Supplementary Table 1. Characteristics of the Icelandic TB cases and population controls with regard to TST and BCG vaccination.

	Pulmonary TB		**Tuberculosis**		***M.tb.* infection (including TB)**	
	Cases	**Controls**	**Cases**	**Controls**	**Cases**	**Controls**
N	3.686	287,427	8,162	284,205	14,724	277,643
TST positive	864	13,356	2,219	12,007	8,781	5,445
TSTnegative	83	6,748	162	6,675	162	6.675
BCG vaccination	46	12,574	90	12,544	90	12,544

Supplementary Table 2. All genome-wide significant sequence variants in the HLA region associating with Tuberculosis and *M.tb.* infection. For each sequence variant the reference SNP ID number (rs#), chromosome (Chr), hg18 position, minor allele frequency (MAF), effect allele, major allele (Amaj) are shown in addition to r^2 with top associating sequence variants. All rs numbers that have been assigned to a variant are given. Multiple rs numbers may have been assigned to a variant in different builds and later merged.

rs#	chr	Position	MAF %	Pvalue	OR	Effect allele	Amaj	r^2 rs9271378	r^2 rs557011
Pulmonary TB									
rs77761176,rs9271378	chr6	32695278	32.5	2.5E-12	0.78	G	A	-	0.32
rs112004002,rs557011,rs78854967	chr6	32694991	40.2	5.8E-12	1.25	T	C	0.32	-
rs114754672,rs33932178	chr6	32692291	40.1	1.7E-10	1.23	G	C	0.32	0.63
rs113548540,rs113924863,rs4959105	chr6	32691124	40.1	1.8E-10	1.23	T	C	0.32	0.63
rs113120809,rs33915496	chr6	32691957	40.1	1.8E-10	1.23	G	A	0.32	0.63
rs3104418	chr6	32688999	39.9	2.4E-10	1.23	A	G	0.32	0.63
rs113389138,rs3129748	chr6	32688993	39.9	2.1E-10	1.23	A	T	0.32	0.63
rs3104417	chr6	32689000	39.9	2.1E-10	1.23	A	T	0.32	0.63
rs113744419,rs3129747	chr6	32688986	39.9	2.1E-10	1.23	C	T	0.32	0.63
rs111272360,rs2395516	chr6	32688635	39.9	2.2E-10	1.23	C	T	0.32	0.63
rs112359949,rs79414789	chr6	32688569	39.9	2.2E-10	1.23	!G	G	0.22	0.37
rs34665982	chr6	32668284	48.9	2.2E-10	1.23	C	T	0.2	0.6
rs111875628,rs1846190,rs76590439	chr6	32691791	35.2	3.5E-10	1.23	A	G	0.26	0.81
rs147773060,rs508318	chr6	32692872	35.2	3.6E-10	1.23	!G	G	0.1	0.32
chr6:32320441:0:T	chr6	32320441	33.3	5.6E-10	0.80	T	!T	0.38	0.14
M.tb. infection									
rs9272785	chr6	32718379	19.1	9.3E-09	1.14	A	G	0.11	0.35
rs112004002,rs557011,rs78854967	chr6	32694991	40.2	3.1E-13	1.14	T	C	0.32	1
rs34665982	chr6	32668284	48.8	3.6E-11	1.13	C	T	0.2	0.6
rs147773060,rs508318	chr6	32692872	35.2	5.2E-11	1.13	!G	G	0.1	0.32
rs111875628,rs1846190,rs76590439	chr6	32691791	35.2	5.3E-11	1.13	A	G	0.26	0.8
rs113744419,rs3129747	chr6	32688986	39.9	5.8E-11	1.13	C	T	0.32	0.62
rs3104418	chr6	32688999	39.9	5.8E-11	1.13	A	G	0.32	0.62

rs3104417	chr6	32689000	39.9	5.9E-11	1.13	A	T	0.32	0.62
rs113389138,rs3129748	chr6	32688993	39.9	5.9E-11	1.13	A	T	0.32	0.62
rs112359949,rs79414789	chr6	32688569	39.9	6.2E-11	1.13	!G	G	0.22	0.37
rs111272360,rs2395516	chr6	32688635	39.9	6.2E-11	1.13	C	T	0.32	0.62
rs113120809,rs33915496	chr6	32691957	40.1	8.1E-11	1.13	G	A	0.32	0.63
rs113548540,rs113924863,rs4959105	chr6	32691124	40.1	8.2E-11	1.13	T	C	0.32	0.63
rs114754672,rs33932178	chr6	32692291	40.1	8.7E-11	1.12	G	C	0.32	0.63

Supplementary Table 3. Association analysis of the three variants among the chip typed individuals and the effect of correcting for principal components. Shown are the results when the first 5 principal components (PC) are used as covariates, demonstrating that correcting for principal components has little impact on the results.

	Chip typed						**Chip typed - PC corrected**					
	Pulmonary TB N=1,188		**Tuberculosis N=2,765**		***M.tb.* Infection N=6,105**		**Pulmonary TB N=1,188**		**Tuberculosis N=2,765**		***M.tb.* Infection N=6,105**	
rs#	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**	**P**	**OR**
rs557011	0.0014	1.16	0.00068	1.11	6.3E-6	1.11	0.0010	1.16	0.00064	1.11	1.2E-5	1.10
rs9271378	0.00010	0.83	0.00014	0.88	0.016	0.95	9.2E-05	0.83	0.00018	0.88	0.018	0.95
rs9272785	0.023	1.14	0.17	1.05	2.9E-5	1.12	0.024	1.13	0.19	1.05	7.5E-5	1.11

Supplementary Table 4. Pairwise r^2 between three genome-wide significant sequence variants.

rs#	**rs9271378**	**rs557011**	**rs9272785**
rs9271378	-	0.32	0.11
rs557011	0.32	-	0.35
rs9272785 (p.Ala210Thr)	0.11	0.35	-

Supplementary Table 5. Association of the sequence variants with PTB, TB and *M.tb.* infected conditioning pairwise on the other significant variants. For each sequence variant tested the reference SNP ID numbe r (rs#), chromosome (Chr), hg18 position, minor allele frequency (MAF), effect allele and the SNP ID number (rs#) for the variant conditioned on are shown in addition to P values and ORs.

						PTB		**TB**		***M.tb. infected***	
rs#	**Chr**	**Pos**	**MAF(%)**	**Effect allele**	**SNP conditioned on**	**Pvalue**	**OR**	**Pvalue**	**OR**	**Pvalue**	**OR**
rs9271378	chr6	32695278	32.5	G	rs557011	8.6E-04	0.86	0.0064	0.92	0.34	1.00
rs9272785/p.Ala210Thr	chr6	32718379	19.1	A	rs557011	0.35	1.05	0.75	0.99	0.14	1.04
rs557011	chr6	32694991	40.2	T	rs9271378	3.5E-04	1.15	0.0015	1.10	7.9E-08	1.13
rs9272785/p.Ala210Thr	chr6	32718379	19.1	A	rs9271378	0.0040	1.12	0.16	1.04	2.6E-05	1.10
rs557011	chr6	32694991	40.2	T	rs9272785/p.Ala210Thr	4.0E-06	1.21	4.0E-06	1.15	4.6E-06	1.11
rs9271378	chr6	32695278	32.5	G	rs9272785/p.Ala210Thr	5.1E-08	0.81	5.1E-07	0.87	8.8E-04	0.93

Supplementary Table 6. Association with microbiologically confirmed PTB cases and microbiologically confirmed TB cases.

				Microbiologically confirmed PTB (N=1,820)		Microbiologically confirmed TB (N=2,440)	
rs#	**chr**	**pos**	**Effect allele**	**P**	**OR**	**P**	**OR**
rs557011	chr6	32,694,991	T	7.2E-9	1.29	4.5E-8	1.24
rs9271378	chr6	32,695,278	G	**3.4E-13**	0.70	**7.3E-13**	0.74
rs9272785	chr6	32,718,379	A	**3.0E-8**	1.34	1.3E-5	1.23

Supplementary Table 8. Association of sequence variants conditioning on the most significant HLA allele. The reference SNP ID number (rs#), chromosome (Chr), hg18 position, minor allele frequency (MAF), coding effect are given. In addition the HLA allele which was used as covariate is given along with the P value and OR for each SNP after conditioning.

rs#	chr	pos	MAF(%)	coding	HLA allele conditioned on	Pvalue	OR
Pulmonary TB							
rs557011	chr6	32694991	40.2	-	*DQA1*03:01*	1.6×10^{-6}	1.16
rs9272785	chr6	32718379	19.1	p.Ala210Thr*	*DQA1*03:01*	0.21	1.07
rs9271378	chr6	32695278	32.5	-	*DQA1*03:01*	9.4×10^{-9}	0.68
Tuberculosis all							
rs557011	chr6	32694991	40.2	-	*DRB1*12:02*	3.9×10^{-8}	1.14
rs9272785	chr6	32718379	19.1	p.Ala210Thr*	*DRB1*12:02*	0.0018	1.09
rs9271378	chr6	32695278	32.5	-	*DRB1*12:02*	7.7×10^{-8}	0.87
M.tb. infected							
rs557011	chr6	32694991	40.2	-	*DQA1*03:01*	7.6×10^{-9}	1.11
rs9272785	chr6	32718379	19.1	p.Ala210Thr*	*DQA1*03:01*	9.7×10^{-4}	1.10
rs9271378	chr6	32695278	32.5	-	*DQA1*03:01*	2.4×10^{-5}	0.91

* p.Thr49Ser (rs1048023), p.Gly79Ser (rs12722072) and p.Met99Val (rs1064944) correlate (r^2=0.99) with rs9272785/p.Ala210Thr and show identical association.

Supplementary Table 9. Association of the TB variants with the expression of *HLA-DQA1*, *HLA-DQB1* and *HLA-DRB1* in white blood cells. The association for the strongest variant in the region is given in bold followed by the association of TB variants. The reference SNP ID number (rs#), chromosome (Chr), hg18 position, minor allele frequency (MAF), effect in standard deviations (Beta), P value, effect allele, major allele (Amaj) and info are given along with the gene name and r^2 with the strongest variant in the region.

rs#	Chrom	Pos	Beta	Pval	Probe	Effect allele	Amaj	MAF	Info	Gene	r^2 with top variant
	chr6	**32735901**	**-1.14237**	**6.91E-198**	**NM_002122**	**G**	**A**	**47.5843**	**0.98876**	HLA-DQA1	
rs9271378	chr6	32695278	-0.60185	1.63E-35	NM_002122	G	A	32.5228	0.99843	HLA-DQA1	0.072
rs9272785/p.Ala210Thr	chr6	32718379	-0.5345	2.09E-22	NM_002122	A	G	19.1178	0.9964	HLA-DQA1	0.28
	chr6	**32741988**	**1.073637**	**1.35E-145**	**NM_002123**	**G**	**A**	**39.9316**	**0.99547**	HLA-DQB1	
rs9271378	chr6	32695278	-0.88338	2.59E-83	NM_002123	G	A	32.5228	0.99843	HLA-DQB1	0.29
rs557011	chr6	32694991	0.38127	2.59E-16	NM_002123	T	C	40.2098	0.99722	HLA-DQB1	0.0012
	chr6	**32657117**	**1.598994**	**1.33E-120**	**NM_002124**	**T**	**C**	**10.0287**	**0.9787**	HLA-DRB1	
rs9271378	chr6	32695278	0.552141	7.85E-32	NM_002124	G	A	32.5228	0.99843	HLA-DRB1	0.23

Supplementary Table 10. Association results from our GWAS scan of sequence variants previously reported to associate with Tuberculosis. For each sequence variant the reference SNP ID number (rs#), chromosome (Chr), hg18 position, minor allele frequency (MAF) , effect allele, its coding effect on a gene, are shown in addition to the tuberculosis odds ratio and the corresponding P values.

								Pulmonary TB		**Tuberculosis all**		***M.tb.* infection**	
								N=3,686		**N=8,162**		**N=14,724**	
rs#	**chr**	**pos**	**MAF(%)**	**Reference**	**coding**	**gene**	**Effect allele**	**Pvalue**	**OR**	**Pvalue**	**OR**	**Pvalue**	**OR**
rs10956514	chr8	131321940	40.2	[1]	-	*ASAP1*	G	0.66	0.99	0.75	0.99	0.64	1.01
rs4733781	chr8	131365949	33.35	[1]	-	*ASAP1*	C	0.99	1.001	0.74	0.99	0.47	1.01
rs4331426	chr18	18444793	4.6	[2]	-	-	G	0.24	1.09	0.10	1.09	**0.016**	1.11
rs2057178	chr11	32320763	15.7	[3]	-	-	G	**0.039**	0.91	**0.0024**	0.91	**0.011**	0.94
rs9469220	chr6	32766288	38.8	[2]	-	-	A	**1.10E-05**	0.86	**0.0044**	0.93	**0.0016**	0.94
rs9272346	chr6	32712350	44.5	[2]	-	-	A	0.56	1.02	0.49	0.98	0.21	1.02

Supplementary Table 11. Associations of the sequence variants from follow up in samples from Russia before and after correcting for first 4 principal components.

	rs557011[T]		rs9271378[G]		p.Ala210Thr	
	P	**OR**	**P**	**OR**	**P**	**OR**
PTB Russia - PC corrected	8.5E-5	1.12	1.9E-5	0.89	0.00054	1.15
PTB Russia - not corrected	7.6E-5	1.12	2.9E-6	0.88	0.00021	1.16

Supplementary Table 12. Accuracy of the HLA imputations. At least three individuals carrying each haplotype were HLA typed for the 6 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%.

	Individuals	Haplotypes	Correct haps	Error haps	Het error	Homo error	Accuracy	Error rate	Freq w. R2	Freq w. R
DQA1	46	92	91	1	1	0	0.989	0.011	0.993	0.996
DQB1	142	284	273	11	7	2	0.961	0.039	0.966	0.983
DRB1	352	704	635	69	59	5	0.902	0.098	0.917	0.957
HLAA	222	444	425	19	17	1	0.957	0.043	0.960	0.979
HLAB	1310	2620	2427	193	149	22	0.926	0.074	0.922	0.959
HLAC	1276	2552	2346	206	142	32	0.919	0.081	0.903	0.950

Supplementary notes

The study populations.

The Icelandic discovery cohort: The Icelandic Tuberculosis Database (ITBDB) is a computerized database containing almost complete information on tuberculosis disease (TB) in Iceland during the 20th century. Information on TB entered into the ITBDB was retrieved from journals and health records from all TB 24 hospitals/health institutions upto 1942, from the National TB registry at the TB and Lung Clinic of the Reykjavik Health Care Center (RHCC) for the period 1942-2005[4,5] and from Landspitali, the National University Hospital of Iceland, Dept. of Bacteriology (*M.tb.* culture, microscopic analysis and drug sensitivity), Dept. of Pathology (histology) and Dept. of Radiology (X-rays) for the period 1924-2005. The ITBDB contains information on personal and family history of TB, diagnosis of TB, major and minor sites of disease, start, end and year of each episode and hospitalization, height, weight, symtoms and signs at admission and discharge, treatment and outcome, as well as familiy history of TB. A total of 11,438 subjects were diagonsed with TB, whereof 3471 were bacteriologically confirmed. A total of 4,951 had confirmed PTB. Of the non-pulmonary TB 2,524 had involvement of the intrathoracic hilar lymph node (LN), 533 other LNs, 445 bone and joints and 159 the central nervous system (CNS), 52 developed disseminated TB and 702 died during TB episode.

Tuberculin skin test (TST) was systematically performed on family members of TB patients in Iceland and also on all 7 and 12 year old school children. The ITBDB also contains information from RHCC on time, dose and outcome of TST for 25,987 subjects during 1935-2007 (with approx. 30% of 7 years old and 50% of the 12 years old responding TST positive). Subjects who were TST positive (used as a surrogate for having been infected with *M.tb.*, since BCG vaccinated subjects are excluded) who never developed TB were 7773 and *M.tb.* infected in total (TB disease and/or TST positive with out developing TB) were 19,211; TST tested individuals who were never TST postitive were 5828. Information on time and outcome of BCG (Bacillus Calmette-Guerin) vaccination of 15,602 subjects during 1925 to 2006 was also retrieved from RHCC and entered into the ITBDB.

Genotypes are available for 3,686 patients with PTB, 8,162 with any TB and 14,723 with *M.tb.* infection with or without TB.

Case control association testing. Logistic regression was used to test for association between SNPs and disease, 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 timespan of phenotype collection.

Given genotype counts for n individuals, $g_1, g_2, \dots, g_n \in \{0, 1, 2\}$, their phenotypes $y_1, y_2, \dots, y_n \in \{0, 1\}$ and a list of vectors of nuisance parameters $x_1, x_2, \dots, x_n$, the logistic regression model states that

$$L_i(\alpha, \beta, \gamma) = P(y_i = 1 | g_i, x_i)$$

$$logit\big(P(y_i = 1 | g_i, x_i)\big) = \alpha + \beta g_i + \gamma^T x_i, \text{for all } i \in \{1, 2, \dots, n\},$$

where α, β and γ are the regression coefficients and L_i is the contribution of the ith indivual to the likelihood function; $L(\alpha, \beta, \gamma) = \prod_{i=1}^{n} L_i(\alpha, \beta, \gamma)$. It is then possible to test for association based on the asymptotic assumption that the likelihood ratio statistic follows a χ^2 distribution with one degree of freedom:

$$2 \log \left(\frac{\max\limits_{\alpha,\beta,\gamma} L(\alpha, \beta, \gamma)}{\max\limits_{\alpha,\gamma} L(\alpha, \beta = 0, \gamma)} \right) \sim \chi_1^2.$$

Maximizing over the nuisance parameters at every marker in the genome would be prohibitively expensive. We therefore choose to maximize the likelihood under the null hypothesis of $\beta = 0$, which is the same for all markers, and use the maximizer of γ, $\tilde{\gamma}$, under the alternative. Since $\max_{\alpha,\beta,\gamma} L(\alpha, \beta, \gamma) \geq \max_{\alpha,\beta} L(\alpha, \beta, \gamma = \tilde{\gamma})$, this will lead to a smaller likelihood ratio than if we would maximize over γ for every marker.

Our analysis is based on imputed genotype values where the values of g_i are not known. Instead we use $P(g_i = j | I_i)$ for $j \in \{0, 1, 2\}$, where I_i stands for the information about g_i. Given the logistic regression model above, this allows us to calculate

$$P(y_i = 1|I_i) = \sum_{j=0}^{2} P(g_i = j|I_i)\, P(y_i = 1|g_i)\,, \text{for all } i \in \{1, 2, \ldots, n\}.$$

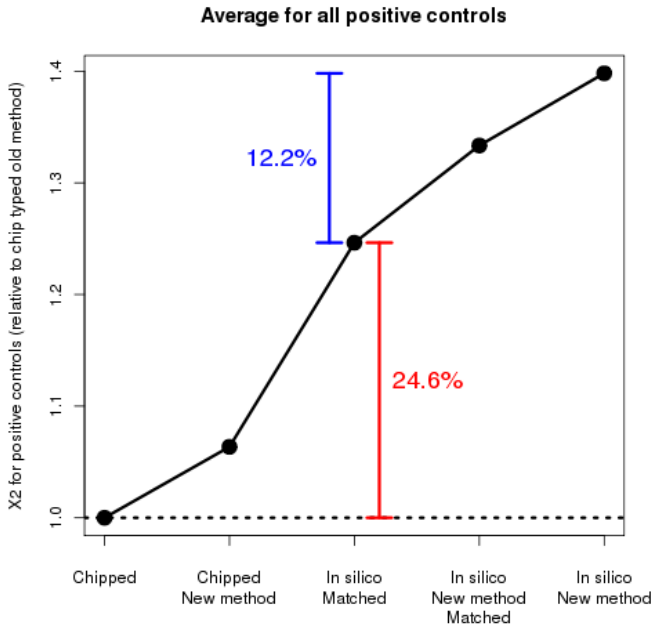
We note that this approach differs from the common approximation of substituting g_i with $Eg_i|I_i$, the expectation of g_i given I_i, in the logistic regression equation above. This approach has a straightforward mathematical justification and requires fewer assumptions than the approximate method and seems to be more robust to very uninformative imputations, such as the *in silico*, imputations we use. This approach requires a special implementation of logistic regression and is slightly more computationally taxing.

We refer to the approach of not including covariates and substituting expected counts as the *old method* and the approach of using the covariates and integrating over the genotype uncertainty as the *new method*. We evaluated the impact of applying these changes to tests for association on nine phenotypes at loci previously shown to associate with these diseases:

Disease	**N *in silico***	**N chipped**	**N ctrls**	**First YOD**	**Last YOB**	**Sex**	**N loci**
Alzheimers disease	945	2,555	158,336	1971	1965	M/F	6
Asthma	2,131	6,117	266,405	1990	2002	M/F	11
Atrial Fibrillation	3,485	5,182	266,093	1983	1999	M/F	10
Atrial Fibrillation under 60	553	820	272,132	1984	1999	M/F	10
Breast Cancer	1,971	3,189	289,890	1955	1998	M/F	12
Hypertrophic Cardiomyopathy	46	107	219,310	1988	1986	M/F	1
Myocardial Infarction	12,700	6,345	230,274	1968	1988	M/F	7
Ovarian Cancer	526	428	135,391	1955	1993	F	6
Prostate Cancer	2,102	2,595	85,848	1957	1967	M	39

Diseases used to evaluate association methods through association at known loci (Catalog of Published Genome-Wide Association Studies)[6].

On average the mean χ^2 statistic for the new method using all controls and both chip typed and *in silico* genotypes was 12% higher than for the old method using matched controls; the statistic for the new method was 40% higher than using just the chip typed individuals and the old method for association.



Performance of different tests of association evaluated based on the average χ^2 statistics for SNPs known to associate with nine different diseases (see table above). All test statistics were adjusted based on genomic control. The five tests for association evaluated are: The commonly used logistic regression method based on using expected allele counts as a covariate (Chipped), the new method of integrating over the uncertain allele count and using other characteristics of individuals as covariates (Chipped New method), the old method applied to all individuals with genotype information (chip typed individuals and *in silico* genotyped individuals) using matched controls (In silico Matched), the new method applied all individuals with genotype information using matched controls (In silico New method Matched) and the new method applied to all individuals with genotype information using all available controls (In silico New method). The performance of each test is shown relative to the old implementation of logistic regression applied to only the chip typed individuals.

Whole-genome SNP and INDEL calling. Multi-sample calling was performed with GATK version 2.3.9 using all the 2,636 BAM files together.

Genotype calls made solely on the basis of next generation sequence data yield errors at a rate that decreases as a function of sequencing depth. Thus, for example, if sequence reads at a heterozygous SNP position carry one copy of the alternative allele and seven copies of the reference allele, then without further information the genotype would be called homozygous for the reference allele. To minimize the number of such errors, we used information about haplotype sharing, taking advantage of the fact that all the sequenced individuals had also been chip-typed and long range phased[7]. Extending the previous example, if the individual shares a haplotype with another who is heterozygous given his sequence reads, then the ambiguous individual would be called as heterozygous. Conversely, if the individual shares both his haplotypes with others who are homozygous for the major allele his genotype would be called homozygous. In order to improve genotype quality and to phase the sequencing genotypes, an iterative algorithm based on the IMPUTE HMM model [8] which uses the LRP haplotypes was employed. Assume a SNP with alleles 0 and 1 is being phased. We let *H* be the long range phased haplotypes of the sequenced individuals and applied the following hidden Markov model (HMM) based algorithm.

Assuming that at each marker *i* the haplotype *h* has a common ancestor with a haplotype in $H\backslash\{h\}$ and denote the variable indicating this with the latent variable $z_i \in H\backslash\{h\}$, the hidden variable in the HMM. Then

$$\gamma_{h,k,i} = P(z_i = k | all\ LRP\ markers),$$

for all $k \in H\backslash\{h\}$. Given a haplotype *h* in *H*, $\gamma_{h,k}$ are calculated simultaneously for all $k \in H\backslash\{h\}$ using the same HMM model as IMPUTE [8]. Given the Markov assumptions of the HMM, the model is fully specified by emission and transition probabilities.

We define the emission probabilities of the HMM at each marker *i* as:

$$P(z_i = k | marker\ i) = \begin{cases} 1-\lambda, \text{if } h \text{ and } k \text{ match at } i \\ \lambda, \text{if } h \text{ and } k \text{ mismatch at } i \end{cases}$$

where λ can be thought of as a penalty for a mismatch. We used $\lambda = 10^{-7}$ in our implementation. We define the transmission probabilities of the HMM model as:

$$P(z_i | z_{i-1}, markers\ 1, \dots, i-1) = \begin{cases} e^{-\frac{\rho_i}{N}} + \frac{1-e^{-\frac{\rho_i}{N}}}{N}, \text{if } z_i = z_{i-1} \\ \frac{1-e^{-\frac{\rho_i}{N}}}{N}, \text{if } z_i \neq z_{i-1} \end{cases}$$

Where N is the number of haplotypes in $k \in H\backslash\{h\}$, which for autosomal chromosomes is $2(2{,}636-1)$ here and $\rho_i = 4N_e r_i$, where r_i is the genetic distance between markers $i-1$ and i according to the most recent version of the deCODE genetic map [9] and N_e was originally meant to be an estimate of the effective number of haplotypes in the population that our sample comes from, we used $N_e = 7{,}000$. These definitions fully specify the probability distribution $P(z_i | all\ markers)$. Calculating $\gamma_{h,k}$ for a single haplotype requires $O(MN)$ operations, where *N* is the number of haplotypes and *M* is the number of markers. Since these calculations can be performed for one haplotype at a time, the calculations can be parallelized across a computer cluster for efficiency. In practice most of the $\gamma_{h,k}$ will be close to zero and can be safely ignored (we used a threshold of 10^{-6} of the largest value at each marker for each *h*) greatly reducing storage requirements.

Now we are set to describe an iterative algorithm for the actual phasing. For every *h* in *H*, initialize the parameter θ_h, which specifies how likely the one allele of the SNP is to occur on the background of *h* from the genotype likelihoods obtained from sequencing. The genotype likelihood L_g is the probability of the observed sequencing data at the SNP for a given

individual assuming g is the true genotype at the SNP. If L_0, L_1 and L_2 are the likelihoods of the genotypes 0, 1 and 2 in the individual that carries h, then set θ_h:

$$\theta_h = \frac{L_2+\frac{1}{2}L_1}{L_2+L_1+L_0}.$$

For every pair of haplotypes h and k in H that are carried by the same individual, use the other haplotypes in H to predict the genotype of the SNP on the backgrounds of h and k:

$$\tau_h = \sum_{l\in H\setminus\{h\}} \gamma_{h,l}\theta_l \text{ and}$$

$$\tau_k = \sum_{l\in H\setminus\{k\}} \gamma_{k,l}\theta_l.$$

Combining these predictions with the genotype likelihoods from sequencing gives un-normalized updated phased genotype probabilities:

$$P_{00} = (1-\tau_h)(1-\tau_k)L_0,$$

$$P_{10} = \tau_h(1-\tau_k)\tfrac{1}{2}L_1,$$

$$P_{01} = (1-\tau_h)\tau_k\tfrac{1}{2}L_1,$$

$$\text{and } P_{11} = \tau_h\tau_k L_2.$$

Now use these values to update θ_h and θ_k to:

$$\theta_h = \frac{P_{10}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}} \text{ and}$$

$$\theta_k = \frac{P_{01}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}.$$

Iterate until the maximum difference between iterations is less than a convergence threshold ε. We used $\varepsilon=10^{-7}$.

Genotype imputation. Given the long range phased haplotypes of the sequenced individuals, H, and θ, the carrier probabilities of the haplotypes in the sequenced set, the probability that a new haplotype n, not in the set of sequenced haplotypes H, is imputed as $\sum_{l\in H}\gamma_{n,l}\theta_l$, where $\gamma_{n,l}$ is calculated as above for every $l \in H$.

In silico (genealogy-based) genotyping:

In addition to imputing sequence variants from the whole-genome sequencing effort into chip-typed individuals, we also performed a second imputation step where genotypes were imputed into relatives of chip genotyped individuals, creating *in silico* genotypes. The inputs into the second imputation step are the fully phased (in particular every allele has been assigned a parent of origin) imputed and chip-type genotypes of the available chip-typed individuals. The algorithm used to perform the second imputation step consists of:

1. For each ungenotyped individual (the proband), find all chip-typed individuals within two meiosis of the individual. The six possible types of two meiosis relatives of the proband are (ignoring more complicated relationships due to pedigree loops): Parents, full and half siblings, grandparents, children and grandchildren. If all pedigree paths from the proband to a genotyped relative go through other genotyped relatives, then that relative is excluded. e.g. if a parent of the proband is genotyped, then the proband's grandparents through that parent are excluded. If the number of meiosis in the pedigree around the proband exceeds a threshold (we used 12), then relatives are removed from the pedigree until the number of meiosis falls below 12, in order to reduce computational complexity.
2. At every point in the genome, calculate the probability for each genotyped relative sharing with the proband based on the autosomal SNPs used for phasing. A multipoint algorithm based on the hidden Markov model Lander-Green multipoint linkage algorithm using fast Fourier transforms is used to calculate these sharing probabilities [10,11]. First single point sharing probabilities are calculated by dividing the genome into 0.5cM bins and using the haplotypes over these bins as alleles. Haplotypes that are the same, except at most at a single SNP, are treated as identical. When the haplotypes in the pedigree are incompatible over a bin, then a uniform probability distribution was

used for that bin. The most common causes for such incompatibilities are recombinations within the pedigree, phasing errors and genotyping errors. Note that since the input genotypes are fully phased, the single point information is substantially more informative than for unphased genotyped, in particular one haplotype of the parent of a genotyped child is always known. The single point distributions are then convolved using the multipoint algorithm to obtain multipoint sharing probabilities at the center of each bin. Genetic distances were obtained from the most recent version of the deCODE genetic map[9].

3. Based on the sharing probabilities at the center of each bin, all the SNPs from the whole-genome sequencing are imputed into the proband. To impute the genotype of the paternal allele of a SNP located at x, flanked by bins with centers at x_{left} and x_{right}. Starting with the left bin, going through all possible sharing patterns v, let I_v be the set of haplotypes of genotyped individuals that share identically by descent within the pedigree with the proband's paternal haplotype given the sharing pattern v and $P(v)$ be the probability of v at the left bin – this is the output from step 2 above – and let e_i be the expected allele count of the SNP for haplotype i. Then $e_v = \frac{\sum_{i \in I_v} e_i}{\sum_{i \in I_v} 1}$ is the expected allele count of the paternal haplotype of the proband given v and an overall estimate of the allele count given the sharing distribution at the left bin is obtained from $e_{left} = \sum_v P(v) e_v$. If I_v is empty then no relative shares with the proband's paternal haplotype given v and thus there is no information about the allele count. We therefore store the probability that some genotyped relative shared the proband's paternal haplotype, $O_{left} = \sum_{v, I_v = \emptyset} P(V)$ and an expected allele count, conditional on the proband's paternal haplotype being shared by at least one genotyped relative: $c_{left} = \frac{\sum_{v, I_{v \neq \emptyset}} P(v) e_v}{\sum_{v, I_{v \neq \emptyset}} P(v)}$. In the same way calculate O_{right} and c_{right}. Linear interpolation is

then used to get an estimates at the SNP from the two flanking bins:

$$O = O_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(O_{right} - O_{left}\right),$$

$$c = c_{left} + \frac{x - x_{left}}{x_{right} - x_{left}}\left(c_{right} - c_{left}\right).$$

If θ is an estimate of the population frequency of the SNP then $Oc + (1 - O)\theta$ is an estimate of the allele count for the proband's paternal haplotype. Similarly, an expected allele count can be obtained for the proband's maternal haplotype.

References:

1. Curtis, J. *et al.* Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat Genet* **advance online publication** (2015).
2. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet* **42**, 739-41 (2010).
3. Thye, T. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat Genet* **44**, 257-9 (2012).
4. Sigurdsson, S. [Tuberculosis in Iceland. 1976]. *Laeknabladid* **91**, 69-102 (2005).
5. Sigurdsson, S. Um berklaveiki á Íslandi. *Læknablaðið* **62**, 3-50 (1976).
6. Hindorff LA, M.J.E.B.I., Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Vol. 2013.
7. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
8. 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).
9. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
10. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* **84**, 2363-7 (1987).
11. Kruglyak, L. & Lander, E.S. Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol* **5**, 1-7 (1998).