HLA class II sequence variants influence tuberculosis risk in populations of European ancestry

Gardar Sveinbjornsson1,2, Daniel F Gudbjartsson1,2, Bjarni V Halldorsson1,3, Karl G Kristinsson4,5, Magnus Gottfredsson4,6, Jeffrey C Barrett7, Larus J Gudmundsson1, Kai Blondal8, Arnaldur Gylfason1, Sigurjon Axel Gudjonsson1, Hafdis T Helgadottir1, Adalbjorg Jonasdottir1, Aslaug Jonasdottir1, Ari Karason1, Ljiljana Bulat Kardum9, Helgi Kristjansson1,4, Mar Kristjansson6, Arthur Love4,12, Yang Luo7, Olafur T Magnusson1, Patrick Sulem1, Augustine Kong1, Gisli Masson1, Unnur Thorsteinsdottir1,4, Zlatko Dembic10, Sergey Nejentsev13, Thorsteinn Blondal8, Ingileif Jonsdottir1,4,14 & Kari Stefansson1,4

Mycobacterium tuberculosis infections cause 9 million new tuberculosis cases and 1.5 million deaths annually. To identify variants conferring risk of tuberculosis, we tested 28.3 million variants identified through whole-genome sequencing of 2,636 Icelanders for association with tuberculosis (TB; 8,162 cases and 277,643 controls), pulmonary tuberculosis (PTB) and M. tuberculosis infection. We found association of three variants in the region harboring genes encoding the class II human leukocyte antigens (HLAs): rs557011[T] (minor allele frequency (MAF) = 40.2%), associated with M. tuberculosis infection (odds ratio (OR) = 1.14, P = 3.1 × 10^-13) and PTB (OR = 1.25, P = 5.8 × 10^-12), and rs9271378[G] (MAF = 32.5%), associated with PTB (OR = 0.78, P = 2.5 × 10^-12)—both located between HLA-DQA1 and HLA-DRB1—and a missense variant encoding p.Ala210Thr in HLA-DQA1 (MAF = 19.1%, rs9272785), associated with M. tuberculosis infection (P = 9.3 × 10^-9, OR = 1.14). We replicated association of these variants with PTB in samples of European ancestry from Russia and Croatia (P < 5.9 × 10^-4). These findings show that the HLA class II region contributes to genetic risk of tuberculosis, possibly through reduced presentation of protective M. tuberculosis antigens to T cells.

Mycobacterium tuberculosis causes 9.0 million new tuberculosis (TB) cases and 1.5 million deaths annually, and M. tuberculosis carriers who are not infected with HIV have a 10% lifetime risk of developing active TB disease1. The majority of infected individuals control the pathogen by mounting a successful, long-lived immune response, leading to clinically latent infection. Immunological impairment caused by malnutrition, diabetes, HIV-AIDS, aging and smoking has a major role in the epidemiology of TB, and heritability studies have implicated genetic susceptibility2,3. It is believed that TB was rare in Iceland until the nineteenth century, when it spread rapidly and reached its peak in the beginning of the 1930s. In 1935, approximately 20.9% of 8-year-olds and 34.2% of 13-year-olds had positive tuberculin skin tests (TST+), indicating that a substantial portion of the population had been exposed to M. tuberculosis4,5. Since then, the incidence of TB in Iceland has decreased to the lowest in Europe6.

We imputed 28.3 million SNPs and insertions-deletions identified through whole-genome sequencing of 2,636 Icelanders into 104,220 chip-type Icelanders and their first- and second-degree relatives7. The Icelandic Tuberculosis Data Registry contains TB diagnosis from 1900 to 2010, including confirmed TB diagnoses for 8,162 people with genotype information, of whom 3,686 had confirmed PTB (Table 1). An additional 6,562 individuals who did not develop TB were recorded as TST+ and had not been Bacillus Calmette-Guerin (BCG) vaccinated, making a total of 14,724 individuals who had contracted M. tuberculosis infection and for whom we have genotype information (Supplementary Table 1). Because TST positivity may also be caused by vaccination with BCG, we excluded those who had been BCG vaccinated and considered the remaining TST+ subjects as M. tuberculosis-infected. We tested the imputed sequence variants for association with PTB, TB and M. tuberculosis infection (and/or TST+) using population controls (n > 277,643). Sequence variants were weighed according to their prior probability of affecting gene function by applying thresholds for genome-wide significance that depend on the variant class. We allocated the type I error rate of 0.05 equally between three classes of variants. We tested 5,955

1deCODE genetics/Angen Inc., Reykjavik, Iceland. 2School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. 3School of Science and Engineering, Reykjavik University, Reykjavik, Iceland. 4Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. 5Department of Clinical Microbiology, Landspitali, the National University Hospital of Iceland, Reykjavik, Iceland. 6Department of Infectious Diseases, Landspitali, the National University Hospital of Iceland, Reykjavik, Iceland. 7Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. 8Division of Communicable Disease Prevention and Control, Primary Health Care of the Capital Area, Reykjavik, Iceland. 9Department of Pulmonology, Clinic of Internal Medicine, Clinical Hospital Center, University of Rijeka, Rijeka, Croatia. 10Laboratory of Molecular Genetics, Department of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway. 11Division of Molecular Medicine, Ružer Bošković Institute, Zagreb, Croatia. 12Department of Virology, Landspitali, the National University Hospital of Iceland, Reykjavik, Iceland. 13Department of Medicine, University of Cambridge, Cambridge, UK. 14Department of Immunology, Landspitali, the National University Hospital of Iceland, Reykjavik, Iceland. Correspondence should be addressed to K.S. (kstefans@decode.is) or I.J. (ingileif.jonsdottir@decode.is).

Received 6 May 2015; accepted 4 January 2016; published online 1 February 2016; doi:10.1038/ng.3498
loss-of-function variants, 157,106 missense variants and 28,113,695 other variants, yielding class-specific Holm-Bonferroni genome-wide significance thresholds of 2.8 × 10^(-6) for loss-of-function variants, 1.1 × 10^(-7) for missense variants and 5.9 × 10^(-10) for other variants. Using these criteria, several sequence variants in the HLA region on chromosome 6p21 showed genome-wide significant association (Table 2, Fig. 1, Supplementary Tables 2 and 3, Supplementary Fig. 1). The strongest association was between rs557011[T] (MAF = 40.2%), located between HLA-DQA1 and HLA-DRB1, and M. tuberculosis infection (P = 3.1 × 10^(-13), OR = 1.14) and PTB (P = 5.8 × 10^(-12), OR = 1.25). rs9271378[G] (MAF = 32.5%), also located between HLA-DQA1 and HLA-DRB1 (Fig. 1), associates with reduced risk of PTB (P = 2.5 × 10^(-12), OR = 0.78). Finally, the missense variant p.Ala210Thr (MAF = 19.1%, rs9272785) in exon 4 of HLA-DQA1 associates with M. tuberculosis infection (P = 9.3 × 10^(-9), OR = 1.14). p.Ala210Thr corresponds to the classical HLA-DQA1*03 superallele. Three missense variants—p.Thr49Ser (rs1048023), p.Gly79Arg (rs12722072 and rs12722074) and p.Met99Val (rs1064944)—that, along with p.Ala210Thr, define HLA-DQA1*03, did not align to the reference sequence. These variants were called by separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for separately aligning each exon to one of the haplotypes found in GATK HLA Caller and imputing for.

Table 1: Characteristics of the Icelandic tuberculosis cases

|                  | Pulmonary TB (n = 3,686) | Tuberculosis (n = 8,162) | M. tuberculosis infection (including TB, n = 14,723) |
|------------------|--------------------------|--------------------------|-----------------------------------------------|
|                  | Chip-typed               | First- and second-degree relatives | Chip-typed                                           | First- and second-degree relatives |
| Cases            | 1,188                    | 2,498                    | 2,765                                               | 5,397                                               |
| Year of birth (mean ± s.d.) | 1931 ± 12               | 1914 ± 16               | 1933 ± 13                                           | 1914 ± 17                                           |
| % alive (2015)   | 53.0                     | 8.4                      | 57.4                                                | 10.0                                                |
| % male           | 40.7                     | 48.0                     | 41.1                                                | 47.5                                                |

Table 2: Sequence variants showing genome-wide significant association with tuberculosis in Iceland

| SNP              | Chr. | Pos. | MAF (%) | Info* | Effect | Gene | Effect allele | P     | OR    | 95% CI  | P     | OR    | 95% CI  |
|------------------|------|------|---------|-------|--------|------|---------------|-------|-------|---------|-------|-------|---------|
| rs557011         | 6    | 362949991 | 40.2 | 0.997 | –      | –    | T             | 5.8 x 12-12 | 1.25  | 1.17-1.33 | 3.7 x 10-9 | 1.15  | 1.10-1.20 | 3.1 x 10-13 | 1.14  | 1.10-1.18 |
| rs9271378        | 6    | 36295278 | 32.5 | 0.998 | –      | –    | G             | 2.5 x 12-12 | 0.78  | 0.73-0.84 | 1.2 x 10-9 | 0.86  | 0.82-0.90 | 1.3 x 10-7 | 0.90  | 0.87-0.94 |
| rs9272785        | 6    | 362718379 | 19.1 | 0.996 | p.Ala210Thr | HLA-DQA1 | A             | 3.5 x 10-12 | 1.22  | 1.13-1.32 | 7.3 x 10-4 | 1.10  | 1.04-1.16 | 9.3 x 10-3 | 1.14  | 1.09-1.19 |

*Information for markers estimated by the ratio of the variance in imputed expected allele counts and the variance in the actual allele counts (Online Methods).
The heterogeneity is driven by the difference between the Icelandic and the Russian samples.

Our study suggests that the DQA1*03 allele may contribute to TB. DQA1*03 is a well-known risk factor for gluten-sensitive enteropathy celiac disease as part of the DQA1*03-DQB1*0302 haplotype, encoding DQ8, and trans-DQA1*03:01 and DQBI*02:01, encoding DQ2.3 (ref. 16). DQ8 and DQ2.3 bind gliadin peptide T cell epitopes, providing functional evidence for the contribution of DQA1*03 to the pathogenesis of the disease17,18. DQA1*03 also increases susceptibility to type 1 diabetes19 and other autoimmune diseases20,21. Many M. tuberculosis–derived epitopes are recognized by HLA-restricted CD4+ and CD8+ T cells in humans infected with M. tuberculosis22. In latent TB infection, the CD4+ T cells recognizing M. tuberculosis epitopes are confined to the CXC3R3CCR6+ type 1 helper T (T H1) cell subset23, and high reactivity is associated with recognition of a few discrete dominant antigenic regions24. The ability of containing a DQA1*03:01–encoded α-chain HLA molecules to present dominant epitopes of critical protective M. tuberculosis antigens is unknown. Cell-surface expression of HLADQ-alleles varies extensively, indicating an allelic hierarchy in the intrinsic stability of HLADQ-molecules25; DQ8, containing a DQA1*03:01–encoded α-chain, is among the most unstable DQ molecules. Alterations in disease-associated amino acids located outside of the peptide-binding groove regulate the stability of DQ molecules25. The p.Ala210Thr-encoding variant is in exon 4 and encodes the transmembrane region of the DQA1 chain, whereas the p.Thr49Ser, p.Gly79Arg and

Table 3 Associations of the sequence variants in samples from Russia and Croatia

|                | rs557011(T)          | rs9271378(G)         | p.Ala210Thr            |
|----------------|----------------------|----------------------|------------------------|
|                | Cases | Controls | P | OR | 95% CI | AF% | P | OR | 95% CI | AF% | P | OR | 95% CI | AF% |
| PTB Iceland    | 3,686 | 287,427  | 5.8 × 10−12 | 1.25 | 1.17–1.33 | 40.2 | 2.5 × 10−12 | 0.78 | 0.73–0.84 | 32.5 | 3.5 × 10−7 | 1.22 | 1.13–1.32 | 19.1 |
| PTB Russia     | 5,530 | 5,607    | 8.5 × 10−5  | 1.12 | 1.06–1.19 | 34.4 | 1.9 × 10−5  | 0.89 | 0.84–0.94 | 46.6 | 5.4 × 10−4 | 1.15 | 1.06–1.24 | 13.9 |
| PTB Croatia    | 438   | 1,009    | 0.0074     | 1.26 | 1.06–1.49 | 36.1 | 0.0526    | 0.854 | 0.72–1.00 | 57.54 | 0.66 | 1.06 | 0.82–1.37 | 10.9 |
| PTB Russia +   | 5,968 | 6,616    | 4.7 × 10−6  | 1.13 | 1.07–1.20 | 30.6 | 3.0 × 10−6  | 0.89 | 0.84–0.93 | 5.9 × 10−4 | 1.14 | 1.06–1.23 | –   |
| Croatia combined|       |          |           |     |     |     |           |     |     |     |     |     |     |     |     |
| PTB all combined| 9,654 | 294,043  | 2.0 × 10−15 | 1.18 | 1.13–1.23 | –   | 3.2 × 10−15 | 0.85 | 0.82–0.89 | –   | 1.9 × 10−9 | 1.18 | 1.12–1.25 | –   |

*P value, OR and AF are given for rs113031369, which correlates with rs9271378 (r2 = 0.8).

Significant heterogeneity was observed between the non-Icelandic and Icelandic samples for rs557011 (P = 0.016) and rs9271378 (P = 0.0024) but not for DQA1*03 (P = 0.21).
**Table 4** Association of the sequence variants with tuberculosis sub-phenotypes

| SNP      | Chr. | Pos   | MAF (%) | Coding Gene | Effect allele | PTB (n = 3,686) vs. TST+ (n = 6,562) | TB (n = 8,162) vs. TST+ (n = 6,562) | Nonpulmonary TB (n = 4,476) | TST+ (n = 6,562) |
|----------|------|-------|---------|-------------|---------------|-------------------------------------|-------------------------------|------------------------|-----------------|
| rs557011 | 6    | 32694991 | 40.2 | T            | 0.0035        | 1.12 1.04–1.21 0.6 1.02 0.95–1.10 0.05 1.06 1.00–1.12 2.5 × 10−6 1.11 1.06–1.16 |
| rs9271378| 6    | 32695278 | 32.5 | G            | 3.7 × 10−6 0.83 0.77–0.90 0.0355 0.91 0.85–0.97 0.028 0.94 0.89–0.99 0.2 0.97 0.93–1.02 |
| rs9272785| 6    | 32718379 | 19.1 | p.Ala DQA1   | 0.061         | 1.10 1.00–1.22 0.13 0.94 0.87–1.02 0.91 1.00 1.00–1.10 1.1 × 10−7 1.16 1.10–1.23 |

TST+, *M. tuberculosis*-infected without TB.

p.Met99Val missense variants are encoded from exon 2, in the vicinity of αβ-chain dimerization interfaces at each end of the peptide binding groove, and may affect peptide binding. Furthermore, the DQ8 molecule containing a DQA1*03:01-encoded α-chain interacts poorly with HLA DRM—which has a critical role in peptide loading during antigen presentation—resulting in reduced loading and DM editing of antigenic peptides. Surface expression of HLA class II peptide complexes on antigen-presenting cells is also regulated by ubiquitination, which affects their assembly, endocytosis, recycling and turnover. The noncoding variants rs557011 and rs9271378 do not overlap with known biologically relevant regions (Supplementary Note). However, a correlated marker, rs1846190 (r² = 0.81 with rs557011), is located in an enhancer site in CD4*CD25*IL-17+ T cells, and in a CTCF binding site that seems to regulate *HLA-DRB1* expression in lymphoblastoid cell lines. It is conceivable that DQA1*03 and the missense variants contribute to the risk of *M. tuberculosis* infection and TB disease through reduced stability of molecules containing DQA1*03 and poor presentation of critical *M. tuberculosis* antigens, resulting in poor activation of protective T cells. The effects of the TB-associated variants on mRNA of DQA1*03 and its surface expression in antigen-presenting cells have not been studied.

The HLA region has a key role in immune responses, is associated with infections and autoimmune diseases, and has been the focus of many tuberculosis candidate gene studies yielding conflicting results. We found sequence variants, located in HLA-DQA1 and between HLA-DQA1 and HLA-DRB1, that associate with TB. rs557011 was associated with PTB and seems to confer susceptibility to both *M. tuberculosis* infection and risk of development of TB disease. By contrast, rs9271378 does not associate with *M. tuberculosis* infection but protects against development of TB disease in *M. tuberculosis*-infected individuals. For all variants, the effects on PTB were the strongest.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

The authors thank the study participants and the staff at the Patient Recruitment Center and the deCODE genetics core facilities. We thank S. Balen (University of Rijeka) and M. Balija (Croatian Institute for Transfusion Medicine) for assistance in the collection of blood samples. S. Grle-Popovic for assistance in collection of blood samples of tuberculosis patients treated at the University Hospital Center, Zagreb, and J. Pavelic (Ruder Bošković Institute) for providing resources and advice. This work was supported by the US National Institute of Allergy and Infectious Diseases grant HHSN2662004000064C (deCODE Genetics, A. Kong, K.G.K., M.G., M.K.), UK Wellcome Trust grants 088838/Z/09/Z and 095198/Z/10/Z (S.N.), EU Framework Programme 7 Collaborative grant 201483 (University of Cambridge and S.N.), European Research Council Starting grant 260177 and the Royal Society grants UF0763346 and RG090638 (S.N.). S.N. is also supported by a Welcome Trust Senior Research Fellowship and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre.

**AUTHOR CONTRIBUTIONS**

G.S., D.F.G., B.V.H., A. Kong, U.T., B.T., I.J. and K.S. designed the study and interpreted the results. T.B., K.G.K., M.G., I.J., A.L., M.K. and K.B. coordinated and managed phenotype data ascertainment and Icelandic subject recruitment. S.N., J.C.B. and Y.L. coordinated, managed, genotyped and analyzed the Russian cohort sample set. I.B.K., I.Z. and D.Z. coordinated and managed the Croatian cohort samples and phenotypes, which were genotyped and analyzed by deCODE. G.S., H.T.H., G.M., S.A.G., O.T.M., U.T. and I.J. performed the sequencing, genotyping and expression analyses. G.S., D.F.G., B.V.H., S.A.G., A.G., Adalbjorg Jonasdottir, Aslaug Jonasdottir, A. Karason, H.K. and I.J. performed HLA typing and analysis of HLA data. G.S., D.F.G., B.V.H., A.G., S.A.G., P.S., A. Kong, G.M. and I.J. performed the statistical and bioinformatics analyses. G.S., D.F.G., S.N., I.J. and K.S. drafted the manuscript. All authors contributed to the final version of the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. World Health Organization. *WHO Global Tuberculosis Report 2014*, WHO/HTM/TB/2014.08 (World Health Organization, 2014).
2. Fox, G.J. & Menzies, D. Epidemiology of tuberculosis immunology. *Adv. Exp. Med. Biol.* 783, 1–32 (2013).
3. Comstock, G.W. Tuberculosis in twins: a re-analysis of the Prophit survey. *Am. Rev. Respir. Dis.* 117, 621–624 (1978).
4. Sigurdsson, S. Tuberculosis in Iceland. 1976. *Laeknabladid* 91, 69–102 (2005).
5. Sigurdsson, S. Um berklaveiki á Íslandi. *Laeknabladid* 62, 3–50 (1976).
6. Bothamley, G.H., Dhillon, S., M'gloin, G.B. & Lange, C. Active case finding of tuberculosis in Europe. *Tuberculosis Network European Trials Group (TBNET) survey. Eur. Respir. J.* 32, 1023–1030 (2008).
7. Gudbjartsson, D.F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* 47, 435–444 (2015).
8. DePristo, M.A. et al. A framework for variant discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498 (2011).
9. Cobat, A. et al. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. *J. Exp. Med.* 206, 2583–2591 (2009).
10. Cobat, A. et al. Tuberculin skin test reactivity is negatively affected by tight genomic control of chromosomal region 11p14-15 in settings with different tuberculosis endemities. *J. Infect. Dis.* 211, 317–321 (2015).
11. Emison, V. et al. Genetics of gene expression and its effect on disease. *Nature* 452, 423–428 (2008).
12. Curtis, J. et al. Susceptibility to tuberculosis is associated with variants in the *ASAP1* gene encoding a regulator of dendritic cell migration. *Nat. Genet.* 47, 523–527 (2015).
13. Thye, T. et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* 44, 257–259 (2012).
14. Thye, T. et al. Genome-wide association analyses identify a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* 42, 739–741 (2010).
15. Chimusa, E.R. et al. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum. Mol. Genet.* 23, 796–809 (2014).
16. Solid, L.M. Coeliac disease: dissecting a complex inflammatory disorder. *Nat. Rev. Immunol.* 2, 647–655 (2002).
17. Sollid, L.M., Qiao, S.W., Anderson, R.P., Gianfrani, C. & Koning, F. Nomenclature and listing of celiac disease relevant gluten T-cell epitopes restricted by HLA-DQ molecules. *Immunogenetics* **64**, 455–460 (2012).

18. Lundin, K.E.A., Scott, H., Fausa, O., Thorsby, E. & Sollid, L.M. T cells from the small intestinal mucosa of a DR4, DQ7/DR4, DQ8 celiac disease patient preferentially recognize gliadin when presented by DQ8. *Hum. Immunol.* **41**, 285–291 (1994).

19. Aly, T.A. *et al.* Extreme genetic risk for type 1A diabetes. *Proc. Natl. Acad. Sci. USA* **103**, 14074–14079 (2006).

20. Zanelli, E., Breedveld, F.C. & de Vries, R.R. HLA class II association with rheumatoid arthritis: facts and interpretations. *Hum. Immunol.* **61**, 1254–1261 (2000).

21. Rider, L.G. The heterogeneity of juvenile myositis. *Autoimmun. Rev.* **6**, 241–247 (2007).

22. Lindestam Arlehamn, C.S., Lewinsohn, D., Sette, A. & Lewinsohn, D. Antigens for CD4 and CD8 T cells in tuberculosis. *Cold Spring Harb. Perspect. Med.* **4**, a018465 (2014).

23. Lindestam Arlehamn, C.S. & Sette, A. Definition of CD4 immunosignatures associated with MTB. *Front. Immunol.* **5**, 124 (2014).

24. Arlehamn, C.S. *et al.* Dissecting mechanisms of immunodominance to the common tuberculosis antigens ESAT-6, CFP10, Rv2031c (hspX), Rv2654c (TB7.7), and Rv1038c (EsxJ). *J. Immunol.* **188**, 5020–5031 (2012).

25. Miyadera, H., Ohashi, J., Lemmark, Å., Kitamura, T. & Tokunaga, K. Cell-surface MHC density profiling reveals instability of autoimmunity-associated HLA. *J. Clin. Invest.* **125**, 275–291 (2015).

26. Busch, R. *et al.* On the perils of poor editing: regulation of peptide loading by HLA-DQ and H2-A molecules associated with celiac disease and type 1 diabetes. *Expert Rev. Mol. Med.* **14**, e15 (2012).

27. Tollefsen, S. *et al.* Structural and functional studies of trans-encoded HLA-DQ2.3(DQA1*03:01/DQB1*02:01) protein molecule. *J. Biol. Chem.* **287**, 13611–13619 (2012).

28. Yin, L., Maben, Z.J., Becerra, A. & Stern, L.J. Evaluating the role of HLA-DM in MHC class II-peptide association reactions. *J. Immunol.* **195**, 706–716 (2015).

29. Cho, K.J., Walseng, E., Ishido, S. & Roche, P.A. Ubiquitination by March-I prevents MHC class II recycling and promotes MHC class II turnover in antigen-presenting cells. *Proc. Natl. Acad. Sci. USA* **112**, 10449–10454 (2015).

30. Chadwick, L.H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317–324 (2012).

31. Sheffield, N.C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23**, 777–788 (2013).

32. Chapman, S.J. & Hill, A.V. Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* **13**, 175–188 (2012).

33. Fernando, M.M. *et al.* Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genet.* **4**, e1000024 (2008).

34. 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**, 1166–1172 (2006).

35. Meyer, C.G. & Thye, T. Host genetic studies in adult pulmonary tuberculosis. *Semin. Immunol.* **26**, 445–453 (2014).
ONLINE METHODS

Icelandic discovery population. Individuals who had been diagnosed with TB or infected with *M. tuberculosis* (TST+) without developing TB during the twentieth century, according to the Icelandic Tuberculosis Database (ITDBB), were invited to participate in the study. The ITDBB contains information on TB diagnosis, major and minor sites of infection based on *M. tuberculosis* culture and microscopic analysis, histology and Roentgen result. It has also records on number and duration of TB episodes and hospitalizations, symptoms and signs, treatment and outcome and family history, as well as nationwide results of TST testing and BCG (Supplementary Note). In this study we used genotype data for 3,686 patients with confirmed pulmonary TB, 8,162 with any TB and 14,723 with *M. tuberculosis* infection with or without developing TB, using TST positivity as a surrogate for *M. tuberculosis* infection (excluding BCG-vaccinated individuals) and ≥2777,643 controls (remaining chip-typed Icelanders and their relatives).

The study was approved by the Icelandic Data Protection Authority (no. 2004120649) and the National Bioethics Committee (no. VSN 04-172 VSNb2004120008-03-1). All participating subjects who donated blood provided informed consent. Personal identities of the participants and biological samples were encrypted by a third-party system approved and monitored by the Icelandic Data Protection Authority.

Russian TB sample set. The study in the Russian cohort was as described previously. Briefly, TB patients were diagnosed through information about TB contact, medical history and clinical symptoms, presence of acid-fast bacilli in sputum smear and symptoms characteristic of pulmonary TB on chest X-rays. For all patients, diagnosis has been confirmed by culture of *M. tuberculosis* from sputum. Patients with extrapulmonary TB and all HIV-positive subjects were excluded. Controls were healthy adult blood-bank donors with no history of TB. *M. tuberculosis* infection status of these controls was unknown.

Croatian TB sample set. DNA was isolated from blood samples of PTB patients (*n* = 244) and contact controls (*n* = 85) treated at the Section of Pulmonology, Department of Internal Medicine, Clinical Hospital Centre Rijeka, Rijeka and Hospital for Lung Diseases “Jordanoivc,” University Hospital Center “Zagreb,” and healthy blood donors (n = 924) collected at the Department of Transfusion Medicine in Rijeka and in Zagreb, Croatia, as previously described. DNA samples from the blood of an additional 194 PTB patients, treated in the Hospital for Lung Disease “Jordanoivc,” University Hospital Center “Zagreb,” Croatia, were isolated at deCODE, Reykjavik, Iceland. The age and gender frequency of the latter group of patients was not significantly different from the former cohort of cases. All study subjects provided oral and written informed consent. The Medical Research Council ethics committees at Zagreb and Rijeka approved the research.

Illumina chip genotyping. Icelandic chip-typed samples were assayed using the Illumina HumanHap300, HumanCNV370, HumanHap610, HumanHapLM, HumanHap660, Omni-1, Omni 2.5 or Omni Express bead chips at deCODE genetics. SNPs were excluded if they (i) had yield <95%, (ii) had MAF <1% in the population, (iii) showed significant deviation from Hardy-Weinberg equilibrium in the controls (P < 0.001) or (iv) produced an excessive inheritance error rate (>0.001), or (v) if there was substantial difference in allele frequency between chip types from just a single chip if that resolved all differences, but from all chips otherwise. All samples with a call rate <97% were excluded from the analysis. For the HumanHap series 304,937 SNPs were used for long-range phasing, whereas for the Omni series 564,196 SNPs were included. The final set of SNPs used for long-range phasing was composed of 707,525 SNPs.

Genotyping of the Russian PTB cases and controls was performed using Affymetrix Genome-Wide Human SNP Array 6.0, and SNPs across the genome were imputed as described previously. Association analysis of the HLA SNPs rs557011, rs9271378 and rs9272785 was performed using PLINK.

Whole-genome sequencing. Whole-genome sequencing was performed for 2,636 Icelanders, selected for various conditions. All individuals were sequenced at a depth of at least 10x (average sequencing depth = 22x).

Template DNA fragments were hybridized to the surface of flow cells (GA PE cluster kit (v2) or HiSeq PE cluster kits (v2.5 or v3)) and amplified to form clusters using the Illumina cBot. In brief, DNA (2.512 pM) was denatured, followed by hybridization to grafted adapters on the flow cell. Isothermal bridge amplification using Phusion polymerase was then followed by linearization of the bridged DNA, denaturation, blocking of 3' ends and hybridization of the sequencing primer. Sequencing-by-synthesis (SBS) was performed on Illumina GAIIx and/or HiSeq 2000 instruments. Paired-end libraries were sequenced at 2 × 101 (HiSeq) or 2 × 120 (GAIIx) cycles of incorporation and imaging using the appropriate TruSeq SBS kits. Each library or sample was initially run on a single GAIIx lane for QC validation followed by further sequencing under either GAIIx (four lanes) or HiSeq (one lane) with targeted raw cluster densities of 500–800 k/mm², depending on the version of the data imaging and analysis packages (SCS2.6.2-9/RTA1.6-1.9, HCS1.3.8-1.4.8/RTA1.10.36-1.12.4.2). Real-time analysis involved conversion of image data to base-calling in real-time.

Generation of whole-genome genotype data. SNP and indel calling from the whole-genome sequence data of the 2,230 Icelanders and generation of imputed genotypes has been described previously (Supplementary Note). Briefly, the genotypes identified were imputed into chip-genotyped and long-range phased Icelanders. Probabilities of genotypes were furthermore predicted for relatives of chip-typed individuals.

**HLA typing.** For each of the six genes, the most common alleles present in the Icelandic population were selected from the Allele Frequency Database using the most ethnically related populations: Norway and Ireland. The exonic sequences of the alleles were downloaded from the Broad Institute's HLA reference (http://www.broadinstitute.org/gatk/media/docs/HLAREFERENCE.zip).

We genotyped in silico a set of 2,615 whole-genome-sequenced (WGS) individuals. For each sequenced individual we selected reads that either (i) mapped to the public reference sequence using BWA or (ii) were unmapped by BWA and could be aligned to one of the haplotypes found in the GATK/HLA Caller database.

For every gene genotyped we align each read in this set to each exon separately. We consider a read to belong to a given haplotype if: (i) the read can be aligned to the exon from GATK with no mismatches or indels, allowing for the possibility that the read only partially overlaps the exon as long as the overlap is at least 40 bp and the overlap does not introduce a mismatch or an indel, or (ii) the mate of this read can be aligned to some sequences occurring within ±1,000 bp from the exon using BWA default parameters. If a read r aligns to an exon of allele A, we arbitrarily say that P(r|A) = 1 – PE, and if r does not align to A we set P(r|A) = PE. We arbitrarily choose PE = 0.001 and have P(r|A) = 0.9999 for reads r that align to the superallele A, and P(r|A) = 0.001 for reads r that do not align to the superallele A. If we let R be the set of reads and assume independence of reads and the two alleles carried by an individual, we can then compute

\[
P(R|A_1, A_2) = \prod_{r \in R} \left( \frac{1}{2} P(r|A_1) + \frac{1}{2} P(r|A_2) \right)
\]

The HLA alleles were imputed into the Icelandic sample set as previously described for genotypes.

To check the accuracy of the imputation, at least three individuals carrying each haplotype were HLA typed for the six genes using All-Set TM Gold SSP (Life Technologies, DQA1, DQB1, DRB1, HLA-A high-resolution typing; HLA-B and HLA-C low-resolution typing).

Accuracy between imputation and wet-lab genotyping was 90–99% and frequency weighted correlation was 95–99.6% (Supplementary Table 11: accuracy was 0.989 for DQA1).

Genotyping of single variants. Single SNP genotyping of rs9271378 and rs9272785 in the Croatian sample set was carried out by deCODE Genetics in Reykjavik, Iceland, applying the Centaurus (NanoGen) platform. Sanger sequencing of rs557011 in the Croatian sample set was performed by deCODE Genetics.
Association testing. Logistic regression was used to test for association between sequence variants and disease (TB), treating disease status as the response and genotype counts as covariates. Other available individual characteristics that correlate with disease status were also included in the model as nuisance variables. These characteristics were: sex, county of birth, current age or age at death (first and second order terms included), blood sample availability for the individual and an indicator function for the overlap of the lifetime of the individual with the time span of phenotype collection (Supplementary Note). Conditional analysis was performed by including the sequence variant being conditioned on as a covariate in the model under the null and the alternative in the generalized linear regression.

Principal component analysis. We calculated principal components using EIGENSTRAT software41 (Supplementary Fig. 2). In the Russian population, the first four principal components were included in the association analysis (Supplementary Table 12). The first 20 principal components only explain 0.6% of the variance of the Icelandic genotype data, and the first two components correlate strongly with the county of birth, already included in the association test42. Including the first five principal components in the association testing of the chip-typed Icelandics had a minimal effect (Supplementary Table 3), and principal components were not included in the final Icelandic association testing, which also included relatives of the individuals.

Genotype imputation information. The informativeness of genotype imputation was estimated by the ratio of the variance of imputed expected allele counts and the variance of the actual allele counts:

$$\frac{\text{Var}(E(\theta | \text{chip data}))}{\text{Var}(\theta)}$$

where $\theta$ is the allele count. $\text{Var}(E(\theta | \text{chip data}))$ was estimated by the observed variance of the imputed expected counts and $\text{Var}(\theta)$ was estimated by $p(1 - p)$, where $p$ is the allele frequency. Sequence variants with imputation information <0.8 were excluded from the analysis.

Gene and variant annotation. For the annotation of the exome data, coordinates of variants were converted between hg18 and hg19 using the liftOver tool from UCSC43. Variants in hg19 coordinates were annotated with information on gene and variant annotation from Ensembl release 70 using Variant Effect Predictor (VEP) 44 version 2.8. Only protein coding transcripts from RefSeq Release 56 were considered45. Individuals in both the Icelandic case and control groups are related, causing the $\chi^2$ test statistic to have a mean >1 and median >0.675. We estimated the inflation factor $\lambda$ on the basis of a subset of about 300,000 common variants, and the $P$ values were adjusted by dividing the corresponding $\chi^2$ values by this factor to adjust for both relatedness and potential population stratification46.

Thresholds for genome-wide significance. We weighed sequence variants according to their prior probability of affecting gene function by applying thresholds for genome-wide significance that depend on the variant class. The Bonferroni correction for multiple testing can be adjusted to account for prior importance of sequence variants. We performed a weighted Holm-Bonferroni correction based on giving equal weight to the classes of LoF, MSSNS, and other variants47. For example, sequence variants in the LoF class get weight 1/(3 × 6, 476). The sum of these weights over all the variants in the genome is 1, and the Bonferroni threshold for significance within a class containing $m$ sequence variants is $0.05/m$.

Expression analysis. Samples of RNA from human peripheral blood were hybridized to Agilent Technologies Human 25K microarrays as described previously36 (GEO: GSE7965). We quantified expression changes between two samples as the mean logarithm ($\log_{10}$) expression ratio (MLR) compared with a reference pool RNA sample. In comparing expression levels between groups of individuals with different genotypes, we denoted the expression level for each genotype as $10^{\text{average MLR}}$, where the MLR is averaged over individuals with the particular genotype. We determined s.e.m. and significance by regressing the MLR values against the number of risk alleles carried. We took into account the effects of age, sex and differential cell type count in blood as explanatory variables in the regression. $P$ values were adjusted for familial relatedness of the individuals by simulation.

Assessment for potential overlap with regulatory regions. To identify TB-associated variants that might have regulatory effects, we took the strongest noncoding signals (rs557011 and rs9271378) and identified all SNPs in LD at $r^2 > 0.8$ (excluding SNPs with low imputation information values). This added rs1846190 and rs508318 (both $r^2 > 0.81$ with rs557011). For each of the four variants identified, we searched for overlaps with known regulatory regions as follows: (i) first we used ENSEMBL to determine whether the variant had been assigned a regulatory region ENSR number then examined the ENCODE data and looked for any evidence of ChIP-seq transcription factor binding and DNAse hypersensitivity sites48; (ii) we looked for enhancer and promoter chromatin segmentation states using the 25-state HMM from the Roadmap consortium30; (iii) we then looked for correlations between DNAse hypersensitive sites and local gene expression using results described by Sheffield et al.31; (iv) we examined Siphon and GERP conservation scores49,50, and (v) we viewed Factorbook and HaploReg v3 data51,52 to search for potential changes in transcription factor binding motifs with ChIP-seq evidence of the cognate transcription factor.

Data access. We are reporting only on the association of variants in the HLA region with TB and TB subphenotypes (Tables 2 and 4), replication of these variants in other populations of European ancestry (Table 3), all SNPs in the HLA region showing genome-wide association with TB phenotypes (Supplementary Table 2) and six classical HLA genes (HLA-A, HLA-B, HLA-C, HLA-DOA1, HLA-DOB1 and HLA-DRB1) for association with PTB, TB and M. tuberculosis infection (Supplementary Table 7). We are not releasing association summary statistics at this time.

36. Etokebe, G.E. et al. Toll-like receptor 2 (P631H) mutant impairs memory internalization and is a dominant negative allele. Scand. J. Immunol. 71, 369–381 (2010).
37. Knezevic, J. et al. Heterozygous carriage of a dysfunctional Toll-like receptor 9 allele affects CpG oligonucleotide responses in B cells. J. Biol. Chem. 287, 24544–24553 (2012).
38. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).
39. Styrkarsdottir, U. et al. Nonsense mutation in the LGRA gene is associated with several human diseases and other traits. Nature 497, 517–520 (2013).
40. Kutyavin, I.V. et al. A novel endonuclease IV post-PCR genotyping system. Nucleic Acids Res. 34, e128 (2006).
41. Price, A.L. et al. Principal components analysiscorrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909 (2006).
42. Price, A.L. et al. The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genet. 5, e1000505 (2009).
43. Kent, W.J. et al. The human genome browser at UCSC. Genome Res. 12, 996–1006 (2002).
44. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26, 2069–2070 (2010).
45. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D130–D135 (2012).
46. Devlin, B. & Roeder, K. Genomic control for association studies. Biometrics 55, 997–1004 (1999).
47. Holm, S. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70 (1979).
48. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
49. Davidov, E.V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP. PLoS Comput. Biol. 6, e1001025 (2010).
50. Garber, M. et al. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 25, i54–i62 (2009).
51. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 22, 1798–1812 (2012).
52. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40, D930–D934 (2012).