis between CD and leprosy in Asian population by estimating genetic correlation between CD and leprosy using LD score regression (LDSC) and genetic overlap using polygenic risk scores (PRS). We found that CD and leprosy shared a substantial number of genetic susceptibility loci in East Asians with significant negative correlation between the two diseases.

First, using LD score regression (LDSC) to evaluate genetic correlation between CD and leprosy, we found that there was a significant negative genetic correlation between CD and leprosy ($r_g$[SE] = -0.30 [0.12], $p = 1.5 \times 10^{-2}$). After removing the most significant loci in CD (*MHC* and *TNFSF15* loci), we found that the correlation estimates and significance of *p*-value were decreased, suggesting that a rather high correlation between CD and leprosy might have been driven by these two loci with population-specific effect.

Second, we used PRS to evaluate how much of variability in the CD phenotype can be explained by PRS based on leprosy. Applying a threshold of the genome-wide significance level, $PRS_{leprosy}$ explained 4.27 % of variance of CD. When the direction of effect of the *LACC1* and *RIPK2* loci was flipped to consider its original directional concordance between CD and leprosy, $PRS_{leprosy}$ could explain up to 8.21 % of variance of CD. Additionally, at the best *p*-value threshold, $PRS_{leprosy}$ could explain 5.28 % of variance of CD, suggesting a significant overlap of genetic architecture between CD and leprosy. Although CD and ulcerative colitis (UC) are both inflammatory bowel diseases with many similarities, $PRS_{leprosy}$ could explain only 0.27% of variance of UC at the best *p*-value threshold (Table 10), supporting key differences of pathophysiology between the two diseases.

Third, to uncover subtypes of CD which show shared genetic architecture with leprosy, we re-calculated CD variance after stratifying CD samples by disease location. $PRS_{leprosy}$ could explain up to 4.40 % of variance of ileocolonic CD and 1.27% of variance of colonic CD. As performance of PRS could depend on sample size,[23] we re-calculated $PRS_{leprosy}$ after random sampling for similar sample size among subtypes of CD location. Although the variance explained by $PRS_{leprosy}$ dropped in ileocolonic and ileal CD, the tendency of higher explained variance of ileocolonic CD than that of colonic CD was maintained (ileocolonic CD: 2.72%, colonic CD: 1.27%), suggesting that CD involving ileum is closer to leprosy than colonic CD.

Finally, to uncover factors driving this significant overlap between CD and leprosy, we re-calculated CD variance after removing the two most significant loci in CD (*MHC* and *TNFSF15*). When we removed each of the two, the variance explained by $PRS_{leprosy}$ dropped. Furthermore, when we removed both of them, the variance explained by $PRS_{leprosy}$ dropped sharply, suggesting that the large variance explained by $PRS_{leprosy}$ might have been driven by these two loci with population specific effect.

Our study is the first to examine genetic correlation between CD and leprosy in East Asians. In the previous studies, several susceptibility loci were identified to be shared between CD and leprosy, however, due to lack of individual level of data, they couldn't perform PRS analysis to estimate overlap of genetic architecture between the two diseases. Using our CD individual-level data, we could perform PRS analysis to estimate genetic overlap between the two diseases.

Our study has several limitations. First, the estimates of genetic correlation between CD and leprosy using LDSC might be overestimated. For estimating genetic correlation, LDSC and genomic restricted maximum likelihood (GREML) are the methods that have been widely used, shedding light on the shared genetic architecture of complex traits based on genome-wide SNPs. Although the accuracy of GREML is reported to be higher than that of LDSC,[10] we had to use LDSC to estimate genetic correlation as we had an access to GWAS summary statistics of leprosy only. Second, leprosy has two distinct clinical manifestations, designated as tuberculoid and lepromatous, it would be interesting to examine which type shares common genetic susceptibility loci with CD, or specific subtype of CD. Due to our modest sample size and lack of detailed clinical information of leprosy, our findings on the shared genetic architectures of CD and leprosy are hardly comprehensive. Larger GWAS in the future may reveal more shared loci between CD and leprosy and shed light on the shared molecular mechanisms in their pathophysiology.

In conclusion, we have identified a negative and significant genetic correlation between CD and leprosy by using the largest available GWAS datasets. Despite of several shortcomings, our study represents the first systemic effort to compare the genetic basis of CD and leprosy. Majority of the susceptibility loci shared between CD and leprosy showed allelic effects in the opposite directions, suggesting their roles in both defense against infection and inflammation. After removing the most significant CD susceptibility loci (*MHC* and *TNFSF15*), genetic

correlation and overlap were decreased, suggesting these two loci were main factors of a rather higher genetic overlap between the two diseases. Of the two loci, T allele at rs6478108/rs6478109 ($r^2 = 1$) in *TNFSF15* increased risk of CD, but decreased risk of leprosy based on fine-mapping analysis.[24] Fine-mapping analysis of the other shared loci warrants further study. We also performed pathway enrichment analysis of 11 genes derived from the 10 shared loci, showing the significant association with 91 biological pathways including regulation of cytokine production involved in immune response, regulation of T helper 1 type immune response, regulation of T-helper cell differentiation. Our findings of negative correlation between CD and leprosy suggest that CD might be caused by over-reaction to pathogenic stimuli.

Web Resources

The URLs for data presented herein are as follows:

The 1000 Genome Project, http://www.1000genomes.org/

STRING database, https://string-db.org/

References


1. Zhang, F. et al. Genomewide association study of leprosy. *N. Engl. J. Med.* **361,** 2609–2618 (2009).
2. Zhang, F. et al. Identification of two loci at IL23R and RAB32 that influence susceptibility to leprosy. *Nat. Genet.* **43,** 1247–1251 (2011).
3. Liu, H. et al. Identification of IL18RAP/IL18R1 and IL12B as leprosy risk genes demonstrates shared pathogenesis between inflammation and infectious diseases. *Am. J. Hum. Genet.* **91,** 935–941 (2012).
4. Liu, H. et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. *Nat. Genet.* **47,** 267–271 (2015).
5. Wang, Z. et al. A large-scale genome-wide association and meta-analysis identified four novel susceptibility loci for leprosy. *Nat. Commun.* **7,** 13760 (2016).
6. Liu, H. et al. Genome-wide analysis of protein-coding variants in leprosy. *J. Invest. Dermatol.* **137,** 2544–2551 (2017).
7. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
8. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet*. **47**, 979–986 (2015).
9. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet*. **47**, 1236–1241 (2015).
10. Bulik-Sullivan, B. et al. Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. *Am J Hum Genet.* **102**, 1185-1194 (2018).

11. Martin AR. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet*. 51, 584-591 (2019).

12. Lambert SA. et al. Towards clinical utility of polygenic risk scores. *Hum Mol Genet.* **28**, R133- 142 (2019).

13. Chen MH, Raffield LM, Mousas A, et al.; VA Million Veteran Program. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**,1198-213.e14 (2020).

14. Choi, S.W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, 1–6 (2019).

15. Yang, S. –K. et al. Identification of loci at 1q21 and 16q23 that affect susceptibility to inflammatory bowel disease in Koreans. *Gastroenterology* **151**, 1096–1099 (2016).

16. Jung, S. et al. Identification of three novel susceptibility loci for inflammatory bowel disease in Koreans in an extended genome-wide association study. *J. Crohns Colitis* (2021) Apr14:jjab060. doi: 10.1093/ecco-jcc/jjab060

17. Han, B. et al. Amino acid position 37 of HLA-DRβ1 affects susceptibility to Crohn's disease in Asians. *Hum. Mole. Genet*. **27**, 3901–3910 (2018).

18. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat*. *Genet*. **39**, 906–913 (2007).

19. Liu, J. Z. et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet*. **42**, 436–440 (2010).

20. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

21. Szklarczyk D. et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. **49**, 605-612 (2021).

22. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, 325-334 (2021).

23. Choi, S.W. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759-2772 (2020).

24. Sun, Y. et al. Fine-mapping analysis revealed complex pleiotropic effect and tissue-specific regulatory mechanism of TNFSF15 in primary biliary cholangitis, Crohn's disease and leprosy. *Sci. Rep.* **6**, 31429; doi: 10.1038/srep31429 (2016).

## 국문 요약

### 연구목적

서양인 크론병과 중국인 나병환자를 이용한 전장 유전체 분석 연구 결과에 따르면 두 질병은 비특이적 선천 면역과 연관된 유전적 감수성을 공유하고 있다. 이들 유전자좌 내의 다면발현성 변이는 크론병과 나병에서 반대되는 대립형질 효과를 나타내고 있다.

### 연구방법

한국인 크론병 환자 2,357 명과 대조군 4,907 명의 연관 분석 자료와 중국인 나병 환자 2,960 명과 대조군 3,747 명의 연관 분석자료를 각각 메타분석하였다. 동아시아인 크론병과 나병의 메타분석 자료를 토대로 유전자위험점수 분석과 연관 불균형 점수 회귀 분석법을 통해 두 질병 사이의 유전적 상관관계를 확인하였다.

### 연구결과

연관 불균형 점수 회귀 분석법을 통해 두 질병 간의 음의 유전적 상관관계가 있는 것으로 확인되었다($r_g$[SE] = -0.30 [0.12], $p$ = 1.5 x $10^{-2}$). 한국인 크론병의 유전자위험점수로 계산한 표현형 분산은 중국인 나병 환자 데이터 사용시 5.27 % 의 설명력을 보였다. 크론병과 나병에서 동일한 위험 방향을 가지고 있는 2 개의 loci 를 뒤집었을 때, 8.21% 의 설명력을 보였다. 크론병에서 가장 유의한 2 개의 유전자위 (*MHC*, *TNFSF15*)를 제거시 두 질병 간의 유전적 상관관계와 설명력이 감소하는 것을 확인하였다.

### 결론

본 연구는 동아시아인에서 크론병과 나병 사이의 유전적 연관성과 중첩정도를 확인한 첫 연구이다. 동아시아인에서 크론병과 나병은 상당수의 유전적 감수성

유전자좌를 공유했으며, 대부분의 공유 유전자좌는 위험 대립 형질이 반대인 것을 확인하였다. 또한 크론병의 가장 유의한 2 개의 유전자위 (*MHC*, *TNFSF15*) 가 크론병과 나병 간의 유전자 구조가 겹치는 주요 요인이 될 수 있음을 확인하였다.

**Key words:** Crohn's disease; leprosy; polygenic risk scores; genetic correlation