Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy

Hong Liu[1,2,3,4,16], Astrid Irwanto[5,6,16], Xi'an Fu[2,4], Gongqi Yu[2,4], Yongxiang Yu[2,4], Yonghu Sun[2,4], Chuan Wang[2,4], Zhenzhen Wang[2,4], Yukinori Okada[7,8,9], Hui Qi Low[5], Yi Li[5], Herty Liany[5], Mingfei Chen[2,4], Fangfang Bao[2,4], Jinghui Li[2,4], Jiabao You[2,4], Qilin Zhang[3], Jian Liu[2,4], Tongsheng Chu[2,4], Anand Kumar Andiappan[10], Na Wang[2,4], Guiye Niu[2,4], Dianchang Liu[2,4], Xiulu Yu[2,4], Lin Zhang[2,4], Hongqing Tian[1,11], Guizhi Zhou[2,4], Olaf Rotzschke[10], Shumin Chen[2,4], Paul W.I.B. de Bakker[8,9,12,13], Xuejun Zhang,[14,15,17], Jianjun Liu[5,6,17], Furen Zhang[2,3,4,11,17]

1. Shandong Provincial Hospital for Skin Diseases, Shandong University, Jinan, China.
2. Shandong Provincial Institute of Dermatology and Venereology, Shandong Academy of Medical Sciences, Jinan, China.
3. School of Medicine, Shandong University, Jinan, China.
4. Shandong Provincial Key Lab for Dermatovenereology, Jinan, China.
5. Human Genetics, Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, Singapore.
6. Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore.
7. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, USA.
8. Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, USA.
9. Program in Medical and Population Genetics, Broad Institute, Cambridge, USA.
10. Singapore Immunology Network, Singapore, Singapore.
11. National Clinical Key Project of Dermatology and Venereology, Jinan, China.
12. Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands.
13. Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands.
14. Department of Dermatology, First Affiliated Hospital, Anhui Medical University, Hefei, China.
15. The Key Laboratory of Gene Resource Utilization for Severe Diseases, Ministry of Education and Anhui Province, Hefei, China.
16. These authors contributed equally to this work.
17. Shared senior authors.

Correspondence should addressed to Furen Zhang (zhangfuren@hotmail.com)

Abstract

Genome-wide association study (GWAS) has led to the discovery of several susceptibility loci for leprosy with robust evidence, providing biological insight about the role of host genetic factors in Mycobacteria infection. However, these loci can explain partially disease heritability, and additional genetic risk factors remain to be discovered. We performed a three-stage GWAS of leprosy in the Chinese population using 8,313 cases and 16,017 controls. Besides confirming all the previously published loci, six new susceptibility loci were discovered and further gene prioritization analysis of these loci implicated *BATF3*, *CCDC88B* and *CIITA*/*SOCS1* as new susceptibility genes for leprosy. A systematic evaluation of pleiotropic effect revealed a high tendency for leprosy susceptibility loci to show association with autoimmune/inflammatory diseases. Further analysis suggests that molecular sensing of infection may play a similar pathogenic role, whereas immune responses play a discordant role between infectious and inflammatory diseases.

Main (1996 words)

Leprosy, a chronic dermatological and neurological disease, is caused by the infection of *Mycobacterium leprae*, whose manifestation, progression and prognosis are strongly associated with patients' immune system[1]. Due to the implementation of multidrug therapy (MDT) in 1980s by the World Health Organization, the prevalence of leprosy has declined dramatically. Nonetheless, with more than 200,000 new cases reported annually throughout the world[2], leprosy remains a public health problem, especially in developing countries.

The molecular nature of genetic susceptibility to leprosy has been intensively investigated by candidate gene studies[3,4], genome-wide linkage analyses[5-8] and association studies (GWAS)[9-11], which have identified 11 leprosy susceptibility loci. Most of the implicated susceptibility genes encoded proteins involved in immunity, while one locus (*RAB32* on 6q24.3) revealed the involvement of autophagocytosis in leprosy pathogenesis. However, these genetic risk loci can only partially explain genetic susceptibility to leprosy; additional genetic risk factors remain to be

discovered.

Here we conducted a large three-stage GWAS of leprosy in Chinese population. The genome-wide discovery analysis (Stage 1) involved two independent datasets, the previously published GWAS dataset of 706 leprosy cases, 1,225 healthy controls and 4,362 immune disease-related subjects as population controls of northern Chinese Han[9] and a new unpublished dataset of 842 leprosy cases and 925 controls from northern (Chinese Han) and southern China (Chinese Han and ethnic minorities) (Supplementary Table 1). Principal components analysis (PCA) confirmed the Chinese ancestry of all the samples. After stringent quality control filtering (Online Methods), the log-additive test was performed to test the association of SNPs by including the first 5 principal components as covariates. The meta-analysis of the two independent datasets investigated 4,577,171 common SNPs (467,552 genotyped, 4,109,619 imputed) in a total of 1,548 cases and 6,512 controls. The genomic inflation factor (λ_{GC} = 1.02) is small, indicating very little confounding effect of population stratification (Supplementary Figure 1). The discovery analysis revealed strong evidences for all the previously identified loci and suggested new loci as well as new independent associations within the known non-MHC susceptibility loci[7,8,9,10] (Supplementary Figure 2 and 3).

To validate new loci, we selected the top SNPs from 91 independent new loci with suggestive association ($P < 5\times10^{-4}$) for a follow-up analysis in an additional 2,761 leprosy cases and 3,038 controls of northern Chinese Han (Stage 2) (Supplementary Table 1). Of the 88 successfully genotyped SNPs, 11 SNPs showed association at $P < 0.05$ in the validation samples, and 5 SNPs showed consistent risk effect between the discovery (Stage 1) and validation (Stage 2) samples. These 16 SNPs were selected for further validation in five additional independent sample series from different regions of China, totaling 4,004 cases and 6,467 controls (Supplementary Table 1, Supplementary Figure 4) (Stage 3). In addition, we also genotyped previously reported SNPs and four new independent secondary SNPs ($P < 5\times10^{-4}$) of the known non-MHC loci in Stage 2 and 3 validation samples

(Supplementary Table 2).

Joint analysis of all the samples of 3 stages, totaling 8,313 cases and 16,017 controls, was performed using a fixed-effects meta-analysis. Six new associations were discovered at genome-wide significance ($P < 5\times10^{-8}$), including rs2221593 on 1q32.3 ($P = 3.09\times10^{-8}$, OR = 1.15), rs73058713 on 5p14.3 ($P = 9.54\times10^{-9}$, OR = 1.19), rs10817758 on 9q32 ($P = 1.15\times10^{-8}$, OR = 1.13), rs58600253 on 10q21.3 ($P = 3.02\times10^{-12}$, OR = 1.22), rs663743 on 11q13.1 ($P = 8.84\times10^{-14}$, OR = 1.24) and rs77061563 on 16p13.13 ($P = 6.23\times10^{-15}$, OR = 0.84) (Figure 1 and 2, Supplementary Table 3). All these SNPs showed significant association in the combined Stage 2 and 3 validation samples (after correction for multiple testing of 88 SNPs). In addition, we discovered two new independent associations within *RIPK2* (rs160451, $P_{conditional} = 7.45\times10^{-12}$, $OR_{conditional} = 0.83$) and *LACC1* (rs8002861, $P_{conditional} = 1.07\times10^{-12}$, $OR_{conditional} = 1.22$) loci (Supplementary Tables 2 and 4). Pairwise interaction analysis among the 18 newly and previously identified SNPs did not discover any significant epistatic interaction (data not shown).

All of the previously reported SNPs in Chinese population showed consistent and genome-wide significant association (Supplementary Table 2) in the current study, except rs2735591 within the *BCL10* locus ($P = 8.0\times10^{-4}$, OR = 1.08, $P_{het} = 0.008$) (Supplementary Figure 5) where significant heterogeneity of association was observed between northern and southern Chinese samples. By only including the northern Chinese Han samples in the joint analysis, we did observe a strong association ($P = 9.27\times10^{-8}$, OR = 1.18, $P_{het} = 0.80$) that is consistent with the previous finding in northern Chinese Han[4]. Intriguingly, the SNP did not show association in the samples of southern Chinese Han and Minorities (P = 0.58), which needs to be confirmed by further studies.

We performed a fine mapping analysis of the MHC region by imputing untyped SNPs, HLA alleles and amino acid polymorphisms in the discovery dataset (excluding the 4,362 immune disease-related subjects). We found that the extensive associations within the MHC region were driven by the variants located around the region of *HLA* class II genes, with *HLA-DRB1*15* to be the most significant *HLA* variant (P =

3.5×10^{-28}, OR = 2.11) and rs9271011 to be the most significant SNP ($P = 4.6\times10^{-29}$, OR = 2.16) (Figure 3a). rs9271011 showed strong LD ($r^2 = 0.8$, D' = 1) only with the *DRB1*15* allele group (Supplementary Table 5). The accuracy of *HLA-DRB1*15* imputation has been confirmed by sequencing-based HLA-typing analysis (Online Methods, Supplementary Table 6), and rs9271011 has been validated by the association of rs9271100 ($r^2 = 0.89$ with rs9271011, see Online Methods) in the Stage 2 and 3 validation samples ($P = 6.46\times10^{-71}$, OR = 1.63) and the combined discovery and validation samples ($P = 7.76\times10^{-95}$, OR = 1.68) (Supplementary Table 2). Our findings are consistent with the previously reported association of *HLA-DRB1*15* in India[12], Brazil[13] and Thai[14] populations. Conditioning on *HLA-DRB1*15*/rs9271100 could abolish the strong and extensive associations within the MHC region, with only *HLA-DQB1*04:01* showing suggestive independent association ($P = 1.7\times10^{-6}$, OR = 0.37) (Figure 3b).

Because all the new associations are located within polygenic LD blocks, we investigated the biological and functional relevance of the genes within the confirmed loci by deriving a functional score for each gene according to whether it carries a nonsynonymous coding variant being in high LD with the validated SNP, has eQTL or mQTL effect with the validated SNP, and is implicated by protein-protein interaction or pathway analysis (see Online Methods). For each locus, the gene with the highest score was considered to be the candidate of susceptibility gene (Supplementary Table 7). For the known loci, the candidate genes identified here are the same to the previously reported ones. For the novel loci, our analysis identified *BATF3* (rs2221593, 1q32.3) and *CCDC88B* (rs694739, 11q13.1) (functional score $\geq$ 3) as well as *CIITA* (rs77061563, 16p13.13) (score = 2) as candidate genes, but did not reveal any eQTL effect. For the loci of rs73058713 on 5p14.3, rs58600253 on 10q21.3 and rs10817758 on 9q33.1, we did not find any genes with a score > 1, although there were biologically interesting candidates within these loci, including *CDH18* (5p14.3), *EGR2* (10q21.3) and *DEC1* (9q33.1). The biological functions of these candidate genes and their potential roles in leprosy development were summarized in Supplementary Table 7.

Because all the validated SNPs are non-coding variants, we investigated whether they have potential regulatory functions. Rs9271100 showed strong eQTL effect on *HLA-DRB1* expression in the lymphoblastoid cell lines of HapMap3 Chinese samples (CHB, $P = 9.9\times10^{-17}$) (Supplementary Figure 6) with the risk 'T' allele being associated with increased *HLA-DRB1* expression and the creation of MEF2A binding motif (Supplementary Table 8a) within the 5' regulatory region of *HLA-DRB1*. The risk 'C' allele of rs9302752 within the *NOD2* locus was also associated with the low expression of *NOD2*[15] and found to disrupt the binding motifs of FOXI1, OCT-x and Oct-1 (Supplementary Table 8b). No evidences of regulatory functionality were found for the rest SNPs.

We carried out a systematic evaluation of the pleiotropic effects of the leprosy susceptibility loci by searching for the reported associations of other diseases within these loci in the NHGRI GWAS catalog[16]. Of the 16 established leprosy loci, 11 have reported associations with other diseases, and interestingly these are largely autoimmune/inflammatory diseases and related quantitative traits (Supplementary Table 9). For example, the association of *CCDC88B* was reported in primary biliary cirrhosis (PBC)[17], sarcoidosis[18], mean platelet volume[19] and Inflammatory Bowel Disease (IBD)[20]. To assess the independence of various disease associations within these loci, we performed conditional and haplotype analyses in our discovery dataset (excluding the 4,362 immune disease-related subjects) and investigated the LD patterns of these SNPs in Asian, European and African populations by using the data from the 1000 Genomes Project. The analysis revealed several shared susceptibility loci between leprosy and autoimmune/inflammatory diseases, such as *RIPK2* on 8q21.3, *LACC1* on 13q14, *IL12B* on 5q33.3, *TNFSF15* on 9q32 and *CCDC88B* on 11q13.1 where the reported SNPs of leprosy and other diseases are in strong LD and carried by the same haplotypes (Supplementary Table 9 and 10), and the SNPs of other diseases also showed significant but non-independent association in our leprosy samples. For example, *TNFSF15* was also reported to be associated with CD, UC, IBD and PBC. The validated leprosy SNP, rs6478109, showed strong association with IBD, and six SNPs reported in CD, UC, IBD and PBC are all in LD with rs6478109

and showed strong but non-independent association in our leprosy samples. The analysis also revealed the independence of several loci (*IL12B*, *IL1RL1*/*IL18RAP*, 10q21.2 and *SOCS1*) from other disease associations, where the leprosy SNPs are in low LD with the SNPs of other diseases, and the SNPs of other diseases did not show association in leprosy. For example, IBD GWAS[20] reported an association at rs10761659 within the same LD block of rs58600253 on 10q21.2. However, rs10761659 did not show association in our leprosy samples, and rs10761659 and rs58600253 are in low LD across different populations. The discovery of causal variants and direct analysis of leprosy SNPs in other diseases can help to confirm the independences of leprosy associations within these loci.

Interestingly, for some shared susceptibility loci, such as *RIPK2* and *LACC1*[20-22], the reported SNPs are in high LD and show the concordant risk effect in leprosy and other diseases (Supplementary Table 9). Meanwhile, some shared loci showed the opposite or discordant risk effects, such as *IL18RAP*/*IL1RL1* where the minor allele showed the risk effect with leprosy, but protective effects with CD[21], IBD[20], celiac disease[23,24] and atopic dermatitis[25] (Supplementary Table 9). Furthermore, we carried out a gene ontology (GO) enrichment analysis of the genes showing concordant (*RIPK2* and *LACC1*) or discordant (*IL1RL1*/*IL18RAP*, *IL12B*, *TNFSF15*, *CCDC88B*) risk effects between immune-related diseases and leprosy. This analysis revealed NOD-signaling pathway (GO:0070427) as the most enriched GO term for concordant genes, and cytokine metabolic processes (GO:0042107) and immune response regulation (GO:002825, 0006955) for the discordant genes. Our analysis has revealed a high tendency for leprosy susceptibility genes to show association with autoimmunity/inflammatory diseases, highlighting the sharing of pathogenic mechanisms between infectious and autoimmunity/inflammatory diseases. Furthermore, our analysis also suggested that genes involved in pathogen-sensing may play a similar pathogenic role in inflammatory and infectious diseases, which is consistent with previous reports of the involvement of bacterial infection in the development of CD[26-28], whereas immune responses play a discordant role in infection and inflammatory diseases.

Overall, our current study has advanced the genetic understanding of leprosy by significantly increasing the number of confirmed genetic susceptibility loci and provided further biological insight into the pleiotropic effects of leprosy susceptibility loci on autoimmunity/inflammatory diseases, demonstrating that while molecular sensing of intracellular infection may play a similar pathogenic role, the immune-related pathways play a discordant role in these diseases, which supports the concept that strong immune and inflammatory responses are good for defending infection, but bad for increasing the risk of autoimmunity/inflammatory diseases. Further sequencing and functional investigations of these loci will likely identify true susceptibility genes and causal mutational events of these susceptibility loci, which can elucidate the molecular mechanisms of leprosy development and understand their impacts on autoimmune/inflammatory diseases.

URLs:

NHGRI GWAS Catalog, http://www.genome.gov/gwastudies/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
SHAPEIT, https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
IMPUTE v2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
GRAIL, http://www.broadinstitute.org/mpg/grail/
DAPPLE v2, http://www.broadinstitute.org/mpg/dapple/dappleTMP.php
MAGENTA, http://www.broadinstitute.org/mpg/magenta/
IPA, http://www.ingenuity.com/
AmiGO, http://geneontology.org/page/go-enrichment-analysis
EMBL-EBI Immunogenetics HLA database, http://www.ebi.ac.uk/imgt/hla/
Omixon target software, www.omixon.com

Acknowledgements The authors have declared that no competing interests exist. We thank the individuals who participated in this project. This work was funded by a grant from the National Natural Science Foundation of China (81371721, 81271746, 31200933, 81101187), National 863 program (2014AA020505), the Shandong Provincial Advanced Taishan Scholar Construction Project and the Agency for Science, Technology, and Research of Singapore. A.I. was supported by the Singapore International Graduate Award.

Author Contributions F.-R.Z. obtained financial support and conceived of the study. F.-R.Z. and J.-J.L. designed the study. H.L. was responsible for sample selection,

genotyping and project management. A.I. was responsible for all the statistical analysis. S.-M.C., T.-S.C., X.-L.Y., L.Z., and D.-C.L undertook recruitment, collected phenotype data. X.-A.F., G.-Q.Y., Y.-X.Y., C.W., F.-F.B., Q.-L.Z., H.-Q.T., M.-F.C., J.-H.L., J.-B.Y., J.L., G.-Z.Z., N.W. and N.-G.Y. conducted sample selection and performed the genotyping of the validation study. A.I., Z.-Z.W., Y.-H.S., Y.L., H.-Q.L, Herty.Liany. and Y.O., undertook data checking, statistical analysis and bioinformatic interrogations. A.-K.A. and O.R. shared their e-QTL database. X.-J.Z. provided the partial control data in the discovery stage. A.I., L.H., F.-R.Z., and J.-J.L wrote the paper. All the authors contributed to the final paper, with F.-R.Z., J.-J.L., H.L. and A.I. playing key roles.

Statement of competing financial interests: There are no competing interests declared.

References:

1. Britton, W.J. & Lockwood, D.N. Leprosy. *Lancet* **363**, 1209-19 (2004).
2. Global leprosy: update on the 2012 situation. *Wkly Epidemiol Rec* **88**, 365-79 (2013).
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**, 935-41 (2012).
4. Liu, H. *et al.* An association study of TOLL and CARD with leprosy susceptibility in Chinese population. *Hum Mol Genet* (2013).
5. Todd, J.R., West, B.C. & McDonald, J.C. Human leukocyte antigen and leprosy: study in northern Louisiana and review. *Rev Infect Dis* **12**, 63-74 (1990).
6. Siddiqui, M.R. *et al.* A major susceptibility locus for leprosy in India maps to chromosome 10p13. *Nat Genet* **27**, 439-41 (2001).
7. Mira, M.T. *et al.* Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature* **427**, 636-40 (2004).
8. Alcaïs, A. *et al.* Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk factor for early-onset leprosy. *Nat Genet* **39**, 517-22 (2007).
9. Zhang, F. *et al.* Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. *Nat Genet* **43**, 1247-51 (2011).
10. Zhang, F.R. *et al.* Genomewide association study of leprosy. *N Engl J Med* **361**, 2609-18 (2009).
11. Wong, S.H. *et al.* Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog* **6**, e1000979 (2010).
12. Rani, R., Fernandez-Vina, M.A., Zaheer, S.A., Beena, K.R. & Stastny, P. Study of HLA class II alleles by PCR oligotyping in leprosy patients from north India. *Tissue Antigens* **42**, 133-7 (1993).
13. Vanderborght, P.R. *et al.* HLA-DRB1*04 and DRB1*10 are associated with resistance and susceptibility, respectively, in Brazilian and Vietnamese leprosy

patients. in *Genes Immun*, Vol. 8 320-4 (England, 2007).
14. Schauf, V. *et al.* Leprosy associated with HLA-DR2 and DQw1 in the population of northern Thailand. *Tissue Antigens* **26**, 243-7 (1985).
15. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
16. Hindorff, L. *et al.*
17. Mells, G.F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat Genet* **43**, 329-32 (2011).
18. Fischer, A. *et al.* A novel sarcoidosis risk locus for Europeans on chromosome 11q13.1. *Am J Respir Crit Care Med* **186**, 877-85 (2012).
19. Qayyum, R. *et al.* A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans. *PLoS Genet* **8**, e1002491 (2012).
20. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
21. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
22. Barrett, J.C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* **40**, 955-62 (2008).
23. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* **42**, 295-302 (2010).
24. Hunt, K.A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* **40**, 395-402 (2008).
25. Hirota, T. *et al.* Genome-wide association study identifies eight new susceptibility loci for atopic dermatitis in the Japanese population. in *Nat Genet*, Vol. 44 1222-6 (United States, 2012).
26. Coulombe, F. & Behr, M.A. Crohn's disease as an immune deficiency? *Lancet* **374**, 769-70 (2009).
27. Casanova, J.L. & Abel, L. Revisiting Crohn's disease as a primary immunodeficiency of macrophages. *J Exp Med* **206**, 1839-43 (2009).
28. Pierce, E.S. Where are all the Mycobacterium avium subspecies paratuberculosis in patients with Crohn's disease? *PLoS Pathog* **5**, e1000234 (2009).

FIGURE LEGENDS

Figure 1 Forest plots of new loci.
ORs, (odds ratio), presented with their 95% confidence intervals within the square brackets. P-het, P-value from Cochran's Q heterogeneity test after Bonferroni correction for multiple testing.

Figure 2 Regional association plots of new loci.
The association results ($-\text{Log}_{10}$(P-value)) of SNPs from the discovery analysis (Stage 1) were shown against their map positions (NCBI build 37).a, 1q32.3 (near *BATF3*); b,

5p14.3 (near *CDH18*); c, 9q32 (near *DEC1*); d, 10q21.3 (near *ZNF365* and *EGR2*); e, 11q13.1 (near *CCDC88B*); f, 16p13.13 (near *CIITA* and *SOCS1*) . Validated SNP is labeled as a purple diamond.

Figure 3 Association plot of the MHC region
Association results ($-\text{Log}_{10}$(P-value)) of SNPs, amino acid polymorphisms and classical HLA alleles from the discovery analysis (Stage 1) on chromosome 6p31-32. a, Unconditional association results ; b, association resultsafter conditioning on rs9271100. All map coordinates are based on NCBI build 37.

Online Methods

Subjects. In the discovery stage (Stage 1), two independent studies were evaluated. The first study consists of 706 individuals with leprosy, 1,225 healthy controls and 4,362 immune disease-related subjects as population controls, all of northern Chinese Han descent as described in our previous GWAS publication[9,10]. The second study consists of 955 individuals with leprosy and 1,040 controls recruited from China between 2006-2011, including 436 cases and 533 controls of Chinese Han from northern China, 289 cases and 305 controls of Chinese Han from southern China, 230 cases and 202 controls of Chinese Chuang and other minorities from southern China.

Two independent sample series were used in the validation stages. In the first validation stage (Stage 2), 2,761 cases and 3,038 controls of Chinese Han were recruited from northern China. In the second replication phase (Stage 3), we recruited samples from multiple regions and ethnic groups in China (Supplementary Figure 4), including 277 cases and 2,626 controls of northern Han, 1,494 cases and 1,474 controls of southwest Han, 418 cases and 306 controls of southeast Han, 418 cases and 395 controls of western Han and 1,397 cases and 1,666 controls of ethnic minorities from southern China. In total, there were 6,765 cases and 9,505 controls used in the validation stages.

The cases and controls were recruited using the uniform criteria with written informed consent and matched regarding to ethnic origin and geographic area. The clinical diagnoses of all the leprosy cases were consistent with the criteria in our previous study[7,8,9]. All the controls used in the validation stage were healthy individuals without leprosy, autoimmune, systemic disorders and family history of

leprosy (including first-, second- and third-degree relatives). The study was approved by the institutional IRB committees at the Shandong Provincial Institute of Dermatology and Venereology, Shandong Academy of Medical Science. Supplementary Table 1 summarizes all the samples used in the three stages.

Genotyping and quality control in the discovery stage. Genotyping and quality control (QC) of the first independent GWAS study has been described previously[9,10].The second independent study was genotyped using Illumina Human 660K-Quad Bead Chips. All the genotyping experiment was conducted at Genergy Bio Technology (Shanghai) Co. Ltd. at Shanghai, China, according to the manufacturer's instructions.

SNP and sample QC criteria for the second independent study were similar with the QC filters in the first study[9,10]. SNP QC was done based on the following criteria: CNV and intensity-only SNPs (95,806 SNPs), SNPs located in the idiochromosome (16,703 SNPs), call-rate < 90% (1,271 SNPs), minor allele frequency (MAF) < 1% in cases and controls (68,327 SNPs), significant deviation from Hardy-Weinberg equilibrium (HWE) in the controls ($P< 1\times10^{-8}$; 998 SNPs) and having undetermined clusters (2 SNPs), were all removed. Finally, a total of 467,552 genotyped SNPs overlapping with the first independent dataset were used as a basis for imputation and genome-wide association analysis.

Samples were assessed by checking their call-rate (6 samples with call-rate < 96% were removed), potential relatedness among samples using pairwise identity-by-state (IBS)-based method implemented in PLINK[29] v1.07 (remove one of the pairs with lower call-rate when detected to have a 1st degree (65 samples were removed) or 2nd degree (107 samples were removed) familial relationships), and based on principal components analysis (PCA) whether they are within sample population outliers (45 samples removed) or ancestry/HapMap-based outliers (5 samples removed). PCA with HapMap samples were based on 206 samples drawn from Yoruba in Ibadan, Nigeria (YRI) (57), Japanese in Tokyo, Japan (JPT) (44), Chinese Han in Beijing, China (CHB) (45), and Utah residents with ancestry from Northern and Western

Europe (CEU) (60). The final number of samples evaluated for the first independent study is 706 cases and 5,587 controls, whereas for the second independent study there were a total of 842 cases and 925 controls after applying the QC filters.

Phasing and imputation. Phasing was performed separately for the first independent study and each of the three ethnic groups of the second study. Phasing was performed using SHAPEIT[30,31] version 2, based on the 467,552 common SNPs across the studies, while imputation was performed by using the IMPUTE[32,33] version 2.2.2 and the reference panel was based on 1000 Genomes Phase I integrated variant set v3 (release March 2012) in NCBI build 37 (hg19) coordinates. Post-imputation QC includes: exclusion of insertions-deletions (indels) and structural variations, SNPs with MAF <5%, SNPs with imputation certainty (info score) of < 0.8 and SNPs with significant deviation from HWE in the controls ($P< 1\times10^{-5}$). Finally, a total of 4,109,619 imputed SNPs and 467,552 genotyped SNPs that were common between studies were used in the association analysis.

Statistical analysis. Association analysis was performed separately for the two independent studies in the discovery stage. Genotyped SNPs were converted to genotype dosages to be analyzed together with the genotype dosages of imputed SNPs. Association analysis was performed in SNPTEST[34] version 2.4.1 by using the frequentist test under the additive model. SNPTEST is a well established program for single SNP analysis of genotype dosages for imputed data. Underlying population structure was assessed using PCA and in each study we included the first 5 principal components as covariates in the association model to account for population stratification.

Fixed-effects meta-analysis of two independent studies in the discovery stage was performed using the inverse variants method implemented in META[35] version 1.3.2, where Cochran's Q statistics P-value was also obtained. As this program allows the adjustment of genomic inflation lambda (λ_{GC}) within each study prior to the

meta-analysis, we specified the λ_{GC} for each study as follows, 1.13 for the first independent study and 1.02 for the second independent study.

To control the impact of population stratification in the replication and combined analysis, we matched cases and controls in terms of ethnic and geographic origins as independent validation samples. In the Stage 2 validation analysis, all of the cases and controls were Northern Chinese Han. In the stage 3, the samples were from five regions or subpopulations: Northern Chinese Han, Southwest Chinese Han, Southeast Chinese Han, Western Chinese Han and Southern Chinese minorities. The matched sample of each region or subpopulation was treated as "independent study" in the validation and combined analysis. Geographic matching has been previously demonstrated as a good proxy for genetic matching due to a high degree of correlation between subpopulation genetic and geographic structures[36]. Log-additive association test of the SNPs in the validation stages was performed using PLINK[29] v1.07. Finally, the combined analysis of 8,313 cases and 16,017 controls was done by a fixed-effects meta-analysis. Assessment of heterogeneity across studies was by evaluating Cochran's Q statistics P-value (Bonferroni-corrected heterogeneity P-values <0.05 were considered significant). A threshold of $P < 5\times10^{-8}$ was considered genome-wide significant. LocusZoom standalone version 1.3[37] tool was used to generate regional association plots of each locus, centered on the top SNP.

Conditional logistic regression was performed in either SNPTEST (for discovery datasets) or PLINK (for replication datasets), followed by meta-analysis in PLINK to assess the combined effect of the top SNP against the association in each locus.

Pairwise interaction analysis was performed on all confirmed associations in all samples combined from three stages (8,313 cases and 16,017 controls). In total there were 153 pairwise interaction tests among 18 SNPs which was performed using logistic regression and likelihood ratio tests. P-values for the interactions were calculated by likelihood ratio tests to compare the two models with and without the interaction term, where SNP1, SNP2, and study variables were included in the model as covariates. Bonferroni threshold for significance was set at P = 0.05/153 tests = 0.00033.

Haplotype analysis was performed to verify the independence of the associations previously reported in other diseases/traits against the leprosy associations within the same locus. For this analysis we used all samples from the discovery stage, but excluding 4,362 immune disease-related subjects. This analysis was done in R after phasing haplotypes using PHASE v.2.1.1 program[38,39]. Haplotype analysis is not performed on SNPs that are in high LD ($r^2 > 0.8$).

HLA allele imputation and association analysis. This analysis excluded the additional 4,362 samples from other immune-related disease subjects to avoid statistical artifact (See Supplementary Table 1 for sample information). Imputation of classical HLA alleles was based on a reference panel of 178 phased haplotypes from the CHB and JPT HapMap samples. This panel comprised dense SNP data and HLA allele types at 2-digit and 4-digit resolution for HLA class I (HLA-A, B, C) and II (DQA1, DQB1, DRB1), whereas amino acid variants (based on codons) were coded as present/absent by following the standard definitions from the EMBL-EBI Immunogenetics HLA database[40]. Imputation was performed using Beagle[41], following a previously described procedure[42].

Association analysis was performed in PLINK by comparing the frequencies of the best-guessed genotypes and allelic dosages in cases and controls using logistic regression assuming an additive model of inheritance. We used the first five principal components of each dataset to correct for population stratification. The final results were meta-analyzed using fixed-effects method. Consistency between the best-guessed and allelic dosage results were checked before presenting the analysis based on best-guessed genotypes.

Accuracy of rs9271100 and HLA-DRB1 imputations. To measure the accuracy of rs9271100 imputation, we randomly chose 600 cases and 600 controls of the discovery dataset and genotype the SNP using a pre-designed TaqMan assay (assay ID: C_4297481_10). The original top SNP in the MHC region, rs9271011, did not have a predesigned TaqMan probe available and the customized probe failed to produce SNP

intensities with distinguishable clusters. Hence, an assay for rs9271100, a SNP in high LD (r^2 = 0.89) with rs9271011 was used instead. Randomization of the samples was done using R software 'sample' command, taking an input of our sample IDs and sample number to select. This was done separately for cases and controls. Comparing the genotype results of the SNP from imputation and TaqMan genotyping revealed 97.3% concordance, showing high accuracy for the SNP imputation.

To measure the accuracy of the HLA-DRB1 imputation, we performed sequencing-based HLA typing in 466 Chinese samples that had been imputed. We chose these samples by the same randomization method mentioned above. Primer design for exon 2 of HLA-DRB1 was prepared according to the PCR-based barcoding method as previously described[43]. Extracted DNA in each sample was amplified by long range PCR for the amplicon spanning exon 2 of HLA-DRB1 gene. Individual libraries of each amplicon were prepared by enzymatic fragmentation, end repair, adenylation and ligation of indexed adaptors. The 12 indexed libraries are pooled prior to paired-end 2x250 bp sequencing on an illumina MiSeq. Sequence reads were aligned to the HLA region using Omixon target software with default settings and a minimal coverage of 5X. We calculated the Pearson's correlation coefficient (*r*) between genotyped and imputed allele dosage in the samples. Finally, we calculated the accuracy of both 2-digit and 4-digit HLA allele resolutions with the same method as described in the published manuscript[44] (Supplementary Table 6).

SNP selection for the validation stages. For SNPs within previously reported loci[3,4,9,10], we selected all of the reported genome-wide and suggestive associations by choosing either the reported SNP or a proxy SNP with perfect LD (r^2 and D' $\geq$ 0.99) to the reported SNP. Besides, we also selected SNPs from a number of known loci showing independent associations (P after conditional analysis on the reported association < 5×10^{-4}).

For SNPs within suggestive new loci, each independent locus was determined through conditional analysis within 1 MB of the most significant SNP. The locus was considered independent if the P-value after conditional analysis remained to be <

1×10^{-4} and hence the most significant SNP in each independent locus or its surrogate ($r^2 > 0.9$) is considered for replication. We also considered selecting independent loci whose top SNP have P-values between 5×10^{-4} and 1×10^{-4} and whose GRAIL P-values are < 0.05. The QQ plot in Supplementary Figure 2 shows where the statistics start to lift-off from the expected null distribution, i.e. between –logP of 3 and 4, which becomes the statistical threshold of our SNP selection ($P < 5\times10^{-4}$). At this threshold and a sample size of 1,548 cases and 6,512 controls in our discovery set we have calculated that we will have >80% power to detect an odds ratio of 1.4 at a minor allele frequency as low as 5%.

Genotyping analysis and quality control in the validation study. SNP genotyping for the validation stages were conducted at Shandong Provincial Key Lab for Dermatovenereology, Jinan, China using the Sequenom MassArray system (Sequenom inc)) and TaqMan Custom Genotyping Assay in a 7900 HT Fast Real-Time PCR System (Applied Biosystems) according to manufacturer's instructions. Ninety-one SNPs were selected for stage 2, but three SNPs were unsuccessfully genotyped due to the following reasons, two SNPs were rejected during the design process and one SNP had a bad genotyping cluster. Quality control measures were employed as follows: SNPs with undetermined clusters and SNPs with call rate < 90% were excluded. Samples with call-rate < 95% were excluded.

Prioritization of candidate genes through *in silico* analysis. Both protein-protein interaction (PPI) and pathway enrichment analyses are based on the observation that genes causing phenotypically-similar diseases tend to lie close to one another in a network more likely than by chance alone. These methods have been widely used to help prioritize potential causal genes in validated associated regions that contain multiple genes[20,45-47]. Each gene within the LD block of each locus was scored based on their biological evidences implicated by our established SNP based on supporting evidences which includes the following: a) A gene will get a score if the risk SNP or any of the SNPs in LD with this SNP ($r^2 > 0.8$) in the Asians (based on 1000 Genomes

data) are classified as a missense or nonsense mutation based on dbSNP functional annotations; b) A gene will also get a score if they were found to show significant cis-expression quantitative trait loci (eQTL) and methylation quantitative trait loci (mQTL) effects ($P < 0.001$) based on the risk SNP or any of the SNPs in LD with this SNP ($r^2 > 0.8$). The datasets for cis-eQTL and mQTL include a recently published peripheral blood mononuclear cells eQTL meta-analysis[15] as well as studies available within the Sanger Genevar database[48], which includes cell- and tissue-specific eQTL analysis of lymphoblastoid cell lines, fibroblasts, T-cells, skin, and adipose; c) If a gene shows significance ($P < 0.05$) in prioritization based on PubMed text mining (GRAIL)[49] by using established and suggestive SNPs as input, it will be scored. Similarly, if the gene is prioritized in protein-protein interaction networks using DAPPLE ($P < 0.05$)[45] and molecular pathway analysis using MAGENTA[46] (FDR $q < 0.05$) and IPA top network genes (network $P < 1\times10^{-10}$), the gene will also be scored. The gene with the highest score will be selected as the candidate gene for that locus.

Gene Ontology Enrichment Analysis. We used GO Enrichment Analysis by Gene Ontology Consortium (AmiGO)[50] to test genes that are concordant (*RIPK2* and *LACC1*) and discordant (*IL1RL1*/*IL18RAP*, *IL12B*, *TNFSF15*, *CCDC88B*) with IBD based on Supplementary Table 9.

Methods-only References:

29. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
30. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).
31. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
32. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-70 (2011).
33. 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).
34. 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**, 906-13 (2007).
35. Liu, J.Z. *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* **42**, 436-40 (2010).
36. Chen, J. *et al.* Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am J Hum Genet* **85**, 775-85 (2009).
37. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
38. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978-89 (2001).
39. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449-62 (2005).
40. de Bakker, P.I. *et al.* A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* **38**, 1166-72 (2006).
41. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
42. Pereyra, F. *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* **330**, 1551-7 (2010).
43. Bentley, G. *et al.* High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens* **74**, 393-403 (2009).
44. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet* **44**, 291-6 (2012).
45. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* **7**, e1001273 (2011).
46. Segrè, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**(2010).
47. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
48. Yang, T.P. *et al.* Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* **26**, 2474-6 (2010).
49. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet* **5**, e1000534 (2009).
50. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25**, 288-9 (2009).