ORIGINAL ARTICLE

Genome-Wide Analysis of Protein-Coding Variants in Leprosy

Hong Liu[1,2,3,4,24], Zhenzhen Wang[1,2,3,24], Yi Li[5,6,24], Gongqi Yu[1,3], Xi'an Fu[1,3,7], Chuan Wang[1,3], Wenting Liu[5], Yongxiang Yu[1,2], Fangfang Bao[1,3], Astrid Irwanto[5], Jian Liu[1,2], Tongsheng Chu[1,2], Anand Kumar Andiappan[8], Sebastian Maurer-Stroh[9,10], Vachiranee Limviphuvadh[9], Honglei Wang[1,3,7], Zihao Mi[1,3], Yonghu Sun[1,3], Lele Sun[1,3], Ling Wang[5], Chaolong Wang[11], Jiabao You[1,2], Jinghui Li[1,2], Jia Nee Foo[5], Herty Liany[5], Wee Yang Meah[5], Guiye Niu[1,2,3], Zhenhua Yue[1,3,7], Qing Zhao[1,3,7], Na Wang[1,3,7], Meiwen Yu[12], Wenjun Yu[1,3,13], Xiujun Cheng[1,3,7], Chiea Chuen Khor[5], Kar Seng Sim[5], Tin Aung[14], Ningli Wang[15], Deyun Wang[16], Li Shi[17], Yong Ning[18], Zhongyi Zheng[19], Rongde Yang[20], Jinlan Li[21], Jun Yang[22], Liangbin Yan[12], Jianping Shen[12], Guocheng Zhang[12], Shumin Chen[1,2], Jianjun Liu[5,25] and Furen Zhang[1,2,3,4,7,13,23,25]

Although genome-wide association studies have greatly advanced our understanding of the contribution of common noncoding variants to leprosy susceptibility, protein-coding variants have not been systematically investigated. We carried out a three-stage genome-wide association study of protein-coding variants in Han Chinese, of whom were 7,048 leprosy patients and 14,398 were healthy control subjects. Seven coding variants of exome-wide significance were discovered, including two rare variants: rs145562243 in *NCKIPSD* ($P = 1.71 \times 10^{-9}$, odds ratio [OR] = 4.35) and rs149308743 in *CARD9* ($P = 2.09 \times 10^{-8}$, OR = 4.75); three low-frequency variants: rs76418789 in *IL23R* ($P = 1.03 \times 10^{-10}$, OR = 1.36), rs146466242 in *FLG* ($P = 3.39 \times 10^{-12}$, OR = 1.45), and rs55882956 in *TYK2* ($P = 1.04 \times 10^{-6}$, OR = 1.30); and two common variants: rs780668 in *SLC29A3* ($P = 2.17 \times 10^{-9}$, OR = 1.14) and rs181206 in *IL27* ($P = 1.08 \times 10^{-7}$, OR = 0.83). Discovered protein-coding variants, particularly low-frequency and rare ones, showed involvement of skin barrier and endocytosis/phagocytosis/autophagy, in addition to known innate and adaptive immunity, in the pathogenesis of leprosy, highlighting the merits of protein-coding variant studies for complex diseases.

Journal of Investigative Dermatology (2017) **137**, 2544–2551; doi:10.1016/j.jid.2017.08.004

INTRODUCTION

Leprosy is an ancient infectious disease caused by *Mycobacterium leprae* (*M. leprae*) that affects the skin and peripheral nerves. Although treatable, it remains as a major cause of disability and social stigma in many parts of the world, particularly in developing countries, with 210,758 new cases reported in 2015 (World Health Organization, 2016). To understand the genetic basis of leprosy susceptibility, candidate gene-based association analysis and genome-wide linkage and association studies (GWASs) have been performed, and 21 common risk variants were discovered, showing the involvement of the innate and adaptive immune responses (Liu et al., 2012, 2013, 2015; Misch et al., 2010; Wang et al., 2016; Wong et al., 2010; Zhang et al., 2009, 2011). However, these common risk variants are largely located within noncoding regions. Protein-coding variants, particularly low-frequency and rare ones, have not been investigated systematically, although these variants may

[1] *Shandong Provincial Institute of Dermatology and Venereology, Shandong Academy of Medical Sciences, Jinan, China;* [2] *Shandong Provincial Hospital for Skin Diseases, Shandong University, Jinan, China;* [3] *Shandong Provincial Key Lab for Dermatovenereology, Jinan, China;* [4] *Shandong Provincial Medical Center for Dermatovenereology, Jinan, China;* [5] *Human Genetics, Genome Institute of Singapore, A*STAR, Singapore;* [6] *Computational Sciences, The Jackson Laboratory, Farmington, Connecticut, USA;* [7] *School of Medicine, Shandong University, Jinan, China;* [8] *Singapore Immunology Network, Agency for Science, Technology and Research Singapore, Singapore;* [9] *Biomolecular Function Discovery Division, Bioinformatics Institute, A*STAR, Singapore;* [10] *School of Biological Sciences, Nanyang Technological University, Singapore;* [11] *Computational and Systems Biology, Genome Institute of Singapore, A*STAR, Singapore;* [12] *Institute of Dermatology, Chinese Academy of Medical Sciences and Peking Union Medical College, Nanjing, China;* [13] *School of Medicine and Life Science, University of Jinan-Shandong Academy of Medical Sciences, Jinan, Shandong, China;* [14] *Singapore National Eye Centre, Glaucoma Department, Singapore;* [15] *Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing Ophthalmology and Visual Science Key Lab, Beijing, China;* [16] *Department of Otolaryngology, National University of Singapore, Singapore;* [17] *Department of Otolaryngology, the Second Affiliated Hospital, Shandong University, Jinan, China;* [18] *Sichuan Provincial Institute of Dermatology, Sichuan, China;* [19] *Honghe Institute of Dermatology, Honghe, Yunnan, China;* [20] *Wenshan Institute of Dermatology, Wenshan, Yunnan, China;* [21] *Guizhou Provincial Center for Disease Control and Prevention, Guizhou, China;* [22] *Yunnan Provincial Center for Disease Control and Prevention, Yunnan, China; and* [23] *National Clinical Key Project of Dermatology and Venereology, Jinan, China*

[24] *These authors contributed equally to this work.*

[25] *These authors share senior authorship.*

Correspondence: Furen Zhang, Shandong Provincial Institute of Dermatology and Venereology, 27397 Jingshi Lu, Jinan 250022, Shandong Province, People's Republic of China. E-mail: zhangfuren@hotmail.com

Abbreviations: GO, gene ontology; GWAS, genome-wide association study; MAF, minor allele frequency; OR, odds ratio; SNP, single nucleotide polymorphism

Received 15 May 2017; revised 9 July 2017; accepted 2 August 2017; accepted manuscript published online 24 August 2017

probably be the causal ones in most cases (Bodmer and Bonilla, 2008; Fearnhead et al., 2005; Frazer et al., 2009).

Here we performed a three-stage GWAS of protein-coding variants in leprosy. In the discovery stage, 40,491 coding variants were tested for single variant association in 1,648 leprosy patients and 2,318 healthy control subjects. We then validated and replicated our findings in four independent cohorts, totaling 5,400 leprosy patients and 12,080 healthy control subjects. In addition, we performed the biological and network analysis of seven genes identified in this study and 26 susceptibility genes implicated in the previous leprosy GWASs to provide more information on leprosy genetic pathogenesis.

RESULTS

Exome-wide discovery analysis

In the discovery stage (stage 1), we successfully genotyped 273,028 variants in 1,670 leprosy patients and 2,321 control individuals. Principal component analysis confirmed the Chinese ancestry of all participant samples, and a good genetic match between the patients and control subjects was observed (see Supplementary Figures S1, S2 online). After a series of sample and variant quality control filtering, 74,764 polymorphic coding variants were retained for association test. According to the result of the power analysis by Genetic Power Calculator (available at http://zzz.bwh.harvard.edu/gpc/cc2.html), the sample size of our discovery analysis was underpowered for detecting single variants with minor allele frequency (MAF) (in all samples) less than 0.1%, hence only 40,491 coding variants with MAF (in all samples) greater than 0.1% (38,068 nonsynonymous and 2,423 synonymous ones) were tested for single variant association in 1,648 leprosy patients and 2,318 healthy control subjects by GMMAT (Breslow and Clayton, 1993; Chen et al., 2016) software.

As expected, the exome-wide analysis showed strong and extensive associations within the major histocompatibility complex region (Figures 1, 2a). After removing all the variants within the major histocompatibility complex region, the quantile-quantile plot of the remaining single nucleotide polymorphisms (SNPs) showed a good fit with the expected null distribution (Figure 2b), indicating a minimal inflation of exome-wide association results due to population stratification ($\lambda_{GC} = 0.99$). Additional quantile-quantile plots of SNP sets with different MAF thresholds (>0.5%, >1%, and >5%) also showed consistent indication of minimal inflation (see Supplementary Figure S3 online).

The strongest association was obtained at rs3200405 ($P = 3.46 \times 10^{-40}$, odds ratio [OR] = 2.15) within *HLA-DRB5*, which is correlated to the previously reported GWAS SNP rs9271100 and HLA allele HLA-DRB1*15:01 (Liu et al., 2015; Rani et al., 1993; Wang et al., 2016) ($P_{\text{conditional}} = 0.95$, $r^2 = 0.74$, $D' = 0.97$; $P_{\text{conditional}} = 0.41$, $r^2 = 0.72$, $D' = 0.99$, respectively) (see Supplementary Table S1 online). Five coding variants (four common and one low frequency) with suggestive association ($P < 1.0 \times 10^{-3}$) were also found within the previously identified GWAS loci (Liu et al., 2012, 2013, 2015; Wang et al., 2016; Zhang et al., 2009, 2011). All four common coding variants were in high linkage disequilibrium with the previously reported SNPs, and only the low-frequency variant (rs76418789 in *IL23R*) showed independent association (see Supplementary Table S1). In addition, a moderate deviation of the quantile-quantile plot from the expected null distribution was observed at the tail of distribution (after removing all the SNPs within the major histocompatibility complex region), indicating the existence of associations (Figure 2, Supplementary Figure S3). There were 38 coding variants in, to our knowledge, previously unreported genes with suggestive association significance ($P < 1.0 \times 10^{-3}$).

To further investigate the role of low-frequency and rare coding variants, we carried out gene-based tests using SKAT-O (Lee et al., 2012b) and burden test methods (Lee et al., 2012a). However, both SKAT-O and burden test analyses failed to provide indication of true genetic association beyond expectation by chance (see Supplementary Figures S4, S5 online).

Validation analysis

A total of 39 independent coding variants were genotyped in an additional 3,169 leprosy patients and 9,814 healthy control subjects from the northern region of China (stage 2). Of the 34 successfully genotyped variants, 15 were found to show consistent associations between the discovery and validation samples ($P < 0.05$ in the validation samples and OR in the same direction as in the discovery stage), and six reached exome-wide significance ($P < 0.05/40{,}491 = 1.23 \times 10^{-6}$ based on Bonferroni correction, which is a conservative way to control for false positives as multiple tests are performed) (Duggal et al., 2008; Gibson, 2012) in the combined discovery and validation samples without evidence of heterogeneity: rs145562243 in *NCKIPSD* ($P = 1.44 \times 10^{-8}$, OR = 4.35), rs149308743 in *CARD9* ($P = 4.99 \times 10^{-10}$, OR = 4.75), rs76418789 in *IL23R* ($P = 6.28 \times 10^{-9}$, OR = 1.37), rs146466242 in *FLG* ($P = 1.44 \times 10^{-10}$, OR = 1.45), rs780668 in *SLC29A3* ($P = 2.89 \times 10^{-7}$, OR = 1.14), and rs181206 in *IL27* ($P = 5.92 \times 10^{-7}$, OR = 0.83) (Table 1, Supplementary Table S2 online). In addition, two coding variants, rs55882956 in *TYK2* ($P = 2.75 \times 10^{-5}$, OR = 1.29) and rs75746803 in *USP49* ($P = 3.45 \times 10^{-6}$, OR = 1.28) showed consistent associations between the discovery and validation samples but were barely below the exome-wide significance in the combined samples (Table 1, Supplementary Table S2).

Replication analysis

Eight nonsynonymous coding variants were further selected for replication analysis using the following criteria: (i) showing consistent association in the validation stage at *P*-value less than 0.005 and (ii) had a *P*-value less than 5×10^{-5} in the meta-analysis of the discovery and validation samples. The replication study (stage 3) was performed by using additional independent samples from three southern regions of China, totaling 2,231 leprosy patients and 2,266 healthy control subjects (see Supplementary Table S3 online). Five of the eight variants obtained significant associations in the replication samples alone (rs76418789 in *IL23R*, $P = 4.35 \times 10^{-3}$; rs146466242 in *FLG*, $P = 6.58 \times 10^{-3}$; rs55882956 in *TYK2*, $P = 1.15 \times 10^{-2}$; rs780668 in *SLC29A3*, $P = 2.05 \times 10^{-3}$; rs145562243 in *NCKIPSD*, $P = 3.16 \times 10^{-2}$). Out of the three nonsignificant variants, one (rs149308743 in *CARD9*, $P = 8.69 \times 10^{-1}$) is very rare, another (rs181206 in *IL27*, $P = 6.09 \times 10^{-2}$) is on the borderline, and the third variant (rs75746803 in *USP49*,

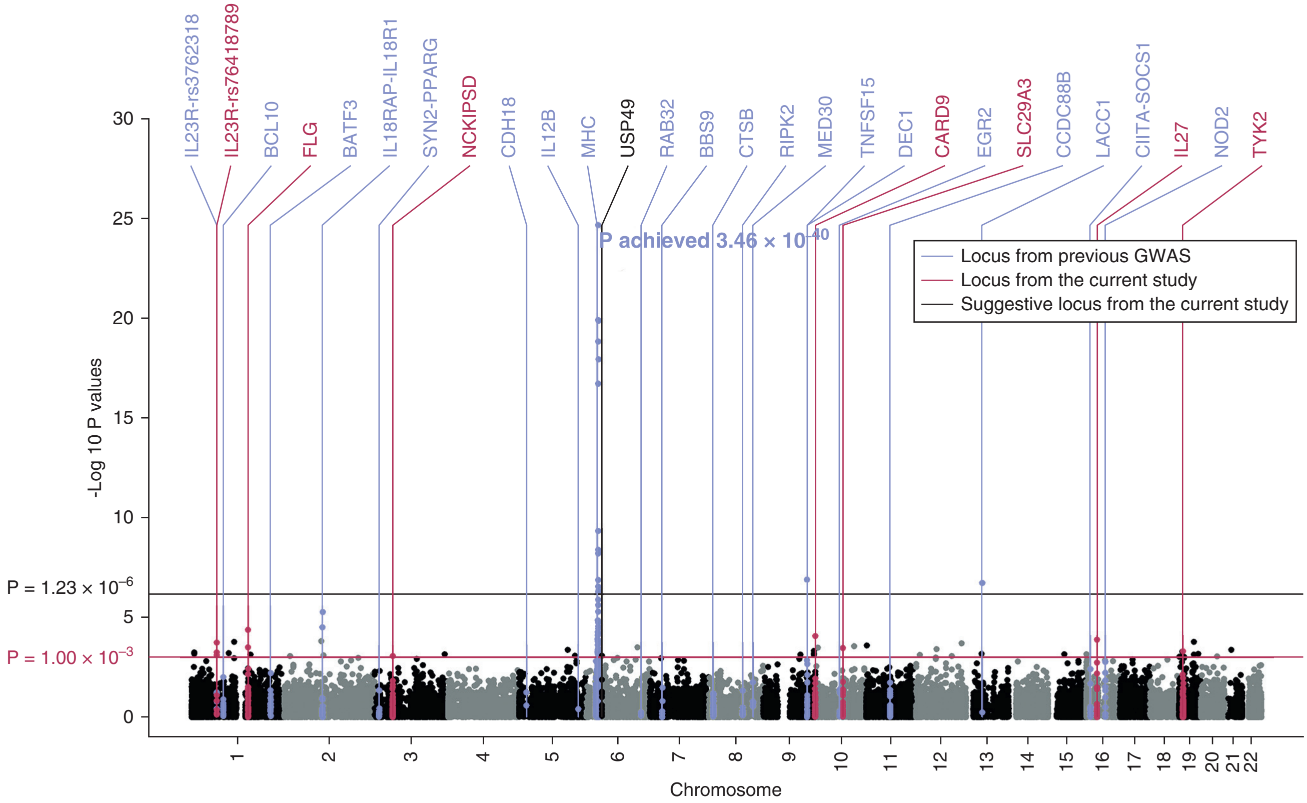

Figure 1. Manhattan plot of all the coding variants with minor allele frequency ≥ 0.001. Previously published loci with genome-wide significance ($P < 5 \times 10^{-8}$), exome-wide significant loci validated in the current study ($P < 1.23 \times 10^{-6}$), and exome-wide suggestive locus validated in the current study ($1.23 \times 10^{-6} < P < 1 \times 10^{-5}$) were shown in the manhattan plot. GWAS, genome-wide association study.

$P = 3.27 \times 10^{-1}$) showed inconsistent genetic effect across the three replication samples.

We then performed the meta-analysis of all the samples from the three stages, consisting of a total of 7,048 patients and 14,398 control subjects, using the Fisher combination method for rare variants and a fixed-effect model for the others. Seven (two rare, three low-frequency, and two common) coding variants showed consistent associations without significant evidence of heterogeneity and reached the exome-wide significance in the combined samples (Table 1). rs75746803 in *USP49* on chromosome locus 6p21.1 remained barely below exome-wide significance (MAF = 4.8%, $P = 3.57 \times 10^{-6}$, OR = 1.25) (Table 1, Figure 1, Supplementary Table S2, and Supplementary Figure S6 online).

We evaluated the impact of sex by performing the sex-adjusted meta association test across three stages. The unadjusted and sex-adjusted ORs of the two common and three low-frequency SNPs were similar (with <10% relative difference) (see Supplementary Table S4 online), indicating that the genetic effects of these variants are independent of the effects of sex (Aschengrau and Seage, 2006). The two rare variants were extremely rare and almost absent in southern

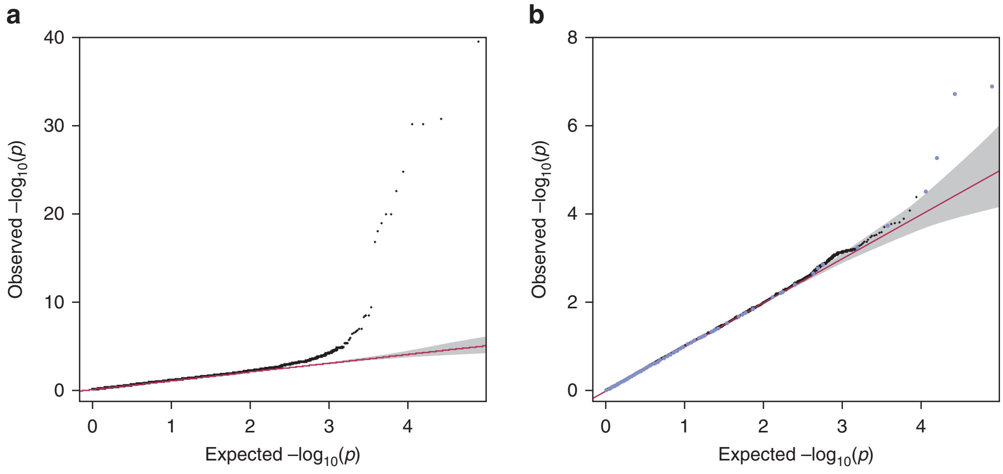

Figure 2. Quantile-quantile (QQ) plot of the association. (**a**) QQ plot ($\lambda_{GC} = 1.01$) of all the coding SNPs with MAF ≥ 0.001. *P*-values were from GMMAT score tests. (**b**) QQ plot ($\lambda_{GC} = 0.99$) of all the coding SNPs with MAF ≥ 0.001. *P*-values were from GMMAT score tests. SNPs within the major histocompatibility complex region were removed, and SNPs in the previously reported GWAS loci are colored as purple. One SNP, rs3764147, which is a known genome-wide association study SNP with *P*-value less than 10^{-20} in this study, was removed for ease of observation of the association signal. MAF, minor allele frequency; SNP, single nucleotide polymorphism.

Table 1. Associations results of eight variants that were studied in all the discovery, validation and replication samples

Variant/AA	Alleles[1]	Gene	Function	Type[2]	Stage	F_A[4]	F_U[4]	OR	L95	U95	*P*[5]	Phet
rs145562243 (chr3:48719549) AA: R176Q	T/C	*NCKIPSD*	missense	rare[4]	1. Discovery	0.0039	0.0004	**8.92**	**2.26**	**35.13**	**8.78E-04**	**NA**
					2. Validation	0.0057	0.0015	3.93	2.28	6.79	1.31E-06	NA
					3.Combined 1+2[3]	**NA**	**NA**	**4.35**	**2.58**	**7.35**	**1.44E-08**	**0.30**
					4. Replication1	0/1/882	0/1/793	NA	NA	NA	1.00E+00	NA
					5. Replication2	0/0/788	0/0/402	NA	NA	NA	1.00E+00	NA
					6. Replication3	0/1/445	0/0/778	NA	NA	NA	3.64E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**NA**	**NA**	**NA**	**3.16E-02**	**NA**
					8. All combined[3]	**NA**	**NA**	**NA**	**NA**	**NA**	**1.71E-09**	**NA**
rs149308743 (chr9:139258965) AA:R494H	T/C	*CARD9*	missense	rare[4]	1. Discovery	0.0055	0.0006	**8.52**	**2.79**	**26.05**	**8.56E-05**	**NA**
					2. Validation	0.0050	0.0012	4.25	2.47	7.29	1.77E-07	NA
					3. Combined 1+2[3]	**NA**	**NA**	**4.75**	**2.87**	**7.86**	**4.99E-10**	**0.31**
					4. Replication1	0/7/889	0/2/870	NA	NA	NA	1.80E-01	NA
					5. Replication2	0/1/825	0/3/578	NA	NA	NA	3.13E-01	NA
					6. Replication3	0/1/493	0/2/796	NA	NA	NA	1.00E+00	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**NA**	**NA**	**NA**	**8.69E-01**	**NA**
					8. All combined[3]	**NA**	**NA**	**NA**	**NA**	**NA**	**2.09E-08**	**NA**
rs76418789 (chr1:67648596) AA: G149R	A/G	*IL23R*	missense	low freq	1. Discovery	0.0610	0.0432	**1.45**	**1.16**	**1.81**	**5.59E-04**	**NA**
					2. Validation	0.0632	0.0479	1.34	1.19	1.52	2.37E-06	NA
					3. Combined 1+2[3]	**NA**	**NA**	**1.37**	**1.23**	**1.53**	**6.28E-09**	**0.55**
					4. Replication1	0.0628	0.0505	1.26	0.94	1.68	1.18E-01	NA
					5. Replication2	0.0624	0.0417	1.53	1.07	2.18	1.87E-02	NA
					6. Replication3	0.0536	0.0445	1.22	0.84	1.76	2.90E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**1.32**	**1.07**	**1.63**	**4.35E-03**	**0.62**
					8. All combined[3]	**NA**	**NA**	**1.36**	**1.24**	**1.49**	**1.03E-10**	**0.84**
rs146466242 (chr1:152275298) AA: K4022*	A/T	*FLG*	stop-gained	low freq	1. Discovery	0.0558	0.0354	**1.59**	**1.26**	**2.00**	**4.17E-05**	**NA**
					2. Validation	0.0552	0.0400	1.40	1.23	1.60	4.39E-07	NA
					3. Combined 1+2[3]	**NA**	**NA**	**1.45**	**1.29**	**1.62**	**1.44E-10**	**0.34**
					4. Replication1	0.0320	0.0186	1.77	1.14	2.76	1.18E-02	NA
					5. Replication2	0.0249	0.0186	1.35	0.79	2.31	2.71E-01	NA
					6. Replication3	0.0243	0.0188	1.28	0.75	2.17	3.59E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**1.49**	**1.09**	**2.04**	**6.58E-03**	**0.60**
					8. All combined[3]	**NA**	**NA**	**1.45**	**1.31**	**1.61**	**3.39E-12**	**0.74**
rs75746803 (chr6:41773726) AA:W332C	G/C	*USP49*	missense	low freq	1. Discovery	0.0664	0.0481	**1.40**	**1.14**	**1.73**	**8.34E-04**	**NA**
					2. Validation	0.0612	0.0500	1.24	1.10	1.40	6.50E-04	NA
					3. Combined 1+2[3]	**NA**	**NA**	**1.28**	**1.15**	**1.43**	**3.45E-06**	**0.29**
					4. Replication1	0.0337	0.0356	0.95	0.66	1.36	7.63E-01	NA
					5. Replication2	0.0339	0.0266	1.29	0.82	2.03	2.72E-01	NA
					6. Replication3	0.0436	0.0352	1.25	0.83	1.87	2.83E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**1.12**	**0.68**	**1.86**	**3.27E-01**	**0.48**
					8. All combined[3]	**NA**	**NA**	**1.25**	**1.14**	**1.38**	**3.57E-06**	**0.46**
rs55882956 (chr19:10469919) AA:R703W	A/G	*TYK2*	missense	low freq	1. Discovery	0.0529	0.0363	**1.51**	**1.18**	**1.92**	**4.96E-04**	**NA**
					2. Validation	0.0464	0.0383	1.22	1.06	1.40	4.89E-03	NA
					3. Combined 1+2[3]	**NA**	**NA**	**1.29**	**1.14**	**1.46**	**2.75E-05**	**0.12**
					4. Replication1	0.0431	0.0262	1.70	1.17	2.48	5.58E-03	NA
					5. Replication2	0.0423	0.0374	1.14	0.77	1.67	5.12E-01	NA
					6. Replication3	0.0333	0.0276	1.22	0.76	1.94	4.07E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**1.35**	**1.04**	**1.75**	**1.15E-02**	**0.30**
					8. All combined[3]	**NA**	**NA**	**1.30**	**1.17**	**1.45**	**1.04E-06**	**0.30**
rs780668 (chr10:73111408) AA:S158F	T/C	*SLC29A3*	missense	common	1. Discovery	0.4648	0.4213	**1.19**	**1.08**	**1.31**	**3.42E-04**	**NA**
					2. Validation	0.4619	0.4338	1.12	1.06	1.18	1.36E-04	NA
					3. Combined 1+2[3]	**NA**	**NA**	**1.14**	**1.08**	**1.19**	**2.89E-07**	**0.28**
					4. Replication1	0.4723	0.4320	1.17	1.03	1.34	1.86E-02	NA
					5. Replication2	0.4730	0.4541	1.08	0.93	1.26	3.31E-01	NA
					6. Replication3	0.5062	0.4658	1.17	1.00	1.37	5.03E-02	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**1.14**	**1.04**	**1.25**	**2.05E-03**	**0.67**
					8. All combined[3]	**NA**	**NA**	**1.14**	**1.09**	**1.19**	**2.17E-09**	**0.74**

(*continued*)

Table 1. Continued

Variant/AA	Alleles[1]	Gene	Function	Type[2]	Stage	F_A[4]	F_U[4]	OR	L95	U95	P[5]	Phet
rs181206 (chr16:28513403) AA:L119P	G/A	*IL27*	missense	common	1. Discovery	0.1171	0.1497	**0.76**	**0.65**	**0.88**	**1.30E-04**	**NA**
					2. Validation	0.1175	0.1346	0.85	0.78	0.93	4.68E-04	NA
					3. Combined 1+2[3]	**NA**	**NA**	**0.83**	**0.77**	**0.89**	**5.92E-07**	**0.16**
					4. Replication1	0.0809	0.0969	0.82	0.65	1.04	9.42E-02	NA
					5. Replication2	0.0720	0.0769	0.94	0.71	1.24	6.38E-01	NA
					6. Replication3	0.0617	0.0721	0.85	0.62	1.17	3.15E-01	NA
					7. Combined 4+5+6[3]	**NA**	**NA**	**0.86**	**0.71**	**1.04**	**6.09E-02**	**0.77**
					8. All combined[3]	**NA**	**NA**	**0.83**	**0.78**	**0.89**	**1.08E-07**	**0.61**

Abbreviations: AA, Amino acid change; chr, chromosome; F_A, minor allele frequency in patients; F_U, minor allele frequency in control subjects; freq, frequency; L95, lower 95% confidence limit; NA, not applicable; OR, odds ratio with respect to the minor allele; Phet, *P*-value of heterogeneity; U95, upper 95% confidence limit.

[1] Minor allele/major allele.

[2] Common: minor allele frequency $\geq$ 5% in patients; low frequency: 1% $\leq$ minor allele frequency $<$ 5% in patients; rare: minor allele frequency $<$ 1% in control subjects.

[3] Indicates results from fixed-effects meta-analysis.

[4] For rare variants in the three replications, F_A is count for A1A1/A1A2/A2A2 in patients; F_U is count for A1A1/A1A2/A2A2 in control subjects. Because of the large deviations of the estimated OR in the small samples, we reported only their association *P*-values, which were calculated by Fisher exact test.

[5] Combined *P*-values were calculated by fixed-effect inverse-variance method if ORs were estimated; otherwise by fixed-effect *z* statistics combination method.

Chinese (stage 3), and as a result, their ORs could not be evaluated in the replication samples and are not included in Supplementary Table S4.

Conditional and haplotype analyses

To verify whether the low-frequency coding variant rs76418789 is independent of the previously reported GWAS SNP rs3762318 (Zhang et al., 2011) within *IL23R*, we performed conditional and haplotype analyses of the two variants using the overlapping samples between the current study and the datasets in the previously published GWASs (Liu et al., 2015; Zhang et al., 2009) (including 3,019 patients and 5,767 control subjects of northern Chinese Han ethnicity). Conditioning on rs3762318 had a minimal impact on the association at rs76418789 ($P_{\text{conditional}} = 9.06 \times 10^{-4}$, $\text{OR}_{\text{conditional}} = 1.26$ vs. $\text{P} = 1.04 \times 10^{-4}$, $\text{OR} = 1.30$) (see Supplementary Table S5 online). Consistently, the haplotype analysis showed that the AA haplotype (i.e., carrying the two risk alleles of the two variants) confers significantly higher risk than the AG haplotype (i.e., carrying only the risk allele of rs3762318), confirming further the independent risk effects of these two SNPs within *IL23R* (see Supplementary Table S6 online). We also probed the association evidence for the two common coding variants (rs780668 in *SLC29A3* and rs181206 in *IL27*) in our previously reported GWAS dataset (1,548 patients and 6,512 control subjects) (Liu et al., 2015) and found that both rs780668 and rs181206 were filtered out in the imputation stage of our previous GWAS because of low imputation information score. All the other low-frequency and rare coding variants identified in this study are either outside the linkage disequilibrium blocks of or independent from the previously reported GWAS SNPs.

Risk explained by identified variants

Commutatively, the seven coding variants could explain 1.20% of risk, which is much lower than the 17.01% of risk explained by the 20 GWAS loci (22 independent GWAS SNPs). Of the seven coding risk variants, only two showed strong genetic effect (OR > 4), but they are extremely rare (MAF in control subjects < 0.1%), and as a result, their contribution to the overall risk of leprosy is very small. The other five variants all showed moderate genetic effect. Together, 18.21% of leprosy risk can be explained by all 29 genetic risk variants that have been discovered so far (see Supplementary Table S7 online).

Biological and network analysis

All seven coding risk variants discovered in this study are nonsynonymous and were found to be deleterious by either *SIFT* tool (available from http://sift.bii.a-star.edu.sg), *PolyPhen-2* tool (available from http://genetics.bwh.harvard.edu/pph2/bgi.shtml), or protein structural analysis (see Supplementary Table S8 and Supplementary Figures S7–S10 online). Except for rs76418789 within the known leprosy gene *IL23R* (Zhang et al., 2011), all coding risk variants are within genes that have not been reported before to our knowledge (Table 1 and Figure 1), justifying the value of investigating exome-wide coding variants.

We performed gene ontology (GO) enrichment analysis of seven genes identified in this study and of those seven genes together with 26 susceptibility genes implicated in the previous leprosy GWASs. The analysis showed that the regulation of immune response (GO: 0050776, $P = 5.00 \times 10^{-6}$) and the immune response (GO: 0006955, $P = 3.10 \times 10^{-14}$) are the most enriched GO term for the two gene sets, respectively (Liu et al., 2012, 2013, 2015; Wang et al., 2016; Zhang et al., 2009, 2011) (see Supplementary Table S9 and Supplementary Figure S11 online). According to the GO hierarchy, the GO: 0050776 is annotated as the "Regulate" of the GO: 0006955, indicating that the leprosy genes discovered in this are closely connected (in terms of molecular function) with the ones implicated by the previous GWASs.

An integrated gene network analysis of all 33 susceptibility genes (identified by previous GWASs and these analyses) using GeneMANIA (Mostafavi et al., 2008) in Cytoscape tool (http://www.cytoscape.org/) was performed. The analysis showed a highly interactive gene network where all the significant subnetworks are found to be associated with specific functions and components of immunity (see Supplementary Figure S12 online). For example, a group of subnetworks that contains susceptibility genes *HLA-DRB1, HLA-DQB1, NOD2, IL27, IL23R,* and *IL12B* plays an important role in the regulation of T-cell activation ($P = 7.09 \times 10^{-9}$), adaptive immune response ($P = 7.68 \times 10^{-8}$) and IFN-γ production ($P = 2.92 \times 10^{-6}$) (see Supplementary Table S10 online and Supplementary Figure S12). Another cluster of subnetworks involving *BCL10, CARD9, RIPK2, NOD2, SOCS1, IL12B, PPARG, CTSB,* and *TYK2* are related to the regulation of innate immune response ($P = 7.09 \times 10^{-9}$) and positive regulation of NF-κB transcription factor activity ($P = 8.95 \times 10^{-6}$) (see Supplementary Table S10 and Supplementary Figure S12).

DISCUSSION

By carrying out a three-stage exome-wide association study of protein-coding variants in the Chinese population, we have discovered seven nonsynonymous risk variants for leprosy, including five low-frequency and rare ones. These coding risk variants showed six, to our knowledge previously unreported, disease susceptibility genes for leprosy. The individual genetic effects (ORs) of low-frequency and rare coding risk variants are stronger than common variants (identified by GWAS), although their contributions to risk are, individually or cumulatively, weaker (because of their low frequency) than the common ones, which did not support the hypothesis that protein-coding risk variants have strong genetic effects, and the "missing of heritability" of complex diseases (failed to be explained by GWAS loci) can be explained by the strong genetic effects of protein-coding variants.

Through the integrated gene network analysis of leprosy susceptibility genes discovered by the current protein-coding variant study and previous GWASs, we have also shown the involvement of both innate and adaptive immunity in leprosy. In particular, our study has highlighted the important roles of several specific functional components of immunity, such as the regulation of T activation, IFN-γ production, NF-κB transcription factor activity, and innate immune response. The highly integrated gene network implicated by leprosy disease genes also contains additional genes (see Supplementary Figure S12) whose involvements in leprosy have not been shown. These genes are strong candidates for future genetic study of leprosy.

FLG was not a part of the gene network, coinciding with non-immunity functionality of this gene shown by our literature review. The *FLG* gene encoded profilaggrin and filaggrin, which play a pivotal role in skin barrier function by affecting formed stratum corneum and water binding (Sandilands et al., 2009; Scott and Harding, 1986). It has long been believed that the upper respiratory tract is the primary route of infection for *M. leprae*, although few case reports of skin injury-induced leprosy lesions suggested the possibility of infection through skin (Abraham et al., 1998; Ghorpade, 2002). Together with the finding that filaggrin is not expressed in the normal upper airway (De Benedetto et al., 2008), our discovery of *FLG* as a disease susceptibility gene for leprosy provides strong genetic evidence that impaired skin barrier plays an important role in leprosy, and direct contact through skin might be an important route of infection for *M. leprae*.

Literature review also showed that *NCKIPSD* is not related to the immunity, although it has been included in the gene network. The protein of *NCKIPSD* belongs to the NCK family of adaptor proteins, major regulators of act in the cytoskeleton. It is a part of FCG receptor-dependent phagocytosis pathway and is implicated in many functional processes, such as assembly and maintenance of sarcomeres and stress fiber formation (Lim et al., 2001; Satoh and Tominaga, 2001). The rare variant rs145562243 (p.R176Q) was located in the proline-rich region (PRD) of the *NCKIPSD* N-terminus, whose overexpression was found to be related to abnormalities in vesicle formation and trafficking, leading to the defective endocytosis of FCG receptor (Oh et al., 2013). Together with the identification of *RAB32* (Zhang et al., 2011), which has been shown to be involved in autophagy and phagocytotic digestion of bacteria (Hirota and Tanaka, 2009; Seto et al., 2011; Spano and Galan, 2012), as well as *LRRK2* (Zhang et al., 2009) as an interacting partner of *RAB32* (Waschbusch et al., 2014) and related to cell death due to autophagic dysfunction and mitochondrial damage (Alegre-Abarrategui et al., 2009; Plowey et al., 2008), the discovery of *NCKIPSD* highlights the potential involvement of endocytosis/phagocytosis/autophagy in host defense against *M. leprae* infection.

We acknowledge that the results from the Gene Mania analysis or GO term enrichment analysis are broad in terms of functional inference, and future integrated analysis with other types of data, for example, gene expression profiles or epigenetic data, may provide more specific functional impacts of the reported variants.

MATERIALS AND METHODS

Study subjects

We performed a three-stage case-control analysis for this study. Stage 1 and stage 2 included 16,974 individuals from the northern region of China. Stage 3 included 2,231 leprosy patients and 2,266 healthy control subjects from southern China. All patients and healthy control subjects with written informed consent were recruited as previously described (Zhang et al., 2009). All individuals were of Chinese descent, and the clinical information of samples is summarized in Supplementary Table S3.

This study was approved by the institutional review board committee at the Shandong Provincial Institute of Dermatology and Venereology, Shandong Academy of Medical Sciences.

Genotyping and quality control in the discovery stage

We carried out this study using Illumina Infinium Human Exome Bead Chip (version 1.0) array (Illumina, San Diego, CA). SNPs went through the following quality control filters: call rate greater than 99%, MAF in all samples greater than 0.1%, and Hardy-Weinberg equilibrium P-value in control subjects greater than 1.0×10^{-8}. We also excluded SNPs located in non-autosomal chromosomes and noncoding regions. A total of 40,491 genotyped SNPs were used for association analysis in the discovery stage.

The samples with call rate of less than 98%, one of related pairs (first-, second- or third-degree familial relationships) with lower call rate (seven), and population outliers based on the principal component analysis method (16 patients and two control subjects) were excluded. Overall, 1,648 patients and 2,318 control subjects passed the sample quality control filters and were used in the discovery analysis.

Statistical analysis

In the discovery stage, we tested associations between phenotypes and single-variant genotypes using GMMAT_v0.7 (Breslow and Clayton, 1993; Chen et al., 2016). In the validation and replication stages, we used a logistic regression model for those SNPs with MAF greater than 1% and Fisher exact test for those SNPs with MAF less than 1%. The meta-analysis was performed using a fixed-effect model. We declared a single variant-trait association significant if the nominal *P*-value was less than 1.23×10^{-6} (0.05/40491). Assessment of OR heterogeneity across independent samples was carried out by evaluating the *P*-values from Cochran *Q* statistics, and heterogeneity *P*-values less than 0.05 were considered significant.

Genotyping analysis and quality control in the validation and replication studies

SNP genotyping for the validation and replication stages were conducted at the Shandong Provincial Key Laboratory for Dermatovenereology using the Sequenom MassARRAY system, OpenArray custom genotyping assays on QuantStudio 12K Flex Real-Time PCR System (Applied Biosystems, Foster City, CA) and TaqMan custom genotyping assays on a 7900 HT Fast Real-Time PCR System (Applied Biosystems) according to the manufacturers' instructions. The following quality control filters were adopted: variants with undetermined clusters or variants with call rate less than 90%, variants with significant deviation from Hardy-Weinberg equilibrium in control samples (Hardy-Weinberg equilibrium, $P < 1 \times 10^{-3}$), and samples with call rate less than 95% were excluded.

Biological and network analysis

GO enrichment analysis. GO enrichment analysis was performed with seven identified genes set from the current study and 33-gene set, including seven genes from the current study and 26 reported genes from previous GWASs (Liu et al., 2012, 2013, 2015; Wang et al., 2016; Zhang et al., 2009, 2011), respectively.

GO enrichment was implemented by TopGO (Alexa et al., 2006), an R package, which calculates GO enrichment *P*-values for a given gene list. The 30 most significant GO terms are listed in the Supplementary Table S9.

Integrated network enrichment. GeneMANIA (Multiple Association Network Integration Algorithm) (Mostafavi et al., 2008) in Cytoscape was applied to the 33-gene set to carry out the integrated gene network analysis (see Supplementary Figure S12).

More detailed information on the study subjects, genotyping, quality control, variant selection, statistical analysis, and biological and network analyses is provided in the Supplementary Materials online.

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

We thank the individuals who participated in this project and Shandong Computer Science Center (National Supercomputer Center in Jinan), which provided us with the platform for statistical analysis. This work was funded by grants from the National Natural Science Foundation of China (81620108025, 81601387, 81602746, 31601040); the National Clinical Key Project of Dermatology and Venereology; the Shandong Provincial Advanced Taishan Scholar Construction Project; the Innovation Project of Shandong Academy of Medical Science; the Shandong Provincial Independent Innovation Project (ZR2015HZ001); the Shandong Province independent innovation and achievement transformation project (2014CGZH1307); the Natural Science Foundation of Shandong Province (BS2013YY010, 2014ZRC03128, ZR2014YL037, ZR2015YL025, JQ201616, ZR2016HB47); the Key research and development program of Shandong Province (2015GSF118126, 2016ZDJS07A06); the Science Development Funding Agency of Shandong Province, China (2010G0020258); a joint BII-GIS cat 3 platform grant from which VL has been partially paid (IAF311010); and the Agency of Science, Technology and Research (A*STAR) of Singapore.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at http://dx.doi.org/10.1016/j.jid.2017.08.004.

REFERENCES

Abraham S, Mozhi NM, Joseph GA, Kurian N, Rao PS, Job CK. Epidemiological significance of first skin lesion in leprosy. Int J Lepr Other Mycobact Dis 1998;66:131–9.

Alegre-Abarrategui J, Christian H, Lufino MM, Mutihac R, Venda LL, Ansorge O, et al. LRRK2 regulates autophagic activity and localizes to specific membrane microdomains in a novel human genomic reporter cellular model. Hum Mol Genet 2009;18:4022–34.

Alexa A, Rahnenfuhrer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006;22:1600–7.

Aschengrau A, Seage G. Essentials of epidemiology for public health. Sudbury, MA: Jones and Bartlett Publishers, Inc.; 2006.

Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 2008;40:695–701.

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993;88:9–25.

Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed Mmodels. Am J Hum Genet 2016;98:653–66.

De Benedetto A, Qualia CM, Baroody FM, Beck LA. Filaggrin expression in oral, nasal, and esophageal mucosa. J Invest Dermatol 2008;128:1594–7.

Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 2008;9:516.

Fearnhead NS, Winney B, Bodmer WF. Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. Cell Cycle 2005;4:521–5.

Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10:241–51.

Ghorpade A. Inoculation (tattoo) leprosy: a report of 31 cases. J Eur Acad Dermatol Venereol 2002;16:494–9.

Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet 2012;13:135–45.

Hirota Y, Tanaka Y. A small GTPase, human Rab32, is required for the formation of autophagic vacuoles under basal conditions. Cell Mol Life Sci 2009;66:2913–32.

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 2012a;91:224–37.

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012b;13:762–75.

Lim CS, Park ES, Kim DJ, Song YH, Eom SH, Chun JS, et al. SPIN90 (SH3 protein interacting with Nck, 90 kDa), an adaptor protein that is developmentally regulated during cardiac myocyte differentiation. J Biol Chem 2001;276:12871–8.

Liu H, Bao F, Irwanto A, Fu X, Lu N, Yu G, et al. An association study of TOLL and CARD with leprosy susceptibility in Chinese population. Hum Mol Genet 2013;22:4430–7.

Liu H, Irwanto A, Fu X, Yu G, Yu Y, Sun Y, et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. Nat Genet 2015;47:267–71.

Liu H, Irwanto A, Tian H, Fu X, Yu Y, Yu G, et al. Identification of IL18RAP/IL18R1 and IL12B as leprosy risk genes demonstrates shared pathogenesis between inflammation and infectious diseases. Am J Hum Genet 2012;91:935−41.

Misch EA, Berrington WR, Vary JC Jr, Hawn TR. Leprosy and the human genome. Microbiol Mol Biol Rev 2010;74:589−620.

Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol 2008;9(Suppl. 1):S4.

Oh H, Kim H, Chung KH, Hong NH, Shin B, Park WJ, et al. SPIN90 knockdown attenuates the formation and movement of endosomal vesicles in the early stages of epidermal growth factor receptor endocytosis. PLoS One 2013;8:e82610.

Plowey ED, Cherra SJ 3rd, Liu YJ, Chu CT. Role of autophagy in G2019S-LRRK2-associated neurite shortening in differentiated SH-SY5Y cells. J Neurochem 2008;105:1048−56.

Rani R, Fernandez-Vina MA, Zaheer SA, Beena KR, Stastny P. Study of HLA class II alleles by PCR oligotyping in leprosy patients from north India. Tissue Antigens 1993;42:133−7.

Sandilands A, Sutherland C, Irvine AD, McLean WH. Filaggrin in the frontline: role in skin barrier function and disease. J Cell Sci 2009;122:1285−94.

Satoh S, Tominaga T. mDia-interacting protein acts downstream of Rho-mDia and modifies Src activation and stress fiber formation. J Biol Chem 2001;276:39290−4.

Scott IR, Harding CR. Filaggrin breakdown to water binding compounds during development of the rat stratum corneum is controlled by the water activity of the environment. Dev Biol 1986;115:84−92.

Seto S, Tsujimura K, Koide Y. Rab GTPases regulating phagosome maturation are differentially recruited to mycobacterial phagosomes. Traffic 2011;12:407−20.

Spano S, Galan JE. A Rab32-dependent pathway contributes to Salmonella typhi host restriction. Science 2012;338:960−3.

Wang Z, Sun Y, Fu X, Yu G, Wang C, Bao F, et al. A large-scale genome-wide association and meta-analysis identified four novel susceptibility loci for leprosy. Nat Commun 2016;7:13760.

Waschbusch D, Michels H, Strassheim S, Ossendorf E, Kessler D, Gloeckner CJ, et al. LRRK2 transport is regulated by its novel interacting partner Rab32. PLoS One 2014;9:e111632.

WHO. Global leprosy update, 2015: time for action, accountability and inclusion. Wkly Epidemiol Rec 2016;91:405−20.

Wong SH, Gochhait S, Malhotra D, Pettersson FH, Teo YY, Khor CC, et al. Leprosy and the adaptation of human toll-like receptor 1. PLoS Pathog 2010;6:e1000979.

Zhang F, Liu H, Chen S, Low H, Sun L, Cui Y, et al. Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. Nat Genet 2011;43:1247−51.

Zhang FR, Huang W, Chen SM, Sun LD, Liu H, Li Y, et al. Genomewide association study of leprosy. N Engl J Med 2009;361:2609−18.